Improving global and catchment estimates of runoff through computationally-intelligent ensemble approaches

Applications of intelligent multi-model combination, cross-scale model comparisons, ensemble analyses, and new model parameterisations

Jamal Zaherpour BSc, MSc

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

March 2018

This thesis is dedicated to
my beloved wife Atefeh
and my parents

Acknowledgment

Primarily, thanks go to my PhD supervisors Dr Simon N Gosling and Dr Nick Mount, who have consistently provided me with their great supports of time, encouragement, knowledge and informative advice, which I know will stay with me in future endeavours. Thanks, especially for introducing me to two new worlds; global hydrology by Simon and hydro-informatics by Nick. Thanks also go to Colin Thorne for comments of my PhD at the end of first and second years. I truly appreciate all of the time that each of you have invested in my academic progression and I very much hope that our collaboration continue in the future.

My PhD was sponsored mainly by Islamic Development Bank (IDB) and partly by University of Nottingham. They provided full support for this research to be completed and to enhance both my personal academic progression and facilitating communication of the research by attending at conferences, workshops and courses. My endless thanks therefore go to them.

A primary source of data in this research has come from the Inter-sectoral Impact Model Inter-comparison Project (ISIMIP) part of the Potsdam Institute for Climate Impact Research (PIK, Potsdam, Germany) who provided simulated data for global and catchment hydrological models. I offer my deepest thanks to the PIK and modellers participated in ISIMIP, phase II for the creation and provision of their world-class dataset. Special thanks must also go to the Global Runoff Data Centre (GRDC, Koblenz, Germany) who allowed me to use their observed runoff datasets and maps free of charge under Academic License. Thanks also go to Petra Döll and Bernhard Lehner for providing the DDM30 (30' global drainage direction map) database. Specific people or data acknowledgements for other datasets, if appropriate, are presented at the end of each paper.

I would also thank the School of Geography at the University of Nottingham for the provision of additional funding. Mainly, though, I thank the people therein, for their many day-to-day contributions that have made for such a fulfilling experience over the past few years.





Table of Contents

ACKNOWLEDGMENT ..... **I**

TABLE OF CONTENTS ..... **II**

LIST OF FIGURES IV
LIST OF TABLES IV
LIST OF ABBREVIATIONS V

1 EXTENDED THESIS ABSTRACT **1**

1.1 INTRODUCTION 1
1.2 RESEARCH AIM, OBJECTIVES AND RATIONALE 5
1.3 THESIS STRUCTURE AND RATIONALE 9
1.3.1 Rationale *9*
1.3.1.1 Paper I 9
1.3.1.2 Paper II 11
1.3.1.3 Paper III 14
1.3.1.4 Paper IV 15
1.3.1.5 Paper V 17

2 METHODS **19**

2.1 GLOBAL HYDROLOGY AND MODEL INTERCOMPARISON PROJECTS (MIPS) 19
2.1.1 Global hydrological models (GHMs); emergence and applications *19*
2.1.2 GHMs classifications *20*
2.1.3 Model intercomparison projects (MIPs) *22*
2.2 MULTI MODEL COMBINATION (MMC) 23
2.2.1 MMC approaches *24*
2.2.2 Intelligent MMC *26*
2.2.3 Symbolic regression (SR) *28*
2.2.4 Gene Expression Programming (GEP) *28*
2.2.4.1 The structure and workflow of a GEP 29
2.3 MODEL PERFORMANCE EVALUATION 32
2.3.1 Quantitative model evaluation; graphical methods *33*
2.3.2 Quantitative model evaluation; numerical methods *35*
2.3.2.1 Single evaluation metrics 35
2.3.2.2 Integrated evaluation metrics 35
2.3.3 Qualitative model evaluation *36*
2.3.4 Evaluation measures applied in this research *36*
2.4 DATA AND SCALE OF THE STUDY 39
2.4.1 ISIMIP data *39*
2.4.2 Study catchments and data used for Papers I and II *40*
2.4.3 Study catchments and data used for Papers III and IV *40*
2.4.1 Study catchments and data used for Paper V *43*

3 SYNTHESIS OF THE RESULTS AND FUTURE DIRECTIONS **46**

3.1 RELIABILITY OF THE EM 46
3.2 MMC; APPLYING MODEL ENSEMBLES MORE INTELLIGENTLY THAN THE EM 49
3.3 WHICH SCALE; GLOBAL OR CATCHMENT? 53
3.4 APPLICATION OF MMC IN FILLING THE GAP BETWEEN GHMS AND CHMS 55
3.5 SHOULD WE USE GHMS WITH HUMAN IMPACT PARAMETRISATIONS (VARSOC)? 59
3.6 CONCLUSION 62
3.7 CURRENT STATUS OF THE PAPERS, AND STATEMENT ON AUTHORSHIP 64

4 REFERENCES FOR THE EXTENDED ABSTRACT 67

PAPER I: WORLDWIDE EVALUATION OF MEAN AND EXTREME RUNOFF FROM SIX GLOBAL-SCALE HYDROLOGICAL MODELS THAT ACCOUNT FOR HUMAN IMPACTS 78

PAPER II: EXPLORING THE VALUE OF MACHINE LEARNING FOR WEIGHTED MULTI-MODEL COMBINATION OF AN ENSEMBLE OF GLOBAL HYDROLOGICAL MODELS..... 79

PAPER III: A COMPARISON OF CHANGES IN RIVER RUNOFF FROM MULTIPLE GLOBAL AND CATCHMENT-SCALE HYDROLOGICAL MODELS UNDER GLOBAL WARMING SCENARIOS OF 1°C, 2°C AND 3°C 80

PAPER IV: MULTI-MODEL COMBINATION WITH A SUPER-ENSEMBLE OF CATCHMENT-SCALE AND GLOBAL-SCALE HYDROLOGICAL MODELS 81

PAPER V: HUMAN IMPACT PARAMETERIZATION IN GLOBAL HYDROLOGICAL MODELS IMPROVES ESTIMATES OF MONTHLY DISCHARGES AND HYDROLOGICAL EXTREMES: A MULTI-MODEL VALIDATION STUDY 82

Appendix A (Supplementary information for Paper I) 83
Appendix B (Supplementary information for Paper II)..... 84
Appendix C (Supplementary information for Paper III)..... 85
Appendix D (Supplementary information for Paper IV) 86
Appendix D (Supplementary information for Paper V) 87

List of figures

Figure	Caption	Page
2.1	Model combination approaches based on the how models are combined. From, Freitas and Rodrigues (2006).	25
2.2	Model combination classification based on model types.	25
2.3	(a) Expression tree for the mathematical expression $\sqrt{(a+b)*(c-d)}$, and (b) the expression tree for two single-gene chromosomes combined by addition.	30
2.4	The GEP algorithm flowchart.	31
2.5	IPE in the domain of evaluation methods.	33
2.6	Locations of the 40 selected catchments used in Papers II and III across the hydrobelt system defined in Meybeck et al. (2013). Shapefiles of the selected catchments from the global runoff data centre (GRDC), Germany, and shapefiles of the hydrobelts freely provided by Meybeck et al., 2013. Hydrobelts name: BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML=southern mid-latitude, SDR=southern dry and SST=southern subtropical.	42
2.7	Location of the study catchments and gauging stations considered in Papers III and IV. Shapefiles of the catchments courtesy of Simon Gosling, University of Nottingham, UK. Projection: EQDCYLIN Equidistant Cylindrical.	44
2.8	Spatial distribution of GRDC stations used for analysis in Paper V. Red dots (n = 471) indicate all GRDC stations that meet the selection criteria, whereas the black dots (n = 92) refer to the stations that are located at the outlet of a catchment. Shapefiles of the stations courtesy of Ted I.E. Veldkamp, VU University Amsterdam, the Netherlands. Projection: EQDCYLIN Equidistant Cylindrical.	45
3.1	Future research suggested based on the study conducted in Paper I.	49
3.2	Future research suggested based on the study conducted in Paper II.	52
3.3	Future research suggested based on the study conducted in Paper IV.	57
3.4	Future research suggested based on the study conducted in Paper V.	62
3.5	The status of the five research papers in the submission process.	65

List of tables

Table	Caption	Page
1.1	Research aim, research questions (RQs) and objectives (Obj) along with the related paper where they are addressed.	7
2.1	Model evaluation measures applied in this research.	38
3.1	Key findings and summary of results obtained for Paper I.	48
3.2	Key findings and summary of results obtained for Paper II.	51
3.3	Key findings and summary of results obtained for Paper III.	54
3.4	Key findings and summary of results obtained for Paper IV.	58
3.5	Key findings and summary of results obtained for Paper V.	61
3.6	Submission information and the status of the five research papers.	66

List of abbreviations

Abbreviations	Full term	Page of first appearance
GHM	Global hydrological models	1
CHM	Catchment hydrological models	1
MIP	Model inter-comparison projects	2
ISIMIP	Inter-sectoral impact model inter-comparison project	2
IPCC	International panel on climate change	2
EM	Ensemble mean	3
MMC	Multi-model combination	3
GEP	Gene expression programing	3
HIP	human impact parameterisations	5
SR	Symbolic regression	6
IPE	Ideal point error	6
KGE	Kling-Gupta efficiency index	6
MAR	Mean annual runoff	7
NOSOC	ISIMIP simulations with naturalised human impact	8
VARSOC	ISIMIP simulations with time-dependent human impact	8
MLA	Machine-learning algorithm	12
EA	Evolutionary algorithm	12
ANN	Artificial neural networks	12
UNFCCC	United Nations Framework Convention on Climate Change	15
LSM	Land surface model	20
DVM	Dynamic vegetation model	20
MHM	Macroscale hydrological model	20
GHWM	Global hydrological and water resources model	20
LESMs	Land earth system models	21
GA	Genetic algorithm	27
GP	Genetic programming	27
ET	expression trees	29
ORF	open reading frame	29
FDC	Flow duration curve	34
GRDC	Global runoff data centre	40
GCM	Global climate models	41

1 Extended thesis abstract

1.1 Introduction

Water related problems (scarcity, availability and hazards) together form one of the three major crises (the two other are food and energy) for today and the future across the globe (World Economic Forum, 2016, Schewe et al., 2014a, Hanasaki et al., 2013, Rockstrom et al., 2009). Water crises are widespread and heterogeneous around the world and climate change and socioeconomic drivers are expected to accelerate these problems (Veldkamp, 2017). To deal with the above concerns, mitigation and adaptation strategies are developed at different scales (global, regional and local). Developing these strategies, as well as selecting the most appropriate one to the problem of interest, should ideally benefit from the highest possible accuracy in estimates of the hydrological cycle and water resources. More reliable decisions can in turn be made by applying tools and techniques that enhance decision makers' perception of the hydrological cycle, particularly extreme events i.e. droughts and floods. These tools should also facilitate insights into the cycle: ideally by reference to hydrological indicators, spatially (globally and locally) and temporally (present day and future). Global and catchment scale hydrological models (GHMs and CHMs) have been used as such tools that along with advances in data acquisition, analytical techniques and computation power offer powerful tools for modelling natural processes and provide useful insights into the hydrological cycle.

GHMs have a shorter history of emergence and application than CHMs. GHMs have been developed and applied from 1986 in recognition of the fact that hydrological processes and water resources are global phenomena and should be treated at global scale (Bierkens, 2015). A GHM is a pragmatic trade-off between a faithful representation of the diversity of hydrological processes found across the world's catchments, and a generalised and simplified representation of hydrological processes that can support multi-decadal, generalised hydrological simulations at global scales. Compared to hydrological models designed for catchment-scale simulations (Arnold et al., 1993; Krysanova et al., 1998; Lindstrom et al., 2010), GHMs employ coarser spatial discretisation and model the global land surface in a single instantiation.

The global scope of GHMs, limited availability and quality of observed discharge data across the global domain and their use of spatially generalised parameters make them more difficult to calibrate than catchment hydrological models. Whilst examples of calibrated GHMs do exist (Müller Schmied et al., 2016), the majority of GHMs are uncalibrated (Gosling et al., 2016; Hattermann et al., 2017). This lack of calibration, coupled with the diversity of simplifications employed in the hydrological process representations, means that there can be large inconsistency in the skill, bias and uncertainty of an individual GHM at different locations, as

well as large inconsistencies between different GHMs at any given location (van Huijgevoort et al., 2013). This spatial inconsistency means that GHMs risk becoming a "jungle of models" (Kundzewicz, 1986) in which it can be difficult to determine where a particular GHM output is likely to be capable of delivering optimal hydrological simulations. It also makes it dangerous to assume that any individual GHM will be an adequate basis for making projections at any given location, even if the model's ability to replicate observed data in particular catchments is enhanced through the acquisition of higher quality input data or efforts to improve process representations (Liu et al., 2007). To an extent, these arguments are also applicable to CHMs because whilst they have been shown to generally perform better than GHMs in model evaluation studies, ensembles of such models still result in an uncertainty range when the models are run with identical inputs (Hattermann et al., 2017; Hattermann et al., 2018).

To minimise the challenge of varying outputs from different models, several model inter-comparison projects (MIPs) have been undertaken around the world (Henderson-Sellers et al., 1995, Entin et al., 1999, Guo and Dirmeyer, 2006, Koster et al., 2006, Harding et al., 2011). These projects usually use standard modelling baselines to deal with discrepancies between model outputs. This results in higher consistency in the climate forcings input to the models (where applicable), their process representations (e.g. the simulation of human impacts such as water abstractions), and the temporal and spatial resolutions of their simulations. This way, model outputs are directly comparable to each other, which supports diagnostic inter-comparisons between them (Bierkens, 2015). One of the largest, ongoing MIPs (whose data are used in this thesis) is the Inter-Sectoral Impact Model Inter-comparison Project (ISIMIP) (Schellnhuber et al., 2014, Warszawski et al., 2014). ISIMIP is a community-driven effort by more than 130 modelling groups, that covers different sectors including water (both global and catchment hydrological modelling communities). Outputs from ISIMIP are widely used in different projects, such as reports of the International Panel on Climate Change (IPCC) (http://www.ipcc.ch/).

MIPs including ISIMIP provide a unique opportunity to access data from different models and to assess their relative performance. It also facilitates continuous model improvement via the inclusion of new schemes (e.g. human impacts such as dams, reservoirs and water abstractions) accompanied by dozens of models, as well as communication between modelling groups working in the same or different sectors. Nonetheless, they do not fully address the challenge of spatial inconsistencies between models, as well as the question of what ensemble representative to select for use when trying to improve the reliability of decision-making. There remain other shortcomings or unexplored aspects within MIPs (particularly ISIMIP as this research's focus MIP), hence areas of further research and potential improvement in model evaluation and application which will be addressed later in this introduction.

The question of how to address the challenges of spatial inconsistency in hydrological models has been a feature of catchment-scale model research for several decades. In answering it, catchment modellers have recognised that reliance on a single, inconsistent model is inherently risky and should be avoided (Marshall et al., 2006; Shamseldin et al., 1997). Instead, they have developed ways to take advantage of the diversity of outputs (Clemen, 1989) generated by different models by using optimised mathematical combination methods to deliver a combined output that performs better than the individual models from which it was created (Hagedorn et al., 2005). This general approach—known as multi-model combination (MMC)—has been an important focus of catchment hydrological modelling studies, especially over the last two decades (Abrahart and See, 2002; Ajami et al., 2006; Arsenault et al., 2015; Azmi et al., 2010; de Menezes et al., 2000; Fernando et al., 2012; Jeong and Kim, 2009; Marshall et al., 2007; Marshall et al., 2006; Moges et al., 2016; Nasseri et al., 2014; Sanderson and Knutti, 2012; Shamseldin et al., 1997). Given its demonstrable potential in catchment studies, it is perhaps surprising that the potential of applying MMC to GHMs has yet to be explored.

A wide range of techniques can be used to generate an MMC solution. The simplest example is the calculation of the arithmetic mean of the input models (commonly referred to as an Ensemble Mean (EM)). More sophisticated techniques employ weighted schemes (Arsenault et al., 2015), with the differential weightings applied to each input model reflecting their relative strengths or weaknesses. The mathematical approach taken to determining the weights depends on the objective of the MMC. Where the primary objective is to minimise the difference between the MMC solution and observed data (i.e. maximise the predictive performance), without explicitly accounting for model or parameter uncertainty, the use of multiple linear regression (Doblas-Reyes et al., 2005) or machine learning algorithms to 'learn' the optimal set weights to apply to each MMC input model is a popular approach (Marshall et al., 2007). The use of algorithms such as artificial neural networks (ANNs) (Shamseldin et al., 1997; Xiong et al., 2001) or gene expression programming (GEP) (Barbulescu and Bautu, 2010; Bărbulescu and Băutu, 2009; Fernando et al., 2012) to define non-linear weighting schemes have proven to be particularly effective. This is down to their ability to generate optimised, non-linear schemes rapidly, without the need for any prior knowledge of the model parameters.

Where there is a desire to account for and minimise model and parameter uncertainty in the weighting scheme, Bayesian averaging methods are required (Ajami et al., 2007; Hoeting et al., 1999). These optimise the weights according to the posterior performance of the MMC solution under the prior probabilities of model parameter values (Duan et al., 2007; Vrugt and Robinson, 2007; Ye et al., 2004). However, these methods require knowledge of the probability density functions (PDFs) for each of the MMC's input model parameters (or at least their maximum

likelihood estimates (Ye et al., 2004)). This makes their use in the MMC of GHMs problematic because the number of parameters used in GHMs is particularly high, the parameters vary considerably between models, and the PDFs of the parameters in a GHM can be extremely difficult to specify over a global domain. Consequently, the PDFs for GHM parameters are seldom specified and, in many cases, remain unknown.

On this basis, this research applies the first MMC approach to tackle the challenge of inconsistency between multitude of models and using them optimally. To this end, itexplores the potential of MMC for addressing the challenge of spatial inconsistency in simulations from several GHMs (for the first time) and CHMs, by combining outputs from diverse sets of GHMs and CHMs using the machine learning tool GEP (Ferreira, 2001; Ferreira, 2006) across up to forty major catchments (Papers II and IV of the thesis). In each catchment, the MMC's ability to replicate the observed monthly runoff is compared against that of the EM, each of the individual models from which the MMC is derived, and the best-performing individual model from the ensemble.

In addition, this thesis addresses four other areas of further research spotted within the framework of the ISIMIP as follow with more detailed rationale behind performing each part of the research and their contribution to knowledge presented in Section 1.3. It should, however, be emphasised that these explorations are applicable to any MIPs and not limited to the ISIMIP.

First, despite the development of modelling protocols as common modelling baselines in different phases of the ISIMIP, its framework lacks an evaluation protocol. This is important towards model appraisal and, in turn, model application and improvement in a comparative manner. This research tackles this problem by setting a more comprehensive evaluation framework to an ensemble of GHMs and their ensemble mean (EM) that informs, via a multi-dimensional assessment, which models are best, where, and according to which hydrological indicator.

Second, as discussed above both global and catchment models provide valuable information for decision-making and scientific understanding, but while GHMs provide output data for almost all parts of the world (applicable for aggregated assessment of e.g. climate change impacts), the applicability of outputs generated by a CHM is usually limited to the catchment for which the model has been calibrated. Given the huge resources required for developing each model type, it is of great importance to understand their relative performance and whether they can be used interchangeably both for current and future climates. Nevertheless, knowledge of the relative spreads in simulations/projections from GHM and CHM ensembles (within and outside ISIMIP) is limited to cross-scale inter-comparison of small ensembles, precluding a robust ensemble

comparison. This research uses, for the first time, large multi-model ensembles of GHMs and CHMs to explore whether there are systematic differences between projections of runoff change from two ensembles comprising different types of hydrological model. It also investigates the effect of different degrees of global-mean warming on runoff for each catchment as indicated by each model type to the end of 21st century.

Third, while cross-scale inter-comparison of large GHM and CHM ensembles is more informative than their small ensembles, application of more robust representatives of each model ensemble (other than EM) was found an area of further study as the EM showed to not be necessarily the best ensemble representative. Given the potential found in the intelligent MMC approach that provided runoff simulations overall better than the EM, its capacity in filling the gap between two large ensembles of GHMs and CHMs is sought. In addition, this study examines the performance of MMC when a group of GHMs and a group of CHMs are pooled together as a 'super-ensemble'. The performance of MMC from the super-ensemble is then compared with the performance of the EM of the super-ensemble and MMCs from GHM and CHM ensembles to unravel potential benefits of the approach.

Fourth and finally, GHM outputs available and applied at the time of this research were from two different phases of ISIMIP i.e. Fast Track and ISIMIP2a of which in the latter GHMs were run under two different conditions of with and without the inclusion of anthropogenic impacts in the GHMs' parametrisation. Human impact parameterisations (HIP) has been an important step forwards to make modelling conditions as close as possible to the real world. Nonetheless, no comprehensive evaluation and comparison of GHMs under the two abovementioned conditions had been carried out, precluding test of models' capacity to represent human activities and their impacts on assessment of freshwater resources and hydrological extremes. The final part of this research (Paper V) explores change in the performance of five GHMs that participated in ISIMIP2a and had human impacts in their parameterisation.

1.2 Research aim, objectives and rationale

This section discusses the aim and objectives of the research followed by the structure of the thesis, more detailed rationale behind performing each part of the research and its contribution to knowledge presented in Section 1.3.

Table 1.1 summarises the research aim, research questions (RQs) and objectives (Objs) and references to the criterion (criteria) that each objective addresses towards fulfilling the degree qualification. Those criteria are as follows:

C1. Creation and interpretation of new knowledge, through original research or other advanced scholarship, of a quality to satisfy peer review, extend the forefront of the discipline.

C2. Systematic acquisition and understanding of a substantial body of knowledge at the forefront of an academic discipline or area of professional practice.

C3. Generation of new knowledge, applications or understanding at the forefront of the discipline.

C4. Detailed understanding of applicable techniques for research and advanced academic enquiry.

In order to fulfil the main research aim of improving estimation and presentation of runoff simulations from ensembles of GHMs and CHMs, the thesis centrally focuses on exploring the potential of an intelligent multi-model combination (MMC) technique and establishing opportunities it provides for improving the utility of runoff simulations from GHMs and CHMs in supporting reliable decision-making in the contexts of current and future water resource management. This technique combines outputs of several models into an output, which is optimised in terms of both the selected input models, and weights applied to each input model. For this purpose, a machine-learning (ML) based tool that applies intelligent algorithms (gene expression programing (GEP)) has been employed. The technique is capable of both enhancing the model ensemble simulations and providing diagnostic insights into the MMC models performance given the application of symbolic regression (SR) which uses explicit equations. This technique also provides opportunities for diagnostically improving models, which is applicable for future modelling applications and climate change forecasting. Prior to this and for an enhanced understanding from the GHMs performance, a comprehensive evaluation of six GHMs is performed applying a multi-dimensional assessment approach.

In addition, the thesis investigates the potential of cross-scale model comparisons to explore relative merits of estimations obtained from the two model scales both for historical hindcasting and for future projections. Further, it investigates whether and to what extent incorporation of anthropogenic impact parameterisations (e.g. dam operations and water withdrawals) in GHMs improves their performance. Better spatial presentation and comparison of the results worldwide is sought by the application of two integrated evaluation metric — *ideal point error, IPE (Dawson et al., 2012)* and Kling-Gupta efficiency (KGE) (Gupta et al., 2009; Kling et al., 2012)— and a recent hydro-climatic zoning system of hy*drobelts* (Meybeck et al., 2013), which is based on more hydro-geographical factors than its previous counterparts such as Koppen system.

Table 1.1 Research aim, research questions (RQs) and objectives (Obj) along with the related paper where they are addressed.

Thesis Main Aim: **Improving global and catchment estimates of runoff through computationally-intelligent ensemble approaches** Applications of intelligent multi-model combination, cross-scale model comparisons, ensemble analyses, and new parameterisations				
Research Questions (RQs)	**Objectives**	**Research sub-Questions (sRQs)**	**Paper**	**Criteria addressed**
RQ1: What are the characteristic strengths and limitations of the 'state-of-the-art' global-scale, hydrological modelling?	**Obj1**: Comprehensive evaluation of mean and extreme runoff from six global-scale hydrological models that account for human-impacts	1) How do GHMS perform in replicating time series of observed monthly runoff? 2) How well do GHMs do in simulating mean annual runoff (MAR) and extreme events (Q5, Q95 and flows of different return periods)? 3) How do models perform in simulating seasonal cycle (both magnitude and timing)? 4) How does the performance of GHMs vary spatially, given the application of a standardised integrated evaluation metric and a more hydro-geographically based climate zoning system? 5) What insight does this study provide concerning the performance of the ensemble mean (EM)? 6) What model improvement insights (in terms of structure, parametrisation, and calibration) does this evaluation provide?	I	C1 and C2
RQ2: Can optimised combinations of multiple GHM or CHM outputs generate improved simulation outputs and provide a mechanism for delivering heuristic insights into the strengths and weaknesses of different models?	**Obj2**: Optimising runoff simulations from an ensemble of global-scale hydrological models through multi-model combination and weighting	1) How do GHMs and EM perform, and what performance gain do MMC solutions deliver over the individual GHMs and EM? 2) What is the spatial pattern of the GHMs, EM and MMC solutions performance across the globe? 3) What insights does the intelligent MMC technique offer into the success or failure of the technique and combination mechanism based on the evaluation criteria set?	II	C3 and C4

Table 1.1 (cont.). Research aim, research questions (RQs) and objectives (Obj) along with the related paper where they are addressed.

Thesis Main Aim: **Improving global and catchment estimates of runoff through computationally-intelligent ensemble approaches** Applications of intelligent multi-model combination, cross-scale model comparisons, ensemble analyses, and new parameterisations				
Research Questions (RQs)	**Objectives**	**Research sub-Questions (RQs)**	**Paper**	**Criteria addressed**
RQ2: Can optimised combinations of multiple GHM or CHM outputs generate improved simulations and provide a mechanism for delivering heuristic insights into the strengths and weaknesses of the models?	**Obj3:** Cross-scale multi-model combination in hydrological modelling	1) Do solutions provided by MMCs of CHMs and GHMs, perform better than their respective EMs? 2) Does MMC improve the performance of one type of ensemble (i.e. comprised of CHMs or GHMs) more than the other type? 3) To what extent does the complexity of the MMC, i.e. the complexity of the weighting and combination of models, affect its performance? 4) Does an MMC developed from a "super-ensemble" of GHMs and CHMs outperform MMCs derived from an ensemble of each model type respectively?	IV	C3 and C4
RQ3: What are the relative performance gains for GHMs and CHMs (both in ensemble and MMC forms)?	**Obj4**: A comparison of changes in river runoff from multiple global and catchment scale hydrological models under global warming scenarios of 1°C, 2°C and 3°C	1) Are there systematic differences between projections of runoff change from the two ensembles in simulating indicators of mean and extreme events (MAR, Q5, Q95, and runoff with different return periods)? 2) What is the effect of different amounts of global-mean warming (1°C, 2°C and 3°C above pre-industrial levels) on GHM and CHM predicted, spatially distributed runoff in different catchments? 3) Could any particular type of model ensemble be preferred over another?	III	C1 and C2
RQ4: Does the inclusion of human impact parameterisations (HIP) such as dams, reservoirs and water abstractions in GHMs improve models ensemble simulations?	**Obj5**: Assessing the impact of inclusion of human activities on GHM performance: a comparison of modelled discharges under a naturalised (NOSOC) and a time-dependent human impact (VARSOC) simulation using five GHMs.	1) How does the incorporation of human impact affect representation of monthly discharges and their hydrological extremes, especially drought events? 2) How does the inclusion of human impact affect simulations of hydrological indicators in (near-)natural and managed catchments? 3) What is the spatial pattern of the changes caused by inclusion of human impact parameterisations? 4) What model improvement opportunities does this study offer?	V	C1 and C2

1.3 Thesis structure and rationale

The thesis is organised into this, introductory chapter, which explains the rationale behind carrying out the research, and seven other chapters. Chapter 2 comprises brief descriptions of the MMC method and the model evaluation methods employed, as well as detailing the sources and types of data used in the research. Chapter 3 presents a synthesis of the results and future research directions. It also details the current status (i.e. at the time of thesis submission) of the five papers that comprise the remaining chapters. The remaining chapters (4-8) are the published/submitted papers that form the body of the doctoral thesis.

1.3.1 Rationale

This section provides a detailed explanation of the rationale for conducting each part of the research. To this end, the problem addressed in each paper, is mapped into a four-step sequence consisting of the: *challenge*, *previous approaches (if any), research approach and contribution to knowledge*.

1.3.1.1 Paper I

Challenge. A comprehensive performance evaluation of the most state-of-the-art GHMs (i.e. those with the inclusion of human impact parameterisations (HIP)) and their ensemble mean (EM) was missing from the research domain. Evaluation of model performance, particularly when new features such as human impacts are added to models, provides an important context for decision-makers when assessing the level of confidence that can be placed in the outputs of different models, and their application in different locations across the globe. For the modelling community, such assessments are vital for indicating how and where future model research and development might deliver the greatest potential benefits to reliable decision-making.

Previous approaches. In practice, and in the absence of a standard GHM evaluation protocol that could formulate a multi-dimensional model assessment, the evaluation approaches taken in previous global model evaluation studies (within or outwith the ISIMIP framework) are highly varied. Variation in previous studies exists in the number of models evaluated (2-14, with a median of 6), the number of catchments included (8-6,192), the size of catchments considered (10 km^2 - 4,758,000 km^2), the evaluation metrics calculated, the hydrological indicators evaluated, and the period of analyses (Oki et al., 1999, Milly et al., 2005, Decharme and Douville, 2006, Nohara et al.,

2006, Decharme and Douville, 2007, Materia et al., 2010, Haddeland et al., 2011, van Dijk et al., 2013, Beck et al., 2016, Hattermann et al., 2017) (also see Table 1 in Paper I, Page 78).

This model evaluation variation, in turn, means there remain several opportunities for improving the way in which model evaluation studies are conducted. Those opportunities include but are not limited to assessing or re-assessing:

- the types of hydrological indicators studied, especially extreme events,
- inclusion of anthropogenic parameterisations,
- spatial representation of the models performance,
- model ensemble representation, especially by the use of EM, and
- quantitative model performance evaluation e.g. application of more robust metrics such as integrated metrics.

Research approach (RQ1, Obj1). In this context, Paper I presents one of the most comprehensive GHM evaluation studies to date. It uses six GHMs that incorporate human activities in their parameterisations to address the above issues and opportunities. This is done using monthly observed and simulated data for the period 1971-2010 for 40 major catchments across the globe, covering all eight globally-defined, hydro-regions (Meybeck et al., 2013). To this end, different hydrological indicators are analysed, including mean, and high and low flows for monthly and annual runoff, as well as different aspects of the seasonal cycle. Using the hydro-climatic system of hydrobelts and a standardised, integrated model evaluation metric (ideal point error, IPE) (Dawson et al., 2012), results are presented in an improved, spatially-comparable manner.

Contribution to Knowledge. Performing the comprehensive model evaluation detailed above delivers:

- The first GHM evaluation to date that uses models run with human influences, as well as the most comprehensive evaluation of extreme runoff from six GHMs, highlighting the importance of conducting a multi-dimensional GHMs assessment, and the need for the development of a more general GHM evaluation protocol, especially within the framework of ISIMIP.
- An informative framework that highlights the relative variability in the performance of GHMs – with respect to both the hydrological evaluation criteria used and the geospatial locations of the catchments modelled. In other words, it indicates where, which model, and which indicator(s) are the best, globally.

- Opportunities for model improvement given the identification of relative performance of the GHMs in different locations or across different indicators, so that their process representations will improve simulations of extreme flows and the representations of the magnitude and timing of the seasonal cycle, where GHMs show opportunities for improvement.
- New insights into the performance of the EM that challenge the commonly held perception that the EM delivers superior performance over that attainable using individual models.

1.3.1.2 Paper II

Challenge. Global hydrological model ensembles are poorly represented by their EM and not optimally combined.

In practice, model deficiencies are inevitable and this causes errors in their simulations of the real world, and performance variability when comparing models with each other, as set out in Paper I. These errors generate the need for continued improvement of hydrological models, which is normally pursued by activities such as acquiring higher quality data and/or improving structural mechanisms and parameterisations (Liu and Gupta, 2007). Model improvement through these pathways is usually useful, but requires considerable resources. In addition, aleatory (random) and epistemic (resulting from deficits of knowledge) uncertainties, as well as natural variability in the hydrological cycle which remains unknowable mean that it is inappropriate to assume that a single model can provide an adequate basis for a simulation. This leads to alternative approaches being sought: ones that make better use of available data and which overcome the current disadvantages associated with relying on a single model or dealing with outputs from multiple models.

Previous approaches. One way overcoming the above problem has been combining outputs from a multitude of diverse models – a process known as multi-model combination (MMC). This should ideally go beyond simply mixing outputs of several models into a single one and it will benefit from some optimisation and intelligence in the combination process. In current applications of GHMs, as a straightforward solution, the arithmetic mean of outputs from ensembles of several models (EM) is usually used to combine multiple model outputs into a single one (Nohara et al., 2006, Haddeland et al., 2011, Beck et al., 2016, Hattermann et al., 2017). The EM is widely used in climate change and impact studies, especially when making medium-term or long-term projections (Phillips and Gleckler, 2006, Nohara et al., 2006, Dirmeyer et al., 2006, Tebaldi and Knutti, 2007,

Cloke and Pappenberger, 2009, Gosling et al., 2010, Sanderson and Knutti, 2012). The EM and its associated statistical parameters are then used to inform assessments of water availability, water scarcity etc. This approach assumes each model output warrants equal influence in the ensemble output, whereas, there is often little, if any, evidence that the EM is an optimal combination strategy beyond the fact that it is often able to outperform individual models. An alternative approach to the EM (that weights all input models equally) could be a mixed-approach that combines the outputs of model ensembles into a single, integrated solution by weighting the input models based on their performance. In catchment scale hydrology, various arbitrary or 'informed' approaches, such as variable weightings, have been explored to weight models (Giorgi and Mearns, 2002, Fowler and Ekström, 2009, Chandler, 2013, Arsenault et al., 2015) (see the latter for various weighting methods). However, the fundamental challenge remains that of finding computational methods that can optimise both the weighting strategy and the specific weights assigned, in a meaningful way.

Research approach (RQ2, Obj2). To these ends, machine-learning algorithms (MLAs) offer enormous potential as agents for optimising both the weighting strategies and the weights themselves. MLAs are capable of formulating computational solutions for mapping multiple inputs to one or more outputs that are optimised both in terms of their numerical structure and parameter values (see Solomatine et al., (2008) for an overview of the application of MLAs in hydrology). Paper II explores, for the first time, applications of an 'intelligent' MMC method that employs gene expression programming (GEP) (Ferreira, 2001, Ferreira, 2006), which is a type of evolutionary algorithm (EA), and symbolic regression (SR) (Zelinka et al., 2005), to optimise outputs from a GHM ensemble. Using this method, the best aspects of the best models are emphasised, and the worst aspects of the worst-performing models are de-emphasised. This novel approach provides the potential to both enhance model outputs *and* gain valuable heuristic insights into the relative capabilities of different models, under different contextual demands. This, unlike black box and implicit methods e.g. artificial neural networks (ANN) (Shamseldin et al., 1997, Shamseldin et al., 2007, Freitas and Rodrigues, 2006), delivers explicit equations enabling the rationality of solutions to be assessed and the numerical mechanisms used in solutions for different climatic and/or physiographic settings to be examined and compared. In addition, Paper II explores whether one, or a group of GHMs, might be better suited than another to application in different regions of the world, given the multiplicity and diversity of different GHMs. Adoption of the MMC approach in GHM studies could provide both improved simulations and heuristic insights. This would be particularly valuable for climate change impact assessment applications that project the impacts of climate change on global hydrology. Paper II exemplifies the approach

through the development of MMC solutions for simulated runoff from five, individual GHMs participating in ISIMIP2a, across 40 major catchments and over the period 1971 to 2010. The catchments are distributed across all of the world's hydrobelt types and application of the standardised integrated metric IPE promotes robust spatial analysis.

Contribution to Knowledge. Performing, for the first time, intelligent multi-model combination of several GHMs that delivers:

- A novel, intelligent MMC technique that optimises runoff simulations from an ensemble of GHMs by applying MLAs and SR to provide explicit equations that have both their numerical structure and parameter values optimised.
- New insights into the weak performance of EM as a widely applied representative of model ensemble.
- Fresh insights into the potential reasons for MMC failure based on re-examining the explicit SR equations that define the MMC solutions and recognising that failure could arise from MMC solutions' higher level of structural complexity relative to the complexity of combinatorial problems, or re-analysing the residual structure of the GHM outputs.
- Opportunities for meaningful interpretation of GEP-based MMC solutions, given their explicit nature. This includes, but is not limited to: the GHM's contribution to the optimal solution, which reveals the suitability of different models in each catchment or hydrobelt; MMC equation size, which may help to reveal the complexity of the challenge of improving GHM simulations in different catchments/regions, and; equation component examination that may help researchers gain qualitative insights into the numerical mechanisms by which MMC optimisation is achieved.
- Diagnostic insights into the relative merits of GHMs based on the extent of their inclusion/exclusion from the selected MMC solutions. This offers an important means by which limitations in process representation may be better identified and understood, which could, in turn, highlight areas of model improvement specific to each geographical location.
- Exemplary analysis and characterisation of the correction mechanisms used by the SR equation components that support improved qualitative understanding of model deficiencies and correction mechanisms.

1.3.1.3 Paper III

Challenge. Knowledge of the relative spreads in simulations/projections from global and catchment hydrological model (GHM and CHM) ensembles is limited to cross-scale inter-comparions of small ensembles, precluding a robust ensemble comparison.

Previous approaches. While GHMs provide output data for almost all parts of the world, the applicability of outputs generated by a CHM is usually limited to the catchment for which the model has been calibrated. Despite this, both global and catchment models provide valuable information for decision-making and scientific understanding. A global perspective on the hydrological cycle, particularly under impacts of climate change, is vital for quantification of the aggregated impacts. It is at the catchment scale, however, that most adaptation measures take place and where interaction with affected stakeholders is strongest (Hattermann et al., 2017). CHMs tend to be tailored to a specific catchment, so they use local input data and calibration (Hattermann et al., 2017). Calibration usually improves model performance (Hattermann et al., 2006, Koch et al., 2013), although, it can be a complex and resource-intensive procedure. GHMs and CHMs are also distinguished in their spatial resolution, which is usually finer for CHMs. With the expectation that results obtained from both scales should agree in trend and direction (e.g. in terms of the magnitude and direction of climate change impact), their relative performance in simulating present day or projected, future hydrologies have rarely been explored, with the exception of few studies (Gosling et al., 2011, Thompson et al., 2013, Piniewski et al., 2014, Gosling et al., 2016, Hattermann et al., 2017) including Hattermann et al. (2018) conducted in parallel with this doctoral research . Cross-scale studies of this type can provide decision-makers with an appreciation of the possible range of potential impacts that can arise from using different tools available to the hydrological modelling community and the sensitivity of model projections to the spatial scale at which climate change and its impacts are simulated. In the latter context, it is significant that this approach enables testing for consistency between different scales of analysis, and identification of the need for improvement. Nonetheless, cross-scale, inter-comparison studies remain very rare, and those studies that have been conducted have usually applied only one or two models in their comparisons. This is, in part, attributable to the significant resources that are required to setup and calibrate CHMs (Hattermann et al., 2017).

Research approach (RQ3, Obj4). Paper III presents results of a cross-scale comparison study that, for the first time, uses large, multi-model ensembles of GHMs and CHMs to explore whether there are systematic differences between projections of runoff change from two ensembles comprising different types of hydrological model. It also investigates the effect of different degrees of global-mean warming on runoff for each catchment as indicated by each model type. The approach

allows an exploration of the extent to which current runoff will be affected if the 2°C target set by United Nations Framework Convention on Climate Change (UNFCCC) under the Paris Agreement is met, and what the impacts might be if it is missed and global-mean warming rises to 3°C.

Contribution to Knowledge. Cross-scale model comparison using large ensembles of GHMs and CHMs for the first time delivers:

- Large ensembles of GHMs and CHMs suitable for conducting a cross-scale study that assesses the impact of climate change (1°C, 2°C and 3°C warming above pre-industrial levels) on projections of runoff for both mean and extreme river flows.
- New insights on the utility of global-scale hydrological models for understanding changes in runoff at the catchment-scale. This is valuable because, ensemble median estimates of change from ensembles of GHMs and CHMs have generally showed similar effects of global warming on all the hydrological indicators considered. There are, however, some distinct differences, such as a wider spread of hydrological impacts from the GHMs for several catchments.
- Findings that highlight the benefits of applying both types of model ensemble when assessing climate change impacts at the catchment scale. This stems from the finding that superior performance of CHM ensembles in simulating present-day discharges does not guarantee higher credibility of CHM compared to GHM-generated projections under climate change. This finding should encourage decision-makers to consider the full range of potential impacts from both the GHM and CHM ensembles, since these represent the spread of possible impacts according to several different types of credible models used by the hydrological modelling scientific community.
- Findings that highlight the implications of changes in future runoff understood at both scales of hydrological modelling should the UNFCCC target of limiting global-mean temperature rise to well below 2°C above pre-industrial levels is missed.
- Demonstrations of the differing sensitivities of individual catchments to varying degrees of global warming. This help identification of catchments where global climate change mitigation, which limits global-mean temperature rise to below 2°C above pre-industrial levels, could avoid some of the hydrological hazards that might otherwise occur with higher magnitudes of global warming e.g. 3°C.

1.3.1.4 Paper IV

Challenge. Reducing inconsistencies between runoff simulations made using GHMs and CHMs. While cross-scale model comparison (e.g. Paper III) assists end-users with understanding the range

of possible impacts and deciding which type of model to believe, the ways in which model results are compared and presented to end-users (usually by their arithmetic mean) can still be improved.

Previous approaches. In practice, the findings from the handful of cross-scale studies performed to date (Gosling et al., 2011, Piniewski et al., 2014, Gosling et al., 2016, Hattermann et al., 2017)) have been presented by their spread, visualising the level of uncertainty, and/or have been qualitatively summarised in terms of ensemble mean (EM) or ensemble median. In this situation, GHMs usually present a wider spread and larger EM due to their being mostly uncalibrated, coarse-resolution models. This is why the EM was found (in Papers I and II) to have limited utility in representing model ensembles.

Research approach (RQ2, Obj3). Paper IV explores whether, and to what extent, MMC brings runoff simulations generated using GHMs closer to those of CHMs, in different catchment contexts. In addition, this paper examines whether the MMC approach can offer a better ensemble model outcome than the EM. Paper IV adopts the novel, GEP-MMC approach described and applied in Paper II. The approach is applied to optimise combination strategies for the runoff simulations from one ensemble of eight GHMs and one of nine CHMs, across eight, large catchments to, for the first time, conduct a cross-scale comparison of multi-model combinations. It also examines the amount of performance gain/loss if a mixture of both ensembles, a super-ensemble, is used in conducting MMC. Model performance is, again, quantitatively assessed using IPE.

Contribution to Knowledge. Performing, for the first time, a cross-scale comparison of intelligent MMCs from a catchment and a global hydrological model ensemble that delivers:

- Optimised runoff simulations from large ensembles of GHMs and CHMs via application of intelligent MMC technique, which uses MLAs and SR to provide explicit equations that have both their numerical structure and parameter values optimised.
- Insights into the relative performance of the two ensembles, in terms of their individual members (EMs and MMCs) under different settings and in the form of a super-ensemble comprising all GHMs and CHMs.
- Insights into the relative performance of MMC solutions under different settings (different level of complexity) that highlight that a more complex MMC will not, necessarily, deliver a better performing model.
- Demonstrations of the impact of the MMC approach on estimations of hydrological indicators and representations of seasonal cycles ~~(both timing and magnitude)~~, which usually lead to improved simulations.

- Diagnostic insights into the relative merits of GHMs and CHMs, based on the extent of their inclusion/exclusion from the selected MMC solutions, offering an important means by which limitations to process representation may be better identified and understood. This, could in turn, highlight areas for model improvements in the context of each particular, geographical location.

1.3.1.5 Paper V

Challenge. No comprehensive evaluation of GHMs that include human impact parameterisations (HIP) had been carried out prior to that reported in Paper I. Up to now, GHMs have been applied and evaluated under naturalised or near-naturalised conditions (Bierkens, 2015). Such studies are unable to test model's capacity to represent human activities and their impacts on assessment of freshwater resources and hydrological extremes. Meeting societal and scientific needs also necessitates addressing model challenges such as accurately representing the hazards associated with floods (Zhao et al., 2017), water scarcity (Liu et al., 2017a) and droughts (van Huijgevoort et al., 2013) at a global scale. This has led to a fair amount of research aimed at representing human activities and their interactions with the hydrological cycle (Bierkens, 2015). In fact, a fairer evaluation of a model's performance necessitates keeping simulation conditions as close as possible to observations and the real-word situation. However, to date, a systematic analysis that evaluates whether and to what extent incorporating human activities into GHMs improves their performance in simulating real-world hydrological conditions is missing. Failure to account for human impacts on the natural water cycle could potentially preclude the incorporation of GHMs and their modelling results into policy-making and decision-support with respect to water resource management.

Previous approaches. The fact is that there is none. All but one of the previous studies have only evaluated model runs under naturalised conditions (Oki et al., 1999, Milly et al., 2005, Decharme and Douville, 2006, Nohara et al., 2006, Decharme and Douville, 2007, Materia et al., 2010, Haddeland et al., 2011, van Dijk et al., 2013, Beck et al., 2016, Hattermann et al., 2017). The only one study to have evaluated models run with human impacts is that reported in Paper I and no study has demonstrated the additional benefit of including human impacts in GHMs, compared to excluding them.

Research approach (RQ4, Obj5). Paper V presents the results of a study on the change in the performance of five GHMs that participated in ISIMIP2a that followed inclusion of human impacts

in their parameterisation. The research was conducted through comparison of monthly discharges under a naturalised and a time-dependent human impact simulations (respectively termed NOSOC and VARSOC in the ISIMIP framework), based on monthly discharge observations from 473 stations over the period 1971–2010.

Contribution to Knowledge. The impact on GHM performance of incorporating human activities in their parameterisations has been investigated and established, to deliver:

- The first comprehensive evaluation of five GHMs by systematic comparison of indicators of mean and extreme values of their modelled, monthly discharges incorporating naturalised and time-dependent human impact simulations, covering not only (near-)natural catchments but also heavily-managed catchments, which show higher model performance when human impacts are included.
- Quantitative and visual representations of how the inclusion of human impact parametrisation generally enhances the performance of GHMs, by distinguishing features such as changes in metrics (e.g. bias ratio or variability), types of catchments (managed or (near-)natural, the catchment's geographical locations, and hydrological indicators of mean and extreme events.
- Detailed demonstration analyses of the changes in hydrological performance in a selection of 12 managed and (near-)natural catchments, that provide new insights into the mechanisms by which inclusion of human impact in GHMs operates in different geographical locations.
- Descriptions of how changes observed in evaluation metrics (e.g. bias ratio, variability and correlation coefficient) for mean and extreme runoff events when human impact is included might be linked to each part of the parametrisations in a VARSOC run. For example, changes in the variability ratio are found to be driven predominantly by the timing of water abstractions and return flows.
- Demonstrations of the overall impact of including human impact on the performance of the selected GHMs, that highlight potential improvements that could be made with regard to human impact parametrisation. These are based on the evaluation measures (such as the integrated evaluation metric Kling-Gupta Efficiency index (KGE)) and differences in model algorithms and parameterisations
- A new emphasis on the care that should be taken in model selection and interpretation of risk assessment due to large variation in the performance of GHMs that do or do not account for human impacts on the natural water cycle.

2 Methods

2.1 Global Hydrology and Model Intercomparison Projects (MIPs)

2.1.1 Global hydrological models (GHMs); emergence and applications

Global hydrological models describe a group of models designed to simulate the effect of climate variability and water use on river discharge and freshwater resources across the global domain (Bierkens, 2015). They encompass different types of global scale models (introduced in Section 2.1.2), but for the sake of simplicity hereafter all of them are called global hydrological models (GHMs). Compared to smaller scale models (catchment or regional scale), GHMs are usually uncalibrated, have coarser spatial resolution (Hattermann et al., 2017), and compromise between the model performance at the scale of individual catchments and overall model performance (Gosling and Arnell, 2011, Müller Schmied et al., 2014).

GHMs emerged in 1986 (REF) in response to the realisation of the fact that climate and human impacts on the hydrological cycle and water resources occur at continental and global scales, and therefore, a global description of terrestrial hydrology was needed. Early attempts to develop GHMs were led by atmospheric scientists, but these models were well received by hydrologists and other earth science communities (Bierkens, 2015). GHMs have proved their capabilities by being increasingly used in a variety of applications, such as assessing the impacts of climate change on water availability and scarcity (Schewe et al., 2014a, Veldkamp et al., 2017, Gosling and Arnell, 2016), droughts (Burke et al., 2006, Field et al., 2012, Prudhomme et al., 2014), flood hazard and risk (Ward et al., 2013, Dankers et al., 2014, Gosling et al., 2016, Arnell and Gosling, 2016). They have been also used for short-term forecasting (Emerton et al., 2016), and examining the role of water in assessments of food security (Fraser et al., 2013, Elliott et al., 2014, Rockstrom et al., 2009). In turn, their results are used to inform global policy decisions on climate change (Milly et al., 2005, Nohara et al., 2006, Arnell et al., 2014).

From their emergence up until now, GHMs have benefitted from considerable development in different aspects including, but not limited to, expansion of their scientific basis and types, improvement of structure and parametrisations, better spatial resolution, increased diversity of forcing data (both measured *in situ* and remotely sensed), and applications, which have mostly been made possible thanks to advances in computational power. A brief description of some aspects of GHM development is provided in this section, but interested readers are referred to (Bierkens, 2015, Sood and Smakhtin, 2015) for reviews of “global hydrology; state, trends and directions”.

Global water resource (availability/stress) assessments are now routine tasks conducted in several global projects, such as the Global Water Systems Project (GWSP), the UNESCO World Water Assessment Program, and the UNEP Environmental Outlook (Bierkens, 2015). More recent applications of GHMs have focused on the following issues:

- Global flood hazard and risk (Pappenberger et al., 2012, Hirabayashi et al., 2013, Ward et al., 2013).
- Modelling global freshwater temperature (van Beek et al., 2012, van Vliet et al., 2012).
- Groundwater depletion (Wada et al., 2010, Gleeson et al., 2011).
- The contribution of terrestrial water stores to global sea level change (Wada et al., 2012, Pokhrel et al., 2012).
- Methane emission (Petrescu et al., 2010, Ringeval et al., 2014), and
- Medium range to seasonal streamflow forecasting (Alfieri et al., 2013, Yossef et al., 2013).

In terms of resolution, early calculations provided by GHMs were originally available with a resolution normally around 1-2˚ (Veldkamp, 2017), but most of the current GHMs run with a spatial resolution of 0.5˚ (i.e. 50 km at the equator) and with daily time steps. The latest versions of some GHMs e.g., WaterGap3 (Verzano et al., 2012) and PCR-GLOBWB2.0 (Wada et al., 2014), run at an even at finer resolution (0.083˚ = 5 arc min) (Bierkens, 2015).

2.1.2 GHMs classifications

Whilst the general term GHM is used throughout this thesis, it may be useful to provide a brief explanation of the different model types that are combined into this collective terminology. GHMs developed, more or less independently, in different geoscientific fields, of which three main domains include atmospheric science, vegetation and carbon, and hydrology and water resources. Models developed in each domain are respectively termed; land surface models (LSMs), dynamic vegetation models (DVMs), and macroscale hydrological models (MHMs) or Global Hydrology and Water Resources Models (GHWMs), featuring the global water cycle among other processes (Bierkens, 2015).

LSMs have been applied to model the exchange of energy, moisture and momentum in the soil-vegetation-atmosphere nexus. They developed from a rather simple, one-layer soil model

(Manabe, 1969) to more sophisticated models accommodating multi-layer soil routines (e.g. TESSEL; (Viterbo and Beljaars, 1995) and LaD; (Milly and Shmakin, 2002)). Over time, several other LSMs have been developed and applied examples of which are VIC (Liang et al., 1994), NoAH (Niu et al., 2011), ORCHIDEE (Guimberteau et al., 2012), JULES (Best et al., 2011, Clark et al., 2011), CLM (Lawrence et al., 2011), and LM3 (Milly et al., 2014). A new generation of LSMs are land earth system models (LESMs) that combine energy and momentum exchange with mature hydrology, vegetation phenology, and dynamics of carbon cycling. CLM and LM3 are examples of this generation (Bierkens, 2015).

Dynamic vegetation models, DVMs, were originally developed by ecologists to study the impacts of climate change on terrestrial vegetation, especially forests. Through time these models developed to dynamically model both vegetation and carbon cycling (Bierkens, 2015). Development of DVMs dates back to the 1990s, with BIOME (Prentice et al., 1992) being one of the first models to include a simple hydrological modelling scheme. The next generation DVMs emerged through development of LPJ (Sitch et al., 2003) and two of its version; LPJ-Guess (Sitch et al., 2003) and LPJmL (Bondeau et al., 2007). By implementing some hydrological improvements (e.g. those by Gerten et al., (2004) and adding irrigation modules and reservoirs, the global hydrological modelling community have applied LPJ versions for water resources assessments (Bierkens, 2015).

Finally, stand-alone, global hydrology models (GHMs), including both macroscale hydrological models (MHMs) and Global Hydrology and Water Resources Models (GHWMs) (for those explicitly including water use), cover the global scale models that emerged and developed in the scientific domain of hydrology and water resources (Bierkens, 2015). Development of the first GHMs dates back to late-1990s (e.g. WaterGap (Alcamo et al., 1997) and WBM (Vörösmarty et al., 1998)), when water shortages at the global scale led to the first detailed assessments of global water resources and, in turn, the need for higher quality data for modelling purposes. GHMs are normally applied to simulate soil moisture variability and runoff generation on a cell-by-cell basis using input data from climatic (temperature and precipitation), physiographic (land cover, soil texture, slope, and elevation) and surface cover (vegetation, land-use) data sources (Veldkamp, 2017).

Inclusion of water use (domestic, industrial, irrigation water withdrawal and consumption) was a novel aspect of the first GHMs (e.g. WaterGap and WBM) enabling explicit modelling of demands and comparing results against freshwater availability. Over time, several other rounds of improvements have been made to GHMs to enhance both their performance and resolution (Bierkens, 2015). Major advances made to GHMs include:

- Inclusion of reservoir operations (Haddeland et al., 2006, Hanasaki et al., 2006).
- Hydrodynamic routing (van Beek et al., 2011).
- Floodplain inundation (Yamazaki et al., 2011).
- Determining water scarcity at a monthly time scale to account for intra-annual variability in availability and demand (Hanasaki et al., 2008, Wada et al., 2011), and
- Improved representation of surface-groundwater interactions (de Graaf et al., 2014, Döll et al., 2012).

2.1.3 Model intercomparison projects (MIPs)

In practice, due to their different structures, parameterisations and forcing data, different GHMs yield a range of outputs. This could be unhelpful in terms of gaining an improved, overall understanding of the processes being modelled and this could, in turn, hamper uptake of GHMs to inform decision-making by non-specialist policy makers. To minimise the challenge posed by varying outputs from different models, in several projects — known as model inter-comparison projects (MIPs) — GHMs have been run under common frameworks that provide standard modelling baselines to deal with the discrepancies between models (Henderson-Sellers et al., 1995, Entin et al., 1999, Guo and Dirmeyer, 2006, Koster et al., 2006, Harding et al., 2011). Use of MIPs, in turn facilitates higher consistency in, for example, the climate forcings input to the models, the process representations (e.g. the simulation of human impacts such as water abstractions) and the temporal and spatial resolutions of their simulations. In this way, model outputs are more directly comparable with each other, which supports diagnostic inter-comparisons between them (Bierkens, 2015). This achievement was not possible through arbitrarily running multiple models.

The MIP approach was one of the important aspects of LSM applications noted by the atmospheric science community from the early days of performing multi-model intercomparison experiments. In the MIPs experiment by Bierkens (2015), LSMs are used either on their own or coupled with climate models applicable for land-atmosphere studies. The earliest examples of MIPs conducted using LSMs are the Project for Intercomparison of Land Surface Parameterisation Schemes (PILPS) (Henderson-Sellers et al., 1995), the Global Soil Wetness Projects I and II (Entin et al., 1999, Guo and Dirmeyer, 2006) and GLACE (Koster et al., 2006). Further examples of MIPs involving the GHM community encompass WATCH (Harding et al., 2011), phases I and II of ISIMIP (Schellnhuber et al., 2014, Warszawski et al., 2014) and E2O

(Schellekens et al., 2017). As a notable example, phase 2 of ISIMIP has brought together more than 130 modelling groups across various sectors including agriculture, water, ecosystems and infrastructure (Warszawski et al., 2014). Consistency of model outputs from ISIMIP underpins the evaluations and assessments conducted in this research, which would not have been possible without this MIP.

Models participating in MIPs are usually run using the same input forcing data and then compared to each other and/or to the observations to explore their performance (Bierkens, 2015). In addition, through MIPs, global models are examined in terms of the extent to which they can represent the hydrological cycle and the robustness of the climate change impacts they indicate (Veldkamp, 2017). As results, MIPs expose the short comings of GHMs (which can be corrected for), facilitate estimation of model uncertainty, and enhance transparent modelling endeavours, all of which helps modellers achieve their goals for improvement (Bierkens, 2015, Veldkamp, 2017). However, the foci of multi-model studies have usually been comparisons of surface runoff or stream discharge (Haddeland et al., 2011, Haddeland et al., 2014, Hagemann et al., 2013, Prudhomme et al., 2014) and estimates of water scarcity (Veldkamp, 2017) or exposure to water scarcity (e.g. Schewe et al., (2014) have received less attention. Results obtained from MIPs have been used in important applications; for example, ISIMIP outcomes have been cited in influential reports such as the IPCC AR5 Working Group 2 report (Bierkens, 2015).

2.2 Multi model combination (MMC)

Currently, there are numerous models of different types that are applied to simulate the hydrological cycle and water resources. Despite providing useful information for scientific communities and end-users, these models leave modellers and end-users faced with the difficult choice of identifying the best model amongst them. This problem remains unsolved even in MIP frameworks where, although all models are run with a level of consistency e.g. in their forcing data, there exist several model outputs to be dealt with. This, in turn, prompts the question of what to do with them all and how to integrate their outputs in an objective and meaningful manner.

Practically, uncertainties associated with the diversity of hydrological models suggest that no single model can be acclaimed as the 'best' model under all circumstances. Neither is it reasonable to rely solely on one model, as it may fail to perform satisfactorily all the time (Shamseldin et al., 1997). Even though acquisition of higher quality input data and/or efforts to improve the

structures of individual models can certainly enhance a model's ability to replicate observed data (Liu et al., 2007), there is increasing recognition of the opportunities provided by combining the outputs from multiple, diverse models; a process known as multi-model combination (MMC). The idea underpinning MMC is to combine the outputs from several, individual models and deliver an output that outperforms all of the individual models (Shamseldin et al., 1997). In practice, this is usually achieved by reporting the mean output of an ensemble of GHMs (Phillips and Gleckler, 2006, Gosling et al., 2010, Sanderson and Knutti, 2012, Gudmundsson et al., 2012a, Gudmundsson et al., 2012b, Masaki et al., 2017); known as the ensemble mean (EM), and also sometimes the ensemble median (Gosling et al., 2016, van Huijgevoort et al., 2013). However, there is often little, if any, evidence that the EM is an optimal combination strategy beyond the empirical observation that it does often outperform the individual models (Gao and Dirmeyer, 2006, Guo et al., 2007, Liu et al., 2017b). Alternatively, the idea of combining an ensemble of models into a single, integrated solution has proved to be effective (Shamseldin and O'Connor, 1999, See and Abrahart, 2001, Ajami et al., 2006, Azmi et al., 2010).

The idea of combining forecasts obtained from different models dates back to 1970s in economics (Bates and Granger, 1969), but has been used in various forms across multiple disciplines, including hydrological modelling. In hydrological modelling, MMC was first used in the 1980s, when McLeod et al. (1987) combined the outputs of three models to improve forecasts of inflow to a reservoir in Canada. Subsequently, the MMC technique has been applied in numerous studies of which a non-inclusive but representative list is: (Shamseldin and O'Connor, 1999, See and Abrahart, 2001, Shamseldin et al., 2002, Ajami et al., 2006, Shamseldin et al., 2007, Ajami et al., 2007, Azmi et al., 2010, Nasseri et al., 2014). Even though in some cases (Mcleod et al., 1987) model combination did not result in significant improvement, it is generally accepted that MMC is a beneficial technique to enhance simulations/projections/predictions (Clemen, 1989, de Menezes et al., 2000).

2.2.1 MMC approaches

MMC encompasses a plethora of techniques designed to integrate the outputs of multiple models into a single, integrated solution. One classification is based on the approach by which results obtained from different models can be combined. These include model mixing, model synthesis and model switching (Freitas and Rodrigues, 2006) (Figure 2.1).

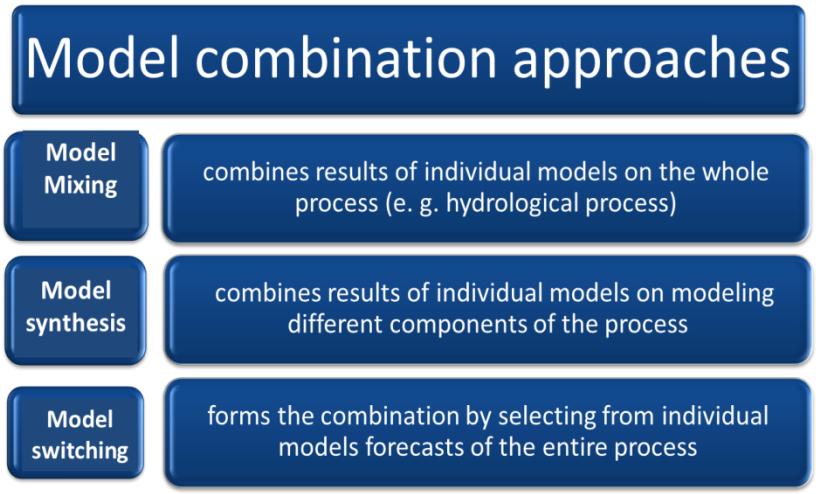


Figure 2.1. Model combination approaches based on the how models are combined. From, Freitas and Rodrigues (2006).

Another distinction is made based on the number of models contributing to the MMC,which could be a single-model or multiple models. In the former case the final forecast is a combination of different instantiations of the same model (Oudin et al., 2006), while in the latter case results from a number of diverse models are combined.

The kinds of models used in conducting MMC have been differently and arbitrarily selected, based on the researchers' justifications and cases of interest. Examples include combining models of the same kind (e.g. data-driven or conceptual models), or models of different types (such as conceptual, statistical, neural and fuzzy) (Figure 2.2). In a review of combined forecasting, based on the degree of complexity, Menezes et al. (2000) categorise combination methods into seven groups. However, simple approaches are claimed to be more accurate (Clemen, 1989).

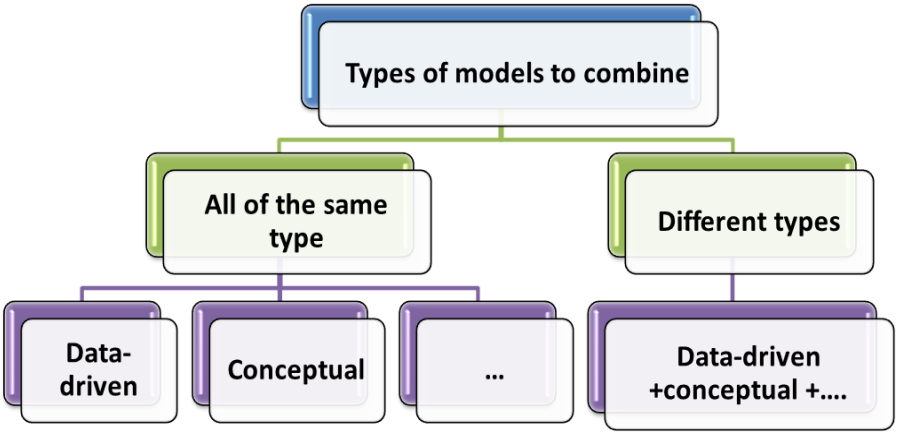


Figure 2.2. Model combination classification based on model types.

In addition, diversity also exists in model combining approaches. Generally, individual models are given weights to generate the optimal combination, based on metrics of goodness-of-fit. The components of combined models can take fixed or variable weights. In the first case, methods used to determine the optimum weight may include simple average, weighted average, artificial neural network (Shamseldin et al., 1997), Bayesian model averaging (Ajami et al., 2007) or fuzzy logic (Abrahart and See, 2002). Variable weighting methods, on the other hand, either give higher weights to models with better performance under certain circumstances or try to minimise model error (Jeong and Kim, 2009).

The ensemble mean (EM) is, in effect, the simplest form of MMC where all model outputs are afforded equal weight in the integrated solution. The EM is widely used in climate change and impact studies, especially when making medium-term or long-term projections (Phillips and Gleckler, 2006, Gosling et al., 2010, Sanderson and Knutti, 2012, Gudmundsson et al., 2012a, Gudmundsson et al., 2012b). However, the implicit assumption underpinning the justification for computing an EM; that overall the diversity of model outputs will contribute towards improving the integrated output cannot be guaranteed. Certain models may be consistently poor performers in certain climatic or physiographic settings, or over certain hydrological response ranges, and their inclusion in a model ensemble may act to limit the performance of the EM (Shamseldin et al., 1997). To overcome this problem, the development of MMC solutions based on identification and application of variable weights has been explored (Giorgi and Mearns, 2002, Fowler and Ekström, 2009, H et al., 2010, Chandler, 2013), see Arsenault et al., (2015) for various weighting methods employed in hydrological modelling. However, these techniques raise equally difficult questions about how the 'best' weighting strategy and combination of weights can be determined *a priori*. Various arbitrary or 'informed' approaches have been explored REF, but the fundamental challenge remains that of finding computational methods that can optimise *both* the weighting strategy and the specific weights in an objective and meaningful way.

2.2.2 Intelligent MMC

While different MMC approaches have shown promising degrees of effectiveness, an alternative method, within which the best aspects of the best models are emphasised, and the worst aspects of the worst-performing models are de-emphasised, represents a novel approach. This approach should have the potential to both enhance model outputs *and* gain valuable heuristic insights into the relative capabilities of different models under different contextual demands (Shamseldin et al., 1997, Beriro, 2015). The fact that this approach has yet to be adopted in GHM studies and

inter-comparison of different scales (i.e. global and catchment models) means that there is a unique opportunity to evaluate whether the optimised combination of multiple GHM outputs (or GHM vs CHM when compared) can provide both improved simulations and heuristic insights. This could be particularly valuable for climate change impact assessment applications that project the impacts of climate change on global hydrology, which is one of the main objectives of this thesis.

In this context, Machine Learning Algorithms (MLAs) offer enormous potential as agents for optimising both the weighting strategies and the weights themselves due to their ability to formulate computational solutions for mapping multiple inputs to one or more outputs that are optimised both in terms of their numerical structure and parameter values (see (Solomatine et al., 2008) for an overview in hydrology). MLAs use iterative, computational processes to generate candidate solutions that are optimised for one or more objective functions; most commonly a minimisation of the statistical difference between the solution and an observed data set. MLAs are stochastic (producing multiple solutions due to the use of random start configurations), heuristic (producing solutions with structures that are not known *a priori*), and non-parametric (incorporating no assumptions about the input or variables distributions) processes (Beriro, 2015). Whilst the numerical solutions generated by some MLAs (e.g. artificial neural networks (Shamseldin et al., 1997, Shamseldin et al., 2007, Freitas and Rodrigues, 2006)) are black box and implicit (Beriro, 2015), others deliver explicit equations. Being explicit enables the rationality of solutions to be assessed and allows the numerical mechanisms used in solutions for different climatic and/or physiographic settings to be examined and compared. In this way, the relative merits of different MMC solutions in different settings is revealed in a more nuanced manner than is possible through the comparison of metrics of fit, alone.

On this basis, in this research, a novel MMC approach is adopted in which Gene Expression Programming (GEP) (Ferreira, 2001, Ferreira, 2006) is employed. GEP is a machine learning algorithm that has demonstrable success in delivering optimal solutions for problems in hydrology (Bărbulescu and Băutu, 2009, Barbulescu and Bautu, 2010) and climatology (Fernando et al., 2012, Aytek et al., 2008, Beriro et al., 2013, Traore and Guven, 2011, Barbulescu and Bautu, 2010, Beriro, 2015). GEP is a highly-adaptive, symbolic regression (SR) algorithm, closely related to genetic algorithm (GA) and genetic programming (GP). GA and GP are generally used as means of optimising the search for the global optimum of a function. Compared to GA, GP allows for optimisation of more complicated structures and it can, therefore, be applied to a wider diversity of problems, especially with regard to optimal control problems involving symbolic regression (Sette and Boullart, 2000).

2.2.3 Symbolic regression (SR)

Symbolic regression (SR) represents a process during which fitness of simulated data to observed data is formulated and measured by means of linear or nonlinear, mathematical formulae. Application of SR by means of evolutionary algorithms (EAs) such as GP was invented by Koza (1998). In fact, in standard linear regression parameters and coefficients are real numbers and an arithmetic, linear fitting model (i.e. least squares) is used to optimise the objective function by finding the values of the coefficients so that the magnitude of the residuals is minimised. Conversely, in SR, the parameters and coefficients do not have to be real numbers (although they can be). They can also be algebraic 'symbols' which represent an underlying set of real numbers and logical and mathematical functions. The aim of SR is the construction of expressions (e.g. $y^3 - c$ and $\cos(x) + e^{-x}$) that use and combine these elements in ways that, once executed, will optimise the fit between input data and observed data. Importantly, unlike linear regression, no *a priori* assumptions are made about the form or complexity of the expression, which can be highly non-linear and can range from very simple to very complex.

When using EAs for problem solving there is not a sole solution, but a class of possible solutions can exist. The main aim of evolutionary algorithms is to find the best solution of all during the evolutionary process. The main attribute of SR is that the whole evolutionary process is working with a so-called functional set and a terminal set. The functional set is a set of all user-defined functions (e.g. SIN and COS), while the terminal set is a set of all constants and variables. SR synthetises mathematical formulae, computer programs or logical expressions to solve user-defined problems (Zelinka et al., 2005). An advantage of explicit SR solutions over traditional machine learning techniques (such as the artificial neural network (ANN) as a black box method) is that they provide insights into the intrinsic work of the technique.

2.2.4 Gene Expression Programming (GEP)

MMCs based upon MLAs may need to accommodate non-linear functions in the combination mechanisms that are learned and which facilitate flexibility in the solutions that must be applied to different data ranges because of non-stationarities in the data. To these ends, GEP offers particularly high potential. GEP is a highly-adaptive algorithm that searches for and learns optimal solutions that can incorporate constants, mathematical operators and non-linear functions to map multiple input data series to an output series using a non-parametric optimisation procedure (Beriro, 2015). GEP is, therefore, well-suited for solving symbolic regression (SR) problems that

need higher computational intelligence than standard regression, in order to deliver solutions of high non-linearity and complexity. This means that, unlike standard regression approaches, *a priori* assumptions about the form or complexity of the solution that is delivered are minimised.

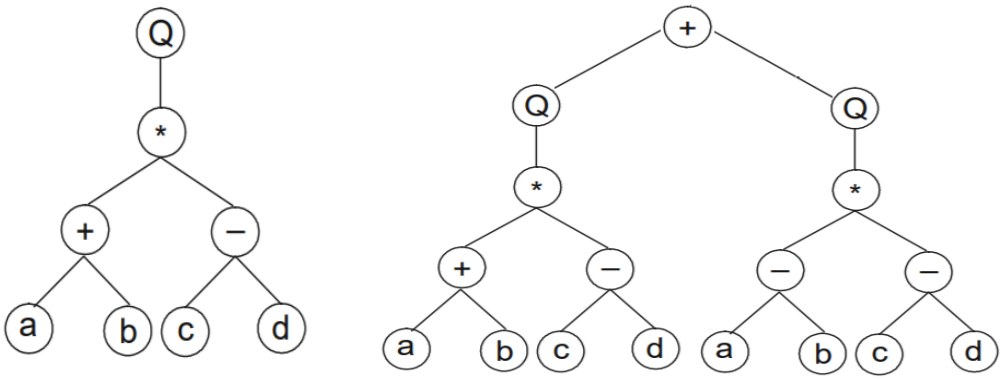
The GEP algorithm is an evolutionary algorithm that is closely related to genetic algorithms and genetic programming as it uses populations of individuals, selects them according to fitness, and introduces genetic variation using one or more genetic operators. In GEP the individuals are encoded as linear, fixed-length chromosomes which are afterwards expressed as nonlinear entities of different sizes and shapes (i.e. simple, diagram representations or expression trees) (Ferreira, 2001). GEP applies Darwinian principles to 'evolve' a set of optimised model solutions, in the form of symbolic regression equations. The process starts with random generation of the chromosomes of the initial population through an iterative, evolutionary learning process that includes both stochastic and systematic computational adjustments. The algorithm and its mathematical basis has been fully described by (Ferreira, 2001, Ferreira, 2006), and the following subsection provides a brief description of the GEP's structure and how it functions.

2.2.4.1 The structure and workflow of a GEP

As stated above, in a GEP, individuals are encoded as linear, fixed-length chromosomes. The chromosomes are themselves composed of one or more genes of equal length, and each GEP gene is formed by a head and a tail. While the head contains symbols that represent elements from both the so-called functions set (F) and the terminal set (T), the tail comprises only terminals (Ferreira, 2001). Each chromosome encodes a user-defined input, a function, a numerical constant or an arithmetic operator. These are the building blocks of the SR model that GEP evolves.

Genes and chromosomes can be presented as expression trees (ETs), which facilitates a better understanding of their organisational structure. Figure 2.3a presents an ET for a gene with the mathematical expression $\sqrt{(a+b)*(c-d)}$ where, Q denotes the square root function. ETs can also be expressed in terms of open reading frames (ORFs), which for the parse tree presented in Figure 2.3a would be 'Q*+-abcd' (a head with the length of 8), reading the ET from left to right and from top to bottom. As a second example '+Q-/b*aaQbaab**aabbaaab**' presents a gene of 21 with 10 and 11 as the length of the head and tail (in bold), respectively, and '-b*b**abbab***Qb+**abbba**' presents a multigenic chromosome of two genes (with tails in bold). In its simplest form, a chromosome can be formed by a single gene. Chromosomes can also be combined (usually by addition) to deliver a final SR solution (often referred to as a program).

Considering $\sqrt{(a-b)*(c-d)}$ (which corresponds to 'Q*--abcd') as a second single-gene chromosome, Figure 2.3b visualises the ETs of the two chromosomes combined by addition.



a b

Figure 2.3. (a) Expression tree for the mathematical expression $\sqrt{(a+b)*(c-d)}$, and (b) the expression tree for two single-gene chromosomes combined by addition.

During reproduction, it is the genes/chromosomes of the individuals, not the ETs, which are reproduced with modification (replicate, mutate, recombine, transpose, etc.) and transmitted to the next generation (Ferreira, 2001). Therefore, the linear, relatively small and easy to manipulate genetic nature of chromosomes makes them advantageous for SR problems. Chromosome elements can be switched on and off; enabling SR solutions of different sizes and complexities to be generated through an adaptive learning process. Importantly, there is also a high degree of flexibility over whether/how often, the inputs, functions, operators and constants are used in the SR models that GEP generates. This gives GEP the unusual property of being able to preferentially 'deselect' inputs and so remove them from its solutions entirely.

In the GEP learning process (Figure 2.4), an initial set of chromosomes encoding a random population of solutions is generated and the fitness of each solution in the set is assessed according to its ability to deliver a predefined objective function – usually a minimisation of the statistical error between the model and a set of observed training data (Fernando et al., 2012). An iterative process then follows whereby the best performing solution is identified as a 'candidate solution' and isolated. The bases of the chromosomes that comprise the remaining solutions are then adjusted through replication, mutation, transposition and recombination (Ferreira, 2001) resulting in the evolution of a set that comprises a new generation of potential solutions. If, at the end of an iteration, the best fitting expression is new, it is added to the candidate solution set which is output at the end of the GEP run. The iterations continue until a

stopping point (usually a pre-defined number of iterations) is met and the set of candidate solutions with different level of fitness (error) is output.

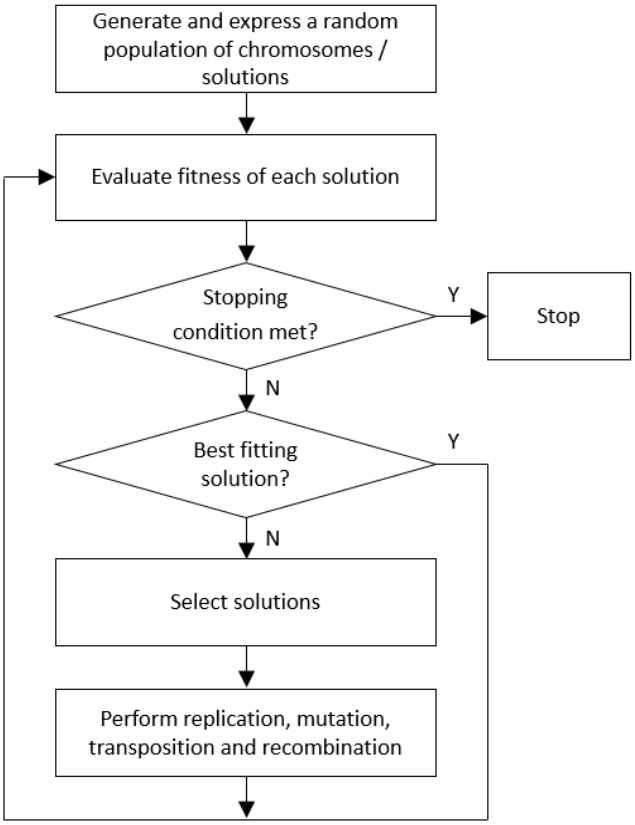


Figure 2.4. The GEP algorithm flowchart.

One important advantage of GEP over other MLAs is the way it structures and expresses its solutions to a modelling problem as component chromosomes (hereafter termed 'components') that, in summation, form the MMC's SR equation. These components are adjusted separately during the evolutionary process; enabling tailored and flexible adjustment of different components of the overall solution. This provides GEP with a computational mechanism for accommodating multiple characteristics that are inherent to the overall modelling problem in its solutions. For example, the need for different solutions at different data ranges can be handled by disaggregation of the GEP solution into multiple components, with the partial solution encoded by each being optimised for a specific data range or for a specific characteristic of the fitting problem. Coupled with its ability to preferentially select / deselect inputs, GEP should have the capacity to support the development of MMC solutions in which the best aspects of the best contributing models are emphasised, and the worst aspects of the worst ones are de-emphasised.

The physical meaning of this is that the approach is capable of picking off low and high flows and their timing towards delivering MMC solutions that outperform individual inputs. In this study, GEP was implemented using GeneXpro Tools 4.0 (GXPT4) (http://www.gepsoft.com/).

2.3 Model performance evaluation

Reliability of (hydrological) models should be tested by evaluating their performance prior to applying them for operational purposes. It is to examine whether their results are reasonable in the modelling context. In order to evaluate the extent to which the main aim of this research— improving runoff simulations from global and catchment models— is achieved, determining the 'fit' of model outputs is fundamental as a key part of the empirical work involved in assessing the relative performance of individual and combined model outputs – both to each other and to observed data. To this end, a sophisticated method for quantifying model performance (and performance gain) is needed. On this basis, this section provides a brief description of some of the model evaluation approaches commonly used in the literature, followed by the evaluation measure(s) applied in each research paper of the thesis.

Model evaluation requires both qualitative and quantitative assessments (Bennett et al., 2013) which together provide acceptability and model communication, and justifies application of a model for both modelers and end-users (Dawson et al., 2012, Bennett et al., 2013) (Figure 2.5). A lot of effort has been dedicated to developing, modifying and standardising metrics of evaluation in the domain of catchment and global scale hydrological models (Dawson et al., 2007a, Dawson et al., 2012, Bennett et al., 2013). For better model performance evaluation which covers different aspects of models it is recommended that both quantitative and qualitative indices be applied (Bennett et al., 2013, Ritter and Carpena, 2013). However, as different sources of error in hydrological models affect metrics differently, having good understanding of what values of different metrics mean is essential (Dawson et al., 2012, Bennett et al., 2013). A good model provides both good predictions i.e. small error, and is not biased i.e. gives a sum of errors equal to zero (Ritter and Carpena, 2013). In addition, both subjective and objective assessments are recommended in evaluation of hydrological models' performance. Subjective assessment of hydrological models deals with systematic behavior like under-prediction and dynamic behavior such as timing. In objective evaluation mathematical estimates of the errors between observed and modelled time series are produced (Dawson et al., 2007b).

In a review paper Bennett et al. (2013) summarise several techniques of model performance evaluation which are categorised in two main classes of quantitative (with two branches; graphical and numerical methods) and qualitative model evaluation (Figure 2.5). These techniques in particular cover approaches that include:

- Direct value comparison
- Comparison through coupling real and modelled values (for studying both graphical methods and quantitative metrics of residual and relative errors—frequently used in data driven modelling field)
- Data pattern preservation
- Indirect metrics based on parameter values
- Data transformation

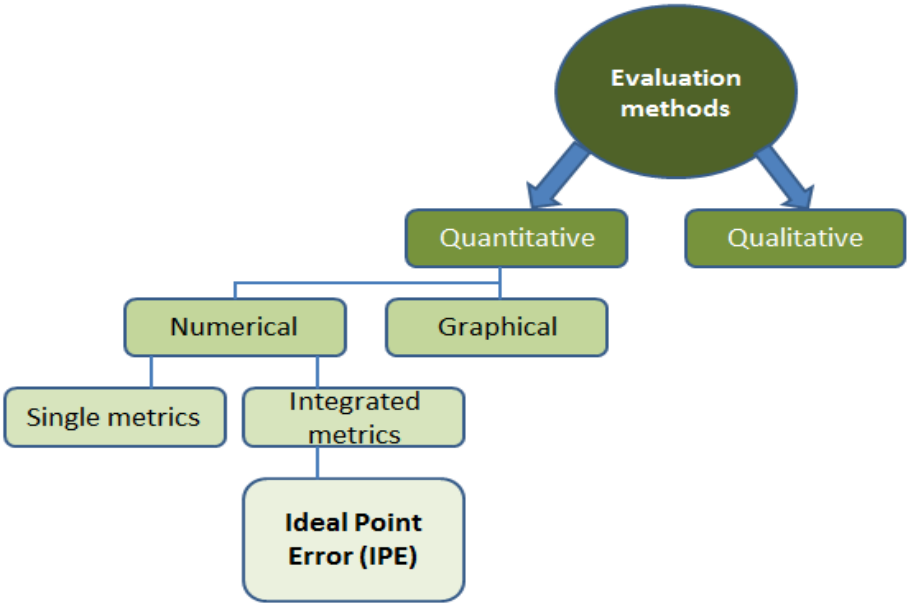


Figure 2.5. IPE in the domain of evaluation methods.

2.3.1 Quantitative model evaluation; graphical methods

Graphical means of model evaluation provide visual insights into the model performance, helpful for subjective model assessment. A description of a few of the graphical tools of model evaluation is provided in the following section.

Scatter plot. A scatter plot of observed and modelled data is a useful means to evaluate them against each other. A lot of information can be obtained from this plot including:

- Possible heteroscedasticity
- If the homogeneity of model performance is in the prediction span or depends on the magnitude of modelled values
- Other unanticipated relationships between datasets like non-linearity or linearity but different to the 1:1 line
- Probable outliers
- Any model bias

Heteroscedasticity—non-constant (increasing) variance in data (Helsel and Hirsch, 2002)—can affect our understanding and interpretation of the nature of time series or relationships between them. In this case, approximation of the points to the identity line (1:1 line) demonstrates the degree of match between observed and modelled values (Ritter and Carpena, 2013).

Flow duration curve. Another widely used graphical method is the flow duration curve, FDC (or similarly probability of exceedance plot). FDC is the graphical representation of all flow records in a given period. FDC provides useful information on the equalled or exceeded runoff in a specified percentage or vice versa. The curve can be plotted for flows in any time interval; daily, monthly or yearly (http://www.ewater.com.au, (2014) and Helsel and Hirsch, (2002)). The order in which data are ranked can be from largest to smallest or vice versa as both approachs are applied in different sources (Helsel and Hirsch, 2002, Brown et al., 2010). There are several formulae for calculating the percentage corresponding to each ranking. One of the most widely used formulae is Weibull (Helsel and Hirsch, 2002) as expressed in equation (1).

$$P = i / (n+1) \quad (1)$$

Helsel and Hirsch (2002) discuss a number of these formulae and suggest Cunnane formula presented in equation (2). They count some reasons for their selection of this formula; for example, normal probability can be acceptably plotted by this formula as a very close formula to Blom. Cunnane formula is also frequently applied for plotting flow duration by hydrologists across Canada and Europe (Helsel and Hirsch, 2002).

$$P = (i-0.4) / (n+0.2) \quad (2)$$

In equations (1) and (2); P is probability, i is the rank of ith value and n the number of all records. In this study, Cunnane formula is used for plotting the probability of exceedence plots.

2.3.2 Quantitative model evaluation; numerical methods

2.3.2.1 Single evaluation metrics

There are several metrics to scrutinise and evaluate hydrological model performance. Because of widespread use of these metrics, several guidelines and pieces of software are developed to facilitate calculation and interpretation of these metrics (see Bennett et al. (2013) for a review). In a classification by Dawson et al. (2007) numerical metrics are categorised into three types:

1) Statistical parameters of observed and modelled time series
2) Statistical parameters of the residual errors like absolute and relative errors
3) Dimensionless coefficient contrasting model performance accepted standards

Statistical parameters of residual errors between observed and modelled data sets comprises absolute errors that assess model errors in terms of the units of variables, and relative errors that are, on the other hand, quantitative assessment of model errors in terms of unbiased, unit free metrics.

One important point to consider while using evaluation metrics is their applicability to the evaluation purposes. Moreover, relying only on one metric may be misleading in model evaluation and selection process, as one model can provide a good score on an absolute or relative error metric, but poor results for other metrics. It is even possible that a model may be discard based on some metrics while it has a good performance for a particular hydrological behavior (Dawson et al., 2007a).

2.3.2.2 Integrated evaluation metrics

In practice, it is not possible to fully determine the performance of a specific model based only on one metric (Willems, 2009). Selecting the best model based on only a superior metric is argued to be "overly simplistic" (Dawson et al., 2012, Pushpalatha et al., 2012), and it is generally accepted that models' performance should be judged based on more than one metric and using more than one method (Bennett et al., 2013, Ritter and Carpena, 2013). In recognition of this, performance analysis using integrated metrics has been acknowledged by scientists (Gupta et al., 1998, Wagener et al., 2001, Willems, 2009, Dawson et al., 2012). This is particularly important in data-driven modelling such as riverflow forecasting for several reasons; first, the limited number of available methods for model evaluation which highlights the role of different available metrics. Second, difficulty in exploring the physical rationale of the modelling procedure due to

the implicit black box feature of majority of the data-driven models. This results in dependency on individual metrics for identifying the optimal model, which is overly simplistic. Third, complexity in choosing evaluation metrics since different sources of error in hydrological models influence assessment metrics differently. This also implies not relying on only one evaluation metric for model evaluation. On this basis, having a single integrated metric is more useful if those metrics are integrated based on a transferrable standardisation (Dawson et al., 2012). Different approaches are suggested in the literature like using a group of metrics focusing on important aspects of model behaviors or creating a purposeful combination of metrics (Dawson et al., 2007a).

2.3.3 Qualitative model evaluation

Despite their usefulness and the amount of information that quantitative metrics provide for models evaluation and comparison, in the end, model acceptance and adoption depend strongly on qualitative, and often subjective, considerations. Qualitative assessments are especially essential in highly complex or data-poor situations. In such cases, qualitative assessment can provide trends and system behavior better than quantitative outputs. These qualitative assessments typically involve expert knowledge which provides, for instance, endpoint estimates to compare against model results or estimates of uncertainty associated with the model (Bennett et al., 2013). Where needed, this thesis has benefited from qualitative evaluation, examples of which are justifying the model selection criterion and application of partial solution spaces in Paper II. In the latter case, by plotting and examining the behavior of the SR equation components of an MMC solution, it is suggested how it may be possible to gain useful, qualitative insights into the mechanisms by which MMC performance gains have been achieved, and the roles of different GHMs in achieving this. In addition, how different SR equation components can be combined to address timing and magnitude errors.

2.3.4 Evaluation measures applied in this research

Table 2.1 summarises the numerical (single and integrated) and graphical model evaluation measures applied in this research. While individual evaluation metrics are presented in some cases, , the overall model performance is mainly assessed by using two integrated metrics; the Ideal Point Error, IPE (Dawson et al., 2012) (in Papers I, II and IV), and the Modified Kling-Gupta

Efficiency Index, KGE (Gupta et al., 2009, Kling et al., 2012) (in Paper V). This helps address several issues identified in applications of individual metrics including:

1) Difficulties in interpreting the overall performance of a model when multiple, separate metrics are used, in particular, when comparing models with each other;
2) The importance of avoiding single metrics when assessing the performance of a model (Abrahart et al., 2011) since different metrics are more or less suited to assessing individual characteristics of a model's fit; and
3) The shortcomings reported for individual evaluation metrics (for example see discussions in (Schaefli and Gupta, 2007, Jain and Sudheer, 2008).

In addition, a benchmarked version of the IPE is developed and applied to facilitate comparisons between models, as well as comparisons across multiple catchments of different scales and hydro-climatic regime. Assessment of relative model performance in a meaningful way is difficult without a transferrable benchmark against which model performance in different catchments can be compared to consistently. Other individual metrics and graphical tools are used depending on the hydrological indicator being assessed (see Table 2.1 and relevant papers therein).

Table 2.1. Model evaluation measures applied in this research.

Paper № in Thesis	Title	Numerical measures applied	Graphical measures applied
I	Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human-impacts.	- Ideal point error, IPE: (formed by Root Mean Square Error (RMSE) as a metric of relative error, Mean Absolute Relative Error (MARE) as a metric of absolute error, and the Nash-Sutcliffe Coefficient of Efficiency (CE) as a dimensionless metric). - PBIAS - Percentage Difference - Pearson correlation coefficient (r)	- Scatter Plot of simulations vs observation - Graphs of seasonal cycle - Plots of residual
II	Optimising runoff simulations from an ensemble of global-scale hydrological models through multi-model combination weighting.	- Integrated metric IPE (as above)	- Scatter Plot of simulations vs observation - Plots of partial solution spaces - Flow duration curve (probability of exceedance)
III	A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1°C, 2°C and 3°C.	- Relative runoff values - Probability of change in runoff - Statistical test (paired-samples t-test)	- Box-whiskers plot
IV	Multi-model combination with a super-ensemble of catchment-scale and global-scale hydrological models.	- Integrated metric IPE (as above)	- Scatter Plot of simulations vs observation - Plots of partial solution spaces - Graphs of seasonal cycle - Plots of residual
V	Human impact parameterisation in global hydrological models improves estimates of monthly discharges and hydrological extremes: a multi-model validation study.	- Modified Kling-Gupta Efficiency Index, KGE: (formed by elements of correlation (r), bias ratio (β) and variability ratio γ) - Nash-Sutcliffe Coefficient of Efficiency (CE)	- Box-whiskers plot - Scatter Plot of simulations vs observation - Graphs of seasonal cycle (river regime) - Flow duration curve (probability of exceedance)

2.4 Data and scale of the study

Different sets of data are used to conduct this research. A brief description of the main sources of data and information used in each paper is presented as follows.

2.4.1 ISIMIP data

The main source of modelled data for this thesis has been simulations for GHMs and CHMs provided by ISIMIP. ISIMIP modelling groups use a standard protocol (available at: https://www.isimip.org/protocol/) to maximise consistency in the temporal and spatial resolutions of their simulations, the input climate forcings to the models, and the process representations (e.g. the simulation of human impacts such as dams, reservoirs and water abstractions). In this way, model outputs are directly comparable with one another, which supports diagnostic inter-comparisons between them. It also means that the outputs from ISIMIP model simulations are ideally suited for use as inputs to an MMC process. The research capitalises on the data provided by two phases of ISIMIP.

The first set was an older phase of ISIMIP known as Fast Track whose water sector simulated data for the period 1980-2010 are used in Papers III and IV. It was selected for the availability of data for both global and catchment models enabling cross-scale studies. Not all CHMs were run for every catchment, however, because unlike GHMs they need to be calibrated on a catchment-by-catchment basis and run for each catchment individually, which is a resource- and time-intensive activity. This means that CHM ensembles may contain a different number of CHM simulations for each catchment. While the CHMs were calibrated, the GHMs were generally not calibrated to catchment-specific conditions. Water management and other human alterations on the fluvial system were not modelled by all GHMs and CHMs. The GHMs and CHMs differ in a fundamental way, which is the spatial scale at which the models represent hydrological processes and in turn the water balance resolved. All the GHMs applied here operate with a 0.5°x0.5° spatial resolution grid across the global land-surface. The CHMs employ various approaches; some of them run on a grid, while others operate by splitting the catchment into sub-catchments and hydrological response units, or treat the entire catchment as a single entity.

The second set of data was provided by the latest simulations from the water sector of the current phase of ISIMIP (ISIMIP2a) applied in Papers I, II and V. These data were selected because they have made available simulated catchment runoff using a protocol that accounts for time-varying human management such as land use, water usage, water withdrawals, and dam

operation (referred to as "VARSOC" simulations in the ISIMIP2a model protocol). For this set of data, all models use the 2015 ISI-MIP2a data release and were run for the period 1971 – 2010 with input climate data provided by the Global Soil Wetness Project 3, GSWP3 (http://hydro.iis.u-tokyo.ac.jp/GSWP3/). Normally, the discharge simulations (amongst other hydrological variables) are available at a daily time resolution and for a global land surface domain at 0.5° x 0.5° grid resolution. Output from the models is openly available from the Earth System Grid Federation (ESGF; https://esgf-data.dkrz.de). For both sets of data, where necessary, conversion of gridded discharge (m^3/s) data to catchment-mean daily/monthly runoff was achieved by using the upstream drainage area.

2.4.2 Study catchments and data used for Papers I and II

Observed data. In these two papers, study catchments are identified setting four criteria. For the sake of data quality consistency, this is done based on observed data from only the Global Runoff Data Centre, GRDC (available from http://grdc.bafg.de). The set criteria resulted in the selection of 40 catchments with data covering the period 1971-2010 (Figure 2.6 and detailed in SI of Papers I and II). Even though the selected catchments provided a good geographic coverage, the availability and quality of observed data resulted in a bias towards catchments in boreal and northern mid-latitude hydrobelts.

Simulated data. In these two papers, simulated data are VARSOC runs of several GHMs from the same source and specifications explained in Section 2.4.1. The only difference is that in Paper I simulation outputs are from five GHMs including DBH, H08, LPJmL, PCR-GLOBWB and WaterGAP2, and in Paper II simulations are from six GHMs including DBH, H08, LPJmL, MATSIRO, and PCR-GLOBWB (see the relevant papers for their specifications). The reason for this difference is due to the availability of GHMs output under VARSOC runs when these studies started.

2.4.3 Study catchments and data used for Papers III and IV

Observed data. For Papers III and IV, eight large (>50,000 km2) river catchments were selected for each study where the Global Runoff Data Centre (GRDC) held observed discharge data. These data were then later used to calibrate the catchment hydrological models used in these papers (Krysanova and Hattermann, 2017). Although of the same number, the location of the selected catchments differ since different data were required to conduct each experiment in Papers III

and IV. The catchments used in Paper III comprise the Upper Amazon, Darling, Ganges, Lena, Upper Mississippi, Upper Niger, Rhine and Tagus, and those included in the study in Paper IV include the Upper Amazon, Lena, Upper Mississippi, upper Niger, Rhine, Tagus, Yangtze and Yellow. For both studies, historical data cover the time period 1971-2001 (see Figure 2.7 for their locations).

Simulated data used in Paper III. The modelled data applied in this study are of the same data source used in Paper III. There are, however, differences in terms of models of the two scales applied and their number. Here, in total, nine GHMs (DBH, H08, LPJmL, Mac-PDM0.9, MATSIRO, MPI-HM, PCR-GLOBWB, VIC, and WBMplus) and eight CHMs (HBV, HYMOD, HYPE, mHM, SWAT, SWIM, VIC and WaterGAP3) are used to conduct the study (see the relevant paper for their specifications).

Global warming scenarios applied in Paper III. Daily input data on the climate variables required by each hydrological model (Paper III, Online Resource 2) were extracted from five global climate models (GCMs) simulations: HadGEM2-ES, IPSL-CM5A-LR, MIROC-ESM-CHEM, GFDL-ESM2 and NorESM1-M; all run under the RCP8.5 emissions pathway (Riahi et al., 2011) for the period 1971-2099 because this is the only pathway for which all five GCMs reach 3°C by the end of their simulation period (year 2100). All variables were bias-corrected towards the WATCH observation-based dataset (Weedon et al., 2011), using an established method (Hempel et al., 2013), specifically designed to preserve long-term trends in temperature and precipitation projections to facilitate climate change impact assessments.

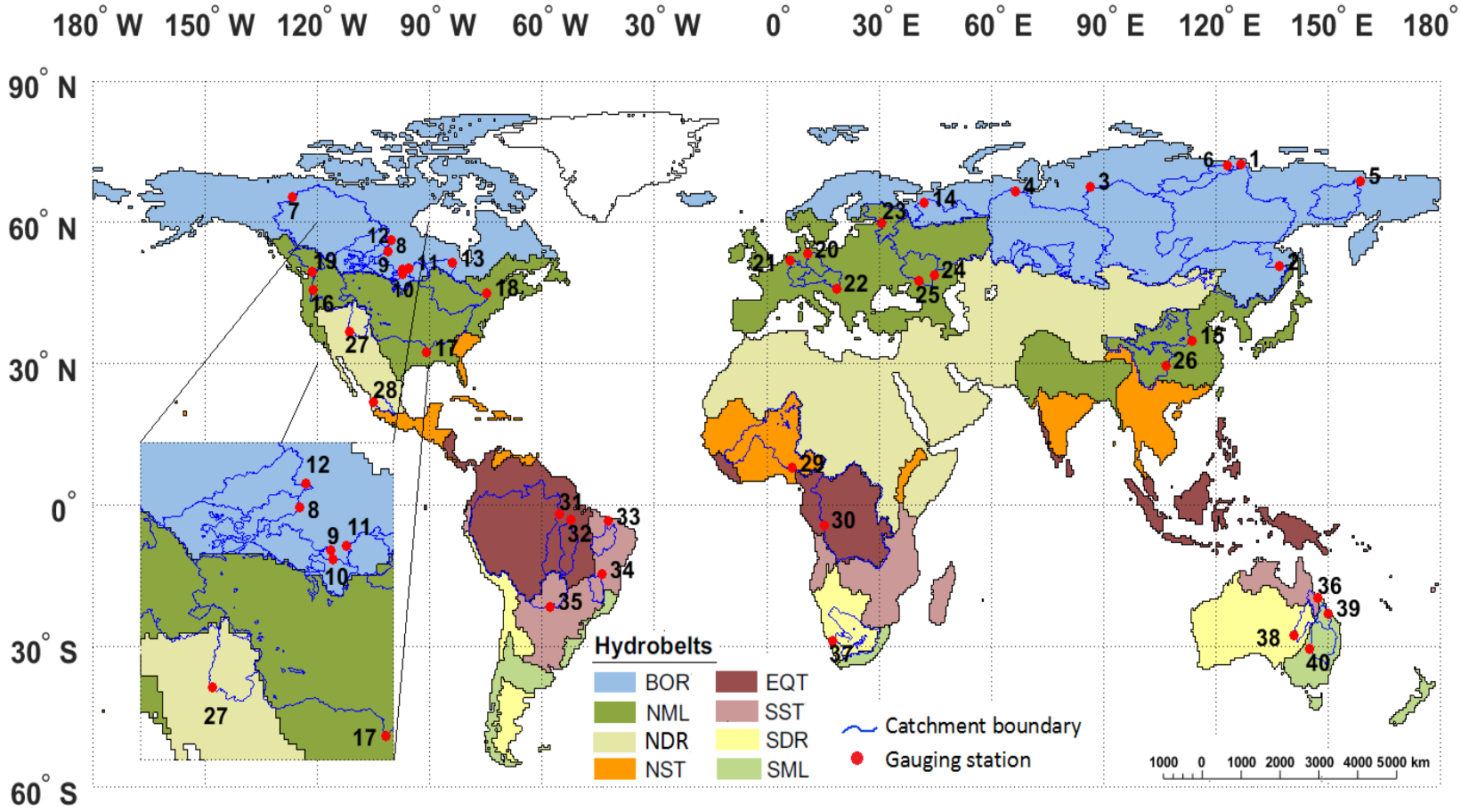


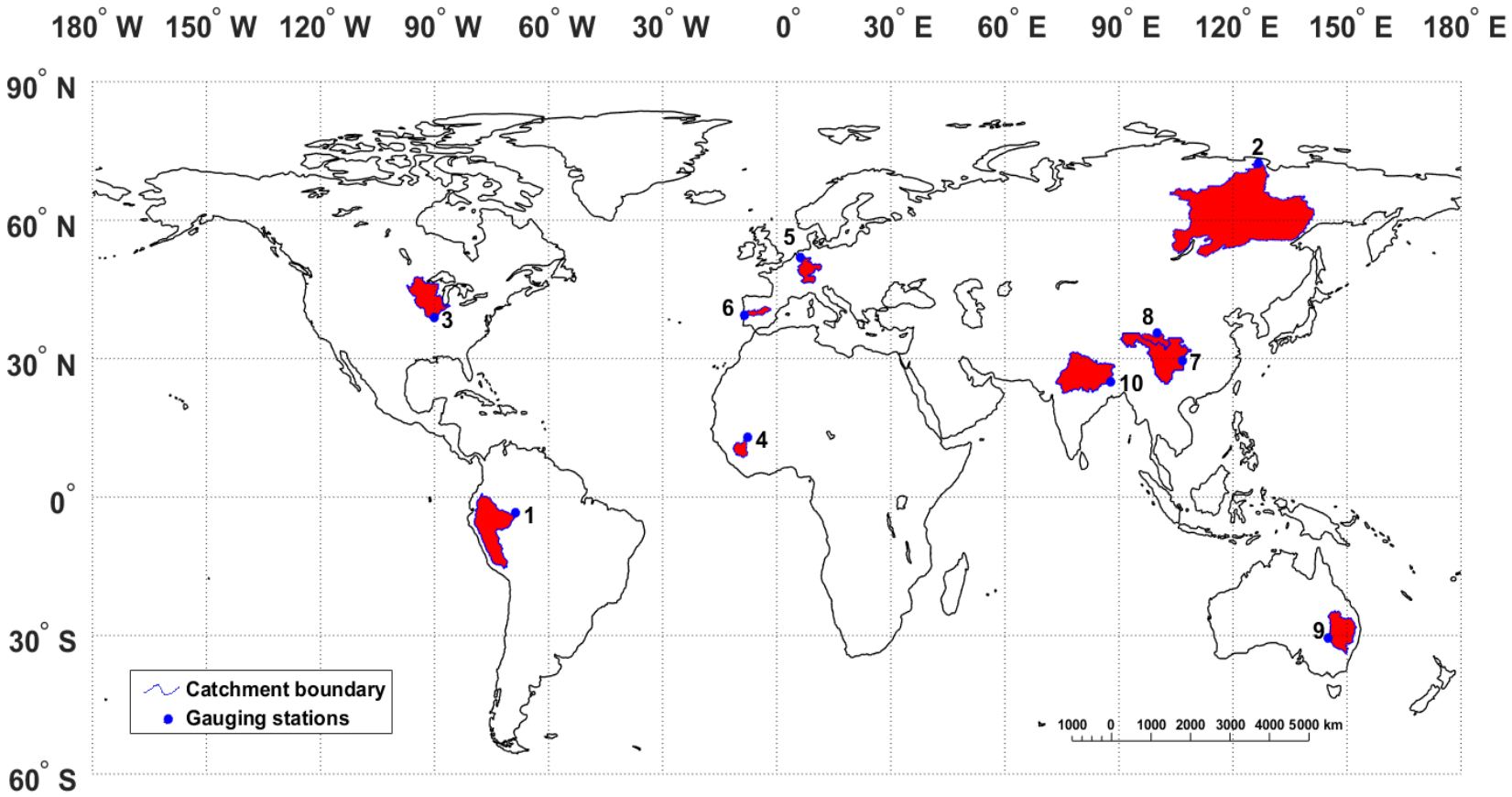
Figure 2.6. Locations of the 40 selected catchments used in Papers II and III across the hydrobelt system defined in Mybeck et al. (2013). Shapefiles of the selected catchments from the global runoff data centre (GRDC), Germany, and shapefiles of the hydrobelts freely provided by Meybeck et al. (2013). Hydrobelts name: BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML=southern mid-latitude, SDR=southern dry and SST=southern subtropical.

Simulated data used in Paper IV. Simulated data for this paper and the next are provided by a large international collaboration of multiple GHM and CHM modelling groups brought together under the framework of the Fast Track phase of ISIMIP over several years (Warszawski et al., 2014). In total, eight GHMs (CLM, DBH, H08, LPJmL, MATSIRO, MPI-HM, PCR-GLOBWB and WaterGAP) and nine CHMs (ECOMAG, HBV, HYMOD, HYPE, mHM, SWAT, SWIM, VIC and WaterGAP3) comprising two ensembles participated in the assessment (see the relevant paper for their specifications).The CHMs were calibrated and the performance of the calibration evaluated in a separate validation period (Huang et al., 2016) while the GHMs were generally not calibrated to catchment-specific conditions. Detailed descriptions of the GHMs and CHMs employed in this study are provided along with the respective references to each model cited in Table 1 and in Table S3 of the paper.

2.4.1 Study catchments and data used for Paper V

Observed data. Observed monthly discharges from the GRDC were used as reference dataset, consisting of > 9,000 stations in 160 countries. With the set of criteria explained in Paper V, 471 stations with data covering the time period 1971-2010 were selected for analyses in this paper (Figure 2.8). Of all stations selected, 226 were identified as managed (i.e. having a managed upstream area), and 245 as (near-)natural. Of 471 stations, 92 are outlet stations which cover in total 19.8% of the global land area. Twelve (highlight) stations are picked up for more detailed analyses and exemplification.

Simulated data. In this paper, simulated data are NOSOC and VARSOC runs of five GHMs from the same source and specifications explained in Section 2.4.1. The GHMs whose data are used in this paper comprise H08, LPJmL, MATSIRO, PCR-GLOBWB and WaterGAP. For a detailed overview of the modelling characteristics of each of the GHMs and the methods for parameterisation of human impacts, refer to **suppl. table 1** and the individual model references.



Catchments: 1. Upper Amazon; 2. Lena; 3. Upper Mississippi; 4. Upper Niger; 5. Rhine; 6. Tagus; 7. Yangtze; 8. Yellow; 9. Darling; 10. Ganges

Figure 2.7: Location of the study catchments and gauging stations considered in Papers III and IV. Shapefiles of the catchments courtesy of Simon Gosling, University of Nottingham, UK. Projection: EQDCYLIN Equidistant Cylindrical.

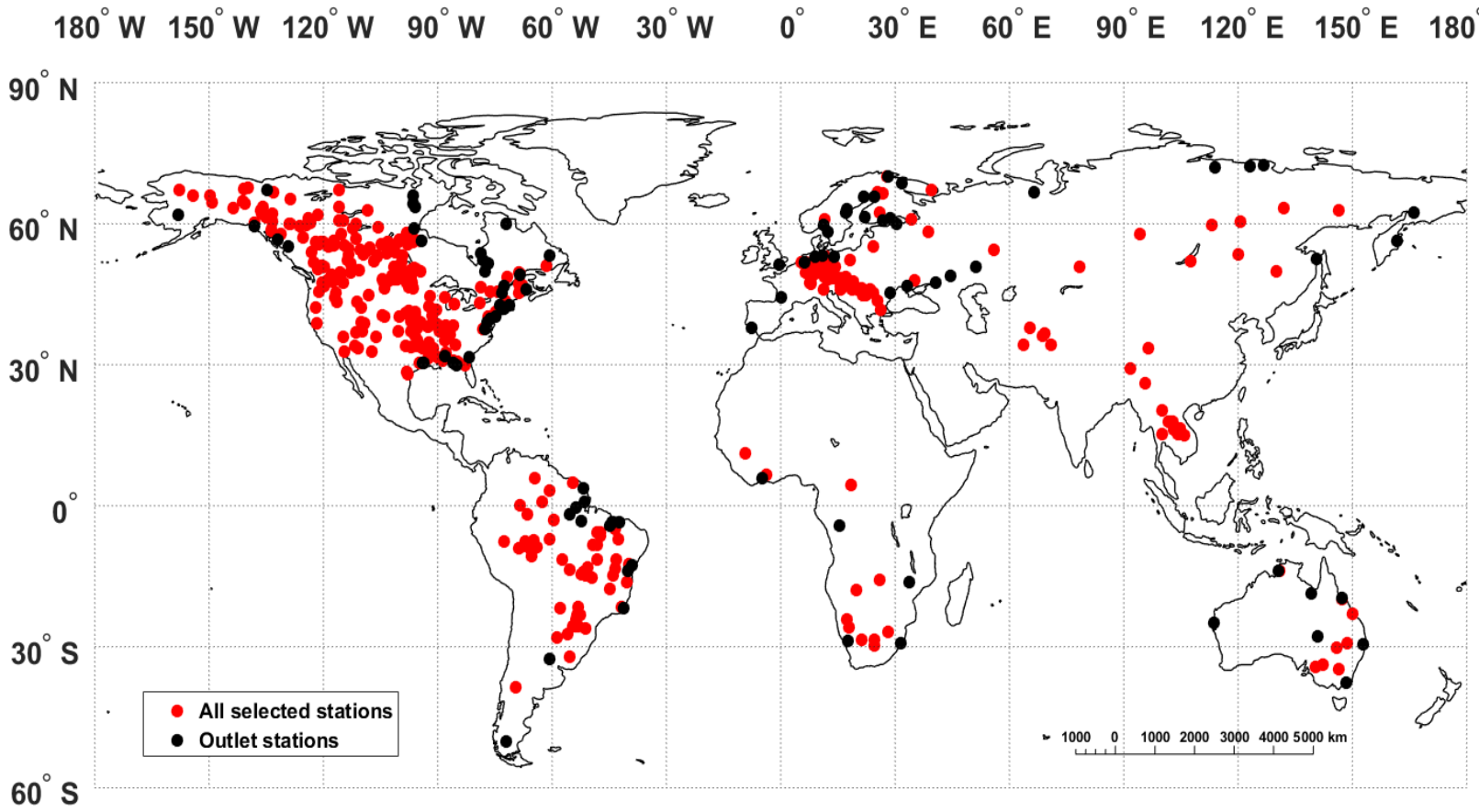


Figure 2.8: Spatial distribution of GRDC stations used for analysis in Paper V. Red dots (n = 471) indicate all GRDC stations that meet the selection criteria, whereas the black dots (n = 92) refer to the stations that are located at the outlet of a catchment. Shapefiles of the stations courtesy of Ted I.E. Veldkamp, VU University Amsterdam, the Netherlands. Projection: EQDCYLIN Equidistant Cylindrical.

3 Synthesis of the results and future directions

This chapter synthesises the main findings of the papers and possible implications of the findings. It also suggests areas of further potential research that would improve the methods applied and results obtained. It is then followed by a concluding remark and the latest status of each paper, to date.

3.1 Reliability of the EM

Inclusion of human impact parametrisations in the GHMs (Paper V) has been shown to generally achieve better performance in the GHMs studied, and revealed some of the models' shortcomings (see Section 3.5). Paper I yielded more detailed insights into the specific strengths and weakness of the GHMs (see Table 3.1 for results), especially with respect to different hydrological indicators. Moreover, applying a comprehensive number of hydrological indictors of extreme events and different tools of performance evaluation and aggregation (IPE and a newly introduced climate zoning system), showed conformities between the findings of Paper I and Paper V, some of which include:

- General tendency of models to overestimate indicators of mean and extreme runoff
- Spatial heterogeneity of performance presented by models for different indicators
- Opportunities for model improvement through making the linkage between indicators and GHMs' structure and parametrisations

In addition, Paper I provided specific insights into:

- The performance of GHMs in replicating the timing and magnitude of the seasonal cycle
- The performance of GHMs in reproducing extreme events of different return periods
- The performance of the EM

Paper I is the first GHM evaluation assessment that uses models run with incorporated human influences and the most comprehensive evaluation of extreme runoff, from six GHMs, highlighting the importance of conducting a multi-dimensional GHMs assessment. It contributed to constructing a more informative framework that says where, which model, and for which indicator(s), is best globally.

Paper I yielded new insights into the performance of the EM and challenged a commonly held perception that the EM delivers superior performance over individual models, as it failed to outperform the best ensemble member in 95% of the catchments when using all six GHMs. The implication is that a decision should not be taken *a priori* to use the EM as the basis of model evaluation and/or climate change impact assessments (Van Loon et al., 2012, Gudmundsson et al., 2012a, Schewe et al., 2014a, Beck et al., 2016) without considering its performance relative to the models it is summarising, because an individual model may perform significantly better. In addition, while excluding outlying models (best or worst performing model(s)) improved the EM's performance to some extent, the question of how to perform the process optimally to minimise the risk of missing information inherited by poor performing models still remained challenging. This, in turn, highlighted the need to investigate alternative, more robust ways of summarising model ensembles.

This study also highlighted opportunities for model improvement such as robust model calibration to GHMs, as well as improved model parametrisation. Calibration of global models might help reduce the performance gap between GHMs and catchment scale models. With regard to parametrisation, the study highlighted which models perform particularly well for certain hydrological indicators, and in which regions (hydrobelts). Thus, the global hydrological modelling community could dedicate greater effort to improving the representation of low runoff, and the magnitude and timing of the seasonal cycle, where GHMs showed higher challenges.

The application of the integrated metric IPE proved to facilitate overcoming problems of using several different metrics, as well as spatial comparison of the results. However, the IPE results obtained for e.g. ephemeral catchments (affected by large MARE due to overestimated precipitation) suggests that it might be better to use other instantiations of IPE for such areas. Finally, the adoption of the hydrobelt classification system (in Papers I and II) provided a feasible means of aggregating catchment-scale results around the axis of hydro-geographical similarities, as well as facilitating a comparison of model performance spatially worldwide. Paper I suggests opportunities for improving GHMs evaluation processes presented in the diagram in Figure 3.1.

Table 3.1. Key findings and summary of results obtained for Paper I.

RQ1: What are the characteristic strengths and limitations of the 'state-of-the-art' global-scale, hydrological modelling?	
Obj1: Comprehensive evaluation of mean and extreme runoff from six global-scale hydrological models that account for human-impacts	
Key findings: • **The EM is not necessarily the best representative of a model ensemble.** • **GHMs performs differently across different indicators and different hydrobelts.** • **GHMs perform better in wetter hydrobelts, but globally tend to overestimate the hydrological indicators.** • **GHMs's struggle to capture the timing and magnitude of the seasonal cycle more highlighted in northern and southern hydrobelts, respectively.**	
1) How do GHMs perform in replicating time series of observed monthly runoff?	1) Model performance varies spatially, but based on IPE scores overall calibrated WaterGAP2 performs best and DBH weakest - Generally, GHMs perform better in the wetter equatorial and northern hydrobelts (EQT, BOR, NML, NDR and NST) than in drier southern hydrobelts (SST, SDR and SML).
2) How well do GHMs do in simulating MAR and extreme events (Q5, Q95 and flows of different return periods)?	2) Majority of models tend to overestimate mean annual runoff and all indicators of upper and lower extreme runoff, particularly: - The models overestimate low flows (Q95) considerably more than they overestimate high flows (Q5). - The models overestimate minimum flow return periods to a lesser degree than they do for maximum flows.
3) How do models perform in simulating seasonal cycle (both magnitude and timing)?	3) There are biases in reproducing both the timing and magnitude of seasonal cycles across all hydrobelts with a general pattern of overestimation and higher timing bias for minimum runoff. - GHMs struggle to capture the timing of the seasonal cycle, particularly in northern hydrobelts, while in southern hydrobelts the models struggle to reproduce the magnitude of the seasonal cycle. - Across the hydrobelts, timing early bias is especially prevalent for minimum flows, while late bias is less evident. - The highest magnitude bias is observed in SDR and SML, and the lowest in NML and EQT.
4) How does the performance of GHMs spatially vary, given the application of a standardised integrated evaluation metric and a more hydro-geographically based climate zoning system?	4) The hydrobelt system facilitated worldwide aggregation and comparison of catchment-scale results for different hydrological indicators around the axis of hydro-geographical similarities. - Better results were seen in the wetter equatorial and northern hydrobelts. - Overall, for time series replication, the best performance of models (in term of IPE) takes place in the EQT, and the weakest in SDR hydrobelt. - For the hydrological indicators, the best performance of models (based on PBIAS) takes place in the BOR, NML and EQT, and the weakest in SML and SDR hydrobelt.
5) What insight does this study provide concerning the performance of the EM?	5) Based on IPE scores, the EM of 6 GHMs outperformed the best individual model in only 5% of the catchments. - When assessed against hydrological indicators (MAR etc.), the performance of the EM improves relative to the individual models, but some individual models still perform better than the EM. - The EM is not necessarily the best way of summarising the performance of a model ensemble. This highlights needs for further study on the application of more reliable model ensemble representatives.
6) What model improvement insights (in terms of structure, parametrisation, and calibration) does this evaluation provide?	6) GHMs performance varies by hydrological indicator and spatially, meaning that not a single model is the best everywhere or in all aspects. Each model might have superior structure and parametrisations for a specific setting that other modellers could look at these to improve their models. Examples of these include: - Incorporating process representations that resulted in better simulation of high flows as seen for DBH. - Incorporating parametrisations that lower timing bias of seasonal cycle as seen for LPJmL and DBH.

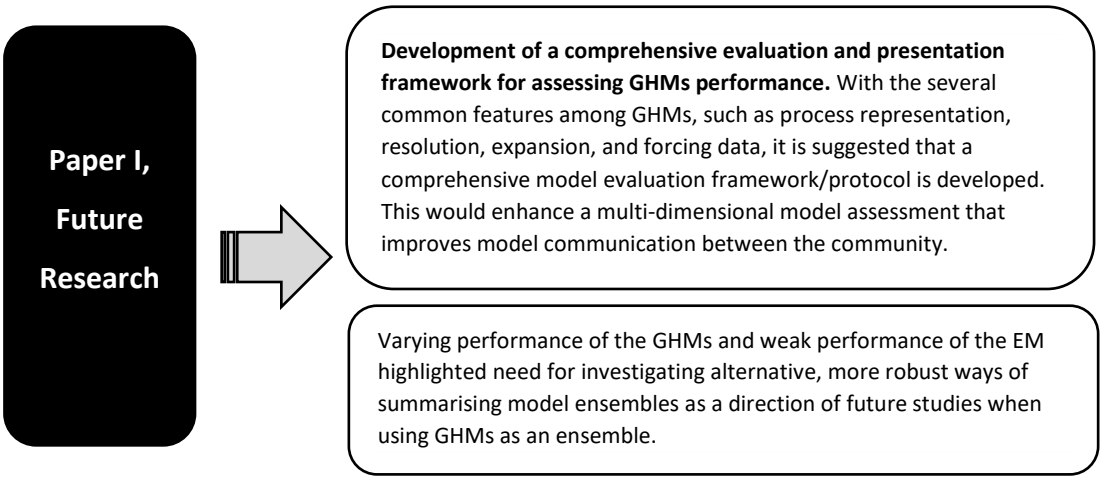


Figure 3.1. Future research suggested based on the study conducted in Paper I.

3.2 MMC; applying model ensembles more intelligently than the EM

Aggregating the results through the application of the hydrobelts system, an integrated evaluation metric— IPE, and the EM as the simplest combination of the GHMs, was one of the endeavours of Paper I to facilitate model evaluation and then decision-making when dealing with several GHMs. In the existence of diversity in models' performance across different indicators and spatially, the EM could act as a means of model aggregation, so that the application of model ensemble is more feasible. However, as seen in Paper I, the EM failed to represent the ensemble in a successful manner. This was found in stark contrast to several other GHM and climate modelling studies, which have shown that for the historical period, the EM is closer to observed data than the simulations from any individual model (Tebaldi and Knutti, 2007, Cloke and Pappenberger, 2009, Sanderson and Knutti, 2012). In addition, the exclusion of the outliers from the ensemble did not improve the EM performance to a considerably reliable level.

In response to the varying performance of the GHMs and weak demonstration of the EM, as the core of this thesis, Paper II sought for an alternative way of combining the models looking for enhanced model simulation and representation. The intelligent MMC approach applied in this paper proved a success in ensemble application of GHMs by overall outperforming the best performing GHM and EM in 85% of the catchments. The MMC performance gain was 38% against the best GHM and excess of 100% against the EM (see Table 3.2).

Paper II contributed to the field the first multi-model combination of several GHMs using machine learning algorithms (MLAs). The method provided explicit equations and enabled

the investigation of potential reasons for MMC failure, as well as opportunities for meaningful interpretation of the MMC solution, including GHMs contribution to the optimal solution, MMC equation size, and equation component examination. It also contributed to diagnostic insights into the relative merits of GHMs based on the extent of their inclusion / exclusion from the selected MMC solutions. This offers an important means by which limitations in process representation may be better identified and understood that could, in turn, highlight areas of model improvement in each geographical location. In addition, the approach helped qualitative understanding of model deficiencies and correction mechanisms by analysing the SR equation components. Finally, it facilitated examining the spatial performance of the MMC strategies across different hydro-climatological regimes, which may help to establish a new route towards diagnostic appraisal of the relative performance of different GHMs in different hydro-climatic settings.

Therefore, parallel to possible improvements that can be made to GHMs performance by activities such as inclusion of human impact in their modelling scheme (Paper V), MMC approach added more gain of the available data (median performance gain of 45% over the best individual model or EM) and provided a more robust ensemble representative than the EM. In light of the significantly improved performance offered by MMC, relative to individual GHMs and also the EM, it is recommended that future multi-model applications consider using MMCs, alongside the EM and intermodal range, to provide end-users of the ensemble with a better informed estimate of what the ensemble shows.

Despite the good performance of the MMCs across the majority of catchments, it should not be seen as a “silver bullet” for counteracting biases and fit residuals of individual GHMs. In six (15%) of the catchments either the EM or an individual GHM performed marginally better than the MMC solution. In addition, care must be taken in presenting MMC results in isolation. Instead, it is recommended that MMC results are presented alongside the range of model outputs from the whole ensemble and the EM. In fact, MMC does not reduce the uncertainty range, as argued in (Giorgi and Mearns, 2002) – but provides a more robust and informative estimate from the ensemble that takes into account the performance of its members. This research offers several directions as ways of future research presented in the diagram in Figure 3.2.

Table 3.2. Key findings and summary of results obtained for Paper II.

RQ2: Can optimised combinations of multiple GHM or CHM outputs generate improved simulation outputs and provide a mechanism for delivering heuristic insights into the strengths and weaknesses of different models?	
Obj2: Optimising runoff simulations from an ensemble of global-scale hydrological models through multi-model combination weighting	
Key findings: • **The MMC outperforms the best performing GHM and EM in 85% of the catchments with the median performance gain of 46%.** • **The EM outperforms the best individual GHMs in only 10% of the catchments** • **The largest MMC performance gains are achieved in hydrobelts located in the southern hemisphere (SST, SDR and SML).** • **MMC consistently performs better than the GHM and EM across MAR, Q95 and Q5, but tends to underestimate Q5.**	
1) How do GHMs and EM perform, and what performance gain do MMC solutions deliver over the individual GHMs and EM?	1) - The MMC performs better than the best performing GHM and EM in 85% of the catchments. The median performance gain is 46% against the best GHM or EM, 38% relative to the best GHM, and in excess of 100% against the EM. - MMC consistently performs better than the GHM and EM across MAR, Q95 and Q5, but tends to underestimate Q5 i.e. hazard flood events. - The EM marginally outperforms the best individual GHMs and MMC in only 10% and 5% of the catchments respectively - Among the GHMs, calibrated WaterGAP2 is the best model in 32 (85%) catchments and LPJmL in none of the catchments.
2) What is the spatial pattern of the GHMs, EM and MMC solutions performance across the globe?	2) When examined by hydrobelt/catchment: - MMC solutions generally deliver substantial improvements over their EM and GHM counterparts in all hydrobelts. The MMC performance gain is largest against the EM than the best-performing GHM in all hydrobelts. - The largest MMC performance gains are achieved in hydrobelts located in the southern hemisphere (SST, SDR and SML). - The hydrobelts system also reveals where MMC provides opportunities for more frequent performance gain (e.g. in BOR where 65% of the catchments are positioned in the top 50% of the MMC performance gain rankings). - None of the GHMs, the EM nor the MMC solution was able to improve upon the naïve benchmark model in catchments in NDR and SDR (all their IPE>0). This indicates that the process representations employed in the GHMs may be deficient for modelling runoff in these hydrobelt.
3) What insights does the intelligent MMC technique offer into the success or failure of the technique and combination mechanism?	3) Given the explicit nature of the MMC solutions, the GEP-MMC technique enabled gaining heuristic insights into: A) Potential reasons for MMC failure which could include: - MMCs high structural complexity; indicating inappropriateness of either or both 'One-size-fit-all' setting and fitness-size trade-off for MMC selection. - GHMs residual structures; having similar structure or no structure limiting the opportunity for GEP learning through complementary residual structures. B) Interpreting MMC solutions in terms of: - GHMs contribution to MMC solutions, with WaterGAP2 and LPJmL respectively contributing most and least in both MMC solutions and their best components. - MMC equation size, as an index of the complexity of optimising GHMs combination, with lowest median in SDR and EQT and highest median in NST. - SR equation component examination, helpful for analysing GHMs performance share in each MMC solution, and qualitative understanding of models process representation deficiencies and correction mechanism e.g. in terms of timing and magnitude to gain performance by an MMC solution.

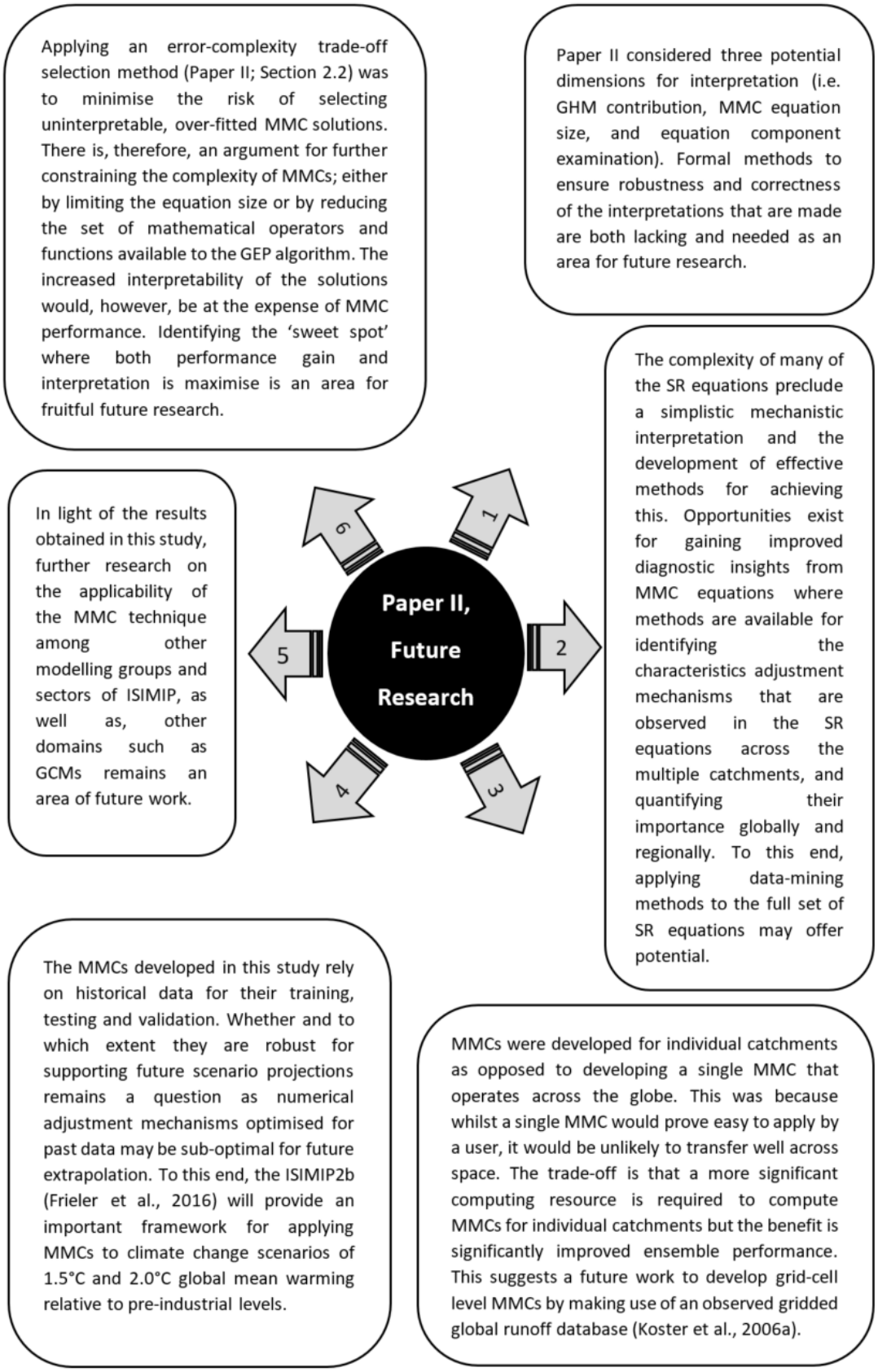


Figure 3.2. Future research suggested based on the study conducted in Paper II.

3.3 Which scale; global or catchment?

Papers I and II presented some features of GHMs' skill in modelling global runoff, some of their strengths and weaknesses, their performance in the simple (EM) and intelligently-weighted combined forms, and areas of potential improvement that can be made to the models. The results were that an optimised combination of GHMs through the means of machine learning and symbolic regression would, generally, outperform the best GHM and EM, and that models' performance varies by indicator and geographical location.

While GHMs have proved to be valuable in exploring hydrological phenomenon at continental and global scales, examining their applicability at catchment scale and their performance relative to catchment scale hydrological models could clarify their relative value even more, and potentially answer the question if they can be used instead of CHMs. This is more highlighted given the amount of resource that CHMs need to be set of and calibrated to each catchment specifications. Cross-scale studies could help towards finding this out, and have been performed previously. However, they have been handful (Gosling et al., 2011, Thompson et al., 2013, Piniewski et al., 2014, Hattermann et al., 2017) and usually applied few number of models of both types, which could preclude making a robust conclusion of their relative performance. In responding to these gaps, Papers III and IV, present two cross-scale studies of large ensembles of GHMs and CHMs using both historical and projections of multiple models.

Along with a parallel study (Hattermann et al., 2017), Paper III, contributed as the first multi-catchment assessment study comprising a large ensemble of 9 GHMs and a large ensemble of 8 CHMs to assess and compare the impacts of climate change from both model ensembles. Studying different hydrological indicators of mean and extreme events, the results shed new light on the application of GHMs to understand changes in runoff at the catchment-scale because the median estimates of change from ensembles of GHMs and CHMs showed similar impacts of global warming on the considered hydrological indicators (see Table 3.3 for more details). There were some distinct differences though, such as a wider spread in impacts from the Glob-HMs for several catchments. The possible reasons are first the catchment models being calibrated (to specific catchment conditions) and second sampling bias introduced by the different sizes of the two ensembles leading the catchment models to perform better than the global ones.

Table 3.3. Key findings and summary of results obtained for Paper III.

RQ3: What are the relative performance gains for GHMs and CHMs (both in ensemble and MMC forms)?	
Obj4: Comparison of changes in river runoff from multiple global and catchment scale hydrological models under global warming scenarios of 1°C, 2°C and 3°C	
Key findings: • **The median estimates of runoff change from both ensembles generally showed similar effects of global warming, but the spreads are generally wider for the GHM ensemble.** • **Both ensembles show that Lena, Tagus and Rhine could see large changes in runoff with increasing warming. For these catchments, stringent mitigation action could have significant benefits in terms of avoiding some of the hydrological hazards. There are large spread in other catchments.** • **Decision-makers are encourages to still consider the full range of spreads from both ensembles.**	
1) Are there systematic differences between projections of runoff change from the two ensembles in simulating indicators of mean and extreme events (MAR, Q5, Q95, and runoff of different return period)?	**1) A) Statistical distributions of changes in MAR, Q5 and Q95 with global warming** - The median estimates from both ensembles generally show that there is similarity in the effect of global warming on the hydrological indicators considered for all global warming scenarios. - The spreads are generally wider for the GHM ensemble than the CHM ensemble for the majority of catchments, with occasional different sign of change between the two ensembles (e.g. in Upper Niger). **B) Probability of changes in MAR, Q5 and Q95 with global warming** - The probability of impacts of different magnitudes (as an indicator of the level of model agreement within an ensemble) occurring from the impact distributions of the two ensembles, indicate that the two ensembles show similar responses to global warming for all three hydrological indicators for the Ganges, Lena and Tagus. In other catchments, the agreement between the ensembles is strong, but not consistently across the three indicators. - The probabilities of large decreases (>20%) in low flows (Q95) are greater with the GHM ensemble than with the CHM ensemble for the majority (5) of catchments. - The probabilities of large increase (>20%) in high flows (Q5) are higher for the CHM ensemble in three catchments. **C) Changes in magnitude of return period runoff** - Both ensembles project changes that are similar in magnitude to each other in most cases; two exceptions are for the Upper Mississippi and Upper Niger.
2) What is the effect of different amounts of global-mean warming (1°C, 2°C and 3°C above pre-industrial levels) on GHMs and CHMs runoff in different catchments, spatially distributed?	2) **The impacts of global warming on each catchment** - Both ensembles clearly show that the Tagus and Rhine could see large increases in the magnitude of impacts with increasing warming by experiencing declines in the magnitude of low and high flows and MAR. - There is clear evidence for increases in MAR and high flows for the Lena and no conclusive effects of global warming on runoff for the other catchments. - There are large spread in Darling (as a challenging catchment to model) and Upper Amazon, Ganges, Upper Niger and Upper Mississippi (due to different approaches to simulate potential evapotranspiration by models affecting runoff simulation in these catchments). - For Rhine, Tagus and Lena it is clear that stringent global climate change mitigation action in line with Article 2 of the Paris Agreement (UNFCCC, 2015) could have significant benefits in terms of avoiding some of the hydrological hazards that could be seen with higher magnitudes of global warming such as 3°C.
3) Could any type of model ensemble be preferred than the other?	3) The median estimates of change from both ensembles generally showed similar effects of global warming on the considered hydrological indicators. Nonetheless, none of the ensemble types could be favoured against. This is because the outperformance of the CHM ensemble over the GHM ensemble in simulating present-day discharge does not guarantee higher credibility of projections under climate change (Hattermann et al., 2017). Therefore, decision-makers are encouraged to consider the full range of spreads from both ensembles.

With some catchments showing more compelling evidence of the impacts of climate change on runoff, the study also demonstrated consequences of missing the UNFCCC's 2°C target, and suggested that stringent global climate change mitigation action could be beneficial in terms of avoiding some of the hydrological hazards that could be seen with higher magnitudes of global warming such as 3°C.

Fulfilling the rationale of the study to enhance the knowledge of the relative performance and spreads in projections from the two ensembles that could help enhanced decision making, the results would not advise against favouring any one type of ensemble. For one reason, while the CHM ensemble has been shown to perform better than the GHM ensemble when simulating present-day discharge (Hattermann et al., 2017), this does not guarantee higher credibility of projections under climate change. On this basis, decision-makers are encouraged to consider the full range of spreads from both the ensembles, since these represent the spread of possible impacts according to several different types of models used by the hydrological modelling scientific community.

3.4 Application of MMC in filling the gap between GHMs and CHMs

Similar to previous cross-scale studies (Gosling et al., 2011, Thompson et al., 2013, Piniewski et al., 2014, Hattermann et al., 2017), findings of Paper III would advise parallel applications of GHMs and CHMs both due to noticeable difference in their spreads and the valuable information that each model type may provide. As a first study of its kind, Paper IV conducted a cross-scale model comparison to explore if results obtained from GHMs can be brought close(r) to those of CHMs by means of intelligent MMC.

At the first instance, the study revealed that generally CHMs and their EM demonstrate better performance than their counterpart GHMs and EM (see Table 3.4 for the results), conforming with findings of previous cross-scale studies (Gosling et al., 2011, Piniewski et al., 2014, Gosling et al., 2016, Hattermann et al., 2017). However, across both ensembles the EM presents weak performance where it outperforms the best individual GHM (CHM) in only 2 (3) catchments. This confirms the implication of Papers I and II that for future studies the ensemble mean is not used as the main (or only) summary of ensemble performance, when presenting the results from an ensemble of GHMs and/or CHMs, as an individual model or an MMC may perform better.

Conducting the intelligent MMC, while IPE scores are usually higher for the CHM ensemble, the MMC approach delivers more improvement over the GHMs or their EM. In addition, while all GHMs (except WaterGAP) and their EM failed to deliver any better performance against the benchmark model (their mean IPE across all catchments>0), all CHMs (except WaterGAP3) and their EM did obtain performance gain (their mean IPE across all catchments <0). Nonetheless, 3-component MMC for GHMs managed to build up significant improvement over weak individual GHMs by overall performing closer to 3-component MMC for CHMs (mean IPE of -1.95 against -2.37) than any individual GHM or their EM. 3-component MMC for GHMs is also, overall (averaged over all catchments), outperforming five individual CHMs (HBV, HYMOD, SWAT, VIC and WATERGAP3). Although, it should be acknowledged that different number of CHMs run across each catchment may bias the results.

When examined under 2-component settings, the MMC improved performance for the CHM ensemble but lowered that of the GHM ensemble contrasting the assumption that a more complex MMC will deliver a better performing model. This highlights that a more flexible and dynamic MMC approach could be used for future applications so that MMCs of different models and/or data types are developed accordingly. In addition, the inclusion of more input models to the MMC in the form of a super-ensemble of GHMs and CHMs proved to improve the performance of the applied MMC approach. Analysing the symbolic regression equations showed that both calibrated and uncalibrated GHMs, as well as poorer models from both ensembles are exploited in the MMC solutions. This could be because of the inherent capabilities that uncalibrated or poor models have to replicate certain aspects of the hydrological regime, and are, therefore, employed by the algorithm and should be kept in model ensembles.

It is also noteworthy that, generally, the MMCs improve the number of catchments where timing bias occurs, as well as the magnitude of the timing bias, which is always smaller than maximum (absolute) values recorded for any individual GHMs or CHMs. This, as analysis of the SR equation components in Paper I shows, could be explained a result of correction mechanism that the algorithm uses to adjust for magnitude and timing bias.

GHMs applied in Paper II and this paper vary in the important feature of the former being run with the inclusion of human impact. Although their MMCs are conducted with different data, the IPE scores of common GHMs in the two papers (e.g. DBH) suggest that GHMs with the inclusion of human impact parametrisation could deliver even better MMC solutions.

Paper II, acknowledged that the results reflect a fundamental error-complexity trade-off for optimal MMC selection that means a higher MMC performance gain could be achieved if MMCs of greater equation size were selected and the range of non-linear functions were increased. It has been sought to minimise the risk of selecting uninterpretable, over-fitted MMC solutions. It was then argued that further constraining the complexity of MMCs; either by limiting the equation size or by reducing the set of mathematical operators and functions available to the GEP algorithm might help increasing interpretability of the solutions, however, at the expense of MMC performance. Paper IV addressed parts of this argument by conducting 2-component MMC, but identifying the 'sweet spot' where both performance gain and interpretation is maximised still remains as an area for fruitful future research.

On all these bases, it could be argued that developing GHMs might be reasonable substitute for CHMs particularly when developing resource-intensive CHMs seems demanding. On the other hand, if both types of model ensemble are available, a pooled MMC of both types of models might be the best option. Future research suggested by this study is presented in Figure 3.3.

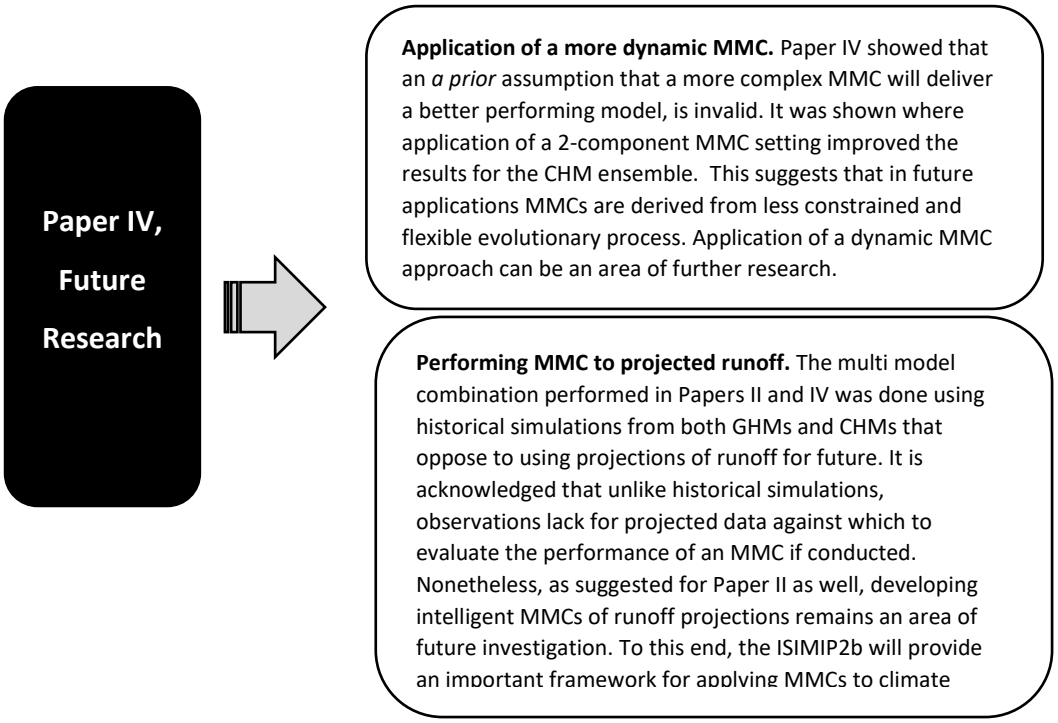


Figure 3.3. Future research suggested based on the study conducted in Paper IV.

Table 3.4. Key findings and summary of results obtained for Paper IV.

RQ3: What are the relative performance gains for GHMs and CHMs (both in ensemble and MMC forms)?	
Obj3: Multi-model combination with a super-ensemble of catchment-scale and global-scale hydrological models	
Key findings: • **MMC delivers results better than the EM for both GHM and CHM ensemble.** • **While IPE scores are usually higher for the CHM ensemble, the MMC approach delivers more improvement over the best GHMs or their EM.** • **A more complex MMC does not necessarily perform better since 2-component MMCs usually outperformed 3-component MMCs.** • **Generally, the inclusion of more input models, calibrated or uncalibrated, and both poorly and well performing models improves MMC performance.** • **Generally, the MMC approach improves the number of catchments where timing and magnitude bias occurs.**	
1) Do solutions provided by MMCs of CHMs and GHMs, perform better than their respective EMs?	1) Solutions provided by MMCs of CHMs and GHMs perform better than their respective EMs; their settings, however, might be different. The implication is that for future studies the ensemble mean is not used as the main (or only) summary of ensemble performance, when presenting the results from an ensemble of GHMs and/or CHMs, as an individual model or an MMC may perform better.
2) Does MMC improve the performance of one type of ensemble (i.e. comprised of CHMs or GHMs) more than the other type?	2) While IPE scores are usually higher for the CHM ensemble, the MMC approach delivers more improvement over the GHMs or their EM.
3) To what extent does the complexity of the MMC, i.e. the complexity of the weighting and combination of models, affect its performance?	3) An *a prior* assumption that a more complex MMC will deliver a better performing model, is invalid. It is recommended that in future applications, therefore, MMCs are either derived from an unconstrained evolutionary process, or a sensitivity analysis is conducted that involves different constraints, such as what presented here. 20
4) Does an MMC developed from a "super ensemble" of GHMs and CHMs outperform MMCs derived from an ensemble of each model type respectively?	4) Generally, the inclusion of more input models to the MMC (and constraining the MMC to 2 components) improves the performance of the MMC relative to including fewer input models. Moreover, the best performing solution from the series of experiments conducted here is on the whole for the 2-component MMC from the super-ensemble. This has important implications for future studies, for which it is recommended: - Studies seek to include as many individual models in their assessment as is feasible, with a means to weighting those models to deliver a MMC solution. - To include both calibrated and uncalibrated models as inputs to MMC solutions because the uncalibrated models may inherently replicate certain aspects of the hydrological regime that the calibrated models do not. - With MMC, value can be gained from even poorly performing individual models. For this reason, all the MMC solutions involve two or more GHMs. 5) Other important findings: - Generally, EMs and MMCs tend to overestimate MAR and Q95, but less Q5. Overestimation is more highlighted by in the EM of GHMs followed by 3-and 2-component MMC for CHMS across the three indicators. GHMs 3-component MMC tends to underestimate the indicators performing better that 3-component MMC for CHMs. - Generally, the MMCs improve the number of catchments where timing bias occurs, as well as the magnitude of the timing bias, which is always smaller than maximum (absolute) values recorded for any individual GHMs or CHMs. - In term of timing early bias, either 3-or 2-component MMC for GHMs show minimum (zero bias) for both maximum and minimum flows. Magnitude of timing late bias is, however, smaller for MMC of CHMs. Super ensemble, however does not improve timing bias over the MMCs of GHMs and CHMs in either case of early or late bias.

3.5 Should we use GHMs with human impact parametrisations (VARSOC)?

One of the main differences between the GHMs used in Papers I and II, and those applied in Papers III and IV was that the former use a time-variant parametrisations of human impacts (VARSOC) in their modelling algorithm, whereas the latter models are run without any human influence (NOSOC). Human impact was incorporated into the GHMs in the framework of ISIMIP2a in order, intuitively but without model experimentation, to improve models' performance. The effect of this was tested in Paper V as a first evaluation study of its kind where discharge simulations for five GHMs under a NOSOC and a VARSOC simulation were evaluated and compared. This was performed quantitatively and visually by distinguishing features such as change (improvement or otherwise) in evaluation metrics (e.g. bias ratio or variability), type of catchments (managed or (near-)natural, catchments spatial spread, and hydrological indicators studied (mean and extreme events in particular hydrological drought).

The main finding here was that, the inclusion of human activities in a modelling scheme generally enhances the ability of GHMs to estimate indicators of mean monthly discharge, and both high and low flows (Q1 and Q99), especially successful identification of hydrological drought events (Q90) (see Table 3.5 for results). However, whilst high-flows estimates mainly decreased, low-flow estimates showed mixed results, and the effect on the low-flow estimates is generally larger than on the high-flow estimates, both in managed and (near-)natural catchments. It was also found that whilst the human impact parametrisations predominantly leads to changes in simulated discharge in the managed catchments, a significant share of the (near-)natural catchments studied are influenced as well. However, the model improvement seen was neither ubiquitous nor consistent among the indicators. Examples of the existing challenges in the current GHMs were that:

- The majority of models failed to gain performance improvement in parts of the world (e.g. Amazon)
- There were tendencies of overestimation of long-term mean monthly discharge (seasonal cycle) across both (near-)natural and managed catchments, and
- There was a mixed pattern in estimating low flows after inclusion of HIP in the GHMs

Therefore, despite the improvements, the limited representation of long-term mean discharge and hydrological variability in managed basins restricts the overall performance of global models.

Paper V contributed to better explaining how changes observed in evaluation metrics (bias ratio, variability and correlation coefficient) for mean and extreme discharge after the inclusion of human impact parametrisations might be linked to different parts of the models' parametrisations in the VARSOC run. For instance, changes in the variability ratio are predominantly driven by the timing of water abstractions and return flows. In addition, based on the measure of overall evaluation (i.e. KGE) and differences in algorithms and parameterisations of models leading to their varying superiority in different indicators, the work highlighted modules of human impact parametrisation that could potentially be improved. For instance, better identification of sectoral water abstractions and return flows, and the share of built-up land area may result in clearer GHMs performance evaluation when distinguishing between managed and natural catchments. Moreover, the large variation in performance between GHMs, regions, and performance indicators, highlights, the importance of a careful selection of models, model components, and evaluation metrics, given varying research interests (e.g. drought or water scarcity). The systematic misrepresentation of hydrological extremes across all GHMs, especially in managed catchments, calls for a careful interpretation of risk assessments dealing with their results.

In summary, the fact that positive impacts of VARSOC runs outweigh their negative effects, rationalises the application of VARSOC data, and helps identify and link regions of lower GHMs performance with the human impact parametrisations in the models. However, the results also call for caution in relying fully upon the VARSOC run of the models and encourages acts of model improvement through improved structure and parametrisations that take into account human impacts. Future research suggested by this study is presented in Figure 3.4.

Table 3.5. Key findings and summary of results obtained for Paper V.

RQ4: Does the inclusion of HIP in GHMs improve models ensemble simulations?	
Obj5: Assessing the impact of inclusion of human activities on GHMs performance: a comparison of modelled discharges under a naturalised (NOSOC) and a time-dependent human impact (VARSOC) simulation of five GHMs.	
Key findings: • **The incorporation of human activities in GHMs parameterisations improves their ability to estimate runoff in 71% and 52% of studied managed and (near-)natural catchments respectively** • **The effect of human impact inclusion in GHMs is generally larger on low flows than high flows.** • **Despite the inclusion of human impacts, models perform significantly better in (near-)natural catchments, and impacts of human activities on both high and low flows are generally smaller in (near-)natural catchments compared to the managed catchments.** • **Significant improvements in performance for the majority of GHMs were found in Latin America, Southern Africa, and Northwest U.S., and significant decreases in (near-)natural area e.g. Amazon.**	
1) How does the incorporation of human impact affect representation of monthly discharges and their hydrological extremes, especially drought events?	1) The inclusion of human impact in the GHMs generally enhances the overall hydrological performance of GHMs, mainly driven by improvements in the bias ratio and variability ratio. - The ability of GHMs to model hydrological extremes in a substantial share of the global land area. The result is that whilst high-flows estimates mainly decreased, low-flow estimates showed mixed results. - The effect of human impact inclusion is generally larger on the low-flow estimates (Q_{99}) than on the high-flow estimates (Q_1), both in managed and (near-)natural catchments.
2) How does the inclusion of human impact affect simulations of hydrological indicators in (near-)natural and managed catchments?	2) The incorporation of human activities in GHMs parameterisations improves their ability to estimate runoff in respectively 71% and 52% of managed and natural catchments studied. - Whilst the human activities predominantly leads to changes in simulated discharge in the managed catchments, a significant share of the (near-)natural catchments studied are influenced as well. - Despite the inclusion of human activities, all models perform significantly better in (near-)natural catchments compared to the managed ones, both looking at the total KGE values and its sub-indicators. - Impacts of human activities on both high- and low-flows are generally smaller in (near-)natural catchments compared to the managed catchments
3) What is the spatial pattern of the changes caused by inclusion of HIP?	3) - Significant improvements for a majority of GHMs were found in Latin America, Southern Africa, and Northwest U.S. - Looking at the spatial explicit distribution of these results, only a limited number of locations globally where found that human impact led to significant decreases in overall hydrological performance for a majority of GHMs. Predominantly in (near-)natural areas (e.g. Amazon). - Looking at the results for the 12 sample catchments spatially distributed, it is clear that not a single model is best in all catchments or for all indicators
4) What model improvement opportunities does this study offer?	4) Looking at the hydrological factors studied and the evaluation tools applied, some model improvement could be suggested as follows: - Improving soil properties to improve generation of runoff, improvement of how reservoirs are incorporated to improve volumetric estimation of runoff, and improving both factors to improve timing of runoff simulations - Improvement of groundwater withdrawals parametrisation to improve estimation of need for surface water abstraction and consequently better stream flow estimation - Improvement of ET, routing and snow algorithm to improve generation of runoff, estimation of extreme events and components of KGE - Performing calibration to improve general model performance

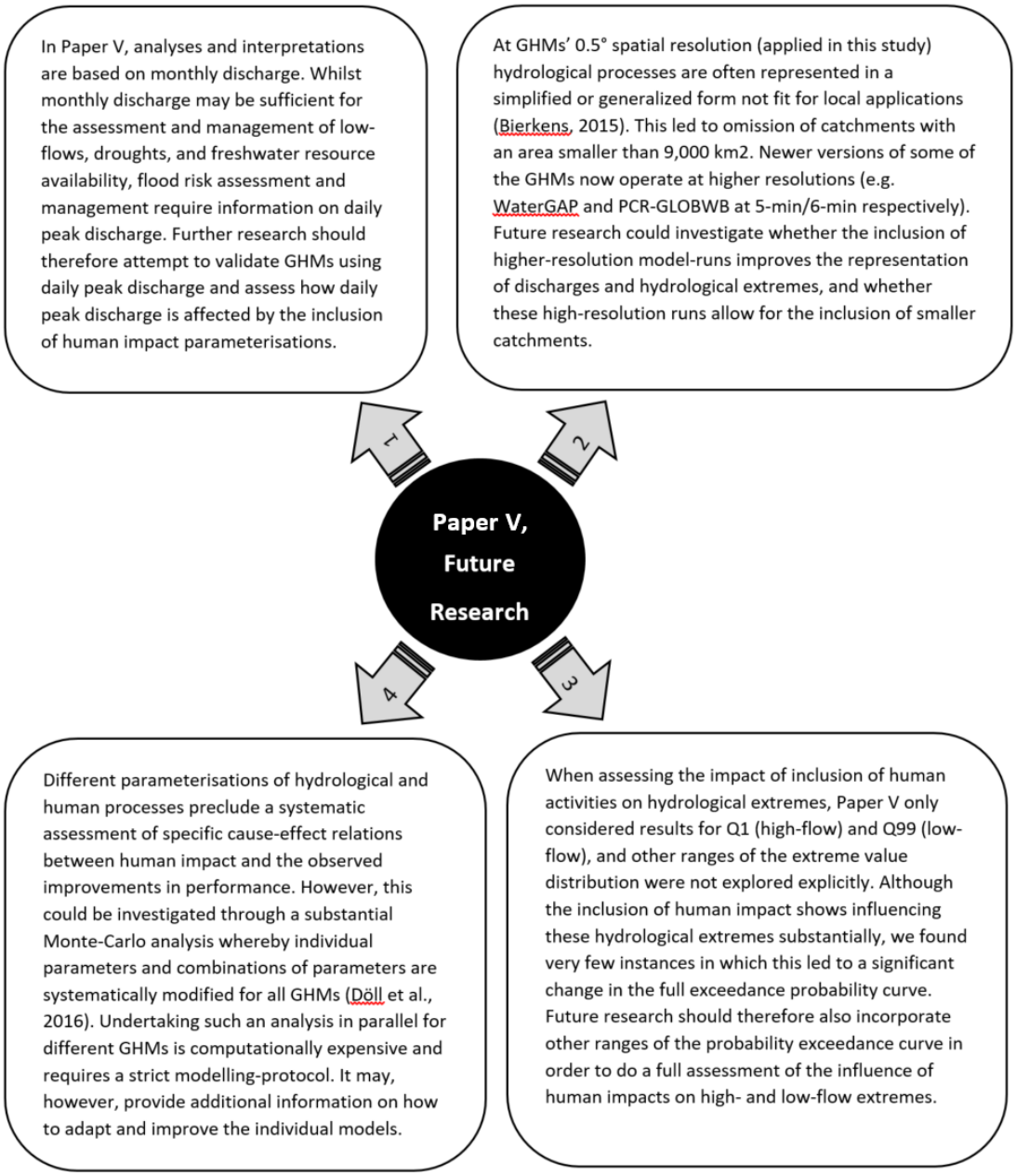


Figure 3.4. Future research suggested based on the study conducted in Paper V.

3.6 Conclusion

This thesis sought to explore whether the estimation of runoff simulations from ensembles of global and catchment hydrological models could be improved through application of a multi-model combination technique. This aim was addressed by employing different tools including the application of intelligent multi-model combination, cross-scale model comparisons, and incorporating new model parameterisations in models, as well as application of integrated evaluation metrics and a more geo-hydrological climate zoning system. The study was conducted and presented in the framework of five papers.

A comprehensive evaluation of six GHMs that include human impacts in their parametrisations was performed (Paper I) to explore characteristic strengths and shortcomings of the most state-of-the-art GHMs available. This revealed that GHM performance varies across hydrological indicators spatially so that none of the models is the best performing for all indicators or everywhere. The finding given the application of the hydrobelt notion for classifying the hydro-climate regions, highlightes opportunities for model improvement. This, however, highlights the need for a standard GHM evaluation framework, especially for GHMs within the same MIP. It was also found that, contrary to common assumption, the EM is *not* necessarily the best representative of a model ensemble. Care, therefore, must be taken in summarising the performance of an ensemble of GHMs by their EM, particularly in isolation. This finding also highlights the need for a more robust alternative than the EM.

MMC was performed in Paper II to explore whether an intelligently optimised combination of GHMs can improve runoff simulations. Going beyond equal weighting (as is the case with the EM) , the MLA-based MMC approach proved to be a successful alternative to the EM as it outperformed the EM in 95% of the study catchments. The explicit solutions offered by the intelligent MMC approach, made possible via using symbolic regression concepts, provided insights into the contribution of individual GHMs and revealed locations where specific GHMs offer greater performance in the MMC solutions. They also helped with interpreting the combination mechanisms. Nonetheless, failure of MMC in six catchments indicates that the technique is not a “silver bullet” and therefore it is suggested that both the MMC solution *and* EM are used along with the range of model ensemble, and not in isolation in future studies. The study also suggested several areas of further study on improving the MMC skill (via improved methodologies), application of MMC with global coverage, and the applicability of the MMC technique in other sectors than water.

Cross-scale studies in Papers III and IV showed signs that the performance of GHMs and CHMs are starting to converge and some gaps in their performance might be filled via techniques such as MMC. Uncommon with previous cross-scale studies, the application of a large ensemble of GHMs and CHMs respectively facilitated a more robust evaluation and performance comparison of the two model types. As presented in Paper III, the median estimates of runoff change from both ensembles generally showed similar effects of global warming, but the spreads are generally wider for the GHM ensemble. There were also similar impacts across the hydrological indicators studied in some of the study catchments. Despite these, the main conclusion of this study was that when studying projections of runoff

(change), one should consider the full range of spreads from both ensembles as different tools of hydrological studies.

Application of the intelligent MMC to ensembles of GHMs and CHMs (Paper IV) was successful in outperforming their EM and best individual model of each ensemble, but the technique showed higher skill across the GHM ensemble. MMC solutions also succeeded in bringing simulations from the two model scales closer together in terms of the overall performance. One unconventional finding of this study was that a more complex MMC does not necessarily perform better, since 2-gene MMCs usually outperformed 3-gene MMCs. In addition, the inclusion of more input models to the MMC, generally, improves the performance of the MMC relative to including fewer input models, hence, a super-ensemble comprising all GHMs and CHMs showed promising results. This highlighted another unconventional finding that value can be gained from even poorly performing individual models, as well as uncalibrated models. The implication is, therefore, to include both poor and uncalibrated models. Further, the MMC approach generally reduces the number of catchments where timing and magnitude bias were seen in the seasonal cycle.

Finally, whether and to what extent inclusion of human activities in GHMs improves their performance in simulation of observed runoff was evaluated in Paper V. This study revealed improvement in the ability of GHMs to estimate runoff, however, that was heterogeneous across the globe. In fact, significant performance improvement was seen for the majority of GHMs in Latin America, Southern Africa, and Northwest U.S., and significant decreases in (near-)natural area e.g. Amazon. In addition, performance gain differs across the managed and (near-)natural catchments (in 71% and 52% of their studied catchments respectively), and across the hydrological indicators. Despite the inclusion of human impacts, models perform significantly better in (near-)natural catchments, and impacts of human activities on both high and low flows are generally smaller in (near-)natural catchments compared to the managed catchments. Making linkage between the findings about the GHMs performance their characteristics, suggested future opportunities model parametrisation and structural improvement.

3.7 Current status of the papers, and statement on authorship

This section provides the important submission information and status of each paper in the publication process (Figure 3.5 and Table 3.6). As summarised in Figure 3.5 and Table 3.6, Papers I, III and V are published, and Papers II and IV are submitted and therefore in review

stage. For the published paper(s) there may be minor discrepancies between the current manuscript and the published version resulting from the copy-editing process. The papers' format complies with the format in which they are submitted. Where possible, Supplementary Information (SI) is provided for each paper at the end of the relevant paper, otherwise in the Appendices section. For the submitted paper(s) for which I have received reviewers' comments and replied to, the response letter is also included in the Appendices section.

Author contributions are included towards the end of each paper and also appear in the published versions, where applicable. I was the lead author on three of the papers (Papers I, II and IV), second author on Paper III and one of the co-authors on Paper V. My contribution to Paper III include all data preparation and coding activities, and commenting through the revision process. In Paper V, I contributed towards the provision of MATLAB codes for data processing and analysis, as well as contributing towards the text of the manuscript through comments, suggestions and guidance through the whole preparation and submission process. Since the data used in Papers I and II, as well as Paper V were not publically available at the time of preparing these papers, as a general rule set by the ISIMIP coordinators and modellers, any papers applying those data needed to invite the modellers as co-authors. Their contribution, however, was not limited to providing simulated data; they were involved in providing comments and guidance on the manuscripts in different stages. Except to the order of the first three authors in Papers I, II and IV, the rest of the co-authors are listed alphabetically for their contribution being more or less equal. Styles may vary according to the journal/target journal.

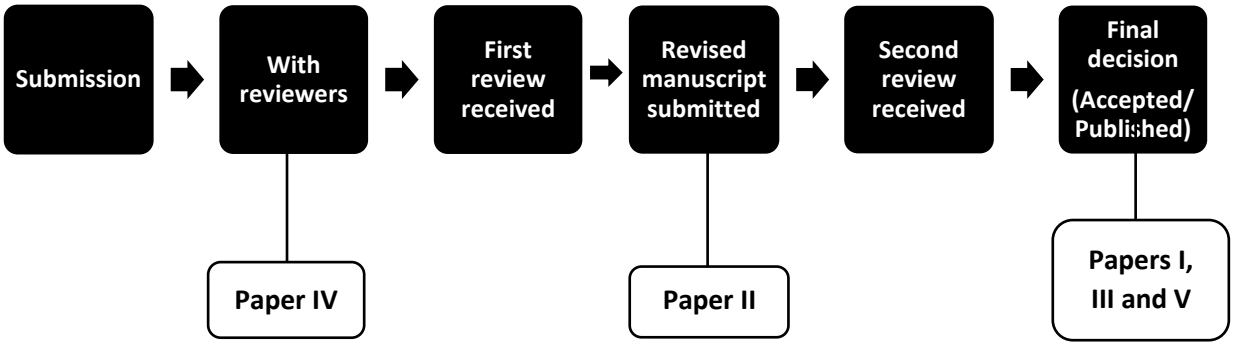


Figure 3.5. The status of the five research papers in the submission process.

Table 3.6. Submission information and the status of the five research papers.

Paper № in Thesis	Title	Authors	Status (Date accepted)	Journal or (Target Journal)	Impact factor (year)	Page № in Thesis
I	Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human-impacts.	**Jamal Zaherpour**, Simon N Gosling, Nick Mount, Hannes M Schmied, Ted I E Veldkamp, Rutger Dankers, Stephanie Eisner, Dieter Gerten, Lukas Gudmundsson, Ingjerd Haddeland, Naota Hanasaki, Hyungjun Kim, Guoyong Leng, Junguo Liu, Yoshimitsu Masaki, Taikan Oki, Yadu Pokhrel, Yusuke Satoh, Jacob Schewe and Yoshihide Wada	Published (2018)	Environmental Research Letters	4.404 (2017)	78
II	Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models.	**Jamal Zaherpour**, Nick Mount, Simon N Gosling, Dieter Gerten, Hannes Müller Schmied, Qiuhong Tang, Rutger Dankers, Stephanie Eisner, Xingcai Liu, Yoshihide Wada , Yoshimitsu Masaki	In review	Environmental Modelling and Software	4.404 (2016)	79
III	A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1°C, 2°C and 3°C. DOI: 10.1007/s10584-016-1773	Simon N. Gosling, **Jamal Zaherpour**, Nick Mount, Fred F. Hattermann, Rutger Dankers, Berit Arheimer, Lutz Breuer, Jie Ding, Ingjerd Haddeland, Rohini Kumar, Dipangkar Kundu, Junguo Liu, Ann van Griensven, Ted I.E. Veldkamp, Tobias Vetter, Xiaoyan Wang, Xinxin Zhang	Published (2016)	Climatic Change	3.496 (2016)	80
IV	Multi-model combination with a super-ensemble of catchment-scale and global-scale hydrological models	**Jamal Zaherpour**, Simon N Gosling, Nick Mount, Fred Hattermann	In review	Journal of Hydrology	3.727 (2017)	81
V	Human impact parameterisation in global hydrological models improves estimates of monthly discharges and hydrological extremes: a multi-model validation study. DOI: 10.1088/1748-9326/aab96f	Ted I E Veldkamp, Fang Zhao, Philip J Ward, Hans de Moel, Jeroen C.J.H., Aerts, Hannes Müller Schmied, Felix T Portmann, Yoshimitsu Masaki, Yadu Pokhrel, Xingcai. Liu, Yusuke Satoh, Dieter Gerten, Simon N. Gosling, **Jamal Zaherpour**, Yoshihide Wada	Published (2018)	Environmental Research Letters	4.404 (2017)	82

4 References for the extended abstract

ABRAHART, R. J., MOUNT, N. J., AB GHANI, N., CLIFFORD, N. J. & DAWSON, C. W. 2011. DAMP: A protocol for contextualising goodness-of-fit statistics in sediment-discharge data-driven modelling. *Journal of Hydrology,* 409**,** 596-611.

ABRAHART, R. J. & SEE, L. 2002. Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments. *Hydrology and Earth System Sciences,* 6**,** 655-670.

AJAMI, N. K., DUAN, Q. Y., GAO, X. G. & SOROOSHIAN, S. 2006. Multimodel combination techniques for analysis of hydrological simulations: Application to Distributed Model Intercomparison Project results. *Journal of Hydrometeorology,* 7**,** 755-768.

AJAMI, N. K., DUAN, Q. Y. & SOROOSHIAN, S. 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research,* 43.

ALCAMO, J., DOLL, P., KASPAR, F. & SIEBERT, S. 1997. Global change and global scenarios of water use and availability: An application of WaterGAP 1.0., . Rep. A9701, Cent. for Environ. Syst. Res., Univ. of Kassel, Kassel, Germany.

ALCAMO, J., FLÖRKE, M. & MÄRKER, M. 2007. Future long-term changes in global water resources driven by socio-economic and climatic changes. *Hydrological Sciences Journal,* 52**,** 247-275.

ALFIERI, L., BUREK, P., DUTRA, E., KRZEMINSKI, B., MURARO, D., THIELEN, J. & PAPPENBERGER, F. 2013. GloFAS – global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences,* 17**,** 1161-1175.

ARNELL, N. W., BROWN, S., GOSLING, S. N., HINKEL, J., HUNTINGFORD, C., LLOYD-HUGHES, B., LOWE, J. A., OSBORN, T., NICHOLLS, R. J. & ZELAZOWSKI, P. 2014. Global-scale climate impact functions: the relationship between climate forcing and impact. *Climatic Change,* 134**,** 475-487.

ARNELL, N. W. & GOSLING, S. N. 2016. The impacts of climate change on river flood risk at the global scale. *Climatic Change,* 134**,** 387-401.

ARSENAULT, R., GATIEN, P., RENAUD, B., BRISSETTE, F. & MARTEL, J. L. 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. *Journal of Hydrology,* 529**,** 754-767.

AYTEK, A., ASCE, M. & ALP, M. 2008. An application of artificial intelligence for rainfall-runoff modeling. *Journal of Earth System Science,* 117**,** 145-155.

AZMI, M., ARAGHINEJAD, S. & KHOLGHI, M. 2010. Multi Model Data Fusion for Hydrological Forecasting Using K-Nearest Neighbour Method. *Iranian Journal of Science and Technology Transaction B-Engineering,* 34**,** 81-92.

BARBULESCU, A. & BAUTU, E. 2010. Mathematical models of climate evolution in Dobrudja. *Theoretical and Applied Climatology,,* 100.

BĂRBULESCU, A. & BĂUTU, E. 2009. Time Series Modeling Using an Adaptive Gene Expression Programming Algorithm. *International journal of mathematical models and methods in applied sciences,* 3**,** 85-93.

BATES, J. M. & GRANGER, C. W. J. 1969. The comination of forecasts. *Operational research,* 20**,** 451-468.

BECK, H. E., VAN DIJK, A. I. J. M., DE ROO, A., DUTRA, E., FINK, G., ORTH, R. & SCHELLEKENS, J. 2016. Global evaluation of runoff from ten state-of-the-art hydrological models. *Hydrology and Earth System Sciences Discussions***,** 1-33.

BENNETT, N. D., CROKE, B. F. W., GUARISO, G., GUILLAUME, J. H. A., HAMILTON, S. H., JAKEMAN, A. J., LIBELLI, S. M., NEWHAMA, L. T. H., NORTON, J. P., PERRIN, C., PIERCE, S. A., ROBSON, B., SEPPELT, R., VOINOV, A. A., FATH, B. D. & ANDREASSIAN, V. 2013. Characterising performance of environmental models. *Environmental Modelling & Software,* 40**,** 1-20.

BERIRO, D. 2015. *Gene Expression Programming Models of Bioaccessible Benzo[a]pyrene in Coking Works Soils.* PhD, Nottingham.

BERIRO, D. J., ABRAHART, R. J., NATHANAIL, C. P., MORENO, J. & BAWAZIR, A. S. 2013. A typology of different development and testing options for symbolic regression modelling of measured and calculated datasets. *Environmental Modelling & Software,* 47**,** 29-41.

BEST, M. J., PRYOR, M., CLARK, D. B., ROONEY, G. G., ESSERY, R. L. H., MÉNARD, C. B., EDWARDS, J. M., HENDRY, M. A., PORSON, A., GEDNEY, N., MERCADO, L. M., SITCH, S., BLYTH, E., BOUCHER, O., COX, P. M., GRIMMOND, C. S. B. & HARDING, R. J. 2011. The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes. *Geoscientific Model Development,* 4**,** 677-699.

BIERKENS, M. F. P. 2015. Global hydrology 2015: State, trends, and directions. *Water Resources Research,* 51**,** 4923-4947.

BONDEAU, A., SMITH, P. C., ZAEHLE, S., SCHAPHOFF, S., LUCHT, W., CRAMER, W., GERTEN, D., LOTZE-CAMPEN, H., MULLER, C., REICHSTEIN, M. & SMITH, B. 2007. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biology,* 13**,** 679-706.

BROWN, C., BAROANG, K. M., CONRAD, E., LYON, B., WATKINS, D., FIONDELLA, F., KAHEIL, Y., ROBERTSON, A., RODRIGUEZ, J., SHEREMATA, M. & WARD, M. N. 2010. Managing Climate Risk in Water Supply Systems.

BURKE, E. J., BROWN, S. J. & CHRISTIDIS, N. 2006. Modeling the recent evolution of global drought and projections for the twenty-first century with the hadley centre climate model. *Journal of Hydrometeorology,* 7**,** 1113-1125.

CHANDLER, R. E. 2013. Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Philos Trans A Math Phys Eng Sci,* 371**,** 20120388.

CLARK, D. B., MERCADO, L. M., SITCH, S., JONES, C. D., GEDNEY, N., BEST, M. J., PRYOR, M., ROONEY, G. G., ESSERY, R. L. H., BLYTH, E., BOUCHER, O., HARDING, R. J., HUNTINGFORD, C. & COX, P. M. 2011. The Joint UK Land Environment Simulator (JULES), model description – Part 2: Carbon fluxes and vegetation dynamics. *Geoscientific Model Development,* 4**,** 701-722.

CLEMEN, R. T. 1989. COMBINING FORECASTS - A REVIEW AND ANNOTATED-BIBLIOGRAPHY. *International Journal of Forecasting,* 5**,** 559-583.

CLOKE, H. L. & PAPPENBERGER, F. 2009. Ensemble flood forecasting: A review. *Journal of Hydrology,* 375**,** 613-626.

DANKERS, R., ARNELL, N. W., CLARK, D. B., FALLOON, P. D., FEKETE, B. M., GOSLING, S. N., HEINKE, J., KIM, H., MASAKI, Y., SATOH, Y., STACKE, T., WADA, Y. & WISSER, D. 2014. First look at changes in flood hazard in the Inter-Sectoral Impact Model Intercomparison Project ensemble. *Proc Natl Acad Sci U S A,* 111**,** 3257-61.

DAWSON, C. W., ABRAHART, R. J. & SEE, L. M. 2007a. HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling & Software,* 22**,** 1034-1052.

DAWSON, C. W., ABRAHART, R. J. & SEE, L. M. 2007b. HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling & Software,* 22**,** 1034-1052.

DAWSON, C. W., MOUNT, N. J., ABRAHART, R. J. & SHAMSELDIN, A. Y. 2012. Ideal point error for model assessment in data-driven river flow forecasting. *Hydrology and Earth System Sciences,* 16**,** 3049-3060.

DE GRAAF, I. E. M., VAN BEEK, L. P. H., WADA, Y. & BIERKENS, M. F. P. 2014. Dynamic attribution of global water demand to surface water and groundwater resources: Effects of abstractions and return flows on river discharges. *Advances in Water Resources,* 64**,** 21-33.

DE MENEZES, L. M., BUNN, D. W. & TAYLOR, J. W. 2000. Review of guidelines for the use of combined forecasts. *European Journal of Operational Research,* 120**,** 190-204.
DECHARME, B. & DOUVILLE, H. 2006. Uncertainties in the GSWP-2 precipitation forcing and their impacts on regional and global hydrological simulations. *Climate Dynamics,* 27**,** 695-713.
DECHARME, B. & DOUVILLE, H. 2007. Global validation of the ISBA sub-grid hydrology. *Climate Dynamics,* 29**,** 21-37.
DIRMEYER, P. A., GAO, X., ZHAO, M., GUO, Z., OKI, T. & HANASAKI, N. 2006. GSWP-2: Multimodel Analysis and Implications for Our Perception of the Land Surface. *Bulletin of the American Meteorological Society,* 87**,** 1381-1397.
DÖLL, P., HOFFMANN-DOBREV, H., PORTMANN, F. T., SIEBERT, S., EICKER, A., RODELL, M., STRASSBERG, G. & SCANLON, B. R. 2012. Impact of water withdrawals from groundwater and surface water on continental water storage variations. *Journal of Geodynamics,* 59-60**,** 143-156.
ELLIOTT, J., DERYNG, D., MULLER, C., FRIELER, K., KONZMANN, M., GERTEN, D., GLOTTER, M., FLORKE, M., WADA, Y., BEST, N., EISNER, S., FEKETE, B. M., FOLBERTH, C., FOSTER, I., GOSLING, S. N., HADDELAND, I., KHABAROV, N., LUDWIG, F., MASAKI, Y., OLIN, S., ROSENZWEIG, C., RUANE, A. C., SATOH, Y., SCHMID, E., STACKE, T., TANG, Q. & WISSER, D. 2014. Constraints and potentials of future irrigation water availability on agricultural production under climate change. *Proc Natl Acad Sci U S A,* 111**,** 3239-44.
EMERTON, R. E., STEPHENS, E. M., PAPPENBERGER, F., PAGANO, T. C., WEERTS, A. H., WOOD, A. W., SALAMON, P., BROWN, J. D., HJERDT, N., DONNELLY, C., BAUGH, C. A. & CLOKE, H. L. 2016. Continental and global scale flood forecasting systems. *Wiley Interdisciplinary Reviews-Water,* 3**,** 391-418.
ENTIN, J. K., ROBOCK, A., VINNIKOV, K. Y., ZABELIN, V., LIU, S., NAMKHAI, A. & ADYASUREN, T. 1999. Evaluation of Global Soil Wetness Project Soil Moisture Simulations. *Journal of the Meteorological Society of Japan. Ser. II,* 77**,** 183-198.
FERNANDO, A. K., SHAMSELDIN, A. Y. & ABRAHART, R. J. 2012. Use of Gene Expression Programming for Multimodel Combination of Rainfall-Runoff Models. *Journal of Hydrologic Engineering,* 17**,** 975-985.
FERREIRA, C. 2001. Gene Expression Programming: A New AdaptiveAlgorithm for Solving Problems. 13**,** 87-129.
FERREIRA, C. 2006. *Gene Expression Programming: Mathematical Model-ing by an Artificial Intelligence,* Verlag: Berlin, Springer.
FIELD, C. B., BARROS, V., STOCKER, T. F., QIN, D., DOKKEN, D. J., EBI, K. L., MASTRANDREA, M. D., MACH, K. J., PLATTNER, G.-K., ALLEN, S. K., TIGNOR, M. & (EDS.), A. P. M. M. 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change.
FOWLER, H. J. & EKSTRÖM, M. 2009. Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes. *International Journal of Climatology,* 29**,** 385-416.
FRASER, E. D. G., SIMELTON, E., TERMANSEN, M., GOSLING, S. N. & SOUTH, A. 2013. "Vulnerability hotspots": Integrating socio-economic and hydrological models to identify where cereal production may decline in the future due to climate change induced drought. *Agricultural and Forest Meteorology,* 170**,** 195-205.
FREITAS, P. S. A. & RODRIGUES, A. J. L. 2006. Model combination in neural-based forecasting. *European Journal of Operational Research,* 173**,** 801-814.
FRIELER, K., BETTS, R., BURKE, E., CIAIS, P., DENVIL, S., DERYNG, D., EBI, K., EDDY, T., EMANUEL, K., ELLIOTT, J., GALBRAITH, E., GOSLING, S. N., HALLADAY, K., HATTERMANN, F., HICKLER, T., HINKEL, J., HUBER, V., JONES, C., KRYSANOVA, V.,

LANGE, S., LOTZE, H. K., LOTZE-CAMPEN, H., MENGEL, M., MOURATIADOU, I., MÜLLER SCHMIED, H., OSTBERG, S., PIONTEK, F., POPP, A., REYER, C. P. O., SCHEWE, J., STEVANOVIC, M., SUZUKI, T., THONICKE, K., TIAN, H., TITTENSOR, D. P., VAUTARD, R., VAN VLIET, M., WARSZAWSKI, L. & ZHAO, F. 2016. Assessing the impacts of 1.5 °C global warming – simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b). *Geoscientific Model Development Discussions***,** 1-59.

GAO, X. & DIRMEYER, P. A. 2006. A multimodel analysis, validation, and transferability study of global soil wetness products. *Journal of Hydrometeorology,* 7**,** 1218-1236.

GEORGAKAKOS, K. P., SEO, D. J., GUPTA, H., SCHAAKE, J. & BUTTS, M. B. 2004. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology,* 298**,** 222-241.

GERTEN, D., SCHAPHOFF, S., HABERLANDT, U., LUCHT, W. & SITCH, S. 2004. Terrestrial vegetation and water balance—hydrological evaluation of a dynamic global vegetation model. *Journal of Hydrology,* 286**,** 249-270.

GIORGI, F. & MEARNS, L. O. 2002. Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging'' (REA) method. *Journal of Climate,* 15**,** 1141-1158.

GLEESON, T., SMITH, L., MOOSDORF, N., HARTMANN, J., DÜRR, H. H., MANNING, A. H., VAN BEEK, L. P. H. & JELLINEK, A. M. 2011. Mapping permeability over the surface of the Earth. *Geophysical Research Letters,* 38**,** n/a-n/a.

GOSLING, S. N. & ARNELL, N. W. 2011. Simulating current global river runoff with a global hydrological model: model revisions, validation, and sensitivity analysis. *Hydrological Processes,* 25**,** 1129-1145.

GOSLING, S. N. & ARNELL, N. W. 2016. A global assessment of the impact of climate change on water scarcity. *Climatic Change,* 134**,** 371-385.

GOSLING, S. N., BRETHERTON, D., HAINES, K. & ARNELL, N. W. 2010. Global hydrology modelling and uncertainty: running multiple ensembles with a campus grid. *Philos Trans A Math Phys Eng Sci,* 368**,** 4005-21.

GOSLING, S. N., TAYLOR, R. G., ARNELL, N. W. & TODD, M. C. 2011. A comparative analysis of projected impacts of climate change on river runoff from global and catchment-scale hydrological models. *Hydrology and Earth System Sciences,* 15**,** 279-294.

GOSLING, S. N., ZAHERPOUR, J., MOUNT, N. J., HATTERMANN, F. F., DANKERS, R., ARHEIMER, B., BREUER, L., DING, J., HADDELAND, I., KUMAR, R., KUNDU, D., LIU, J., VAN GRIENSVEN, A., VELDKAMP, T. I. E., VETTER, T., WANG, X. & ZHANG, X. 2016. A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C. *Climatic Change,* 141**,** 577-595.

GUDMUNDSSON, L., TALLAKSEN, L. M., STAHL, K., CLARK, D. B., DUMONT, E., HAGEMANN, S., BERTRAND, N., GERTEN, D., HEINKE, J., HANASAKI, N., VOSS, F. & KOIRALA, S. 2012a. Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe. *Journal of Hydrometeorology,* 13**,** 604-620.

GUDMUNDSSON, L., WAGENER, T., TALLAKSEN, L. M. & ENGELAND, K. 2012b. Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe. *Water Resources Research,* 48.

GUIMBERTEAU, M., DRAPEAU, G., RONCHAIL, J., SULTAN, B., POLCHER, J., MARTINEZ, J. M., PRIGENT, C., GUYOT, J. L., COCHONNEAU, G., ESPINOZA, J. C., FILIZOLA, N., FRAIZY, P., LAVADO, W., DE OLIVEIRA, E., POMBOSA, R., NORIEGA, L. & VAUCHEL, P. 2012. Discharge simulation in the sub-basins of the Amazon using ORCHIDEE forced by new datasets. *Hydrology and Earth System Sciences,* 16**,** 911-935.

GUO, Z. & DIRMEYER, P. A. 2006. Evaluation of the Second Global Soil Wetness Project soil moisture simulations: 1. Intermodel comparison. *Journal of Geophysical Research,* 111.

GUO, Z. C., DIRMEYER, P. A., GAO, X. & ZHAO, M. 2007. Improving the quality of simulated soil moisture with a multi-model ensemble approach. *Quarterly Journal of the Royal Meteorological Society,* 133**,** 731-747.

GUPTA, H. V., KLING, H., YILMAZ, K. K. & MARTINEZ, G. F. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology,* 377**,** 80-91.

GUPTA, H. V., SOROOSHIAN, S. & YAPO, P. O. 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *WATER RESOURCES RESEARCH,* 34.

H, C. J., E, K., F, G., G, L. & M, R. 2010. Weight assignment in regional climate models. *CLIMATE RESEARCH,* 44**,** 179–194.

HADDELAND, I., CLARK, D. B., FRANSSEN, W., LUDWIG, F., VOSS, F., ARNELL, N. W., BERTRAND, N., BEST, M., FOLWELL, S., GERTEN, D., GOMES, S., GOSLING, S. N., HAGEMANN, S., HANASAKI, N., HARDING, R., HEINKE, J., KABAT, P., KOIRALA, S., OKI, T., POLCHER, J., STACKE, T., VITERBO, P., WEEDON, G. P. & YEH, P. 2011. Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results. *Journal of Hydrometeorology,* 12**,** 869-884.

HADDELAND, I., HEINKE, J., BIEMANS, H., EISNER, S., FLORKE, M., HANASAKI, N., KONZMANN, M., LUDWIG, F., MASAKI, Y., SCHEWE, J., STACKE, T., TESSLER, Z. D., WADA, Y. & WISSER, D. 2014. Global water resources affected by human interventions and climate change. *Proc Natl Acad Sci U S A,* 111**,** 3251-6.

HADDELAND, I., SKAUGEN, T. & LETTENMAIER, D. P. 2006. Anthropogenic impacts on continental surface water fluxes. *Geophysical Research Letters,* 33.

HAGEMANN, S., CHEN, C., CLARK, D. B., FOLWELL, S., GOSLING, S. N., HADDELAND, I., HANASAKI, N., HEINKE, J., LUDWIG, F., VOSS, F. & WILTSHIRE, A. J. 2013. Climate change impact on available water resources obtained using multiple global climate and hydrology models. *Earth System Dynamics,* 4**,** 129-144.

HANASAKI, N., FUJIMORI, S., YAMAMOTO, T., YOSHIKAWA, S., MASAKI, Y., HIJIOKA, Y., KAINUMA, M., KANAMORI, Y., MASUI, T., TAKAHASHI, K. & KANAE, S. 2013. A global water scarcity assessment under Shared Socio-economic Pathways - Part 2: Water availability and scarcity. *Hydrology and Earth System Sciences,* 17**,** 2393-2413.

HANASAKI, N., KANAE, S. & OKI, T. 2006. A reservoir operation scheme for global river routing models. *Journal of Hydrology,* 327**,** 22-41.

HANASAKI, N., KANAE, S., OKI, T., MASUDA, K., MOTOYA, K., SHIRAKAWA, N., SHEN, Y. & TANAKA, K. 2008. An integrated model for the assessment of global water resources –Part 2: Applications and assessments. *Hydrol. Earth Syst. Sci.,* 12**,** 1027-1037.

HARDING, R., BEST, M., BLYTH, E., HAGEMANN, S., KABAT, P., TALLAKSEN, L. M., WARNAARS, T., WIBERG, D., WEEDON, G. P., LANEN, H. V., LUDWIG, F. & HADDELAND, I. 2011. WATCH: Current Knowledge of the Terrestrial Global Water Cycle. *Journal of Hydrometeorology,* 12**,** 1149-1156.

HATTERMANN, F. F., KRYSANOVA, V., GOSLING, S. N., DANKERS, R., DAGGUPATI, P., DONNELLY, C., FLÖRKE, M., HUANG, S., MOTOVILOV, Y., BUDA, S., YANG, T., MÜLLER, C., LENG, G., TANG, Q., PORTMANN, F. T., HAGEMANN, S., GERTEN, D., WADA, Y., MASAKI, Y., ALEMAYEHU, T., SATOH, Y. & SAMANIEGO, L. 2017. Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins. *Climatic Change,* 141**,** 561-576.

HATTERMANN, F. F., KRYSANOVA, V., HABECK, A. & BRONSTERT, A. 2006. Integrating wetlands and riparian zones in river basin modelling. *Ecological Modelling,* 199**,** 379-392.

HELSEL, D. R. & HIRSCH, R. M. 2002. Techniques of Water-Resources Investigations of the United States Geological Survey.

HEMPEL, S., FRIELER, K., WARSZAWSKI, L., SCHEWE, J. & PIONTEK, F. 2013. A trend-preserving bias correction – the ISI-MIP approach. *Earth Syst. Dynam.,* 4**,** 219-236.

HENDERSON-SELLERS, A., PITMAN, A. J., LOVE, P. K., IRANNEJAD, P. & CHEN, T. H. 1995. The Project for Intercomparison of Land Surface Parameterization Schemes (PILPS): Phases 2 and 3. *Bulletin of the American Meteorological Society,* 76**,** 489-503.

HIRABAYASHI, Y., MAHENDRAN, R., KOIRALA, S., KONOSHIMA, L., YAMAZAKI, D., WATANABE, S., KIM, H. & KANAE, S. 2013. Global flood risk under climate change. *Nature Climate Change,* 3**,** 816-821.

HTTP://WWW.EWATER.COM.AU. 2014. *UC Innovation Centre, Building 22 University Drive South Bruce, ACT, 2617, Australia* [Online].

HUANG, S., KUMAR, R., FLÖRKE, M., YANG, T., HUNDECHA, Y., KRAFT, P., GAO, C., GELFAN, A., LIERSCH, S., LOBANOVA, A., STRAUCH, M., VAN OGTROP, F., REINHARDT, J., HABERLANDT, U. & KRYSANOVA, V. 2016. Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide. *Climatic Change,* 141**,** 381-397.

JAIN, S. K. & SUDHEER, K. P. 2008. Fitting of hydrologic models: A close look at the Nash-Sutcliffe index. *Journal of Hydrologic Engineering,* 13**,** 981-986.

JEONG, D. & KIM, Y.-O. 2009. Combining single-value streamflow forecasts - A review and guidelines for selecting techniques. *Journal of Hydrology,* 377**,** 284-299.

KLING, H., FUCHS, M. & PAULIN, M. 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology,* 424-425**,** 264-277.

KOCH, H., LIERSCH, S. & HATTERMANN, F. F. 2013. Integrating water resources management in eco-hydrological modelling. *Water Sci Technol,* 67**,** 1525-33.

KOSTER, R. D., FEKETE, B. M., HUFFMAN, G. J. & STACKHOUSE, P. W. 2006a. Revisiting a hydrological analysis framework with International Satellite Land Surface Climatology Project Initiative 2 rainfall, net radiation, and runoff fields. *Journal of Geophysical Research-Atmospheres,* 111.

KOSTER, R. D., SUD, Y. C., GUO, Z., DIRMEYER, P. A., BONAN, G., OLESON, K. W., CHAN, E., VERSEGHY, D., COX, P., DAVIES, H., KOWALCZYK, E., GORDON, C. T., KANAE, S., LAWRENCE, D., LIU, P., MOCKO, D., LU, C.-H., MITCHELL, K., MALYSHEV, S., MCAVANEY, B., OKI, T., YAMADA, T., PITMAN, A., TAYLOR, C. M., VASIC, R. & XUE, Y. 2006b. GLACE: The Global Land–Atmosphere Coupling Experiment. Part I: Overview. *Journal of Hydrometeorology,* 7**,** 590-610.

KRYSANOVA, V. & HATTERMANN, F. F. 2017. Intercomparison of climate change impacts in 12 large river basins: overview of methods and summary of results. *Climatic Change,* 141**,** 363-379.

LAWRENCE, D. M., OLESON, K. W., FLANNER, M. G., THORNTON, P. E., SWENSON, S. C., LAWRENCE, P. J., ZENG, X., YANG, Z.-L., LEVIS, S., SAKAGUCHI, K., BONAN, G. B. & SLATER, A. G. 2011. Parameterization improvements and functional and structural advances in Version 4 of the Community Land Model. *Journal of Advances in Modeling Earth Systems,* 3.

LIANG, X., LETTENMAIER, D. P., WOOD, E. F. & BURGES, S. J. 1994. A Simple Hydrologically Based Model of Land-Surface Water and Energy Fluxes for General-Circulation Models. *Journal of Geophysical Research-Atmospheres,* 99**,** 14415-14428.

LIU, J., YANG, H., GOSLING, S. N., KUMMU, M., FLÖRKE, M., PFISTER, S., HANASAKI, N., WADA, Y., ZHANG, X., ZHENG, C., ALCAMO, J. & OKI, T. 2017a. Water scarcity assessments in the past, present, and future. *Earth's Future,* 5**,** 545-559.

LIU, X., TANG, Q., CUI, H., MU, M., GERTEN, D., GOSLING, S. N., MASAKI, Y., SATOH, Y. & WADA, Y. 2017b. Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations. *Environmental Research Letters,* 12**,** 025009.

LIU, Y., GUO, H., ZHANG, Z., WANG, L., DAI, Y. & FAN, Y. 2007. An optimization method based on scenario analysis for watershed management under uncertainty. *Environ Manage,* 39**,** 678-90.

LIU, Y. & GUPTA, H. V. 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research,* 43.

MANABE, S. 1969. Climate and the ocean circulation: The atmospheric circulation and the hydrology of the earth's surface, . *Mon. Weather Rev., ,* 97, **,** 739–774.

MASAKI, Y., HANASAKI, N., BIEMANS, H., MÜLLER SCHMIED, H., TANG, Q., WADA, Y., GOSLING, S., TAKAHASHI, K. & HIJIOKA, Y. 2017. Intercomparison of global river discharge simulations focusing on dam operation --- Part II: Multiple models analysis in two case-study river basins, Missouri-Mississippi and Green-Colorado. *Environmental Research Letters*.

MATERIA, S., DIRMEYER, P. A., GUO, Z. C., ALESSANDRI, A. & NAVARRA, A. 2010. The Sensitivity of Simulated River Discharge to Land Surface Representation and Meteorological Forcings. *Journal of Hydrometeorology,* 11**,** 334-351.

MCLEOD, A. I., NOAKES, D. J., HIPEL, K. W. & THOMPSTONE, R. M. 1987. Combining Hydrologic Forecasts. *Journal of Water Resources Planning and Management-Asce,* 113**,** 29-41.

MEYBECK, M., KUMMU, M. & DÜRR, H. H. 2013. Global hydrobelts and hydroregions: improved reporting scale for water-related issues? *Hydrology and Earth System Sciences,* 17**,** 1093-1111.

MILLY, P. C., DUNNE, K. A. & VECCHIA, A. V. 2005. Global pattern of trends in streamflow and water availability in a changing climate. *Nature,* 438**,** 347-50.

MILLY, P. C. D., MALYSHEV, S. L., SHEVLIAKOVA, E., DUNNE, K. A., FINDELL, K. L., GLEESON, T., LIANG, Z., PHILLIPPS, P., STOUFFER, R. J. & SWENSON, S. 2014. An Enhanced Model of Land Water and Energy for Global Hydrologic and Earth-System Studies. *Journal of Hydrometeorology,* 15**,** 1739-1761.

MILLY, P. C. D. & SHMAKIN, A. B. 2002. Global Modeling of Land Water and Energy Balances. Part I: The Land Dynamics (LaD) Model. *Journal of Hydrometeorology,* 3**,** 283-299.

MÜLLER SCHMIED, H., EISNER, S., FRANZ, D., WATTENBACH, M., PORTMANN, F. T., FLÖRKE, M. & DÖLL, P. 2014. Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. *Hydrology and Earth System Sciences,* 18**,** 3511-3538.

NASSERI, M., ZAHRAIE, B., AJAMI, N. K. & SOLOMATINE, D. P. 2014. Monthly water balance modeling: Probabilistic, possibilistic and hybrid methods for model combination and ensemble simulation. *Journal of Hydrology,* 511**,** 675–691.

NIU, G.-Y., YANG, Z.-L., MITCHELL, K. E., CHEN, F., EK, M. B., BARLAGE, M., KUMAR, A., MANNING, K., NIYOGI, D., ROSERO, E., TEWARI, M. & XIA, Y. 2011. The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research,* 116.

NOHARA, D., KITOH, A., HOSAKA, M. & OKI, T. 2006. Impact of climate change on river discharge projected by multimodel ensemble. *Journal of Hydrometeorology,* 7**,** 1076-1089.

OKI, T., NISHIMURA, T. & DIRMEYER, P. 1999. Assessment of annual runoff from land surface models using Total Runoff Integrating Pathways (TRIP). *Journal of the Meteorological Society of Japan,* 77**,** 235-255.

OUDIN, L., ANDREASSIAN, V., MATHEVET, T., PERRIN, C. & MICHEL, C. 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. *Water Resources Research,* 42.

PAPPENBERGER, F., DUTRA, E., WETTERHALL, F. & CLOKE, H. L. 2012. Deriving global flood hazard maps of fluvial floods through a physical model cascade. *Hydrology and Earth System Sciences,* 16**,** 4143-4156.

PETRESCU, A. M. R., VAN BEEK, L. P. H., VAN HUISSTEDEN, J., PRIGENT, C., SACHS, T., CORRADI, C. A. R., PARMENTIER, F. J. W. & DOLMAN, A. J. 2010. Modeling regional to global CH4emissions of boreal and arctic wetlands. *Global Biogeochemical Cycles,* 24**,** n/a-n/a.

PHILLIPS, T. J. & GLECKLER, P. J. 2006. Evaluation of continental precipitation in 20th century climate simulations: The utility of multimodel statistics. *Water Resources Research,* 42.

PINIEWSKI, M., LAIZE, C. L., ACREMAN, M. C., OKRUSZKO, T. & SCHNEIDER, C. 2014. Effect of climate change on environmental flow indicators in the narew basin, poland. *J Environ Qual,* 43**,** 155-67.

POKHREL, Y. N., HANASAKI, N., YEH, P. J. F., YAMADA, T. J., KANAE, S. & OKI, T. 2012. Model estimates of sea-level change due to anthropogenic impacts on terrestrial water storage. *Nature Geoscience,* 5**,** 389-392.

PRENTICE, I. C., CRAMER, W., HARRISON, S. P., LEEMANS, R., MONSERUD, R. A. & SOLOMON, A. M. 1992. Special Paper: A Global Biome Model Based on Plant Physiology and Dominance, Soil Properties and Climate. *Journal of Biogeography,* 19**,** 117.

PRUDHOMME, C., GIUNTOLI, I., ROBINSON, E. L., CLARK, D. B., ARNELL, N. W., DANKERS, R., FEKETE, B. M., FRANSSEN, W., GERTEN, D., GOSLING, S. N., HAGEMANN, S., HANNAH, D. M., KIM, H., MASAKI, Y., SATOH, Y., STACKE, T., WADA, Y. & WISSER, D. 2014. Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment. *Proc Natl Acad Sci U S A,* 111**,** 3262-7.

PUSHPALATHA, R., PERRIN, C., MOINE, N. L. & ANDRÉASSIAN, V. 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology,* 420–421**,** 171–182.

RIAHI, K., RAO, S., KREY, V., CHO, C., CHIRKOV, V., FISCHER, G., KINDERMANN, G., NAKICENOVIC, N. & RAFAJ, P. 2011. RCP 8.5—A scenario of comparatively high greenhouse gas emissions. *Climatic Change,* 109**,** 33-57.

RINGEVAL, B., HOUWELING, S., VAN BODEGOM, P. M., SPAHNI, R., VAN BEEK, R., JOOS, F. & RÖCKMANN, T. 2014. Methane emissions from floodplains in the Amazon Basin: challenges in developing a process-based model for global applications. *Biogeosciences,* 11**,** 1519-1558.

RITTER, A. & CARPENA, R. M. 2013. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology,* 480**,** 33–45.

ROCKSTROM, J., FALKENMARK, M., KARLBERG, L., HOFF, H., ROST, S. & GERTEN, D. 2009. Future water availability for global food production: The potential of green water for increasing resilience to global change. *Water Resources Research,* 45.

SANDERSON, B. M. & KNUTTI, R. 2012. On the interpretation of constrained climate model ensembles. *Geophysical Research Letters,* 39.

SCHAEFLI, B. & GUPTA, H. V. 2007. Do Nash values have value? *Hydrological Processes,* 21**,** 2075-2080.

SCHELLEKENS, J., DUTRA, E., MARTÍNEZ-DE LA TORRE, A., BALSAMO, G., VAN DIJK, A., SPERNA WEILAND, F., MINVIELLE, M., CALVET, J.-C., DECHARME, B., EISNER, S., FINK, G., FLÖRKE, M., PEßENTEINER, S., VAN BEEK, R., POLCHER, J., BECK, H., ORTH, R., CALTON, B., BURKE, S., DORIGO, W. & WEEDON, G. P. 2017. A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset. *Earth System Science Data,* 9**,** 389-413.

SCHELLNHUBER, H. J., FRIELER, K. & KABAT, P. 2014. The elephant, the blind, and the intersectoral intercomparison of climate impacts. *Proc Natl Acad Sci U S A,* 111**,** 3225-7.

SCHEWE, J., HEINKE, J., GERTEN, D., HADDELAND, I., ARNELL, N. W., CLARK, D. B., DANKERS, R., EISNER, S., FEKETE, B. M., COLON-GONZALEZ, F. J., GOSLING, S. N., KIM, H., LIU, X., MASAKI, Y., PORTMANN, F. T., SATOH, Y., STACKE, T., TANG, Q., WADA, Y., WISSER, D., ALBRECHT, T., FRIELER, K., PIONTEK, F., WARSZAWSKI, L. & KABAT, P. 2014. Multimodel assessment of water scarcity under climate change. *Proc Natl Acad Sci U S A,* 111**,** 3245-50.

SEE, L. & ABRAHART, R. J. 2001. Multi-model data fusion for hydrological forecasting. *Computers & Geosciences,* 27**,** 987-994.

SETTE, S. & BOULLART, L. 2000. An implementation of genetic algorithms for rule based machine learning. *Engineering Applications of Artificial Intelligence,* 13**,** 381-390.

SHAMSELDIN, A. Y., NASR, A. E. & O'CONNOR, K. M. 2002. Comparison of different forms of the Multi-Layer Feed-Forward Neural Network method used for river flow forecasting. *Hydrology and Earth System Sciences,* 6**,** 671-684.

SHAMSELDIN, A. Y. & O'CONNOR, K. M. 1999. A real-time combination method for the outputs of different rainfall-runoff models. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques,* 44**,** 895-912.

SHAMSELDIN, A. Y., O'CONNOR, K. M. & NASR, A. E. 2007. A comparative study of three neural network forecast combination methods for simulated river flows of different rainfall-runoff models. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques,* 52**,** 896-916.

SHAMSELDIN, A. Y., OCONNOR, K. M. & LIANG, G. C. 1997. Methods for combining the outputs of different rainfall-runoff models. *Journal of Hydrology,* 197**,** 203-229.

SITCH, S., SMITH, B., PRENTICE, I. C., ARNETH, A., BONDEAU, A., CRAMER, W., KAPLAN, J. O., LEVIS, S., LUCHT, W., SYKES, M. T., THONICKE, K. & VENEVSKY, S. 2003. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology,* 9**,** 161-185.

SOLOMATINE, D. P., SEE, L. M. & ABRAHART, R. J. 2008. *Data-driven modelling: concepts, approaches and experiences, in: Abrahart, R.J., See, L.M., Solomatine, D.P. (Eds.), Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, Springer-Verlag, Berlin, pp. 17-33.*

SOOD, A. & SMAKHTIN, V. 2015. Global hydrological models: a review. *Hydrological Sciences Journal,* 60**,** 549-565.

TEBALDI, C. & KNUTTI, R. 2007. The use of the multi-model ensemble in probabilistic climate projections. *Philos Trans A Math Phys Eng Sci,* 365**,** 2053-75.

TRAORE, S. & GUVEN, A. 2011. New algebraic formulations of evapotranspiration extracted from gene-expression programming in the tropical seasonally dry regions of West Africa. *Irrigation Science,* 31**,** 1-10.

UNFCCC 2015. Adoption of the Paris Agreement. Proposal by the President. Proposal by the President. Available from: http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf. Geneva, Switzerland: United Nations.

VAN BEEK, L. P. H., EIKELBOOM, T., VAN VLIET, M. T. H. & BIERKENS, M. F. P. 2012. A physically based model of global freshwater surface temperature. *Water Resources Research,* 48.

VAN BEEK, L. P. H., WADA, Y. & BIERKENS, M. F. P. 2011. Global monthly water stress: 1. Water balance and water availability. *Water Resources Research,* 47.

VAN DIJK, A. I. J. M., PEÑA-ARANCIBIA, J. L., WOOD, E. F., SHEFFIELD, J. & BECK, H. E. 2013. Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. *Water Resources Research,* 49**,** 2729-2746.

VAN HUIJGEVOORT, M. H. J., HAZENBERG, P., VAN LANEN, H. A. J., TEULING, A. J., CLARK, D. B., FOLWELL, S., GOSLING, S. N., HANASAKI, N., HEINKE, J., KOIRALA, S., STACKE, T., VOSS, F., SHEFFIELD, J. & UIJLENHOET, R. 2013. Global Multimodel Analysis of Drought in Runoff for the Second Half of the Twentieth Century. *Journal of Hydrometeorology,* 14**,** 1535-1552.

VAN LOON, A. F., VAN HUIJGEVOORT, M. H. J. & VAN LANEN, H. A. J. 2012. Evaluation of drought propagation in an ensemble mean of large-scale hydrological models. *Hydrology and Earth System Sciences,* 16**,** 4057-4078.

VAN VLIET, M. T. H., YEARSLEY, J. R., LUDWIG, F., VÖGELE, S., LETTENMAIER, D. P. & KABAT, P. 2012. Vulnerability of US and European electricity supply to climate change. *Nature Climate Change,* 2**,** 676-681.

VELDKAMP, T. I. E. 2017. *Water scarcity at the global and regional scales unravelling its dominant drivers in historical and future time periods,.* PhD, Vrije Universiteit Amsterdam, .

VELDKAMP, T. I. E., WADA, Y., AERTS, J., DOLL, P., GOSLING, S. N., LIU, J., MASAKI, Y., OKI, T., OSTBERG, S., POKHREL, Y., SATOH, Y., KIM, H. & WARD, P. J. 2017. Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century. *Nat Commun,* 8**,** 15697.

VERZANO, K., BÄRLUND, I., FLÖRKE, M., LEHNER, B., KYNAST, E., VOß, F. & ALCAMO, J. 2012. Modeling variable river flow velocity on continental scale: Current situation and climate change impacts in Europe. *Journal of Hydrology,* 424-425**,** 238-251.

VITERBO, P. & BELJAARS, A. C. M. 1995. An Improved Land Surface Parameterization Scheme in the ECMWF Model and Its Validation. *Journal of Climate,* 8**,** 2716-2748.

VÖRÖSMARTY, C. J., FEDERER, C. A. & SCHLOSS, A. L. 1998. Potential evaporation functions compared on US watersheds: Possible implications for global-scale water balance and terrestrial ecosystem modeling. *Journal of Hydrology,* 207**,** 147-169.

WADA, Y., VAN BEEK, L. P. H., SPERNA WEILAND, F. C., CHAO, B. F., WU, Y.-H. & BIERKENS, M. F. P. 2012. Past and future contribution of global groundwater depletion to sea-level rise. *Geophysical Research Letters,* 39**,** n/a-n/a.

WADA, Y., VAN BEEK, L. P. H., VAN KEMPEN, C. M., RECKMAN, J. W. T. M., VASAK, S. & BIERKENS, M. F. P. 2010. Global depletion of groundwater resources. *Geophysical Research Letters,* 37**,** n/a-n/a.

WADA, Y., VAN BEEK, L. P. H., VIVIROLI, D., DÜRR, H. H., WEINGARTNER, R. & BIERKENS, M. F. P. 2011. Global monthly water stress: 2. Water demand and severity of water stress. *Water Resources Research,* 47.

WADA, Y., WISSER, D. & BIERKENS, M. F. P. 2014. Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth System Dynamics,* 5**,** 15-40.

WAGENER, T., BOYLE, D. P., LEES, M. J., WHEATER, H. S., GUPTA, H. V. & SOROOSHIAN, S. 2001. A framework for development and application of hydrological models. *Hydrology and Earth System Sciences,* 5**,** 13–26.

WARD, P. J., JONGMAN, B., WEILAND, F. S., BOUWMAN, A., VAN BEEK, R., BIERKENS, M. F. P., LIGTVOET, W. & WINSEMIUS, H. C. 2013. Assessing flood risk at the global scale: model setup, results, and sensitivity. *Environmental Research Letters,* 8.

WARSZAWSKI, L., FRIELER, K., HUBER, V., PIONTEK, F., SERDECZNY, O. & SCHEWE, J. 2014. The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework. *Proceedings of the National Academy of Sciences,* 111**,** 3228-3232.

WEEDON, G. P., GOMES, S., VITERBO, P., SHUTTLEWORTH, W. J., BLYTH, E., ÖSTERLE, H., ADAM, J. C., BELLOUIN, N., BOUCHER, O. & BEST, M. 2011. Creation of the WATCH Forcing Data and Its Use to Assess Global and Regional Reference Crop Evaporation over Land during the Twentieth Century. *Journal of Hydrometeorology,* 12**,** 823-848.

WILLEMS, P. 2009. A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models. *Environmental Modelling & Software,* 24**,** 311–321.

WORLD ECONOMIC FORUM 2016. The global risks report 2016. 11th edition. World Economic Forum,Geneva, Switzerland.

YAMAZAKI, D., KANAE, S., KIM, H. & OKI, T. 2011. A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resources Research,* 47.

YOSSEF, N. C., WINSEMIUS, H., WEERTS, A., VAN BEEK, R. & BIERKENS, M. F. P. 2013. Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. *Water Resources Research,* 49**,** 4687-4699.

ZELINKA, I., OPLATKOVA, Z. & NOLLE, L. 2005. Analytic Programming - Symbolic Regression by Means of Arbitrary Evolutionary Algorithms. *Special Issue on Inteligent Systems, International Journal of Simulation, Systems, Science and Technology,* 6**,** 44 - 56.

ZHAO, F., VELDKAMP, T. I. E., FRIELER, K., SCHEWE, J., OSTBERG, S., WILLNER, S., SCHAUBERGER, B., GOSLING, S. N., SCHMIED, H. M., PORTMANN, F. T., LENG, G., HUANG, M., LIU, X., TANG, Q., HANASAKI, N., BIEMANS, H., GERTEN, D., SATOH, Y., POKHREL, Y., STACKE, T., CIAIS, P., CHANG, J., DUCHARNE, A., GUIMBERTEAU, M., WADA, Y., KIM, H. & YAMAZAKI, D. 2017. The critical role of the routing scheme in simulating peak river discharge in global hydrological models. *Environmental Research Letters,* 12**,** 075003.

Paper I: Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts

Environmental Research Letters

LETTER • **OPEN ACCESS**

Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts

To cite this article: Jamal Zaherpour *et al* 2018 *Environ. Res. Lett.* **13** 065015

View the article online for updates and enhancements.



This content was downloaded from IP address 128.243.2.30 on 12/06/2018 at 12:26

Environmental Research Letters

LETTER

Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts

Jamal Zaherpour[1,19], Simon N Gosling[1], Nick Mount[1], Hannes Müller Schmied[2,3], Ted I E Veldkamp[4,18], Rutger Dankers[5], Stephanie Eisner[6], Dieter Gerten[7,8], Lukas Gudmundsson[9], Ingjerd Haddeland[10], Naota Hanasaki[11], Hyungjun Kim[12], Guoyong Leng[13], Junguo Liu[14], Yoshimitsu Masaki[15], Taikan Oki[12,16], Yadu Pokhrel[17], Yusuke Satoh[18], Jacob Schewe[7] and Yoshihide Wada[18]

1 School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom
2 Institute of Physical Geography, Goethe-University, Frankfurt, Germany
3 Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Frankfurt, Germany
4 VU Amsterdam, de Boelelaan 1087, 1081 HV Amsterdam, the Netherlands
5 Met Office, Fitzroy Road, Exeter, EX1 3PB, United Kingdom
6 Center for Environmental Systems Research, University of Kassel, Kassel, Germany
7 Potsdam Institute for Climate Impact Research, Telegrafenberg, 14473 Potsdam, Germany
8 Geography Department, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
9 ETH Zurich, Institute for Atmospheric and Climate Science, Universitätstrasse 16, 8092 Zürich, Switzerland
10 Norwegian Public Roads Administration, Statens vegvesen Region vest, Postboks 43, 6861 Leikanger, Norway
11 National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, 305–8506, Japan
12 Institute of Industrial Science, University of Tokyo, 4–6-1 Meguro-ku, Komaba, Tokyo 153–8505, Japan
13 Environmental Change Institute, University of Oxford, Oxford, OX1 3QY, United Kingdom
14 School of Environmental Science and Engineering, Southern University of Science and Technology of China, Shenzhen, 518055, People's Republic of China
15 Hirosaki University, Bunkyocho-3, Hirosaki, Aomori, 036-8561, Japan
16 United Nations University, 5–53–70 Jingumae, Shibuya-ku, Tokyo 150-8925, Japan
17 Department of Civil and Environmental Engineering, Michigan State University, East Lansing, Michigan 48824, United States of America
18 International Institute for Applied Systems Analysis (IIASA) - Schlossplatz 1 - A-2361 Laxenburg, Austria
19 Author to whom any correspondence should be addressed.

E-mail: lgxjz1@nottingham.ac.uk **and** zaherpour@gmail.com

Keywords: global hydrological models, land surface models, human impacts, extreme events, model evaluation, model validation

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
11 October 2017

REVISED
14 May 2018

ACCEPTED FOR PUBLICATION
16 May 2018

PUBLISHED
12 June 2018

Original content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Abstract

Global-scale hydrological models are routinely used to assess water scarcity, flood hazards and droughts worldwide. Recent efforts to incorporate anthropogenic activities in these models have enabled more realistic comparisons with observations. Here we evaluate simulations from an ensemble of six models participating in the second phase of the Inter-Sectoral Impact Model Inter-comparison Project (ISIMIP2a). We simulate monthly runoff in 40 catchments, spatially distributed across eight global hydrobelts. The performance of each model and the ensemble mean is examined with respect to their ability to replicate observed mean and extreme runoff under human-influenced conditions. Application of a novel integrated evaluation metric to quantify the models' ability to simulate timeseries of monthly runoff suggests that the models generally perform better in the wetter equatorial and northern hydrobelts than in drier southern hydrobelts. When model outputs are temporally aggregated to assess mean annual and extreme runoff, the models perform better. Nevertheless, we find a general trend in the majority of models towards the overestimation of mean annual runoff and all indicators of upper and lower extreme runoff. The models struggle to capture the timing of the seasonal cycle, particularly in northern hydrobelts, while in southern hydrobelts the models struggle to reproduce the magnitude of the seasonal cycle. It is noteworthy that over all hydrological indicators, the ensemble mean fails to perform better than any individual model—a finding that challenges the commonly held perception that model ensemble

© 2018 The Author(s). Published by IOP Publishing Ltd

estimates deliver superior performance over individual models. The study highlights the need for continued model development and improvement. It also suggests that caution should be taken when summarising the simulations from a model ensemble based upon its mean output.

1. Introduction

Global hydrological models (GHMs) and land surface models (LSMs) are used for assessing the impacts of climate change on water availability and scarcity [1–5], droughts [6, 7], flood hazard and risk [8–11], the response of the global hydrological cycle to climate change mitigation [12], forecasting at short timescales [13] and examining the role of water in assessments of food security [14–16]. In turn, their results are used to inform global policy decisions on climate change [17, 18]. Therefore, it is important to understand the strengths and limitations of these models in simulating global hydrological variability.

The aims of this study are to: provide new understanding on how global-scale hydrological models perform in different hydro-geographical locations of the globe; to highlight the current strengths and limitations of global-scale hydrological modelling; to identify opportunities for the community to improve the models; and to explore the potential implications of our research for future work in the field.

Previous global evaluation studies (table 1) differ from each other in several ways: in the number of models evaluated (2–14, with a median of 6); the number of catchments included (8–6192); the size of catchments considered ($10\,km^2$–$4\,758\,000\,km^2$); the evaluation metrics calculated; the hydrological indicators evaluated; and the period of analyses. The highly varied approaches taken in previous studies means there remain several opportunities for improving the way in which model evaluation studies are conducted.

Firstly, only a handful of studies [19–22] have explored spatial patterns of model performance, all of them by using the Köppen climate classification system. Here, for the first time, we evaluate the performance of several global-scale hydrological models across hydrobelts [23] (figure 1 and table S2; tables and figures in the supplementary information available at stacks.iop.org/ERL/13/065015/mmedia, hereafter called SI, are denoted by an S in their numbering). This offers a more appropriate classification scheme for catchment hydrology than the Köppen system because it takes into account a greater number, and diversity of, hydro-geographical factors in defining the boundaries of the spatial units.

Secondly, almost all previous evaluation studies compare simulations of *naturalised* discharge to observations. This means that the effects of human impacts on runoff and river flows remain unaccounted for by the models. This is despite the fact that the majority of catchments across the globe have been severely influenced by human activities [24, 25]. We capitalise on the latest model simulations conducted as part of the second phase of the Inter-Sectoral Impact Model Inter-comparison Project (ISIMIP2a), which provides a set of simulations that include time-varying human impacts, such as water abstractions and reservoir operations by dams [26]. In so doing and in parallel with a companion study presented in this journal issue [25], we conduct the first multi-model evaluation of runoff under human-impacted conditions as simulated by global-scale hydrological models.

Thirdly, with only a very limited number of exceptions [20, 21, 25, 27], it is uncommon for global-scale hydrological model evaluation studies to assess the ability of models to simulate hydrological extremes, particularly high and low runoff or specific return periods. We address this paucity of evidence by evaluating indicators of hydrological extremes, including specific return periods.

Fourthly, a multitude of different performance metrics have been employed within and between studies. The ranking of models, by performance, changes according to the metrics that are employed because different metrics emphasise different characteristics of model performance. To overcome this issue we use a novel integrated metric [28–31].

Finally, the ensemble mean (EM) is often used to summarise the performance of several models because it has been shown that the EM often performs better than the majority of the individual models from which it is derived [32–38]. However, not all studies have provided a comprehensive analysis of the performance of the EM relative to individual models, at global (table 1) and continental scales [39, 40]. Therefore, we evaluate the performance of six models individually, as well as the EM.

By capitalising on the five opportunities discussed above we provide a distinct contribution to understanding how global-scale hydrological models perform in different parts of the globe and for different hydrological indicators, including extremes.

2. Methods

2.1. Study catchments, models and data

2.1.1. *Study catchments and observed data*

Forty study catchments (figure 1 and table S1) were identified following a comprehensive set of four criteria (see SI) applied to observed data from the Global Runoff Data Centre, GRDC (available from http://grdc.bafg.de). The catchments provide a reasonable

Table 1. Overview of selected global-scale studies evaluating multiple models, including the present study. **GHM**: Global Hydrological Model, **LSM**: Land Surface Model, **DGVM**: Dynamic Global Vegetation Model, **GCM**: Global Climate Model, **MAR**: mean annual runoff/discharge, **MMR**: mean monthly runoff/discharge and/or seasonality, **MDR**: mean daily runoff/discharge, **HFP**: extreme high flow percentiles (e.g. Q5), **LFP**: extreme low flow percentiles (e.g. Q95), **HFR**: high flow return periods, **LFR**: low flow return periods.

Study (ordered by publication date)	No. of models evaluated				No. of catchments included (area in parentheses; km^2)	Inclusion of:		Hydrological indicators							Evaluation metrics								Spatial analysis classification	No. of years (year range in parentheses)
								Mean flows			Extreme events														
	GHM	LSM	DGVM	GCM		Human impacts	Ensemble mean	MAR	MMR	MDR	HFP	LFP	HFR	LFR	Integrated metric	CE: Coefficient of Efficiency	PBIAS: Percent bias	RMSE: Root mean square error	MARE: Mean absolute relative error	R^2: Coefficient of determination	r: Pearson correlation coefficient	Statistical tests	SD: Standard deviation	CV: Coefficient of variation		
[41]		11			250 (mean 157,000)		✓	✓	✓								✓	✓				✓			Latitude, rainauge density	2 (87-88)
[42]				12	165 (> 50,000)		✓	✓										✓								28 (28-99)
[43]		6			33 (100,000 - 4758000)				✓	✓						✓	✓									10 (86–19)
[38]				19	24 (100,000-4,640,000)		✓	✓	✓								✓	✓				✓			Latitude	20 (81-00)
[44]		6			80 (100,000 - 4,758,000)				✓							✓	✓									10 (86–95)
[45]		2			80 (100,000 - 4,758,000)				✓							✓	✓	✓			✓					3 (82-85)
[46]		13			30 (82,000 - 4,677,000)				✓									✓								10 (86–95)
[27]		4			66 (19,000 - 4,600,000)				✓	✓	✓	✓						✓			✓		✓		Continental	28 (79-07)
[19]	5	6			8 (650,000 - 4,600,000)		✓		✓															✓	Köppen	15 (85-99)
[47]		14			150 (>10,000)		✓	✓								✓	✓			✓						10 (86-95)
[48]		5			6192 (10 - 10,000)				✓												✓	✓				30 (79-08)
[20]	2	2			4079 (10 - 10,000)					✓	✓	✓							✓	✓					Köppen	31 (79–10)
[49]		1	6		16 (135,757 - 3,475,000)			✓	✓									✓	✓		✓				Latitude	30 (81–10)
[22]	1	1			644 (>2,000)			✓	✓							✓	✓				✓				Köppen	31 (80-10)
[21]	6	4			966 (1,000 - 5,000)		✓	✓	✓	✓	✓	✓			✓						✓	✓			Köppen	34 (79–12)
[50]	6	3			11 (67,490 - 2,460,000)		✓	✓	✓									✓			✓	✓				30 (71-00)
[25]	4	1			471 (>9,000)	✓			✓		✓	✓			✓	✓	✓				✓	✓				40 (71-10)
This study	5	1			40 (104,000 – 4,640,300)	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓				Hydrobelts	40 (71-10)

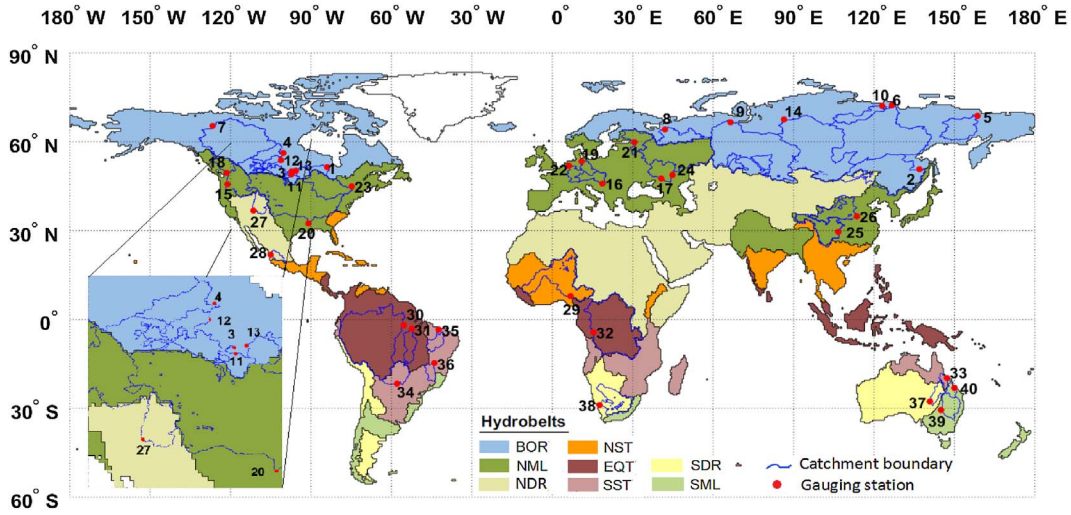


Figure 1. Locations of the 40 catchments across hydrobelts. Catchment details are provided in table S1. Hydrobelts are named: BOR= boreal, NML = northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML=southern mid-latitude, SDR = southern dry and SST = southern subtropical.

geographic coverage, however, the availability and quality of observed data in the GRDC database led to a selection bias that resulted in the number of boreal and northern mid-latitude catchments being proportionately high (table S2).

2.1.2. Model ensemble and forcing data

The model ensemble comprises a suite of GHMs and LSMs that participated in the water (global) sector of ISIMIP2a [51]. The models are DBH [52], H08 [53], LPJmL [54], MATSIRO [55], PCR-GLOBWB [56] and WaterGAP2 [57] (see SI for full model names and details of the hydrological processes represented in the models).

All of the models have been developed to represent and account for the impacts of historical time-varying human management such as land use, water use and dam operation. Apart from WaterGAP2 the models were not calibrated for the ISIMIP2a simulations. The models simulate runoff (amongst other hydrological variables) across the global land surface at $0.5° \times 0.5°$ spatial resolution. Following the method described by [19], monthly observed and simulated discharge data was converted to catchment-mean monthly runoff by using the area upstream of the gauge according to the DDM30 river network. Thus an area correction factor is applied to the GRDC discharge data to account for the fact that the river network, which is at 0.5 ° spatial resolution, may not perfectly overlap with the GRDC river catchment boundaries. Output from the models is openly available from the Earth System Grid Federation (ESGF; https://esgf-data.dkrz.de; [51]).

All models were run for the period 1971–2010 with input climate data provided by the Global Soil Wetness Project Phase 3, (GSWP3; [58]). GSWP3 has been used as a forcing dataset in several other recent GHM and LSM studies [2, 26, 57, 59, 60] (see SI, section 1.3, for an explanation of why we used this forcing dataset specifically).

2.2. Evaluating model performance

2.2.1. The integrated evaluation metric

Assessment of relative model performance in a meaningful way is difficult without a transferrable benchmark against which model performance in different catchments can be compared to consistently. In addition, different metrics are more or less suited to assessing individual characteristics of a model's fit. To overcome this we use a ratiometric integrated metric, the ideal point error (IPE) [31], equation 1. Our configuration of IPE has three components because it combines three commonly used individual evaluation metrics: root mean square error (RMSE), Mean absolute relative error (MARE) and the Nash-Sutcliffe coefficient of efficiency (CE). These are used to assess the relative performance of each model and the EM to replicate observed data. IPE is standardised against a benchmarked model. The benchmark model can be a simple statistical model, so it is sometimes called a naïve model [61]. It can be as simple as observed runoff shifted backwards by different time steps [61]. In our application of IPE, we adopted a naïve model benchmark such as this, where the observed runoff is shifted backwards by one month. Hence our naïve model benchmark assumes runoff in month t is equal to runoff in month $t-1$ and therefore essentially assumes persistence (IPEn, equation 1). The three IPE components are evaluated against their benchmark model counterparts.

The IPE equation presented in equation (1) is adapted from the original formula by (Dawson *et al* 2012). The negative reciprocal of the IPE score is used (equation 3) where the performance of a model

Table 2. Indicators of mean and extreme runoff calculated in this study for model evaluation.

Indicator abbreviation	Description of indicator
MAR	Mean annual runoff[a]
MMR	Mean monthly runoff[a]
Q5	The magnitude of monthly runoff that is exceeded 5 % of the time in the monthly time series (indicator of high flows)[a]
Q95	The magnitude of monthly runoff that is exceeded 95 % of the time in the monthly time series (indicator of low flows)[a]
None	The magnitude of maximum monthly runoff associated with the 2-, 5-, 10-, 20-, and 25 year return periods respectively (see SI for calculation method)[b]
None	The magnitude of minimum 3 month moving average of runoff associated with the 2-, 5-, 10-, 20-, and 25 year return periods respectively (see SI for calculation method)[b]

[a] For these indicators the EM is calculated as a timeseries by calculating the average across all individual models for each month, to allow calculation of timeseries-based evaluation metrics such as the NSE and IPE.

[b] For the maximum and minimum return period flows the EM was calculated as the average of the maximum/minimum monthly runoff for each return period calculated across all GHMs.

exceeds that of the benchmark to maintain proportionality in comparisons between the IPE scores of models that fail to perform as well as the benchmark and those where performance exceeds it. The IPE scores range between −1 and −∞ (performance improvement over the benchmark model) and 1 and +∞ (performance loss over benchmark model). The IPE score is ratiometric—for example, a model that performs twice as well as the benchmark model will have an IPE score of −2 and a model that performs twice as badly will have a score of 2. IPE would be 1 if a model performs the same as the benchmark, whilst a model infinitely better than the benchmark would have an IPE of −∞.

$$\mathrm{IPEn} = \left\{ \begin{array}{l} [0.333 * ((\mathrm{RMSE_m}/\mathrm{RMSE_b})^2 + \\ (\mathrm{MARE_m}/\mathrm{MARE_b})^2 + \\ ((\mathrm{CE}_m - 1)/(\mathrm{CE}_b - 1))^2)]^{\frac{1}{2}} \end{array} \right\} \quad (1)$$

$$\mathrm{IPE} = \mathrm{IPEn}. \quad (2)$$

If IPEn ≥1 i.e. where a model fails to outperform the benchmark model

$$\mathrm{IPE} = \frac{-1}{\mathrm{IPEn}}. \quad (3)$$

If IPEn <1 i.e. where a model outperforms the benchmark model

where:

RMSE = root mean squared error

MARE = mean absolute relative error

CE = coefficient of efficiency (Nash-Sutcliffe Efficiency)

m = model simulated data

b = benchmark (the naïve model).

2.2.2. Weighted performance measures and performance ranking

Where measures of performance are aggregated for an entire hydrobelt we do so by calculating a weighted mean, to resolve spatial biases introduced by having a different number of catchments in each hydrobelt. For each catchment, observed mean annual runoff (MAR), representing the effect of both catchment size and flow, is applied as the relative weight, so that any given weighted performance metric (W_m) can be calculated as:

$$\mathrm{W\_m_{HB}} = \frac{\sum_{c=1}^{n} \mathrm{MAR_c} * \mathrm{m_c}}{\sum_{c=1}^{n} \mathrm{MAR_c}} \quad (4)$$

where *m* denotes metric, HB and *c* respectively denote hydrobelt and catchment, and *n* the number of catchments in each hydrobelt.

Measures of performance that we calculated and weighted in this way include IPE, percent bias (PBIAS) and the relative difference between simulated and observed values for the seasonal cycle (see table S4 for formulae). The median percentage difference (MPD) was calculated across all catchments but was not weighted.

For consistency throughout the paper, metrics derived for the EM are used to rank and/or facilitate comparisons between catchments or hydrobelts. This is not to say, however, that the EM is a reliable indicator of overall model performance.

2.2.3. Hydrological indicators

In addition to IPE and measures described in section 2.2.2 we calculated six indicators of mean and extreme runoff (table 2).

3. Results

3.1. Models' ability to replicate observed monthly runoff time series

IPE scores estimated from the monthly runoff timeseries for individual models and the EM respectively, across hydrobelts, are presented in figure 2 (see table S6 for individual catchments and figure S1 for the individual metrics that comprise the IPE). Model performance is generally better in the equatorial

Table 3. Mean weighted IPE for each model by hydrobelt. Negative (positive) values indicate performance improvement (loss) over the naïve model benchmark. Best performing individual models for each hydrobelt are in bold. Number of catchments in each hydrobelt are in parentheses. Hydrobelts are ranked according to the perfomance of the EM. Rows are ordered according to the mean latitude of each hydrobelt from north to south (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML= southern mid-latitude, SDR= southern dry and SST= southern subtropical).

Hydrobelt (No. of catchments)	Model							Hydrobelt rank (based on EM)	Hydrobelt rank (based on best individual model)
	DBH	H08	LPJmL	MATSIRO	PCR-GLOBWB	WaterGAP2	EM		
BOR (14)	25.63	5.33	8.12	2.52	3.53	**1.12**	2.07	2	5
NML (12)	53.08	12.76	9.87	7.20	5.87	**−0.30**	3.90	4	1
NDR (2)	14.53	4.09	4.71	4.32	7.46	**0.95**	2.96	3	3
NST (1)	12.10	12.89	12.10	1.82	5.39	**1.12**	4.32	5	4
EQT (3)	4.82	3.28	3.65	1.93	2.95	**0.57**	1.67	1	2
SST (4)	26.91	21.86	19.53	2.55	13.40	**1.38**	10.58	6	6
SDR (2)	5305.19	96.43	2051.53	326.04	520.61	**78.19**	1393.72	8	8
SML (2)	2780.76	109.56	1440.15	128.58	958.59	**42.15**	909.93	7	7
Median weighted IPE across all hydrobelts	26.27	12.82	10.99	3.44	6.67	1.12	4.11	—	—
Model rank (based on median weighted IPE across all hydrobelts)	6	5	4	2	3	1			

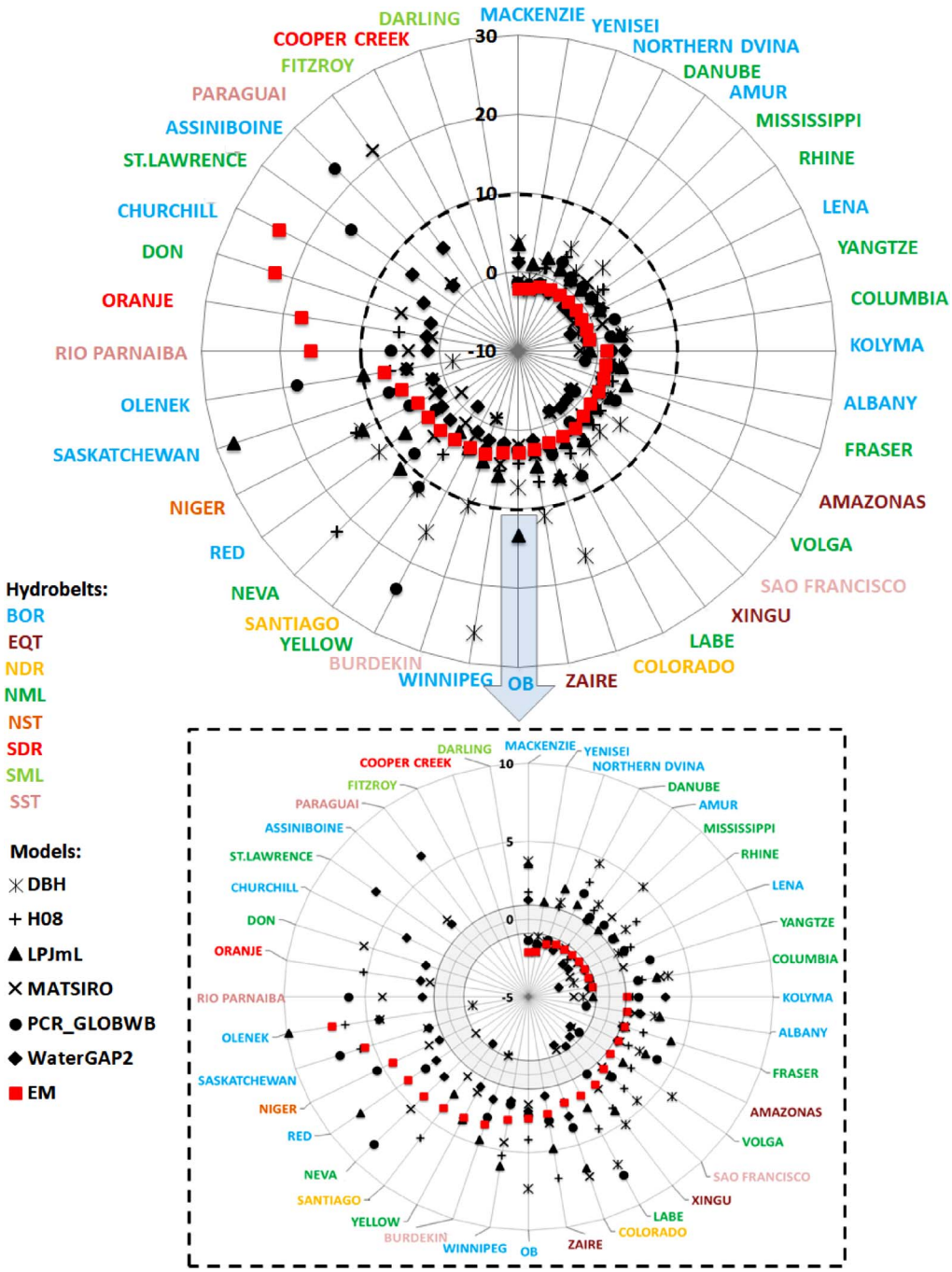


Figure 2. Catchments ranked clockwise (starting with Mackenzie) according to the IPE score for the ensemble mean (EM, red squares). IPE scores range between $(-1, -\infty]$ and $[1, +\infty)$. The bottom panel focuses on IPE≤10, with the range (−1,1) in grey, representing the boundary of performance improvements (IPE ≤ -1) or loss (IPE ≥ 1) relative to the naïve benchmark model. Catchment names are colored by hydrobelts (BOR = boreal, NML = northern mid-latitude, NDR = northern dry, NST = northern subtropical, EQT = equatorial, SML= southern mid-latitude, SDR= southern dry and SST = southern subtropical).

and northern hydrobelts (EQT, BOR, NML, NDR and NST) than the southern hydrobelts (SST, SDR and SML). When ranked by the EM, the 13 highest ranked catchments are located in BOR and NML hydrobelts. This is in part the result of a bias in the number of catchments in these hydrobelts. However, weighted IPE scores (table 3) indicate that model performance in these two hydrobelts is particularly favourable when compared to that of southern hemisphere hydrobelts. The two SML catchments are ranked 38th and 40th and the two SDR catchments are ranked 32nd and 39th. The relatively lower performance of the majority of models and the EM in the SDR and SML hydrobelts, which include seasonally dry ephemeral rivers, is the result of the models slightly overestimating low runoff values (typically less than 1 mm month^{-1}) during dry periods. This serves to decrease the MARE (figure S1), which in turn delivers poor IPE values. We considered lowering the weighting of the MARE component of the IPE equation, since the three components can be weighted differently [31] so that the IPE is not disproportionally affected by the

low MARE scores in these four catchments. However, this would be somewhat counter-intuitive to the objectives of our assessment because the low MARE values highlight that the models struggle to simulate low flows in these catchments, which is an important result of the evaluation exercise.

The EM and all individual models except WaterGAP2 and marginally MATSIRO, deliver their best IPE in the EQT hydrobelt. Here, the three contributing catchments cover 51% of its area; indicating that they are representative of the hydrobelt as a whole. However, the strong model performance is, in part, an artefact of using the naïve ($t-1$) model as a benchmark for computing IPE scores. In EQT catchments the month-to-month variability of hydrological inputs and responses is low, which makes the modelling challenge easier and the likelihood of significant deviation from the naïve benchmark model relatively low.

Examining the performance of individual models, it is evident that H08 performs relatively well (second to WaterGAP2) in the NDR hydrobelt as well as the two lowest ranked hydrobelts i.e. SDR and SML, suggesting it has particular capabilities in modelling dry regions. The calibrated WaterGAP2 outperforms the EM and other models in all hydrobelts, however, it is not always the best performing individual model. Indeed, MATSIRO achieves the best IPE scores in six catchments (table S6). In terms of the IPE, the EM performs better than the best model only in two catchments, the Amur and Mackenzie—for the other 38 catchments an individual model performs better than the EM.

To further investigate the performance of the EM, we calculated the EM under two different cases:

1. excluding the weakest performing model for each catchment in turn; and
2. excluding the first and second weakest models for each catchment in turn.

The best/weakest models that were left in the ensemble were identified based on the models' IPE scores.

Under the two cases, the EM outperformed the best individual model that was left in the ensemble in ten and 12 catchments respectively (table S6). Thus the relative performance of the EM, compared to the models that are left in the ensemble, improves when the weakest performing model(s) are removed. However, in the majority of catchments, even when the weakest performing model(s) are removed from the ensemble, an individual model still performs better than the EM.

3.2. Models' ability to reproduce key aggregated hydrological indicators

The PBIAS and MPD across catchments indicate a general trend towards the over-estimation of mean annual runoff (MAR), Q5 (high runoff) and Q95 (low runoff) amongst the models, especially for low(er) runoff values (figure 3, table 4 and table 5). In terms of individual model performance, WaterGAP2 and MATSIRO are ranked first or second by both MPD (table 4) and PBIAS (table 5) for all three indicators, whereas DBH is consistently ranked the lowest.

In terms of model performance across hydrobelts, the three best estimates of weighted PBIAS for all three hydrological indicators are, according to the performance of the EM, consistently BOR, NML and EQT, (table 5). The EQT hydrobelt is consistently highly ranked (top 3) according to the performance of both the EM and the best performing individual model, whilst the SML and SDR hydrobelts are always ranked low. The ranking of hydrobelts changes slightly when evaluating performance with MPD (table S5). Although BOR and NML are still ranked highly (top 3 for all three hydrological indicators) EQT is lower down the ranking (ranked 6th for all three hydrological indicators).

3.3. Models' ability to estimate runoff of different return periods

Table 6 presents the MPD between modelled and observed runoff calculated over all catchments for each of the five return periods (2-, 5-, 10-, 20- and 25-year) of maximum monthly runoff and minimum 3 month runoff respectively. For maximum runoff, PCR-GLOBWB performs particularly well, with MATSIRO and WaterGAP2 also delivering MPD values of $<25\%$. The remaining models perform relatively poorly with the MPD $>50\%$ across all return periods. For minimum runoff, MATSIRO again performs well, with DBH also achieving MPD values of $<20\%$. Despite being calibrated, WaterGAP2 (and PCR-GLOBWB) struggles to reproduce observed low runoff, with MPD values generally $>50\%$.

The magnitude of runoff associated with each return period for both maximum and minimum runoff is overestimated by the EM and by the majority of individual models. The EM fails to perform better than the best performing model for all maximum and minimum flow return periods. In half of all cases, MPD values are generally consistent across the different return periods, while MPD decreases or increases with higher return periods in other cases.

3.4. Models' ability to replicate seasonal cycles

Figure 4 displays for each hydrobelt the average weighted relative difference (using equation 4) between the modelled and observed seasonal cycle (long-term MMR) for all models. There is a general pattern of overestimation of MMR across the model ensemble. The largest relative differences occur in the months of peak runoff. No single model performs consistently better or worse than all other models throughout the whole seasonal cycle, since the month in which the maximum relative difference occurs varies across models.

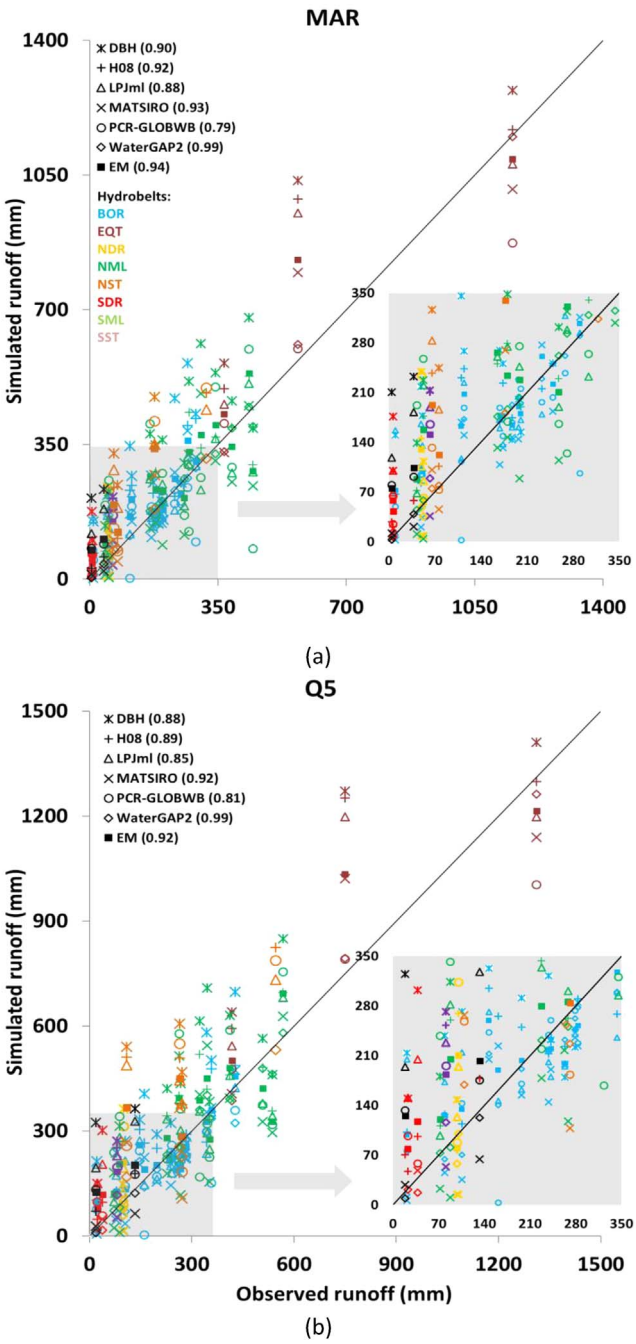
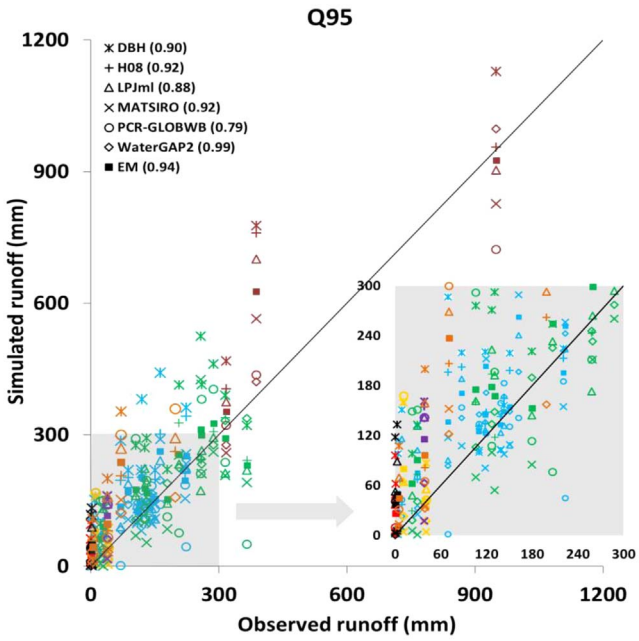


Figure 3. Simulated versus observed MAR (*a*); Q5 (*b*); and Q95, (*c*) for each catchment (n = 40). The numbers in parentheses indicate the Pearson correlation coefficients between simulated and observed data (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML = southern mid-latitude, SDR = southern dry and SST= southern subtropical).

The highest magnitude differences are observed in the SDR and SML hydrobelts. The lowest are in NML and EQT. The models' poor performance in simulating long-term MMR in the SDR hydrobelt confirms their poor performance in replicating the time series of runoff in this hydrobelt (table 3). This suggests that timing errors may be responsible for much of the error in simulating runoff timeseries in the SDR hydrobelt.

Table 7 shows the duration, in months, of the difference between simulated and observed timing of the month of maximum/minimum runoff, i.e. timing bias, as an average for each hydrobelt (figure S2 displays the simulated and observed MMR for each catchment). Early bias is common in all hydrobelts and it is especially prevalent for minimum flows. Late bias is less evident. Two potential explanations for early bias are the models' inability to capture late snowmelt in snow-dominant regions [62] and challenges in accurately representing groundwater or baseflow [55].

Across all models and catchments, DBH shows the least early (−0.60 months) and LPJmL the least late (0.03 months) biases for maximum flow. For



(c)

Figure 3. *Continued.*

Table 4. Median percentage difference, MPD (%) between modelled and observed runoff calculated over all catchments for MAR, Q5 and Q95 (shaded cells denote underestimation of runoff by the model; ranking of the individual models across each indicator are in parantheses). The best performing individual model (or EM) is in bold.

Indicator	DBH	H08	LPJmL	MATSIRO	PCR-GLOBWB	WaterGAP2	EM
MAR	33.7 (6)	20.1 (5)	13.5 (4)	−12.6 (3)	7.8 (2)	**0.9 (1)**	13.9
Q5	26.2 (6)	15.3 (5)	12.2 (4)	−11.8 (3)	5.1 (2)	**−2.9 (1)**	9.6
Q95	37.3 (6)	19.4 (5)	19.1 (4)	−8.8 (2)	14.6 (3)	**5.0 (1)**	21.0

minimum flow, WaterGAP2 presents the least early bias (−0.43 months) and DBH shows the least late bias (0.28 months). Despite the good performance of DBH and LPJmL in terms of timing, these two models do show high relative difference values compared to other models (figure 4).

4. Discussion

4.1. Models' performance across hydrobelts

We found high variability between models in their ability to simulate MAR, Q5, Q95 and the magnitude of return period runoff values. The majority of models overestimate these hydrological indicators, with positive biases particularly acute in southern hydrobelts (SML and SDR). This can, in part, be explained by the general overestimation of precipitation values in climate forcing data in these regions [57], which in turn means the models overestimate runoff [21]. Nonetheless, previous studies emphasise that large ensemble spreads from GHMs and LSMs are not primarily due to errors in the (common) forcing data, but instead due to model structural uncertainty [19, 39]. Missing physical process representations in the models such as transmission loss [62] explain some of the differences between simulated and observed runoff. The underestimation of runoff by certain models, particularly in the NDR hydrobelt, may be a result of excessive evapotranspiration (as reported for MATSIRO by [19]). Moreover, the simulation of evapotranspiration has been shown to vary widely between the ISIMIP2a global-scale hydrological models [63].

Several models struggle to accurately simulate the timing and magnitude of long-term MMR in all months of the year in NDR and the first and/or last months in other hydrobelts. The relatively low levels of season-to-season variability in tropical and equatorial catchments EQT, SST and NST (i.e. the absence of a strong, predictable signal that can be modelled) may explain the weak performance in these hydrobelts. Additional factors may include the ability of models to sufficiently represent the range of soil properties influencing the generation and timing of runoff [64], as well as human-induced factors such as the operation of different reservoir management schemes [25]. In the BOR hydrobelt, the simulation and representation of snowmelt is likely to be the main cause of the early bias we reported. Temporal bias in snow-dominant regions (observed in [20, 21, 27, 40, 43, 65]) has previously been related to general errors in forcing data

Table 5. Mean weighted PBIAS (%) for MAR, Q5 and Q95, across hydrobelts. The best performing model (or EM) according to weighted PBIAS for each hydrobelt is in bold. Shaded cells denote where the runoff indiciator is underestimated by the model or EM. Hydrobelts are ranked according to the perfomance of the EM. Rows are ordered according to the mean latitude of each hydrobelt from north to south (BOR = boreal, NML = northern mid-latitude, NDR = northern dry, NST = northern subtropical, EQT = equatorial, SML = southern mid-latitude, SDR = southern dry and SST = southern subtropical).

Indicator	Hydrobelt (No. of catchments)	Mean weighted PBIAS (%)							Rank (based on EM)	Rank (based on best individual model)
		DBH	H08	LPJmL	MATSIRO	PCR-GLOBWB	WaterGAP2	EM		
MAR	BOR (14)	69	29	12	**-3**	-8	**3**	17	3	5
	NML (12)	61	16	4	-17	8	**1**	12	1	2
	NDR (2)	246	95	88	-84	191	**-18**	87	5	7
	NST (1)	242	237	204	**-41**	166	43	142	6	8
	EQT (3)	37	27	19	2	-10	**0**	13	2	1
	SST (4)	156	85	98	20	75	**2**	73	4	4
	SDR (2)	1964	166	1168	50	566	**9**	654	8	6
	SML (2)	936	100	603	-21	301	**-1**	320	7	3
	Median over all catchments	214	85	111	-11	103	**2**	79		
	Model rank	6	3	5	2	4	1			
Q5	BOR (14)	47	14	**-1**	-9	-13	-7	3	1	1
	NML (12)	52	16	3	-15	9	**-1**	8	2	2
	NDR (2)	153	70	59	-85	97	**-21**	43	4	6
	NST (1)	236	213	182	**-33**	142	44	127	7	8
	EQT (3)	32	25	17	**2**	-9	**-2**	10	3	3
	SST (4)	114	87	78	-6	23	**-3**	47	5	4
	SDR (2)	631	132	501	**-27**	186	-31	229	8	7
	SML (2)	338	61	235	-40	97	**-12**	110	6	5
	Median over all catchments	134	66	68	-21	60	-5	45		
	Model rank	6	4	5	2	3	1			
Q95	BOR (14)	134	55	54	10	**6**	21	50	3	3
	NML (12)	72	8	5	-20	12	**0**	14	1	1
	NDR (2)	734	241	262	-76	715	**27**	315	5	4
	NST (1)	321	304	271	**-55**	270	65	202	4	6
	EQT (3)	46	31	22	**2**	-10	3	17	2	2
	SST (4)	1108	64	696	86	317	**52**	432	6	5
	SDR (2)	25356	704	10547	794	4695	**440**	7432	8	8
	SML (2)	15396	918	7814	309	5200	**125**	5026	7	7
	Median over all catchments	528	153	266	6	293	40	258		
	Model rank	6	3	4	1	5	2			

Table 6. Median percentage difference, MPD (%) between modelled and observed runoff calculated over all catchments for different return periods (**shaded cells** denote underestimation of runoff by the model). The best performing model (or EM) is in bold.

Extreme flow	Return period	DBH	H08	LPJmL	MATSIRO	PCR-GLOBWB	WaterGAP2	EM
Maximum flow	2	122.7	52.4	75.9	-22.1	19.2	**-14.0**	39.5
	5	94.6	62.5	76.9	-20.8	13.2	**-12.8**	41.5
	10	90.0	68.4	77.5	-19.0	**6.8**	-12.3	38.9
	20	86.7	73.8	74.3	-18.1	**4.8**	-13.3	39.0
	25	85.9	75.4	73.3	-18.0	**4.3**	-13.6	39.3
Minimum flow	2	20.8	17.1	-10.8	**-2.9**	48.4	36.8	33.1
	5	12.6	15.6	-20.4	**-1.5**	58.7	54.6	39.1
	10	5.9	19.6	-25.5	**1.3**	53.7	57.8	43.5
	20	7.1	28.1	-17.8	**4.5**	50.1	59.8	45.0
	25	9.7	30.9	-18.4	**3.4**	51.1	61.4	47.5

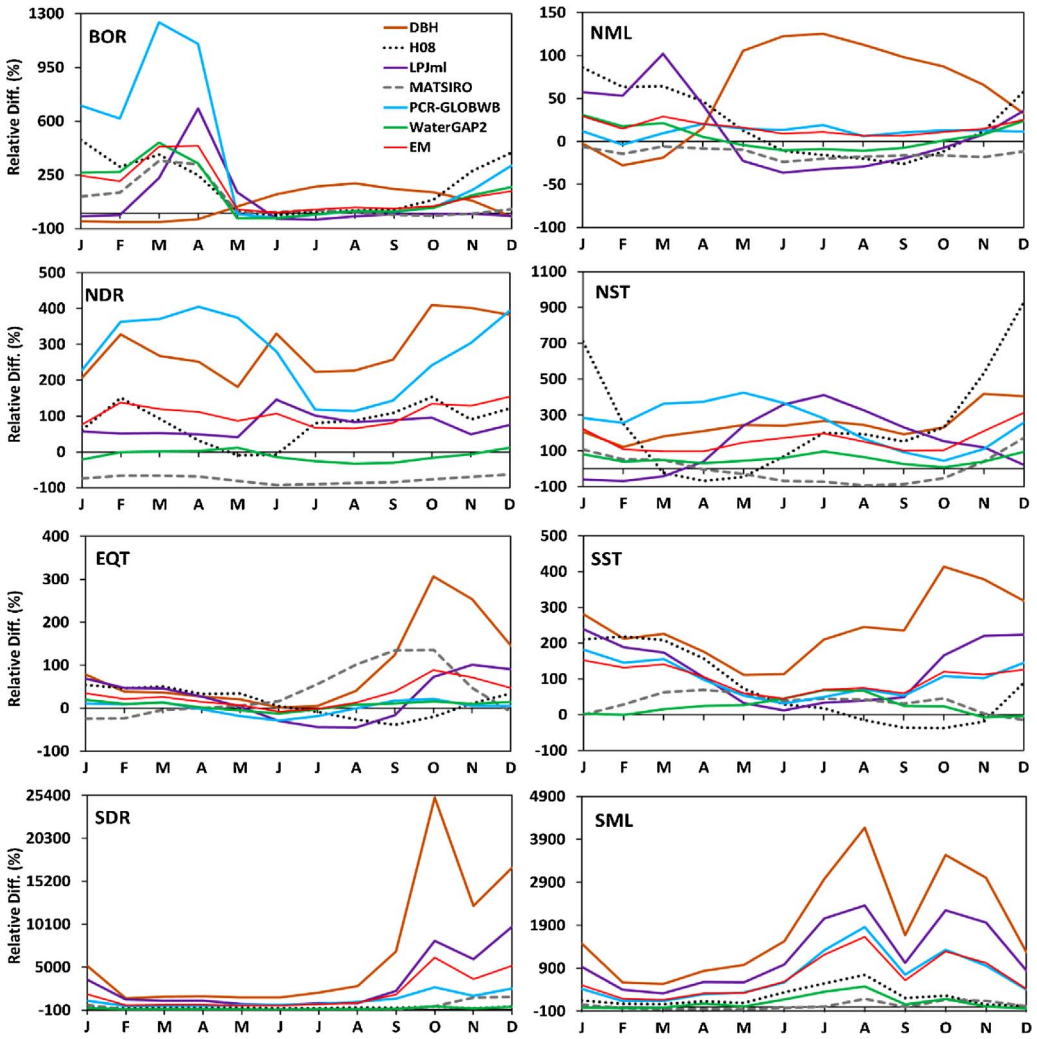


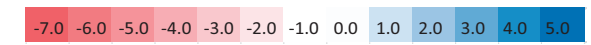
Figure 4. Mean weighted relative difference between simulated and observed MMR (seasonal cycle) for each hydrobelt (note that vertical scales differ) (BOR = boreal, NML = northern mid-latitude, NDR = northern dry, NST = northern subtropical, EQT = equatorial, SML= southern mid-latitude, SDR = southern dry and SST= southern subtropical).

(especially the underestimation of precipitation) or the absence/misrepresentation of processes that delay snowmelt in models. Some of these processes include the infiltration of meltwater into soils, the refreezing of meltwater over cold periods in the diurnal cycle, and ice-jams in rivers [21].

The higher timing bias we found for minimum runoff is related to the observed data's higher sensitivity to the timing of water abstractions and reservoir operation during low flow periods [25]. Nonetheless, it should be emphasised that the degree to which monthly flow is influenced by reservoirs depends on the ratio of reservoir storage and the annual flow. With large reservoirs, any model's ability to simulate storage and release of water can be more important than, for example, the timing of snowmelt.

Table 7. Hydrobelt mean (not weighted) early and late runoff timing bias (units are number of months). Negative (positive) values denote an early (late) bias. Values calculated over all catchments across each hydrobelt, or globally. The number of catchments affected by timing bias is in parantheses. Rows are ordered according to the mean latitude of each hydrobelt from north to south (BOR = boreal, NML = northern mid-latitude, NDR = northern dry, NST = northern subtropical, EQT = equatorial, SML = southern mid-latitude, SDR = southern dry and SST = southern subtropical).

		Maximum flow							Minimum flow						
Timing bias	Hydrobelt (No. Catchments)	DBH	H08	LPJmL	MATSIRO	PCR-GLOBWB	WaterGAP2	EM	DBH	H08	LPJmL	MATSIRO	PCR-GLOBWB	WaterGAP2	EM
Early bias	BOR (14)	-0.14 (2)	-0.21 (2)	-1.00 (12)	-0.43 (3)	-0.21 (3)	-1.00 (4)	-0.21 (3)	-1.86 (9)	-0.71 (2)	-2.07 (12)	-2.07 (11)	-1.36 (4)	-0.29 (2)	-1.57 (7)
	NML (12)	-0.08 (1)	-1.25 (7)	-1.67 (9)	-0.67 (6)	-0.58 (7)	-0.75 (7)	-0.75 (8)	-2.42 (4)	-0.08 (1)	-0.33 (4)	-0.17 (2)	-1.33 (2)	-0.67 (1)	-0.67 (1)
	NDR (2)	0.00 (0)	-2.00 (1)	-1.50 (2)	0.00 (0)	-1.00 (1)	-1.00 (1)	-1.00 (1)	-4.00 (1)	-1.50 (1)	-2.00 (1)	-4.00 (1)	-4.00 (1)	-0.50 (1)	-4.00 (1)
	NST (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	-1.00 (1)	0.00 (0)	-1.00 (1)	0.00 (0)	-1.00 (1)	0.00 (0)	0.00 (0)
	EQT (3)	-3.67 (3)	-0.67 (2)	-1.67 (3)	-2.67 (2)	-4.00 (3)	-0.67 (2)	-0.67 (2)	-1.33 (2)	-0.33 (1)	-1.00 (3)	0.00 (0)	-0.33 (1)	-0.33 (1)	-0.33 (1)
	SST (4)	-1.50 (1)	-1.00 (1)	-1.50 (1)	-0.75 (1)	-1.25 (2)	-1.00 (1)	-1.00 (1)	-1.50 (2)	-0.75 (1)	-2.00 (2)	-0.25 (1)	-1.25 (2)	-0.75 (1)	-1.00 (2)
	SDR (2)	-0.50 (1)	0.00 (0)	0.00 (0)	-0.50 (1)	0.00 (0)	-0.50 (1)	0.00 (0)	0.00 (0)	0.00 (0)	-1.00 (2)	-7.00 (2)	0.00 (0)	0.00 (0)	0.00 (0)
	SML (2)	-1.50 (2)	-0.50 (1)	-1.50 (2)	-1.50 (2)	-1.50 (2)	-0.50 (1)	-1.50 (2)	-0.50 (1)	0.00 (0)	-0.50 (1)	-1.00 (2)	-0.50 (1)	0.00 (0)	-0.50 (1)
	Mean of all catchments	-0.60 (10)	-0.73 (14)	-1.28 (29)	-0.73 (15)	-0.80 (18)	-0.83 (17)	-0.58 (17)	-1.88 (19)	-0.45 (6)	-1.30 (26)	-1.40 (19)	-1.28 (12)	-0.43 (6)	-1.10 (13)
Late bias	BOR (14)	1.07 (10)	0.36 (4)	0.00 (0)	0.07 (1)	1.14 (7)	0.50 (2)	0.21 (2)	0.07 (1)	0.93 (6)	0.07 (1)	0.50 (2)	0.64 (3)	1.71 (6)	0.07 (1)
	NML (12)	0.92 (5)	0.42 (3)	0.08 (1)	0.75 (5)	0.25 (3)	0.25 (2)	0.33 (4)	0.67 (6)	3.75 (9)	2.67 (6)	1.75 (7)	2.00 (6)	2.67 (5)	2.25 (7)
	NDR (2)	1.00 (1)	0.00 (0)	0.00 (0)	1.00 (2)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.50 (1)	0.50 (1)	1.00 (1)	0.00 (0)	0.50 (1)	0.50 (1)
	NST (1)	0.00 (0)	1.00 (1)	0.00 (0)	2.00 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (1)	0.00 (0)	5.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)
	EQT (3)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.33 (1)	0.00 (0)	2.00 (3)	0.00 (0)	0.00 (0)	0.00 (0)
	SST (4)	0.00 (0)	0.00 (0)	0.00 (0)	0.25 (1)	0.00 (0)	0.25 (1)	0.00 (0)	0.25 (1)	0.50 (2)	0.00 (0)	0.50 (1)	0.50 (2)	1.50 (2)	0.50 (2)
	SDR (2)	0.00 (0)	0.50 (1)	0.00 (0)	0.00 (0)	0.50 (1)	0.00 (0)	0.00 (0)	0.50 (1)	2.00 (2)	0.00 (0)	0.00 (0)	1.00 (2)	1.00 (1)	0.50 (1)
	SML (2)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	2.50 (2)	0.00 (0)	0.00 (0)	0.00 (0)	2.00 (2)	0.00 (0)
	Mean of all catchments	0.70 (16)	0.30 (9)	0.03 (1)	0.38 (10)	0.50 (11)	0.28 (5)	0.18 (6)	0.28 (9)	1.80 (24)	0.85 (8)	1.08 (15)	0.93 (13)	1.75 (18)	0.80 (12)



Whilst when aggregating across catchments WaterGAP2 is consistently the best performing model in terms of overall fit (assessed by mean weighted IPE, table 3), consistently the best for MAR, Q5 and Q95 when assessed by MPD (table 4), and almost always the best model for MAR, Q5 and Q95 when assessed by mean weighted PBIAS (table 5), it is not the best model for the magnitude of flows associated with different return periods (table 6) nor for the timing of seasonal flows (table 7). For these latter two hydrological indicators, the best performing models include MATSIRO, H08 and PCR-GLOBWB. This is in part due to the global-scale models having spatially generalised parameters. Thus they may perform differently in different locations and for different indicators. In addition, although they all use the same forcing data, differences in model structure and parameterisation (table S3) lead to different performances across different indicators [64, 66, 67]. Thus caution should be applied in presenting the results from only one model (including the ensemble mean) in future model applications, where changes in multiple hydrological indicators are presented. Instead, we recommend the ensemble spread is presented and/or models are weighted according to their performance for each hydrological indicator respectively [68].

Generally, the superior performance of WaterGAP2 can be attributed to its calibration with long term annual average river discharge. However, our results show that this approach does not necessarily guarantee optimal performance for all hydrological indicators, e.g. high and low flows. Moreover, due to limitations associated with the quality, length and global coverage of observed discharge data, WaterGAP2 was only calibrated for 1319 catchments, covering 54% of the global land area except Antarctica and Greenland [57]. Thus the model is unlikely to achieve optimal performance globally, for all hydrological indicators. The physically-based snow and soil scheme in MATSIRO may be the reason for its superior performance in snow-dominated areas. DBH and PCR-GLOBWB presented higher skill in simulating maximum runoff of higher return periods. DBH also achieved good performance in simulating the timing of the seasonal cycle.

4.2. Opportunities for model improvement

WaterGAP2, the only calibrated model in the ensemble, presents the greatest overall model skill. However, WaterGAP2 is not the best-performing model (with respect to IPE) in 13 of the 40 catchments—ten of these are in the BOR hydrobelt. The result arises from the calibration of WaterGAP2 to match observed long-term annual river discharge, meaning that the model is prone to timing errors in snow dominant regions [69]. Conversely, the superior, physically-based snow and soil scheme in (uncalibrated) MATSIRO is likely to be a key reason for its good performance in snow-dominated hydrobelts. That said, H08 uses a similar snow modelling scheme to MATSIRO and does not attain a similar level of performance in the BOR hydrobelt. This suggests that H08's vegetation, evapotranspiration and soil representation schemes (all of which are different to MATSIRO) may be limiting its performance.

Our results imply that robust calibration methods may improve model performance. However, a smaller error in runoff estimates for a present-day hydrological simulation (from an uncalibrated or calibrated model) does not necessarily mean that projections for the future from that model will be more certain than projections from a model with larger error. Whilst the calibration procedure for WaterGAP2 does lead to better performance for present-day hydrology in many cases, we note that the calibration of any hydrological model can encourage over-fitting and so act to compensate for structural errors and errors in atmospheric forcing data [69]. This could pose a problem if the models were used within a climate modelling framework. Therefore, whilst this study highlights the benefits of model calibration for improving simulations of present-day runoff, we also urge caution towards interpreting it as a definitive route to higher model certainty; especially when the models are used for delivering future projections [70].

DBH and PCR-GLOBWB were able to achieve a reduction in the magnitude of error in simulated maximum runoff with increasing return period. It was more common for the other models to show little difference in error across return periods. This suggests that these two models may include process representations that work better towards achieving very extreme high flows, compared with other models.

Our results show that even when including human impacts, there remain challenges with the accurate simulation of low runoff magnitudes and long term MMR. Therefore we recommend a comprehensive and systematic evaluation of the specific methods used to represent human impacts in models, for multiple rivers across the globe and at different locations along the river network, including: irrigation; dam operation rules [26]; the representation of reservoirs; the way in which withdrawn water is returned to the river network (including groundwater, surface water bodies and soils); and the sources where water is withdrawn from to meet water needs (i.e. groundwater, surface water, and desalinisation).

4.3. Performance of the EM and implications for future applications

In climate modelling and meteorological forecasting, the ensemble mean is often reported as outperforming individual models [34–36]. Our analysis shows that this is not the case with global-scale hydrological models. Even when excluding the weakest performing

model(s) from the ensemble an individual model still outperforms the EM in the majority of catchments. We only excluded up to two weakest performing models when computing the EM under different cases, so the generally consistent weak performance of the EM may be the result of other outlying models (in terms of their performance) having a large disproportionate influence on the EM.

Certain models may be consistently poor performers in certain climatic or physiographic settings, or over certain hydrological response ranges, and their inclusion in a model ensemble may act to limit the performance of the EM [71]. However, totally excluding the weakest model might be at the risk of missing other skills of that model. For example, whilst DBH performed poorly for some hydrological indicators, it performed well in simulating timing of the seasonal cycle. The approach of including/excluding the best/weakest performing models to calculate the EM could be extended to weighting methods [68, 72–74]. This, however, raises difficult questions about how the 'best' weighting strategy and combination of weights can be determined *a priori*.

The variable performance of the EM that we report here means that a decision should not be taken *a priori* to use the EM as the basis of model evaluation and/or climate change impact assessments [1, 21, 39, 75] without considering its performance relative to the models it is summarising, because an individual model may perform significantly better.

5. Conclusion and recommendations

We have presented a worldwide comparative evaluation of the performance of six global-scale hydrological models to simulate mean and extreme monthly runoff. In parallel with a companion study presented in this journal issue [25], it is the first such evaluation to use models run with human impacts and it is also the most comprehensive evaluation of extreme runoff (table 1). Our adoption of the hydrobelt classification system provided a feasible means of aggregating catchment-scale results around the axis of hydro-geographical similarities, as well as facilitating a comparison of model performance spatially worldwide.

We found a tendency for the majority of models to overestimate MAR and all indicators of upper and lower extreme runoff. The models overestimate low flows (Q95) considerably more than they overestimate high flows (Q5) but on the other hand, the models overestimate minimum flow return periods to a lesser degree than they do for maximum flows. *Either way, the overestimation of runoff is a key issue that we recommend is addressed by the global-scale hydrological modelling community.* Whilst the incorporation of human activities into global-scale hydrological models has been shown to enhance model simulation capabilities [25], our evaluation leads us to *recommend that the global-scale hydrological modelling community pursue efforts to improve the representation of low runoff and the models' ability to predict the magnitude and timing of seasonal cycles.*

We have highlighted which models perform particularly well for certain hydrological indicators, and in which hydrobelts, and discussed potential solutions to improving model performance. Whilst calibration can deliver some improvements it is particularly challenging for global-scale models due to the paucity of global coverage of long-term and complete observed runoff records. Therefore *we recommend that efforts are made towards improving the quality [76, 77] and global coverage of observed runoff records and in turn technical approaches to model calibration are explored.*

Other model improvements can be achieved through better quality input and evaluation datasets, the inclusion of missing physical processes, and better representation of existing processes in the models. *We recommend that the global-scale hydrological modelling community explore the process parametrisations within their models to help inform their decisions on future model development and that they consider running perturbed parameter ensembles [78] to explore the uncertainties associated with those parametrisations.*

While the EM is a straightforward, widely used means of summarising the performance of an ensemble of hydrological models, our results highlight the limitations of this approach. Therefore we recommend the exploration of alternatives to the EM such as weighting models based upon their performance [68]. Nevertheless, the models that comprise the ensemble we evaluated here, represent the state-of-the-art, and multi-model ensembles may well be the best way to capture some of the existing uncertainties in representing the hydrological cycle at the global scale. Therefore we recommend that future studies adopt an ensemble approach so that the spread of possible outcomes, based upon current scientific modelling state-of-the-art, is known. The value of ensembles lie in their offering of a suite of models that can be evaluated with respect to a specific question that might need addressing. For example, if droughts were of interest, then one may possibly select a subset of models that is different from a subset that might be used if peak flows were of interest. Such a procedure can of course only be undertaken if the full range of model simulations are available in the first place.

The models evaluated here are under continuous development and we expect that their performance will improve as developers address known shortcomings and as observed data improves in quality. Model improvements will then in turn lead to more precise and accurate representation of hydrological patterns across the globe.

Acknowledgments

This work has been conducted under the framework of the Inter-Sectoral Impact Model Intercomparison Project, phase 2a (ISIMIP2a), so our thanks go to the modellers who submitted results to this project. The ISIMIP2a was funded by the German Ministry of Education and Research, with project funding reference number 01LS1201A. Data is available from [51]. We also thank the Global Runoff Data Centre (GRDC) for making available the observed runoff data. JZ was supported by the Islamic Development Bank and a 2018 University of Nottingham Faculty of Social Sciences Research Outputs Award. IH was supported by grant no. 243803/E10 from the Norwegian Research Council. JS was supported within the framework of the Leibniz Competition (SAW-2013 PIK-5) and by the EU FP7 HELIX project (grant no. 603864). JL was supported by the National Natural Science Foundation of China (41625001, 41571022), the Beijing Natural Science Foundation Grant (8151002), and the Southern University of Science and Technology (Grant no. G01296001). GL was supported by the Office of Science of the US Department of Energy as part of the Integrated Assessment Research Program. PNNL is operated by Battelle Memorial Institute for the US DOE under contract DE-AC05-76RLO1830. NH and YM were supported by the Environment Research and Technology Development Fund (S-10) of the Ministry of the Environment, Japan. HK and TK were supported by Japan Society for the Promotion of Science KAKENHI (16H06291).

ORCID iDs

Jamal Zaherpour https://orcid.org/0000-0001-5497-4589
Simon N Gosling https://orcid.org/0000-0001-5973-6862
Hannes Müller Schmied https://orcid.org/0000-0001-5330-9923
Ted I E Veldkamp https://orcid.org/0000-0002-2295-8135
Rutger Danker https://orcid.org/0000-0003-2375-5468
Lukas Gudmundsson https://orcid.org/0000-0003-3539-8621
Naota Hanasaki https://orcid.org/0000-0002-5092-7563
Junguo Liu https://orcid.org/0000-0002-5745-6311
Taikan Oki https://orcid.org/0000-0003-4067-4678
Jacob Schewe https://orcid.org/0000-0001-9455-4159
Yoshihide Wada https://orcid.org/0000-0003-4770-2539

References

[1] Schewe J *et al* 2014 Multimodel assessment of water scarcity under climate change *Proc. Natl Acad. Sci. USA* **111** 3245–50
[2] Veldkamp T I E *et al* 2017 Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century *Nat. Commun.* **8** 15697
[3] Gosling S N and Arnell N W 2016 A global assessment of the impact of climate change on water scarcity *Clim. Change* **134** 371–85
[4] Arnell N W *et al* 2016 The impacts of climate change across the globe: A multi-sectoral assessment *Clim. Change* **134** 457–74
[5] Liu J *et al* 2017 Water scarcity assessments in the past, present, and future *Earth's Future* **5** 545–59
[6] van Huijgevoort M H J *et al* 2013 Global multimodel analysis of drought in runoff for the second half of the twentieth century *J. Hydrometeorol.* **14** 1535–52
[7] Prudhomme C *et al* 2014 Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment *Proc. Natl Acad. Sci. USA* **111** 3262–7
[8] Ward P J, Jongman B, Weiland F S, Bouwman A, van Beek R, Bierkens M F P, Ligtvoet W and Winsemius H C 2013 Assessing flood risk at the global scale: model setup, results, and sensitivity *Environ. Res. Lett.* **8** 044019
[9] Dankers R *et al* 2014 First look at changes in flood hazard in the inter-sectoral impact model intercomparison project ensemble *Proc. Natl Acad. Sci. USA* **111** 3257–61
[10] Gosling S N *et al* 2016 A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C *Clim. Change* **141** 577–95
[11] Arnell N W and Gosling S N 2016 The impacts of climate change on river flood risk at the global scale *Clim. Change* **134** 387–401
[12] Irvine P J *et al* 2017 Towards a comprehensive climate impacts assessment of solar geoengineering *Earth's Future* **5** 93–106
[13] Emerton R E *et al* 2016 Continental and global scale flood forecasting systems *Wires-Water* **3** 391–418
[14] Fraser E D G, Simelton E, Termansen M, Gosling S N and South A 2013 Vulnerability hotspots: Integrating socio-economic and hydrological models to identify where cereal production may decline in the future due to climate change induced drought *Agric. Forest. Meteorol.* **170** 195–205
[15] Elliott J *et al* 2014 Constraints and potentials of future irrigation water availability on agricultural production under climate change *Proc. Natl Acad. Sci. USA* **111** 3239–44
[16] Rockstrom J, Falkenmark M, Karlberg L, Hoff H, Rost S and Gerten D 2009 Future water availability for global food production: The potential of green water for increasing resilience to global change *Water Resour. Res.* **45** 1–16
[17] Arnell N W, Brown S, Gosling S N, Hinkel J, Huntingford C, Lloyd-Hughes B, Lowe J A, Osborn T, Nicholls R J and Zelazowski P 2014 Global-scale climate impact functions: the relationship between climate forcing and impact *Clim. Change* **134** 475–87
[18] Warren R *et al* 2013 The AVOID programmes new simulations of the global benefits of stringent climate change mitigation *Clim. Change* **120** 55–70
[19] Haddeland I *et al* 2011 Multimodel estimate of the global terrestrial water balance: setup and first results *J. Hydrometeorol.* **12** 869–84
[20] Beck H E, de Roo A and van Dijk A I J M 2015 Global maps of streamflow characteristics based on observations from several thousand catchments *J. Hydrometeorol.* **16** 1478–501
[21] Beck H E, van Dijk A I J M, de Roo A, Dutra E, Fink G, Orth R and Schellekens J 2016 Global evaluation of runoff from ten state-of-the-art hydrological models *Hydrol. Earth Syst. Sci. Discuss.* **21** 2889–903
[22] Zhang Y Q, Zheng H X, Chiew F H S, Pena-Arancibia J and Zhou X Y 2016 Evaluating regional and global hydrological models against streamflow and evapotranspiration measurements *J. Hydrometeorol.* **17** 995–1010

[23] Meybeck M, Kummu M and Dürr H H 2013 Global hydrobelts and hydroregions: improved reporting scale for water-related issues? *Hydrol. Earth Syst. Sci.* **17** 1093–111

[24] Vorosmarty C J *et al* 2010 Global threats to human water security and river biodiversity *Nature* **467** 555–61

[25] Veldkamp T I E *et al* 2018 Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study *Environ. Res. Lett.* **13** 055008

[26] Masaki Y, Hanasaki N, Biemans H, Müller Schmied H, Tang Q, Wada Y, Gosling S N, Takahashi K and Hijioka Y 2017 Intercomparison of global river discharge simulations focusing on dam operation—multiple models analysis in two case-study river basins, Missouri–Mississippi and Green–Colorado *Environ. Res. Lett.* **12** 055002

[27] Zaitchik B F, Rodell M and Olivera F 2010 Evaluation of the Global Land Data Assimilation System using global river discharge data and a source-to-sink routing scheme *Water Resour. Res.* **46** 1–17

[28] Gupta H V, Kling H, Yilmaz K K and Martinez G F 2009 Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling *J. Hydrol.* **377** 80–91

[29] Elshorbagy A, Corzo G, Srinivasulu S and Solomatine D P 2010 Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology *Hydrol. Earth Syst. Sci.* **14** 1931–41

[30] Elshorbagy A, Corzo G, Srinivasulu S and Solomatine D P 2010 Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 2: Application *Hydrol. Earth Syst. Sci.* **14** 1943–61

[31] Dawson C W, Mount N J, Abrahart R J and Shamseldin A Y 2012 Ideal point error for model assessment in data-driven river flow forecasting *Hydrol. Earth Syst. Sci.* **16** 3049–60

[32] Phillips T J and Gleckler P J 2006 Evaluation of continental precipitation in 20th century climate simulations: The utility of multimodel statistics *Water Resour. Res.* **42** 1–10

[33] Gosling S N, Bretherton D, Haines K and Arnell N W 2010 Global hydrology modelling and uncertainty: running multiple ensembles with a campus grid *Phil. Trans. A Math. Phys. Eng. Sci.* **368** 4005–21

[34] Sanderson B M and Knutti R 2012 On the interpretation of constrained climate model ensembles *Geophys. Res. Lett.* **39** 1–6

[35] Tebaldi C and Knutti R 2007 The use of the multi-model ensemble in probabilistic climate projections *Phil. Trans. A Math. Phys. Eng. Sci.* **365** 2053–75

[36] Cloke H L and Pappenberger F 2009 Ensemble flood forecasting: a review *J. Hydrol.* **375** 613–26

[37] Dirmeyer P A, Gao X, Zhao M, Guo Z, Oki T and Hanasaki N 2006 GSWP-2: Multimodel analysis and implications for our perception of the land surface *Bull. Am. Meteorol. Soc.* **87** 1381–97

[38] Nohara D, Kitoh A, Hosaka M and Oki T 2006 Impact of climate change on river discharge projected by multimodel ensemble *J. Hydrometeorol.* **7** 1076–89

[39] Gudmundsson L *et al* 2012 Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe *J. Hydrometeorol.* **13** 604–20

[40] Gudmundsson L, Wagener T, Tallaksen L M and Engeland K 2012 Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe *Water Resour. Res.* **48** 1–20

[41] Oki T, Nishimura T and Dirmeyer P 1999 Assessment of annual runoff from land surface models using Total Runoff Integrating Pathways (TRIP) *J. Meteorol. Soc. Jpn.* **77** 235–55

[42] Milly P C, Dunne K A and Vecchia A V 2005 Global pattern of trends in streamflow and water availability in a changing climate *Nature* **438** 347–50

[43] Decharme B and Douville H 2006 Uncertainties in the GSWP-2 precipitation forcing and their impacts on regional and global hydrological simulations *Clim. Dyn.* **27** 695–713

[44] Decharme B and Douville H 2007 Global validation of the ISBA sub-grid hydrology *Clim. Dyn.* **29** 21–37

[45] Decharme B 2007 Influence of runoff parameterization on continental hydrology: Comparison between the Noah and the ISBA land surface models *J. Geophys. Res. Atmos.* **112** 1–13

[46] Materia S, Dirmeyer P A, Guo Z C, Alessandri A and Navarra A 2010 The sensitivity of simulated river discharge to land surface representation and meteorological forcings *J. Hydrometeorol.* **11** 334–51

[47] Zhou X Y, Zhang Y Q, Wang Y P, Zhang H Q, Vaze J, Zhang L, Yang Y H and Zhou Y C 2012 Benchmarking global land surface models against the observed mean annual runoff from 150 large basins *J. Hydrol.* **470** 269–79

[48] van Dijk A I J M, Peña-Arancibia J L, Wood E F, Sheffield J and Beck H E 2013 Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide *Water Resour. Res.* **49** 2729–46

[49] Yang H *et al* 2015 Multicriteria evaluation of discharge simulation in dynamic global vegetation models *J. Geophys. Res-Atmos.* **120** 7488–505

[50] Hattermann F F *et al* 2017 Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins *Clim. Change* **141** 561–76

[51] Gosling S *et al* 2017 *ISIMIP2a Simulation Data from Water (global) Sector* (Potsdam: GFZ Data Services) (http://doi.org/10.5880/PIK.2017.010)

[52] Tang Q H, Oki T, Kanae S and Hu H P 2007 The influence of precipitation variability and partial irrigation within grid cells on a hydrological simulation *J. Hydrometeorol.* **8** 499–512

[53] Hanasaki N, Kanae S, Oki T, Masuda K, Motoya K, Shirakawa N, Shen Y and Tanaka K 2008 An integrated model for the assessment of global water resources Part 1: Model description and input meteorological forcing *Hydrol. Earth Syst. Sci.* **12** 1007–25

[54] Rost S, Gerten D, Bondeau A, Lucht W, Rohwer J and Schaphoff S 2008 Agricultural green and blue water consumption and its influence on the global water system *Water Resour. Res.* **44** 1–17

[55] Pokhrel Y N, Koirala S, Yeh P J F, Hanasaki N, Longuevergne L, Kanae S and Oki T 2015 Incorporation of groundwater pumping in a global Land Surface Model with the representation of human impacts *Water Resour. Res.* **51** 78–96

[56] Wada Y, Wisser D and Bierkens M F P 2014 Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources *Earth Sys. Dyn.* **5** 15–40

[57] Müller Schmied H *et al* 2016 Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use *Hydrol. Earth Syst. Sci.* **20** 2877–98

[58] Kim H 2017 Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1) [Data set]. Data Integration and Analysis System (DIAS) (https://doi.org/10.20783/DIAS.501)

[59] Liu X, Tang Q, Cui H, Mu M, Gerten D, Gosling S N, Masaki Y, Satoh Y and Wada Y 2017 Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations *Environ. Res. Lett.* **12** 025009

[60] Tangdamrongsub N, Han S-C, Decker M, Yeo I-Y and Kim H 2018 On the use of the GRACE normal equation of inter-satellite tracking data for estimation of soil moisture and groundwater in Australia *Hydrol. Earth Syst. Sci.* **22** 1811–29

[61] Seibert J 2001 On the need for benchmarks in hydrological modelling *Hydrol. Process.* **15** 1063–4

[62] Gosling S N and Arnell N W 2011 Simulating current global river runoff with a global hydrological model: model revisions, validation, and sensitivity analysis *Hydrol. Process.* **25** 1129–45

[63] Wartenburger R *et al* 2018 *Environ. Res. Lett.* in press (https://doi.org/10.1088/1748-9326/aac4bb)

[64] Döll P, Douville H, Güntner A, Müller Schmied H and Wada Y 2015 Modelling freshwater resources at the global scale: challenges and prospects *Surv. Geophys.* **37** 195–221

[65] Lohmann D 2004 Streamflow and water balance intercomparisons of four land surface models in the North American land data assimilation system project *J. Geophys. Res.* **109** 1–22
[66] Hagemann S *et al* 2013 Climate change impact on available water resources obtained using multiple global climate and hydrology models *Earth Sys. Dyn.* **4** 129–44
[67] Beck H E, van Dijk A I J M, de Roo A, Miralles D G, McVicar T R, Schellekens J and Bruijnzeel L A 2016 Global-scale regionalization of hydrologic model parameters *Water Resour. Res.* **52** 3599–622
[68] Zaherpour J *et al* Optimising the combination of global-scale hydrological models from a multi-model ensemble *Environ. Soft. Model.* in review
[69] Müller Schmied H, Eisner S, Franz D, Wattenbach M, Portmann F T, Flörke M and Döll P 2014 Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration *Hydrol. Earth Syst. Sci.* **18** 3511–38
[70] Krysanova V, Donnelly C, Gelfan A, Gerten D, Arheimer B, Hattermann F and Kundzewicz Z W 2018 How the performance of hydrological models relates to credibility of projections under climate change *Hydrol. Sci. J.* **63** 696–720
[71] Shamseldin A Y, OConnor K M and Liang G C 1997 Methods for combining the outputs of different rainfall-runoff models *J. Hydrol.* **197** 203–29
[72] Giorgi F and Mearns L O 2002 Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the 'reliability ensemble averaging' (REA) method *J. Clim.* **15** 1141–58
[73] Qi Y, Qian C and Yan Z 2017 An alternative multi-model ensemble mean approach for near-term projection *Int. J. Clim.* **37** 109–22
[74] Gillett N P 2015 Weighting climate model projections using observational constraints *Phil. Trans. A Math. Phys. Eng. Sci.* **373** 1–8
[75] Do H X, Gudmundsson L, Leonard M and Westra S 2018 The global streamflow indices and metadata archive (GSIM)—Part 1: the production of a daily streamflow archive and metadata *Earth Syst. Sci. Data* **10** 765–85
[76] Gudmundsson L, Do H X, Leonard M and Westra S 2018 The global streamflow indices and metadata archive (GSIM)—Part 2: quality control, time-series indices and homogeneity assessment *Earth Syst. Sci. Data* **10** 787–804
[77] van Loon A F, van Huijgevoort M H J and van Lanen H A J 2012 Evaluation of drought propagation in an ensemble mean of large-scale hydrological models *Hydrol. Earth Syst. Sci.* **16** 4057–78
[78] Gosling S N 2013 Systematic quantification of climate change impacts modelling uncertainty. Impacts World 2013 *International Conference on Climate Change Effects* (Potsdam: PIK) pp 268–274

Paper II: Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models

Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models

Zaherpour Jamal[a*], Mount Nick[a], N Gosling Simon[a]

Rest of co-authors in alphabetic order:

Dankers Rutger [b], Eisner Stephanie [c], Gerten Dieter [d, e], Liu Xingcai [f], Masaki Yoshimitsu [g],

Müller Schmied Hannes [h, i], Tang Qiuhong [f], Wada Yoshihide [j]

[a] School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom

[b] Met Office, FitzRoy Road, Exeter, EX1 3PB, United Kingdom

[c] Center for Environmental Systems Research, University of Kassel, Kassel, Germany

[d] Potsdam Institute for Climate Impact Research, Telegrafenberg, 14473 Potsdam, Germany

[e] Geography Dept., Humboldt-Universität zu Berlin, 10099 Berlin, Germany

[f] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

[g] Hirosaki University, Bunkyocho-3, Hirosaki, Aomori, 36-8561, Japan

[h] Institute of Physical Geography, Goethe-University, Frankfurt, Germany

[i] Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Frankfurt, Germany

[j] International Institute for Applied Systems Analysis (IIASA) - Schlossplatz 1 - A-2361 Laxenburg, Austria

*Corresponding author:
Tel.: +44 115 951 5428, Fax: +44 (0)115 951 5249

E-mail addresses: lgxjz1@nottingham.ac.uk, zaherpour@gmail.com

Postal Address: Sir Clive Granger Building, School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom

Abstract

This study presents a novel application of machine learning to deliver optimised, multi-model combinations (MMCs) of Global Hydrological Model (GHM) simulations. We exemplify the approach using runoff simulations from five GHMs across 40 large global catchments. The benchmarked, median performance gain of the MMC solutions is 45% compared to the best performing GHM and exceeds 100% when compared to the EM. The performance gain offered by MMC suggests that future multi-model applications consider reporting MMCs, alongside the EM and intermodal range, to provide end-users of GHM ensembles with a better contextualised estimate of runoff. Importantly, the study highlights the difficulty of interpreting complex, non-linear MMC solutions in physical terms. This indicates that a pragmatic approach to future MMC studies based on machine learning methods is required, in which the allowable solution complexity is carefully constrained.

Keywords:

Machine Learning; Model Weighting; Gene Expression Programming; Global Hydrological Models; Optimization

Highlights:

- First use of intelligent multi-model combination (MMC) applied to a global hydrological model ensemble.
- MMC performs better than both any individual model from the ensemble and the ensemble mean, overall.
- The physical interpretation of MMCs is limited by the complexity of the non-linear weighting schemes that are developed.
- MMC based on Gene Expression Programming facilities controllable model mixing and provides explicit solutions.

Software and/or data availability:

The software applied in this study is GeneXpro Tools 4.0 (GXPT4) available at http://www.gepsoft.com/ and provided by Gepsoft Limited. Gepsoft is a predictive modelling software company located in Bristol, United Kingdom. Gepsoft was founded in 2000 to market the Gene Expression Programming (GEP) technique invented by Dr. Candida Ferreira, founder and currently director of Gepsoft. The first product to be released was a COM component (GEPSR 1.0 and 2.0) followed by the desktop application APS (Automatic Problem Solver). Gepsoft continues to develop this product, which was renamed GeneXproTools (Gene Expression Programming Tools) after version 4.0. Observed discharge data are available from the Global Runoff Data Centre (GRDC, http://grdc.bafg.de) and the global hydrological model simulations are available from Gosling et al. (2017).

Address:

Redwood House
65 Bristol Road, Keynsham
Bristol BS31 2WB
United Kingdom

Phone/Fax: +44 (0)117 325 1468
Email: sales@gepsoft.com

Conflict of interest

None

1. Introduction

Global Hydrological Models (GHMs) is a category of hydrological model that has been developed to facilitate simulations of runoff and river discharge at continental and global scales. They are designed to support assessments of the impact of climate variability and water management on freshwater resources across the global domain (Bierkens, 2015). GHMs can be instantiated as stand-alone hydrological models (Gosling and Arnell, 2011; Hanasaki et al., 2008b), but are also integral components of land surface models, LSMs (Guimberteau et al., 2018; Koirala et al., 2014) and dynamic global vegetation models, DGVMs (Jägermeyr et al., 2015; Thiery et al., 2017).

A GHM is a pragmatic trade-off between a faithful representation of the diversity of hydrological contexts and processes found across the world's catchments, and a generalised and simplified representation of hydrological processes that can support multi-decadal, generalised hydrological simulations at global scales. Compared to hydrological models designed for catchment-scale simulations (Arnold et al., 1993; Krysanova et al., 1998; Lindstrom et al., 2010), GHMs employ a coarser spatial discretisation (most commonly a 0.5 x 0.5 degree grid) and model the global land surface in a single instantiation. This means that they must use large numbers of spatially generalised parameters and employ a variety of simplifications to their representations of fundamental hydrological processes (Gosling and Arnell, 2011; Müller Schmied et al., 2014). For example, GHMs use conceptually-based soil moisture schemes that include probability distributed models (Moore, 2007) as well as 'leaky bucket' (Huang et al., 1996) methods (Hanasaki et al., 2008a, b) rather than the physically-based equations that underpin many catchment-scale models (Arnold et al., 1993; Graham and Butts, 2005). Similarly, GHMs may use a variety of simplified methods to estimate evapotranspiration (Wartenburger et al., 2018). Simplification is also evident in the snowmelt schemes used by GHMs, which can include degree-day methods (Gosling and Arnell, 2011) as well as more advanced energy balance approaches (Van Beek et al., 2008).

The global scope of GHMs, limited availability and quality of observed discharge data across the global domain and their use of spatially generalised parameters make them more difficult to calibrate than catchment hydrological models. Whilst examples of calibrated GHMs do exist (Müller Schmied et al., 2016), the majority of GHMs are uncalibrated (Gosling et al., 2016; Hattermann et al., 2017). This lack of calibration, coupled with the diversity of simplifications

employed in the hydrological process representations, means that there can be large inconsistency in the skill, bias and uncertainty of an individual GHM at different locations, as well as large inconsistencies between different GHMs at any given location (van Huijgevoort et al., 2013; Zaherpour et al., 2018). This spatial inconsistency means that GHMs risk becoming a "jungle of models" (Kundzewicz, 1986) in which it can be difficult to determine where a particular GHM output is likely to be capable of delivering optimal hydrological simulations. It also makes it dangerous to assume that any individual GHM will be an adequate basis for making projections at any given location, even if the model's ability to replicate observed data in particular catchments is enhanced through the acquisition of higher quality input data or efforts to improve process representations (Liu et al., 2007). To an extent, these arguments are also applicable to catchment hydrological models because whilst they have been shown to generally perform better than GHMs in model evaluation studies, ensembles of such models still result in an uncertainty range when the models are run with identical inputs (Hattermann et al., 2017; Hattermann et al., 2018).

The question of how to address the challenges of spatial inconsistency in hydrological models has been a feature of catchment-scale model research for several decades. In answering it, catchment modellers have recognised that reliance on a single, inconsistent model is inherently risky and should be avoided (Marshall et al., 2006; Shamseldin et al., 1997). Instead, they have developed ways to take advantage of the diversity of outputs (Clemen, 1989) generated by different models by using optimised mathematical combination methods to deliver a combined output that performs better than the individual models from which it was created (Hagedorn et al., 2005). This general approach—known as multi-model combination (MMC)—has been an important focus of catchment hydrological modelling studies over the last two decades (Abrahart and See, 2002; Ajami et al., 2006; Arsenault et al., 2015; Azmi et al., 2010; de Menezes et al., 2000; Fernando et al., 2012; Jeong and Kim, 2009; Marshall et al., 2007; Marshall et al., 2006; Moges et al., 2016; Nasseri et al., 2014; Sanderson and Knutti, 2012; Shamseldin et al., 1997). Given its demonstrable potential in catchment studies, it is perhaps surprising that the potential of applying MMC to GHMs has yet to be explored.

A wide range of techniques can be used to generate an MMC solution. The simplest example is the calculation of the arithmetic mean of the input models (commonly referred to as an Ensemble Mean (EM)). More sophisticated techniques employ weighted schemes (Arsenault et al., 2015), with the differential weightings applied to each input model reflecting their

relative strengths or weaknesses. The mathematical approach taken to determining the weights depends on the objective of the MMC. Where the primary objective is to minimise the difference between the MMC solution and observed data (i.e. maximise the predictive performance), without explicitly accounting for model or parameter uncertainty, the use of multiple linear regression (Doblas-Reyes et al., 2005) or machine learning algorithms to 'learn' the optimal set weights to apply to each MMC input model is a popular approach (Marshall et al., 2007). The use of algorithms such as artificial neural networks (ANNs) (Shamseldin et al., 1997; Xiong et al., 2001) or gene expression programming (GEP) (Barbulescu and Bautu, 2010; Bărbulescu and Băutu, 2009; Fernando et al., 2012) to define non-linear weighting schemes have proven to be particularly effective. This is down to their ability to generate optimised, non-linear schemes rapidly, without the need for any prior knowledge of the model parameters.

Where there is a desire to account for and minimise model and parameter uncertainty in the weighting scheme, Bayesian averaging methods are required (Ajami et al., 2007; Hoeting et al., 1999). These optimise the weights according to the posterior performance of the MMC solution under the prior probabilities of model parameter values (Duan et al., 2007; Vrugt and Robinson, 2007; Ye et al., 2004). However, these methods require knowledge of the probability density functions (PDFs) for each of the MMC's input model parameters (or at least their maximum likelihood estimates (Ye et al., 2004)). This makes their use in the MMC of GHMs problematic because the number of parameters used in GHMs is particularly high, the parameters vary considerably between models, and the PDFs of the parameters in a GHM can be extremely difficult to specify over a global domain. Consequently, the PDFs for GHM parameters are seldom specified and, in many cases, remain unknown.

In this study we explore the potential of MMC for addressing the challenge of spatial inconsistency in simulations by GHMs, by combining outputs from a diverse set of five GHMs using GEP (Ferreira, 2001; Ferreira, 2006). 40 optimised MMC solutions of monthly mean runoff are generated for the period 1971 – 2010, one for each of 40 large catchments that are distributed throughout the world's eight hydrobelts (Meybeck et al., 2013) (Figure 1). In each catchment, the MMC's ability to replicate the observed monthly runoff is compared against that of the EM and each of the five GHMs from which the MMC is derived, as well as, the best-performing individual GHM from the ensemble.

The objectives of the paper are, therefore, twofold: 1) to assess the levels of performance gain that GEP-based MMC solutions can deliver to GHMs in different hydro-climatic settings and; 2) to critique the extent to which interpretation of GEP expressions can provide useful insights about the relative strengths and weaknesses of the different input models. Our experiments provide a clear demonstration that optimised MMCs of GHMs can deliver substantial performance gains in all hydrobelts when compared to the EM or individual GHMs. However, they also highlight the challenges associated with delivering GEP-based MMCs that can be usefully and meaningfully interpreted.

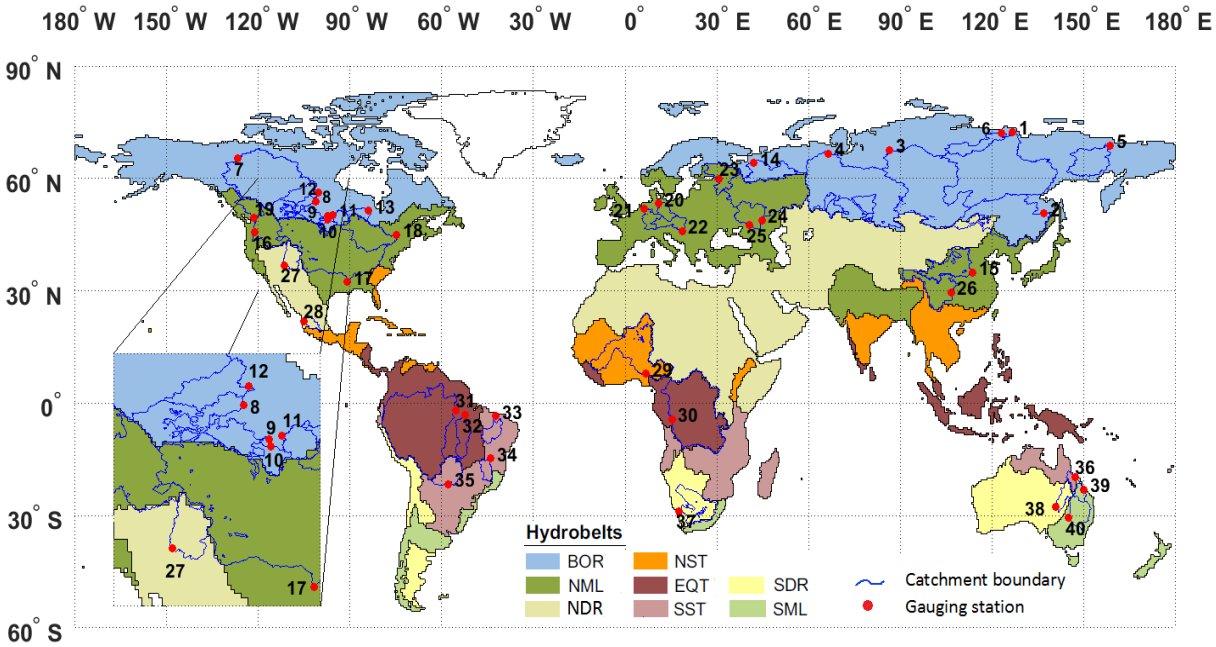


Figure 1. Locations of the 40 catchments (details in Table 1 and Table S1 in Supplementary Information) across the hydrobelt system defined in Meybeck et al. (2013). The hydrobelts are BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML=southern mid-latitude, SDR=southern dry and SST=southern subtropical.

2. MMC model inputs and study catchments

2.1. The GHMs

The study capitalises on the recent release of historical GHM simulations through the second phase of the Inter Sectoral Impacts Model Intercomparison Project (ISIMIP2a) (http://www.isimip.org; (Gosling et al., 2017)). ISIMIP2a provides a consistent modelling framework that ensures any inconsistencies between model outputs are a result of differences in the GHMs' structures or parameters. However, the GHMs providing ISIMIP2a

simulation products are not generally calibrated and are not accompanied by detailed information about the aleatory or epistemic uncertainties associated with each simulation, or the PDFs of model parameters from which it was generated. Consequently, this study is focused on the use of MMC to maximise predictive performance gain and not to minimise model or parameter uncertainty.

ISIMIP2a modelling groups used a standard protocol (available at: https://www.isimip.org/protocol/#isimip2a) to maximise consistency in the temporal and spatial resolutions of their simulations, the input climate forcings to the models, and the process representations (e.g. the simulation of human impacts such as dams, reservoirs and water abstractions (Masaki et al., 2017; Veldkamp et al., 2018)). The MMC solutions in the present study combine the simulation outputs from an ensemble of five input models: DBH, H08, LPJmL, PCR-GLOBWB (hereafter called PCRGLOBWB in the main text in order to avoid confusion by '-' in MMC expressions) and WaterGAP2 (Table S2).

All five input models to the MMC use the 2015 ISI-MIP2a data release and provide discharge simulations for the period 1971 – 2010 with input climate data provided by the Global Soil Wetness Project 3, GSWP3 (Kim, 2017). In all cases, the simulations are available at a daily time resolution and for a global land surface domain at 0.5^{o} x 0.5^{o} grid resolution. Conversion of gridded discharge data to catchment-mean monthly runoff was achieved by applying an area correction factor to the catchment area following the method detailed in Haddeland et al. (2011). It is important to note that, of the five models, only WaterGAP2 was calibrated against long-term mean annual runoff for a selection of catchments (Müller Schmied et al., 2016). The inclusion of calibrated WaterGAP2 may highlight the benefits (or otherwise) of calibrating global scale models.

2.2. Study catchments and observed data

For consistency and quality control we only selected catchments for which observed data is held by the Global Runoff Data Centre (GRDC; available from http://grdc.bafg.de). We identified study catchments based upon four selection criteria:

1- Catchments had to be larger than 100,000 km^2 to conform with the World Meteorological Organisation's definition of 'major' catchments (WMO, 2006). This

ensured that the catchments were of sufficient size to accommodate the output resolution of the models (0.5° x 0.5°).

2- The selected catchments had to cover all eight hydrobelts defined by Meybeck et al. (2013) (see Table S3).

3- Observed monthly discharge for the catchment had to be available for 25 years or longer, within 1971-2010 (the period over which the models were run) and without missing data. Other studies have allowed missing data (Beck et al., 2015; Beck et al., 2016; Milly et al., 2005), enabling them to include more catchments. We, however, preferred higher data quality, at the expense of number of catchments, because the use of longer, complete time-series facilitates more robust analyses.

4- Multiple gauges in individual catchments were excluded so that observed data from only one gauge, located at the most downstream location was used for each catchment.

The criteria resulted in the selection of 40 catchments. For each catchment, mean monthly river discharge was obtained for the most downstream gauge (Table 1), with mean monthly runoff subsequently derived by dividing the mean monthly discharge values by the area upstream of the gauge. Even though the selected catchments provided a good geographic coverage, the availability and quality of observed data resulted in a bias towards catchments in boreal and northern mid-latitude hydrobelts (Table 1). The least number of catchments in each hydrobelt is one (Niger basin in northern subtropical region), although this catchment does cover 20% of its hydrobelt. Two catchments were identified in NDR, SST, SDR, and SML hydrobelts. The low(er) number of catchments, or more precisely the area represented, particularly for NDR, SST, SDR, and SML hydrobelts, limits the extent to which our analyses and conclusions can be generalised across entire hydrobelts and the global domain.

Table 1. The 40 study catchments and their gauging sites.

No	GRDC Reference	River	Gauging Station	Total data length (years)	Catchment Area (km^2)	Hydro-belt
1	2903430	LENA	STOLB	32	2,460,000	BOR
2	2906900	AMUR	KOMSOMOLSK	26	1,730,000	BOR
3	2909150	YENISEI	IGARKA	32	2,440,000	BOR
4	2912600	OB	SALEKHARD	39	2,949,998	BOR
5	2998510	KOLYMA	KOLYMSKAYA	28	526,000	BOR
6	2999910	OLENEK	7.5KM DOWNSTREAM OF MOUTH OF RIVER PUR	39	198,000	BOR
7	4208150	MACKENZIE RIVER	NORMAN WELLS	30	1,570,000	BOR
8	4213550	SASKATCHEWAN	THE PAS	40	347,000	BOR
9	4213650	ASSINIBOINE	HEADINGLEY	40	153,000	BOR
10	4213680	RED RIVER	EMERSON	40	104,000	BOR
11	4213800	WINNIPEG RIVER	SLAVE FALLS	38	126,000	BOR
12	4214260	CHURCHILL RIVER	ABOVE GRANVILLE FALLS	36	228,000	BOR
13	4214520	ALBANY RIVER	NEAR HAT ISLAND	31	118,000	BOR
14	6970250	NORTHERN DVINA	UST-PINEGA	31	348,000	BOR
15	2180800	YELLOW	HUAYUANKOU	40	730,036	NML
16	4115200	COLUMBIA	THE DALLES, OREG.	40	613,830	NML
17	4127800	MISSISSIPPI	VICKSBURG, MISS.	37	2,964,252	NML
18	4143550	ST.LAWRENCE	CORNWALL(ONTARIO), NEAR MASSENA, N.Y.	40	773,892	NML
19	4207900	FRASER RIVER	HOPE	40	217,000	NML
20	6340110	LABE	NEU-DARCHAU	40	131,950	NML
21	6435060	RHINE RIVER	LOBITH	40	160,800	NML
22	6442600	DANUBE	MOHACS	29	209,064	NML
23	6972430	NEVA	NOVOSARATOVKA	40	281,000	NML
24	6977100	VOLGA	VOLGOGRAD POWER PLANT	39	1,360,000	NML
25	6978250	DON	RAZDORSKAYA	38	378,000	NML
26*	7222222	YANGTZE	CUNTAN	31	804,859	NML
27	4152450	COLORADO	LEES FERRY, ARIZ.	40	289,562	NDR
28	4356100	SANTIAGO	EL CAPOMAL	31	128,943	NDR
29	1834101	NIGER	LOKOJA	25	2,074,171	NST
30	1147010	ZAIRE	KINSHASA	40	3,475,000	EQT
31	3629000	AMAZONAS	OBIDOS	27	4,640,300	EQT
32	3630050	XINGU	ALTAMIRA	35	446,570	EQT
33	3650481	RIO PARNAIBA	LUZILANDIA	26	322,823	SST
34	3651805	SAO FRANCISCO	MANGA	37	200,789	SST
35	3667060	PARAGUAI	PORTO MURTINHO (FB/DNOS)	37	474,500	SST
36	5101200	BURDEKIN	CLARE	40	129,660	SST
37	1159100	ORANJE	VIOOLSDRIF	38	850,530	SDR
38	5410100	COOPER CREEK	CALLAMURRA	33	230,000	SDR
39	5101301	FITZROY	THE GAP	40	135,860	SML
40	5204250	DARLING RIVER	LOUTH	26	489,300	SML

*not included in GRDC database, obtained from local authorities.

3. Developing MMC solutions via Gene Expression Programming

3.1. GEP

GEP, which is detailed fully in Ferreira (2001, 2006), is an automated, machine learning algorithm that searches for optimal symbolic regression expressions to relate one or more

series of input data to an independent, observed series. Unlike standard linear regression, where the expression structure is limited to the input and output variables, numerical constants (the regression coefficients) and addition and multiplication operators; GEP expressions can incorporate the full range of arithmetic operators, as well as, mathematical functions (which are selected by the modeller). This makes it possible for GEP to relate input and observed data series via non-linear expressions. GEP expressions are modular; they are comprised of component trees (hereafter simply termed components) which are themselves made up of *bases* - the individual inputs, functions, constants and operators that comprise the component. Components are aggregated together using mathematical operators (usually addition) to form more complex expressions that can be readily translated into standard algebraic equations (Figure 2).

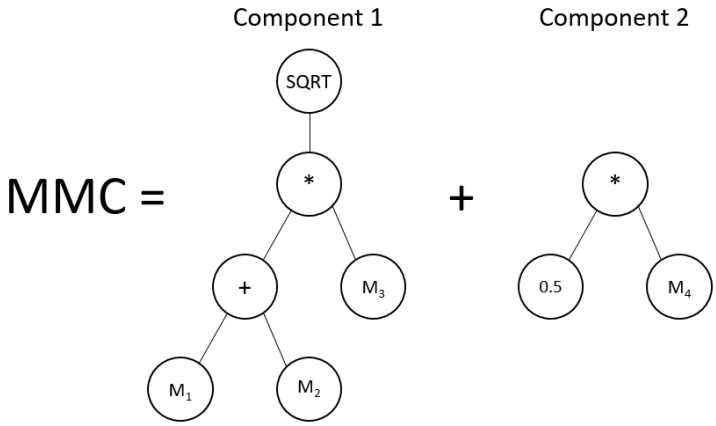


Figure 2. A GEP-based MMC solution (MMC) expressed as two components. The first component is made up of six bases and the second is made up of three. The MMC solution combines the four input models (M_1 to M_4) into an expression that includes a constant (0.5), operators (+ and *) and a non-linear function (SQRT). The equivalent algebraic expression for the solution is:

$$MMC = \sqrt{(M_1 + M_2) \times M_3} + 0.5 \times M_4$$

The GEP algorithm is an example of an iterative evolutionary algorithm that evolves a set of expressions to relate the input data series to the observed series (Figure 3). The algorithm begins by creating a random set of expressions which are then evolved in subsequent iterations. The set of expressions that GEP develops in each iteration are analogous to the genetic codes of biological 'organisms'. Each organism's likelihood of survival to the next iteration of the algorithm is dependent upon the extent to which its genetic code (i.e. the GEP expression) optimises the fit between the input data series and the observed data according to a pre-determined metric (a process known as 'training'). In this study we use the ideal point error metric (Dawson et al., 2012) to determine fitness, (see Section 3.4), due to its

incorporation of multiple error metrics into a single fitness measure. Each expression is then applied to an independent set of model input and the fit is validated to ensure that the expression can be generalised beyond the specific data from which it was learnt. If, at the end of an iteration, the best fitting expression is new, it is added to the candidate solution set which is output at the end of the GEP run. It is also preserved in the expression set (known as replication) whilst the remaining expressions are modified through adjustments to the bases in each component. These modifications can include mutation (where bases are randomly replaced with an alternative function, operator, input or constant) or transposition (where the arrangement of bases in the component is changed). In addition, entire components can be recombined by pairing them and exchanging their locations in the overall expression. The degree of modification allowed by each in any iteration is controlled by a rate set by the user. The number of iterations of the algorithm is also determined by a stopping point that is controlled by the user. This is usually a fixed number of iterations that is a large multiple of the number of data points in the observed series (i.e. to ensure adequate sampling of input data during training). Similarly, the user controls the complexity (equation size) of the expression by setting how many components it should include and the set of operators, functions and number of constants that can be included in the GEP expressions. The user settings applied in this study are provided in Table 2 and more detailed in Table S4.

Table 2. User settings for the GEP.

Control	**Setting used**
Number of components	3
Allowable operators	+, -, *, /
Allowable functions	Sqrt, Exp, x^2, x^3, Natural Log, Sine, Cosine
Number of constants allowed per component	2
Mutation rate	0.044
Transposition rate	0.1
Recombination rate	0.7
Stopping condition	100,000 iterations
Fit measure	IPE (see Section 3.4 below)

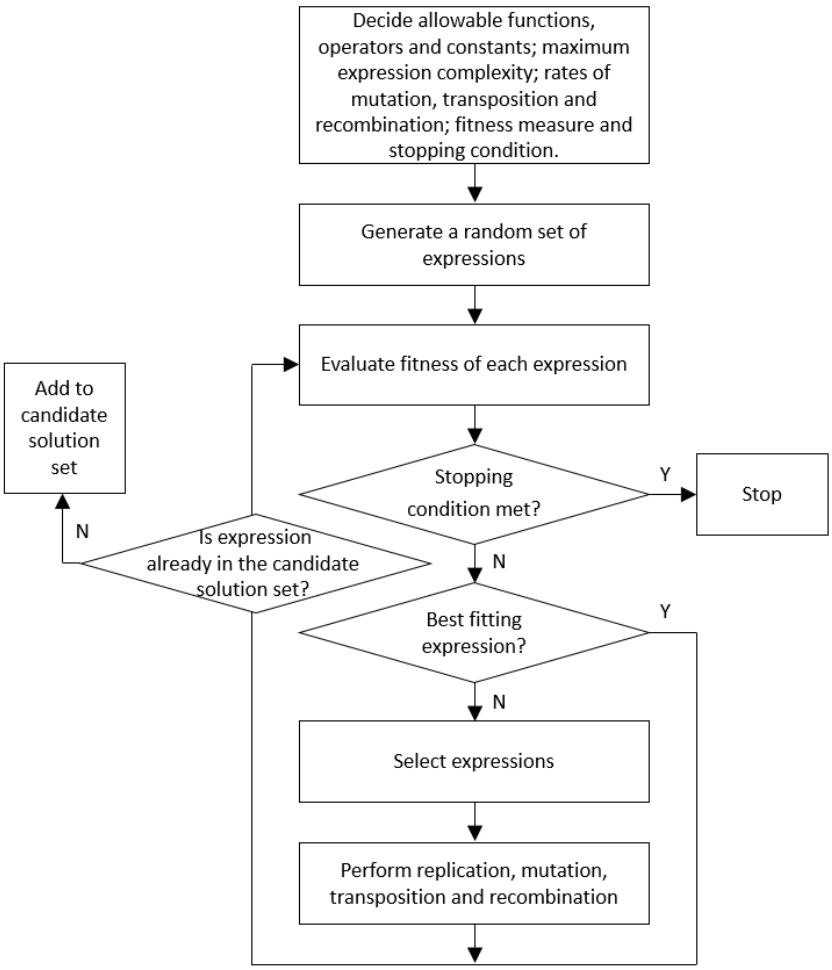


Figure 3. The GEP algorithm.

It is important to recognise that GEP expressions can provide MMC solutions that are more sophisticated than differential weighting schemes. The inclusion of non-linear functions and the relative lack of constraint on the form of the expression compared to multiple linear regression, for example, means that individual input models can be adjusted and combined in complex ways to exploit characteristic differences between model inputs. For example, Figure 4 shows an example of a GEP expression in which the difference between two input models (M_1 and M_2) is non-linearly weighted before being added back to M_2 in order to correct a substantial underestimation of peak discharge magnitude by both of the two input models. However, the extent to which the adjustments are purely mechanistic or informative about the advantages and limitations of different hydrological process representations in the models involved, will depend on the nature and complexity of the MMC solution.

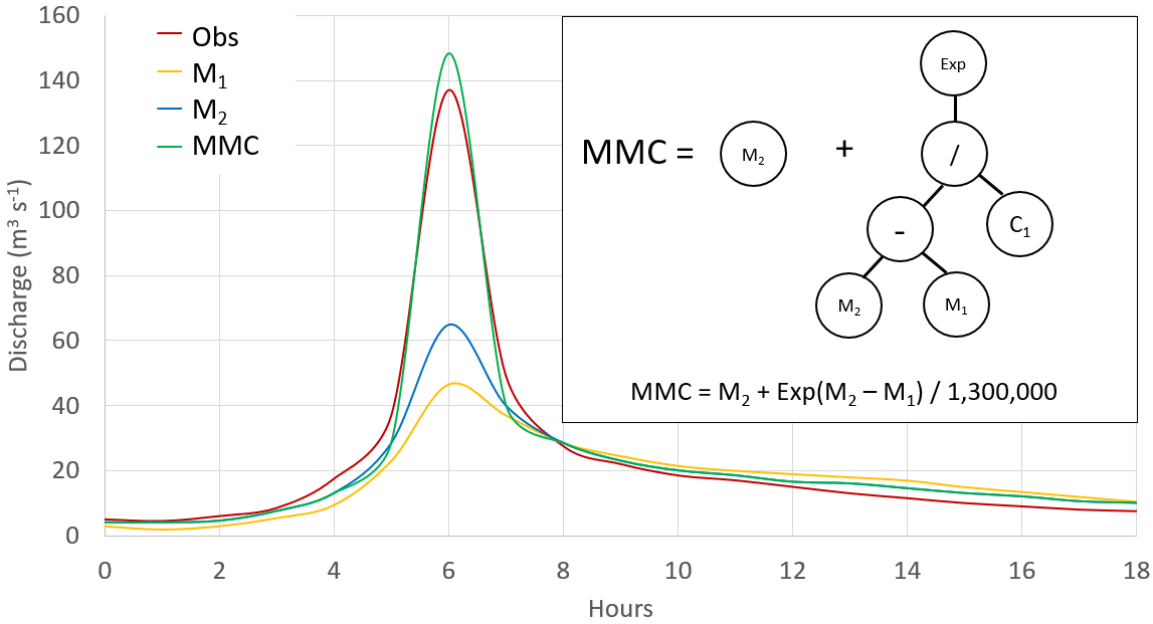


Figure 4. An example of a non-linear, GEP-based MMC solution in which the difference between two poorly performing models (M_1 and M_2) is used to correct the underestimation of peak discharge. C_1 in the second MMC component is a constant equal to 1,300,000.

3.2. Data splitting for GEP expression development

GEP's requirement for independent fit assessments during training and validation (see Section 3.1 above) means that the model input and observed data series from which the expressions will be evolved must be split into subsets. This is standard practice in machine learning methods (Phukoetphim et al., 2016; Wu et al., 2012; Wu et al., 2014). The way that the data are split is important. The GEP expressions that are developed will inevitably reflect the statistical characteristics of the in-sample, training data subsets. Conversely, their validity will depend on the statistical characteristics of the out-of-sample validation data subsets. It is, therefore, important to ensure that training and validation subsets are representative of the observed data and of each other.

Arbitrary data splitting approaches (e.g. taking the first 50% of a dataset for training and second for validation) cannot be guaranteed to achieve this. Therefore, a range of splitting methods have been developed (May et al., 2010; Snee, 1977; Wu et al., 2012) that are based on variations of cluster-based sampling or data proximity considerations. Tests of the effectiveness of alternative splitting techniques (Wu et al., 2012) have shown the DUPLEX method (Snee, 1977) to be particularly well suited to delivering representative data splits for use in model development by machine learning algorithms. It is, therefore, used throughout this study as the method for generating the data subsets required by GEP.

DUPLEX partitions data based on data proximity by sequential assignment of most distal data pairs to alternate sets so that consistency in the statistical characteristics of the subsets (e.g. equal representation of high and low flows) is maintained and bias during model development is minimised (Wu et al., 2012). We were consistent across all 40 catchments in the size of the training data subset which comprised 20 years in total for each catchment. The size of the validation data subset varied from catchment-to-catchment according to the length of the observed data series that was available (Table 1). However, it was never less than 60 months (5 years) and extended up to 240 months (20 years) in some catchments (Table S5).

3.3. Selecting a final MMC solution from the GEP candidate solution set

The end point of GEP is a set of "candidate" MMC solutions that contains the best-fitting expressions developed during iteration (Figure 3). These will vary in terms of their fit to the training and validation data, as well as, in their complexity. As a general rule, best-fitted expressions added to the candidate solution set from later iterations will be more complex than those added from earlier iterations. Similarly, the more complex solutions will tend to have higher levels of fit. However, more complex MMC solutions are harder to interpret and high levels of fit may indicate overfitting, which will limit the extent to which it can be generalised. Therefore, it is necessary to employ a procedure to select a final MMC solution from the candidate set that ensures it has both a good degree of fit and is parsimonious with respect to its complexity.

In the absence of a generally accepted method for doing this (Sudheer et al., 2002; Wagener et al., 2001), we devised a simple trade-off between candidate solution size (computed according to the number of inputs, constants, operators and functions in the expression) and fitness (Figure 5). Firstly, the fitness and equation size of each candidate solution was normalised to an error range between 0 and 1 by applying a linear maximum/minimum stretch. This enabled a normalised fitness/equation size coordinate to be defined for each solution. The Euclidean distance between this coordinate and the coordinate space origin (0, 0) was then computed, and the solution with the smallest Euclidean distance was selected as the final solution from the candidate set.

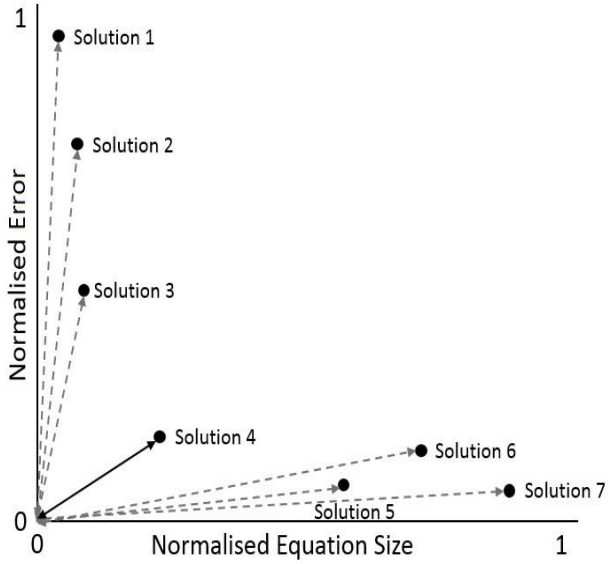


Figure 5. Selecting the GEP solution from a normalised fitness-equation space.
Solution 4 is selected because it has the smallest Euclidean distance from the origin.

3.4. Fit metrics

In this study, the fitness of each GEP expression during iteration, as well as, the performance of the final MMC solutions, is assessed using an integrated metric, called the ideal point error (IPE) (Dawson et al., 2012). IPE combines multiple error measures into a single metric so that multiple characteristics of fit are evaluated and summarised into a single value. The use of an integrated metric is particularly helpful during GEP's development of MMC solutions because it prevents the preferential development of expressions that minimise a specific characteristic of fit (Dawson et al., 2012; Pushpalatha et al., 2012). In order to improve the meaningfulness of comparisons of MMC performance across multiple catchments of varying sizes and located in different hydro-climatic zones, our instantiation of IPE also incorporates a consistent and transferrable benchmark. In this study, we follow Seibert (2001) and Zaherpour et al. (2018) and use the naïve t-1 model.

IPE delivers a single value that expresses the ratio of performance gain / loss of a MMC solution compared to the benchmark. In other words, it details how much better (or worse) the MMC solution has performed compared to the naïve model. The benchmarked IPE equation is presented in (1), IPEn, and is adapted from the original formula in Dawson et al. (2012). The negative reciprocal of the IPE score is used (3), where the performance of an MMC solution exceeds that of the benchmark. This maintains proportionality in comparisons

between IPE scores of MMC solutions that fail to perform as well as the benchmark and those whose performance exceeds it. In this study, Root Mean Square Error (RMSE), Mean Absolute Relative Error (MARE) and the Nash-Sutcliffe Coefficient of Efficiency (CE) were selected due to their different emphases on the overall pattern of fit (CE), low flows (MARE) and high flows (RMSE). Although IPE supports the use of differential weights to emphasise / de-emphasise individual metrics in the overall score, we here use equal weightings for all three metrics.

The IPE scores can range between -1 and -∞ (performance improvement over the benchmark model) and 1 and +∞ (performance loss over benchmark model). The IPE score is ratiometric – for example, an MMC solution that performs twice as well as the benchmark model will have an IPE score of -2 and a solution that performs twice as badly will have a score of 2. IPE would be 1 if MMC performs the same as the benchmark, whilst a model infinitely better than the benchmark would have an IPE of $-\infty$.

$$\mathrm{IPEn} = \left\{[1/3 * ((\mathrm{RMSE}/\mathrm{RMSE_b})^2 + (\mathrm{MARE}/\mathrm{MARE_b})^2 + ((\mathrm{CE}-1)/(\mathrm{CE_b}-1))^2)]^{\frac{1}{2}}\right\} \quad (1)$$

$$\mathrm{IPE} = \mathrm{IPEn} \qquad \text{IF IPEn} > 1 \quad (2)$$

$$\mathrm{IPE} = -1/\mathrm{IPEn} \qquad \text{IF IPEn} < 1 \quad (3)$$

Where:
IPEn = benchmarked IPE
RMSE = root mean squared error
MARE = mean absolute relative error
CE = Coefficient of Efficiency
b = benchmark data from the naïve (t-1) model

The IPE performance gain (PG) of an MMC solution (*A*) relative to either an individual GHM output or the GHM EM (*B*) can be expressed in percentage terms. The way that this is computed depends on the respective signs of the IPE scores for the solutions being compared (4-6). PG values are 0% where there is no difference in the performance gain / loss relative to the benchmark delivered by *A* over *B*. PG values are negative where performance gain is evident and positive where there is a loss of performance. For example, a PG value of -50% will indicate a gain in performance over the benchmark that is 50% larger for the MMC than its counterpart EM or best-performing GHM. Similarly, a PG value of 120% indicates that there is a 1.2 times reduction in performance of the MMC solution relative to its counterpart.

Where both *A* and *B* are either positive, or both negative:

$$\mathrm{MMC_{PG}} = 0 - (\mathrm{IPE_A} - \mathrm{IPE_B}) \times 100 \quad (4)$$

Where *A* is negative and *B* is positive:

$$\mathrm{MMC_{PG}} = 0 - \big((\mathrm{IPE_A} - 1) - (\mathrm{IPE_B} + 1)\big) \times 100 \quad (5)$$

Where *A* is positive and *B* is negative:

$$\mathrm{MMC_{PG}} = 0 - \big((\mathrm{IPE_A} + 1) - (\mathrm{IPE_B} - 1)\big) \times 100 \quad (6)$$

4. GHM, EM and MMC Performance

In the following section, we summarise the performance of individual GHMs and the EM, and present the performance gain/loss delivered by the MMC solutions. We pay particular attention to differences in performance gain across different hydrobelts to explore the spatial variability of MMC. All results pertain to validation data unless otherwise stated. Catchment-by-catchment results are detailed in the Supplementary Information. This includes performance metrics for all models for both training and validation data subsets (Table S8). In addition, observed versus simulated plots for mean annual runoff, the exceedance probability curves for each GHM, the EM and the MMC solution, and plots for each GEP expression component, are all provided in the Supplementary Information, Section S2.

4.1. GHM performance

To assess the performance of the different GHMs, the fit of the monthly simulated and observed runoff time series was computed against the validation data for each model as well as the EM and the MMC solution in each of the 40 catchments. The IPE metrics for each catchment are reported in Table 3 and the spatial distribution of the best individual GHM and the best overall model is mapped in Figure 6. This reveals that WaterGAP2 is the GHM most able to improve upon the naïve model benchmark. It outperforms the other GHMs in 32 catchments, and also performs better than the EM for the majority of catchments (34). This finding is perhaps unsurprising given that this is the only calibrated model in the ensemble. However, it is noteworthy that the dominant performance of WaterGAP2 is considerably less evident in the boreal hydrobelt compared to the other hydrobelts. Here both PCRGLOBWB and DBH are the best performing individual models in 5 of the 14 catchments. Across the

remaining hydrobelts, calibrated WaterGAP2 is out-performed by its uncalibrated counterparts in only 3 out of 26 catchments and these are spread across south sub-tropical, north dry belt and north mid-latitude without any apparent spatial pattern.

In several catchments (Assiniboine, Churchill, Yellow, St Lawrence, Neva, Don, Colorado, Rio Parnaiba, Paraguai, Oranje, Cooper Creek, Fitzroy and Darling) the IPE scores of one or more GHMs exceeds 10, indicating a failure to deliver a performance anywhere close to that of the naïve model benchmark. In the ephemeral catchments of Cooper Creek and Fitzroy the IPE scores for all GHMs are extremely high. This reflects the metric's sensitivity to proportionally large errors in runoff estimation which are particularly likely when runoff depths are close to zero. This is because a high ratio between the MARE of the individual GHMs and those of the naïve model benchmark translates directly into high overall IPE scores. Consequently, it is important to recognise that the exceptionally large IPE scores for the ephemeral Cooper Creek and the Fitzroy River are a result of periods of zero runoff having a disproportionate influence on their IPE scores.

4.2. EM Performance

Table 3 reveals that the ability of the EM to improve upon the naïve model benchmark exceeds that of any individual GHM in only 4 catchments. The failure of the EM to deliver significant performance gains in the majority of the study catchments implies that the specific sequencing of beneficial cancelling of relative over- and under-estimation of runoff (e.g. Figure 4) by individual GHMs necessary to facilitate the gains is not present in the ensemble of GHM outputs. Indeed, the tendency of the four uncalibrated GHMs to over-estimate runoff, both for mean runoff and hydrological extremes, is evident in observed versus simulated plots of mean annual, and Q5 (high flow) and Q95 (low flow) runoff (Figure 7).

Table 3. IPE scores for individual GHMs, EM and MMC for the validation period in each catchment. Models that outperformed the naïve model benchmark are shaded in grey. The best performing model in each catchment is indicated in bold.

Catchment No.	River	Hydrobelt	DBH	H08	LPJmL	PCRGLOBWB	WaterGAP2	EM	MMC
1	LENA	BOR	1.58	2.04	1.42	1.51	-1.22	1.15	**-2.00**
2	AMUR	BOR	3.06	1.91	1.33	1.34	1.17	1.07	**-1.49**
3	YENISEI	BOR	1.18	-1.54	1.25	-1.54	-1.72	-1.69	**-2.33**
4	OB	BOR	8.42	4.75	13.92	2.61	2.50	3.53	**-1.32**
5	KOLYMA	BOR	-1.23	1.10	1.18	1.27	2.30	-1.19	**-2.38**
6	OLENEK	BOR	**-1.47**	6.32	12.45	17.70	3.94	8.12	-1.15
7	MACKENZIE RIVER	BOR	4.50	1.85	3.37	-1.30	1.07	**-1.39**	-1.33
8	SASKATCHEWAN	BOR	61.42	5.75	27.03	8.16	1.43	8.97	**1.03**
9	ASSINIBOINE	BOR	384.84	44.46	512.25	28.94	1.57	85.79	**1.06**
10	RED RIVER	BOR	6.56	1.62	4.83	2.12	1.52	2.77	**-1.25**
11	WINNIPEG RIVER	BOR	24.16	4.85	5.05	**1.55**	1.67	2.29	1.63
12	CHURCHILL RIVER	BOR	297.53	50.12	32.22	25.65	3.60	17.08	**3.10**
13	ALBANY RIVER	BOR	2.82	-1.03	2.76	-1.33	1.73	-1.22	**-1.67**
14	NORTHERN DVINA	BOR	1.48	-1.04	2.14	-1.15	-1.52	-1.54	**-2.27**
15	YELLOW	NML	23.41	5.50	7.42	44.87	1.49	9.75	**1.16**
16	COLUMBIA	NML	4.25	2.12	3.11	1.75	-1.11	**-1.28**	-1.20
17	MISSISSIPPI	NML	4.98	-1.56	1.07	1.70	-1.89	1.16	**-2.04**
18	ST.LAWRENCE	NML	375.18	75.36	56.89	13.97	7.09	31.61	**2.47**
19	FRASER RIVER	NML	1.18	2.53	4.06	1.15	1.16	1.30	**-1.61**
20	LABE	NML	6.70	4.11	2.98	7.67	**-1.47**	3.10	-1.45
21	RHINE RIVER	NML	2.63	3.29	1.50	1.39	-1.96	1.15	**-2.50**
22	DANUBE	NML	4.02	2.72	1.25	2.07	-1.89	-1.08	**-2.22**
23	NEVA	NML	83.42	25.58	12.19	8.94	2.42	4.74	**1.09**
24	VOLGA	NML	6.80	2.79	1.89	-1.35	-1.75	1.52	**-2.00**
25	DON	NML	83.47	39.91	58.79	100.12	1.54	37.14	**1.23**
26	YANGTZE	NML	-2.44	-1.10	-1.05	2.81	-3.03	-1.15	**-4.17**
27	COLORADO	NDR	52.90	2.50	12.10	8.50	4.59	6.44	**2.22**
28	SANTIAGO	NDR	15.13	8.26	3.84	14.97	1.35	7.33	**1.16**
29	NIGER	NST	9.67	10.65	10.04	3.61	-1.37	4.86	**-1.79**
30	ZAIRE	EQT	8.28	5.92	3.89	2.47	1.78	2.40	**1.42**
31	AMAZONAS	EQT	2.05	1.46	2.60	3.44	-1.09	1.27	**-1.85**
32	XINGU	EQT	5.89	4.65	4.89	1.12	1.16	2.65	**1.04**
33	RIO PARNAIBA	SST	48.77	70.84	63.41	8.39	1.46	25.41	**-2.27**
34	SAO FRANCISCO	SST	4.81	3.48	1.89	2.25	-1.64	1.94	**-1.92**
35	PARAGUAI	SST	136.88	153.69	108.09	98.44	**8.00**	78.53	8.51
36	BURDEKIN	SST	6.87	1.44	3.13	2.03	1.65	2.92	**-1.35**
37	ORANJE	SDR	83.15	7.09	81.10	46.42	2.26	31.15	**2.04**
38	COOPER CREEK	SDR	6993.0	149.00	2578.0	625.00	107.00	2089.0	**20.05**
39	FITZROY	SML	641.17	52.61	447.46	270.32	38.47	290.00	**30.64**
40	DARLING RIVER	SML	200.58	6.95	92.30	35.20	-1.54	41.93	**-1.64**

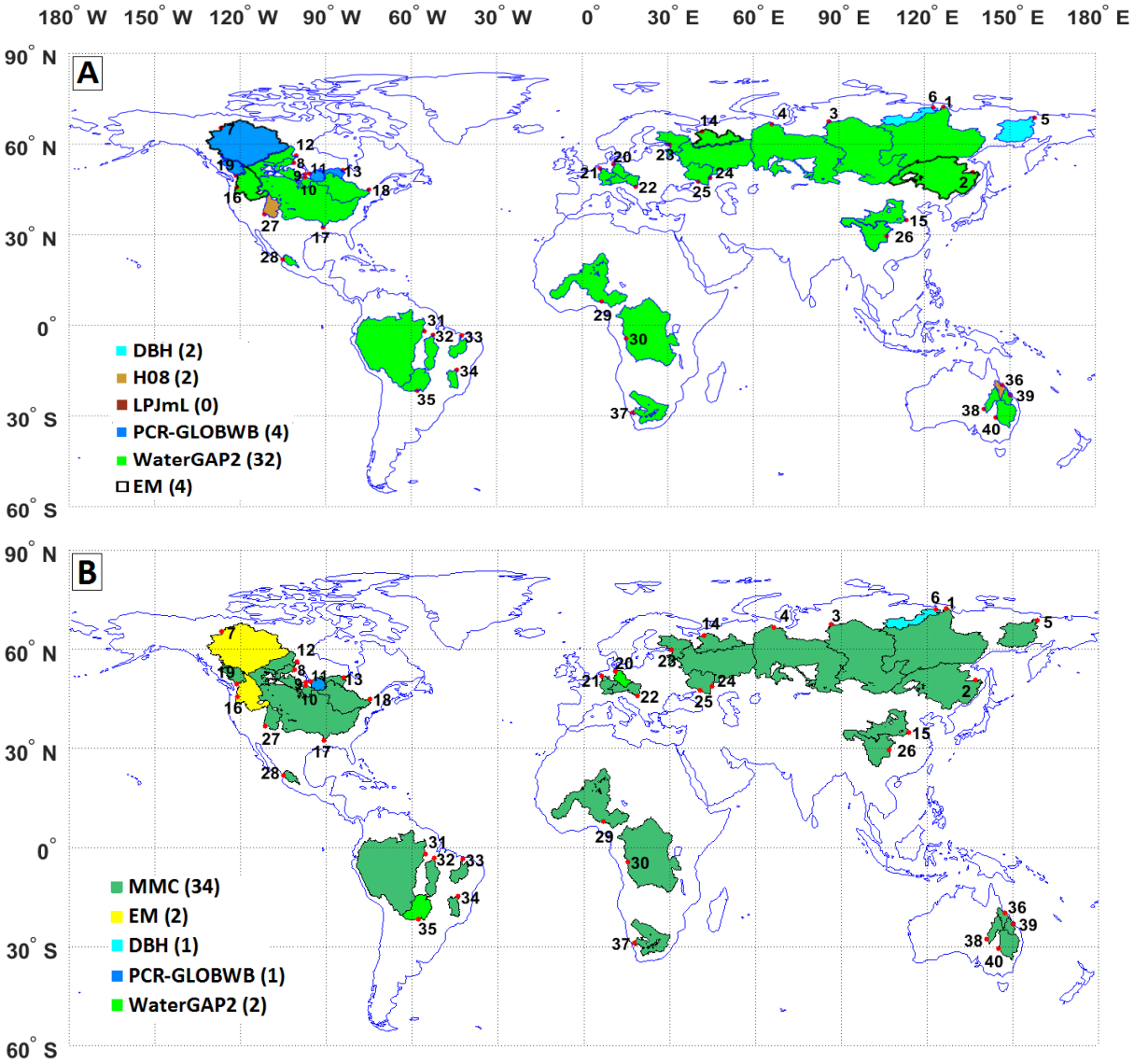


Figure 6. The best performing individual GHM (A); four catchments (2, 7, 14 and 16) where the EM outperforms the individual models have borders in bold black lines (in these cases the catchment is still shaded according to the best performing individual GHM). The best performing overall model/MMC (B); the two catchments where the EM is the best are shaded in yellow. Numbers in parentheses denote number of catchments where each model performs best.

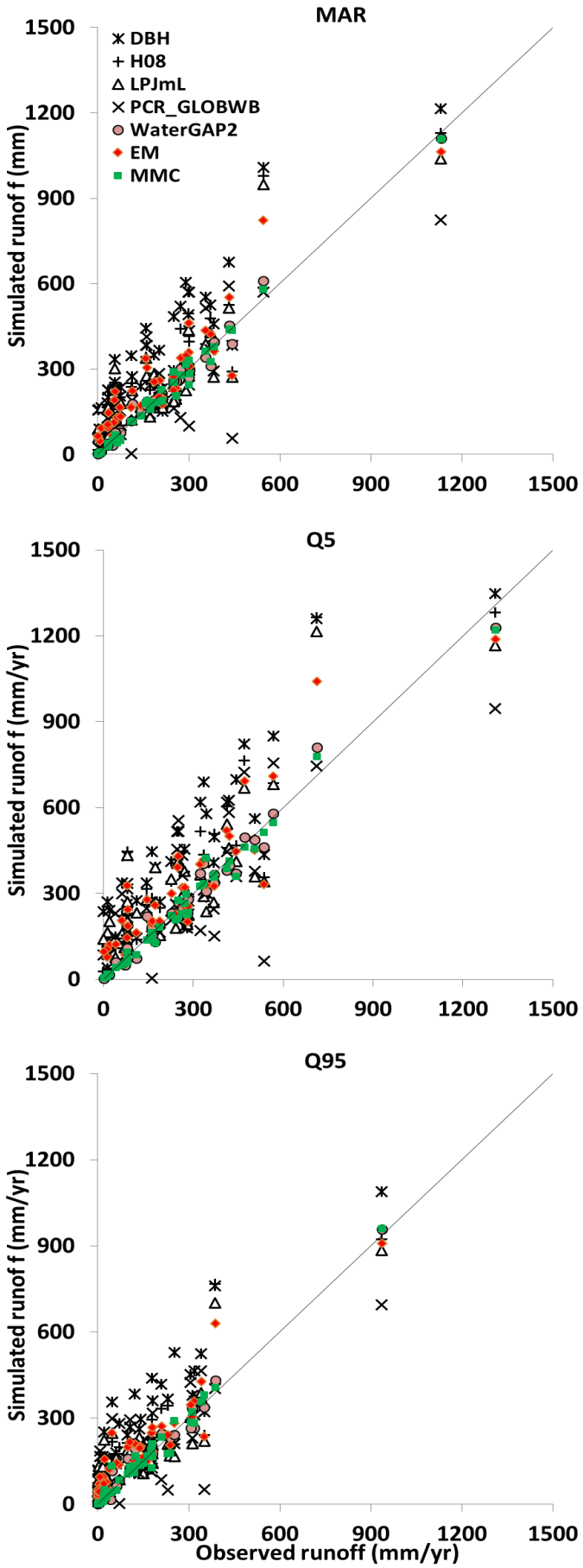


Figure 7. Plots of observed versus simulated runoff for each GHM, the EM and the MMC for mean annual runoff, Q5 and Q95.

The positive biases amongst the GHMs from which the EM is calculated also precludes a better performance by the EM relative to the best performing GHM for each catchment. Even in the four catchments where the EM outperforms the best GHM (Amur, Mackenzie, Northern Dvina and Columbia), the differences in IPE between the EM (IPE_{EM}) and the best performing GHM (IPE_{GHM}) are marginal (see Table 3): Amur 1.07 (IPE_{EM}) and 1.17 ($IPE_{WaterGAP2}$); Mackenzie -1.39 (IPE_{EM}) and -1.30 ($IPE_{PCRGLOBWB}$); Northern Dvina -1.54 (IPE_{EM}) and -1.52 ($IPE_{WaterGAP2}$); Columbia -1.28 (IPE_{EM}) and -1.11 ($IPE_{WaterGAP2}$). This highlights the importance of recognising that the potential performance gains that can be realised through the use of the EM is limited to the specific configuration of relative directional biases within the outputs from the individual models from which it is computed. Indeed, we would argue that the EM, where computed, should always be contextualised with respect to such biases.

4.3. MMC Performance

IPE scores for the validation data subset for individual GHMs, the EM and the MMC solution are presented for each catchment in Table 3. The MMC solutions, and their GEP expressions for each catchment are detailed in Table 4 along with the performance gain of the MMC solutions (MMC_{PG}).

The tables demonstrate the substantial improvements in IPE that are achieved by MMC relative to individual GHMs and the EM. Indeed, MMC solutions attain the best IPE scores in 34 of the 40 catchments. Observed versus simulated plots (Figure 7) highlight the consistency of the better MMC performance across mean and extreme hydrological indicators. Significant outliers amongst the MMC data are few and the magnitude is generally small. There is also little evidence of systematic over or underestimation bias in the mean annual runoff and Q95 data, although the tendency of the MMC data to plot just beneath the 1:1 line in the Q5 plot does indicate that the MMC solutions produce a general underestimation of the largest hydrological events across the study catchments. i.e. flood hazard events.

MMC performance gain (MMC_{PG}) scores reveal the MMC solutions deliver performance gains of > 50% in half (20) of the catchments and a median performance gain of 46% across all 40 catchments. If the outliers of Cooper Creek, Darling and Fitzroy River are omitted, the median MMC_{PG} is 40% and performance gains of > 50% are recorded in 17 of 37 catchments.

MMC performance gains are, however, not ubiquitous. In four catchments (Olenek, Winnipeg, Labe and Paraguai) the performance gain for the best performing GHM is 15% greater than for the MMC on average. Similarly, in 2 catchments (Mackenzie and Columbia) the EM delivers performance gains over the MMC equal to 5% and 7% respectively. These results highlight the fact that GEP-based MMC performance gain is dependent on the availability of a range of model inputs with relative inconsistencies that can be exploited by the optimisation algorithm. It also indicates that the success (or otherwise) of GEP-based MMC is dependent on the selection of appropriate constraints on expression size and structure, as well as the range of functions that are allowed. It is also noteworthy that there is a discrepancy in the magnitude of the MMC performance gains for the northern and southern hemisphere catchments. The median and mean MMC_{PG} relative to the best performing GHM for the southern hemisphere catchments (Fitzroy and Cooper Creek omitted) are -29% and -217% respectively. This is considerably smaller than their northern hemisphere equivalents; -41% and -119%.

When summarised by hydrobelt (Table 5), it is evident from the median MMC_{PG} score that MMC solutions generally deliver substantial improvements over their EM and GHM counterparts in all hydrobelts. The MMC performance gain is largest against the EM than the best-performing GHM in all hydrobelts. It is always several orders of magnitude greater and reflects the limiting impact that positive biases in GHM outputs have on the performance of the EM. When compared against the best-performing GHM, the median MMC performance gain is lowest in the northern dry hydrobelt (-24%) and highest in southern sub-tropical (-254%) and the boreal (-51%) hydrobelts. Northern mid-latitude catchments see performance gains of -32%. However, it is important to acknowledge that whilst IPE facilitates comparison of MMCs across hydrobelts, the robustness of the comparison is limited by the lower proportion of the total hydrobelt area represented by catchments in NDR, SST, SDR and SML hydrobelts. Addressing this will require data from a greater number of study catchments to be made available, with the temporally-extensive runoff records needed to support robust application of the machine learning algorithms that underpin MMC development. This highlights the importance of improving data collection systems in these hydrobelts in particular.

When the hydrobelt performance is examined with respect to the performance rankings of the catchments that comprise them, it is evident that MMC solutions achieve a disproportionately high performance gain in boreal catchments compared to other hydrobelts.

Here, 65% of the catchments are positioned in the top 50% of the MMC performance gain rankings (Table 4). This suggests there may be particular opportunities for achieving performance gain through MMC in boreal catchments. In northern mid latitude (NML) catchments no discernible trends in the performance rankings are evident – catchments are split approximately evenly between the top and bottom halves of the rankings. Catchments in both of the northern dry (NDR) hydrobelt catchments, as well as SDR's, are noteworthy because none of the GHMs, the EM nor the MMC solution was able to improve upon the naïve benchmark model (all their IPE scores are positive) in either of the catchments (see Table 3). This indicates that the process representations employed in our suite of GHMs may be deficient for modelling runoff in this hydrobelt, although as a caveat we note that there are only two NDR catchments in the data set.

Table 4. MMC solution and equations ranked by MMC performance gain (MMC_{PG}) in the validation data set. MMC_{PG} is measured against *either* the best performing GHM *or* the EM, whichever of the two performs better.

No	River	Hydro-belt	MMC IPE score	Best performing model (GHM or EM) and IPE score	MMC_{PG} (%)	Rank	MMC solution separated into its GEP-expression components. MMC = C1 + C2 + C3. Components are ordered according to their explanatory power as assessed by their IPE.	Eqn. size[1]
38	COOPER CREE	SDR	20.05	WaterGAP2 IPE = 107.00	-8674	1	C1: 0 C2: + (-0.143) * H08 * (WaterGAP2 +1) * cos(cos(WaterGAP2)) C3: + 0.436*H08*sqrt WaterGAP2	18
40	DARLING RIVER	SML	-1.64	WaterGAP2 IPE = -1.54	-1350	2	C1: 0.174*H08^2/DBH C2: + (-0.06/DBH) C3: + H08/DBH	11
39	FITZROY	SML	30.64	WaterGAP2 IPE = 38.47	-784	3	C1: sin(H08/-4.91) C2: + WaterGAP2 C3: + sin((LPJmL - sqrt DBH-8.45)*(WaterGAP2+H08)/(DBH *PCRGLOBWB))	20
4	Ob	BOR	-1.32	WaterGAP2 IPE = 2.50	-581	4	C1: 2*DBH/(log(sin H08)+6247.9) C2: + sqrt H08 C3: + WaterGAP2/H08^2	15
33	RIO PARNAIBA	SST	-2.27	WaterGAP2 IPE = 1.46	-574	5	C1: 3.695 C2: + 0.625*((cos(0.227/H08))^6*(log(WaterGAP2))^4) C3: + 1.472 / (log(1/PCRGLOBWB) – 1.08396)	20
36	BURDEKIN	SST	-1.35	H08 IPE = 1.44	-479	6	C1: 0 C2: + sqrt H08 C3: + H08 * sin(log(log(PCRGLOBWB/2)))	10
10	RED RIVER	BOR	-1.25	WaterGAP2 IPE = 1.52	-478	7	C1: H08*WaterGAP2/10.045 C2: + sin PCRGLOBWB^3/(DBH^3*H08+H08-LPJmL-5.44) C3: + sin(cos(WaterGAP2))^3	23
19	FRASER RIVER	NML	-1.61	PCRGLOBWB IPE = 1.15	-477	8	C1: 0.33*DBH*sqrt(log(PCRGLOBWB)) C2: + cos((H08+1.63)/LPJmL)+8.12 C3: + cos H08	17
18	ST. LAWRENCE	NML	2.47	WaterGAP2 IPE = 7.09	-462	9	C1: 23.04 C2: + 0.67*sqrt WaterGAP2 * cos(sqrt WaterGAP2+ 1.42/H08) C3: + 1.1*sqrt(DBH/PCRGLOBWB)	19
2	AMUR	BOR	-1.49	EM IPE = 1.07	-356	10	C1: 2.534*(DBH-H08-LPJmL-LPJmL/H08)/PCRGLOBWB C2: + WaterGAP2-4.33 C3: + sin DBH	18

23	NEVA	NML	1.09	WaterGAP2 IPE =2.42	-133	11	C1: PCRGLOBWB C2: + log(DBH^3) C3: + WaterGAP2/PCRGLOBWB + 0.5*log(log(WaterGAP2))	13
5	KOLYMA	BOR	-2.38	DBH IPE = -1.23	- -114	12	C1: DBH C2: + sqrt LPJmL C3: + DBH*(-2.74*DBH+LPJmL-3.133)/WaterGAP2	14
26	YANGTZE	NML	-4.17	WaterGAP2 IPE = -3.03	-108	13	C1: WaterGAP2 C2: + sqrt LPJmL C3: + cos(PCRGLOBWB +0.039*H08*PCRGLOBWB/DBH)	13
1	LENA	BOR	-2.00	WaterGAP2 IPE = -1.22	-78	14	C1: WaterGAP2-sqrt DBH C2: + LPJmL/(2*LPJmL/WaterGAP2^2+5.575) C3: + (-0.626)	15
31	AMAZONAS	EQT	-1.85	WaterGAP2 IPE = -1.09	-75	15	C1: WaterGAP2 C2: + (H08-DBH+LPJmL+0.77)* (WaterGAP2-LPJmL- 0.77)/(PCRGLOBWB+24.9) C3: + (-2.98)	19
14	NORTHERN DVINA	BOR	-2.27	EM IPE= -1.54	-70	16	C1: WaterGAP2 C2: + PCRGLOBWB C3: + (-9.29)	3
3	YENISEI	BOR	-2.32	WaterGAP2 IPE = -1.72	-58	17	C1: WaterGAP2 C2: + (-0.742) C3: + 7.0*sin(sqrt H08)	7
9	ASSINIBOINE	BOR	1.06	WaterGAP2 IPE = 1.57	-51	18	C1: WaterGAP2^2 C2: + sin(0.5*log(0.268*H08+cosWaterGAP2/WaterGAP2+0.003)) C3: + 0.064	17
21	RHINE RIVER	NML	-2.50	WaterGAP2 IPE = -1.96	-51	19	C1: WaterGAP2 C2: + 5.813 C3: + (-0.153)*H08	5
12	CHURCHILL RIVER	BOR	3.10	WaterGAP2 IPE = 3.60	-50	20	C1: WaterGAP2 C2: + sin PCRGLOBWB C3: + cos(sqrt H08)	6
29	NIGER	NST	-1.79	WaterGAP2 IPE = -1.37	-41	21	C1: 0.062* log(DBH)^4*(cos(4.647/PCRGLOBWB))^6 C2: + cos(sin LPJmL/WaterGAP2) C3: + 0.556	17
8	SASKATCHEWAN	BOR	1.03	WaterGAP2 IPE = 1.43	-40	22	C1: WaterGAP2 C2: + (cos(cos(DBH + log WaterGAP2 + 0.31))-sin(sqrt PCRGLOBWB^3))^3 C3: + -sin((log LPJmL^3)/8-sin(cos(0.401*LPJmL)+1.723)	29

30	ZAIRE	EQT	1.42	WaterGAP2 IPE = 1.78	-36	23	C1: WaterGAP2 C2: + cos(sqrt DBH) C3: + cos(sqrt DBH)	7
15	YELLOW	NML	1.16	WaterGAP2 IPE = 1.49	-33	24	C1: sqrt(DBH) C2: + DBH*WaterGAP2^5/4/(DBH^2*WaterGAP2-0.043*PCRGLOBWB) C3: + (sin WaterGAP2)^2*sin(sqrt(PCRGLOBWB+DBH))	26
13	ALBANY RIVER	BOR	-1.66	PCRGLOBWB IPE = -1.33	-33	25	C1: PCRGLOBWB C2: + log(0.106*DBH) C3: + log(0.041*DBH)	9
22	DANUBE	NML	-2.22	WaterGAP2 IPE = -1.89	-32	26	C1: WaterGAP2 C2: + DBH/H08- H08/(PCRGLOBWB-1) C3: + 7.93/H08	13
25	DON	NML	1.23	WaterGAP2 IPE = 1.54	-32	27	C1: WaterGAP2 C2: + 1 C3: + (-0.325)*WaterGAP2	5
34	SAO FRANCISCO	SST	-1.92	WaterGAP2 IPE = -1.64	-29	28	C1: sqrt(WaterGAP2) C2: + 1.46*(PCRGLOBWB+WaterGAP2-5.75)/log(PCRGLOBWB) C3: + cos(H08/LPJmL)	16
27	COLORADO	NDR	2.22	H08 IPE = 2.50	-29	29	C1: log(DBH) C2: + log(PCRGLOBWB) C3: + WaterGAP2/PCRGLOBWB	7
24	VOLGA	NML	-2.00	WaterGAP2 IPE = -1.75	-23	30	C1: WaterGAP2-0.978 C2: + 3.35/DBH C3: + 0.999/LPJmL	9
37	ORANJE	SDR	2.04	WaterGAP2 IPE = 2.26	-22	31	C1: WaterGAP2 C2: + 0.808 C3: + (-0.672)	3
28	SANTIAGO	NDR	1.16	WaterGAP2 IPE = 1.35	-19	32	C1: sin(LPJmL^2*(0.319-LPJmL/DBH))/DBH C2: + WaterGAP2 C3: + sin((sin(((sin((LPJmL))-(((LPJmL)/(WaterGAP2))^3))^2))-(WaterGAP2)))	24
17	MISSISSIPPI	NML	-2.04	WaterGAP2 IPE = -1.89	-14	33	C1: WaterGAP2 C2: + (log(WaterGAP2^3)-WaterGAP2)/PCRGLOBWB C3: + (-1.70-DBH)/PCRGLOBWB	13
32	XINGU	EQT	1.04	WaterGAP2 IPE = 1.16	-9	34	C1: WaterGAP2 C2: + (-0.494) C3: + (-0.204)*LPJmL/sqrt WaterGAP2	8

20	LABE	NML	-1.45	WaterGAP2 IPE = -1.47	2	35	C1: WaterGAP2 C2: + 4.32/DBH C3: + 0.962*sin((DBH-H08)/PCRGLOBWB+ cos(0.15*WaterGAP2))	17
7	MACKENZIE RIVER	BOR	-1.33	EM IPE = -1.39	5	36	C1: PCRGLOBWB C2: + 0.107*DBH C3: + (-0.978)	5
16	COLUMBIA	NML	-1.20	EM IPE = -1.28	7	37	C1: WaterGAP2 C2: + sin(cos(LPJmL)^3)^2*sin(PCRGLOBWB*cos(3.78*PCRGLOBWB)) C3:+ exp(cos(cos(LPJmL)*sin(WaterGAP2)))* sin(0.479+0.166*WaterGAP2)	28
11	WINNIPEG RIVER	BOR	1.63	PCRGLOBWB IPE = 1.55	8	38	C1: WaterGAP2 C2: + H08/DBH C3: + (-4.91+log(PCRGLOBWB))	8
35	PARAGUAI	SST	8.51	WaterGAP2 IPE = 8.00	19	39	C1: WaterGAP2 C2: log(9.84/LPJmL) C3: 0.99- (LPJmL/(PCRGLOBWB-((LPJmL+WaterGAP2)/945.48)))	16
6	OLENEK	BOR	-1.15	DBH IPE = -1.47	33	40	C1: -sin(0.004* LPJmL^2*PCRGLOBWB-LPJmL+9.04) C2: + PCRGLOBWB/(-0.31*DBH^2*cosec(PCRGLOBWB) -7.71) C3: + WaterGAP2	22

1-As defined in Section 3.1, equation size is calculated according to the number of inputs (GHMs), constants, operators and functions in an equation.

Table 5. Median MMC performance gain (MMC_{PG}) for each hydrobelt, for the validation data set.

Hydrobelt	No. of catchments	Median MMC_{PG} over best-performing GHM (%)	Median MMC_{PG} over EM (%)
BOR	14	-51	-415
NML	12	-32	-434
NDR	2	-24	-520
NST	1	-41	-764
EQT	3	-36	-161
SST	4	-254	-1698
SDR	2	-4348*	-104900*
SML	2	-1067*	-703068*

** Denotes a median MMC_{PG} score significantly influenced by the individual result for Cooper Creek, Darling or Fitzroy River.*

5. Discussion

5.1. Interpretability of MMC solutions

Our rationale for developing weighted MMC solutions from an ensemble of GHMs was in part a response to a question frequently asked by modellers, decision-makers, and the public: *why not weight / adjust the models according to their performance*? We acknowledge that in other disciplines (Gillett, 2015; Giorgi and Mearns, 2002; Qi et al., 2017), including climate modelling (Christensen et al., 2010; Fowler and Ekström, 2009) and catchment hydrological modelling (Abrahart and See, 2002; Ajami et al., 2006; Arsenault et al., 2015; Shamseldin et al., 1997), weighting strategies have been highly effective in improving the performance of a model ensemble. However, the question cannot be answered adequately unless the best approach to determining the weighting strategies is known. In past examples, the strategy has been to apply simple constants (Arsenault et al., 2015; Christensen et al., 2010; Shamseldin et al., 1997) which may be optimised using linear constraints (e.g. the multiple linear regression approach of Doblas-Reyes et al. (2005)). By contrast, in this paper, we have examined what happens when the constraints are relaxed and more complex optimisation of non-linear weighting schemes is allowed (Table 4). Superficially, relaxing the constraints imposed on the weighting scheme is appealing because it increases the likelihood of improving the performance of the MMC solution. However, it introduces several critical shortcomings.

Firstly, the interpretation of the weights (and therefore MMC equations; Table 4) in physical terms becomes increasingly difficult as the constraints on the form and complexity of the

weighting scheme are relaxed. Where there is little or no attempt to constrain it, GEP-based MMC can become nothing more than a curve fitting exercise whose solution complexity makes it difficult to quantify the relative power of each model in the overall solution and precludes meaningful physical interpretation of the expressions that are generated. There is, therefore, a strong argument for a more pragmatic approach that applies careful constraint to the allowable complexity of GEP-based MMCs. This can be achieved by limiting the number of components and/or bases by reducing the set of mathematical operators and non-linear functions available to the GEP algorithm. In this study, we have used the GEP parameters to constrain the solution to three components and a relatively small set of seven non-linear functions (Table 2). Constraint has also been achieved by the selection of the final MMC solution from the candidate set based on a trade-off between complexity and performance (Figure 5). Despite this, several of the MMC solutions remain very complex and preclude meaningful interpretation (see Table 4). However, knowing how much to constrain the GEP expressions is vital because the benefits of increased interpretability of highly constrained solutions can be offset by reductions in overall MMC performance. Identifying the 'sweet spot' where both performance gain and interpretability is maximised will be an area fruitful for future research. To this end, Bayesian optimisation methods such as those underpinning model mixing studies (Marshall et al., 2006; Moges et al., 2016) are of interest because they indicate how it might be possible to optimise the values of the GEP parameter set (which constrain the solution) through Bayesian updating procedures. However, to this end the non-numerical nature of certain GEP parameters (e.g. the allowable operators and functions) are likely to be highly problematic because they will prevent the quantification of the PDFs required by Bayesian approaches. Therefore, more realistic approaches could include the dynamic configuration of the GEP algorithm parameters during training.

Secondly, with greater complexity comes a tendency towards overfitting of the MMC solutions. Whilst we sought to minimise the risk of selecting over-fitted MMC solutions by applying an error-complexity trade-off selection method (Figure 5), the high degree of complexity in some of the weighting schemes presented in Table 4 suggests that the MMCs may still be over-fitted.

Thirdly, we acknowledge that any attempt to weight models may be viewed by some as futile so long as the current generation of GHMs (or any model) are far from being empirically

adequate for purpose (Stainforth et al., 2007). Other work has shown that the GHMs applied here are imperfect (Zaherpour et al., 2018) and in this sense it can be argued that applying weights to any type of model that is known to contain errors is counter-intuitive because the errors in even well performing models will be weighted inherently in the approach. Where weights are applied in a simple manner (e.g. each GHM output is multiplied by a single coefficient), this is certainly the case. However, a key advantage of GEP is that it develops more complex schemes in which the products of more than one model can be weighted (e.g. the difference in performance between two or more models at different hydrological response ranges - see Figure 4). In this respect, GEP offers the significant advantage of being able to exploit the characteristic differences in the capabilities and/or failings of the models it combines – allowing MMC solutions to deliver performance gains based on non-linear adjustments made to the characteristic differences between each model input. Where GEP is concerned, it can be argued that it is its counter-intuitive ability to exploit model failings in the MMC solutions that provides a strong argument for using it rather than simple weighting – especially where the objective is to combine models known to be lacking with respect to their empirical fitness-for-purpose.

Current model combination approaches in hydrological modelling include simple model averaging (Arsenault et al., 2015; Cloke and Pappenberger, 2009) and complex weighting approaches (Ajami et al., 2006; Arsenault et al., 2015; Shamseldin et al., 2007) comprising machine learning algorithms, as described here. The critique above implies that the application of complex weights via machine learning algorithms offers little towards helping to understand individual model performance. For example, it is difficult to understand the relative weightings of individual models (i.e. which models are weighted more/less than others, e.g. see the solution for the Columbia river in Table 4), let alone why those weights have been applied (e.g. are the weights applied due to a model's ability to simulate high flows well?) and why some models are excluded altogether. Therefore, whilst we have demonstrated that generally a complex MMC solution can perform better than the EM, the interpretability of the MMC is highly limited. This suggests that a more interpretable, but still intelligent, approach to model combination is needed. An alternative approach would be to follow the framework described by Krysanova et al. (2018) for global- and catchment models. They recommend first evaluating model performance for several hydrological variables over

various time periods, as in a classical model evaluation (Zaherpour et al., 2018), and if performance is considered to be acceptable then the models can be weighted, otherwise they are excluded from the ensemble. Although there is value in the approach, no specific recommendations are provided on how to weight the models, other than weighting based upon model performance. In addition, identification of a threshold for "good performance" is not straightforward, and the approach rejects, *a priori*, poorly performing models. One of the arguable advantages of GEP is that it can exploit the characteristic error patterns of poorly performing models by using them as mechanisms to adjust other models through the MMC development, as we have demonstrated. Merging a more interpretable MMC approach with that of Krysanova et al. (2018) may be a pragmatic way forwards for future model combination and weighting studies.

5.2. MMC does not always deliver optimal solutions

It is important to note that machine learning-based MMC methods may not always deliver solutions that outperform the EM/best individual model despite their inherent optimisation capabilities. In six of our study catchments, we found that GEP failed, even though mostly marginally, to optimise its MMC solutions sufficiently to outperform either the EM (Mackenzie and Columbia catchments) or the best performing GHM (Olenek, Winnipeg, Labe and Paraguai catchments) (Table 3). Two potential causes are likely.

Firstly, the GEP algorithm's ability to learn an optimised MMC solution depends on it being able to learn expressions that capitalise on characteristic differences between the error structures and magnitudes of the different input models. If all model inputs have the same characteristic errors, or if their errors are all random, there will be insufficient 'raw material' for the GEP algorithm to learn from. Cross-correlation of the model residuals for these six catchments (Table S6) indicates that this may be a reason for the failure of the MMC solution in the Olenek and Paraguai catchments. Here high cross-correlation between the residuals of the majority of model inputs exists – limiting opportunities for the GEP algorithm to use the characteristic differences between input models in the weighting scheme optimisation.

Secondly, deficiencies in our error-complexity trade-off method to select the final MMC solution from the candidate set (Figure 5) could be a factor. Whilst the trade-off is necessary to limit the complexity of the final GEP-based MMC solution, it does mean that the best

performing MMC solution in the candidate set can be overlooked in favour of a simpler, lower-performing counterpart. This means that although the GEP algorithm may have developed a candidate MMC solution that outperforms either the EM or best-performing GHM, if its complexity is high relative to other solutions, it will not be selected as the final MMC solution. To check whether this is a factor behind MMC's poor performance in the six catchments, the best performing solutions from GEP's candidate solution set, irrespective of their complexity, are compared to the EM and best performing GHM in each catchment (Table S7). In the Mackenzie and Labe catchments, the best-performing MMC solution from the candidate set does outperform both the EM and the best-performing GHM. In the Paraguai catchment it equals it. However, in the Columbia, Olenek and Winnipeg even the best-performing candidate MMC solution fails to outperform the best individual GHM and the reasons for MMC failure remain unclear – particularly in the Columbia and Winnipeg catchments.

5.3. Accounting for and presenting uncertainty in MMC development

Compared to the model mixing approaches being used in catchment-scale modelling (Marshall et al., 2006; Moges et al., 2016), the MMC approach applied here is inferior because the lack of knowledge of the PDFs (and maximum likelihoods) of the model parameters prevents the minimisation of MMC uncertainty. In fact, the lack of knowledge about the PDFs associated with the highly generalised parameters of the individual GHMs, and the sheer number of parameters that they use means that it is going to be difficult to get beyond the performance optimisation approach taken in this study in the short to medium term. However, compared to other performance optimisation approaches that use machine learning (especially ANNs (Shamseldin et al., 2007)), GEP has the advantage that it is at least explicit. It also has the advantage that the user can easily control the form of the MMC solutions through the allowable expression complexity and allowable non-linear functions. Therefore, it is a step forward towards improved MMC development and interpretability. Nonetheless, the big challenge remains the application of more advanced, maximum likelihood model mixture approaches to GHMs.

In addition, even though our study highlights how MMC outputs generally out-perform individual GHMs and the EM, we caution against presenting MMC results in isolation. Instead,

we recommend that MMC results are presented alongside the range of model outputs from the whole ensemble and the EM (e.g. Figures 6 and 7, and Table 3). Even though MMC techniques employed in other disciplines have been claimed to result in a "reduction of the uncertainty range" (Giorgi and Mearns, 2002; Marshall et al., 2006), we argue that the original uncertainty range should still be presented because it has been computed from a set of physically-based models specifically designed to simulate relevant environmental processes and feedbacks. Indeed, we would go further and argue that MMC does not reduce the inherent uncertainty. It does, however, provide a more robust and informative estimate from the ensemble that takes into account the performance of its members. To not explicitly present the uncertainty in the models that contribute to an MMC solution risks masking an important dimension of the data that underpin it.

6. Conclusions

This study has, for the first time, applied a set of intelligently defined weights to a state-of-the-art ensemble of global-scale hydrological models. The GEP-based MMC applied, is shown to employ a diverse array of linear and non-linear adjustments to exploit information in runoff estimates from the individual GHMs. The result is that in 34 catchments (85%) the MMC performs better than the best performing GHM and EM with the median performance gain over a naïve benchmark model being 45% across all 40 catchments. The EM performs better than individual GHMs in only 10% (4) of our catchments.

Despite the good performance of the MMCs across the majority of catchments, it should not be seen as a "silver bullet" for counteracting biases and fit residuals of individual GHMs. In six (15%) of the catchments either the EM or an individual GHM performed marginally better than the MMC solution, with GHMs' lack of insufficient 'raw material' for the GEP algorithm to exploit, or deficiency in our error-complexity trade-off method for selecting final MMC being potentially responsible for this.

More importantly, the GEP approach applied here includes weighting schemes whose complexity prevents meaningful physical interpretation of the MMCs solutions and realisation of the absolute and relative power and contribution of individual GHMs. More research is,

therefore, needed to explore the effect of application of different levels of constraints on GEP-based algorithm performance in providing more interpretable MMC solutions.

In addition, the MMC approach applied here does not account for uncertainty within input models or their parameters due to the lack of information on their PDFs. Hence, the approach does not go beyond optimising their predictive performance. However, there could be potential in applying more realistic approaches that include dynamic configurations of the GEP algorithm parameters during training.

Despite shortcomings of the GEP-based MMC in the current level of functionality, its explicit outputs and controllability is a step forward towards unravelling the black box nature of approaches such as ANNs and increasing MMC interpretability. In addition, in light of the significantly improved performance offered by MMC, relative to individual GHMs and also the EM, we recommend that future multi-model applications consider using MMCs alongside the EM and intermodal range, to provide end-users of the ensemble with a better informed estimate of what the ensemble shows.

Acknowledgements

This work has been conducted under the framework of the Inter-Sectoral Impact Model Inter-comparison Project, phase 2 (ISIMIP2a), so our thanks go to the modellers who submitted results to this project. ISIMIP2a was funded by the German Ministry of Education and Research, with project funding reference number 01LS1201A. ISIMIP2 data are available from Gosling et al. (2017). We also thank the Global Runoff Data Centre (GRDC) for making available the observed runoff data. JZ was supported by the Islamic Development Bank and a 2018 University of Nottingham Faculty of Social Sciences Research Outputs Award. YM was supported by the Environmental Research and Technology Fund (S-10) of the Ministry of the Environment, Japan.

References

Abrahart, R.J., See, L., 2002. Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments. Hydrology and Earth System Sciences 6(4) 655-670.

Ajami, N.K., Duan, Q.Y., Gao, X.G., Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: Application to Distributed Model Intercomparison Project results. Journal of Hydrometeorology 7(4) 755-768.

Ajami, N.K., Duan, Q.Y., Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resources Research 43(1).

Arnold, J.G., Allen, P.M., Bernhardt, G., 1993. A comprehensive surface-groundwater flow model. Journal of Hydrology 142(1–4) 47-69.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J.L., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. Journal of Hydrology 529 754-767.

Azmi, M., Araghinejad, S., Kholghi, M., 2010. Multi Model Data Fusion for Hydrological Forecasting Using K-Nearest Neighbour Method. Iranian Journal of Science and Technology Transaction B-Engineering 34(B1) 81-92.

Barbulescu, A., Bautu, E., 2010. Mathematical models of climate evolution in Dobrudja. Theoretical and Applied Climatology, 100(29-44).

Bărbulescu, A., Băutu, E., 2009. Time Series Modeling Using an Adaptive Gene Expression Programming Algorithm. International journal of mathematical models and methods in applied sciences 3(2) 85-93.

Beck, H.E., de Roo, A., van Dijk, A.I.J.M., 2015. Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments. Journal of Hydrometeorology 16(4) 1478-1501.

Beck, H.E., van Dijk, A.I.J.M., de Roo, A., Dutra, E., Fink, G., Orth, R., Schellekens, J., 2016. Global evaluation of runoff from ten state-of-the-art hydrological models. Hydrology and Earth System Sciences Discussions 21 2881-2903.

Christensen, J.H., Kjellstrom, E., Giorgi, F., Lenderink, G., Rummukainen, M., 2010. Weight assignment in regional climate models. CLIMATE RESEARCH 44(2-3) 179-194.

Clemen, R.T., 1989. COMBINING FORECASTS - A REVIEW AND ANNOTATED-BIBLIOGRAPHY. International Journal of Forecasting 5(4) 559-583.

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. Journal of Hydrology 375(3-4) 613-626.

Dawson, C.W., Mount, N.J., Abrahart, R.J., Shamseldin, A.Y., 2012. Ideal point error for model assessment in data-driven river flow forecasting. Hydrology and Earth System Sciences 16(8) 3049-3060.

de Menezes, L.M., W. Bunn, D., Taylor, J.W., 2000. Review of guidelines for the use of combined forecasts. European Journal of Operational Research 120(1) 190-204.

Doblas-Reyes, F.J., Hagedorn, R., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. Tellus A 57(3) 234-252.

Duan, Q.Y., Ajami, N.K., Gao, X.G., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Advances in Water Resources 30(5) 1371-1386.

Fernando, A.K., Shamseldin, A.Y., Abrahart, R.J., 2012. Use of Gene Expression Programming for Multimodel Combination of Rainfall-Runoff Models. Journal of Hydrologic Engineering 17(9) 975-985.

Ferreira, C., 2001. Gene Expression Programming: A New AdaptiveAlgorithm for Solving Problems. 13(2) 87-129.

Ferreira, C., 2006. Gene Expression Programming: Mathematical Model-ing by an Artificial Intelligence, 2nd ed. Springer, Verlag: Berlin.

Fowler, H.J., Ekström, M., 2009. Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes. INTERNATIONAL JOURNAL OF CLIMATOLOGY 29(3) 385-416.

Gillett, N.P., 2015. Weighting climate model projections using observational constraints. Philos Trans A Math Phys Eng Sci 373(2045) 1-8.

Giorgi, F., Mearns, L.O., 2002. Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging'' (REA) method. Journal of Climate 15(10) 1141-1158.

Gosling, S., Müller Schmied, H., Betts, R., Chang, J., Ciais, P., Dankers, R., Döll, P., Eisner, S., Flörke, M., Gerten, D., Grillakis, M., Hanasaki, N., Hagemann, S., Huang, M., Huang, Z., Jerez, S., Kim, H., Koutroulis, A., Leng, G., Liu, X., Masaki, Y., Montavez, P., Morfopoulos, C., Oki, T., Papadimitriou, L., Pokhrel, Y., Portmann, F.T., Orth, R., Ostberg, S., Satoh, Y., Seneviratne, S., Sommer, P., Stacke, T., Tang, Q., Tsanis, I., Wada, Y., Zhou, T., Büchner, M., Schewe, J., Zhao, F., 2017. ISIMIP2a Simulation Data from Water (global) Sector. GFZ Data Services. . PIK: Potsdam, Germany.

Gosling, S.N., Arnell, N.W., 2011. Simulating current global river runoff with a global hydrological model: model revisions, validation, and sensitivity analysis. Hydrological Processes 25(7) 1129-1145.

Gosling, S.N., Zaherpour, J., Mount, N.J., Hattermann, F.F., Dankers, R., Arheimer, B., Breuer, L., Ding, J., Haddeland, I., Kumar, R., Kundu, D., Liu, J., van Griensven, A., Veldkamp, T.I.E.,

Vetter, T., Wang, X., Zhang, X., 2016. A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C. Climatic change 141(3) 577-595.

Graham, D.N., Butts, M.B., 2005. Flexible, Integrated Watershed Modelling with MIKE SHE. CRC Press.

Guimberteau, M., Zhu, D., Maignan, F., Huang, Y., Yue, C., Dantec-Nédélec, S., Ottlé, C., Jornet-Puig, A., Bastos, A., Laurent, P., Goll, D., Bowring, S., Chang, J., Guenet, B., Tifafi, M., Peng, S., Krinner, G., Ducharne, A., Wang, F., Wang, T., Wang, X., Wang, Y., Yin, Z., Lauerwald, R., Joetzjer, E., Qiu, C., Kim, H., Ciais, P., 2018. ORCHIDEE-MICT (v8.4.1), a land surface model for the high latitudes: model description and validation. Geoscientific Model Development 11(1) 121-163.

Hagedorn, R., Doblas-Reyes, F.J., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. Tellus Series a-Dynamic Meteorology and Oceanography 57(3) 219-233.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., Tanaka, K., 2008a. An integrated model for the assessment of global water resources – Part 2: Applications and assessments. Hydrology and Earth System Sciences 12(4) 1027-1037.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., Tanaka, K., 2008b. An integrated model for the assessment of global water resources Part 1: Model description and input meteorological forcing. Hydrology and Earth System Sciences 12(4) 1007-1025.

Hattermann, F.F., Krysanova, V., Gosling, S.N., Dankers, R., Daggupati, P., Donnelly, C., Flörke, M., Huang, S., Motovilov, Y., Buda, S., Yang, T., Müller, C., Leng, G., Tang, Q., Portmann, F.T., Hagemann, S., Gerten, D., Wada, Y., Masaki, Y., Alemayehu, T., Satoh, Y., Samaniego, L., 2017. Cross‐scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins. Climatic change 141(3) 561-576.

Hattermann, F.F., Vetter, T., Breuer, L., Su, B.D., Daggupati, P., Donnelly, C., Fekete, B., Florke, F., Gosling, S.N., Hoffmann, P., Liersch, S., Masaki, Y., Motovilov, Y., Muller, C., Samaniego, L., Stacke, T., Wada, Y., Yang, T., Krysnaova, V., 2018. Sources of uncertainty in hydrological climate impact assessment: a cross-scale study. ENVIRONMENTAL RESEARCH LETTERS 13(1).

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial. Statistical Science 14(4) 382-401.

Huang, J., vandenDool, H.M., Georgakakos, K.P., 1996. Analysis of model-calculated soil moisture over the United States (1931-1993) and applications to long-range temperature forecasts. Journal of Climate 9(6) 1350-1362.

Jägermeyr, J., Gerten, D., Heinke, J., Schaphoff, S., Kummu, M., Lucht, W., 2015. Water savings potentials of irrigation systems: global simulation of processes and linkages. Hydrology and Earth System Sciences 19(7) 3073-3091.

Jeong, D.I., Kim, Y.-O., 2009. Combining single-value streamflow forecasts – A review and guidelines for selecting techniques. Journal of Hydrology 377(3-4) 284-299.

Kim, H., 2017. Global SoilWetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1) [Data set]. Data Integration and Analysis System (DIAS).

Koirala, S., Yeh, P.J.F., Hirabayashi, Y., Kanae, S., Oki, T., 2014. Global-scale land surface hydrologic modeling with the representation of water table dynamics. Journal of Geophysical Research: Atmospheres 119(1) 75-89.

Krysanova, V., Müller-Wohlfeil, D.-I., Becker, A., 1998. Development and test of a spatially distributed hydrological/water quality model for mesoscale watersheds. Ecological Modelling 106(2–3) 261-289.

Kundzewicz, Z.W., 1986. The Hydrology of Tomorrow. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 31(2) 223-235.

Lindstrom, G., Pers, C., Rosberg, J., Stromqvist, J., Arheimer, B., 2010. Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. Hydrology Research 41(3-4) 295-319.

Liu, Y., Guo, H., Zhang, Z., Wang, L., Dai, Y., Fan, Y., 2007. An optimization method based on scenario analysis for watershed management under uncertainty. Environ Manage 39(5) 678-690.

Marshall, L., Nott, D., Sharma, A., 2007. Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework. Hydrological Processes 21(7) 847-861.

Marshall, L., Sharma, A., Nott, D., 2006. Modeling the catchment via mixtures: Issues of model specification and validation. Water Resources Research 42(11).

Masaki, Y., Hanasaki, N., Biemans, H., Schmied, H.M., Tang, Q., Wada, Y., Gosling, S.N., Takahashi, K., Hijioka, Y., 2017. Intercomparison of global river discharge simulations focusing on dam operation—multiple models analysis in two case-study river basins, Missouri–Mississippi and Green–Colorado. ENVIRONMENTAL RESEARCH LETTERS 12(5) 055002.

May, R.J., Maier, H.R., Dandy, G.C., 2010. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Netw 23(2) 283-294.

Meybeck, M., Kummu, M., Dürr, H.H., 2013. Global hydrobelts and hydroregions: improved reporting scale for water-related issues? Hydrology and Earth System Sciences 17(3) 1093-1111.

Milly, P.C., Dunne, K.A., Vecchia, A.V., 2005. Global pattern of trends in streamflow and water availability in a changing climate. Nature 438(7066) 347-350.

Moges, E., Demissie, Y., Li, H.-Y., 2016. Hierarchical mixture of experts and diagnostic modeling approach to reduce hydrologic model structural uncertainty. Water Resources Research 52(4) 2551-2570.

Moore, R.J., 2007. The PDM rainfall-runoff model. Hydrology and Earth System Sciences 11(1) 483-499.

Müller Schmied, H., Adam, L., Eisner, S., Fink, G., Flörke, M., Kim, H., Oki, T., Portmann, F.T., Reinecke, R., Riedel, C., Song, Q., Zhang, J., Döll, P., 2016. Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use. Hydrology and Earth System Sciences 20(7) 2877-2898.

Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F.T., Flörke, M., Döll, P., 2014. Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. Hydrology and Earth System Sciences 18(9) 3511-3538.

Nasseri, M., Zahraie, B., Ajami, N.K., Solomatine, D.P., 2014. Monthly water balance modeling: Probabilistic, possibilistic and hybrid methods for model combination and ensemble simulation. Journal of Hydrology 511 675-691.

Phukoetphim, P., Shamseldin, A.Y., Adams, K., 2016. Multimodel Approach Using Neural Networks and Symbolic Regression to Combine the Estimated Discharges of Rainfall-Runoff Models. Journal of Hydrologic Engineering 21(8) 04016022.

Pushpalatha, R., Perrin, C., Le Moine, N., Andreassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. Journal of Hydrology 420 171-182.

Qi, Y., Qian, C., Yan, Z., 2017. An alternative multi-model ensemble mean approach for near-term projection. INTERNATIONAL JOURNAL OF CLIMATOLOGY 37(1) 109-122.

Sanderson, B.M., Knutti, R., 2012. On the interpretation of constrained climate model ensembles. Geophysical Research Letters 39(16) 1-6.

Shamseldin, A.Y., O'Connor, K.M., Nasr, A.E., 2007. A comparative study of three neural network forecast combination methods for simulated river flows of different rainfall-runoff models. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 52(5) 896-916.

Shamseldin, A.Y., OConnor, K.M., Liang, G.C., 1997. Methods for combining the outputs of different rainfall-runoff models. Journal of Hydrology 197(1-4) 203-229.

Snee, R.D., 1977. Validation of Regression-Models - Methods and Examples. Technometrics 19(4) 415-428.

Stainforth, D.A., Allen, M.R., Tredger, E.R., Smith, L.A., 2007. Confidence, uncertainty and decision-support relevance in climate predictions. Philos Trans A Math Phys Eng Sci 365(1857) 2145-2161.

Sudheer, K.P., Gosain, A.K., Rangan, D.M., Saheb, S.M., 2002. Modelling evaporation using an artificial neural network algorithm. Hydrological Processes 16(16) 3189-3202.

Thiery, W., Davin, E.L., Lawrence, D.M., Hirsch, A.L., Hauser, M., Seneviratne, S.I., 2017. Present-day irrigation mitigates heat extremes. Journal of Geophysical Research: Atmospheres 122(3) 1403-1422.

Van Beek, L.P.H., Bierkens, M.F.P., 2008. The Global Hydrological Model PCR-GLOBWB: Conceptualization, Parameterization and Verification, Report

Department of Physical Geography, Utrecht University, Utrecht, The Netherlands,.

van Huijgevoort, M.H.J., Hazenberg, P., van Lanen, H.A.J., Teuling, A.J., Clark, D.B., Folwell, S., Gosling, S.N., Hanasaki, N., Heinke, J., Koirala, S., Stacke, T., Voss, F., Sheffield, J., Uijlenhoet, R., 2013. Global Multimodel Analysis of Drought in Runoff for the Second Half of the Twentieth Century. Journal of Hydrometeorology 14(5) 1535-1552.

Veldkamp, T.I.E., Zhao, F., Ward, P.J., de Moel, H., Aerts, J.C.J.H., Schmied, H.M., Portmann, F.T., Masaki, Y., Pokhrel, Y., Liu, X., Satoh, Y., Gerten, D., Gosling, S.N., Zaherpour, J., Wada, Y., 2018. Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study. ENVIRONMENTAL RESEARCH LETTERS 13(5) 055008.

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resources Research 43(1).

Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrology and Earth System Sciences 5(1) 13-26.

Wartenburger, R., Seneviratne, S.I., Hirschi, M., Chang, J., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Gosling, S.N., Gudmundsson, L., Henrot, A.-J., Hickler, T., Ito, A., Khabarov, N., Kim, H., Leng, G., Liu, J., Liu, X., Masaki, Y., Morfopoulos, C., Müller, C., Schmied, H.M., Nishina, K., Orth, R., Pokhrel, Y., Pugh, T.A.M., Satoh, Y., Schaphoff, S., Schmid, E., Sheffield, J., Stacke, T., Steinkamp, J., Tang, Q., Thiery, W., Wada, Y., Wang, X., Weedon, G.P., Yang, H., Zhou, T., 2018. Evapotranspiration simulations in ISIMIP2a—Evaluation of spatio-temporal characteristics with a comprehensive ensemble of independent datasets. ENVIRONMENTAL RESEARCH LETTERS 13(7) 075001.

WMO, 2006. Technical Regulations, Volume III: Hydrology. Available online at: http://library.wmo.int/pmb_ged/wmo_49-v3-2006_en.pdf.

Wu, W., May, R., Dandy, G.C., Maier, H.R., 2012. A method for comparing data splitting approaches for developing hydrological ANN models, In: R. Seppelt, A.A.V., S. Lange, D.

Bankamp (Eds.) (Ed.), International Environmental Modelling and Software Society (iEMSs), 2012 International Congress on Environmental Modelling and Software, Managing Resources of a Limited Planet, Sixth Biennial Meeting: Leipzig, Germany.

Wu, W.Y., Dandy, G.C., Maier, H.R., 2014. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environmental Modelling & Software 54 108-127.

Xiong, L.H., Shamseldin, A.Y., O'Connor, K.M., 2001. A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi-Sugeno fuzzy system. Journal of Hydrology 245(1-4) 196-217.

Ye, M., Neuman, S.P., Meyer, P.D., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resources Research 40(5).

Zaherpour, J., Gosling, S.N., Mount, N., Schmied, H.M., Veldkamp, T.I.E., Dankers, R., Eisner, S., Gerten, D., Gudmundsson, L., Haddeland, I., Hanasaki, N., Kim, H., Leng, G., Liu, J., Masaki, Y., Oki, T., Pokhrel, Y., Satoh, Y., Schewe, J., Wada, Y., 2018. Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts. ENVIRONMENTAL RESEARCH LETTERS 13(6) 065015.

Paper III: A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1°C, 2°C and 3°C

Climatic Change
DOI 10.1007/s10584-016-1773-3

A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C

Simon N. Gosling[1] · **Jamal Zaherpour**[1] · **Nick J. Mount**[1] · **Fred F. Hattermann**[2] · **Rutger Dankers**[3] · **Berit Arheimer**[4] · **Lutz Breuer**[5,6] · **Jie Ding**[7,8] · **Ingjerd Haddeland**[9] · **Rohini Kumar**[10] · **Dipangkar Kundu**[11] · **Junguo Liu**[7,12] · **Ann van Griensven**[13,14] · **Ted I. E. Veldkamp**[15] · **Tobias Vetter**[2] · **Xiaoyan Wang**[16] · **Xinxin Zhang**[12]

Received: 21 December 2015 / Accepted: 24 July 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract We present one of the first climate change impact assessments on river runoff that utilises an ensemble of global hydrological models (Glob-HMs) and an ensemble of catchment-scale hydrological models (Cat-HMs), across multiple catchments: the upper Amazon, Darling, Ganges, Lena, upper Mississippi, upper Niger, Rhine and Tagus. Relative changes in simulated mean annual runoff (MAR) and four indicators of high and low extreme

This article is part of a Special Issue on "Hydrological Model Intercomparison for Climate Impact Assessment" edited by Valentina Krysanova and Fred Hattermann.

Electronic supplementary material The online version of this article (doi:10.1007/s10584-016-1773-3) contains supplementary material, which is available to authorized users.

✉ Simon N. Gosling
simon.gosling@nottingham.ac.uk

[1] School of Geography, University of Nottingham, Nottingham NG7 2RD, UK

[2] Potsdam Institute for Climate Impact Research, Telegrafenberg, Potsdam A62, D-14473, Germany

[3] Met Office, Exeter, UK

[4] Swedish Meteorological and Hydrological Institute (SMHI), Norrköping 60176, Sweden

[5] Institute for Landscape Ecology and Resources Management (ILR), Research Centre for BioSystems, Land Use and Nutrition (iFZ), Justus Liebig University, Giessen, Germany

[6] Centre for International Development and Environmental Research (ZEU), Justus Liebig University, Giessen, Germany

[7] School of Environmental Science and Engineering, South University of Science and Technology of China, Shenzhen 518055, China

Published online: 09 November 2016

flows are compared between the two ensembles. The ensemble median values of changes in runoff with three different scenarios of global-mean warming (1, 2 and 3 °C above pre-industrial levels) are generally similar between the two ensembles, although the ensemble spread is often larger for the Glob-HM ensemble. In addition the ensemble spread is normally larger than the difference between the two ensemble medians. Whilst we find compelling evidence for projected runoff changes for the Rhine (decrease), Tagus (decrease) and Lena (increase) with global warming, the sign and magnitude of change for the other catchments is unclear. Our model results highlight that for these three catchments in particular, global climate change mitigation, which limits global-mean temperature rise to below 2 °C above preindustrial levels, could avoid some of the hydrological hazards that could be seen with higher magnitudes of global warming.

1 Introduction

Article 2 of the 2015 United Nations Framework Convention on Climate Change (UNFCCC) Paris Agreement includes an action to limit any future increase in global-mean temperature to well below 2 °C above pre-industrial levels and to pursue efforts to limit to 1.5 °C, recognising that this would significantly reduce risks and impacts of climate change (UNFCCC 2015). However, if the latest Government climate action pledges from 185 countries were implemented then global-mean warming would still reach 2.7 °C (CAT 2015). Whilst this is almost 1 °C lower than an alternative future in which only existing policies remain enacted (CAT 2015), it misses the UNFCCC targets. There is therefore significant interest in understanding the potential impacts of different amounts of global-mean warming.

Whilst the UNFCCC target is framed in terms of global-mean temperature, the impacts will be felt heterogeneously across the world (Arnell et al. 2016) and one of the key impacts of global warming will be on water resources (Arnell and Gosling 2013). Within this context, we present an assessment of the impact of different levels of global warming on river runoff, focusing on eight major river catchments across the world. The assessment has two main aims.

The first aim is to understand if there are systematic differences between projections of runoff change from two ensembles that are comprised of different types of hydrological model. Here, 'type' of model refers to whether a model is designed to operate at the global-scale

[8] Institute of Water Resources Management, Hydrology and Agricultural Hydraulic Engineering, Leibniz Hannover University, Hannover 30167, Germany

[9] Norwegian Water Resources and Energy Directorate, Oslo, Norway

[10] UFZ-Helmholtz Centre for Environmental Research, Leipzig, Germany

[11] Department of Environmental Sciences, Faculty of Agriculture and Environment, The University of Sydney, Sydney, Australia

[12] School of Nature Conservation, Beijing Forestry University, Beijing 10083, China

[13] Department of Hydrology and Hydraulic Engineering, Vrije Universiteit Brussel, Brussels, Belgium

[14] UNESCO-IHE Institute for Water Education, Delft, Netherlands

[15] Vrije Universiteit, Institute for Environmental Studies (IVM), Amsterdam, Netherlands

[16] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Center for Global Change and Water Cycle, Hohai University, Nanjing 210098, China

(global-scale hydrological model, Glob-HM) or at the catchment-scale (catchment-scale hydrological model, Cat-HM). Previous assessments have employed *either* ensembles comprised exclusively of Cat-HMs (e.g. Dams et al. 2015) *or* Glob-HMs (including global land surface models; e.g. Haddeland et al. 2011; Prudhomme et al. 2014) to assess the impact of climate change on runoff. Whilst some studies have employed one Glob-HM and one Cat-HM (Gosling et al. 2011), or one Glob-HM with two Cat-HMs (Thompson et al. 2013), none have compared impacts across two large multi-model ensembles of Glob-HMs and Cat-HMs. This is in part due to the significant resources that are required to setup and calibrate Cat-HMs. Knowledge of the relative spreads in projections from the two types of ensemble can help users decide which types of model to use in their assessments and can also provide decision-makers with an appreciation of the possible range in impacts that can arise from using different tools available to the hydrological modelling scientific community.

The second aim is to investigate the effect of different amounts of global-mean warming on runoff for each catchment. We investigate the impacts of three global warming scenarios from a consistent set of global climate model (GCM) projections: 1, 2 and 3 °C warming above pre-industrial levels. The eight catchments are of significance to human settlements and major ecosystems: the Upper Amazon, Darling, Ganges, Lena, Upper Mississippi, Upper Niger, Rhine and Tagus. The approach allows us to explore the extent to which current runoff will be affected if the UNFCCC 2 °C target is met, and what the impacts might be if it is missed and global-mean warming rises to 3 °C.

In this context the application of a Glob-HM ensemble and Cat-HM ensemble to assess global warming impacts across several catchments is a significant advancement because with a companion study that we conducted in parallel (Hattermann et al. 2016) it is the first time that impacts across two large multi-model ensembles of Glob-HMs and Cat-HMs have been compared. Up until now the two types of ensemble have been considered only independently. This advancement has been afforded by a large international collaboration of multiple Glob-HM and Cat-HM modelling groups brought together under the framework of the Inter-Sectoral Impact Model Inter-comparison Project (ISIMIP) over several years (Warszawski et al. 2014).

2 Methods

2.1 Study catchments

The eight large (>50,000 km^2) river catchments (Table 1; see Online Resource 1 for a map) cover all continents and several global hydro-regions defined by Meybeck et al. (2013), including Equatorial (Upper Amazon), Southern Mid-Latitudinal (Darling), Northern Mid-Latitudinal (Ganges, Upper Mississippi, Rhine, Tagus), Boreal (Lena) and Northern Sub-tropical (Upper Niger).

Gauging stations were identified for each catchment for which the Global Runoff Data Centre (GRDC) held observed discharge data, which were then later used to calibrate the Cat-HMs used in this analysis (Krysanova and Hattermann, this special issue). Availability of this data meant that sub-catchments had to be modelled instead of entire systems (Online Resource 1) and for three catchments (Upper Amazon, Upper Mississippi and Upper Niger) only the upper parts were modelled because of complex geomorphological structures and numerous human alterations further downstream, which would require extended input datasets and

Table 1 Catchments, gauges, upstream area of gauge, and the Glob-HMs and Cat-HMs that comprised each ensemble (shaded)

	Catchment	Upper Amazon	Darling	Ganges	Lena	Upper Mississippi	Upper Niger	Rhine	Tagus
	Gauge	Sao Paulo De Olivenca	Louth	Farakka	Stolb	Alton	Koulikoro	Lobith	Almourol
	Upstream drainage area (km^2)	990,781	489,300	835,000	2,460,000	444,185	120,000	160,800	67,490
Cat-HM ensemble	**HBV** (Bergstrom and Forsman 1973)	■		■		■	■	■	
	HYMOD (Boyle 2001)	■	■	■		■	■	■	
	mHM (Kumar et al. 2013; Samaniego et al. 2010)	■	■	■		■	■	■	
	SWIM (Krysanova et al. 1998)	■		■		■	■	■	
	VIC-Cat-HM * (Liang et al. 1994)	■	■	■	■	■	■	■	■
	WaterGAP3 + (Verzano 2009)	■		■	■	■	■	■	■
	HYPE (Lindstrom et al. 2010)				■			■	■
	SWAT (Arnold et al. 1993)	■				■			
	Number of Cat-HM simulations	7	3	6	3	7	6	7	3
Glob-HM ensemble	**LPJmL** (Bondeau et al. 2007)	■	■	■	■	■	■	■	■
	Mac-PDM.09 (Gosling and Arnell 2011)	■	■	■	■	■	■	■	■
	MPI-HM (Hagemann and Dümenil 1997)	■	■	■	■	■	■	■	■
	PCR-GLOBWB (Wada et al. 2014)	■	■	■	■	■	■	■	■
	WBMplus (Wisser et al. 2010)	■	■	■	■	■	■	■	■
	H08 (Hanasaki et al. 2008)	■	■	■	■	■	■	■	■
	MATSIRO (Pokhrel et al. 2011)	■	■	■	■	■	■	■	■
	VIC-Glob-HM * (Liang et al. 1994)	■	■	■	■	■	■	■	■
	DBH (Tang et al. 2007)	■	■	■	■	■	■	■	■
	Number of Glob-HM simulations	9	9	9	9	9	9	9	9
	Total number of simulations	**16**	**12**	**15**	**12**	**16**	**15**	**16**	**12**

* VIC-Cat-HM was developed from a Glob-HM counter-part (VIC-Glob-HM). VIC-Cat-HM was calibrated for each catchment and run in "catchment mode" similar to previous applications (Abdulla et al. 1996), whereas VIC-Glob-HM was not calibrated to the study catchments and it was run in "global mode". Thus they can be considered as different types of models. However, the Cat-HM and Glob-HM versions of this model are not completely independent but neither are any of the hydrological models because they all share some similar model parameterisations (e.g. method for calculating potential evapotranspiration)

+ The WaterGAP model can operate at the global-scale but the version employed here (WaterGAP3) operates on a 5 arcminute global grid, specifically so that it can be run at the catchment-scale. This is a much finer resolution than what the global-scale models used in the analysis operate at (0.5°x0.5°), so it is classed as a catchment-scale model in this analysis

significantly more resource intensive modelling approaches (Krysanova and Hattermann, this special issue).

2.2 Hydrological models

A total of nine Glob-HMs and eight Cat-HMs participated in the assessment (Table 1). The Cat-HMs comprised one ensemble and the Glob-HMs another. Not all Cat-HMs were run for every catchment, however, because unlike Glob-HMs they need to be calibrated on a catchment-by-catchment basis and run for each catchment individually, which is a resource- and time-intensive activity. This means that the Glob-HM ensembles contain nine simulations per catchment/scenario/GCM combination while the Cat-HM ensembles contain from 3 to 7 (all eight Cat-HMs were not run for any of the catchments - the number of Cat-HM simulations for each catchment is listed in Table 1). The Cat-HMs were calibrated and the performance of the calibration evaluated in a separate validation period (Huang et al. 2016) while the Glob-HMs were generally not calibrated to catchment-specific conditions. Water management and other human alterations on the fluvial system were not modelled by all Glob-HMs and Cat-HMs. Flows for every catchment were extracted and processed from the model simulated fields of daily discharge (m^3/s) time series and then converted to daily runoff (mm/day) using the upstream drainage area for each simulation.

We did not weight or exclude individual hydrological models based upon their performance at simulating present-day runoff (Hattermann et al. 2016) and we treated all hydrological models as being independent even though many share similar model parameterisations (e.g. method for calculating potential evapotranspiration; Online Resource 2).

All the models share similarities in their fundamental approach to modeling land-surface hydrological processes, such as simulating the land surface water balance, representing soil and vegetation across the catchment, calculating evapotranspiration, and applying routing schemes to transfer locally generated runoff over the catchment to the catchment outlet. However, the Glob-HMs and Cat-HMs differ in a fundamental way, which is the spatial scale at which the models represent hydrological processes and in turn the water balance resolved. All the Glob-HMs applied here operate with a 0.5°x0.5° spatial resolution grid across the global land-surface. The Cat-HMs employ various approaches. Three Cat-HMs run on a grid (mHM, VIC-Cat-HM, WaterGAP3), while four operate by splitting the catchment into sub-catchments and hydrological response units (HBV, HYPE, SWAT and SWIM), and one treats the entire catchment as a single entity (HYMOD).

An evaluation of the performance of the Glob-HMs and Cat-HMs across 11 catchments (Hattermann et al. 2016) shows that the correlation coefficients between simulated and observed long-term average seasonal dynamics, averaged over all models, is greater than 0.9 in 10 catchments for the Cat-HMs and in 4 catchments for the Glob-HMs (the coefficients are greater than 0.8 in 11 (Cat-HMs) and 8 (Glob-HMs) catchments). While the sensitivity of the Glob-HMs and Cat-HMs to observed climate variability is in general similar, the Glob-HM ensemble mean shows a large positive bias to observed data in annual flows for almost all 11 catchments and the spread across Glob-HMs is wider than for the Cat-HMs in the historical period (Hattermann et al. 2016). Detailed descriptions of the Glob-HMs and Cat-HMs employed in this study are provided by the respective references to each model cited in Table 1 and in Online Resource 2, whilst brief summaries are provided by Krysanova and Hattermann (this special issue).

2.3 Global warming scenarios

Daily input data on the climate variables required by each hydrological model (Online Resource 2) were extracted from five GCM simulations: HadGEM2-ES, IPSL-CM5A-LR, MIROC-ESM-CHEM, GFDL-ESM2 and NorESM1-M; all run under the RCP8.5 emissions pathway (Riahi et al. 2011) for the period 1971-2099 because this is the only pathway for which all five GCMs reach 3 °C by the end of their simulation period (year 2100). All variables were bias-corrected towards the WATCH observation-based dataset (Weedon et al. 2011), using an established method (Hempel et al. 2013), specifically designed to preserve long-term trends in temperature and precipitation projections to facilitate climate change impact assessments.

Daily hydrological simulations were performed with each Glob-HM and Cat-HM, for each catchment, using the daily climate data from each GCM as input (five simulations per Glob-HM/Cat-HM). Summaries of temperature and precipitation change are reported by Krysanova and Hattermann (this special issue). Daily discharge for the 31-year periods centred on the year whose global-mean temperature corresponds to different levels of global-mean warming relative to pre-industrial (1 °C, 2 °C, 3 °C) were extracted from each Glob-HM/Cat-HM simulation (Online Resource 3). At 3 °C the periods are centred around 2050 for three GCMs and around 2075 for two GCMs. Also extracted from each Glob-HM/Cat-HM simulation was the period 1980–2010 to represent the "present-day" (which corresponds to 0.6 °C above pre-industrial). This approach allowed us to estimate the effect of different amounts of global-mean warming (relative to pre-industrial) on catchment runoff, with the runoff changes characterised as changes relative to present-day (note that global-mean warming translates into different levels of warming across the different catchments).

2.4 Hydrological indicators

Five hydrological indicators were calculated from the daily timeseries of runoff:

1. Mean annual runoff (MAR): the mean over 31-years of the total daily runoff for each year in the timeseries.
2. Q5: the magnitude of daily runoff that is exceeded 5 % of the time in the timeseries, and thus an indicator of high flow.
3. Q95: the magnitude of daily runoff that is exceeded 95 % of the time in the timeseries, and thus an indicator of low flow.
4. The magnitude of maximum daily runoff associated with the 2, 5, 10, 20, 25, and 50-year return periods (see Online Resource 4 for specific details).
5. The magnitude of the minimum 7-day moving average of runoff associated with the 2, 5, 10, 20, 25, and 50-year return periods (see Online Resource 4 for specific details).

3 Results

3.1 Statistical distributions of changes in MAR, Q5 and Q95 with global warming

The median values of the two ensembles, corresponding to Glob-HMs and Cat-HMs, mostly respond consistently with each other to global warming (Fig. 1). We applied a Wilcoxon rank sum test to the 31-year median values of change in each hydrological indicator relative to

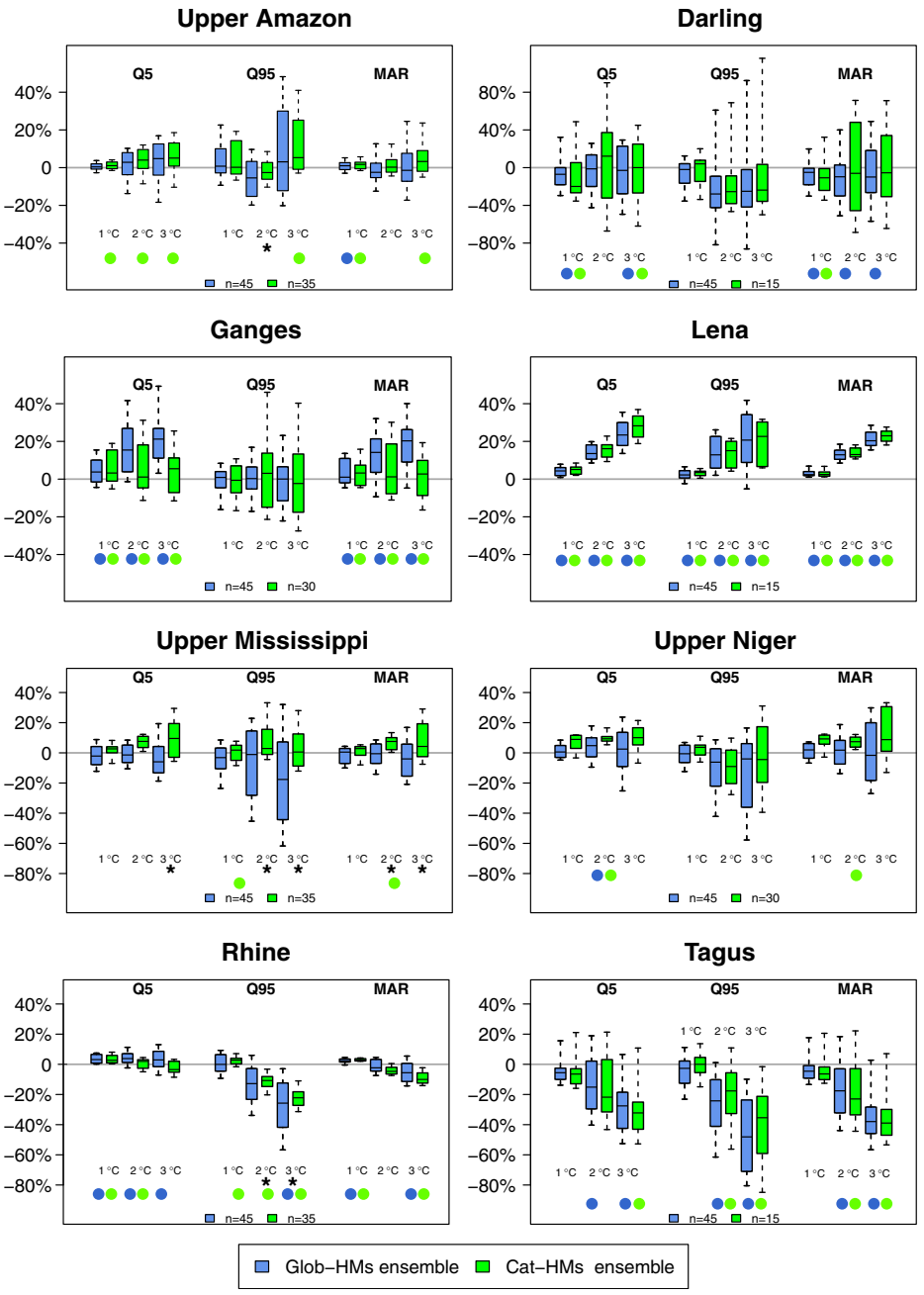


Fig. 1 Change (%; vertical axis) from present-day in three hydrological indicators (Q5, Q95, MAR) with three global warming scenarios (horizontal axis) for eight catchments. The box-whiskers show the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of changes in each hydrological indicator, for the Glob-HM and Cat-HM ensembles respectively (n denotes size of the ensemble, i.e. the number of GCM-hydrological model combinations). Asterisks denote where the Wilcoxon rank sum test rejects the null hypothesis of equal medians between the two ensembles at the 0.05 significance level. Filled circles denote where the projections under climate change for the models that comprise each ensemble represent a significant change from present-day according to a paired-samples t-test (0.05 significance level)

present (MAR, Q5 and Q95, respectively) as estimated by each ensemble. This was to test whether the median values for each ensemble were statistically different from each other. Each ensemble was treated independently (Muerth et al. 2013) because although the driving GCMs were the same, the models that yield the runoff simulations are independent. In the majority of catchments (5/8) there is no statistically significant ($p < 0.05$) difference between the medians of the two ensembles, for all three hydrological indicators and all global warming scenarios. Only for one catchment (Upper Mississippi) is there a significant difference for all three hydrological indicators.

Whilst there is some consistency in the median values between the two ensembles, the spreads are generally wider for the Glob-HM ensemble than the Cat-HM ensemble for the majority of catchments (Upper Amazon, Lena, Upper Mississippi, Upper Niger, Rhine, Tagus). It should also be noted that in some cases the medians of the two model ensembles differ in sign, so that one ensemble projects an increase in runoff and the other a decrease (for instance for MAR in Upper Amazon, Upper Mississippi, and Upper Niger).

Some catchments show a clear trend in their response to increases in global warming, with large differences between 2 and 3 °C that are consistent across both ensembles, as well as significant changes in hydrological indicators from present-day (assessed by a paired-samples t-test between the ensemble projections for each magnitude of global warming and the present-day simulations; Fig. 1). Examples include the Rhine, where there is an increased risk of decreases in low flows, with the median Q95 change escalating from around −11 % at 2 °C to −23 % at 3 °C (the change relative to present-day at 3 °C is statistically significant for both ensembles for Q95, $p < 0.05$). The risk of increases in high flows increases for the Lena, where the change in median Q5 intensifies from around +17 % (2 °C) to +26 % (3 °C), with a similar magnitude increase in MAR also (the changes are statistically significant for both ensembles at all magnitudes of global warming for all hydrological indicators, $p < 0.05$). The Tagus experiences declines in MAR and an enhanced risk of decreases in low flows, with changes in both MAR and Q95 of around −20 % at 2 °C, and then declining even further to around −40 % at 3 °C (the changes relative to present-day at 2 and 3 °C are statistically significant for both ensembles for MAR and Q95, $p < 0.05$). On the other hand, the Upper Mississippi and Upper Niger experience relatively small changes in the medians of hydrological indicators between 2 and 3 °C and the magnitude of changes relative to present are infrequently statistically significant.

3.2 Probability of changes in MAR, Q5 and Q95 with global warming

As an alternative approach to investigate the distribution of impacts from the Glob-HM and Cat-HM ensembles (as in Fig. 1), we calculated the probability of impacts of different magnitudes occurring from the impact distributions of the two ensembles (Fig. 2). For example, if 18 out of 45 simulations in an ensemble show that MAR increases by up to 30 % with global warming, then the probability of this magnitude of change is 40 %. Strictly, the probabilities should be interpreted as an indicator of the level of model agreement within an ensemble, not the level of certainty of a particular outcome occurring.

The two ensembles show similar responses to global warming for all three hydrological indicators for the Ganges, Lena and Tagus. However, for some catchments there is strong agreement between the ensembles for some indicators (such as increases in MAR and Q5 for the Upper Amazon; and decreases in MAR for the Rhine) but not for other indicators in the same catchment.

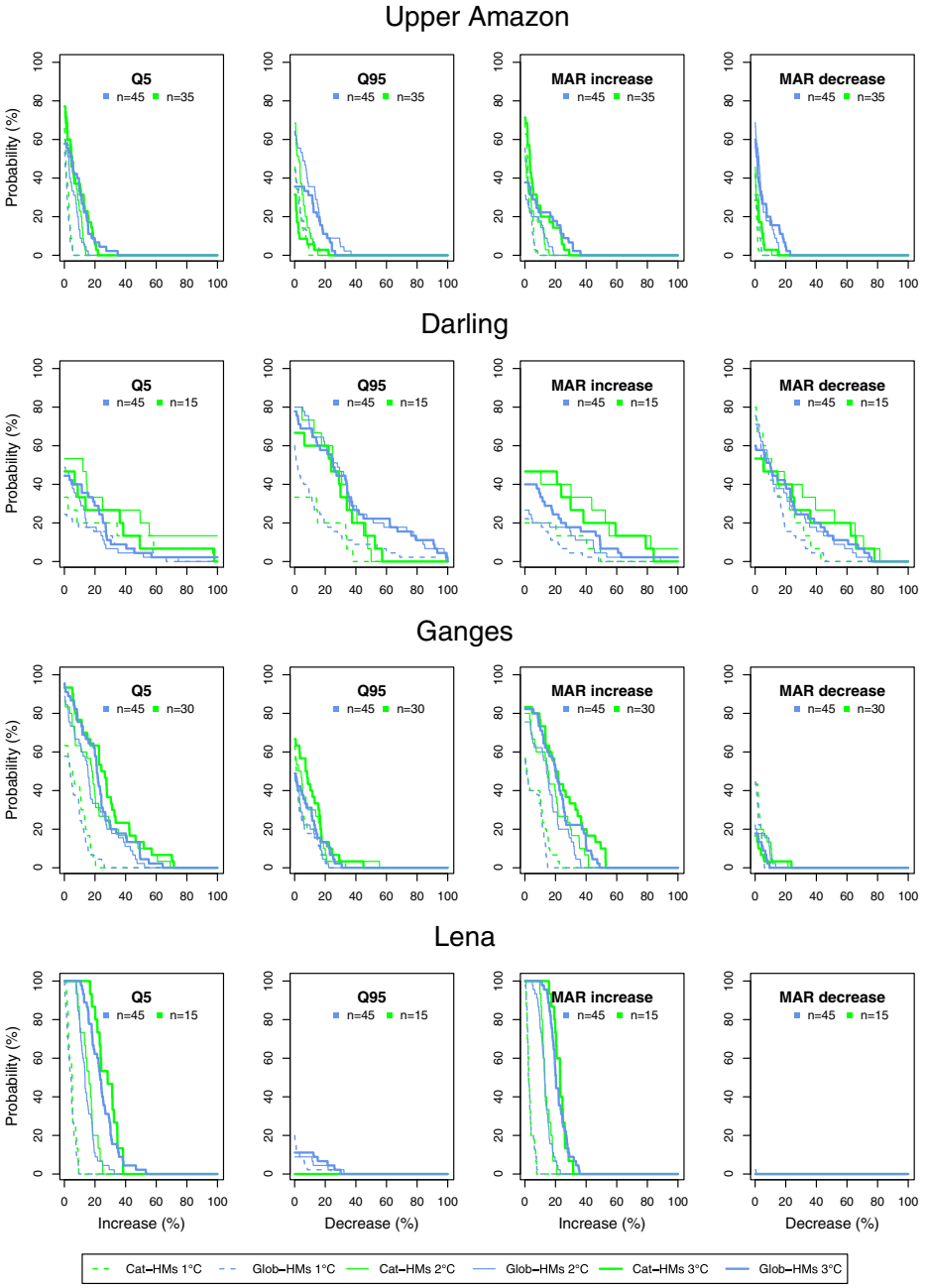


Fig. 2 The probability (%, vertical axes) under three global warming scenarios, of present-day Q5 increasing, Q95 decreasing, MAR increasing, and MAR decreasing, by different magnitudes (1–100 % of present day values, in 0.1 % increments; horizontal axes). Probability is calculated from the distribution of changes in each hydrological indicator as simulated by the Glob-HM and Cat-HM ensembles respectively (n denotes size of the ensemble)

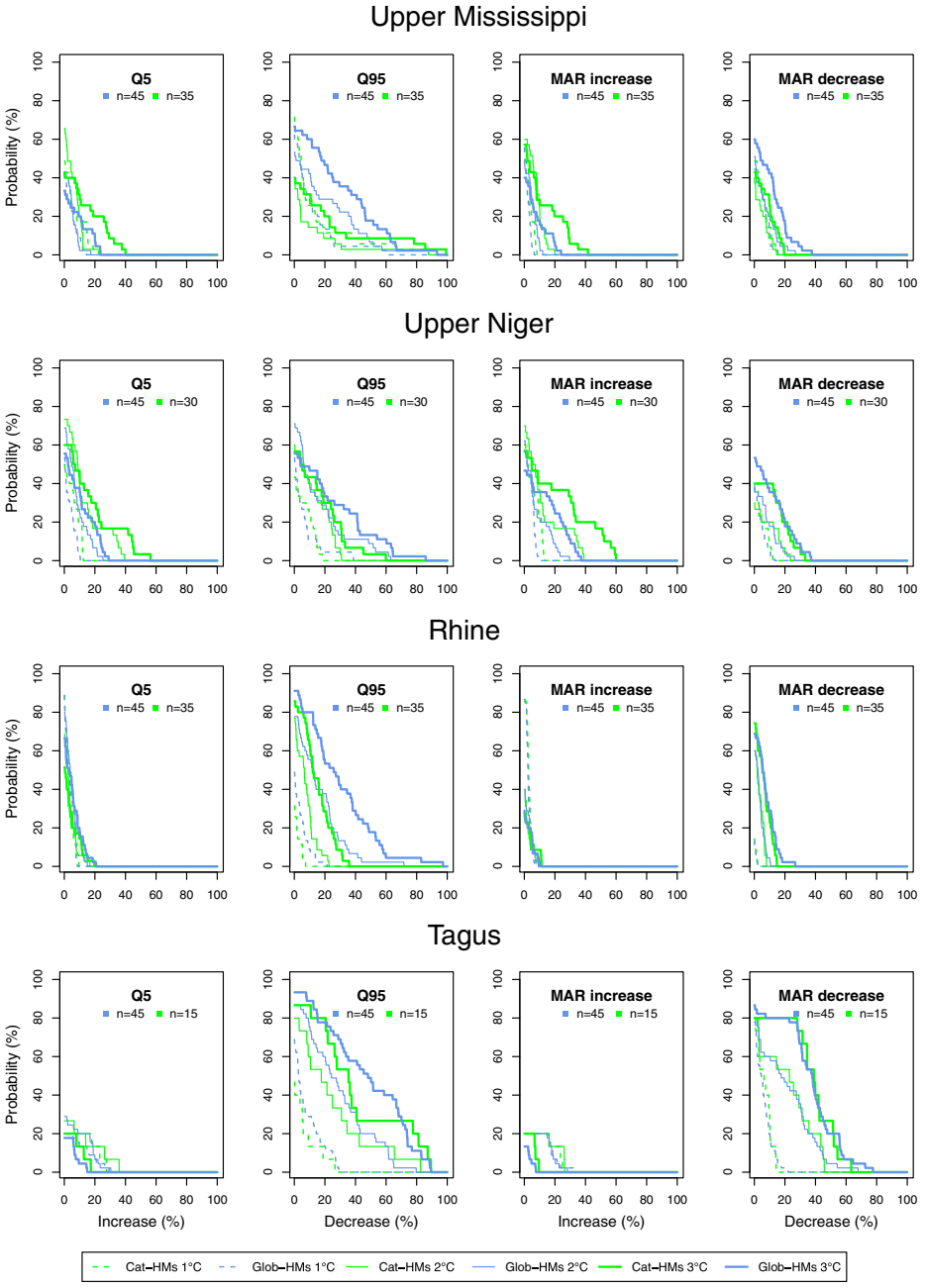


Fig. 2 (continued)

The probabilities of large decreases (>20 %) in Q95 are greater with the Glob-HM ensemble than with the Cat-HM ensemble for the majority (5) of catchments. For example at 3 °C, for the Darling, Upper Mississippi, Upper Niger, and Rhine, the probabilities of Q95

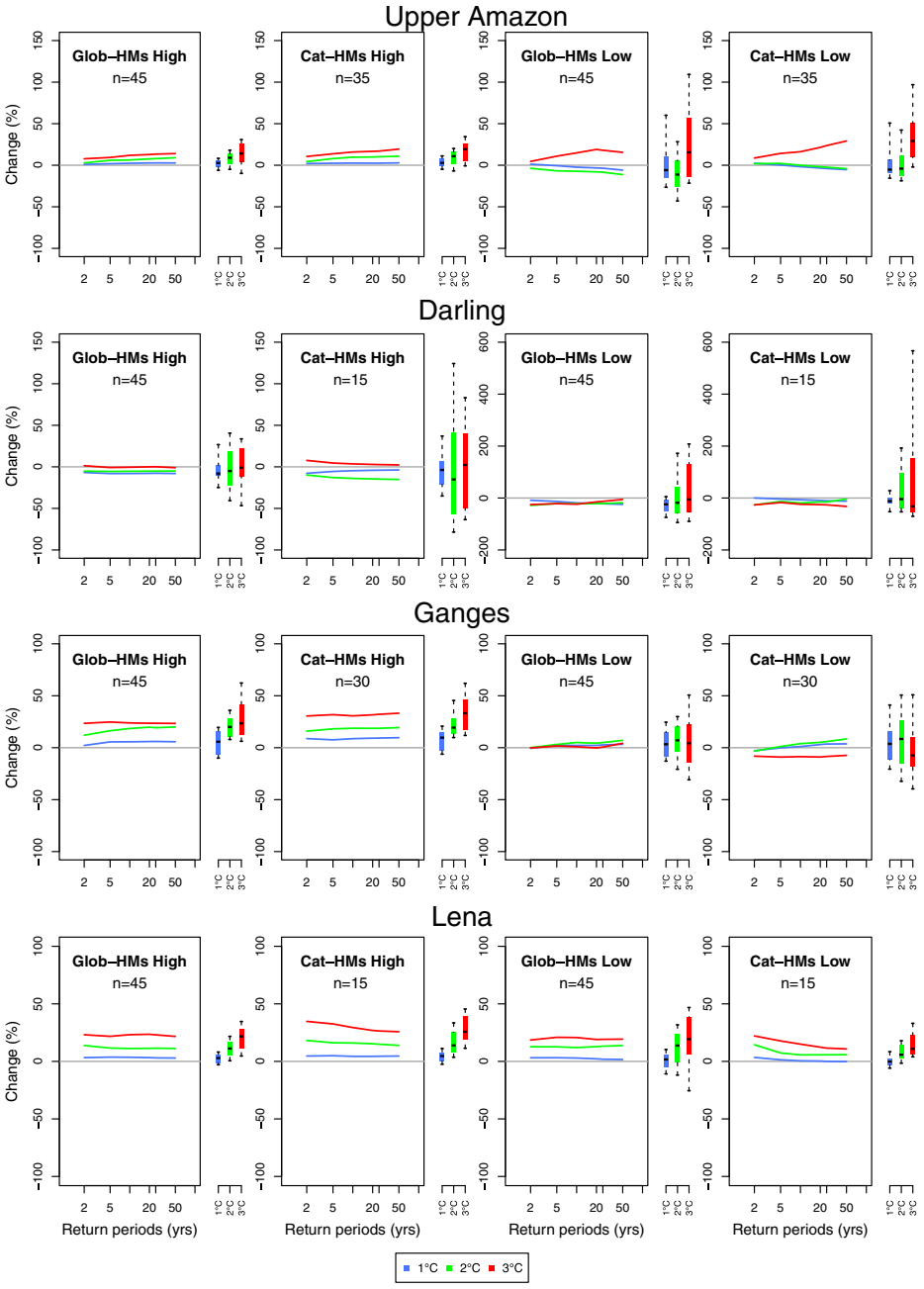


Fig. 3 The change relative to present-day (vertical axes) of maximum daily runoff ("High") and minimum 7-day moving average of runoff ("Low") associated with 2, 5, 10, 20, 25, and 50-year return periods (horizontal axes; log-scale) for the Glob-HM ensemble and Cat-HM ensemble. The lines show the median for each ensemble and the box-whiskers show the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of changes for the 50-year return period. The absolute values for each scenario and present-day (i.e. not change) are shown in Online Resource 5

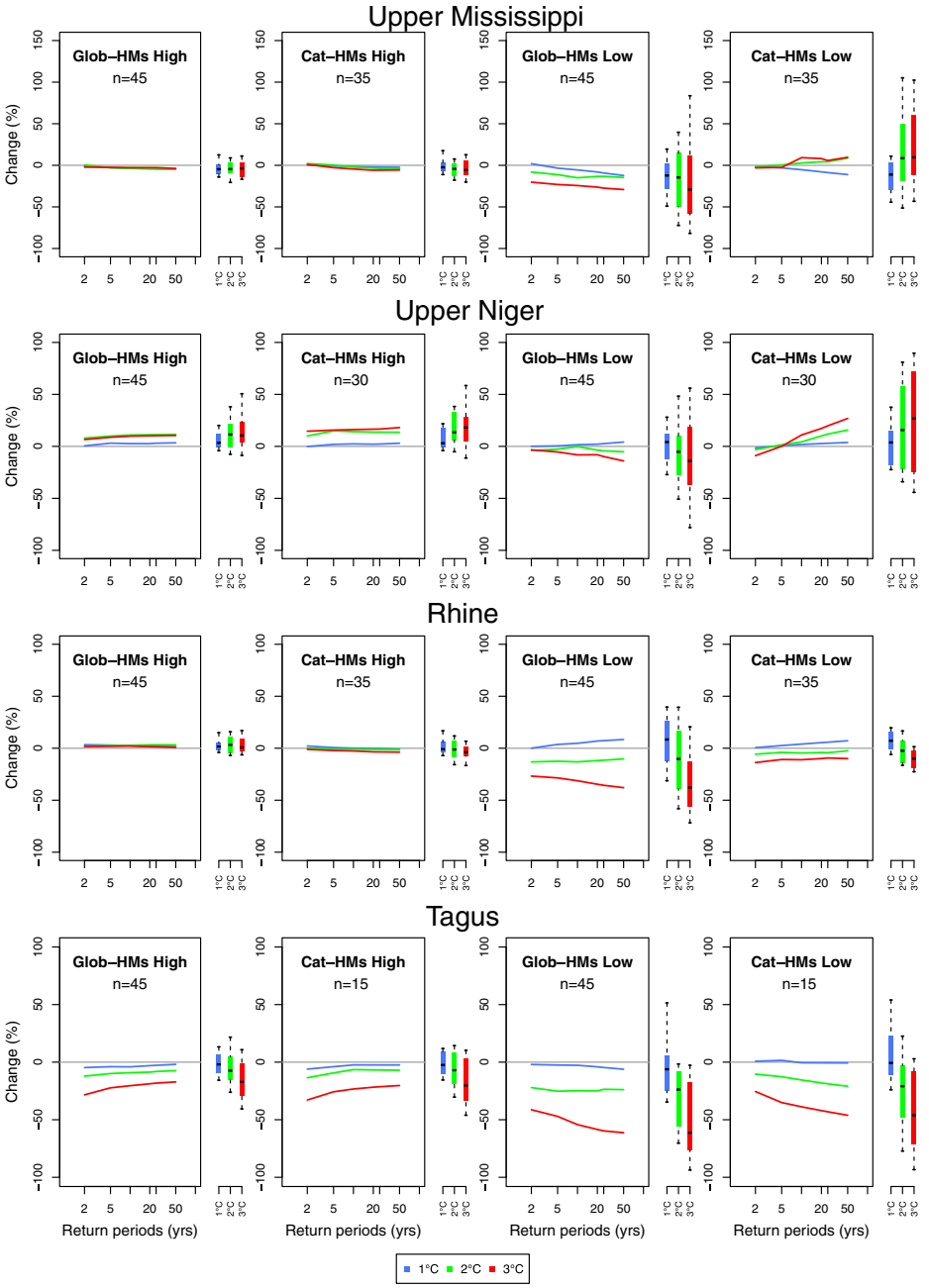


Fig. 3 (continued)

decreasing by up to 40 % according to the Glob-HM ensemble (Cat-HM ensemble) are: 26 % (20 %), 31 % (9 %), 24 % (7 %), and 29 % (0 %), respectively. This reverses in some cases for

large (>20 %) increases in Q5 though, where the probabilities are higher for the Cat-HM ensemble (Darling, Upper Mississippi, Upper Niger).

The probability of large decreases in MAR (>20 %) with 3 °C warming simulated by the Glob-HMs (Cat-HMs) are substantial for the Darling, Upper Niger and Tagus: 40 % (40 %), 18 % (20 %), and 80 % (80 %), respectively. These probabilities decline considerably when global warming is limited to 2 °C for the Upper Niger (from around 18 % at 3 °C to 4 % at 2 °C) and Tagus (from around 80 % at 3 °C to 47 % at 2 °C) (but not for the Darling where at 2 °C the probability is similar to at 3 °C). Likewise, substantial probabilities of large increases in MAR (>20 %) for the Ganges, Lena and Upper Amazon are considerably reduced at lower levels of global warming.

3.3 Changes in magnitude of return period runoff

The box-whiskers in Fig. 3 show that the Glob-HM and Cat-HM ensembles in most cases project changes that are similar in magnitude to each other. Two exceptions are for the Upper Mississippi and Upper Niger, where the Glob-HM ensemble projects declines in the magnitude of minimum flows associated with return periods greater than 5-years, while the Cat-HM ensemble projects increases. Also, for some catchments the spread of the Glob-HM ensemble is wider than for the Cat-HM ensemble (Upper Amazon, Rhine, Tagus). Whilst it appears for some catchments as though global warming has little effect on minimum (Darling, Ganges) and maximum flows (Upper Mississippi, Rhine), the medians disguise the large spreads of the ensembles.

Both ensembles project clear increases in the magnitude of maximum flows with global warming for the Upper Amazon, Ganges, and Lena, and decreases in the magnitude of minimum flows under 2 and 3 °C for the Rhine, and Tagus (Fig. 3), which are consistent with projected changes in Q5 and Q95 (Fig. 1). Some of the impacts intensify substantially with each degree of global warming. For example, the median changes from the Glob-HM ensemble in the 50-year return period minimum flow for 2 °C (and 3 °C) warming is −30 % (−60 %) for the Tagus and −10 % (−35 %) for the Rhine. Notably, the Upper Amazon shows a small decrease in the magnitude of minimum flows at the lower levels of global warming but a strong increase at 3 °C warming, due to a significant increase in precipitation between 2 and 3 °C warming (Krysanova and Hattermann, this special issue).

4 Discussion

4.1 Similarities and differences between the impacts simulated by the two ensembles

Our primary aim was to compare projections of runoff change from two ensembles of different types of hydrological models. We tested this by comparing impacts from an ensemble of Glob-HMs with an ensemble of Cat-HMs, both run with consistent input data from five GCMs.

The median estimates from both ensembles generally show that there is similarity in the effect of global warming on the hydrological indicators considered. This supports earlier work that showed climate change projections from a single Glob-HM were generally similar to those from a single Cat-HM (Gosling et al. 2011) but here the result is based upon a significantly more robust approach that employed ensembles of multiple Glob-HMs and Cat-HMs. There are a number of subtleties that

underlay the general conclusion that the two ensembles produce similar projections, however.

Firstly, the spread in projections of any one hydrological indicator for each ensemble is usually larger than the difference between the two ensemble medians. This means that the choice of hydrological models that comprise an ensemble, and the GCMs that provide the input data, are an important factor in determining the spread of the projections from the ensemble. Other work has quantified the relative contributions of hydrological models and GCMs to ensemble spread (Hagemann et al. 2013) with evidence suggesting that their relative contributions vary spatially across the globe (Dankers et al. 2014). Only three Cat-HMs comprised the Cat-HM ensemble for three catchments (Darling, Lena, Tagus). Whilst the inter-ensemble comparison would have been more consistent had we omitted these three catchments so that a similar number of models were included in each ensemble, our preference was to include as many catchments as possible to cover as many hydro-regions as possible. Clearly there is a loss of potential information where only three Cat-HMs are used compared to where six or seven are used. Thus we acknowledge that for these three catchments the Cat-HM ensemble spread could have been different if other Cat-HMs had been included in the ensemble.

Secondly, the difference in medians between the two ensembles is comparable to the climate change impact (difference between 1 and 3 °C) for the Ganges, Upper Mississippi and Upper Niger and there is sometimes a clear difference in sign of change between the ensembles at 3 °C warming (Upper Amazon MAR, Upper Mississippi MAR and Q5, Upper Niger MAR and Rhine Q5).

Thirdly, there are a small number of cases where the difference between the two ensembles could be large in absolute terms (mm), despite the Wilcoxon rank sum test indicating that the differences are not significant ($p < 0.05$), namely for the projections of Q5 and MAR for the Ganges.

Whilst Hattermann et al. (2016) investigated the relative abilities of the Glob-HMs and Cat-HMs to simulate present-day runoff in terms of long-term average seasonal dynamics, we did not weight and/or include/exclude individual models accordingly from each ensemble. This is because we wanted to assess impacts irrespective of individual model performance. A separate study is currently weighting models according to their performance. All hydrological models were treated as being independent even though many share similar model parameterisations (e.g. method for calculating potential evapotranspiration; Online Resource 2).

Despite these caveats the results highlight that the two modelling communities (Glob-HM and Cat-HM) seem to be approaching each other, both in terms of simulation results and also in their conceptual approach to model development. This is because some Cat-HMs can now be run at the continental scale with high resolution, such as a version of HYPE (E-HYPE, different from what was applied here; Donnelly et al. 2016; Hundecha et al. 2016) that can be run across around 35,000 sub-catchments in Europe, whilst some Glob-HMs are now starting to develop rigorous calibration procedures (Müller Schmied et al. 2014).

4.2 Comparing ensemble spread

Whilst the medians of the two ensembles generally display similar responses to global warming, the spread in projections from the Glob-HM ensemble is wider than from the Cat-HM ensemble in the majority of cases. A possible reason is that the Cat-HMs are constrained (calibrated) to specific catchment conditions (e.g., to reproduce observed discharge

characteristics) (Krysanova and Hattermann, this special issue), which means that they perform better than the Glob-HMs under present-day climate conditions (Hattermann et al. 2016).

It is also possibly the result of sampling bias introduced by the different sizes of the two ensembles, where the relatively smaller number of Cat-HMs represents only a limited section of the true hydrological model spread that would arise from using all the Cat-HMs that are in existence (Butts et al. 2004). Indeed, our study is not exhaustive in its inclusion of all available Glob-HMs and Cat-HMs, so the spread of the ensembles are likely underestimated.

However, even if we had included every Glob-HM and Cat-HM in existence, it remains possible that all models may miss (or not represent well enough) certain key processes, for instance the response of glaciers to global warming and in turn on runoff – thus even a "complete" ensemble may not sample the true spread (this also applies to the ensemble of GCMs that provide input to the hydrological models). Nevertheless, our study and a companion study that we conducted in parallel (Hattermann et al. 2016) are the first multi-catchment assessments to date of the impacts of climate change derived from multiple catchment-scale and global-scale hydrological models.

4.3 The impacts of global warming

We sought to demonstrate the effects of different amounts of global-mean warming on runoff for each catchment. Both ensembles clearly show that the Tagus and Rhine could see large increases in the magnitude of impacts between 1, 2, and 3 °C warming for all the considered hydrological indicators. Both catchments may experience declines in the magnitude of low and high flows and MAR with increasing warming (only at 3 °C with the Cat-HMs for high flows for the Rhine) due to large declines in precipitation (Tagus) or little change (Rhine) combined with increases in evapotranspiration due to warmer temperatures that are all simulated consistently by the five GCMs (Krysanova and Hattermann, this special issue). This supports earlier impact assessment studies conducted in these catchments (Kilsby et al. 2007; Kwadijk and Rotmans 1995; Middelkoop et al. 2001).

We also found clear evidence for increases in MAR and high flows for the Lena. The application of different snowmelt schemes by the hydrological models (see Online Resource 2) has little bearing on the projections. Temperature for the Lena rises most out of the eight study catchments under each 1 °C of global-mean warming (Krysanova and Hattermann, this special issue) – this in turn increases snowmelt runoff, confirming previous climate change assessments (Arora and Boer 2001) and observed changes in response to recent warming (Yang et al. 2002).

Our analysis shows no conclusive effects of global warming on runoff for the other catchments because the median changes are often close to zero with the 25th and 75th percentiles of the impacts distributions on either side.

The high spread seen for the Darling arises from it being a challenging catchment to model as it has a very low runoff coefficient and transmission losses are not well represented by many hydrological models (Gosling and Arnell 2011). Moreover, the performance of the Glob-HM ensemble has not been evaluated for its low flow simulation performance in the study catchments. Elsewhere it has been shown that Glob-HMs tend to struggle to simulate low flows (Giuntoli et al. 2015; Gudmundsson et al. 2012).

The large spread for the Ganges may be explained by the models' varying abilities to represent and account for (or not) dynamical changes in glacier mass balance with global

warming, for instance earlier work showed that Glob-HMs differ significantly in their simulations of snow accumulation in the Himalayan region (Haddeland et al. 2011).

It is well documented that the method for calculating potential evapotranspiration plays a significant role in determining simulated surface runoff changes (Haddeland et al. 2011; Thompson et al. 2013) and this can partly explain the large spread in simulated runoff for the Upper Amazon, Ganges, Upper Niger and Upper Mississippi. The hydrological models employ different approaches to simulate potential evapotranspiration (see Online Resource 2) which means that whilst the GCMs consistently simulate relative increases in both precipitation and temperature with global warming for these catchments (Krysanova and Hattermann, this special issue), there is diversity in the simulated surface runoff response.

The overall approach tied the analysis to prescribed global temperature changes so that the time periods for each magnitude of warming vary across GCMs instead of being fixed (Online Resource 3). This yields more robust regional patterns of change than if fixed periods are used because some of the uncertainty related to the GCMs' climate sensitivity is removed (Vautard et al. 2014). Although the approach is established (Arnell et al. 2014; Gosling et al. 2010; Roudier et al. 2015; Schewe et al. 2014) it does come with the caveat that it largely overlooks the rate at which, and when, each global-warming level is reached. Timing, as well as the magnitude of warming, is important (Lenton 2011), because adaptation is generally more straightforward and less costly with more time and natural systems may respond differently to slower rates of climate change. Some studies (Scholze et al. 2006) have overcome this by grouping simulations from multiple GCMs into categories of warming (e.g. <2 °C, 2–3 °C and >3 °C) for a fixed time period (e.g. 2071–2100) but this can reduce the precision and range for which global warming impacts can be assessed, especially where simulations from only a limited number of GCMs are available.

5 Conclusion

Along with a companion study that we conducted in parallel (Hattermann et al., this special issue) we have presented the first multi-catchment assessment of the impacts of climate change derived from multiple catchment- and global-scale hydrological models. Our results shed new light on the utility of global-scale hydrological models to understand changes in runoff at the catchment-scale because the median estimates of change from ensembles of Glob-HMs and Cat-HMs generally showed similar effects of global warming on the considered hydrological indicators. We did find some distinct differences though, such as a wider spread in impacts from the Glob-HMs for several catchments.

Although the sign and magnitude of change with global warming for some catchments is unclear (Upper Amazon, Darling, Ganges, Upper Niger and Upper Mississippi) we found more compelling evidence for the Rhine, Tagus and Lena. For these latter catchments in particular, it is clear that stringent global climate change mitigation action in line with Article 2 of the Paris Agreement (UNFCCC 2015) could have significant benefits in terms of avoiding some of the hydrological hazards that could be seen with higher magnitudes of global warming such as 3 °C.

A rationale for our study was that knowledge of the relative spreads in projections from the two types of ensemble could provide decision-makers with an appreciation of the possible range in impacts that can arise from using different tools available to the hydrological

modelling scientific community. In this context a decision-maker might ask: "If I am to conduct a climate impact study in e.g. the Mississippi catchment, which model type would I favor for assessing future runoff?" In light of our results, we would advise against favoring any one type of ensemble. Whilst the Cat-HM ensemble has been shown to perform better than the Glob-HM ensemble when simulating present-day discharge (Hattermann et al., this special issue), this does not guarantee higher credibility of projections under climate change. To this end we would encourage decision-makers to consider the full range of spreads from both the Glob-HM and Cat-HM ensembles, since these represent the spread of possible impacts according to several different types of models used by the hydrological modelling scientific community.

Acknowledgments This work has been conducted under the framework of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP), using Glob-HM and Cat-HM simulations conducted as part of the ISIMIP Fast Track, so our thanks go to the Glob-HM and Cat-HM modellers who submitted results to this project. The ISIMIP Fast Track was funded by the German Ministry of Education and Research, with project funding reference number 01LS1201A. To obtain access to data, please refer to the "data archive" section of the project portal www.isimip.org. JD, JL and XZ were supported by the Beijing Natural Science Foundation Grant (8151002), the National Natural Science Foundation of China (41571022), and the National Science and Technology Major Project (2015ZX07203-005). IH was supported by grant no. 243803/E10 from the Norwegian Research council. JZ was supported by the Islamic Development Bank.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Abdulla FA, Lettenmaier DP, Wood EF, Smith JA (1996) Application of a macroscale hydrologic model to estimate the water balance of the Arkansas-Red River Basin. J Geophys Res: Atmos 101:7449–7459

Arnell NW, Gosling SN (2013) The impacts of climate change on river flow regimes at the global scale. J Hydrol 486:351–364

Arnell NW, Brown S, Gosling SN, Hinkel J, Huntingford C, Lloyd-Hughes B, Lowe JA, Osborn T, Nicholls RJ, Zelazowski P (2014) Global-scale climate impact functions: the relationship between climate forcing and impact. Clim Chang 134:475–487

Arnell NW, Brown S, Gosling SN, Gottschalk P, Hinkel J, Huntingford C, Lloyd-Hughes B, Lowe JA, Nicholls RJ, Osborn TJ, Osborne TM, Rose GA, Smith P, Wheeler TR, Zelazowski P (2016) The impacts of climate change across the globe: a multi-sectoral assessment. Clim Chang 134:457–474

Arnold JG, Allen PM, Bernhardt G (1993) A comprehensive surface-groundwater flow model. J Hydrol 142:47–69

Arora VK, Boer GJ (2001) Effects of simulated climate change on the hydrology of major river basins. J Geophys Res: Atmos 106:3335–3348

Bergstrom S, Forsman A (1973) Development of a conceptual deterministic rainfall-runoff model. Hydrol Res 4:147–170

Bondeau A, Smith PC, Zaehle S, Schaphoff S, Lucht W, Cramer W, Gerten D, Lotze-Campen H, MÜLler C, Reichstein M, Smith B (2007) Modelling the role of agriculture for the 20th century global terrestrial carbon balance. Glob Chang Biol 13:679–706

Boyle DP (2001) Multicriteria calibration of hydrological models. PhD Dissertation, University of Arizona, Tucson, Arizona, USA

Butts MB, Payne JT, Kristensen M, Madsen H (2004) An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. J Hydrol 298:242–266

CAT (2015) Climate Action Tracker

Dams J, Nossent J, Senbeta TB, Willems P, Batelaan O (2015) Multi-model approach to assess the impact of climate change on runoff. J Hydrol 529(Part 3):1601–1616

Donnelly, C, Andersson, J.C.M. and Arheimer, B., 2016. Using flow signatures and catchment similarities to evaluate a multi-basin model (E-HYPE) across Europe. Hydr. Sciences Journal 6:255-273

Giuntoli I, Villarini G, Prudhomme C, Mallakpour I, Hannah DM (2015) Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States. J Geophys Res: Atmos 120:9138–9159

Gosling SN, Arnell NW (2011) Simulating current global river runoff with a global hydrological model: model revisions, validation, and sensitivity analysis. Hydrol Process 25:1129–1145

Gosling SN, Bretherton D, Haines K, Arnell NW (2010) Global hydrology modelling and uncertainty: running multiple ensembles with a campus grid. Philos Trans Royal Soc Lond A: Math Phys Eng Sci 368:4005–4021

Gosling SN, Taylor RG, Arnell NW, Todd MC (2011) A comparative analysis of projected impacts of climate change on river runoff from global and catchment-scale hydrological models. Hydrol Earth Syst Sci 15:279–294

Gudmundsson L, Tallaksen LM, Stahl K, Clark DB, Dumont E, Hagemann S, Bertrand N, Gerten D, Heinke J, Hanasaki N, Voss F, Koirala S (2012) Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe. J Hydrometeorol 13:604–620

Haddeland I, Clark DB, Franssen W, Ludwig F, Voß F, Arnell NW, Bertrand N, Best M, Folwell S, Gerten D, Gomes S, Gosling SN, Hagemann S, Hanasaki N, Harding R, Heinke J, Kabat P, Koirala S, Oki T, Polcher J, Stacke T, Viterbo P, Weedon GP, Yeh P (2011) Multimodel estimate of the global terrestrial water balance: setup and first results. J Hydrometeorol 12:869–884

Hagemann S, Dümenil L (1997) A parametrization of the lateral waterflow for the global scale. Clim Dyn 14:17–31

Hagemann S, Chen C, Clark DB, Folwell S, Gosling SN, Haddeland I, Hanasaki N, Heinke J, Ludwig F, Voss F, Wiltshire AJ (2013) Climate change impact on available water resources obtained using multiple global climate and hydrology models. Earth Syst Dyn 4:129–144

Hanasaki N, Kanae S, Oki T, Masuda K, Motoya K, Shirakawa N, Shen Y, Tanaka K (2008) An integrated model for the assessment of global water resources – Part 1: model description and input meteorological forcing. Hydrol Earth Syst Sci 12:1007–1025

Hattermann FF, Krysanova V, Gosling SN, Dankers R, Daggupati P, Donnelly C, Flörke M, Huang Sh, Motovilov Yu, Buda S, Yang T, Müller C, Leng G, Tang Q, Portmann FT, Hagemann S, Gerten D, Wada Y, Masaki Y, Alemayehu T, Satoh Y, Samaniego L (2016) Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins. Clim Chang. doi.10.1007/s10584-016-1829-4

Hempel S, Frieler K, Warszawski L, Schewe J, Piontek F (2013) A trend-preserving bias correction-the ISI-MIP approach. Earth Syst Dyn 4:219–236

Huang S, Kumar R, Flörke M, Yang T, Hundecha Y, Kraft P, Gao C, Gelfan A, Liersch S, Lobanova A, Strauch M, Ogtrop FV, Reinhardt J, Haberlandt U, Krysanova V (2016) Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide. Clim Chang. doi:10.1007/s10584-016-1841-8

Hundecha Y, Arheimer B, Donnelly C, Pechlivanidis I (2016) A regional parameter estimation scheme for a pan-European multi-basin model. Journal of Hydrology: Regional Studies, 6: 90-111 doi:10.1016/j.ejrh.2016.04.002

Kilsby CG, Tellier SS, Fowler HJ, Howels TR (2007) Hydrological impacts of climate change on the Tejo and Guadiana Rivers. Hydrol Earth Syst Sci 11:1175–1189

Krysanova V, Müller-Wohlfeil D-I, Becker A (1998) Development and test of a spatially distributed hydrological/water quality model for mesoscale watersheds. Ecol Model 106:261–289

Kumar R, Samaniego L, Attinger S (2013) Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. Water Resour Res 49:360–379

Kwadijk J, Rotmans J (1995) The impact of climate change on the river rhine: a scenario study. Clim Chang 30: 397–425

Lenton TM (2011) Beyond 2°C: redefining dangerous climate change for physical systems. Wiley Interdiscip Rev Clim Chang 2:451–461

Liang X, Lettenmaier DP, Wood EF, Burges SJ (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J Geophys Res: Atmos 99:14415–14428

Lindstrom G, Pers C, Rosberg J, Stromqvist J, Arheimer B (2010) Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. Hydrol Res 41:295–319

Meybeck M, Kummu M, Dürr HH (2013) Global hydrobelts and hydroregions: improved reporting scale for water-related issues? Hydrol Earth Syst Sci 17:1093–1111

Middelkoop H, Daamen K, Gellens D, Grabs W, Kwadijk JCJ, Lang H, Parmet BWAH, Schädler B, Schulla J, Wilke K (2001) Impact of climate change on hydrological regimes and water resources management in the Rhine Basin. Clim Chang 49:105–128

Muerth MJ, Gauvin St-Denis B, Ricard S, Velázquez JA, Schmid J, Minville M, Caya D, Chaumont D, Ludwig R, Turcotte R (2013) On the need for bias correction in regional climate scenarios to assess climate change impacts on river runoff. Hydrol Earth Syst Sci 17:1189–1204

Müller Schmied H, Eisner S, Franz D, Wattenbach M, Portmann FT, Flörke M, Döll P (2014) Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. Hydrol Earth Syst Sci 18:3511–3538

Pokhrel Y, Hanasaki N, Koirala S, Cho J, Yeh PJF, Kim H, Kanae S, Oki T (2011) Incorporating anthropogenic water regulation modules into a land surface model. J Hydrometeorol 13:255–269

Prudhomme C, Giuntoli I, Robinson EL, Clark DB, Arnell NW, Dankers R, Fekete BM, Franssen W, Gerten D, Gosling SN, Hagemann S, Hannah DM, Kim H, Masaki Y, Satoh Y, Stacke T, Wada Y, Wisser D (2014) Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment. Proc Natl Acad Sci 111:3262–3267

Riahi K, Rao S, Krey V, Cho C, Chirkov V, Fischer G, Kindermann G, Nakicenovic N, Rafaj P (2011) RCP 8.5—A scenario of comparatively high greenhouse gas emissions. Clim Chang 109:33–57

Roudier P, Andersson JM, Donnelly C, Feyen L, Greuell W, Ludwig F (2015) Projections of future floods and hydrological droughts in Europe under a +2°C global warming. Climatic Change:1-15

Samaniego L, Kumar R, Attinger S (2010), Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resour Res 46: W05523

Schewe J, Heinke J, Gerten D, Haddeland I, Arnell NW, Clark DB, Dankers R, Eisner S, Fekete BM, Colón-González FJ, Gosling SN, Kim H, Liu X, Masaki Y, Portmann FT, Satoh Y, Stacke T, Tang Q, Wada Y, Wisser D, Albrecht T, Frieler K, Piontek F, Warszawski L, Kabat P (2014) Multimodel assessment of water scarcity under climate change. Proc Natl Acad Sci 111:3245–3250

Scholze M, Knorr W, Arnell NW, Prentice IC (2006) A climate-change risk analysis for world ecosystems. Proc Natl Acad Sci 103:13116–13120

Tang Q, Oki T, Kanae S, Hu H (2007) The influence of precipitation variability and partial irrigation within grid cells on a hydrological simulation. J Hydrometeorol 8:499–512

Thompson JR, Green AJ, Kingston DG, Gosling SN (2013) Assessment of uncertainty in river flow projections for the Mekong River using multiple GCMs and hydrological models. J Hydrol 486:1–30

UNFCCC (2015) Adoption of the Paris Agreement. Proposal by the President. Proposal by the President. Available from: http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf. United Nations, Geneva, Switzerland

Vautard R, Gobiet A, Sobolowski S, Kjellström E, Stegehuis A, Watkiss P, Mendlik T, Landgren O, Nikulin G, Teichmann C, Jacob D (2014) The European climate under a 2 °C global warming. Environ Res Lett 9:034006

Verzano K (2009) Climate change impacts on flood related hydrological processes: Further development and application of a global scale hydrological model, no. 71-2009 in Reports on Earth System Science. Max Planck Institute for Meteorology, Hamburg, Germany

Wada Y, Wisser D, Bierkens MFP (2014) Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. Earth Syst Dyn 5:15–40

Warszawski L, Frieler K, Huber V, Piontek F, Serdeczny O, Schewe J (2014) The inter-sectoral impact model intercomparison project (ISI–MIP): project framework. Proc Natl Acad Sci 111:3228–3232

Weedon GP, Gomes S, Viterbo P, Shuttleworth WJ, Blyth E, Österle H, Adam JC, Bellouin N, Boucher O, Best M (2011) Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century. J Hydrometeorol 12:823–848

Wisser D, Fekete BM, Vörösmarty CJ, Schumann AH (2010) Reconstructing 20th century global hydrography: a contribution to the Global Terrestrial Network- Hydrology (GTN-H). Hydrol Earth Syst Sci 14:1–24

Yang D, Kane DL, Hinzman LD, Zhang X, Zhang T, Ye H (2002) Siberian Lena River hydrologic regime and recent change. J Geophys Res: Atmos 107:4694

Paper IV: Multi-model combination with a super-ensemble of catchment-scale and global-scale hydrological models

Multi-model combination with a super-ensemble of catchment-scale and global-scale hydrological models

Jamal Zaherpour[1*], Simon N Gosling[1], Nick Mount[1], Fred Hattermann[2]

[1]School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom.
[2]Potsdam Institute for Climate Impact Research, Telegrafenberg, 14473 Potsdam, Germany

*Corresponding author:
Tel.: +44 115 951 5428, Fax: +44 (0)115 951 5249
E-mail addresses: zaherpour@gmail.com , lgxjz1@nottingham.ac.uk
Postal address: School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom

Abstract

In this study we capitalise on the timely opportunity to bring together two communities in the field of hydrological modelling, who have respectively developed ensembles of catchment- and global-scale hydrological models (CHMs and GHMs). Through doing so we construct one of the few "super-ensembles" to ever be developed, which comprises both model types. A novel machine learning technique, gene expression programming (GEP), is applied to the super-ensemble. This facilitates the computationally intelligent selection and weighting of models – based upon the ability of individual models to replicate historical runoff – into a single solution known as a multi-model combination (MMC). The MMC solution is represented as a regression equation that includes constants and runoff values from the selected input models. We find that a MMC solution developed from the super-ensemble performs better on average across catchments than the ensemble means of the separate GHM and CHM ensembles. Thus we identify particular value in pooling multiple hydrological models from both communities into a single super-ensemble. Moreover, we demonstrate value in hydrological models that perform relatively weakly in terms of their ability to replicate historical runoff because they still inherently replicate certain aspects of the hydrological regime that the better performing models might not. Through gene expression programming, these aspects of the weaker models can be used to adjust aspects of the better performing models that require improvement. Whilst we identify the value of the MMC technique and the construction of super-ensembles at the aggregated catchment level, i.e. as an average across multiple catchments, we find that the improvements in performance delivered by MMC can vary at the individual catchment level, by month of the year, and also by type of model and hydrological indicator. Furthermore, the constrained complexity of the MMCs can affect their performance.

Key words: super ensemble; hydrological modelling; model evaluation; model weighting; global hydrology; catchment hydrology

Declarations of interest:

None

1. Introduction

Within the field of hydrological modelling there exist a multitude of catchment- and global-scale hydrological models (CHMs and GHMs, respectively). Such models have been developed to enhance scientific understanding of hydrological processes and their interactions with each other, as well as to act as tools that support decision-making. They are often used to assess flood hazards (Zhao et al., 2017), water scarcity (Liu et al., 2017), droughts (van Huijgevoort et al., 2013), dam operations (Masaki et al., 2017) and the effects of water management on catchment hydrology (Veldkamp et al., 2017).

GHMs typically operate across continental and global scales. They include a trade-off between maximising overall model performance and the performance that can be achieved for any individual catchment because model parameters have to be spatially generalised and because limited availability of global datasets of observed discharge with good spatial coverage and quality precludes comprehensive model calibration (Gosling and Arnell, 2011, Müller Schmied et al., 2014). On the other hand, CHMs are tailored to specific catchments, which means that they use local observed data and to this end can be calibrated comprehensively. The spatial resolution at which physical processes are represented is also a distinguishing feature between GHMs and CHMs: the resolution is often finer for CHMs.

Despite the rapid growth of hydrological models in the respective GHM and CHM communities over the past few decades, there have been very few collaborative efforts between the two communities. The two types of model are typically developed and applied in isolation of each other (Zhou et al., 2012, van Dijk et al., 2013, Beck et al., 2015, Beck et al., 2016, Zaherpour et al., 2018a, Breuer et al., 2009, Bosshard et al., 2013, Chen et al., 2012, Vetter et al., 2015), which has led to the development of separate ensembles of GHMs and CHMs, which are rarely combined (Gosling et al., 2011, Piniewski et al., 2014, Gosling et al., 2016, Hattermann et al., 2017, Hattermann et al., 2018, Thompson et al., 2013). The timely opportunity exists to explore the potential of using what we call “super-ensembles”, where multiple CHMs and GHMs are combined into a single ensemble.

In this study we bring together the two communities to develop one such super-ensemble.

Past studies of CHM and GHM ensembles present simulated runoff as a range across models and/or in terms of the ensemble mean (EM) or median (Hattermann et al., 2018, Gosling et al., 2010, Hattermann et al., 2017, van Huijgevoort et al., 2013, Gosling et al., 2016, Thompson et al., 2013). Although the EM provides a simple indicator of how an ensemble of models performs, it may not necessarily be the ideal indicator to use. The implicit assumption underpinning the justification for computing an EM, that overall the diversity of model outputs will contribute towards improving the integrated output, cannot be guaranteed, even though some studies have shown that the EM can be closer to observations (Gao and Dirmeyer, 2006, Guo et al., 2007). Certain models may be consistently poor performers in certain climatic or physiographic settings, or over certain hydrological response ranges, and their inclusion in a model ensemble may act to limit the performance of the EM (Zaherpour et al., 2018a).

An alternative approach to the EM is to combine the ensemble members by assigning them different weights, instead of equal weights (as is in the case with the EM). This approach is known as multi-model combination (MMC). Intelligent MMC methods combine several models' outputs into a single, integrated solution so that the best aspects of the best models are emphasised, and the worst aspects of the worst-performing models are de-emphasised and/or omitted. MMC was first applied in economics (Bates and Granger, 1969). It has since been applied to CHM modelling (Shamseldin et al., 1997, Ajami et al., 2006, Arsenault et al., 2015, Li et al., 2015) but apart from one study (Zaherpour et al., 2018b) its application is unprecedented with GHMs, and unique for a super-ensemble of GHMs and CHMs.

Accordingly, here, we present the first comparison of MMCs developed from GHMs and CHMs respectively, as well as the first application of MMC to a super-ensemble of both model types. This is achieved by applying Gene Expression Programming (GEP) (Ferreira, 2001, Ferreira, 2006) to an ensemble of eight GHMs and an ensemble of nine CHMs across eight study catchments (Figure 1). GEP is a machine

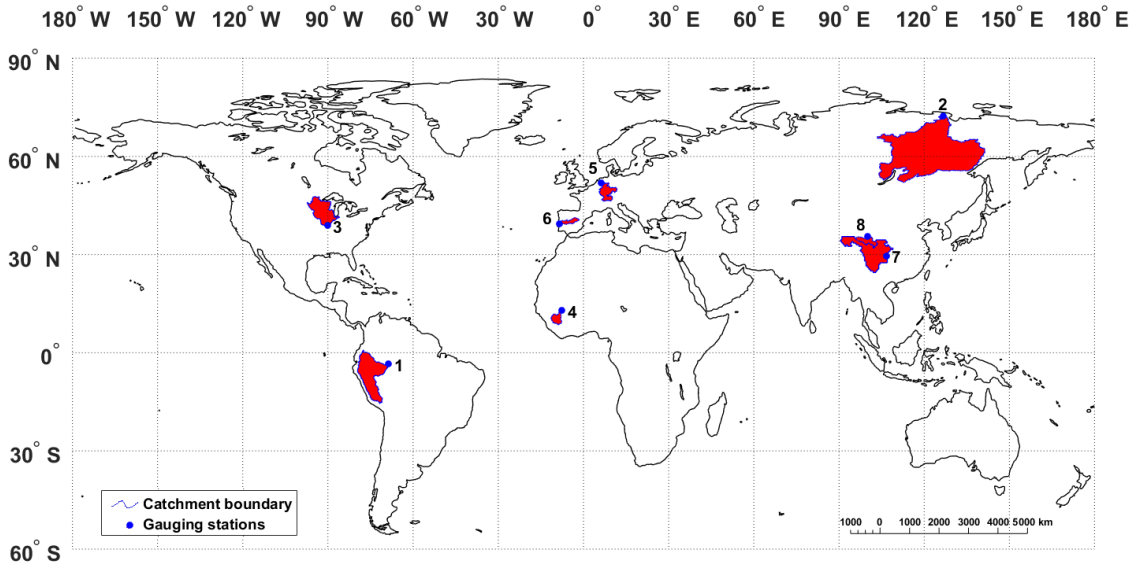
learning algorithm that has demonstrable success in delivering MMCs for hydrological (Bărbulescu and Băutu, 2009, Barbulescu and Bautu, 2010) and climatological modelling (Fernando et al., 2012, Aytek et al., 2008, Beriro et al., 2013, Traore and Guven, 2011, Barbulescu and Bautu, 2010, Beriro, 2015).

Our study aimed to answer the following two research questions:

1) Do MMCs of GHMs and CHMs respectively, improve the performance of the ensemble, relative to the EM and/or the best performing individual model; and is the magnitude of improvement the same for each ensemble?

2) Does an MMC developed from a "super-ensemble" of GHMs and CHMs outperform MMCs derived from an ensemble of each model type respectively?

This research aimed, therefore, to demonstrate the utility and value (or limited nature) of constructing super-ensembles and also of applying MMC to such an ensemble.



Catchments: 1. Upper Amazon 2. Lena 3. Upper Mississippi 4. Upper Niger
5. Rhine 6. Tagus 7. Yangtze 8. Yellow

Figure 1: Location of the eight study catchments and gauging stations.

2. Methods

2.1 Experimental design

We addressed our research questions in turn through an experimental design in which we:

1) Derived MMCs for the GHM and CHM ensembles respectively, then compared them with the EMs and best performing individual models from each respective ensemble and between ensembles.

2) Derived MMCs from a "super-ensemble" that includes all GHMs and CHMs, then compared their performance to MMCs derived from only GHMs and CHMs, respectively.

2.2. Study catchments and observed data

We capitalised on GHM and CHM simulations conducted for fourteen catchments under the framework of the Inter-Sectoral Impact Model Intercomparison Project, ISIMIP (Warszawski et al., 2014). Our study focussed on a sub-set of eight of these catchments (Figure 1 and Table 1) because, for the purposes of robust model evaluation, it was important that a complete and substantial time-series (at least 25 years) of observed discharge data was available for each catchment. Gauging stations were identified for each catchment for which the Global Runoff Data Centre (GRDC, http://grdc.bafg.de) held observed discharge data for a period of at least this duration. The same discharge data were used to calibrate the CHMs that take part in this study (Krysanova and Hattermann, 2017). For each catchment, mean monthly runoff was derived by dividing the mean monthly discharge values by the area upstream of the gauging station, for the models and observed data. The selected catchments cover several continents and several global hydro-regions as defined by Meybeck et al. (2013) including Equatorial (Upper Amazon), Northern Mid-Latitudinal (Upper Mississippi, Rhine, Tagus), Boreal (Lena) and Northern Subtropical (Upper Niger).

2.3. Input models to the MMC

In total, eight GHMs and nine CHMs, comprising two separate ensembles, were used in the study (Table 1). However, not all CHMs were run for every catchment, because unlike GHMs they need to be calibrated on a catchment-by-catchment basis and run for each catchment individually, which is a resource and time-intensive activity. This means that the GHM ensembles contain eight simulations per catchment, while the CHM ensembles contain between 4 and 8 (all nine CHMs were not run for any of the catchments; Table 1). The CHMs were calibrated (Huang et al., 2016) while all the GHMs, except for WaterGAP2, were generally not calibrated to catchment-specific conditions. Water management and other human alterations on the fluvial system were not modelled by all GHMs and CHMs.

All GHMs and CHMs were run for the period 1971–2001 with input climate data provided by the Global Soil Wetness Project 3, GSWP3. We used GSWP3 because it represents realistic sub-monthly variability over the entire twentieth century (1901-2010), as it has been generated based on the Twentieth Century Reanalysis project (20CR; Compo et al., 2011). Output from the models is openly available from the Earth System Grid Federation (ESGF; https://esgf-data.dkrz.de) for the GHMs (Gosling et al., 2017) and the CHMs (Krysanova et al., 2017).

We treated all GHMs and CHMs as being independent even though many share some model parameterisations (e.g. method for calculating potential evapotranspiration; Table S1). However, the GHMs and CHMs differ in a fundamental way, which is the spatial scale at which the models represent hydrological processes and, in turn, the water balance. All the GHMs applied here operate with a 0.5°x0.5° spatial resolution grid across the global land-surface. The CHMs employ various approaches; three CHMs run on a grid (mHM, VIC, WaterGAP3), while five operate by splitting the catchment into sub-catchments and hydrological response units (ECOMAG, HBV, HYPE, SWAT and SWIM), and one treats the entire catchment as a single entity (HYMOD). Detailed descriptions of the GHMs and CHMs employed in this study are provided by the respective references to each model cited in Table 1 and in Table S1, whilst brief summaries are provided by Krysanova and Hattermann (2017).

Table 1: Catchments, gauges, upstream area of gauge, and the GHMs and CHMs that comprised each ensemble (shaded).

	Catchment	Upper Amazon	Lena	Upper Mississippi	Upper Niger	Rhine	Tagus	Yangtze	Yellow
	Gauging station	Sao Paulo de Olivenca	Stolb	Alton	Koulikoro	Lobith	Almourol	Cuntan	Tangnaihai
	Upstream drainage area (km^2)	990781	2460000	444185	120000	160800	67490	866559	121000
	Lat (degree)	-3.45	72.37	38.89	12.87	51.84	39.47	29.62	35.5
	Long (degree)	-68.75	126.8	-90.18	-7.55	6.11	-8.37	106.6	100.15
	Start year	1973*	1971	1971	1971	1971	1973*	1971	1971
	End year	2001	2001	2001	2001	2001	2001	2001	2001
	Missing data (%)	3.9	0	0.8	0.06	0	3.7	0	0
	Analysed data (year)	26	31	29	29	29	27	30	30
CHM ensemble	**ECOMAG** (Motovilov, 1999)								
	HBV (Bergstrom and Forsman, 1973)								
	HYMOD (Boyle, 2001)								
	HYPE (Lindstrom et al., 2010)								
	mHM (Kumar et al., 2013)								
	SWAT (Arnold et al., 1993)								
	SWIM (Krysanova et al., 1998)								
	VIC (Liang et al., 1994)								
	WaterGAP3 ** (Verzano, 2009)								
	Number of CHM simulations	7	5	7	7	7	4	4	6
GHM ensemble	**CLM** (Oleson KW. et al., 2010)								
	DBH (Tang et al., 2007)								
	H08 (Hanasaki et al., 2008)								
	LPJmL (Bondeau et al., 2007)								
	MATSIRO (Pokhrel et al., 2011)								
	MPI-HM (Hagemann and Dümenil, 1997)								
	PCR-GLOBWB (Wada et al., 2014)								
	WaterGAP2 (D¨oll et al., 2003)								
	Number of GHM simulations	8	8	8	8	8	8	8	8

*Start year differs from 1971 due to missing observed data.

**The WaterGAP model can operate at the global-scale but the version employed here (WaterGAP3) operates on a 5 arcminute global grid, specifically so that it can be run at the catchment-scale. This is a much finer resolution than what the global-scale models used in the analysis operate at (0.5°x0.5°), so it is classed as a catchment-scale model in this analysis.

2.4 Performing MMC using GEP

GEP is a highly-adaptive machine learning algorithm that applies symbolic regression (SR) and searches for and learns optimal regression equations that can incorporate constants, mathematical operators and non-linear functions to map multiple input data series to an output series using a non-parametric, stochastic optimisation procedure (Ferreira, 2001, Ferreira, 2006, Beriro, 2015). For each problem of interest, the GEP algorithm 'evolves' a set of optimised model solutions— known as 'candidate solutions' and in the form of SR equations— from a random start through an iterative, evolutionary learning process that includes both stochastic and systematic computational adjustments. The algorithm and its mathematical basis is fully described in Ferreira (2001) and Ferreira (2006).

GEP solutions are encoded in one or more linear structures (often referred to as 'chromosomes', hereafter called 'components') of fixed length where each structural element encodes user-defined inputs, functions, numerical constants or arithmetic operators. These components are the building blocks of the SR equations that GEP evolves and are afterwards expressed as nonlinear entities of different sizes and shapes (i.e., simple diagram representations or expression trees). The components can be combined (usually by addition, see Figure S1) to deliver a final symbolic regression equation (sometimes referred to as a 'program') Ferreira (2001) and Ferreira (2006). There is a high degree of flexibility over whether / how often the inputs, functions, operators and constants are used in the SR equations that GEP generates. This gives GEP the unusual property of being able to preferentially 'deselect' inputs entirely from its solutions.

One important advantage of GEP over other machine learning algorithms is the way it structures and expresses its solutions to a modelling problem as components. That is, the components that form the MMC's symbolic regression equation are adjusted separately during the evolutionary process; enabling tailored and flexible adjustment of different components of the overall solution. Coupled with its ability to preferentially select / deselect inputs, GEP has the capacity to support the development of MMC solutions in which the best aspects of the best contributing

models are emphasised, and the worst aspects of the worst models are de-emphasised or excluded.

In this study, GEP was implemented using GeneXpro Tools 4.0 (GXPT4) (http://www.gene expression programmingsoft.com/). GXPT4 includes a number of default settings (such as genetic operators, Table S2) that we adapted following Ferreira (2001) and Ferreira (2006). The general setting (e.g. the number of components) and the user-defined function set (including maximum generations that GXPT4 evolves, mathematical functions (e.g. sine and log), operators (e.g. addition and multiplication) and the allowable constants (numerical values)) (Table S2) were chosen to maximise the parsimony of the symbolic regression equations that GEP delivered whilst supporting a limited set of non-linear functions. The rationale behind this was to maximise the potential for meaningful interpretation of the regression equations.

2.5. Selection of the final MMC solutions

The 'candidate solutions' generated by GEP vary in terms of their complexity, i.e. the number of components and mathematical functions, operators, model inputs and constants that comprise it; as well as the level of fit to observed data. Intuitively, it is often assumed that the more complex the solution, the greater the degree of fit between modelled and observed data. However, the more complex the solution the greater the chance that it is over-fitted and the harder it is to interpret the symbolic regression equations that define it in a meaningful manner. Thus, it is necessary to employ a procedure for selecting a final solution from the candidate set of solutions so that it has both a good degree of fit and is parsimonious with respect to its numerical complexity so that overfitting is prevented and interpretability is highest. Following Zaherpour et al., 2018b we here apply a simple trade-off between the goodness of fit of the MMC equation and the equation size to identify a final solution from a candidate set. That is, on a normalised fitness/equation size coordinate system of all solutions (Figure 2), the solution with the smallest Euclidean distance was selected as the final solution.

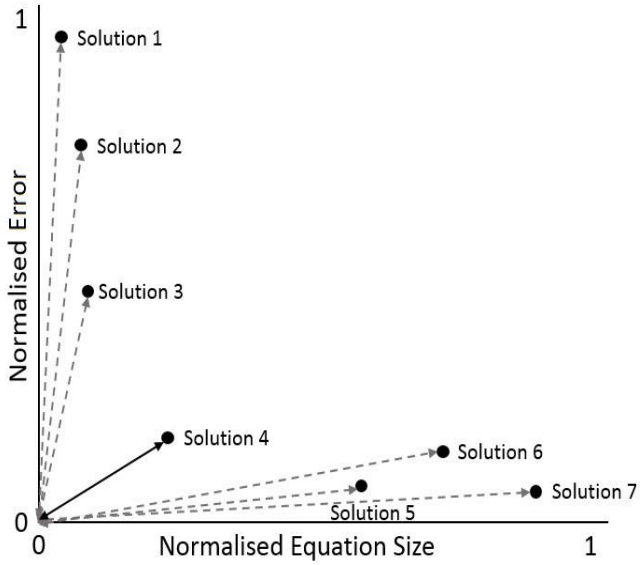


Figure 2. Selecting the gene expression programming solution from a normalised fitness-equation space. Solution 4 is selected because it has the smallest Euclidean distance from the origin.

2.6. Assessing model performance

2.6.1. The integrated performance metric

For the purposes of evaluating performance, we adopted an integrated performance metric called the Ideal Point Error (IPE) (Dawson et al., 2012). IPE was used to inform the learning process during the development of the GEP solutions, i.e. the minimisation of IPE is the objective function used for training the MMCs by GXPT4. It was also used to assess the relative ability of individual models, i.e. the GHMs and CHMs, as well as the EMs, to simulate observed runoff. Furthermore, IPE was used to evaluate the performance of the selected MMC solutions.

IPE applied here expresses the ratio of performance gain / loss of a model being evaluated compared to a model benchmark. In this study, we identified the benchmark as the simple model where runoff in time *t* is predicted by observed runoff in *t-1*, following Seibert (2001), Zaherpour et al. (2018a) and Zaherpour et al. (2018b). This is essentially using persistence as the benchmark. Our application of IPE integrates three metrics often used in hydrological model evaluations: the Coefficient of Efficiency (CE, also known as the Nash Sutcliffe Efficiency, NSE), mean

absolute relative error (MARE) and root mean squared error (RMSE) (Table S3). The benchmarked IPE formula presented in equation (1) is adapted from the original formula by Dawson et al. (2012). The negative reciprocal of the IPE score is used (equation 2) where the performance of a model exceeds that of the benchmark to maintain proportionality in comparisons between the IPE scores of models that fail to perform as well as the benchmark and those where performance exceeds it. The IPE scores range between -1 and -∞ (performance improvement over the benchmark model) and 1 and +∞ (performance loss over the benchmark model). The IPE score is ratiometric – for example, a model that performs twice as well as the benchmark model will have an IPE score of -2 and a model that performs twice as badly will have a score of +2. IPE would be 1 if a model performs the same as the benchmark, whilst a model infinitely better than the benchmark would have an IPE of -∞. The formula to calculate IPE reads:

$$\mathrm{IPE}_b = \left\{[1/3 * ((\mathrm{RMSE_m}/\mathrm{RMSE_b})^2 + (\mathrm{MARE_m}/\mathrm{MARE_b})^2 + ((\mathrm{CE_m} - 1)/(\mathrm{CE_b} - 1))^2)]^{\frac{1}{2}}\right\}$$

$$\mathrm{IPE} = \mathrm{IPE}_b \qquad \text{IF } \mathrm{IPE}_b \geq 1 \tag{1}$$

$$\mathrm{IPE} = -1/\mathrm{IPE}_b \qquad \text{IF } \mathrm{IPE}_b < 1 \tag{2}$$

Where:

RMSE = Root Mean Squared Error
MARE = Mean Absolute Relative Error
CE = Coefficient of Efficiency
m = modelled data
b = benchmark data from the simple model

2.6.2. Quantifying gains in performance from MMC

In order to assess the improvement or otherwise from developing MMC models, their performance is evaluated relative to the individual models from each ensemble, as well as the EMs for each model ensemble. The performance gain (PG) of one model output (e.g. model *A*, which might be the MMC for the GHM ensemble) relative to another (e.g. model *B*, which might be the EM for the GHM ensemble) can be expressed in percentage terms (equations 3-5). For example, a PG

of +50% indicates a gain in performance over the benchmark that is 50% larger for model *A* than model *B*. Similarly, a PG of -120% would indicate that there is a 1.2 times reduction in performance of model *A* relative to model *B*.

Where both model *A* and model *B* are either positive, or both negative:

$$\mathrm{MMC_{PG}} = (\mathrm{IPE_{modelB}} - \mathrm{IPE_{modelA}}) \times 100 \quad (3)$$

Where model *A* is negative and model *B* is positive:

$$\mathrm{MMC_{PG}} = \big((\mathrm{IPE_{modelB}} + 1) - (\mathrm{IPE_{modelA}} - 1)\big) \times 100 \quad (4)$$

Where model *A* is positive and model *B* is negative:

$$\mathrm{MMC_{PG}} = \big((\mathrm{IPE_{modelB}} - 1) - (\mathrm{IPE_{modelA}} + 1)\big) \times 100 \quad (5)$$

2.6.3. Data splitting for MMC development and testing

In order to implement the MMC technique, here we split time-series of data to three subsets of training, testing and validation datasets. Symbolic regression equations that comprise each MMC are developed from a training dataset during the GEP process. The regressions are refined iteratively by evaluating them against a testing dataset (Wu et al., 2014, Phukoetphim et al., 2016, Wu et al., 2012). As well as this training/testing data subset (effectively a calibration subset) an independent 'out-of-sample' validation data subset needs to be defined for assessing how well the final MMCs perform.

The DUPLEX method (Snee, 1977) was used for generating the three data subsets. We were consistent across all eight catchments in the size of the training/testing data subset, which comprised 144 months (12 years) and 96 months (8 years) for the training and testing data subsets respectively. The size of the out-of-sample validation data subset, which was used to assess the performance of the MMCs, varied from catchment-to-catchment according to the length of the observed data series that was available. However, it was never less than 72 months (6 years) and extended up to 132 months (11 years) with Lena (Table 2).

Table 2. The training/testing and out-of-sample validation subsets used for MMC development and validation respectively.

River	Total length (months)	Training/testing data subset		Out-of-sample validation data subset	
		Period	Length (months)	Period	Length (months)
Upper Amazon	312	1/1/1974-31/12/1993	240	1/1/1994-31/12/1999	72
Lena	372	1/1/1971-31/12/1990	240	1/1/1991-31/12/2001	132
Upper Mississippi	348	1/1/1972-31/12/1992	240	1/1/1993-31/12/2000	108
Upper Niger	348	1/1/1972-31/12/1991	240	1/1/1992-31/12/2000	108
Rhine	348	1/1/1972-31/12/1991	240	1/1/1992-31/12/2000	108
Tagus	324	1/1/1974-31/12/1993	240	1/1/1994-31/12/2000	84
Yangtze	360	1/1/1972-31/12/1991	240	1/1/1992-31/12/2001	120
Yellow	360	1/1/1972-31/12/1991	240	1/1/1992-31/12/2000	120

2.7. Hydrological indicators

To explore how the derived MMCs captured different aspects of hydrological performance in each catchment, we calculated four hydrological indicators from the monthly time-series of runoff:

1) Mean annual runoff (MAR): the mean total annual runoff across all years in the time-series.
2) Q5: the magnitude of monthly runoff that is exceeded 5 % of the time in the time-series, and thus an indicator of high flow.
3) Q95: the magnitude of monthly runoff that is exceeded 95 % of the time in the time-series, and thus an indicator of low flow.
4) Magnitude of long term mean monthly runoff (MMR).

3. Results and discussion

3.1. The performance gap between GHMs and CHMs

Our first objective was to assess whether MMCs of GHMs and CHMs perform better than their respective EMs and individual models, and to compare the performance of MMCs between the GHM and CHM ensembles. Therefore, first we calculated the performance of the EM as well as of the individual models that comprise each ensemble, to understand what might be driving any differences in EM performance between the GHMs and CHMs.

Table 3 shows the IPE metrics for the GHM and CHM ensembles and their EM, computed from monthly runoff time-series for the out-of-sample validation period (see Table S8 for IPE scores of individual GHMs and CHMs). There is a clear performance gap, where overall, the GHMs do not perform as well as the CHMs. Based on the mean IPE statistics across all CHMs and GHMs respectively, the CHMs perform better than the benchmark (simple model) for all model-catchment combinations, which is in stark contrast to the GHMs, where the majority of model-catchment combinations perform below the benchmark. The Yangtze is the only catchment where the mean IPE for the GHMs is better than the benchmark.

The performance gap is clearly evident from the EMs of the two types of model. The IPE scores are -2.48 and 0.75 for the CHM and GHM ensembles respectively, as a mean across all catchments (Table 3). The EM for the GHMs performs better than the benchmark in only three of the catchments (shaded grey in Table 3) whereas the EM for the CHMs performs better than the benchmark in all eight catchments. However, the EM for the CHMs rarely performs better than the best performing individual CHM (only for three catchments; shaded in Table 3).

The performance gap can in part be explained by model calibration: all the CHMs were calibrated whereas only one GHM was calibrated (WaterGAP2). Other comparative GHM-CHM evaluation studies (Piniewski et al., 2014, Hattermann et al., 2018) also attribute the performance gap to model calibration (or the lack of it). However, it should be noted that model calibration is not a guarantee that a model will perform better than the benchmark, as the *Max* statistics indicate that there are several model-catchment combinations where the CHMs do not outperform the benchmark (unshaded cells in Table 3). Simply concluding that all GHMs should be calibrated is not the solution to closing the performance gap. The practical way forwards is more nuanced; the calibration of GHMs is precluded by a lack of globally available and complete time-series of runoff data. For instance, WaterGAP2 was only calibrated for 1,319 catchments, covering 54% of the global land area except Antarctica and Greenland (Müller Schmied et al., 2016). Thus there is a need for improved datasets of observed runoff for comprehensively calibrating GHMs. Once this paucity of data is addressed, GHM modelling teams can then work towards the

routine calibration of their models, which may in turn help to reduce the performance gap.

Furthermore, the GHM-CHM differences are in part due to the representation of physical processes at different spatial resolutions (finer for CHMs). Whilst not the focus of this study, it will be interesting to see whether the performance gap reduces over the next decade in line with increases in the resolution of GHMs, because the two scales of model are beginning to merge in terms of spatial resolution and areas of applicability (Gosling et al., 2016), for example, versions of WaterGAP have been applied globally with WaterGAP2 (Müller Schmied et al., 2016) and regionally with WaterGAP3 (Eisner, 2016); DBH has been run both regionally and globally (Tang et al., 2007); and different versions of HYPE (Donnelly et al., 2015) have been run at catchment- and continental-scales.

Sampling bias, introduced by the different sizes of the two ensembles may also explain the performance gap. The relatively smaller number of CHMs may represent only a limited sample of the true hydrological model spread that would arise from using all the CHMs that are in existence (Gosling et al., 2016). Thus the spread of the CHM ensemble is likely underestimated (as may be the GHM ensemble).

Overall the results suggest that the EM may not be the most appropriate statistic to use when presenting results from an ensemble of individual models. The EM is often used in ensemble studies (Phillips and Gleckler, 2006, Gosling et al., 2010, Sanderson and Knutti, 2012, Hattermann et al., 2017) but our results show that doing so inherently overlooks the fact that some of the models that make up the ensemble perform better than the EM. It may, in some circumstances, be more appropriate to present the results of the best performing individual model and/or to compute a more robust statistic than the EM which accounts for the relatively better and poorer performing models in the ensemble. This can be achieved through MMC. Furthermore, the variable performance of the EM means that a decision should not be taken *a priori* to use the EM as the basis of a model evaluation exercise (Van Loon et al., 2012, Gudmundsson et al., 2012, Schewe et al., 2014, Beck et al., 2016), without considering its performance relative to the models it is summarising. We also acknowledge that the performance of EM can depend on the evaluation metric applied (Rougier, 2016).

Table 3. Statistics of the IPE scores for the GHM and CHM ensembles, and IPE sores for their EM and MMC Solutions (all computed for the out-of-sample validation period). Grey shading denotes where the models, the EM, or MMC solution outperforms the benchmark model. The best performing of all individual models (i.e. GHMs or CHMs), the EM and all MMC solutions, for each catchment is in bold. Red text denotes where an MMC solution does not perform better than the best individual model or EM, while green denotes where an MMC does.

GHMs							
Catchment	**IPE statistics over GHMs**				**EM**	**MMC**	**Super-ensemble MMC**
	Min	**Max**	**Mean**	**SD of Mean**			
Upper Amazon	-1.43	2.99	1.46	1.32	1.24	-1.78	**-1.95**
Lena	-1.67	1.92	0.40	1.47	-1.15	-2.14	**-3.36**
Upper Mississippi	-1.40	6.57	1.62	2.44	1.20	-1.42	**-2.60**
Upper Niger	-1.30	12.98	5.99	4.43	4.42	-1.66	**-2.06**
Rhine	-2.51	5.34	1.88	2.25	1.53	-2.53	**-3.66**
Tagus	-1.26	4.43	1.78	1.63	1.70	-1.68	**-2.13**
Yangtze	-1.73	1.44	-0.77	1.31	-1.79	-3.27	**-3.05**
Yellow	1.14	4.41	2.19	1.18	-1.18	-1.15	**-1.53**
Mean*	-0.88	3.52	1.82	1.45	0.75	-1.95	**-2.54**
CHMs							
Catchment	**IPE statistics over CHMs**				**EM**	**MMC**	**Super-ensemble MMC**
	Min	**Max**	**Mean**	**SD of Mean**			
Upper Amazon	**-1.99**	1.46	-1.06	1.15	-1.48	-1.98	-1.95
Lena	**-3.40**	1.29	-1.83	1.90	-1.99	-2.17	-3.36
Upper Mississippi	-2.18	-1.14	-1.60	0.33	-2.47	-1.73	**-2.60**
Upper Niger	**-2.22**	1.46	-0.86	1.52	-1.68	-1.58	-2.06
Rhine	-3.77	-2.15	-2.99	0.66	**-5.58**	-4.36	-3.66
Tagus	-1.50	1.11	-0.15	1.41	-1.32	-1.61	**-2.13**
Yangtze	-3.59	-2.31	-2.93	0.70	-3.86	**-3.86**	-3.05
Yellow	**-1.72**	1.11	-0.97	1.05	-1.44	-1.69	-1.53
Mean*	-2.27	0.06	-1.75	0.80	-2.48	-2.37	**-2.54**

*For the four first columns are statistics for mean IPE scores calculated across all catchments for individual models (see Table S8)

Table 4. Performance gain (PG) achieved by the EM and MMCs (both relative to the best individual model), and the PG achieved by MMC relative to the EM. Grey shading denotes performance gain. The best performing model combination, i.e. the EM or one of the MMCs, is shown in bold for each catchment.

GHMs

Catchment	IPE				PG over best individual GHM (%)			PG over GHM EM (%)	
	Best GHM	EM	MMC	Super-nsemble MMC	EM	MMC	Super-ensemble MMC	MMC	Super-ensemble MMC
Upper Amazon	-1.43	1.24	-1.78	-1.95	-466.9	5.4	**51.5**	402.3	**418.4**
Lena	-1.67	-1.15	-2.14	-3.36	-51.9	46.9	**168.8**	98.8	**220.6**
Upper Mississippi	-1.40	1.20	-1.42	-2.60	-459.2	2.4	**120.6**	361.7	**479.8**
Upper Niger	-1.30	4.42	-1.66	-2.06	-772.4	35.7	**75.9**	708.1	**748.2**
Rhine	-2.51	1.53	-2.53	-3.66	-603.2	2.1	**115.5**	505.3	**618.7**
Tagus	-1.26	1.70	-1.68	-2.13	-495.7	42.3	**86.8**	438.1	**482.5**
Yangtze	-1.73	-1.79	-3.27	-3.05	5.7	**154.2**	132.1	**148.5**	126.4
Yellow	1.14	-1.18	-1.15	-1.53	331.9	328.5	**367.0**	-3.4	**35.1**
Mean		0.75	-1.95	-2.54	-314.0	81.0	**139.8**	332.4	**391.2**

CHMs

Catchment	IPE				PG over best individual CHM (%)			PG over CHM EM (%)	
	Best CHM	EM	MMC	Super-ensemble MMC	EM	MMC	Super-ensemble MMC	MMC	Super-ensemble MMC
Upper Amazon	-1.99	-1.48	-1.98	-1.95	-51.3	**-0.8**	-4.5	**50.6**	46.9
Lena	-3.40	-1.99	-2.17	-3.36	-141.1	-123.2	**-4.2**	17.9	**136.8**
Upper Mississippi	-2.18	-2.47	-1.73	-2.60	29.3	-44.7	**42.1**	-74.0	**12.8**
Upper Niger	-2.22	-1.68	-1.58	-2.06	-54.2	-63.6	**-15.9**	-9.4	**38.3**
Rhine	-3.70	-5.58	-4.36	-3.66	**187.5**	66.0	-3.8	**-121.6**	-191.4
Tagus	-1.50	-1.32	-1.61	-2.13	-18.4	10.7	**62.7**	29.1	**81.1**
Yangtze	-3.59	-3.86	-3.86	-3.05	**28.2**	26.8	-54.0	**-1.4**	-82.2
Yellow	-1.72	-1.44	-1.69	-1.53	-27.6	**-2.6**	-19.0	**25.0**	8.6
Mean		-2.48	-2.37	-2.54	-5.9	-16.4	**0.4**	-10.5	**6.4**

3.2. Comparison of MMC performance to the EM, individual models and between ensembles

Table S4 and Table S5 show the MMC equations for the GHM and CHM ensembles respectively. The MMCs derived from the GHMs outperform the best individual GHM as well as the EM in all catchments except for the Yellow (Table 3). This contrasts with the CHM ensemble, where in six catchments, the MMC for the CHMs fails to outperform either the best individual CHM or the ensemble mean. Generally, better (smaller) IPE scores are delivered by the individual CHMs, the EM of the CHMs, and the MMCs of the CHMs, than their global counterparts; the Upper Niger and Tagus are exceptions.

When considering the aggregated results across all catchments, while the MMCs derived from the GHMs deliver large performance gains relative to both the best individual GHM (81%) and the GHM EM (332%), the MMCs derived from the CHMs yield *lower* performance relative to both the best individual CHM (-16.4%) and the CHM EM (-10.5%) (Table 4). Thus, according to our results, the improvement in performance that can be delivered by MMC varies according to the performance of the individual models that comprise the ensemble. In our case, MMC offers limited ability to improve the performance of an ensemble that performs well relative to a benchmark (here the CHMs, which have a mean IPE of -2.48 for the EM; Table 3).

The inability of the MMCs derived from the CHMs to deliver improvements in performance over the best individual CHM and the EM could in part stem from the calibration of the CHMs. An ensemble of calibrated models, which typically result in a tight(er) range of simulations that are all close to observed runoff with little inter-model spread (Gosling et al., 2016), provides less opportunity for the GEP algorithm to learn from the simulations than if the models are uncalibrated, because all individual models have been tuned to closely match observed data. They will therefore all perform in a relatively similar manner, perhaps all modelling certain aspects of hydrology well, but others poorly. Thus we see value in including both calibrated and uncalibrated models as inputs to MMC solutions because the uncalibrated models may inherently replicate certain aspects of the hydrological regime that the calibrated models do not.

The finding that MMC can show lower performance relative to both the EM and the best individual model (as it did with the CHMs) may also in part be due to the excessive structural complexity of the MMC equations (Tables S4 and Table S5), which comprises 3 components. The evolutionary algorithm was constrained to deriving a MMC solution that had three components and no less. If left unconstrained, the algorithm may have converged on an MMC solution that comprised fewer components. To explore this, we repeated the evolutionary MMC process for the GHM and CHM ensembles except this time constraining the MMC to 2 components instead of 3, thereby constraining the GEP process to deliver a less complex solution.

The 2-component MMC for the CHM ensemble resulted in modest performance gains over the original 3-component MMC (Table S8). The IPE scores improved in 5 catchments relative to the 3-component MMC. Overall, as a mean across all catchments for the CHM ensemble, the 2-component MMC resulted in performance gains over the best individual CHM of 16.4% and the EM of 22.4% (Table S9) compared to respective performance losses of -16.4% and -10.5% achieved with the 3-component MMC (Table 4). However, reducing the complexity did not yield a performance gain relative to the EM in every catchment: the Niger, Rhine and Yellow are exceptions. Residual plots of observed versus simulated runoff for these catchments (Figure S3) reveal a lack of structural diversity (Upper Niger) or similar residual structures (Rhine and Yellow) for the CHMs, which may explain why MMC was unable to deliver runoff estimates that perform better than the EM in these three catchments.

Whilst constraining the MMC to a higher level of complexity (3 components) generally resulted in better MMC performance for the GHM ensemble relative to the less complex MMC, it generally lowered performance for the CHM ensemble (Table S8). This confirms that in some cases (here, for the CHM ensemble), constraining the evolutionary GEP algorithm to converge on an MMC solution with too many components (i.e. a higher complexity) can actually deteriorate the performance of the MMC when compared to constraining it to fewer components (i.e. lower complexity) and/or letting a solution converge from an unconstrained

evolutionary algorithm. An important conclusion, therefore, is that whilst intuitive, an *a priori* assumption that a more complex MMC will deliver a better performing model, is invalid. We recommend that in future applications MMCs are either derived from an unconstrained evolutionary process, or a sensitivity analysis is conducted that involves different constraints, such as we demonstrated here.

3.3. Performing MMC with the super-ensemble

The MMCs derived from the super-ensemble generally performed better than the respective EMs from the GHMs and CHMs. The mean IPE across all catchments for the super-ensemble MMC is -2.54, which contrasts with IPEs for the GHM EM of 0.75 and -2.48 for the CHM EM (Table 3). In terms of performance gain, this represents a significant improvement relative to the best individual GHM and GHM EM: the mean improvement (performance gain) is 139.8% and 391.2% respectively (Table 4). The performance gains relative to the best individual CHM and CHM EM are more modest: 0.4% and 6.4% respectively. Nevertheless, whilst small, these *are* performance gains and it is worth recalling that deriving MMCs from only the CHM ensemble yielded *reductions* in performance relative to the CHM EM and best individual CHM.

At the aggregate level across multiple catchments the inclusion of more input models (i.e. GHMs and CHMs) to the MMC improves the performance of the MMC relative to including fewer input models. This is demonstrated by the lower (i.e. better) IPE values observed with the super-ensemble (mean IPE of -2.54 across all catchments) compared with the IPE scores for the MMCs of GHM and CHM ensembles (-1.95 and -2.37 respectively).

The quantification of model performance is to an extent dependent upon the metric used to calculate performance as well as the hydrological indicators assessed (Zaherpour et al., 2018a). Here we used IPE with the time-series of monthly runoff. Therefore, we also calculated the Percentage Bias (PBIAS; Table 5) for MAR, Q5 and Q95, to explore the robustness of our finding that including more input models to the MMC through a super-ensemble generally improves MMC performance. As a

mean across all catchments, the super-ensemble MMC delivers a PBIAS that is lower (i.e. better) than the GHM MMCs and CHM MMCs for hydrological extremes, i.e. Q95 and Q5 (Table 5). However, for MAR, the GHM MMC performs better than the super-ensemble MMC. Therefore, the extent to which the inclusion of more models to the MMC development process improves performance is, in our case, dependent upon the hydrological indicator of interest. The results also vary by catchment.

We also explored how the super-ensemble performed compared to the GHM and CHM MMCs, by calculating the relative differences between the simulated and observed seasonal cycle (MMR) for the out-of-sample validation period (Figure 3; absolute monthly runoff is displayed in Figure S2). The spread in relative difference for the GHM ensemble is noticeably wider than that of the CHM ensemble, and in general, CHMs reproduce the observed seasonal cycle well, with narrow ranges of uncertainty. This is partly the result of targeting the minimisation of volumetric error as a calibration objective for the CHMs (Hattermann et al., 2017). Despite the wider spread for the GHM ensemble, its MMC compares favourably to that of the CHM EM in the majority of the catchments, especially the Rhine, Tagus and Yellow. Across the majority of the catchments the super-ensemble delivers smaller errors than the EMs and MMCs from the GHM and CHM ensembles. This is particularly evident for the Lena, Niger, Rhine, Tagus and Yellow. The super-ensemble generally has the effect of reducing the over-estimation of runoff seen in the EMs and MMCs from the GHMs and CHMs. The super-ensemble MMC does not necessarily deliver the best performance in every month of the year, however. In part, this is the result of calibrating the MMCs on annual monthly time-series of data – better performance might be possible by developing MMCs for each month in turn, but this would require a significant increase in computation resource.

Table 5. PBIAS (%) calculated for the out-of-sample validation period for MAR, Q5 and Q95, for individual catchments and as a mean across all catchments. The EMs and MMCs are ranked according to their mean PBIAS.

Indicator	Catchment	GHMs EM	GHMs MMC	CHMs EM	CHMs MMC	Super-ensemble EM	Super-ensemble MMC
MAR	Upper Amazon	-22.6	-0.4	-3.5	1.4	-13.7	1.8
	Lena	-13.3	12.8	0.9	14.3	-7.8	-5.5
	Upper Mississippi	36.1	-9.9	-3.3	11.1	17.7	-0.9
	Upper Niger	154.8	15.7	29.4	25.8	96.3	11.9
	Rhine	37.9	2.4	-1.9	2.2	18.0	-1.7
	Tagus	68.6	-18.6	13.3	5.2	50.1	5.4
	Yangtze	-11.8	-3.2	0.3	5.7	-7.8	9.3
	Yellow	21.2	7.6	18	12.2	19.8	18.9
	Mean	33.9	0.8	6.7	9.7	21.6	4.9
	Rank	6	1	3	4	5	2
Q5	Upper Amazon	-20.1	0.5	-2.7	2.2	-12.8	2.7
	Lena	-17.5	14.3	-3.7	11.5	-12.2	-6.6
	Upper Mississippi	17.8	-22.0	-9.9	-1.6	4.9	-6.7
	Upper Niger	133.7	15.0	25.6	27.7	83.3	16.3
	Rhine	33.8	0.4	-1.4	3.4	16.2	-0.4
	Tagus	43.7	-26.9	4.5	3.9	30.6	5.6
	Yangtze	-14.3	-5.5	0.8	1.1	-9.3	4.0
	Yellow	-1.9	-4.8	-1	-3.9	-1.5	0.2
	Mean	21.9	-3.6	1.5	5.5	12.4	1.9
	Rank	6	3	1	4	5	2
Q95	Upper Amazon	-22.5	1.2	-2.7	2.1	-13.3	2.3
	Lena	-14.5	1.9	-1.5	9.6	-9.5	-10.5
	Upper Mississippi	47.7	-10.0	-9.5	13.0	21.0	-7.7
	Upper Niger	206.6	34.8	36.8	25.0	127.3	12.3
	Rhine	39.7	2.5	-3.4	-0.6	18.1	-7.1
	Tagus	127.2	-41.2	42.4	-11.9	98.9	-24.7
	Yangtze	-11.0	0.5	3.1	8.9	-6.3	13.2
	Yellow	23.6	-18.9	25	21.2	24.2	27.8
	Mean	49.6	-3.6	11.3	8.4	32.6	0.7
	Rank	6	2	4	3	5	1

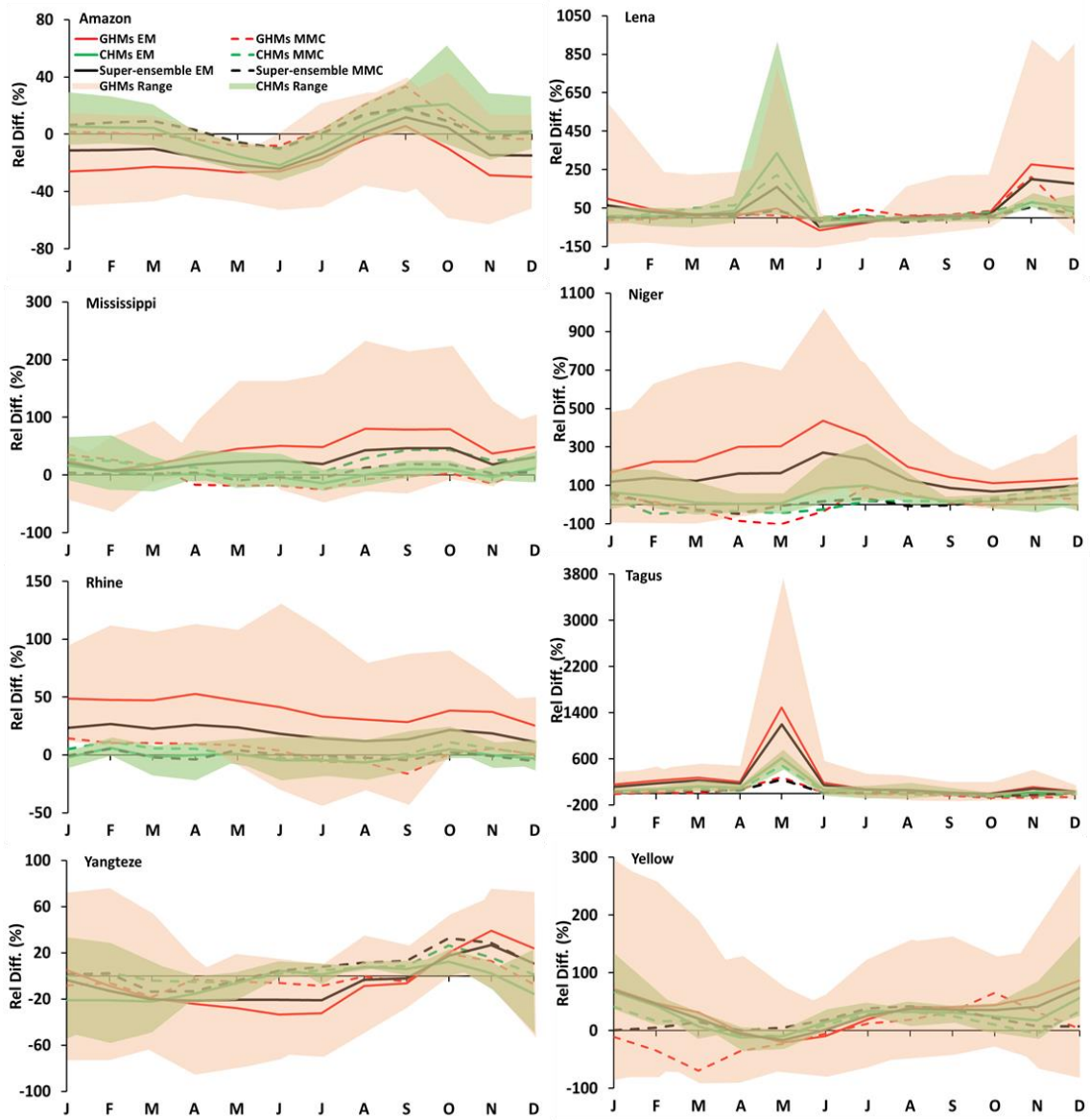


Figure 3. Relative differences (%) between the simulated and observed seasonal cycles (MMR) for the out-of-sample validation period.

Our results broadly indicate that the inclusion of more input models (i.e. GHMs and CHMs) to the MMC improves the performance of the MMC relative to including fewer input models, albeit with some subtle variations that have been noted. The main reason for this is that there is enough diversity, i.e. good-fitting simulations of different magnitudes of runoff amongst different models (GHMs and CHMs; demonstrated by the residual plots of observed versus simulated runoff in Figure S3 and Figure S4) that are input to the GEP process. This in turn means that the GEP process can learn and exploit the additional information to improve the overall

performance of the MMC relative to when fewer models are included. This is especially apparent for extreme runoff (i.e. Q5 and Q95). GHMs and CHMs find it harder to simulate extremes of runoff than MAR, with ability varying across catchments and by model (Zaherpour et al., 2018a). This creates a complex matrix of models' ability to represent extremes, e.g. one model might be the best for Q95 in one catchment but then perform weaker in a different catchment; and that model might not perform so well for Q5. Therefore, including more models in the GEP process delivers greater improvement for Q5 and Q95 than it does for MAR because the GEP algorithm is able to draw upon a more diverse set of simulations.

Our observation that MMCs perform better when more models are used as input has important implications for future studies, for which we recommend that: 1) the EM is not used as the main (or only) summary of ensemble performance, when presenting the results from an ensemble of GHMs and/or CHMs, as an individual model or a MMC may perform better; and 2) studies seek to include as many individual models in their assessment as is feasible, with a means to weighting those models to deliver a MMC solution.

However, we acknowledge that a pragmatic decision should be made on whether to derive MMCs from an ensemble. Deriving MMCs requires a computational and time resource that is significantly greater than selecting the best individual model from an ensemble. Our results show only a small improvement in performance from using the super-ensemble MMC relative to using the best individual CHM (performance gain of 0.4%). It may therefore in some cases be more efficient to simply select the best individual model as one indicator of ensemble performance, as opposed to deriving an MMC. Nevertheless, in such cases we would still recommend including the range of projections from the ensemble so that the (known) uncertainty range is displayed.

3.4. Harnessing value from GHMs and CHMs of variable performance

GEP preferentially selects or deselects individual models based upon their performance for inclusion in the symbolic regression equations that comprise the MMC. Table 6 shows the relative contributions of each GHM and CHM respectively

to the MMCs developed for each catchment. The MATSIRO GHM contributes to all MMCs, even though it is not the best performing individual GHM (Table S8). WaterGAP2 is the most frequent contributor to the first of the three components that make up each symbolic regression equation because the model's superior performance relative to the other GHMs makes it a good candidate as the basis of developing a symbolic regression equation. However, the inclusion (or not) of a model in a MMC should not be wholly interpreted in terms of its 'stand-alone' performance. Indeed, WaterGAP2, which is the best performing GHM in terms of IPE (Table S8), does not appear in every MMC. Models with relatively weak performance may be included in the MMC because their runoff simulations have specific structures that can be exploited to deliver numerical adjustments to consistent weaknesses in better performing models. CLM, H08 and PCR-GLOBWB appear to be regularly used in this way because in some catchments they contribute only to the second or third components of the symbolic regression equations that comprise each MMC. The infrequent contribution of H08 GHM to the MMCs and its lack of contribution to the first component across the MMCs may reflect general weaknesses in H08's stand-alone performance relative to the other models and/or indicate patterns of fit that are difficult to exploit as adjustment mechanisms for the other GHMs. Nevertheless, our results show that generally the MMC approach is able to harness value from models that overall perform weakly, including uncalibrated models. The CHM MMCs support this, as the weakest performing CHM, in terms of IPE, WaterGAP3, appears in the MMCs for around 50% of the catchments. Likewise, the MMCs from the super-ensemble (Table S10) use both GHMs and CHMs, even though overall the CHM ensemble performs better than the GHM ensemble; overall, 10 GHMs and 16 CHMs contribute to the MMCs derived from the super-ensemble across all eight catchments.

An alternative, less objective approach than MMC, where models are not computationally and intelligently selected/deselected might result in the abandonment of some models from the ensemble based upon their evaluation metrics. This might typically occur following human decision-making following an assessment of model evaluation metrics. MMC on the other hand is able to harness

value from such models. This is noteworthy because considerable resources have been invested in hydrological models over recent decades. An approach such as MMC that is able to identify strengths and value from models of variable performance is therefore of significance to, and encouraging for, the hydrological modelling community as a whole.

Table 6. Contribution of the individual GHMs and CHMs to the MMCs derived for each catchment. Crosses in underlined bold (and totals in brackets) indicate the inclusion of the GHM in the first component of the symbolic regression equation.

	GHMs									
Catchment	**CLM**	**DBH**	**H08**	**LPJmL**	**MATSIRO**	**MPI-HM**	**PCR-GLOBWB**	**WaterGAP2**	**No. of models included**	
Upper Amazon	x	x			x			**x**	4(1)	
Lena		x		**x**	**x**		x		4(2)	
Upper Mississippi		**x**		x	x	x		**x**	5(2)	
Upper Niger				x	x		x	**x**	4(1)	
Rhine	x			x	x	x		**x**	5(1)	
Tagus				x	**x**				2(1)	
Yangtze	x			**x**	x	**x**			4(2)	
Yellow			x		**x**				2(2)	
Total	3(0)	3(1)	1(0)	6(2)	8(3)	3(1)	2(0)	4(4)		
	CHMs									
Catchment	**ECOMAG**	**HBV**	**HYMOD**	**HYPE**	**mHM**	**SWAT**	**SWIM**	**VIC**	**WATERGAP3**	**No. of models included**
Upper Amazon		**x**	**x**		x		**x**	**x**		5(4)
Lena				**x**			x		x	3(1)
Upper Mississippi					**x**	x		x	**x**	4(1)
Upper Niger							**x**	**x**	x	3(2)
Rhine		**x**			**x**	x			x	4(2)
Tagus		**x**		x				x	x	4(1)
Yangtze		x					x	**x**		3(1)
Yellow		x			**x**		x			3(1)
Total	0(0)	5(3)	1(1)	2(1)	4(3)	2(0)	5(2)	5(3)	5(1)	

4. Summary

Our results lead us to two unconventional conclusions that contrast significantly with predominant thinking on how to deal with: 1) poorly performing hydrological models in an ensemble; and 2) uncalibrated models. As regards the former, it is often suggested, and intuitively assumed (Krysanova et al., 2018), that poorly performing models should be excluded from an ensemble because they lower the performance of the ensemble mean and because they are unreliable. However, here we have demonstrated value in such models because they can be used to adjust the simulations from other models that are input to the MMC development process. We recommend that future applications of MMC include as many hydrological models as input, as is feasible, since our results show that generally, when aggregated across multiple catchments, MMCs developed from a super-ensemble perform better than individual models and the ensemble means of only GHMs and CHMs.

Our other main conclusion is that we see significant value in simulations from both calibrated *and* uncalibrated models. GHMs are often criticised for not being calibrated – a fact that is in part due to the paucity of complete and globally distributed observed runoff time-series. Here we showed that deriving a MMC from a super-ensemble that includes both calibrated (CHMs) and predominantly uncalibrated models (GHMs) performs better on average than the ensemble means of the calibrated and uncalibrated models. This is because the uncalibrated models may inherently replicate certain aspects of the hydrological regime that the calibrated models do not.

Whilst we are able to draw the above two conclusions from our analyses we note that there are a number of subtleties that underlie these general conclusions that are applicable at the aggregated level across multiple catchments. The improvements in performance delivered by MMC can vary at the individual catchment level, in different months of the year, and also by hydrological indicator. Furthermore, the constrained complexity of the MMCs can affect their performance and research into assessing the optimal balance between complexity and parsimony of the symbolic regression equations delivered by GEP is encouraged.

We encourage the hydrological community to embrace the multitude of GHMs and CHMs that exist in the two modelling communities and to collaborate more on cross-scale super-ensemble studies. This will lead to enhanced understanding of the hydrological cycle, the strengths and limitations of the current state-of-the-art models used in hydrological modelling, and in turn better informed simulations of hydrological change.

Acknowledgements

This work has been conducted under the framework of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP), so our thanks go to the modellers who submitted results to this project. To obtain access to data, please refer to the "data archive" section of the project portal www.isimip.org. We also thank the Global Runoff Data Centre (GRDC) for making available the observed runoff data. JZ was supported by the Islamic Development Bank and a 2018 University of Nottingham Faculty of Social Sciences Research Outputs Award.

References

AJAMI, N. K., DUAN, Q. Y., GAO, X. G. & SOROOSHIAN, S. 2006. Multimodel combination techniques for analysis of hydrological simulations: Application to Distributed Model Intercomparison Project results. *Journal of Hydrometeorology,* 7**,** 755-768.

ARNOLD, J. G., ALLEN, P. M. & BERNHARDT, G. 1993. A comprehensive surface-groundwater flow model. *Journal of Hydrology,* 142**,** 47-69.

ARSENAULT, R., GATIEN, P., RENAUD, B., BRISSETTE, F. & MARTEL, J. L. 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. *Journal of Hydrology,* 529**,** 754-767.

AYTEK, A., ASCE, M. & ALP, M. 2008. An application of artificial intelligence for rainfall-runoff modeling. *Journal of Earth System Science,* 117**,** 145-155.

BARBULESCU, A. & BAUTU, E. 2010. Mathematical models of climate evolution in Dobrudja. *Theoretical and Applied Climatology,,* 100.

BĂRBULESCU, A. & BĂUTU, E. 2009. Time Series Modeling Using an Adaptive Gene Expression Programming Algorithm. *International journal of mathematical models and methods in applied sciences,* 3**,** 85-93.

BATES, J. M. & GRANGER, C. W. J. 1969. The comination of forecasts. *Operational research,* 20**,** 451-468.

BECK, H. E., DE ROO, A. & VAN DIJK, A. I. J. M. 2015. Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments. *Journal of Hydrometeorology,* 16**,** 1478-1501.

BECK, H. E., VAN DIJK, A. I. J. M., DE ROO, A., DUTRA, E., FINK, G., ORTH, R. & SCHELLEKENS, J. 2016. Global evaluation of runoff from ten state-of-the-art hydrological models. *Hydrology and Earth System Sciences Discussions,* 21**,** 2881-2903.

BERGSTROM, S. & FORSMAN, A. 1973. Development of a conceptual deterministic rainfall-runoff model. *Hydrology Research,* 4**,** 147-170.

BERIRO, D. 2015. *Gene Expression Programming Models of Bioaccessible Benzo[a]pyrene in Coking Works Soils.* PhD, Nottingham.

BERIRO, D. J., ABRAHART, R. J., NATHANAIL, C. P., MORENO, J. & BAWAZIR, A. S. 2013. A typology of different development and testing options for symbolic regression modelling of measured and calculated datasets. *Environmental Modelling & Software,* 47**,** 29-41.

BONDEAU, A., SMITH, P. C., ZAEHLE, S., SCHAPHOFF, S., LUCHT, W., CRAMER, W., GERTEN, D., LOTZE-CAMPEN, H., MÜLLER, C., REICHSTEIN, M. & SMITH, B. 2007. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biology,* 13**,** 679-706.

BOSSHARD, T., CARAMBIA, M., GOERGEN, K., KOTLARSKI, S., KRAHE, P., ZAPPA, M. & SCHÄR, C. 2013. Quantifying uncertainty sources in an ensemble of hydrological climate-impact projections. *Water Resources Research,* 49**,** 1523-1536.

BOYLE, D. P. 2001. *Multicriteria calibration of hydrological models. PhD Dissertation.* PhD, University of Arizona, Tucson.

BREUER, L., HUISMAN, J. A., WILLEMS, P., BORMANN, H., BRONSTERT, A., CROKE, B. F. W., FREDE, H. G., GRÄFF, T., HUBRECHTS, L., JAKEMAN, A. J., KITE, G., LANINI, J., LEAVESLEY, G., LETTENMAIER, D. P., LINDSTRÖM, G., SEIBERT, J., SIVAPALAN, M. & VINEY, N. R. 2009. Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use. *Advances in Water Resources,* 32**,** 129-146.

CHEN, X., YANG, T., WANG, X., XU, C.-Y. & YU, Z. 2012. Uncertainty Intercomparison of Different Hydrological Models in Simulating Extreme Flows. *Water Resources Management,* 27**,** 1393-1409.

D¨OLL, P., KASPAR, F. & LEHNER, B. 2003. A global hydrological model for deriving water availability indicators: model tuning and validation. *Journal of Hydrology,* 270**,** 105–134.

DAWSON, C. W., MOUNT, N. J., ABRAHART, R. J. & SHAMSELDIN, A. Y. 2012. Ideal point error for model assessment in data-driven river flow forecasting. *Hydrology and Earth System Sciences,* 16**,** 3049-3060.

DONNELLY, C., ANDERSSON, J. C. M. & ARHEIMER, B. 2015. Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. *Hydrological Sciences Journal,* 61**,** 255-273.

EISNER, S. 2016. *Comprehensive evaluation of the WaterGAP3 model across climatic, physiographic, and anthropogenic gradients. PhD thesis.* PhD thesis. University of Kassel, Kassel, Germany.

FERNANDO, A. K., SHAMSELDIN, A. Y. & ABRAHART, R. J. 2012. Use of Gene Expression Programming for Multimodel Combination of Rainfall-Runoff Models. *Journal of Hydrologic Engineering,* 17**,** 975-985.

FERREIRA, C. 2001. Gene Expression Programming: A New AdaptiveAlgorithm for Solving Problems. 13**,** 87-129.

FERREIRA, C. 2006. *Gene Expression Programming: Mathematical Model-ing by an Artificial Intelligence,* Verlag: Berlin, Springer.

GAO, X. & DIRMEYER, P. A. 2006. A multimodel analysis, validation, and transferability study of global soil wetness products. *Journal of Hydrometeorology,* 7**,** 1218-1236.

GOSLING, S., MÜLLER SCHMIED, H., BETTS, R., CHANG, J., CIAIS, P., DANKERS, R., DÖLL, P., EISNER, S., FLÖRKE, M., GERTEN, D., GRILLAKIS, M., HANASAKI, N., HAGEMANN, S., HUANG, M., HUANG, Z., JEREZ, S., KIM, H., KOUTROULIS, A., LENG, G., LIU, X., MASAKI, Y., MONTAVEZ, P., MORFOPOULOS, C., OKI, T., PAPADIMITRIOU, L., POKHREL, Y., PORTMANN, F. T., ORTH, R., OSTBERG, S.,

SATOH, Y., SENEVIRATNE, S., SOMMER, P., STACKE, T., TANG, Q., TSANIS, I., WADA, Y., ZHOU, T., BÜCHNER, M., SCHEWE, J. & ZHAO, F. 2017. *ISIMIP2a Simulation Data from Water (global) Sector. GFZ Data Services.* [Online]. Potsdam, Germany: PIK. [Accessed].

GOSLING, S. N. & ARNELL, N. W. 2011. Simulating current global river runoff with a global hydrological model: model revisions, validation, and sensitivity analysis. *Hydrological Processes,* 25**,** 1129-1145.

GOSLING, S. N., BRETHERTON, D., HAINES, K. & ARNELL, N. W. 2010. Global hydrology modelling and uncertainty: running multiple ensembles with a campus grid. *Philos Trans A Math Phys Eng Sci,* 368**,** 4005-21.

GOSLING, S. N., TAYLOR, R. G., ARNELL, N. W. & TODD, M. C. 2011. A comparative analysis of projected impacts of climate change on river runoff from global and catchment-scale hydrological models. *Hydrology and Earth System Sciences,* 15**,** 279-294.

GOSLING, S. N., ZAHERPOUR, J., MOUNT, N. J., HATTERMANN, F. F., DANKERS, R., ARHEIMER, B., BREUER, L., DING, J., HADDELAND, I., KUMAR, R., KUNDU, D., LIU, J., VAN GRIENSVEN, A., VELDKAMP, T. I. E., VETTER, T., WANG, X. & ZHANG, X. 2016. A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C. *Climatic Change,* 141**,** 577-595.

GUDMUNDSSON, L., TALLAKSEN, L. M., STAHL, K., CLARK, D. B., DUMONT, E., HAGEMANN, S., BERTRAND, N., GERTEN, D., HEINKE, J., HANASAKI, N., VOSS, F. & KOIRALA, S. 2012. Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe. *Journal of Hydrometeorology,* 13**,** 604-620.

GUO, Z. C., DIRMEYER, P. A., GAO, X. & ZHAO, M. 2007. Improving the quality of simulated soil moisture with a multi-model ensemble approach. *Quarterly Journal of the Royal Meteorological Society,* 133**,** 731-747.

HAGEMANN, S. & DÜMENIL, L. 1997. A parametrization of the lateral waterflow for the global scale. *Climate Dynamics,* 14**,** 17-31.

HANASAKI, N., KANAE, S., OKI, T., MASUDA, K., MOTOYA, K., SHIRAKAWA, N., SHEN, Y. & TANAKA, K. 2008. An integrated model for the assessment of global water resources – Part 1: Model description and input meteorological forcing. *Hydrol. Earth Syst. Sci.,* 12**,** 1007-1025.

HATTERMANN, F. F., KRYSANOVA, V., GOSLING, S. N., DANKERS, R., DAGGUPATI, P., DONNELLY, C., FLÖRKE, M., HUANG, S., MOTOVILOV, Y., BUDA, S., YANG, T., MÜLLER, C., LENG, G., TANG, Q., PORTMANN, F. T., HAGEMANN, S., GERTEN, D., WADA, Y., MASAKI, Y., ALEMAYEHU, T., SATOH, Y. & SAMANIEGO, L. 2017. Cross - scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins. *Climatic Change,* 141**,** 561-576.

HATTERMANN, F. F., VETTER, T., BREUER, L., SU, B. D., DAGGUPATI, P., DONNELLY, C., FEKETE, B., FLORKE, F., GOSLING, S. N., HOFFMANN, P., LIERSCH, S., MASAKI, Y., MOTOVILOV, Y., MULLER, C., SAMANIEGO, L., STACKE, T., WADA, Y., YANG, T. & KRYSNAOVA, V. 2018. Sources of uncertainty in hydrological climate impact assessment: a cross-scale study. *Environmental Research Letters,* 13.

HUANG, S., KUMAR, R., FLÖRKE, M., YANG, T., HUNDECHA, Y., KRAFT, P., GAO, C., GELFAN, A., LIERSCH, S., LOBANOVA, A., STRAUCH, M., VAN OGTROP, F., REINHARDT, J., HABERLANDT, U. & KRYSANOVA, V. 2016. Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide. *Climatic Change,* 141**,** 381-397.

KRYSANOVA, V., DONNELLY, C., GELFAN, A., GERTEN, D., ARHEIMER, B., HATTERMANN, F. & KUNDZEWICZ, Z. W. 2018. How the performance of hydrological models relates to credibility of projections under climate change. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques,* 63**,** 696-720.

KRYSANOVA, V., HATTERMANN, F., AICH, V., ALEMAYEHU, T., ARHEIMER, B., AYZEL, G., BREUER, L., CHAMORRO-CHAVEZ, A., DAGGUPATI, P., DONNELLY, C., EISNER, S., FLÖRKE, M., GAO, C., GELFAN, A., GUSEV, Y., GUSTAFSSON, D., HABERLANDT, U., HUANG, J., HUANG, S., HUNDECHA, Y., JIANG, T., KALUGIN, A., KOCH, H., KOVALEV, E. E., KRAFT, P., KRYLENKO, I., KUMAR, R., KUNDU, D., LIERSCH, S., LIU, S., LOBANOVA, A., LAVRENOV, A., MISHRA, V., MO, X., MOTOVILOV, Y., NASONOVA, O., VAN OGTROP, F., PECHLIVANIDIS, I., PINIEWSKI, M., PLÖTNER, S., REINHARDT, J., SAMANIEGO, L., SEIDOU, O., SHAH, H., SHI, P., STRAUCH, M., SU, B., TECKLENBURG, J., TEKLESADIK, A., VAN GRIENSVEN, A., VERVOORT, W., VETTER, T., WALLNER, M., WANG, X., WORTMANN, M., YANG, T., ZENG, X., BÜCHNER, M., SCHEWE, J. & ZHAO, F. 2017. ISIMIP2a Simulation Data from Water (regional) Sector. GFZ Data Services. .

KRYSANOVA, V. & HATTERMANN, F. F. 2017. Intercomparison of climate change impacts in 12 large river basins: overview of methods and summary of results. *Climatic Change,* 141**,** 363-379.

KRYSANOVA, V., MÜLLER-WOHLFEIL, D.-I. & BECKER, A. 1998. Development and test of a spatially distributed hydrological/water quality model for mesoscale watersheds. *Ecological Modelling,* 106**,** 261-289.

KUMAR, R., SAMANIEGO, L. & ATTINGER, S. 2013. Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research,* 49**,** 360-379.

LI, W., SANKARASUBRAMANIAN, A., RANJITHAN, R. S. & SINHA, T. 2015. Role of multimodel combination and data assimilation in improving streamflow prediction over multiple time scales. *Stochastic Environmental Research and Risk Assessment,* 30**,** 2255-2269.

LIANG, X., LETTENMAIER, D. P., WOOD, E. F. & BURGES, S. J. 1994. A simple hydrologically based model of land surface water and energy fluxes for

general circulation models. *Journal of Geophysical Research: Atmospheres,* 99**,** 14415-14428.

LINDSTROM, G., PERS, C., ROSBERG, J., STROMQVIST, J. & ARHEIMER, B. 2010. Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. *Hydrology Research,* 41**,** 295-319.

LIU, J., YANG, H., GOSLING, S. N., KUMMU, M., FLÖRKE, M., PFISTER, S., HANASAKI, N., WADA, Y., ZHANG, X., ZHENG, C., ALCAMO, J. & OKI, T. 2017. Water scarcity assessments in the past, present, and future. *Earth's Future,* 5**,** 545-559.

MASAKI, Y., HANASAKI, N., BIEMANS, H., SCHMIED, H. M., TANG, Q., WADA, Y., GOSLING, S. N., TAKAHASHI, K. & HIJIOKA, Y. 2017. Intercomparison of global river discharge simulations focusing on dam operation—multiple models analysis in two case-study river basins, Missouri–Mississippi and Green–Colorado. *Environmental Research Letters,* 12**,** 055002.

MOTOVILOV, G. L., ENGELAND KOLBJØRN, BELOKUROV ALEXANDER 1999. ECOMAG – regional model of hydrological cycle. Application to the NOPEX region. Department of Geophysics, University of Oslo, Institute Report Series no.105, 1999, ISBN 82-91885-04-4, ISSN 1501–6854.

MÜLLER SCHMIED, H., ADAM, L., EISNER, S., FINK, G., FLÖRKE, M., KIM, H., OKI, T., PORTMANN, F. T., REINECKE, R., RIEDEL, C., SONG, Q., ZHANG, J. & DÖLL, P. 2016. Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use. *Hydrology and Earth System Sciences,* 20**,** 2877-2898.

MÜLLER SCHMIED, H., EISNER, S., FRANZ, D., WATTENBACH, M., PORTMANN, F. T., FLÖRKE, M. & DÖLL, P. 2014. Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human

water use and calibration. *Hydrology and Earth System Sciences,* 18**,** 3511-3538.

OLESON KW. ET AL. 2010. Technical description of version 4.0 of the Community Land Model (CLM). NCAR Tech. Note NCAR/TN-4781STR, 257 pp. .

PHILLIPS, T. J. & GLECKLER, P. J. 2006. Evaluation of continental precipitation in 20th century climate simulations: The utility of multimodel statistics. *Water Resources Research,* 42**,** 1-10.

PHUKOETPHIM, P., SHAMSELDIN, A. Y. & ADAMS, K. 2016. Multimodel Approach Using Neural Networks and Symbolic Regression to Combine the Estimated Discharges of Rainfall-Runoff Models. *Journal of Hydrologic Engineering,* 21**,** 04016022.

PINIEWSKI, M., LAIZE, C. L., ACREMAN, M. C., OKRUSZKO, T. & SCHNEIDER, C. 2014. Effect of climate change on environmental flow indicators in the narew basin, poland. *J Environ Qual,* 43**,** 155-67.

POKHREL, Y., HANASAKI, N., KOIRALA, S., CHO, J., YEH, P. J. F., KIM, H., KANAE, S. & OKI, T. 2011. Incorporating Anthropogenic Water Regulation Modules into a Land Surface Model. *Journal of Hydrometeorology,* 13**,** 255-269.

ROUGIER, J. 2016. Ensemble Averaging and Mean Squared Error. *Journal of Climate,* 29**,** 8865-8870.

SANDERSON, B. M. & KNUTTI, R. 2012. On the interpretation of constrained climate model ensembles. *Geophysical Research Letters,* 39**,** 1-6.

SCHEWE, J., HEINKE, J., GERTEN, D., HADDELAND, I., ARNELL, N. W., CLARK, D. B., DANKERS, R., EISNER, S., FEKETE, B. M., COLON-GONZALEZ, F. J., GOSLING, S. N., KIM, H., LIU, X., MASAKI, Y., PORTMANN, F. T., SATOH, Y., STACKE, T., TANG, Q., WADA, Y., WISSER, D., ALBRECHT, T., FRIELER, K., PIONTEK, F., WARSZAWSKI, L. & KABAT, P. 2014. Multimodel assessment of water scarcity under climate change. *Proc Natl Acad Sci U S A,* 111**,** 3245-50.

SHAMSELDIN, A. Y., OCONNOR, K. M. & LIANG, G. C. 1997. Methods for combining the outputs of different rainfall-runoff models. *Journal of Hydrology,* 197**,** 203-229.

SNEE, R. D. 1977. Validation of Regression-Models - Methods and Examples. *Technometrics,* 19**,** 415-428.

TANG, Q., OKI, T., KANAE, S. & HU, H. 2007. The Influence of Precipitation Variability and Partial Irrigation within Grid Cells on a Hydrological Simulation. *Journal of Hydrometeorology,* 8**,** 499-512.

THOMPSON, J. R., GREEN, A. J., KINGSTON, D. G. & GOSLING, S. N. 2013. Assessment of uncertainty in river flow projections for the Mekong River using multiple GCMs and hydrological models. *Journal of Hydrology,* 486**,** 1-30.

TRAORE, S. & GUVEN, A. 2011. New algebraic formulations of evapotranspiration extracted from gene-expression programming in the tropical seasonally dry regions of West Africa. *Irrigation Science,* 31**,** 1-10.

VAN DIJK, A. I. J. M., PEÑA-ARANCIBIA, J. L., WOOD, E. F., SHEFFIELD, J. & BECK, H. E. 2013. Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. *Water Resources Research,* 49**,** 2729-2746.

VAN HUIJGEVOORT, M. H. J., HAZENBERG, P., VAN LANEN, H. A. J., TEULING, A. J., CLARK, D. B., FOLWELL, S., GOSLING, S. N., HANASAKI, N., HEINKE, J., KOIRALA, S., STACKE, T., VOSS, F., SHEFFIELD, J. & UIJLENHOET, R. 2013. Global Multimodel Analysis of Drought in Runoff for the Second Half of the Twentieth Century. *Journal of Hydrometeorology,* 14**,** 1535-1552.

VAN LOON, A. F., VAN HUIJGEVOORT, M. H. J. & VAN LANEN, H. A. J. 2012. Evaluation of drought propagation in an ensemble mean of large-scale hydrological models. *Hydrology and Earth System Sciences,* 16**,** 4057-4078.

VELDKAMP, T. I. E., WADA, Y., AERTS, J., DOLL, P., GOSLING, S. N., LIU, J., MASAKI, Y., OKI, T., OSTBERG, S., POKHREL, Y., SATOH, Y., KIM, H. & WARD, P. J. 2017. Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century. *Nat Commun,* 8**,** 15697.

VERZANO, K. 2009. Climate change impacts on flood related hydrological processes: Further development and application of a global scale hydrological model, no. 71-2009 in Reports on Earth System Science. Max Planck Institute for Meteorology, Hamburg, Germany.

VETTER, T., HUANG, S., AICH, V., YANG, T., WANG, X., KRYSANOVA, V. & HATTERMANN, F. 2015. Multi-model climate impact assessment and intercomparison for three large-scale river basins on three continents. *Earth System Dynamics,* 6**,** 17-43.

WADA, Y., WISSER, D. & BIERKENS, M. F. P. 2014. Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth Syst. Dynam.,* 5**,** 15-40.

WARSZAWSKI, L., FRIELER, K., HUBER, V., PIONTEK, F., SERDECZNY, O. & SCHEWE, J. 2014. The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): project framework. *Proc Natl Acad Sci U S A,* 111**,** 3228-32.

WU, W., MAY, R., DANDY, G. C. & MAIER, H. R. A method for comparing data splitting approaches for developing hydrological ANN models. *In:* R. SEPPELT, A. A. V., S. LANGE, D. BANKAMP (EDS.), ed. International Environmental Modelling and Software Society (iEMSs), 2012 International Congress on Environmental Modelling and Software, Managing Resources of a Limited Planet, Sixth Biennial Meeting, 2012 Leipzig, Germany.

WU, W. Y., DANDY, G. C. & MAIER, H. R. 2014. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. *Environmental Modelling & Software,* 54**,** 108-127.

ZAHERPOUR, J., GOSLING, S. N., MOUNT, N., SCHMIED, H. M., VELDKAMP, T. I. E., DANKERS, R., EISNER, S., GERTEN, D., GUDMUNDSSON, L., HADDELAND, I., HANASAKI, N., KIM, H., LENG, G., LIU, J., MASAKI, Y., OKI, T., POKHREL, Y., SATOH, Y., SCHEWE, J. & WADA, Y. 2018a. Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts. *Environmental Research Letters,* 13**,** 065015.

ZAHERPOUR, J., MOUNT, N., GOSLING, S. N., RUTGER, D., GERTEN, D., EISNER, S., YOSHIMITSU, M., MÜLLER SCHMIED, H., TANG, Q., YOSHIHIDE, W. & LIU, X. 2018b. Optimising the combination of global-scale hydrological models from a multi-model ensemble. *Environmental software and modelling, in review*.

ZHAO, F., VELDKAMP, T. I. E., FRIELER, K., SCHEWE, J., OSTBERG, S., WILLNER, S., SCHAUBERGER, B., GOSLING, S. N., SCHMIED, H. M., PORTMANN, F. T., LENG, G., HUANG, M., LIU, X., TANG, Q., HANASAKI, N., BIEMANS, H., GERTEN, D., SATOH, Y., POKHREL, Y., STACKE, T., CIAIS, P., CHANG, J., DUCHARNE, A., GUIMBERTEAU, M., WADA, Y., KIM, H. & YAMAZAKI, D. 2017. The critical role of the routing scheme in simulating peak river discharge in global hydrological models. *Environmental Research Letters,* 12**,** 075003.

ZHOU, X. Y., ZHANG, Y. Q., WANG, Y. P., ZHANG, H. Q., VAZE, J., ZHANG, L., YANG, Y. H. & ZHOU, Y. C. 2012. Benchmarking global land surface models against the observed mean annual runoff from 150 large basins. *Journal of Hydrology,* 470**,** 269-279.

Paper V: Human impact parameterization in global hydrological models improves estimates of monthly discharges and hydrological extremes: a multi-model validation study

LETTER • **OPEN ACCESS**

Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study

To cite this article: T I E Veldkamp *et al* 2018 *Environ. Res. Lett.* **13** 055008

View the article online for updates and enhancements.

Related content

- The critical role of the routing scheme in simulating peak river discharge in global hydrological models
Fang Zhao, Ted I E Veldkamp, Katja Frieler et al.
- Intercomparison of global river discharge simulations focusing on dam operation—multiple models analysis in two case-study river basins, Missouri–Mississippi and Green–Colorado
Yoshimitsu Masaki, Naota Hanasaki, Hester Biemans et al.
- Towards a global water scarcity risk assessment framework: incorporation of probability distributions and hydro-climatic variability
T I E Veldkamp, Y Wada, J C J H Aerts et al.

Recent citations

- Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts
Jamal Zaherpour *et al*



This content was downloaded from IP address 128.243.2.30 on 21/06/2018 at 17:16

Environmental Research Letters

LETTER

Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study

T I E Veldkamp[1,9,13], F Zhao[2], P J Ward[1], H de Moel[1], J C J H Aerts[1,3], H Müller Schmied[4,5], F T Portmann[4], Y Masaki[6], Y Pokhrel[7], X Liu[8], Y Satoh[9], D Gerten[2,10], S N Gosling[11], J Zaherpour[11] and Y Wada[9,12]

1 Institute for Environmental Studies (IVM), VU Amsterdam, the Netherlands
2 Potsdam Institute for Climate Impact Research, Potsdam, Germany
3 Department of Geography, University of California, Santa Barbara, Santa Barbara, United States of America
4 Institute of Physical Geography, Goethe-University Frankfurt, Frankfurt, Germany
5 Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Frankfurt, Germany
6 National Institute for Environmental Studies, Tsukuba, Japan
7 Department of Civil and Environmental Engineering, Michigan State University, Michigan, United States of America
8 Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, People's Republic of China
9 International Institute for Applied Systems Analysis, Laxenburg, Austria
10 Department of Geography, Humboldt-Universität zu Berlin, Berlin, Germany
11 School of Geography, University of Nottingham, Nottingham, United Kingdom
12 Department of Physical Geography, Utrecht University, the Netherlands
13 Author to whom any correspondence should be addressed.

E-mail: ted.veldkamp@vu.nl

Keywords: hydrological extremes, human impact, validation, global hydrological modeling, multi-model, fresh water resources

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
16 January 2017

REVISED
21 March 2018

ACCEPTED FOR PUBLICATION
26 March 2018

PUBLISHED
4 May 2018

Original content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Abstract

Human activity has a profound influence on river discharges, hydrological extremes and water-related hazards. In this study, we compare the results of five state-of-the-art global hydrological models (GHMs) with observations to examine the role of human impact parameterizations (HIP) in the simulation of mean, high- and low-flows. The analysis is performed for 471 gauging stations across the globe for the period 1971–2010. We find that the inclusion of HIP improves the performance of the GHMs, both in managed and near-natural catchments. For near-natural catchments, the improvement in performance results from improvements in incoming discharges from upstream managed catchments. This finding is robust across the GHMs, although the level of improvement and the reasons for it vary greatly. The inclusion of HIP leads to a significant decrease in the bias of the long-term mean monthly discharge in 36%–73% of the studied catchments, and an improvement in the modeled hydrological variability in 31%–74% of the studied catchments. Including HIP in the GHMs also leads to an improvement in the simulation of hydrological extremes, compared to when HIP is excluded. Whilst the inclusion of HIP leads to decreases in the simulated high-flows, it can lead to either increases or decreases in the low-flows. This is due to the relative importance of the timing of return flows and reservoir operations as well as their associated uncertainties. Even with the inclusion of HIP, we find that the model performance is still not optimal. This highlights the need for further research linking human management and hydrological domains, especially in those areas in which human impacts are dominant. The large variation in performance between GHMs, regions and performance indicators, calls for a careful selection of GHMs, model components and evaluation metrics in future model applications.

© 2018 IOP Publishing Ltd

1. Introduction

Human activity has a profound influence on river discharges, hydrological extremes and water-related hazards, like flooding, droughts, water scarcity and water quality issues (van Loon *et al* 2016, Liu *et al* 2017, Padowski *et al* 2015, Veldkamp *et al* 2017, Wada *et al* 2011, Winsemius *et al* 2016). As a result, research efforts have been made to parameterize human activity in global hydrological models (hereafter GHMs; a full list of abbreviations is presented in supplementary table 2 available at stacks.iop.org/ERL/13/055008/mmedia) (Bierkens 2015, Pokhrel *et al* 2016). These model parameterizations include the incorporation of dam and reservoir operations, the representation of human water use and return flows, and representations of land use, land management and land cover change (Pokhrel *et al* 2016, Wada *et al* 2016a, 2017).

GHMs are widely used in scientific studies. For example, they have been used to assess the historical and future impacts of socioeconomic developments and/or hydro-climatic variability and change on freshwater resources, droughts and water scarcity (Biemans *et al* 2011, Döll *et al* 2009, Döll and Müller Schmied 2012, Fujimori *et al* 2017, Gosling *et al* 2017, Haddeland *et al* 2006, 2007, 2014, Hanasaki *et al* 2013, Van Huijgevoort *et al* 2013, Kummu *et al* 2016, Müller Schmied *et al* 2016, Munia *et al* 2016, Rost *et al* 2008, Veldkamp *et al* 2015a, 2015b, 2016, 2017, Wada *et al* 2011, 2013a, 2013b, 2014a, Wanders *et al* 2015). They are also increasingly used in practice. Global institutions are relying on GHMs more and more to conduct first-order assessments of water-related hazards, because data, time or resources are in short-supply for setting-up and executing multiple in-depth local studies. For example, GHMs have provided input into a multitude of high-level policy documents, such as the UN World Water Development Reports (e.g. Alcamo and Gallopin 2009), Global Environmental Outlooks (UNEP 2007), the World Bank series on climate change and development (Hallegatte *et al* 2016, 2017 and IPCC assessment reports (IPCC 2007, 2013).

As GHMs continue to improve in terms of detail, granularity and speed, their importance for global, regional and local applications is likely to increase further (Bierkens 2015). Therefore, it is essential to have a thorough understanding of how well these GHMs represent real-world hydrological conditions. However, most GHM validation studies are limited to near-natural river catchments and make use of naturalized discharge data (Beck *et al* 2016, Gudmundsson *et al* 2011, 2012). Studies that have validated GHM simulations where human activity is included have either focused on a single GHM and/or a few selected river catchments (Biemans *et al* 2011, Döll *et al* 2003, 2009, De Graaf *et al* 2014, Haddeland *et al* 2006, Masaki *et al* 2017, Müller Schmied *et al* 2014, Pokhrel *et al* 2012, Wada *et al* 2011, 2013a, 2014a).

To date, a comprehensive validation of the ability of multiple GHMs to represent the influence of human activity on discharge and hydrological extremes in near-natural and managed catchments is missing. As a result, there is a limited understanding of whether (and where) the parameterization of human activity in GHMs leads to an increase (or decrease) in model performance. To address this issue, the main objectives of this study are: (a) to evaluate the performance of five state-of-the-art GHMs that include the parameterization of human activity in their modeling scheme; and (b) to compare the performance of these GHMs when run with and without human impact parameterizations.

2. Data and methods

The overall methodological framework used in this study is shown in figure 1. In brief, the method involves three main steps: (1) obtaining modeled river discharges from GHMs with human impact parameterizations (HIP) and without human impact parameterization (NOHIP); (2) selecting the observed river discharge data; and (3) evaluating model performance. Each of these steps is explained in the following subsections.

2.1. Obtaining modeled river discharge from GHMs with and without HIP

We used a modeled monthly discharge ($0.5° \times 0.5°$ spatial resolution) for the period 1971–2010 from five GHMs: H08 (Hanasaki *et al* 2008a, 2008b), LPJmL (Bondeau *et al* 2007, Rost *et al* 2008, Schaphoff *et al* 2013), MATSIRO (Pokhrel *et al* 2012, 2015, Takata *et al* 2003), PCR-GLOBWB (van Beek *et al* 2011, Wada *et al* 2011, 2014b) and WaterGAP2 (Müller Schmied *et al* 2016). All simulations were carried out under the modeling framework of phase 2a of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2a: www.isimip.org/protocol/#isimip2a). For each GHM, we used two simulations: (1) HIP: a model run including time-varying land use and land cover change, historical dam construction and operation, irrigation and upstream consumptive water abstractions; and (2) NOHIP: a 'naturalized' model run without HIP.

An overview of the model characteristics of each of the GHMs, and the methods used to parameterize hydrological processes and human impact, can be found in supplementary table 1, and details on each GHM can be found in the individual model references provided therein. In the following subsections, we briefly outline the most important characteristics of the hydrological and human impact parameterizations.

2.1.1. *Parameterizations of hydrological processes*

Each GHM in this study is forced with daily (MATSIRO: three-hourly) inputs from the GSWP3 historical climate dataset (http://hydro.iis.u-tokyo.ac.jp/GSWP3). The GHMs applied in this study differ

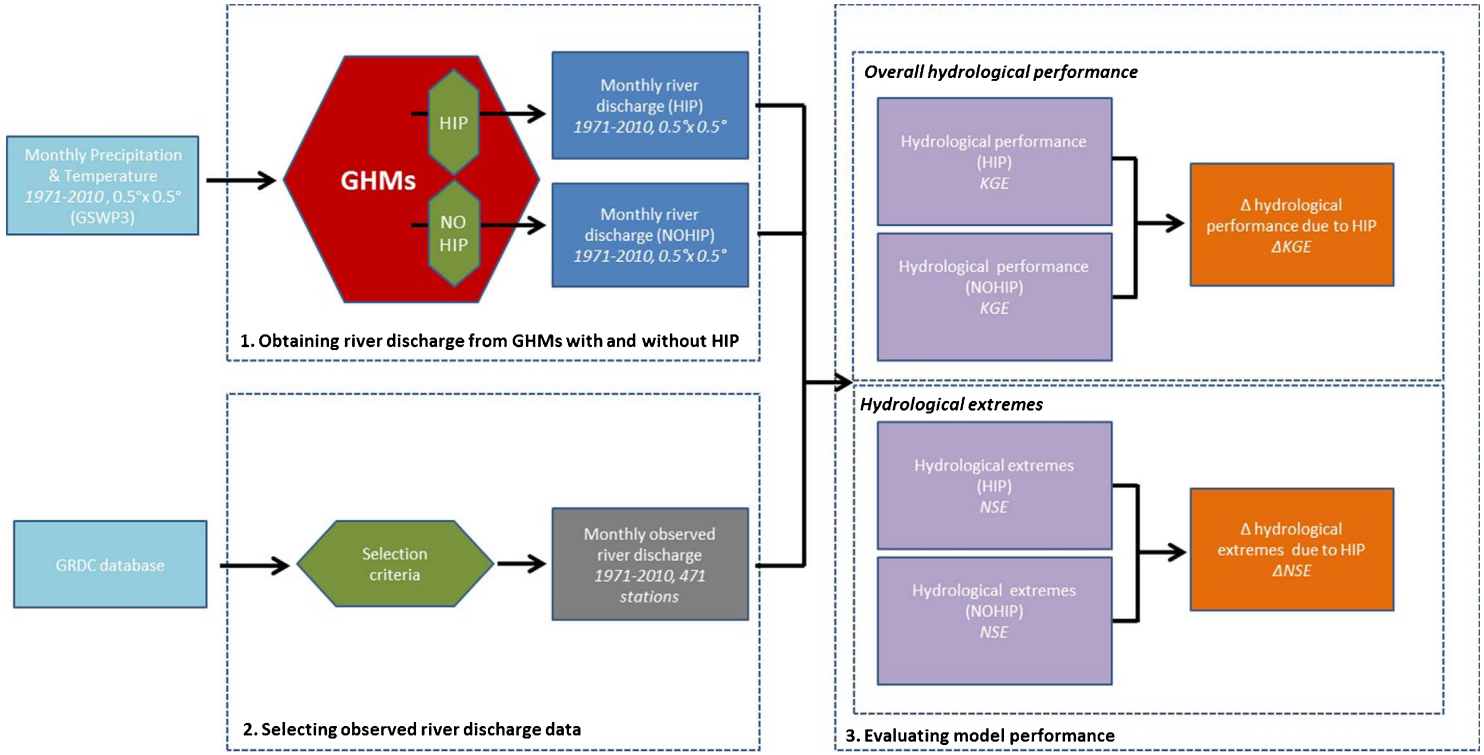


Figure 1. A flowchart of the methodological steps taken in this study. Steps 1, 2 and 3 correspond to paragraphs 2.1, 2.2 and 2.3.

in hydrological representation and parameterization (supplementary table 1A). H08 and MATSIRO model the energy balance explicitly and use the bulk formula in the evaporation scheme (Hanasaki *et al* 2008a, 2008b, Pokhrel *et al* 2012, 2015, Takata *et al* 2003). LPJmL, PCR-GLOBWB and WaterGAP2 do not include the energy balance explicitly and use the Priestley–Taylor and Hammon formulas in their evapotranspiration schemes (van Beek *et al* 2011, Bondeau *et al* 2007, Müller Schmied *et al* 2014, 2016, Schaphoff *et al* 2013, Verzano *et al* 2012, Wada *et al* 2011).

To generate runoff, all GHMs use a saturation excess formula, although the formula is integrated differently in the various GHMs. Snow accumulation and melt are integrated in the modeling framework via the energy balance (H08, MATSIRO) or by means of a degree-day calculation method (LPJmL, PCR-GLOBWB, WaterGAP2). All GHMs use a linear reservoir method in their routing scheme. Whilst H08, LPJmL and MATSIRO route with a constant flow velocity (based on Manning–Strickler), PCR-GLOBWB and WaterGAP2 use variable flow velocities. The number of soil layers and their depths vary significantly between GHMs, from one layer with varying depth (e.g. WaterGAP2, H08) to 12 fully resolved layers.

2.1.2. *Parameterization of human impact*

All GHMs use a combination of socioeconomic and hydro-climatological parameters to estimate sectoral water demands (Hanasaki *et al* 2008a, 2008b, Müller Schmied *et al* 2016, Pokhrel *et al* 2015, Rost *et al* 2008, Schaphoff *et al* 2013, Takata *et al* 2003, Van Beek *et al* 2011, Wada *et al* 2014b). Livestock water needs (supplementary table 1B) are estimated by combining historical gridded livestock density maps with their species-specific water demands. Domestic water demands (supplementary table 1C) are derived by applying a time-series regression at the country-scale, accounting for drivers like population and per capita GDP, and in some cases (PCR-GLOBWB) total electricity production, energy consumption and temperature. Industrial water demands (supplementary table 1D) are based on historical country-scale estimates from the WWDR-II dataset (Shiklomanov 1997, Vörösmarty *et al* 2005, WRI 1998) and the FAO-AQUASTAT database (www.fao.org/nr/water/aquastat/dbase/index.stm), for PCR-GLOBWB and H08 respectively. WaterGAP2 simulates global thermoelectric water use using spatially explicit information on the location of power plants. Manufacturing water demand is simulated in WaterGAP2 for each country using its yearly gross value added (GVA), and factors representing technological change and water use intensity. The models estimate irrigation water use (supplementary table 1E) by multiplying the area equipped for irrigation with its utilization intensity, the total crop-specific water requirements—determined by the hydro-climatic conditions (temperature, precipitation, potential evapotranspiration, soil moisture, crop-growth curves, length and timing of the crop-growth season), and a parameter that accounts for irrigation water use efficiency.

LPJmL, H08 and MATSIRO use surface water (first) to accommodate the sectoral water needs (supplementary table 1F). WaterGAP2 uses the groundwater to fulfill water demands, and surface water is only used if enough is available. PCR-GLOBWB applies a share of readily available groundwater reserves, based on the ratio between simulated daily base-flow and long-term mean river discharge, to be used for consumptive water needs. The remainder of the water needs are fulfilled in PCR-GLOBWB by means of surface water. Whilst all GHMs deal consistently with return flows (supplementary table 1G) for industry (surface water, same day), domestic (surface water, same day) and livestock (no return flow), returns from irrigation water use are incorporated differently. PCR-GLOBWB and H08 allow excess irrigation water return to the soil and groundwater layers by means of infiltration and additional recharge. LPJmL and MATSIRO return directly to the rivers, for which LPJmL uses a fixed ratio of 50%. Excess irrigation water in WaterGAP2 is returned to the surface waters using a cell-specific artificial drainage fraction, while the rest of the excess water is returned to the groundwater.

All GHMs include either irrigation and/or non-irrigation purposes in their reservoir schemes (supplementary table 1H), and PCR-GLOBWB also includes flood control and navigation. The retrospective operation schemes of Hanasaki *et al* (2006), Biemans *et al* (2011) and Haddeland *et al* (2006) form the basis of the reservoir operation schemes in most models. PCR-GLOBWB uses a prospective reservoir operation scheme that integrates the efforts of Haddeland *et al* (2006) and Adam *et al* (2007). H08 is the only GHM that does not account for increased evapotranspiration over reservoirs.

2.2. Selecting observed river discharge data

Observed monthly river discharge data was taken from the Global Runoff Data Centre (GRDC, 56068 Koblenz, Germany). From the 9051 gauging stations in the GRDC database, we selected stations that meet the following criteria: (1) a minimum of 5 years' coverage (not necessarily consecutive) during the period 1971–2010 with a completeness of observations of $\geq$95%; and (2) a minimum catchment area of 9000 km^2, to omit catchments whose hydrological processes cannot be adequately represented by models operating at 0.5° × 0.5° (Hunger and Döll 2008). We discarded the stations for which the difference in the catchment area in the GRDC database and as estimated by using the DDM30 river routing network (Döll and Lehner 2002) is >25%.

We then made a distinction between near-natural and managed catchments. Following Beck *et al* (2016), a catchment is classified as near-natural if the share of land-area subject to irrigation is <2% and the

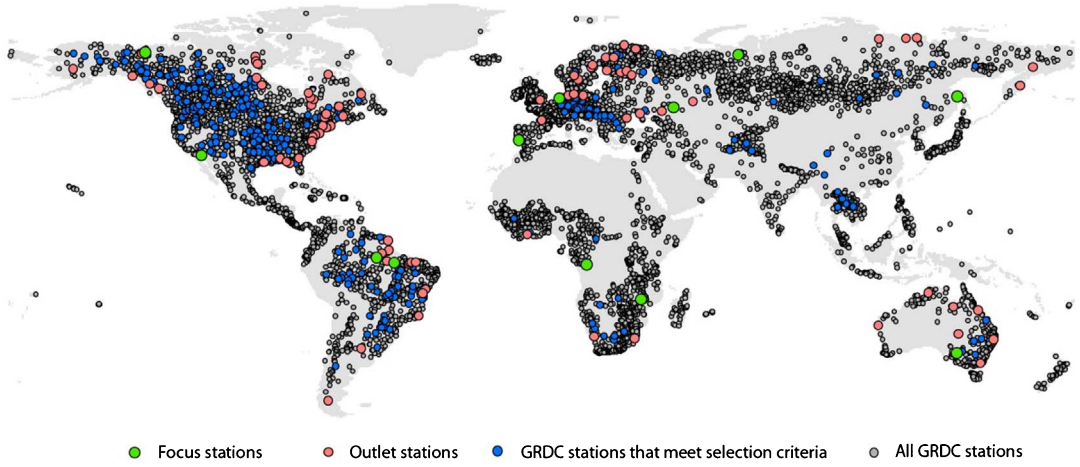


Figure 2. The spatial distribution of GRDC stations used for this study. Each dot shows a GRDC station ($n = 9051$) from the station catalogue. Blue dots indicate all GRDC stations ($n = 471$) that meet the selection criteria, whereas the red dots refer to the stations ($n = 92$) that are located at the outlet of a catchment. The green dots indicate those stations ($n = 12$) that were selected for detailed analyses.

total reservoir capacity is <10% of its long-term mean annual discharge. If these conditions were not met the catchment was classified as managed. The classification was based on the HYDE 3/MIRCA land cover dataset (Fader *et al* 2010, Klein Goldewijk and Van Drecht 2006, Portmann *et al* 2010, Ramankutty *et al* 2008) together with the Global Reservoir and Dam database (Lehner *et al* 2011). Two stations shifted from near-natural to human-impacted conditions between 1971 and 2010 and were discarded from further analysis.

The aforementioned steps resulted in 471 stations with a total catchment area covering 19.8% of the global land (figure 2), of which 92 are located at the outlet of a catchment area. The mean length of observations is 32.8 years for all stations. Of all the stations, 226 are located in managed and 245 in near-natural catchments. Of the stations located at the outlet of a catchment, 45 are managed (4.8% of the global land area) and 47 are near-natural (15.1% of the global land area).

Figure 2 shows that the majority of selected stations (blue) are located in northern and Latin America, Europe, southern Africa and Australia. The number of stations in northern and central Africa and Asia is relatively small. We selected 12 stations in river basins located in different geographic regions (green circles in figure 2: Amazonas, Amur, Colorado, Congo, Guadiana, Mackenzie, Murray, Ob, Rhine, Tocantins, Volga and the Zambezi) for which a detailed analysis is provided in the supplementary results section (supplementary).

2.3. Evaluating model performance

To evaluate the GHM simulation of monthly discharge and hydrological extremes under HIP and NOHIP conditions, we compared the modeled results with observed river discharge data using several evaluation metrics described below. To ensure a consistent comparison between the modeled and observed data, we only used modeled data for the same years for which observations were available. We also corrected modeled discharges for potential over- and underestimations caused by the difference in catchment size between the model and the GRDC. To do this, we used a multiplier that represented the difference in the upstream area as reported by the GRDC and as estimated from the DDM30 network.

First, we applied the modified Kling–Gupta efficiency index (KGE) with its sub-components: the linear correlation coefficient (rKGE), the bias ratio (βKGE) and the variability ratio (γKGE) (Gupta *et al* 2009, Kling *et al* 2012). The KGE is a widely applied indicator for the validation of hydrological performance in modeling studies at the global and regional scale and provides a good representation of the 'closeness' of simulated discharges to observations (Huang *et al* 2017, Kuentz *et al* 2013, Nicolle *et al* 2014, Revilla-Romero *et al* 2015, Thiemig *et al* 2013, 2015, Thirel *et al* 2015, Wöhling *et al* 2013). Moreover, use of its three sub-components enables the identification of reasons for sub-optimal model performance (Gupta *et al* 2009, Kling *et al* 2012, Thiemig *et al* 2013). This was achieved by estimating for each sub-parameter its distance to optimal performance, and by subsequently comparing these distances across the different sub-parameters. The statistical significance of the change in KGE outcomes due to the inclusion of HIP was tested by means of regular bootstrapping ($n = 1000$, $p \leq 0.05$ (two-tailed)), following the method of Livezey and Chen (1982) and Wilks (2006).

Second, we applied the Nash–Sutcliffe efficiency test (NSE, Nash and Sutcliffe 1970) to evaluate the representation of Q_1 (high-flow) and Q_{99} (low-flow) conditions (e.g. Beck *et al* 2017a, Blösch *et al* 2013, Hejazi and Moglen 2008, Mohamoud 2008), obtained

under fixed threshold level settings (van Loon 2015). By means of a two-sample Kolmogorov–Smirnov (KS) test (Massey 1951, $p \leq 0.05$) we tested how often HIP leads to significant changes in the fit of the full modeled exceedance probability curve for hydrological extremes compared to the full observed exceedance probability curve.

3. Results

3.1. Validation and influence of human impact parameterization on overall model performance

Including the parameterization of human impacts in the GHMs leads to a large improvement in overall model performance. The hydrological performance under the HIP simulations shows a significant improvement compared to the NOHIP simulations for between 40.8% and 72.3% of the land area studied, depending on the GHM (figure 3(*a*)). For most GHMs, the positive effects of including HIP in the simulations outweigh the negative effects. This is the case for both near-natural and managed catchments, although the positive effects are more pronounced for the managed catchments (figures 3(*a*)–(*d*)). Near-natural catchments are only indirectly impacted by HIP, for example by receiving improved or altered water simulations from upstream managed catchments. The KGE sub-components show significant improvement in performance in large shares of the land area studied, especially for the bias and variability ratio. The bias ratio improves significantly for 36.1%–73.0% of the total land area for all catchments, compared to 64.8%–90.6% and 24.3%–70.4% in managed and near-natural catchments respectively (figure 3(*b*)). For the variability ratio, improvements were found for 31.4%–74.4% of the land area for all catchments (48.9%–92.6% for the managed, 23.0%–73.2% for the near-natural) (figure 3(*c*)). The lowest improvements are found for the correlation coefficient, with improvements for 15.9%–58.1% of the total land area for all catchments (22.1%–75.1% for the managed, 13.9%–61.4% for the near-natural) (figure 3(*d*)).

The results are shown for each station in figure 4 for the overall model performance (KGE), and in supplementary figure 1 for the KGE sub-parameters. The results show particularly strong improvements in overall performance in Latin America, southern Africa and the northwest US. There are only a limited number of stations for which the inclusion of HIP leads to a significant decrease in overall hydrological performance for the majority of GHMs or where no to limited changes occur, for example in near-natural areas (e.g. the Amazonas).

When considering overall hydrological performance for each GHM under HIP conditions (figure 3(*e*)), WaterGAP2 and MATSIRO show the best performance globally. Even though the simulations with HIP include human impact parameterizations by definition, all GHMs still show better performance in near-natural catchments than in managed catchments (figures 3(*e*)–(*h*)). The KGE bias ratio values >1 indicate that all models systematically overestimate long-term mean monthly discharge (figure 3(*f*)), up to five-fold for LPJmL in managed catchments. For the variability ratio (figure 3(*g*)), WaterGAP2 is the only GHM that tends to slightly underestimate variability (variability ratio <1) in the monthly discharge, in both the managed and near-natural catchments. All other GHMs show overestimations, up to 1.55 fold for LPJmL for near-natural catchments. All GHMs show a reasonable correlation with the observed monthly discharge estimates (figure 3(*h*)), with values ranging between 0.49–0.69 in the managed catchments and 0.50–0.79 in the near-natural catchments. The highest correlation coefficients including HIP are found for WaterGAP2, with a global mean value across all catchments of 0.76 (0.69 for the managed catchments and 0.78 for the near-natural catchments).

For each catchment (and therefore its associated land area), it is possible to distinguish which of the KGE sub-parameters contributes most to sub-optimal performance. These results are summarized in figure 5. The results show that under HIP conditions, the bias ratio contributes most to sub-optimal performance in managed catchments for most GHMs, except WaterGAP2 (for which the correlation coefficient contributes most). For near-natural catchments, sub-optimal performance is most often caused by the variability ratio for H08, LPJmL and WaterGAP2, by the bias ratio for MATSIRO and by the correlation coefficient for PCR-GLOBWB.

Spatially explicit results vary per GHM and are shown in supplementary figure 3. The distribution of dominant contributors to the sub-optimal overall hydrological performance is similar for H08, LPJmL, and PCR-GLOBWB. For these GHMs, we find dominant contributions from the bias ratio in southern Africa, Australia and inland US, and dominant contributions of the variability ratio and the correlation coefficient in Latin America as well as at higher latitude and altitude regions. For Europe, the dominant contributions for H08, LPJmL and PCR-GLOBWB are the variability ratio, the correlation coefficient, and the bias ratio respectively. The dominant contributors that cause sub-optimal overall hydrological performance for MATSIRO and WaterGAP2 are more equally distributed across the globe. While all sub-components contribute to sub-optimal overall hydrological model performance for MATSIRO, it is predominantly the correlation coefficient and the variability ratio that determines the sub-optimal performance in WaterGAP2.

3.2. Validation and influence of human impact parameterizations on the simulation of hydrological extremes

The inclusion of HIP in the simulations affects the ability of GHMs to estimate hydrological extremes cor-

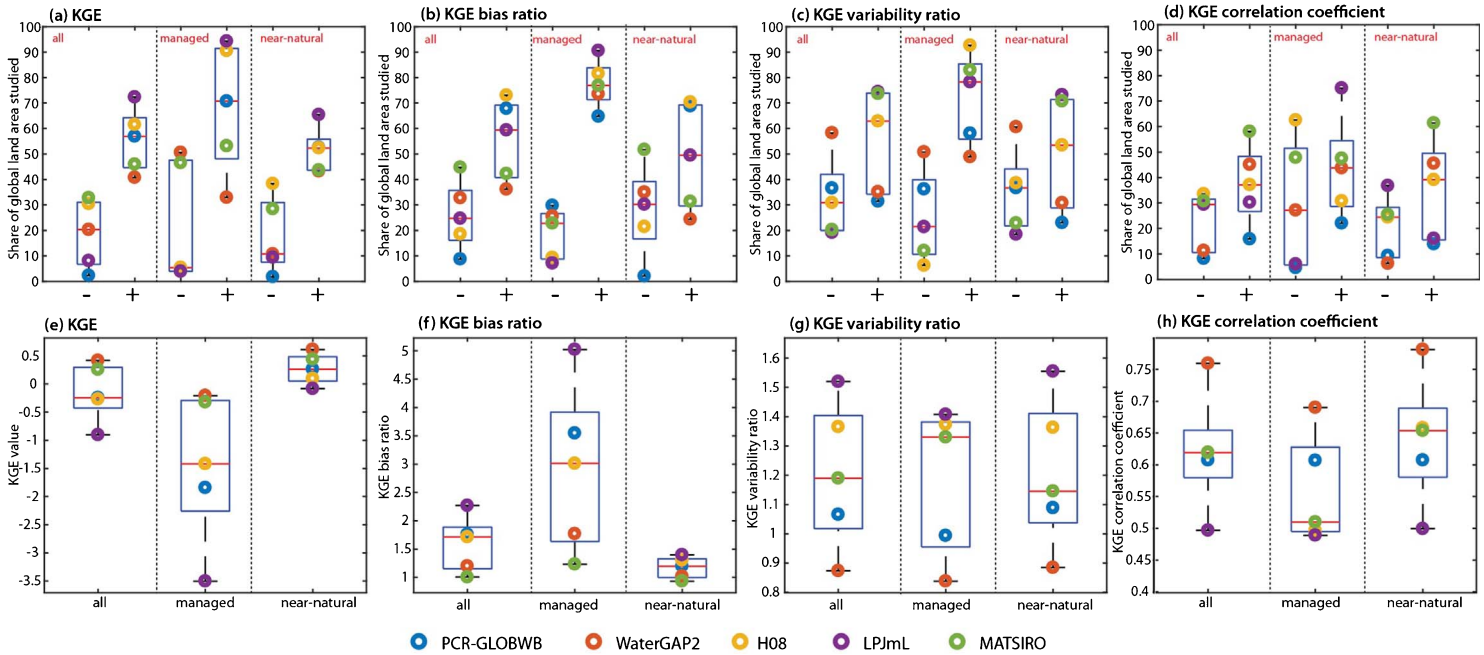


Figure 3. Global weighted-mean (improvement ('+') or deterioration ('−')) in the representation of hydrological performance due to HIP for all catchments, managed catchments and near-natural catchments. Figures 3(a)–(d) visualize for each GHM the share of land area with a significant change in overall hydrological performance due to the inclusion of HIP. Figures 3(e)–(h) indicate the globally weighted-mean hydrological performance after the inclusion of HIP. On each box, the red mark indicates the median. The bottom and top edges of the box indicate the 25th and 75th percentiles of the model ensemble, respectively.

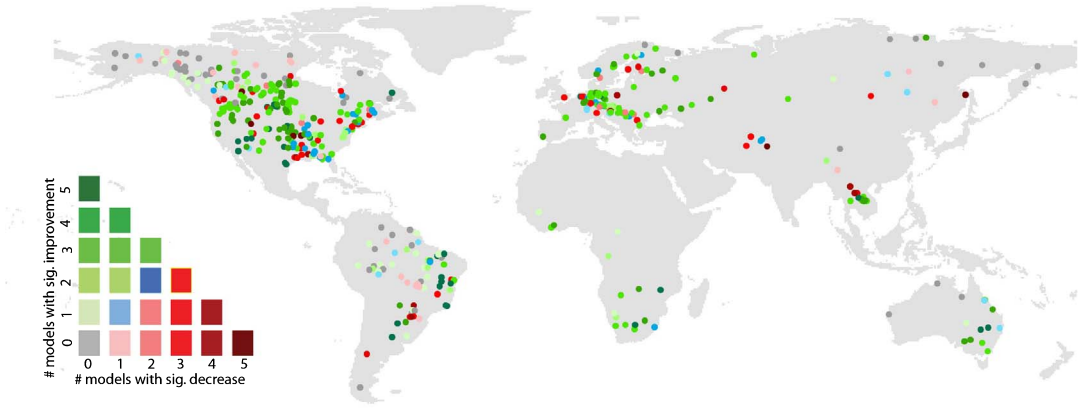


Figure 4. The number of GHMs with a significant improvement or deterioration in overall hydrological performance (KGE) due to the inclusion of HIP. The figures for the underlying KGE sub-parameters (bias ratio, variability ratio, correlation coefficient) are presented in supplementary figure 1; supplementary figure 2 shows the KGE performance values per GHM under HIP conditions.

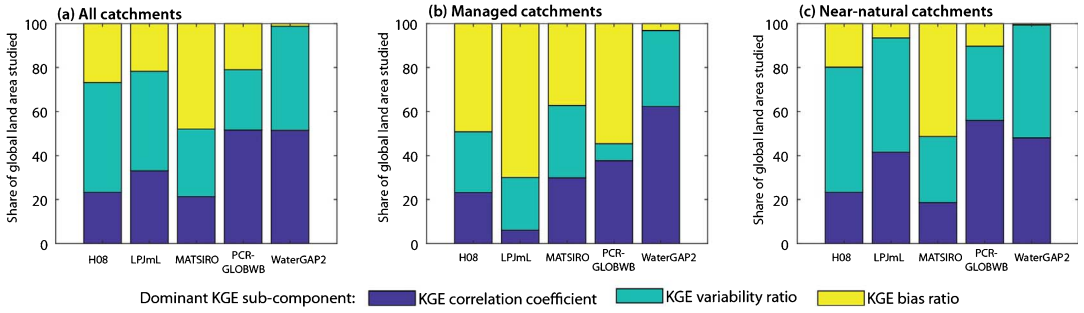


Figure 5. The share of the land area with the dominant contribution of the different KGE sub-components (KGE correlation coefficient, KGE variability ratio, KGE bias ratio) to sub-optimal overall hydrological performance under HIP conditions. Supplementary figure 3 shows per model the spatial distribution of dominant KGE sub-components.

rectly in the majority of the land area studied (figure 6). The inclusion of HIP leads to better model performance for all GHMs, across a substantial share of the land area studied (figures 6(*a*)–(*b*)). For high-flows, HIP improves model performance significantly across 34.6%–77.0% of the land area for all catchments (36.4%–94.7% for managed, 24.1%–79.2% for near-natural). For low-flows, HIP improves model performance significantly across 39.4%–80.4% of the land area for all catchments (29.3%–81.8% for managed, 42.7%–90.3% for near-natural). The KS-test results (supplementary figure 4) show that HIP only leads to significant changes in the representation of the exceedance probability curve in a limited number of cases for H08 and LPJmL (up to 14.1% of the land area studied), predominantly in managed catchments.

Overall, hydrological extremes are represented reasonably well under HIP conditions, with globally weighted mean NSE values ranging between 0.8–0.98 for high-flows, and 0.84–0.98 for low-flows (figures 6(*c*)–(*d*)). However, there is a significant difference in the ability of the GHMs to represent hydrological extremes between managed and near-natural catchments.

Figure 7 indicates that for the majority of stations, the inclusion of HIP leads to an improvement in the representation of hydrological extremes, for most GHMs. A deterioration in the representation of hydrological extremes across the majority of GHMs as a result of the inclusion of HIP was only found in selected areas, for example at higher latitudes and along the east coast of the US. When comparing the results for Q_1 high-flows with Q_{99} low-flows, no large differences in the spatial distribution of the number of GHMs are found with a significant improvement or deterioration.

The effects of HIP on the magnitude of extreme discharge differ for low-flows and high-flows (supplementary figure 5). Whilst the magnitude of high-flows mostly decreases with the inclusion of HIP, the effects on the magnitude of low-flows are both positive and negative. The convergence of results towards higher observed discharges, in both high- and low-flow estimates (as identified for all models in supplementary figure 5), indicates that HIP becomes less important for the correct representation of hydrological extremes with increasing discharge volumes.

Table 1. The performance metrics used in this study and their calculation procedure. Here, s_i and o_i are the simulated and observed monthly discharge at station i; μ_s and μ_o are the simulated and observed mean monthly discharge at station i; σ_s and σ_o are the standard deviation of the simulated and observed discharge at station i, respectively; Q_s and Q_o are the simulated and observed hydrological extremes.

Abbreviation	Name	Calculation procedure	Range and ideal value
KGE	Modified Kling–Gupta efficiency index	$\mathrm{KGE} = 1 - \sqrt{(r\mathrm{KGE}^* - 1)^2 + (\beta\mathrm{KGE}^* - 1)^2 + (\gamma\mathrm{KGE}^* - 1)^2}$	$-\infty - 1$ (ideal value: 1)
rKGE	KGE correlation coefficient (Pearson)	$r\mathrm{KGE} = \frac{\sum_{i=1}^{n}(s_i-\mu_{s,i})(o_i-\mu_{o,i})}{\sqrt{\sum_{i=1}^{n}(s_i-\mu_{s,i})^2}\sqrt{\sum_{i=1}^{n}(o_i-\mu_{o,i})^2}}$	$-1 - 1$ (ideal value: 1)
βKGE	KGE bias ratio	$\beta\mathrm{KGE} = \mu_{s,i}/\mu_{o,i}$	$0 - \infty$ (ideal value: 1)
γKGE	KGE variability ratio	$\gamma\mathrm{KGE} = \frac{\sigma_{s,i}/\mu_{s,i}}{\sigma_{o,i}/\mu_{o,i}}$	$0 - \infty$ (ideal value: 1)
NSE	Nash–Sutcliffe model efficiency	$\mathrm{NSE} = 1 - \frac{\sum(Q_s-Q_o)^2}{\sum(Q_o-\overline{Q_o})^2}$	$-\infty - 1$ (ideal value: 1)
Q_1	High-flow indicator	Monthly discharge ($\mathrm{m\,s^{-3}}$) that is exceeded on average in 1 out of 100 months	
Q_{99}	Low-flow indicator	Monthly discharge ($\mathrm{m\,s^{-3}}$) that is exceeded on average in 99 out of 100 months	
KS	Two-sample Kolmogorov–Smirnov test	$[h, p]$ = kstest2 (cdf (Q_s), cdf (Q_o), 'Alpha', 0.05)[a]	For $p > 0.05$, H_0 (the two cdfs coming from the same distribution) is not rejected.

[a] The calculation procedure for the two-sample Kolmogorov–Smirnov test presented in the table is the Matlab function for the KS-test.

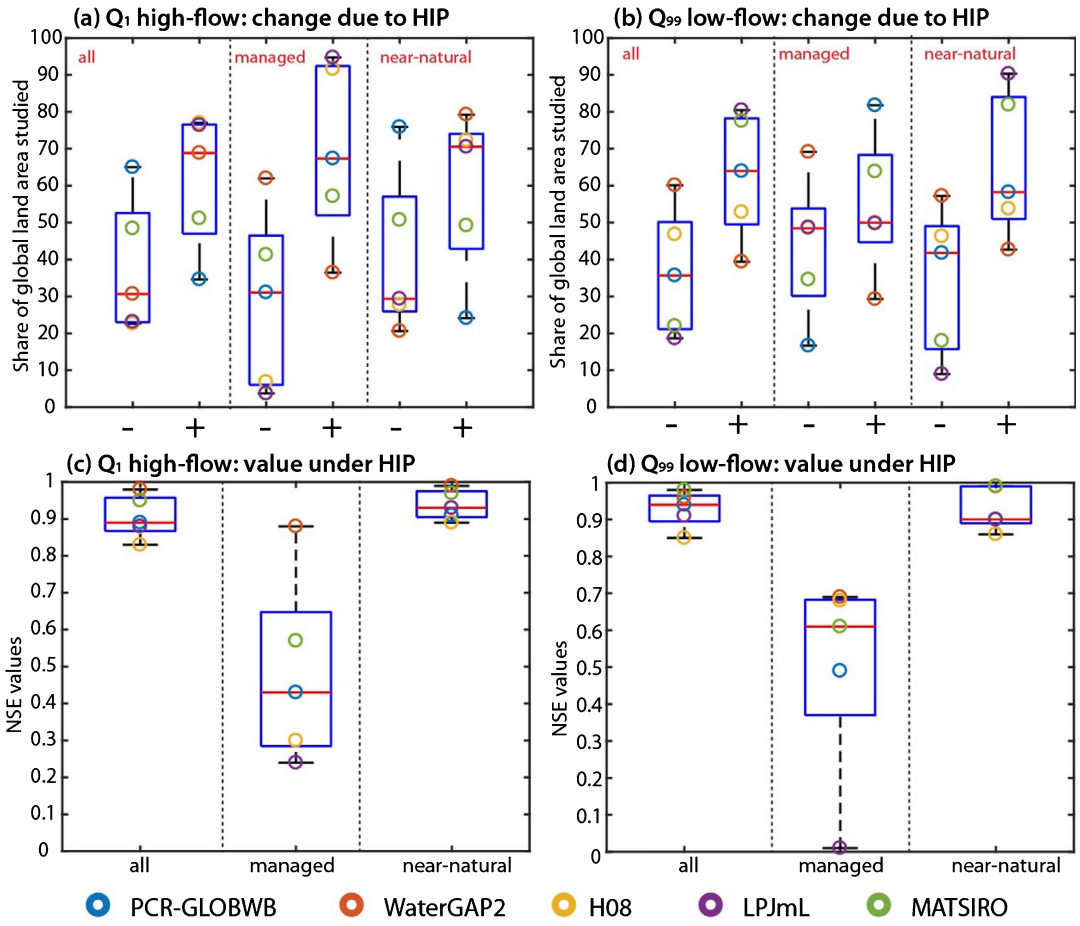


Figure 6. Global weighted-mean (improvement ('+') or deterioration ('−')) in the representation of hydrological extremes (Q_1 high-flow and Q_{99} low-flows) due to HIP, for all catchments, managed catchments, and near-natural catchments respectively. On each box, the red mark indicates the median. The bottom and top edges of the box indicate the 25th and 75th percentiles of the model ensemble, respectively.

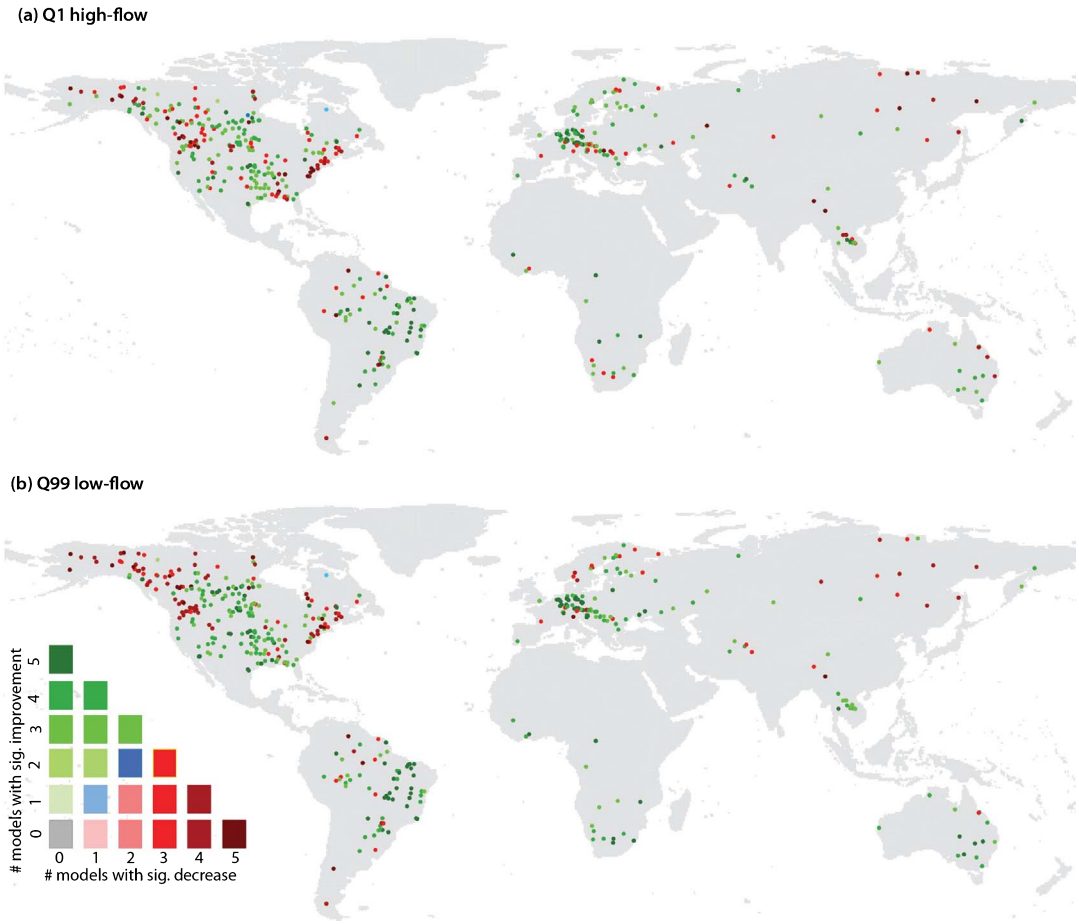


Figure 7. The number of GHMs with a significant improvement or deterioration in the representation of hydrological extremes due to the inclusion of HIP.

4. Discussion

Our results show that including HIP in GHMs generally improves the overall hydrological performance of the GHMs, as well as their representation of hydrological extremes. However, we also show that further improvements are needed. In this section, we discuss: (1) possible reasons for the improved model performance due to HIP; (2) the main limitations of the current modeling frameworks and their representation of HIP, as well as potential ways to improve them; and (3) we reflect on the general limitations in the current study design and provide suggestions for further research.

4.1. Improvements in model performance due to HIP and challenges ahead

Whilst the inclusion of HIP predominantly leads to the largest improvements in simulated discharge in the managed catchments, simulated discharge is also improved in a large share of the near-natural catchments. Improvements in model performance associated with the inclusion of HIP can be attributed to improvements in the different KGE sub-components, and in turn to different model components parameterizing hydrological and human processes. In addition, insights into those factors bounding the optimal hydrological model performance under HIP conditions may help to identify priorities for further model improvement.

4.1.1. Representation of long-term mean discharges (bias ratio)

Our study shows that the representation of long-term mean discharges significantly improved with the inclusion of HIP, especially in the managed catchments. The inclusion of HIP generally results in lower simulated discharges. As most GHMs systematically overestimate river discharges in the NOHIP simulation, this results in an improved performance. When HIP is included, we only find a deterioration in the bias ratio in selected higher latitude/altitude regions, where discharges are underestimated; this finding is in line with the outcomes of single-model studies performed by Döll *et al* (2009), De Graaf *et al* (2014) and Haddeland *et al* (2006). Improvements in the bias ratios due to the inclusion of HIP can be attributed to the inclusion of water abstractions and return flows (supplementary table 1B–G), and the incorporation of irrigated areas and irrigation rules, which influence

evapotranspiration rates and the generation of runoff (supplementary table 1E).

However, despite improvement in the bias ratio with the inclusion of HIP, this KGE sub-indicator contributes most to sub-optimal performance in managed catchments for H08, LPJmL, MATSIRO and PCR-GLOBWB under HIP conditions. As the GHMs continue to overestimate long-term mean discharges in most cases under HIP conditions, future model improvements should target the correction of this bias in these locations. This may be achieved by critically revisiting the methods used to represent evapotranspiration rates (supplementary table 1A), runoff generation processes (supplementary table 1A) and the level of water abstractions in managed catchments (supplementary table 1B–E). The relatively good performance of WaterGAP2, in which biases in long-term mean annual discharge are adjusted using a parameter that determines the portion of effective precipitation that becomes surface runoff (Müller Schmied *et al* 2014), highlights the potential importance of including a calibration routine (supplementary table 1I). Calibration is also performed for H08, but this calibration routine aims to minimize runoff bias by modifying two parameters of subsurface flow for four climatic groups (Hanasaki *et al* 2008a, 2008b); it is therefore less effective at minimizing the bias ratio under HIP conditions.

4.1.2. Representation of hydrological variability (variability ratio)

The inclusion of HIP leads to mixed results regarding the representation of hydrological variability. Whilst HIP improved the representation of variability in some catchments and for some GHMs, it deteriorated the representation of variability for others. For example, it led to improvements in the west coast US, southern Africa and Australia, but a deterioration for most GHMs in Europe and inland US. Similar results were found by Biemans *et al* (2011), De Graaf *et al* (2014) and Masaki *et al* (2017) for a selection of catchments. Changes in the variability ratio due to the inclusion of HIP are predominantly driven by the timing of water abstractions and return flows, as well as by reservoir operation rules (supplementary table 1F–H). These human activities influence the relative size of high- and low-flows compared to their long-term mean discharge values.

The variability ratio is the KGE sub-parameter that contributes most to the sub-optimal performance in near-natural catchments with the inclusion of HIP, for H08, LPJmL and WaterGAP2. These GHMs significantly overestimate hydrological variability in near-natural catchments (except WaterGAP2, which underestimates variability in managed and near-natural catchments), and model improvement should therefore focus on better representing the speed of hydrological response, e.g. through an improved representation of the soil moisture storage capacity or the ratio between surface and sub-surface runoff (supplementary table 1A). In those cases where the variability ratio is also the KGE sub-parameter that contributes most to sub-optimal performance in managed catchments, model improvement should target the timing of water abstractions, return flows and reservoir management (supplementary table 1F–H).

4.1.3. Representation of the goodness-of-fit (correlation coefficient)

The inclusion of HIP only led to improved correlation coefficients in limited cases, and often resulted in a deterioration, even in managed catchments. Correlation coefficients between observed and modeled discharges, which are predominantly determined by the hydro-meteorological forcing data (Döll *et al* 2016, Beck *et al* 2016), were found to be generally high under both HIP and NOHIP conditions. Perturbations of the hydrological cycle due to human activity leading to changes in the timing of discharges and in the shape of the hydrograph, like return flows and reservoir operations, explain the observed decrease in the correlation coefficient in a substantial share of catchments and models globally (supplementary table 1F–H).

Under HIP conditions, the correlation coefficient is the KGE sub-parameter that contributes most to sub-optimal performance only in PCR-GLOBWB for near-natural catchments and WaterGAP2 for managed catchments. It should be acknowledged, though, that correlation coefficients for PCR-GLOBWB and WaterGAP2 are relatively high, especially compared to the other GHMs. The relatively low correlation coefficients in near-natural catchments found at higher latitudes in all models may be addressed by critically reviewing the snow accumulation and melt processes in the GHMs (supplementary table 1A). Higher correlation coefficients in the managed catchments may be established by improving the timing and quantification of return flow estimates and the representativeness of reservoir operations (supplementary table 1F–H).

4.1.4. Representation of hydrological extremes

The inclusion of HIP also led to significant changes in the ability of most GHMs to represent hydrological extremes (both high- and low-flows), although the strength of this change is very much dependent on the location and GHM in question. Whilst the magnitude of high-flow estimates mainly decreased due to the inclusion of HIP, low-flow estimates showed mixed results. This is because the impact of human activity tends to be greater for lower discharges, as the relative 'size' of human perturbations (such as water abstractions, return flows or delayed releases of water via reservoir operations) is higher as a percentage of the overall discharge when flows are low. Both De Graaf *et al* (2014) and Wada *et al* (2013a) found similar results when investigating hydro-climatic extremes. However, even with the inclusion of

HIP, the representation of hydrological extremes is sub-optimal. Future model improvements should aim to better characterize these extremes and to improve the representation of human activity during extreme hydrological conditions.

4.2. Limitations and further research

As the GHMs have very different parameterizations of hydrological and human processes, the current study does not allow a systematic assessment of specific cause–effect relations between HIP and the observed improvements in performance (Döll *et al* 2016, Haddeland *et al* 2014, Hagemann *et al* 2013, Schewe *et al* 2014, Beck *et al* 2016). To do this, a substantial Monte Carlo analysis would be required, whereby individual parameters and combinations of parameters are systematically modified for all GHMs (Döll *et al* 2016). Undertaking such an analysis in parallel to the different GHMs incorporated is computationally expensive and requires a strict modeling protocol. It may, however, provide additional information on how to adapt and improve the individual GHMs and would be a valuable addition to the results presented in this study.

When interpreting the results of this study, one must take into account that we only evaluated the GHMs with respect to the monthly discharge. Whilst the monthly discharge may be sufficient for the assessment and management of low-flows, droughts and freshwater resource availability, flood risk assessment and management require information on the daily peak discharge. Further research should therefore attempt to validate GHMs using daily peak discharge and assess how it is affected by the inclusion of HIP.

The spatial resolution of the GHMs applied in this study is $0.5° \times 0.5°$ ($\sim$50 km $\times$ 50 km at the equator), dictated by the resolution of the GSWP3 input dataset. At a $0.5° \times 0.5°$ spatial resolution, hydrological processes are often represented by GHMs in a simplified or generalized form which are not fit for local applications (Bierkens 2015). To account for this, we applied a minimum catchment size of 9000 km^2, thereby omitting catchments that are too small to be adequately represented by GHMs (Hunger and Döll 2008). Newer versions of several of the GHMs now operate at higher resolutions; for example WaterGAP and PCR-GLOBWB have recently been published in 5 m/6 m versions respectively (Verzano *et al* 2012, Wada *et al* 2016b). Future research can investigate whether the inclusion of these high-resolution model-runs improves the representation of discharges and hydrological extremes in the selected catchments and whether these high-resolution runs also allow for the inclusion of smaller catchments.

In this study, a relatively simple distinction was made between managed and near-natural catchments using two parameters: irrigated agriculture and reservoirs. These parameters were chosen as they have been reported to be the most significant human parameters on river hydrology (Beck *et al* 2016, 2017a). However, to make a more detailed distinction between catchments that are impacted by human activity and those that are not, future studies could consider incorporating additional criteria, such as the share of sectoral water abstractions and return flows, and the share of built-up land area. Additional catchment descriptors (Eisner 2016), like climate conditions and the physiographic properties of the drainage area, could also be applied to further assess the important controls on modeled discharges.

When evaluating the impact of HIP on hydrological extremes, we only incorporated results for the Q_1 high-flow and Q_{99} low-flow. In this study, we did not consider other ranges of the extreme value distribution explicitly. Although the inclusion of HIP influences these hydrological extremes substantially, we found very few instances for which this led to a significant change in the full exceedance probability curve. Future research should therefore also incorporate other ranges of the probability exceedance curve in order to do a full assessment of the influence of HIP on high- and low-flow extremes.

Next to the parameterizations and representation of hydrological processes and human impacts, other sources contribute to uncertainty in the modeling of discharges and hydrological extremes. These include the quality of the data, the uncertainties in the input data and observation datasets and the calibration/validation strategy (Döll *et al* 2016, Sood and Smakhtin 2015). The quality of the selected forcing data, for example, may limit the representation of monthly discharges and hydrological extremes significantly (Döll *et al* 2016, Beck *et al* 2016), although this not been evaluated explicitly here. However, climate forcing uncertainty is probably a dominant driver for model outputs (Müller Schmied *et al* 2014, 2016). Benchmarking of the GSWP3 dataset against historical observations of precipitation and temperature, or against other forcing datasets (e.g. similar to Beck *et al* 2017b, Sun *et al* 2018), may therefore be of added value.

Differences in the quality and trustworthiness of the historical discharge observations (e.g. due to sampling, measurement and interpretation errors), may potentially result in artificial biases in the validation results (Renard *et al* 2010). The spatial representativeness of our results is limited by the availability of consistent publicly available *in situ* observations of sufficient quality. Future research should therefore consider extending the GRDC data-points to the regional repositories of observed discharges, as recently attempted by Beck *et al* (2016), Do *et al* (2017) and Gudmundsson *et al* (2017). However, increasing spatial representation comes at the cost of consistency, and special attention should be paid to the harmonization of these different databases. The use of remotely sensed data could also provide a valuable way of carrying out calibration and validation in ungauged

regions (Döll *et al* 2014a, 2014b, Scanlon *et al* 2018). Remotely sensed data can also be of added value in the assessment of water consumed by agricultural irrigation (Peña-Arancibia *et al* 2016), operational drought monitoring and early warning (Ahmadalipour *et al* 2017), as well as the estimation of terrestrial water budgets (Zhang *et al* 2017). Moreover, clear potential exists for assimilating remotely sensed data into the models (Eicker *et al* 2014).

Calibration and validation are essential to compensate for factors such as the impossibility of measuring all required model parameters at the applied scale, the lack of process understanding, the simplistic process representation in GHMs and errors in forcing data (Beck *et al* 2016, Bierkens 2015, Döll *et al* 2016, Liu *et al* 2017). Hence, calibration/validation is key for obtaining realistic model performance. It should be acknowledged, though, that the representation of hydrological and/or human processes is artificially altered by means of calibration/validation processes and that limited calibration may introduce uncertainties to the model output (Sood and Smakhtin 2015). Before using any calibrated/validated model data, one should therefore critically reflect on whether the calibration/validation procedure executed—together with their optimization objectives—are fit for the specific application in mind.

5. Summary and conclusions

This study shows that the inclusion of human activity in GHMs can significantly improve the simulation of monthly discharges and hydrological extremes, for the majority of catchments studied. The finding is robust across both managed and near-natural catchments. The global and spatially distributed results presented in this study indicate that the inclusion of human impact parameterizations (HIP) is associated with improvements in the bias ratio and the variability ratio. Whilst the biases in long-term mean monthly discharge decrease significantly in 36.1%–73.0% of the studied catchments due to the inclusion of HIP, the modeling of hydrological variability improves significantly in 31.4%–74.4% of the catchments. Estimates of hydrological extremes are also significantly influenced by the inclusion of HIP, although the influence is highly dependent on the location and GHM in question. While HIP generally leads to a decrease (and thus improvement) in the absolute magnitude of simulated high-flows, its impact on low-flows is mixed.

Even when human activity is included in GHMs, their performance is still limited; this is particularly the case in managed catchments. Moreover, the systematic misrepresentation of hydrological extremes across all GHMs calls for a careful interpretation of risk assessments based on their results, and further study into the overarching research theme of water resources, hydrological extremes, human interventions, and feedback linkages. The large variation in performance between GHMs, regions and performance indicators, highlights the importance of carefully selecting models, model components and evaluation metrics in future model applications. For example, for a study of droughts it is essential to correctly represent hydrological variability, whilst to study water scarcity it is crucial to minimize biases.

Sub-KGE results, which were presented in this study for each GHM, allow for the attribution of different hydrological and human impact model-components limiting optimal hydrological performance. In most GHMs, model performance is limited due to the overestimation of long-term mean discharges. The correlation coefficient is the limiting factor for optimal model performance for WaterGAP2, despite the high correlation coefficients that were found for this model relative to the other GHMs studied. A better understanding of these factors, as provided by this study, may assist in the identification of priorities for further model improvement.

Acknowledgments

The Global Runoff Data Centre (GRDC, 56068 Koblenz, Germany) is thanked for providing the observed discharge data. This work has been conducted under the framework of phase two of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2a: www.isimip.org) and the authors want to thank the coordination team responsible for bringing together the different global hydrological modeling groups and for coordinating the research agenda, which resulted in this manuscript. The research leading to this article is partly funded by the EU 7th Framework Programme through the project Earth2Observe (grant agreement no. 603608). JZ was funded by the Islamic Development Bank. PJW received additional funding from the Netherlands Organisation for Scientific Research (NWO) in the form of a Vidi grant (016.161.324). JCJHA received funding from the Netherlands Organisation for Scientific Research (NWO) Vici (grant no. 453-14-006). YM was supported by the Environment Research and Technology Development Fund (S-10) of the Ministry of the Environment.

ORCID iDs

T I E Veldkamp https://orcid.org/0000-0002-2295-8135
F Zhao https://orcid.org/0000-0002-4819-3724
P J Ward https://orcid.org/0000-0001-7702-7859
H Müller Schmied https://orcid.org/0000-0001-5330-9923
X Liu https://orcid.org/0000-0001-5726-7353
S N Gosling https://orcid.org/0000-0001-5973-6862

References

Adam J C, Haddeland I, Su F and Lettenmaier D P 2007 Simulation of reservoir influences on annual and seasonal streamflow changes for the Lena, Yenisei, and Ob' rivers *J. Geophys. Res. Atmos.* **112** 1–22

Ahmadalipour A, Moradkhani H, Hongxiang Y and Zarekarizi M 2017 Remote sensing of drought: vegetation, soil moisture, and data assimilation *Remote Sensing of Hydrological Extremes* ed V Lakshmi (Switzerland: Springer) pp 121–49

Alcamo J and Gallopín G 2009 Building a 2nd generation of world water scenarios *Side Publications Series—Insights* (Paris: United Nations World Water Assessment Programme)

Beck H E, Van Dijk A I J, De Roo A, Miralles D G, McVicar T M, Schellekens J and Bruijnzeel L A 2016 Global-scale regionalization of hydrologic model parameters *Water Resour. Res.* **52** 3599–622

Beck H E, Van Dijk A I J, De Roo A, Dutra E, Fink G, Orth R and Schellekens J 2017a Global evaluation of runoff from 10 state-of-the-art hydrological models *Hydrol. Earth Syst. Sci.* **21** 2881–903

Beck H E, Vergopolan N, Pan M, Levizzani V, Van Dijk A I J, Weedom G P, Brocca L, Pappenberger F, Huffman G J and Wood E F 2017b Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modelling *Hydrol. Earth Syst. Sci.* **21** 6201–17

van Beek L P H, Wada Y and Bierkens M F P 2011 Global monthly water stress: I. Water balance and water availability *Water Resour. Res.* **47** W07517

Biemans H, Haddeland I, Kabat P, Ludwig F, Hutjes R W A, Heinke J, Von Bloh W and Gerten D 2011 Impact of reservoirs on river discharge and irrigation water supply during the 20th century *Water Resour. Res.* **47** W03509

Bierkens M F P 2015 Global hydrology 2015: state, trends, and directions *Water Resour. Res.* **51** 4923–47

Blösch G, Sivapalan M, Wagener T, Viglione A and Savenije H 2013 *Runoff Predictions in Ungauged Basins: Synthesis Across Processes, Places and Scales* (Cambridge: Cambridge University Press)

Bondeau A *et al* 2007 Modelling the role of agriculture for the 20th century global terrestrial carbon balance *Glob. Change Biol.* **13** 679–706

Do H X, Gudmundsson L, Leonard M, Westra S and Seneviratne S I 2017 The global streamflow indices and metadata archive (GSIM)—Part 1: the production of daily streamflow archive and metadata *Earth Syst. Sci. Data Discuss.* in review

Döll P and Lehner B 2002 Validation of a new global 30 m in drainage direction map *J. Hydrol.* **258** 214–31

Döll P and Siebert S 2002 Global modelling of irrigation water requirements *Water Resour. Res.* **38** W1037

Döll P, Kaspar F and Lehner B 2003 A global hydrological model for deriving water availability indicators: model tuning and validation *J. Hydrol.* **270** 105–34

Döll P, Fiedler K and Zhang J 2009 Global-scale analysis of river flow alterations due to water withdrawals and reservoirs *Hydrol. Earth Syst. Sci.* **13** 2413–32

Döll P and Müller Schmied H 2012 How is the impact of climate change on river flow regimes related to the impact on mean annual runoff? A global-scale analysis *Environ. Res. Lett.* **7** 014037

Döll P, Fritsche M, Eicker A and Müller Schmied H 2014a Seasonal water storage variations as impacted by water abstractions: comparing the output of a global hydrological model with GRACE and GPS observations *Surv. Geophys.* **35** 1311–31

Döll P, Müller Schmied H, Schuh C, Portmann E T and Eicker A 2014b Global-scale assessment of groundwater depletion and related groundwater abstractions: combining hydrological modeling with information from well observations and GRACE satellites *Water Resour. Res.* **50** 5698–720

Döll P, Douville H, Güntner A, Müller Schmied H and Wada Y 2016 Modelling freshwater resources at the global scale: challenges and prospects *Surv. Geophys.* **37** 195–221

Eicker A, Schumacher M, Kusche J and Döll P and Müller Schmied H 2014 Calibration/data assimilation approach for integrating GRACE data into the WaterGAP global hydrology model (WGHM) using an ensemble Kalman filter: first results *Surv. Geophys.* **35** 1285–309

Eisner S 2016 Comprehensive evaluation of the WaterGAP3 Model across climatic, physiographic, and anthropogenic gradients *PhD Thesis* University of Kassel, Germany (http//nbn-resolving.de/urnnbndehebis34-2016031450014)

Fader M, Rost S, Muller C, Bondeau A and Gerten D 2010 Virtual water content of temperate cereals and maize: Present and potential future patterns *J. Hydrol.* **384** 218–31

Flörke M, Kynast E, Bärlund I, Eisner S, Wimmer F and Alcamo J 2013 Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: A global simulation study *Glob. Environ. Change* **23** 144–56

Fujimori S, Hanasaki N and Masui T 2017 Projections of industrial water withdrawal under shared socioeconomic pathways and climate mitigation scenarios *Sustain. Sci.* **12** 275–92

Gosling S N *et al* 2017 A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C, and 3 °C *Clim. Change* **141** 577–95

Gupta H V, Kling H, Yilmaz K K and Martinez G F 2009 Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling *J. Hydrol.* **377** 80–91

de Graaf I E M, van Beek L P H, Wada Y and Bierkens M F P 2014 Dynamic attribution of global water demand to surface water and groundwater resources: effects of abstractions and return flows on river discharges *Adv. Water Resour.* **64** 21–33

Gudmundsson L *et al* 2011 Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe *J. Hydrometeor.* **13** 604–20

Gudmundsson L, Wagener T, Tallaksen L M and Engeland K 2012 Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe *Wat. Resour. Res.* **48** W11504

Gudmundsson L, Do H X, Leonard M, Westra S and Seneviratne S I 2017 The global streamflow indices and metadata archive (GSIM)—Part 2: quality control, time-series indices and homogeneity assessment *Earth Syst. Sci. Data Discuss.* in review

Haddeland I, Skaugen T and Lettenmaier D P 2006 Anthropogenic impacts on continental surface water fluxes *Geophys. Res. Lett.* **33** L08406

Haddeland I, Skaugen T and Lettenmaier D P 2007 Hydrologic effects of land and water management in North America and Asia: 1700–1992 *Hydrol. Earth Syst. Sci.* **11** 1035–45

Haddeland I *et al* 2014 Global water resources affected by human interventions and climate change *Proc. Natl Acad. Sci. USA* **111** 3251–6

Hagemann S *et al* 2013 Climate change impact on available water resources obtained using multiple global climate and hydrology models *Earth Syst. Dyn.* **4** 129–44

Hallegatte S, Bangalore M, Bonzanigo L, Fay M, Kane T, Narloch U, Rozenberg J, Treguer D and Vogt-Schilb A 2016 Shock waves: managing the impacts of climate change on poverty *Climate Change and Development* (Washington, DC: World Bank)

Hallegatte S, Vogt-Schilb A, Bangalore M and Rozenberg J 2017 Unbreakable: building the resilience of the poor in the face of natural disasters *Climate Change Development* (Washington, DC: World Bank)

Hanasaki N, Kanaeo S and Oki T 2006 A reservoir operation scheme for global river routing models *J. Hydrol.* **327** 22–41

Hanasaki N, Kanae S, Oki T, Masuda K, Motoya K, Shirakawa N, Shen Y and Tanaka K 2008a An integrated model for the assessment of global water resources—Part 1: model description and input meteorological forcing *Hydrol. Earth Syst. Sci.* **12** 1007–25

Hanasaki N, Kanae S, Oki T, Masuda K, Motoya K, Shirakawa N, Shen Y and Tanaka K 2008b An integrated model for the

assessment of global water resources—Part 2: applications and assessments *Hydrol. Earth Syst. Sci.* **12** 1027–37
Hanasaki N *et al* 2013 A global water scarcity assessment under shared socio-economic pathways—Part 2: water availability and scarcity *Hydrol. Earth Syst. Sci.* **17** 2393–413
Hejazi M I and Moglen G E 2008 The effect of climate and land use change on flow duration in the Maryland Piedmont region *Hydrol. Proc.* **22** 4710–22
Huang S *et al* 2017 Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide *Clim. Change* **141** 381–97
Hunger M and Döll P 2008 Value of river discharge data for global-scale hydrological modelling *Hydrol. Earth Syst. Sci.* **12** 841–61
van Huijgevoort M H J *et al* 2013 Global multimodel analysis of drought in runoff for the second half of the twentieth century *J. Hydrometeor.* **14** 1535–52
IPCC 2007 Climate change 2014: synthesis report *Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Geneva: IPCC)
IPCC 2013 *Climate Change 2013: The Physical Science Basis* (Geneva: IPCC)
Klein Goldewijk K and Van Drecht G 2006 HYDE 3: Current and historical population and land cover. MNP 2006 *Integrated Modelling of Global Environmental Change. An Overview of IMAGE 2.4* ed A F Bouwman, T Kram and K K Goldewijk (Bilthoven: Netherlands Environmental Assessment Agency (MNP))
Kling H, Fuchs M and Paulin M 2012 Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios *J. Hydrol.* **424** 264–77
Kuentz A, Mathevet T, Gailhard J, Perret C and Andreassian V 2013 Over 100 years of climatic and hydrologic variability of a Mediterranean and mountainous watershed: the Durance river Cold and Mountain region hydrological systems under climate change: towards improved projections proceedings of H02, *IAHS-IAPSO-IASPEI Assembly* (Gothenburg: IAHS) p 360
Kummu M, Guillaume J H A, De Moel H, Eisner S, Flörke M, Porkka M, Siebert S, Veldkamp T I E and Ward P J 2016 The world's road to water scarcity: shortage and stress in the 20th century and pathways towards sustainability *Sci. Rep.* **6** 38495
Lehner B *et al* 2011 High resolution mapping of the world's reservoirs and dams for sustainable river flow management *Front. Ecol. Environ.* **9** 494–502
van Loon A F 2015 Hydrological drought explained *WIREs Water* **2** 359–92
van Loon A F *et al* 2016 Drought in the anthropocene *Nat. Geosci.* **9** 89–91
Livezey R E and Chen W Y 1982 Statistical field significance and its determination by Monte Carlo techniques *Mon. Weather Rev.* **111** 46–59
Liu X, Tang Q, Cui H, Mu M, Gerten D, Gosling S, Masaki Y, Wada Y and Satoh Y 2017 Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations *Environ. Res. Lett.* **12** 025009
Masaki *et al* 2017 Intercomparison of regulated river discharge among multiple global hydrological models under multiple forcings—Part II: multiple models in two case-study river basins, Missouri-Mississippi and Green-Colorado *Environ. Res. Lett.* **12** 055002
Massey F J 1951 The Kolmogorov-Smirnov test for goodness of fit *J. Am. Stat. Assoc.* **46** 68–78
Mohamoud Y M 2008 Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves *Hydrol. Sci. J.* **53** 706–24
Müller Schmied H, Eisner S, Franz D, Wattenbach M, Portmann F T, Flörke M and Döll P 2014 Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration *Hydrol. Earth. Syst. Sci.* **18** 3511–38
Müller Schmied H *et al* 2016 Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use *Hydrol. Earth Syst. Sci.* **20** 2877–98
Munia H, Guillaume J H A, Mirumachi N, Porkka M, Wada Y and Kummu M 2016 Water stress in global transboundary river basins: significance of upstream water use on downstream stress *Environ. Res. Lett.* **11** 014002
Nash J E and Sutcliffe J V 1970 River flow forecasting through conceptual models part I—A discussion of principles *J. Hydrol.* **10** 282–90
Nicolle P *et al* 2014 Benchmarking hydrological models for low-flow simulation and forecasting on French catchments *Hydrol. Earth Syst. Sci.* **18** 2829–57
Padowski J C, Gorelick S M, Thompson B H, Rozelle S and Fendorf S 2015 Assessment of human-natural system characteristics influencing global freshwater supply vulnerability *Environ. Res. Lett.* **10** 104014
Peña-Arancibia J L, Mainuddin M, Kirby J M, Chiew F H S, McVicar T R and Vaze J 2016 Assessing irrigated agriculture's surface water and groundwater consumption by combining satellite remote sensing and hydrologic modelling *Sci. Total Environ.* **542** 372–82
Pokhrel Y, Hanasaki N, Koirala S, Cho J, Yeh P J-F, Kim H, Kanae S and Oki T 2012 Incorporating anthropogenic water regulation modules into a land surface model *J. Hydrometeorol.* **13** 255–69
Pokhrel Y N, Koirala S, Yeh P J-F, Hanasaki N, Longuevergne L, Kanae S and Oki T 2015 Incorporation of groundwater pumping in a global land surface model with the representation of human impacts *Water Resour. Res.* **51** 7896
Pokhrel Y N, Hanasaki N, Wada Y and Kim H 2016 Recent progresses in incorporating human land-water management into global land surface models toward their integration into Earth system models *Wiley Interdiscip. Rev. Water* **3** 548–74
Portmann F T, Siebert S and Döll. P 2010 MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: a new high-resolution data set for agricultural and hydrological modelling *Glob. Biogeochem. Cycles* **24** GB1011
Ramankutty N, Evan A T, Monfreda C and Foley J A 2008 Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000 *Glob. Biogeochem. Cycles* **22** GB1003
Renard B, Kavetski D, Kuczera G, Thyer M and Franks S W 2010 Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors *Water Resour. Res.* **46** W05521
Revilla-Romero B, Beck H E, Burek P, Salamon P, De Roo A and Thielen J 2015 Filling the gaps: calibrating a rainfall-runoff model using satellite-derived surface water extent *Remote Sens. Environ.* **171** 118–31
Rost S, Gerten D, Bondeau A, Lucht W, Rohwer J and Schaphoff S 2008 Agricultural green and blue water consumption and its influence on the global water system *Water Resour. Res.* **44** W09405
Scanlon B R *et al* 2018 Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data *Proc. Natl Acad. Sci.* **2017** 04665
Schaphoff S, Heyder U, Ostberg S, Gerten D, Heinke J and Lucht W 2013 Contribution of permafrost soils to the global carbon budget *Environ. Res. Lett.* **8** 014026
Schewe J *et al* 2014 Multimodel assessment of water scarcity under climate change *Proc. Natl Acad. Sci. USA* **111** 3245–50
Shiklomanov I A 1997 *Assessment of Water Resources and Water Availability in the World, Comprehensive Assessment of the Freshwater Resources of the World* (Stockholm: Stockholm Environment Institute)
Sood A and Smakhtin V 2015 Global hydrological models: a review *Hydrol. Sci. J.* **60** 549–65
Sun Q, Miao C, Duan Q, Ashouri H, Sorooshian S and Hsu K-L 2018 A review of global precipitation data sets: data sources, estimation, and intercomparisons *Rev. Geophys.* **56**
Takata K, Emori S and Watanabe T 2003 Development of minimal advanced treatments of surface interaction and runoff *Glob. Planet. Change* **38** 209–22

Thiemig V, Bisselink B, Pappenberg F and Thielen J 2015 A pan-African medium-range ensemble flood forecast system *Hydr. Earth. Syst. Sci.* **19** 3365–85

Thiemig V, Rojas R, Zambrano-Bigiarini M and De Roo A 2013 Hydrological evaluation of satellite-based rainfall estimates over the Volta and Baro-Akobo basin *J. Hydrol.* **499** 324–38

Thirel G *et al* 2015 Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments *Hydrol. Sci. J.* **60** 1184–99

UNEP 2007 *Global Environment Outlook 4: environment for development* (Valletta: Progress Press) (http://drustage.unep.org/geo/assessments/global-assessments/global-environment-outlook-4)

Veldkamp T I E, Eisner S, Wada Y, Aerts J C J H and Ward P J 2015a Sensitivity of water scarcity events to ENSO-driven climate variability at the global scale *Hydrol. Earth Syst. Sci.* **19** 4081–98

Veldkamp T I E *et al* 2017 Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century *Nat. Comm.* **8** 15697

Veldkamp T I E, Wada Y, Aerts J C J H and Ward P J 2016 Towards a global water scarcity risk assessment framework: incorporation of probability distributions and hydro-climatic variability *Environ. Res. Lett.* **11** 024006

Veldkamp T I E, Wada Y, de Moel H, Kummu M, Eisner S, Aerts J C J H and Ward P J 2015b Changing mechanisms of global water scarcity events: Impacts of socioeconomic changes and inter-annual hydro-climatic variability *Glob. Environ. Change* **32** 18–29

Verzano K, Barlund I, Flörke M, Lehner B, Kynast E and Voß F 2012 Modeling variable river flow velocity on continental scale: current situation and climate change impacts in Europe *J. Hydrol.* **424** 238–51

Vörösmarty C J, Leveque C and Revenga C 2005 Freshwater ecosystems *Millennium Ecosystem Assessment Volume 1: Conditions and Trends* (Washington, DC: Island Press) pp 165–207

Wada Y, Van Beek L P H and Bierkens M F P 2011 Modelling global water stress of the recent past: on the relative importance of trends in water demand and climate variability *Hydrol. Earth Syst. Sci.* **15** 3785–808

Wada Y, Van Beek L P H, Wanders N and Bierkens M F P 2013a Human water consumption intensifies hydrological drought worldwide *Environ. Res. Lett.* **8** 034036

Wada Y *et al* 2013b Multimodel projections and uncertainties of irrigation water demand under climate change *Geophys. Res. Lett.* **40** 4626–32

Wada Y, Gleeson T and Esnault L 2014a Wedge approach to water stress *Nat. Geosc.* **7** 615–7

Wada Y, Wisser D and Bierkens M F P 2014b Global modelling of withdrawal, allocation and consumptive use of surface water and groundwater resources *Earth Syst. Dyn.* **5** 15–40

Wada Y *et al* 2016a Modeling global water use for the 21st century: the water futures and solutions (WFaS) initiative and its approaches *Geosci. Model Dev.* **9** 175–222

Wada Y, De Graaf I E M and Van Beek L P H 2016b High-resolution modeling of human and climate impacts on global water resources *J. Adv. Model. Earth Syst.* **8** 735–63

Wada Y *et al* 2017 Human–water interface in hydrological modelling: current status and future directions *Hydrol. Earth Syst. Sci.* **21** 4169–93

Wanders N, Wada Y and van Lanen H A J 2015 Global hydrological droughts in the 21st century under a changing hydrological regime *Earth Syst. Dyn.* **6** 1–15

Wilks D S 2006 On 'field significance' and the false discovery rate *J. Appl. Meteorol. Climatol.* **45** 1181–9

Winsemius H C *et al* 2016 Global drivers of future flood risk *Nat. Clim. Change* **6** 381–5

Wöhling T, Samaniego L and Kumar R 2013 Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment *Env. Earth. Sci.* **59** 453–68

WRI, UNEP, UND, and Bank World 1998 *World Resources 1998–99: A Guide to the Global Environment—Environmental Change and Human Health* (Washington, DC)

Yoshikawa S, Cho J, Yamada H G, Hansaki N and Kanae S 2014 An assessment of global net irrigation water requirements from various water supply sources to sustain irrigation: rivers and reservoirs 1960–2050 *Hydrol. Earth Syst. Sci.* **18** 4289–310

Zhang L, Dobslaw H, Stacke T, Güntner A, Dill R and Thomas M 2017 Validation of terrestrial water storage variations as simulated by different global numerical models with GRACE satellite observations *Hydrol. Earth Syst. Sci.* **21** 821–37

Zhao F *et al* 2017 The critical role of the routing scheme in simulating peak river discharge in global hydrological models *Environ. Res. Lett.* **12** 075003

Appendix A (Supplementary information for Paper I)

Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts

Jamal Zaherpour[1,19], Simon N Gosling[1], NickMount[1], HannesM¨uller Schmied[2,3], Ted I E Veldkamp[4,18], RutgerDankers[5], Stephanie Eisner[6], Dieter Gerten[7,8], Lukas Gudmundsson[9], Ingjerd Haddeland[10], NaotaHanasaki[11], Hyungjun Kim[12], Guoyong Leng[13], Junguo Liu[14,] Yoshimitsu Masaki[15], Taikan Oki[12,16], Yadu Pokhrel[17], Yusuke Satoh[18], Jacob Schewe[7] and YoshihideWada[18]

[1] School of Geography, University of Nottingham, NottinghamNG7 2RD, United Kingdom
[2] Institute of Physical Geography, Goethe-University, Frankfurt, Germany
[3] Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Frankfurt, Germany
[4] VU Amsterdam, de Boelelaan 1087, 1081 HV Amsterdam, the Netherlands
[5] Met Office, Fitzroy Road, Exeter, EX1 3PB, United Kingdom
[6] Center for Environmental Systems Research, University of Kassel, Kassel, Germany
[7] Potsdam Institute for Climate Impact Research, Telegrafenberg, 14473 Potsdam, Germany
[8] Geography Department, Humboldt-Universit¨at zu Berlin, 10099 Berlin, Germany
[9] ETH Zurich, Institute for Atmospheric and Climate Science, Universit¨atstrasse 16, 8092 Z¨urich, Switzerland
[10] Norwegian Public Roads Administration, Statens vegvesen Region vest, Postboks 43, 6861 Leikanger, Norway
[11] National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, 305–8506, Japan
[12] Institute of Industrial Science, University of Tokyo, 4–6-1 Meguro-ku, Komaba, Tokyo 153–8505, Japan
[13] Environmental Change Institute, University of Oxford, Oxford, OX1 3QY, United Kingdom
[14] School of Environmental Science and Engineering, Southern University of Science and Technology of China, Shenzhen, 518055, People's Republic ofChina
[15] Hirosaki University, Bunkyocho-3, Hirosaki, Aomori, 36-8561, Japan
[16] United Nations University, 5–53–70 Jingumae, Shibuya-ku, Tokyo 150-8925, Japan
[17] Department of Civil and Environmental Engineering, Michigan State University, East Lansing, Michigan 48824, United States of America
[18] International Institute for Applied Systems Analysis (IIASA) - Schlossplatz 1 - A-2361 Laxenburg, Austria
[19] Author to whom any correspondence should be addressed.

E-mail: lgxjz1@nottingham.ac.uk and zaherpour@gmail.com

The supplementary materials are structured into three main sections:

- Section 1: additional details on methods employed in the study.
- Section 2: supplementary tables and figures to provide further results and details.
- Section 3: supplementary discussion.

Section 1 : Methods

1.1. Study catchments

We identify study catchments setting the four following criteria. For the sake of data quality consistency, this is done based on observed data from only the Global Runoff Data Centre, GRDC (available from http://grdc.bafg.de).

1- Catchments are larger than 100,000 km^2 to conform with the World Meteorological Organisation's definition of 'major' catchments [1]. This ensured that the catchments were of sufficient size to accommodate the output resolution of the models.

2- The selected catchments cover all eight hydrobelts defined by Meybeck et al. 2013 [2].

3- Observed monthly discharge of the catchment was available for 25 years or longer, within 1971-2010 (the period that the models were run for) and without missing data. Other studies have allowed missing data (e.g. reference [3-5] in Table1, enabling them to include more catchments. We, however, preferred higher data quality, at the expense of number of catchments, because the use of longer, complete time-series, facilitated the calculation of more robust hydrological indicators, especially the return period magnitudes.

4- The catchments are independent, i.e. observed data from only one gauge was used for each catchment. If multiple gauges were available, then the most downstream gauge was selected for which all the criteria were satisfied. The use of independent catchments limits the introduction of statistical biases when calculating mean statistics across hydrobelts. However, we acknowledge that neighbouring catchments are more likely to have similar wet or dry seasons.

The criteria resulted in the selection of 40 catchments for model evaluation (Figure 1 and Table S1). Even though the selected catchments provided a good geographic coverage, the availability and quality of observed data resulted in a bias towards catchments in boreal and northern mid-latitude hydrobelts (Table S2). The least number of catchments in each hydrobelt is one (Niger basin in northern subtropical region), although this catchment does cover 20% of its hydrobelt.

Table S1. The 40 GRDC selected catchments; catchments ordered alphabetically (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML= southern, mid-latitude, SDR= southern dry and SST= southern subtropical).

No	GRDC Code	River	Gauging Station	Relative Hydrobelt	Catchments GRDC specifications			
					Area (km²)	Longitude	Latitude	Data length (yr) and range
1	4214520	ALBANY	NEAR HAT ISLAND	BOR	118000	-83.87	51.33	31 (73-10)
2	2906900	AMUR	KOMSOMOLSK	BOR	1730000	137.12	50.63	26 (71-06)
3	4213650	ASSINIBOINE	HEADINGLEY	BOR	153000	-97.4	49.87	40 (71-10)
4	4214260	CHURCHILL	ABOVE GRANVILLE FALLS	BOR	228000	-100.45	56.15	36 (71-10)
5	2998510	KOLYMA	KOLYMSKAYA	BOR	526000	158.72	68.73	28 (78-08)
6	2903430	LENA	STOLB	BOR	2460000	126.8	72.37	32 (71-02)
7	4208150	MACKENZIE	NORMAN WELLS	BOR	1570000	-126.85	65.28	30 (71-10)
8	6970250	NORTHERN DVINA	UST-PINEGA	BOR	348000	42.17	64.1	31 (71-10)
9	2912600	OB	SALEKHARD	BOR	2949998	66.53	66.57	39 (71-10)
10	2999910	OLENEK	7.5KM DOWNSTREAM OF MOUTH OF RIVER PUR	BOR	198000	123.22	72.12	39 (71-10)
11	4213680	RED	EMERSON	BOR	104000	-97.22	49.0	40 (71-10)
12	4213550	SASKATCHEWAN	THE PAS	BOR	347000	-101.18	53.83	40 (71-10)
13	4213800	WINNIPEG	SLAVE FALLS	BOR	126000	-95.57	50.22	38 (71-10)
14	2909150	YENISEI	IGARKA	BOR	2440000	86.5	67.48	32 (75-10)
15	4115200	COLUMBIA	THE DALLES, OREG.	NML	613830	-121.17	45.61	40 (71-10)
16	6442600	DANUBE	MOHACS	NML	209064	18.67	46.0	29 (71-99)
17	6978250	DON	RAZDORSKAYA	NML	378000	40.67	47.5	38 (71-10)
18	4207900	FRASER	HOPE	NML	217000	-121.45	49.38	40 (71-10)
19	6340110	LABE	NEU-DARCHAU	NML	131950	10.88	53.23	40 (71-10)
20	4127800	MISSISSIPPI	VICKSBURG, MISS.	NML	2964252	-90.9	32.32	37 (71-10)
21	6972430	NEVA	NOVOSARATOVKA	NML	281000	30.72	59.8	40 (71-10)
22	6435060	RHINE	LOBITH	NML	160800	6.11	51.84	40 (71-10)
23	4143550	ST.LAWRENCE	CORNWALL(ONTARIO), NEAR MASSENA, N.Y.	NML	773892	-74.78	45.01	40 (71-10)
24	6977100	VOLGA	VOLGOGRAD POWER PLANT	NML	1360000	44.72	48.77	39 (71-10)
25	7222222*	YANGTZE	Cuntan	NML	804859	106.6	29.62	31 (71-01)
26	2180800	YELLOW	Huayuankou	NML	730036	113.65	34.92	40 (71-10)
27	4152450	COLORADO	LEES FERRY, ARIZ.	NDR	289562	-111.58	36.87	40 (71-10)
28	4356100	SANTIAGO	EL CAPOMAL	NDR	128943	-105.12	21.83	31 (71-03)
29	1834101	NIGER	LOKOJA	NST	2074171	6.77	7.8	25 (71-05)
30	3629000	AMAZONAS	OBIDOS	Equator	4640300	-55.5	-1.9	27 (71-97)
31	3630050	XINGU	ALTAMIRA	Equator	446570	-52.22	-3.2	35 (71-08)
32	1147010	ZAIRE	KINSHASA	Equator	3475000	15.3	-4.3	40 (71-10)
33	5101200	BURDEKIN	CLARE	SST	129660	147.24	-19.76	40 (71-10)
34	3667060	PARAGUAI	PORTO MURTINHO (FB/DNOS)	SST	474500	-57.89	-21.7	37 (71-07)
35	3650481	RIO PARNAIBA	LUZILANDIA	SST	322823	-42.37	-3.45	26 (82-06)
36	3651805	SAO FRANCISCO	MANGA	SST	200789	-43.93	-14.76	37 (71-08)
37	5410100	COOPER CREEK	CALLAMURRA	SDR	230000	140.87	-27.7	33 (74-06)
38	1159100	ORANJE	VIOOLSDRIF	SDR	850530	17.73	-28.76	38 (71-10)
39	5204250	DARLING RIVER	LOUTH	SML	489300	145.11	-30.54	26 (71-07)
40	5101301	FITZROY	THE GAP	SML	135860	150.1	-23.1	40 (71-10)

*not included in GRDC database, obtained from local authority

Table S2. Characteristics of hydrobelts and catchments within each hydrobelt (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML= southern mid-latitude, SDR= southern dry and SST= southern subtropical).

Hydrobelt	Area of hydrobelt (10^3 km²) [2]	Hydrobelt runoff (mm/yr) [2]	No. of catchments in each hydrobelt	Catchments area in each hydrobelt (10^3 km2)	Catchments area to hydrobelt area (%)
BOR	25995	223	14	13298	51.2
NML	24199	343	12	7940.8	32.8
NDR	30234	36	2	418.5	1.4
NST	10579	383	1	2074.2	19.6
EQT	16826	960	3	8561.9	50.9
SST	10599	233	4	1127.8	10.6
SDR	8677	31	2	1080.5	12.5
SML	4008	292	2	625.2	15.6
Sum/Ave.	131119*	277*	40	4390.9	24.3

*Total of non-glaciated land

1.2 Models

The model ensemble comprises a suite of GHMs and LSMs that participated in the water (global) sector of ISIMIP2a [6]. The models are DBH (Distributed Biosphere Hydrological model) [7], H08 (following the initial of the developer's surname, H, and year of release, 2008) [8], LPJmL (Lund-Potsdam-Jena managed Land) [9], MATSIRO (Minimal Advanced Treatments of Surface Interaction and Runoff) [10], PCR-GLOBWB (PCRaster GLOBal Water Balance model) [11] and WaterGAP2 (Water Global Assessment and Prognosis version 2) [12] (see table S3 for details).

The calibration of GHMs and LSMs, which is typically performed at the catchment-scale, is challenging due to a lack of accurate, long-term, global-scale observations of the hydrological cycle. Only WaterGAP2 was calibrated for the ISIMIP2a simulations, in a very simple way to match long-term mean annual observed discharge for 1,319 catchments where time series were available for a reasonable period and quality [12]. Whilst we acknowledge that catchment-scale model calibration can improve model ensemble performance, particularly for present-day conditions, it has been shown that GHMs and LSMs are able to simulate realistic sensitivity to climate variability regardless of calibration [13-15].

Table S3. Key characteristics of the models evaluated. tas = Surface air temperatures, Pr = precipitation, rhs= near-surface relative humidity, rsds/rlds= Surface radiation (shortwave/longwave downwelling), wind = Near-surface wind speed, pet = potential evapotranspiration, ps = Surface air pressure

Model	Class of Model	Input climate parameters	Resolution (space and time)	Representation of soils	Representation of vegetation	Method: potential evapotranspiration	Method: snow melt	Human water use/Reservoirs	Carbon cycling
DBH [7]	GHM	Pr, tas, wind, rhs, rlds, rsds, ps	Grid cells with sub-grid heterogeneity accounting method	Three soil layers and one underlying groundwater layer at the bottom	Prescribed spatial distribution of natural vegetation and agricultural land cover	Energy balance	Energy balance	Yes/No	No
H08 [8]	GHM	tas, Pr, Snowfall, wind, rhs, rsds, rlds, ps	0.5°; 3 hourly	1-layer leaky bucket soil. Its runoff properties vary with climate zones.	Natural use: Globally uniform. No-specific land type is assigned, as known as Manabe's bucket.	Bulk formula	Energy balance	Yes/Yes	No
LPJmL [9]	GHM	Pr, tas, rsds	0.5°; daily	Five hydrologically active soil layers, coupled to carbon and thermal balance	Dynamic simulation of growth and productivity (with prescribed spatial distribution of crops and pasture); daily	Priestley-Taylor (modified for transpiration)	Degree-day with precipitation factor	Irrigation only/Yes	Yes
MATSIRO[10]	LSM	tas, Pr, Snowfall, rhs, rlds, rsds, wind, ps	0.5°; hourly	13 layers with a dynamical ground water schime. 13 soil type with 7 parameters.	Monthly LAI, and fixed vegetation characteristics for 12 types of natural land cover	Penman-Monteith	Energy balance	Yes/Yes	No
PCR-GLOBWB [11]	GHM	tas, Pr, pet	0.5°; daily	Two soil layers and one underlying groundwater layer at the bottom	Prescribed vegetation, agriculture, and land use cover	Hamon	Degree-day	Yes/Yes	No
WaterGAP2 [12]	GHM	tas, Pr, rsds, rlds	0.5°; daily	One soil layer with varying rooting depth, dependent on land cover	IGBP land cover classes based on MODIS, temperature and precipitation based LAI-model, fixed rooting depth	Priestley-Taylor with two alpha factors depending on the aridity of the grid cell	Degree-day	Yes/Yes	No

1.3. Model forcing data

All models were run for the period 1971–2010 with input climate data provided by the Global Soil Wetness Project 3, GSWP3. We used GSWP3 because it represents realistic sub-monthly variability over the entire 20th century (1901-2010), as it has been generated based on the Twentieth Century Reanalysis project (20CR) [16]. 20CR atmospheric pressure and air temperature are correlated highly with a long-term observational record (1905-2006) that is comparable to the forecast skill of a state-of-the-art forecasting system at 3 days lead time [16]. GSWP3 has already been used as the sole input forcing dataset in a GHM and LSM study by Zhao et al. (2017), for instance, and it has been included as a forcing dataset in several other recent GHM and LSM studies [17-20]. GSPW3 has also been used in the climate modelling community. It is the primary input for offline experiments conducted as part of the Land Surface, Snow and Soil Moisture Model Intercomparison Project (LS3MIP; [21]) contribution to CMIP6, and a climate driver used in the ORCHIDEE-MICT land surface model development [22]. Since our analysis is the first comprehensive evaluation of an ensemble of GHMs and LSMs run with human impacts, and as it represents one of the latest GHM and LSM ensembles in the hydrological modelling community, we decided to use one of the latest forcing datasets too. A dynamical downscaling technique that retains synoptic features in the higher spatial resolution was employed to 20CR, to develop GSWP3. Further improvements were achieved by additional bias correction, using observations, vertical damping, and single ensemble member correction (see van den Hurk et al. (2016) and references therein for further details). Whilst all global climate forcing datasets will contain some errors due to issues of spatial representativeness, Slater (2016) has shown that the solar radiation field of GSWP3 outperforms several other forcing datasets. Also, GSWP3 shows comparable performance against another forcing data in the validation of ORCHIDEE-MICT, for a comprehensive set of variables over high latitudes [22]. Further, a comparison of 2m surface temperature and precipitation from GSWP3 with 20 globally distributed in situ FLUXNET sites shows that GSWP3 performs better than Princeton and broadly similar to WFD (see Figure 5 in van den Hurk et al. (2016)).

1.4. Evaluation metrics

RMSE, MARE and CE were chosen as the three components of IPE because they have been used in past studies individually to assess different aspects of hydrological behaviour, including overall pattern of fit (CE) [23-25], low flows (MARE) [26-28] and high flows (RMSE) [4, 29, 30]; and because Dawson et al. (2007) recommends that the components of IPE include metrics of absolute error (in our study, MARE), relative error (RMSE) and a dimensionless coefficient (CE).

Table S4. Single metrics applied in this study

Metric Name	Formula	Range	Ideal value
Root mean squared error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Q_i - \widehat{Q}_i)^2}{n}}$	[0-∞)	0
Mean absolute relative error (MARE)	$\mathrm{MARE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Q_i - \widehat{Q}_i\rvert}{Q_i}$	[0-∞)	0
Coefficient of determination (RSqr)	$\mathrm{RSqr} = \left[\frac{\sum_{i=1}^{n}(Q_i - \overline{Q})\left(\widehat{Q} - \tilde{Q}\right)}{\sqrt{\sum_{i=1}^{n}(Q_i - \overline{Q})^2 \sum_{i=1}^{n}\left(\widehat{Q}_i - \tilde{Q}\right)^2}}\right]^2$	[0-1]	1
Nash-Sutcliffe (NSE) or Coefficient of efficiency (CE)	$\mathrm{CE} = 1 - \frac{\sum_{i=1}^{n}(Q_i - \widehat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \overline{Q})^2}$	(-∞, 1]	1
Correlation of coefficient (r)	$\mathrm{r} = \frac{\sum_{i=1}^{n}(Q_i - \overline{Q})\left(\widehat{Q} - \tilde{Q}\right)}{\sqrt{\sum_{i=1}^{n}(Q_i - \overline{Q})^2 \sum_{i=1}^{n}\left(\widehat{Q}_i - \tilde{Q}\right)^2}}$	[-1,1]	1
Percentage bias (PBIAS)	$PBIAS = \frac{\sum_{i=1}^{n}(Q - \hat{Q}) * 100}{\sum_{i=1}^{n} Q}$	(-∞,∞)	0*
Percentage difference	$PD = \frac{(\hat{Q} - Q) * 100}{(\widehat{Q} + Q)/2}$	(-∞,∞)	0

*Result of zero does not necessarily indicate low error due to cancellation.

In the above equations, Q is the observed value, $\hat{Q}$ denotes modelled value (where *i = 1* to *n* data points), $\bar{Q}$ is the mean of the observed data, and $\tilde{Q}$ is the mean of the modelled values.

1.5. Calculation of return periods

Maximum and minimum monthly runoff values corresponding to 2-, 5-, 10-, 20- and 25-year return periods were estimated in this study using extreme value analysis techniques.

To cover maximum runoff values, a Gumbel distribution [31] was fitted to each time series of maximum monthly runoff, for each model and the corresponding observed data, for each catchment. Subsequently, we applied the estimated Gumbel parameters to calculate the maximum monthly runoff values for the two return periods. We thereby accounted continuously for the probability of zero runoff [32]. To ensure that the runoff time-series were statistically independent, a pre-requisite for such an extreme value analysis [33, 34], we used hydrological years rather than calendar years to find the maximum monthly runoff values per year [35, 36]. For each model and

corresponding observed dataset, for each catchment, we assessed for the available time-period the timing of long-term mean maximum monthly runoff values. If this maximum monthly long-term mean runoff value of a model and observed dataset in a specific catchment, fell in the months September, October, or November, we used the period July-June to characterise the hydrological year. In all other cases, we used the standard hydrological year convention for peak flows that runs from October-September [37].

We covered extreme minimum runoff with the use of minimum 3-month flows [38], using a moving average of 3-months runoff. Subsequently, we fitted a Weibull distribution [39] – often used in low flow studies due to its lower bound and flexibility – through each of the time-series of minimum 3-month flow. We used the estimated Weibull parameters to return minimum 3-month runoff values for the different return periods. In doing so, we accounted for the occurrence of zero flows by the use of a conditional probability model [32]. The use of calendar years might not be always the proper choice to select the minimum 3-month flows for input of the extreme value analysis [33, 34, 38]. We, therefore, assessed for each model and observed dataset, for every catchment, whether the use of calendar years would possibly split the long-term mean low-flow season in the time-period [38]. If the timing of the long-term mean low flow period fell in the months December, January, or February, we applied a hydrological year that runs from August-July [38]. In all other cases, we used the standard hydrological year distinction for low flows, which runs from January-December [38].

Section 2: Results

2.1. Supplementary tables and figures

Table S5. Median percentage difference, MPD (%) between modelled and observed runoff calculated over all catchments for MAR, Q5 and Q95 (shaded cells denote underestimation of runoff by the model). The best performing model (or EM) for each hydrobelt is in bold. Hydrobelts are ranked according to the performance of the EM. Rows are ordered according to the mean latitude of each hydrobelt from north to south (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML= southern mid-latitude, SDR= southern dry and SST= southern subtropical)

Indicator	Hydrobelt (No. of catchments)	Median percentage difference, MPD (%)							Rank (based on EM)	Rank (based on best individual model)
		DBH	H08	LPJmL	MATSIRO	PCR-GLOBWB	WaterGAP2	EM		
MAR	BOR (14)	28.4	15.3	-1.0	-13.2	-0.9	**-0.1**	4.3	1	1
	NML (12)	20.9	14.8	10.6	-5.4	2.6	**-0.2**	7.8	2	2
	NDR (2)	51.2	24.9	30.3	-71.9	36.8	**-10.1**	25.7	4	7
	NST (1)	30.1	7.7	6.4	-16.0	6.3	**1.7**	10.2	3	4
	EQT (3)	54.7	54.2	50.4	-26.2	45.3	**17.6**	41.4	6	8
	SST (4)	90.0	44.0	84.9	-10.4	69.7	**4.0**	76.1	8	5
	SDR (2)	83.7	45.7	78.9	**8.6**	65.1	-9.7	67.3	7	6
	SML (2)	49.1	26.9	37.4	12.7	27.9	**1.5**	27.4	5	3
	Median over all catchments	33.7	20.1	13.5	-12.6	7.8	0.9	13.9		
	Model rank	6	5	4	3	2	1			
Q5	BOR (14)	21.6	5.3	-6.2	-11.8	-4.1	-5.7	**0.9**	1	1
	NML (12)	20.9	17.3	13.0	**-1.5**	2.7	-1.9	9.1	2	2
	NDR (2)	38.4	16.7	21.6	-74.3	22.9	-12.9	**12.6**	4	6
	NST (1)	26.8	9.9	5.6	-16.9	6.2	**-1.7**	9.6	3	4
	EQT (3)	54.1	51.5	47.6	-20.1	41.5	**17.9**	38.8	6	7
	SST (4)	75.7	39.4	71.3	-31.9	41.6	**-20.1**	53.4	8	8
	SDR (2)	67.9	36.4	62.4	**-7.1**	44.6	-19.4	47.5	7	5
	SML (2)	32.9	25.9	20.9	10.1	26.5	**-1.7**	20.1	5	3
	Median over all catchments	26.2	15.3	12.2	-11.8	5.1	-2.9	9.6		
	Model rank	6	5	4	3	2	1			
Q95	BOR (14)	36.6	17.4	6.1	**-5.5**	8.3	8.4	16.7	3	4
	NML (12)	19.2	12.1	8.3	-6.9	**0.5**	2.4	5.2	1	1
	NDR (2)	61.8	32.0	43.3	-62.7	45.2	**-1.6**	38.5	4	3
	NST (1)	33.7	3.4	3.5	-20.5	4.4	**1.5**	10.0	2	2
	EQT (3)	61.6	60.3	57.5	-37.5	57.4	**24.4**	50.2	6	6
	SST (4)	98.4	77.2	97.3	**26.8**	95.9	63.6	96.2	8	7
	SDR (2)	98.0	75.6	96.9	62.2	94.2	**43.4**	94.6	7	8
	SML (2)	67.1	26.8	59.6	15.2	45.3	**8.6**	48.3	5	5
	Median over all catchments	37.3	19.4	19.1	-8.8	14.6	5.0	21.0		
	Model rank	6	5	4	2	3	1			

Table S6. IPE scores for individual models and the EMs. Best performing model in each catchment is underlined. Models that perform better than the naïve benchmark model are shaded and bordered cells indicated where EM outperforms the best individual model. The number of catchments where each model (or the EM) performs best is in parentheses (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML= southern mid-latitude, SDR= southern dry and SST= southern subtropical).

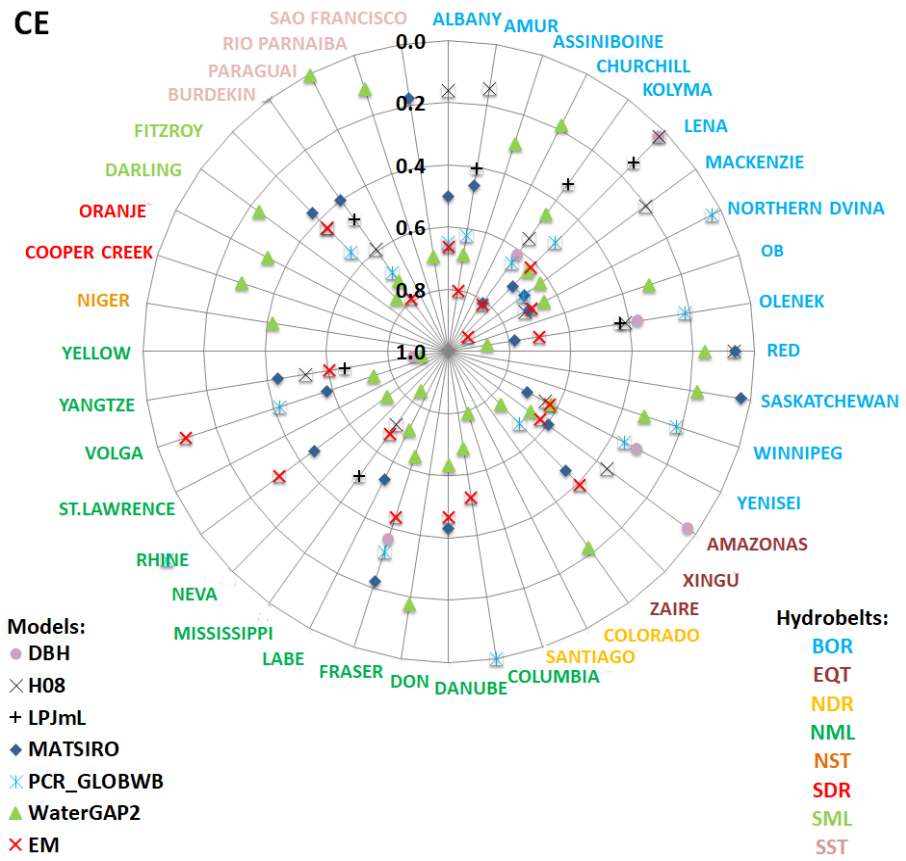
NO	Catchment	Hydrobelt	DBH (1)	H08 (1)	LPJmL (0)	MATSIRO (6)	PCR-GLOBWB (1)	WaterGAP2 (29)	EM of all 6 models (2)	EM with the exclusion of the weakest model (10)	EM with the exclusion of the two weakest models (12)
1	ALBANY	BOR	2.86	1.31	3.16	1.61	-1.43	1.97	1.13	1.12	1.22
2	AMUR	BOR	2.34	1.86	1.20	-1.10	1.36	1.09	-1.25	1.00	-1.16
3	ASSINIBOINE	BOR	273.74	47.10	504.30	2.02	22.58	1.63	49.61	26.82	7.78
4	CHURCHILL	BOR	404.36	62.35	37.91	49.15	33.38	3.36	23.90	7.79	7.31
5	KOLYMA	BOR	-1.63	1.39	-1.03	-2.21	1.73	3.41	1.06	-1.56	-2.13
6	LENA	BOR	1.18	2.06	1.49	-2.14	1.57	-1.16	-1.12	-1.42	-1.53
7	MACKENZIE	BOR	3.74	1.73	3.56	-1.25	-1.37	1.23	-2.12	-1.31	-1.45
8	N. DVINA	BOR	1.38	1.06	2.30	-1.23	-1.14	-1.44	-1.44	-1.82	-1.70
9	OB	BOR	7.32	4.19	13.36	1.91	2.62	2.35	2.82	2.35	2.21
10	OLENEK	BOR	-1.56	6.35	9.85	4.19	18.15	4.26	7.16	4.96	3.79
11	RED	BOR	11.66	2.83	7.72	-1.04	2.78	1.95	4.11	2.78	1.88
12	SASKATCHEWAN	BOR	50.24	5.81	27.70	1.50	7.14	1.40	5.53	3.15	1.64
13	WINNIPEG	BOR	26.16	5.33	6.02	4.46	1.96	1.77	3.00	2.32	2.15
14	YENISEI	BOR	-1.07	-1.52	1.18	-2.08	-1.56	-1.67	-2.07	-1.87	-2.06
15	COLUMBIA	NML	3.68	2.06	2.97	3.21	1.76	-1.26	-1.01	1.03	-1.33
16	DANUBE	NML	4.62	3.28	1.64	-1.28	2.48	-1.64	-1.25	1.53	1.23
17	DON	NML	78.67	36.77	58.18	5.59	94.15	1.61	22.32	1.87	1.77
18	FRASER	NML	1.13	2.41	4.19	1.29	1.15	1.06	1.21	1.91	2.31
19	LABE	NML	7.09	4.53	2.99	-1.20	7.84	-1.52	2.13	2.80	1.31
20	MISSISSIPPI	NML	4.97	-1.61	1.06	2.23	1.53	-2.03	-1.19	1.48	1.36
21	NEVA	NML	67.20	22.43	11.13	5.16	8.39	2.26	4.05	-1.27	-1.74
22	RHINE	NML	2.38	3.20	1.53	-1.29	1.37	-1.96	-1.14	14.94	10.04
23	ST.LAWRENCE	NML	442.57	87.98	65.71	64.77	16.04	6.52	34.47	-1.10	-1.51
24	VOLGA	NML	5.89	2.80	2.16	-1.70	-1.15	-1.81	1.22	1.52	1.21
25	YANGTZE	NML	-2.08	-1.04	-1.06	1.14	2.82	-3.04	-1.11	-2.04	-1.52
26	YELLOW	NML	15.76	3.74	3.87	1.89	23.77	1.46	3.71	3.93	2.74
27	COLORADO	NDR	17.16	2.09	6.57	7.14	3.82	3.05	2.15	-1.24	-1.62
28	SANTIAGO	NDR	11.74	6.23	2.72	1.32	11.34	-1.29	3.82	15.84	11.44
29	NIGER	NST	12.10	12.89	12.10	1.82	5.39	1.12	4.32	-1.28	-1.74
30	AMAZONAS	EQT	2.66	1.89	3.08	1.32	3.83	1.13	1.21	-1.49	-1.71
31	XINGU	EQT	5.14	3.83	4.02	2.33	1.09	-1.09	1.99	2.21	1.34
32	ZAIRE	EQT	11.08	6.79	4.86	3.21	3.03	1.42	2.60	3.45	2.60
33	BURDEKIN	SST	10.66	-1.11	4.66	-1.02	3.12	1.92	3.60	1172.43	484.30
34	PARAGUAI	SST	142.30	146.92	102.61	21.38	101.63	6.17	62.11	12.34	8.73
35	RIO PARNAIBA	SST	47.31	52.82	50.38	3.93	6.02	1.48	16.23	3528.56	2013.72
36	SAO FRANCISCO	SST	4.44	3.31	2.07	1.36	2.23	-1.41	1.54	162.86	97.37
37	COOPER CREEK	SDR	56.50	5.20	43.21	1.08	31.42	1.68	17.79	2.20	1.63
38	ORANJE	SDR	10201.36	181.53	3924.97	629.16	976.94	149.57	2677.24	52.39	38.15
39	DARLING	SML	20193.06	592.13	9588.33	777.49	6597.31	88.20	6305.78	12.65	7.66
40	FITZROY	SML	611.01	49.42	424.80	47.72	255.94	36.41	237.55	1.39	1.03

Table S7. IPE and its component values for the models and the EM.

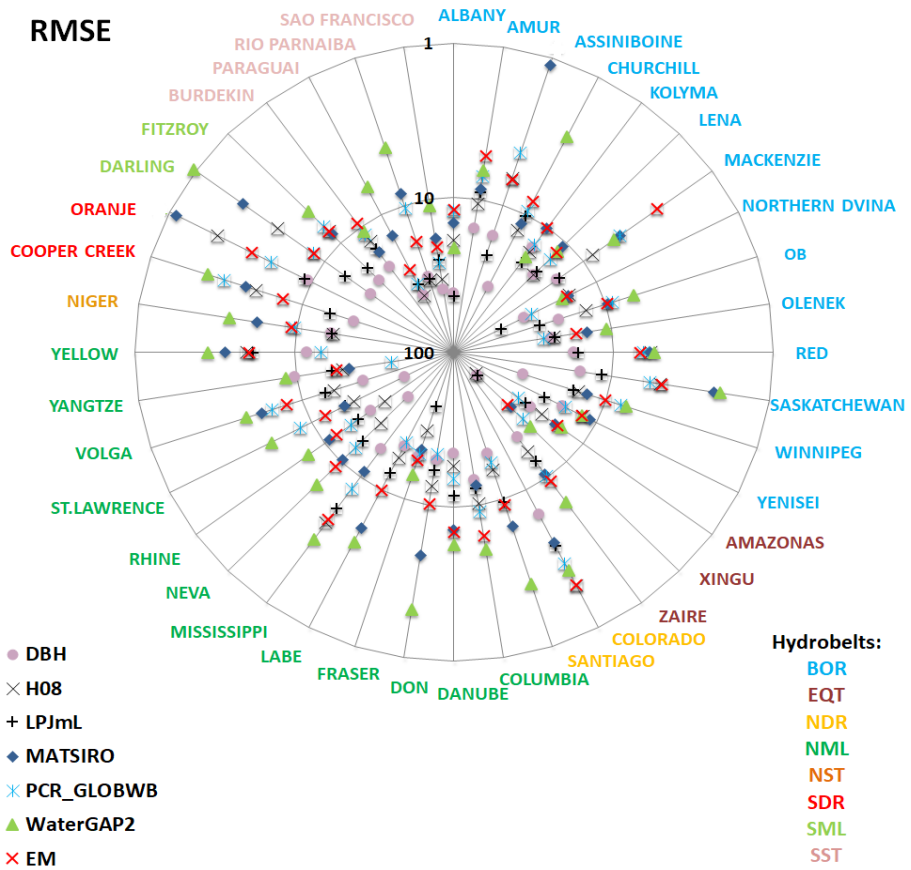
Catchment No.	Metrics	DBH	H08	LPJmL	MATSIRO	PCR_GLOBWB	WaterGap2	EM	Catchment No.	DBH	H08	LPJmL	MATSIRO	PCR_GLOBWB	WaterGap2	EM
1	IPE	2.9	1.3	3.2	1.6	-1.4	2.0	1.1	11	11.7	2.8	7.7	-1.0	2.8	1.9	4.1
	CE	-3.1	0.2	-3.5	0.5	0.6	-0.1	0.7		-7.5	0.1	-6.6	0.1	-0.1	0.2	-0.2
	RMSE	41.6	18.9	43.6	14.6	12.2	21.2	12.0		17.6	5.8	16.7	5.9	6.5	5.5	6.8
	MARE	1.4	1.1	1.6	1.6	0.6	1.9	1.1		12.0	3.1	7.2	0.7	3.0	2.1	4.5
2	IPE	2.3	1.9	1.2	-1.1	1.4	1.1	-1.2	12	50.2	5.8	27.7	1.5	7.1	1.4	5.5
	CE	-0.8	0.1	0.4	0.5	0.6	0.7	0.8		-46.7	-3.6	-25.0	0.0	-5.3	0.2	-3.5
	RMSE	15.5	10.7	8.9	8.5	7.1	6.5	5.1		15.6	4.9	11.6	2.2	5.7	2.1	4.8
	MARE	0.7	1.3	0.7	0.4	1.1	0.9	0.6		2.7	1.2	1.8	0.4	1.0	0.4	1.0
3	IPE	273.7	47.1	504.3	2.0	22.6	1.6	49.6	13	26.2	5.3	6.0	4.5	2.0	1.8	3.0
	CE	-262.5	-42.6	-487.1	-0.3	-19.1	0.3	-44.9		-14.0	-1.8	-2.3	-1.2	0.2	0.3	-0.3
	RMSE	16.0	6.5	21.8	1.1	4.4	0.8	6.7		34.8	15.2	16.2	13.2	8.0	7.4	10.1
	MARE	21.0	8.6	19.7	0.8	5.5	0.9	9.1		2.1	0.6	0.6	0.7	0.4	0.4	0.6
4	IPE	404.4	62.4	37.9	49.1	33.4	3.4	23.9	14	-1.1	-1.5	1.2	-2.1	-1.6	-1.7	-2.1
	CE	-121.2	-17.6	-10.3	-13.7	-8.8	0.2	-6.0		0.3	0.6	-0.2	0.7	0.4	0.6	0.6
	RMSE	33.4	13.1	10.2	11.6	9.5	2.7	8.0		17.1	12.2	23.0	11.1	16.5	12.6	12.5
	MARE	2.8	1.2	0.6	0.8	1.0	0.2	0.7		0.6	0.4	0.7	0.3	0.3	0.4	0.3
5	IPE	-1.6	1.4	-1.0	-2.2	1.7	3.4	1.1	15	3.7	2.1	3.0	3.2	1.8	-1.3	-1.0
	CE	0.6	0.5	0.3	0.8	0.6	0.5	0.8		-1.6	-0.3	-1.0	-1.2	0.0	0.7	0.5
	RMSE	14.4	15.7	19.1	10.2	13.9	17.2	10.1		14.7	10.3	12.9	13.4	9.1	5.1	6.3
	MARE	0.6	1.8	1.1	0.5	2.3	4.6	1.4		0.6	0.4	0.5	0.6	0.4	0.2	0.2
6	IPE	1.2	2.1	1.5	-2.1	1.6	-1.2	-1.1	16	4.6	3.3	1.6	-1.3	2.5	-1.6	-1.2
	CE	0.0	0.0	0.1	0.7	0.5	0.6	0.6		-4.5	-2.8	-0.6	0.4	-1.6	0.6	0.5
	RMSE	19.6	19.6	18.4	10.8	14.0	12.0	12.3		22.1	18.3	11.9	7.1	15.2	5.7	6.9
	MARE	1.0	2.1	1.4	0.3	1.6	0.8	0.9		0.7	0.5	0.3	0.2	0.5	0.1	0.2
7	IPE	3.7	1.7	3.6	-1.3	-1.4	1.2	-2.1	17	78.7	36.8	58.2	5.6	94.1	1.6	22.3
	CE	-1.8	0.2	-1.7	0.7	0.7	0.6	0.9		-74.7	-33.7	-54.4	-3.3	-89.8	0.2	-19.3
	RMSE	15.8	8.4	15.3	5.2	5.2	5.7	2.6		19.7	13.3	16.8	4.7	21.6	2.1	10.2
	MARE	0.8	0.7	0.8	0.3	0.2	0.5	0.2		3.7	2.6	3.6	0.9	4.2	0.4	2.2
8	IPE	1.4	1.1	2.3	-1.2	-1.1	-1.4	-1.5	18	1.1	2.4	4.2	1.3	1.2	1.1	1.2
	CE	-0.2	0.7	-1.6	0.7	0.0	0.6	0.7		0.4	-0.4	-2.0	0.2	0.3	0.6	0.4
	RMSE	32.0	15.4	46.4	15.7	28.5	17.2	16.0		19.7	29.2	42.7	21.8	20.3	14.7	18.5
	MARE	1.1	1.0	1.8	0.7	0.6	0.6	0.6		0.4	1.0	1.3	0.4	0.4	0.6	0.5
9	IPE	7.3	4.2	13.4	1.9	2.6	2.3	2.8	19	7.1	4.5	3.0	-1.2	7.8	-1.5	2.1
	CE	-5.6	-1.9	-11.1	-0.5	-0.3	0.3	-0.5		-6.1	-3.8	-1.9	0.5	-7.0	0.7	-0.6
	RMSE	20.2	13.4	27.4	9.7	9.1	6.5	9.7		20.9	17.2	13.3	5.3	22.3	4.2	9.9
	MARE	1.4	1.6	2.8	0.4	1.2	1.3	1.3		1.7	0.6	0.7	0.3	1.8	0.3	0.7
10	IPE	-1.6	6.4	9.9	4.2	18.2	4.3	7.2	20	5.0	-1.6	1.1	2.2	1.5	-2.0	-1.2
	CE	0.4	0.4	0.4	0.8	0.2	0.9	0.7		-3.4	0.7	0.5	-0.9	0.0	0.8	0.7
	RMSE	24.1	23.3	22.9	14.3	26.9	10.8	16.6		16.9	4.3	5.7	11.1	8.2	3.2	4.6
	MARE	1.5	20.3	31.6	13.4	58.2	13.7	22.9		1.3	0.2	0.4	0.5	0.5	0.2	0.3

Table S7 (Cont.). IPE and its component values for the models and the EM.

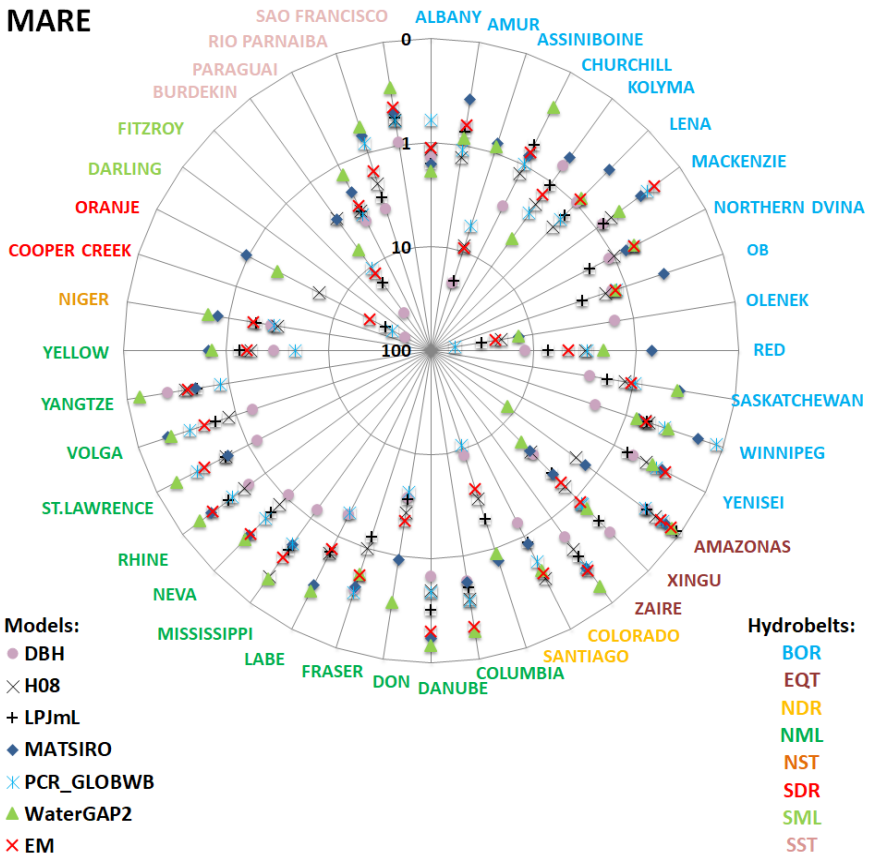
Catchment No.	Metrics	DBH	H08	LPJmL	MATSIRO	PCR_GLOBWB	WaterGap2	EM	Catchment No.	DBH	H08	LPJmL	MATSIRO	PCR_GLOBWB	WaterGap2	EM
	IPE	67.2	22.4	11.1	5.2	8.4	2.3	4.1		5.1	3.8	4.0	2.3	1.1	-1.1	2.0
	CE	-41.0	-12.7	-5.6	-2.0	-3.9	0.0	-1.2		-1.1	-1.2	-1.1	0.5	0.7	0.8	0.4
21	RMSE	39.3	22.5	15.6	10.4	13.4	6.1	9.0	31	62.2	63.9	62.0	31.5	24.5	21.0	33.4
	MARE	1.1	0.8	0.6	0.3	0.5	0.3	0.3		3.8	1.4	2.2	2.1	0.8	0.7	1.6
	IPE	2.4	3.2	1.5	-1.3	1.4	-2.0	-1.1		11.1	6.8	4.9	3.2	3.0	1.4	2.6
	CE	-1.4	-2.8	-0.5	0.5	-0.1	0.8	0.3		-7.3	-4.0	-2.4	-1.2	-1.0	0.2	-0.7
22	RMSE	22.9	29.1	18.2	10.9	15.8	7.4	12.3	32	20.9	16.1	13.4	10.6	10.2	6.4	9.3
	MARE	0.6	0.6	0.4	0.2	0.4	0.2	0.2		0.6	0.4	0.4	0.3	0.3	0.2	0.2
	IPE	442.6	88.0	65.7	64.8	16.0	6.5	34.5		4.4	3.3	2.1	1.4	2.2	-1.4	1.5
	CE	-223.9	-43.5	-32.1	-31.6	-6.9	-2.0	-16.3		-2.8	-1.9	-0.6	0.2	-0.7	0.7	-0.1
23	RMSE	44.8	19.9	17.2	17.1	8.4	5.2	12.4	33	38.6	33.4	24.7	17.9	25.9	11.0	20.5
	MARE	1.3	0.6	0.6	0.6	0.3	0.2	0.3		0.9	0.6	0.5	0.5	0.6	0.3	0.4
	IPE	5.9	2.8	2.2	-1.7	-1.1	-1.8	1.2		47.3	52.8	50.4	3.9	6.0	1.5	16.2
	CE	-8.3	-2.7	-1.9	0.6	0.4	0.7	0.1		-48.7	-55.0	-52.3	-2.7	-4.9	0.1	-15.8
24	RMSE	25.4	16.2	14.2	5.4	6.4	4.3	7.9	34	30.5	32.3	31.5	8.3	10.5	4.1	17.7
	MARE	1.5	0.8	0.6	0.2	0.3	0.2	0.5		3.7	2.0	2.8	0.7	0.8	0.5	1.5
	IPE	-2.1	-1.0	-1.1	1.1	2.8	-3.0	-1.1		10.7	-1.1	4.7	-1.0	3.1	1.9	3.6
	CE	0.9	0.5	0.7	0.4	-1.0	0.9	0.6		0.0	0.6	0.5	0.4	0.7	0.7	0.8
25	RMSE	9.7	19.7	16.8	21.6	40.1	8.5	18.0	35	20.6	13.0	14.8	15.9	11.4	10.8	9.3
	MARE	0.3	0.4	0.5	0.5	0.8	0.1	0.4		35.6	2.7	15.5	2.8	10.4	6.3	12.0
	IPE	15.8	3.7	3.9	1.9	23.8	1.5	3.7		142.3	146.9	102.6	21.4	101.6	6.2	62.1
	CE	-18.2	-2.4	-3.0	-0.8	-28.0	-0.1	-2.6		-34.8	-36.1	-24.8	-4.1	-24.5	0.0	-14.5
26	RMSE	12.0	5.1	5.5	3.7	14.8	2.9	5.2	36	38.0	38.7	32.3	14.3	32.1	6.3	25.0
	MARE	2.9	1.8	1.4	0.7	4.7	0.7	1.6		4.0	3.1	3.2	2.0	3.4	1.3	2.7
	IPE	17.2	2.1	6.6	7.1	3.8	3.1	2.2		10201.4	181.5	3925.0	629.2	976.9	149.6	2677.2
	CE	-10.8	-0.1	-3.2	-3.6	-1.3	-0.8	-0.1		-60.1	-2.1	-28.3	-1.3	-0.2	0.3	-6.1
27	RMSE	6.6	2.0	3.9	4.2	2.9	2.6	2.0	37	22.0	5.0	15.2	4.3	3.0	2.4	7.5
	MARE	1.4	0.3	0.9	0.8	0.5	0.4	0.4		881950.5	15693.7	339328.9	54393.9	84460.7	12930.8	231459.7
	IPE	11.7	6.2	2.7	1.3	11.3	-1.3	3.8		56.5	5.2	43.2	1.1	31.4	1.7	17.8
	CE	-11.5	-6.4	-1.7	-0.3	-8.5	0.8	-1.5		-77.1	-3.2	-68.9	-0.1	-22.9	0.3	-11.8
28	RMSE	20.6	15.8	9.5	6.5	17.9	2.7	9.1	38	9.4	2.2	8.9	1.1	5.2	0.9	3.8
	MARE	8.8	3.2	2.0	0.8	11.0	0.9	4.0		52.0	6.0	31.6	1.0	37.5	2.1	21.4
	IPE	12.1	12.9	12.1	1.8	5.4	1.1	4.3		20193.1	592.1	9588.3	777.5	6597.3	88.2	6305.8
	CE	-10.4	-11.1	-10.6	-0.3	-2.9	0.4	-2.5		-419.2	-14.6	-171.1	-3.5	-53.3	0.2	-54.3
29	RMSE	16.7	17.2	16.8	5.6	9.8	3.8	9.3	39	22.6	4.4	14.4	2.3	8.1	1.0	8.2
	MARE	2.7	3.0	1.9	0.8	2.8	0.6	1.7		189815.3	5565.7	90132.2	7308.9	62018.8	829.1	59278.2
	IPE	2.7	1.9	3.1	1.3	3.8	1.1	1.2		611.0	49.4	424.8	47.7	255.9	36.4	237.5
	CE	0.0	0.4	-0.2	0.6	-0.5	0.7	0.6		-3.4	0.4	-1.7	0.4	0.6	0.8	0.4
30	RMSE	25.3	20.6	28.0	16.3	31.9	14.8	15.7	40	22.0	7.8	17.0	8.3	7.0	5.1	7.8
	MARE	0.3	0.2	0.2	0.1	0.3	0.1	0.1		29447.1	2381.8	20473.2	2299.7	12335.2	1754.8	11448.5


CE
0.0
0.2
0.4
0.6
0.8
1.0
ALBANY
AMUR
ASSINIBOINE
CHURCHILL
KOLYMA
LENA
MACKENZIE
NORTHERN DVINA
OB
OLENEK
RED
SASKATCHEWAN
WINNIPEG
YENISEI
AMAZONAS
XINGU
ZAIRE
COLORADO
SANTIAGO
COLUMBIA
DANUBE
DON
FRASER
LABE
MISSISSIPPI
NEVA
RHINE
ST.LAWRENCE
VOLGA
YANGTZE
YELLOW
NIGER
COOPER CREEK
ORANJE
DARLING
FITZROY
BURDEKIN
PARAGUAI
RIO PARNAIBA
SAO FRANCISCO
Models:
DBH
H08
LPJmL
MATSIRO
PCR_GLOBWB
WaterGAP2
EM
Hydrobelts:
BOR
EQT
NDR
NML
NST
SDR
SML
SST

a


RMSE
1
10
100
ALBANY
AMUR
ASSINIBOINE
CHURCHILL
KOLYMA
LENA
MACKENZIE
NORTHERN DVINA
OB
OLENEK
RED
SASKATCHEWAN
WINNIPEG
YENISEI
AMAZONAS
XINGU
ZAIRE
COLORADO
SANTIAGO
COLUMBIA
DANUBE
DON
FRASER
LABE
MISSISSIPPI
NEVA
RHINE
ST.LAWRENCE
VOLGA
YANGTZE
YELLOW
NIGER
COOPER CREEK
ORANJE
DARLING
FITZROY
BURDEKIN
PARAGUAI
RIO PARNAIBA
SAO FRANCISCO
DBH
H08
LPJmL
MATSIRO
PCR_GLOBWB
WaterGAP2
EM
Hydrobelts:
BOR
EQT
NDR
NML
NST
SDR
SML
SST

b



c

Figure S1. Radar plots of the three components of IPE (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML= southern mid-latitude, SDR= southern dry and SST= southern subtropical).

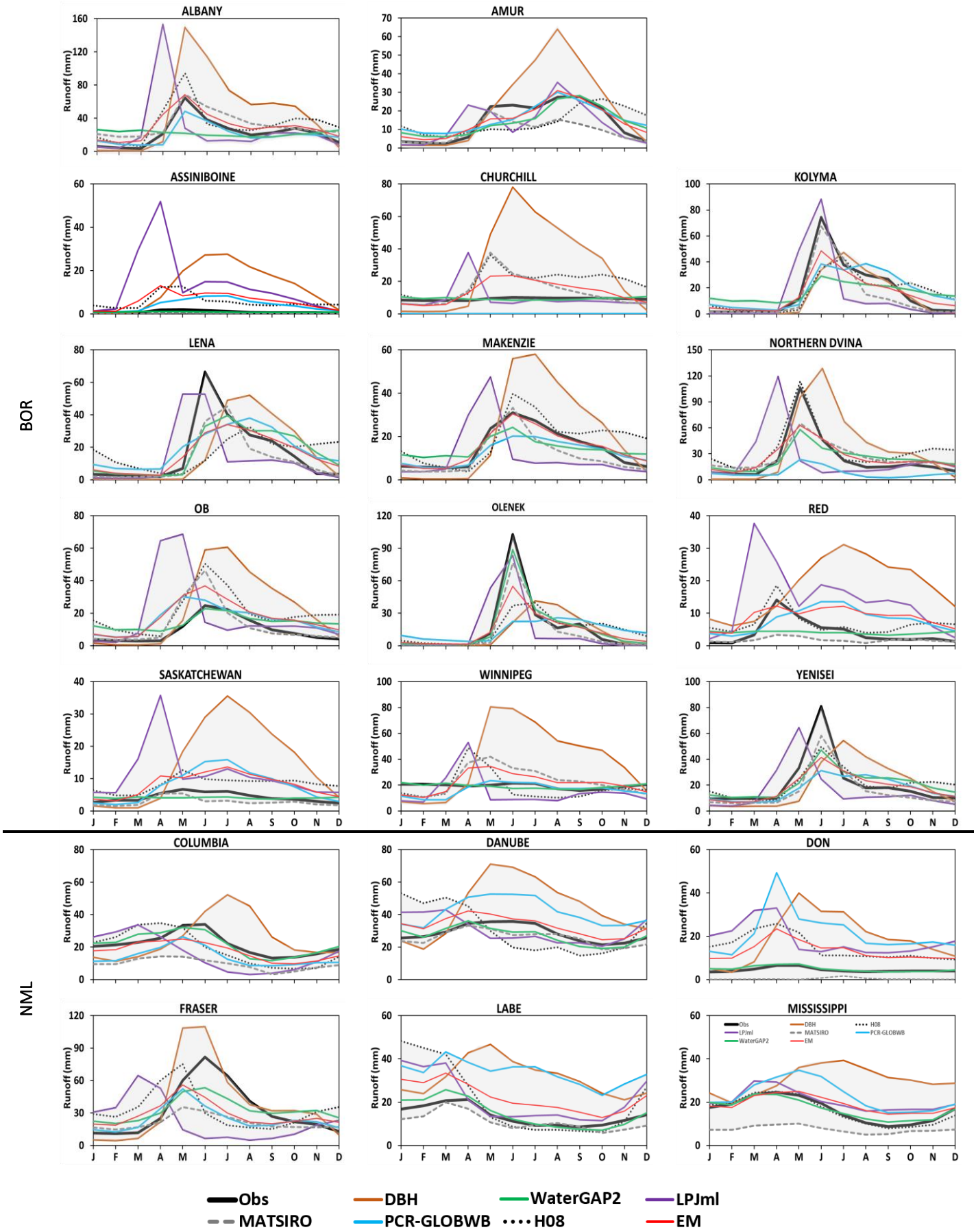


Figure S2. Seasonal cycle (long-term mean monthly runoff, MMR). The shaded region denotes the models' performance range (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML= southern mid-latitude, SDR= southern dry and SST= southern subtropical).

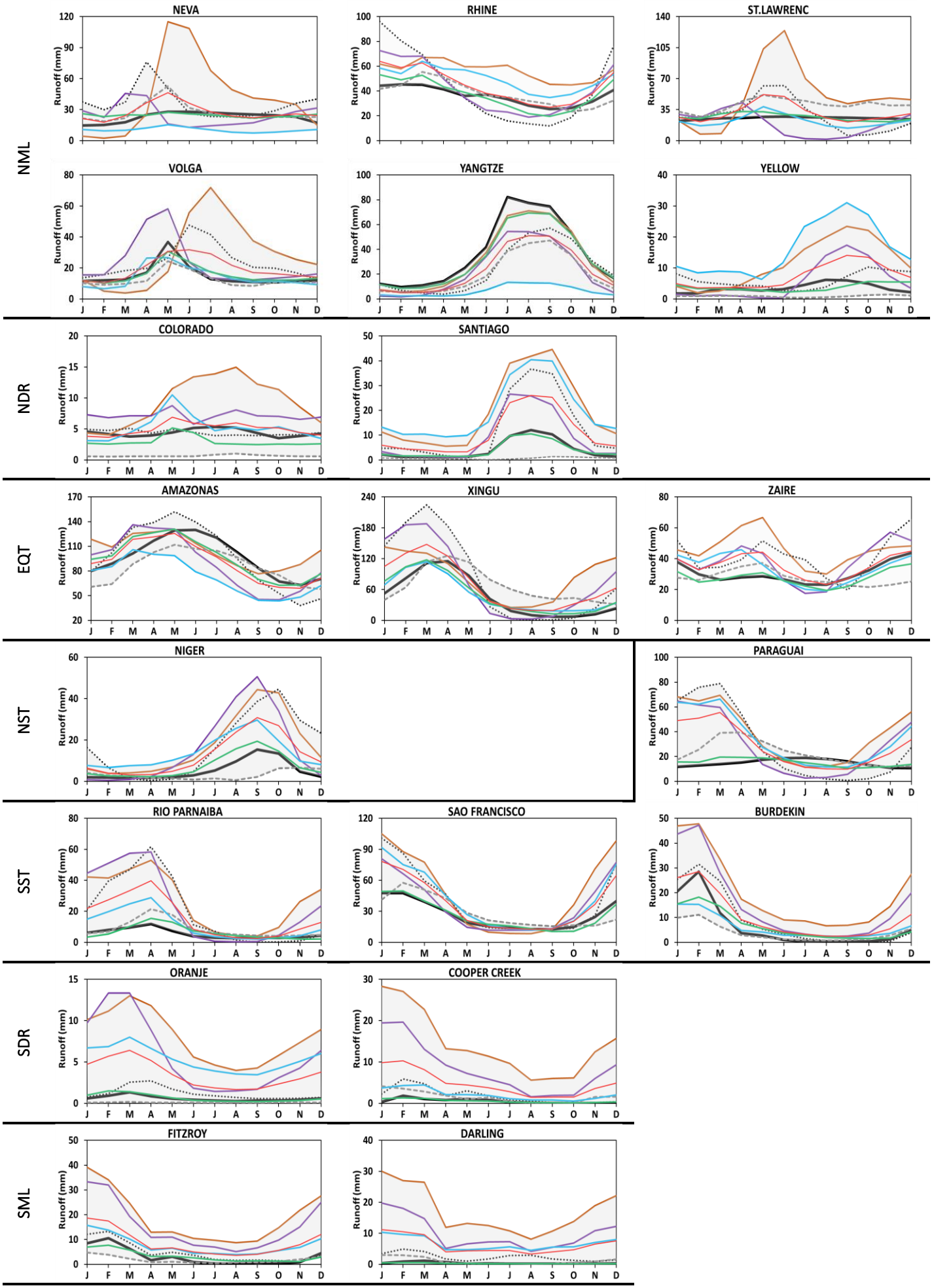

NML
NEVA
RHINE
ST.LAWRENC
VOLGA
YANGTZE
YELLOW
NDR
COLORADO
SANTIAGO
EQT
AMAZONAS
XINGU
ZAIRE
NST
NIGER
PARAGUAI
SST
RIO PARNAIBA
SAO FRANCISCO
BURDEKIN
SDR
ORANJE
COOPER CREEK
SML
FITZROY
DARLING
Runoff (mm)
J F M A M J J A S O N D

Figure S2 (Cont.). Seasonal cycle (long-term mean monthly runoff, MMR). The shaded region denotes the models' performance range (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML= southern mid-latitude, SDR= southern dry and SST= southern subtropical).

2.2. Test of the return periods fitting method

To investigate the reason for the patterns observed in runoff of different return periods, especially decreases in performance at larger return intervals (shown in Section 3.3 and discussed in Section 4.1 of the main text), we tested the Gumbel fitting method applied for maximum runoff. To this end, we examined two modified GRDC datasets for the Zaire catchment and compared the deviations of these datasets with the original data under the different return periods. The modified datasets are:

- Mod1: increased GRDC time series with 5
- Mod2: multiplied GRDC time series with 1.1 (increase of 10%)

It can be seen from the results in Table S8 that:

1. Absolute differences between Mod1 and GRDC are equal under the different return periods which makes sense since the whole Gumbel-curve shifted to the right.
2. Absolute differences between Mod2 and GRDC increase with increasing RP. This also makes sense because a 10% deviation is a bigger difference in absolute terms when taking over higher return periods
3. Relative difference between Mod1 and GRDC decreases which makes sense because this overall shift of +5 is relatively smaller compared to the higher original discharges at higher RPs.
4. Relative difference between Mod1 and GRDC are equal. This also makes sense because the initial modification was a relative one (+10%).

The results suggest that the reason lies in the distribution of peak runoff from the models which skews the Gumbel-fitted distribution and causes relatively larger deviations from the GRDC observations under higher return periods. In addition, the hydrological models are not only on average higher in their results but also skewed differently compared to the GRDC distribution. As a result, there is larger uncertainty in the tails, and the deviation in the extremes is larger than the deviation in the centre-mass of the probability distribution (Figure S3).

Table S8. Results of the Gumbel fitting method test with two modified datasets for the Zaire catchment.

RP	PEAK_GRDC	PEAK_GRDCmod1 (+5)	PEAK_GRDCmod2 (+10%)	ABS_DIFF_mod1	ABS_DIFF_mod2	REL_DIFF_mod1	REL_DIFF_mod2
RP2	43.2	48.2	47.5	5	4.3	11.6	10
RP5	47.8	52.8	52.5	5	4.8	10.5	10
RP10	50.8	55.8	55.9	5	5.1	9.8	10
RP20	53.7	58.7	59.1	5	5.4	9.3	10
RP25	54.6	59.6	60.1	5	5.5	9.2	10
RP50	57.5	62.5	63.2	5	5.7	8.7	10
RP1000	60.3	65.3	66.3	5	6.0	8.3	10
RP500	66.8	71.8	73.5	5	6.7	7.5	10
RP1000	69.6	74.6	76.5	5	7.0	7.2	10

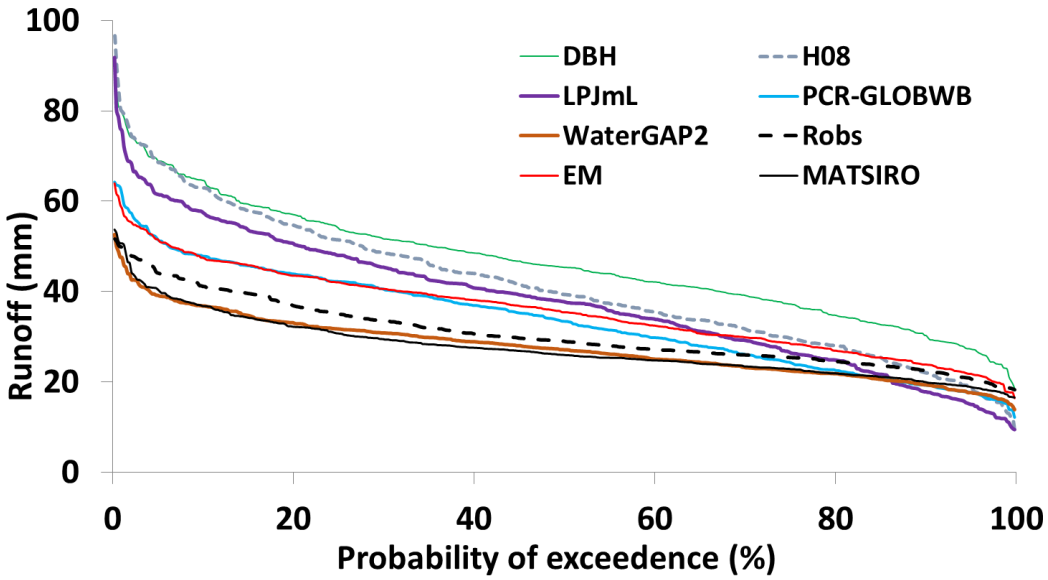


Figure S3. The plot of Exceedence probability observed and modelled data in the Zaire catchment.

Section 3: Supplementary discussion

In this section we expand on some of the themes discussed in the paper.

3.1. On the quality of forcing data

In the discussion in the main text we acknowledge that model improvement is one way of improving model representation of the global hydrological cycle. Another is better observations of climate and hydrology. Our assessed performance of the models is to an extent a function of the quality of the observed runoff data used to evaluate the models. We used data from the GRDC for this purpose. To reduce bias in model performance that may be introduced by the inclusion of broken observed runoff timeseries (e.g. long gaps in the timeseries) and the use of short timeseries (e.g. less than 25 years), we applied a rigorous gauge selection criteria (see SI for the criteria). The performance of the models is also a function of the quality of the input data used to force them. We chose to use the latest input dataset available, GSWP3, which has been used in several other recent global-scale hydrological modelling studies and model intercomparison projects [17-20, 22, 40] and which has been shown to sit well within the range of observations from other observed products [21, 41]. To this end it is unlikely that the overestimation of MAR and all indicators of upper and lower extreme runoff by the majority of models in our study can be attributed to forcing the models with overestimates of observed precipitation. Nevertheless, we acknowledge that some aspects of model performance will be affected by the quality of the forcing data [42-44], with precipitation and radiation being strong drivers of runoff and storage, and temperature having a strong effect on snow and ice [44]. Therefore improvements in model performance will come with improvements in observations. High resolution remotely sensed datasets offer a complementary and timely source for use as input and evaluation data in future studies [45].

3.2. On the calibration of global-scale hydrological models

The generally superior performance of WaterGAP2 relative to the other models suggests that robust calibration methods may improve model performance. However, a smaller error in runoff estimates for a present-day hydrological simulation (from an uncalibrated or calibrated model) does not necessarily mean that projections for the future from that model will be more certain than projections from a model with larger error. Whilst the catchment-scale calibration procedure for WaterGAP2 does lead to better performance for present-day hydrology in many cases, we note that the calibration of any hydrological model can encourage over-fitting and so act to compensate for structural errors and errors in atmospheric forcing data [42]. This could pose a problem if the models were used within a climate modelling framework. Therefore, whilst this study highlights the benefits of model calibration for improving simulations of present-day runoff, we also urge caution towards interpreting it as a definitive route to higher model certainty; especially when the models are used for delivering future projections. Indeed, the calibration of WaterGAP2 does not noticeably improve its ability to simulate the timing of seasonal

runoff relative to other models. This is most likely a result of it being calibrated against long-term mean annual discharge [42] and it highlights that model calibration does not necessarily result in a better performing model for all aspects of the hydrological regime. Calibration can deliver some improvements in performance, but others can only be achieved through better quality input datasets, the inclusion of missing physical processes, and better representation of existing processes in the models.

Catchment-scale models have been shown to perform relatively better than their global-scale counterparts, largely as a result of the former being fully calibrated [13]. Calibration of global models might help to reduce this performance gap but the sensitivity of simulated runoff to changes in precipitation will remain similar between global and regional models, irrespective of whether they are calibrated [13-15]. However, we emphasise that the paucity of long-term globally distributed observations of discharge precludes complete calibration of GHMs and LSMs, and global monitoring studies are needed to address this.

3.3. On the representation of human impacts in global-scale hydrological models

Our study is the first to focus on evaluating an ensemble of hydrological models that include representation of human impacts (table 1). This decision was informed by a parallel study that we conducted [46] that showed inclusion of human impact parameterisation in the models improved simulations of historical runoff, relative to excluding human impact parameterisations. Our results show that even when including human impacts, there remain challenges with the accurate simulation of low runoff magnitudes and long term MMR. Therefore we provide timely encouragement to the global hydrological modelling community to dedicate greater effort to the representation of low runoff, and the magnitude and timing of the seasonal cycle. This in turn should lead to better performing models, particularly in the southern hydrobelts (SML and SDR) where the models generally do not perform well. In particular, we recommend a comprehensive and systematic evaluation of the specific methods used to represent human impacts in models, for multiple rivers across the globe and at different locations along the river network, including: irrigation; dam operation rules; the representation of reservoirs; the way in which withdrawn water is returned to the river network; and the sources where water is withdrawn from to meet water needs (i.e. groundwater, surface water, and desalinisation).

References

[1] WMO 2006 Technical Regulations, Volume III: Hydrology *Available online at: http://librarywmoint/pmb_ged/wmo_49-v3-2006_enpdf:*

[2] Meybeck M, Kummu M and Dürr H H 2013 Global hydrobelts and hydroregions: improved reporting scale for water-related issues? *Hydrology and Earth System Sciences* **17(3)**: 1093-111

[3] Milly P C, Dunne K A and Vecchia A V 2005 Global pattern of trends in streamflow and water availability in a changing climate *Nature* **438(7066)**: 347-50

[4] Beck H E, de Roo A and van Dijk A I J M 2015 Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments *Journal of Hydrometeorology* **16(4)**: 1478-501

[5] Beck H E, van Dijk A I J M, de Roo A, Dutra E, Fink G, Orth R and Schellekens J 2016 Global evaluation of runoff from ten state-of-the-art hydrological models *Hydrology and Earth System Sciences Discussions*: 1-33

[6] Gosling S, et al. 2017 ISIMIP2a Simulation Data from Water (global) Sector. GFZ Data Services.

[7] Tang Q H, Oki T, Kanae S and Hu H P 2007 The influence of precipitation variability and partial irrigation within grid cells on a hydrological simulation *Journal of Hydrometeorology* **8(3)**: 499-512

[8] Hanasaki N, Kanae S, Oki T, Masuda K, Motoya K, Shirakawa N, Shen Y and Tanaka K 2008 An integrated model for the assessment of global water resources Part 1: Model description and input meteorological forcing *Hydrology and Earth System Sciences* **12(4)**: 1007-25

[9] Rost S, Gerten D, Bondeau A, Lucht W, Rohwer J and Schaphoff S 2008 Agricultural green and blue water consumption and its influence on the global water system *Water Resources Research* **44(9)**: 1-17

[10] Pokhrel Y N, Koirala S, Yeh P J F, Hanasaki N, Longuevergne L, Kanae S and Oki T 2015 Incorporation of groundwater pumping in a global Land Surface Model with the representation of human impacts *Water Resources Research* **51(1)**: 78-96

[11] Wada Y, Wisser D and Bierkens M F P 2014 Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources *Earth System Dynamics* **5(1)**: 15-40

[12] Müller Schmied H, et al. 2016 Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use *Hydrology and Earth System Sciences* **20(7)**: 2877-98

[13] Hattermann F F, et al. 2017 Cross - scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins *Climatic Change* **141(3)**: 561-76

[14] Gosling S N, et al. 2016 A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C *Climatic Change* **141(3)**: 577-95

[15] Hattermann F F, et al. 2018 Sources of uncertainty in hydrological climate impact assessment: a cross-scale study *Environmental Research Letters* **13(1)**

[16] Compo G P, et al. 2011 The Twentieth Century Reanalysis Project *Q J Roy Meteor Soc* **137(654)**: 1-28

[17] Masaki Y, Hanasaki N, Biemans H, Schmied H M, Tang Q, Wada Y, Gosling S N, Takahashi K and Hijioka Y 2017 Intercomparison of global river discharge simulations focusing on dam operation—multiple models analysis in two case-study river basins, Missouri–Mississippi and Green–Colorado *Environmental Research Letters* **12(5)**: 055002

[18] Veldkamp T I E, et al. 2017 Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century *Nat Commun* **8**: 15697

[19] Liu X, Tang Q, Cui H, Mu M, Gerten D, Gosling S N, Masaki Y, Satoh Y and Wada Y 2017 Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations *Environmental Research Letters* **12(2)**: 025009

[20] Tangdamrongsub N, Han S-C, Decker M, Yeo I-Y and Kim H 2018 On the use of the GRACE normal equation of inter-satellite tracking data for estimation of soil moisture and groundwater in Australia Hydrology and Earth System Sciences **22(3)**: 1811-29

[21] van den Hurk B, et al. 2016 LS3MIP (v1.0) contribution to CMIP6: the Land Surface, Snow and Soil moisture Model Intercomparison Project – aims, setup and expected outcome *Geoscientific Model Development* **9(8)**: 2809-32

[22] Guimberteau M, et al. 2018 ORCHIDEE-MICT (v8.4.1), a land surface model for the high latitudes: model description and validation *Geoscientific Model Development* **11(1)**: 121-63

[23] Decharme B 2007 Influence of runoff parameterization on continental hydrology: Comparison between the Noah and the ISBA land surface models *Journal of Geophysical Research-Atmospheres* **112(D19)**

[24] Zhou X Y, Zhang Y Q, Wang Y P, Zhang H Q, Vaze J, Zhang L, Yang Y H and Zhou Y C 2012 Benchmarking global land surface models against the observed mean annual runoff from 150 large basins *Journal of Hydrology* **470**: 269-79

[25] Zhang Y Q, Zheng H X, Chiew F H S, Pena-Arancibia J and Zhou X Y 2016 Evaluating Regional and Global Hydrological Models against Streamflow and Evapotranspiration Measurements *Journal of Hydrometeorology* **17(3)**: 995-1010

[26] Dawson C W, Abrahart R J and See L M 2007 HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts *Environmental Modelling & Software* **22**: 1034-52

[27] Pushpalatha R, Perrin C, Moine N L and Andréassian V 2012 A review of efficiency criteria suitable for evaluating low-flow simulations *Journal of Hydrology* **420–421**: 171–82

[28] Bennett N D, et al. 2013 Characterising performance of environmental models *Environmental Modelling & Software* **40**: 1-20

[29] Materia S, Dirmeyer P A, Guo Z C, Alessandri A and Navarra A 2010 The Sensitivity of Simulated River Discharge to Land Surface Representation and Meteorological Forcings *Journal of Hydrometeorology* **11(2)**: 334-51

[30] Yang H, et al. 2015 Multicriteria evaluation of discharge simulation in Dynamic Global Vegetation Models *Journal of Geophysical Research-Atmospheres* **120(15)**: 7488-505

[31] Gumbel E J 1941 The Return Period of Flood Flows **12**: 163-90

[32] Stedinger J R, Vogel R M and Foufoula-Georgiou E. Frequency analysis of extreme events. In: Maidment DR, editor. Handbook of Hydrology. New York: McGraw-Hill; 1993. p. 18.1- .66

[33] Katz R W, Parlange M B and Naveau P 2002 Statistics of extremes in hydrology *Advances in Water Resources* **25(8–12)**: 1287-304

[34] Leadbetter M R, Lindren G and Rootzen H. 1983 *Extremes and related properties of random sequences and processes*. New York, USA: Springer-Verlag. 336 p

[35] Dettinger M D and Diaz H F 2000 Global Characteristics of Stream Flow Seasonality and Variability *Journal of Hydrometeorology* **1(4)**: 289-310

[36] Ward P J, Beets W, Bouwer L M, Aerts J C J H and Renssen H 2010 Sensitivity of river discharge to ENSO *Geophysical Research Letters* **37(12)**: L12402

[37] Ward P J, Eisner S, Flörke M, Dettinger M D and Kummu M 2014 Annual flood sensitivities to El Niño–Southern Oscillation at the global scale *Hydrol Earth Syst Sci* **18(1)**: 47-66

[38] WMO. Manual on low-flow estimation and prediction (http://www.wmo.int/pages/prog/hwrp/publications/low-flow_estimation_prediction/WMO%201029%20en.pdf). 2009

[39] Weibull W 1951 Wide applicability *Journal of Applied Mechanics* **13**: 293-7

[40] Zhao F, et al. 2017 The critical role of the routing scheme in simulating peak river discharge in global hydrological models *Environmental Research Letters* **12(7)**

[41] Slater A G 2016 Surface Solar Radiation in North America: A Comparison of Observations, Reanalyses, Satellite, and Derived Products *Journal of Hydrometeorology* **17(1)**: 401-20

[42] Müller Schmied H, Eisner S, Franz D, Wattenbach M, Portmann F T, Flörke M and Döll P 2014 Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration *Hydrology and Earth System Sciences* **18(9)**: 3511-38

[43] Nasonova O N, Gusev Y M and Kovalev Y E 2011 Impact of uncertainties in meteorological forcing data and land surface parameters on global estimates of terrestrial water balance components *Hydrological Processes* **25(7)**: 1074-90

[44] Döll P, Douville H, Güntner A, Müller Schmied H and Wada Y 2015 Modelling Freshwater Resources at the Global Scale: Challenges and Prospects *Surveys in Geophysics* **37(2)**: 195-221

[45] Scanlon B R, et al. 2018 Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data *Proc Natl Acad Sci U S A* **115(6)**: E1080-E9

[46] Veldkamp T I E, et al. 2018 Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study *Environmental Research Letters* **13(5)**: 055008

Appendix B (Supplementary information for Paper II)

Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models

Zaherpour Jamal[a*], Mount Nick[a], N Gosling Simon[a]

Rest of co-authors in alphabetic order:

Dankers Rutger [b], Eisner Stephanie [c], Gerten Dieter [d, e], Liu Xingcai [f], Masaki Yoshimitsu [g], Müller Schmied Hannes [h, i], Tang Qiuhong [f], Wada Yoshihide [j]

[a] School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom

[b] Met Office, FitzRoy Road, Exeter, EX1 3PB, United Kingdom

[c] Center for Environmental Systems Research, University of Kassel, Kassel, Germany

[d] Potsdam Institute for Climate Impact Research, Telegrafenberg, 14473 Potsdam, Germany

[e] Geography Dept., Humboldt-Universität zu Berlin, 10099 Berlin, Germany

[f] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

[g] Hirosaki University, Bunkyocho-3, Hirosaki, Aomori, 36-8561, Japan

[h] Institute of Physical Geography, Goethe-University, Frankfurt, Germany

[i] Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Frankfurt, Germany

[j] International Institute for Applied Systems Analysis (IIASA) - Schlossplatz 1 - A-2361 Laxenburg, Austria

*Corresponding author:
Tel.: +44 115 951 5428, Fax: +44 (0)115 951 5249

E-mail addresses: lgxjz1@nottingham.ac.uk, zaherpour@gmail.com

Postal Address: Sir Clive Granger Building, School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom

Table of contents

Table S1. Attributes of the 40 selected catchments…….. 3
Table S2. GHMs and their major parameters and modelling methods…….. 4
Table S3. Characteristics of hydrobelts..…… 5
Table S4. The GEP function set, constants and gene configuration…….. 5
Table S5. Training and validation subsets used to inform MMC development …… 6
Table S6. Cross-correlation of the model residuals for the six catchments where their trade-off-based MMC solutions fail……. 7
Table S7. Changes in IPE score and performance gain of the best performing MMS solutions for the six catchments where their trade-off-based MMC solutions fail…… 7
Table S8. IPE and its components values for the GHMs, the EM and MMC…….. 8
Panels of plots and tabulated data for on a catchment by catchment basis…… 12
LENA 12
AMUR 13
YENISEI 14
OB 15
KOLYMA 16
OLENEK 17
MACKENZIE RIVER 18
SASKATCHEWAN 19
ASSINIBOINE 20
RED RIVER 21
WINNIPEG RIVER 22
CHURCHILL RIVER 23
ALBANY RIVER 24
NORTHERN DVINA(SEVERNAYA DVINA) 25
YELLOW 26
COLUMBIA 27
MISSISSIPPI 28
ST.LAWRENCE 29
FRASER RIVER 30
LABE 31
RHINE RIVER 32
DANUBE 33
NEVA 34
VOLGA 35
DON 36
YANGTZE 37
COLORADO 38
SANTIAGO 39
NIGER 40
ZAIRE 41
AMAZONAS 42
XINGU 43
RIO PARNAIBA 44
SAO FRANCISCO 45
PARAGUAI 46
BURDEKIN 47
ORANJE 48
COOPER CREEK 49
FITZROY 50
DARLING RIVER 51
References…….. 52

Section S1

Table S1. 40 selected catchments from the GRDC reference database.

No	GRDC_NR	RIVER	STATION	Data availability (yrs)	Area (KM^2) (GRDC)	Area (KM^2) (ISIMIP)	Long_GRDC (degree)	Lat_GRDC (degree)	Hydro-belt
1	2903430	LENA	STOLB	32	2460000	2443190	126.8	72.37	BOR
2	2906900	AMUR	KOMSOMOLSK	26	1730000	1742103	137.12	50.63	BOR
3	2909150	YENISEI	IGARKA	32	2440000	2443587	86.5	67.48	BOR
4	2912600	OB	SALEKHARD	39	2949998	2438012	66.53	66.57	BOR
5	2998510	KOLYMA	KOLYMSKAYA	28	526000	539526	158.72	68.73	BOR
6	2999910	OLENEK	7.5KM DOWNSTREAM OF MOUTH OF RIVER PUR	39	198000	197649	123.22	72.12	BOR
7	4208150	MACKENZIE RIVER	NORMAN WELLS	30	1570000	1580100	-126.85	65.28	BOR
8	4213550	SASKATCHEWAN	THE PAS	40	347000	349670	-101.18	53.83	BOR
9	4213650	ASSINIBOINE	HEADINGLEY	40	153000	154323	-97.4	49.87	BOR
10	4213680	RED RIVER	EMERSON	40	104000	99974	-97.22	49	BOR
11	4213800	WINNIPEG RIVER	SLAVE FALLS	38	126000	124681	-95.57	50.22	BOR
12	4214260	CHURCHILL RIVER	ABOVE GRANVILLE FALLS	36	228000	233423	-100.45	56.15	BOR
13	4214520	ALBANY RIVER	NEAR HAT ISLAND	31	118000	121442	-83.87	51.33	BOR
14	6970250	NORTHERN DVINA(SEVERNAYA DVINA)	UST-PINEGA	31	348000	350844	42.17	64.1	BOR
15	2180800	YELLOW	Huayuankou	40	730036	723657	113.65	34.92	NML
16	4115200	COLUMBIA	THE DALLES, OREG.	40	613830	616452	-121.17	45.61	NML
17	4127800	MISSISSIPPI	VICKSBURG, MISS.	37	2964252	2947486	-90.9	32.32	NML
18	4143550	ST.LAWRENCE	CORNWALL(ONTARIO), NEAR MASSENA, N.Y.	40	773892	771449	-74.78	45.01	NML
19	4207900	FRASER RIVER	HOPE	40	217000	218162	-121.45	49.38	NML
20	6340110	LABE	NEU-DARCHAU	40	131950	132028	10.88	53.23	NML
21	6435060	RHINE RIVER	LOBITH	40	160800	161606	6.11	51.84	NML
22	6442600	DANUBE	MOHACS	29	209064	212558	18.67	46	NML
23	6972430	NEVA	NOVOSARATOVKA	40	281000	284945	30.72	59.8	NML
24	6977100	VOLGA	VOLGOGRAD POWER PLANT	39	1360000	1356910	44.72	48.77	NML
25	6978250	DON	RAZDORSKAYA	38	378000	364497	40.67	47.5	NML
26*	7222222	YANGTZE	Cuntan	31	121000	117543	100.15	35.5	NML
27	4152450	COLORADO	LEES FERRY, ARIZ.	40	289562	358663	-111.58	36.87	NDR
28	4356100	SANTIAGO	EL CAPOMAL	31	128943	129623	-105.12	21.83	NDR
29	1834101	NIGER	LOKOJA	25	2074171	1761002	6.77	7.8	NST
30	1147010	ZAIRE	KINSHASA	40	3475000	3637502	15.3	-4.3	EQT
31	3629000	AMAZONAS	OBIDOS	27	4640300	4678378	-55.5	-1.9	EQT
32	3630050	XINGU	ALTAMIRA	35	446570	442067	-52.22	-3.2	EQT
33	3650481	RIO PARNAIBA	LUZILANDIA	26	322823	309374	-42.37	-3.45	SST
34	3651805	SAO FRANCISCO	MANGA	37	200789	205950	-43.93	-14.76	SST
35	3667060	PARAGUAI	PORTO MURTINHO (FB/DNOS)	37	474500	487030	-57.89	-21.7	SST
36	5101200	BURDEKIN	CLARE	40	129660	129785	147.24	-19.76	SST
37	1159100	ORANJE	VIOOLSDRIF	38	850530	851280	17.73	-28.76	SDR
38	5410100	COOPER CREEK	CALLAMURRA	33	230000	230043	140.87	-27.7	SDR
39	5101301	FITZROY	THE GAP	40	135860	132642	150.1	-23.1	SML
40	5204250	DARLING RIVER	LOUTH	26	489300	472878	145.11	-30.54	SML

*not included in GRDC database, obtained from local authority

Table S2. The 5 GHMs and their major parameters and modelling methods. tas = surface air temperatures, pr = precipitation, rhs= near-surface relative humidity, rsds/rlds = surface radiation (shortwave/longwave downwelling), wind = near-surface wind speed, pet = potential evapotranspiration, ps = surface air pressure.

Model	Class of Model	Input climate parameters	Resolution (space and time)	Representation of soils	Representation of vegetation	Method: potential evapotranspiration	Method: snow melt	Human water use/ Reservoirs	Carbon cycling	Calibration status
DBH (Tang et al., 2007)	GHM	Pr, tas, wind,rhs, rlds, rsds, ps	Grid cells with sub-grid heterogeneity accounting method	Three soil layers and one underlying groundwater layer at the bottom	Prescribed spatial distribution of natural vegetation and agricultural land cover	Energy balance	Energy balance	Yes/No	No	No
H08 (Hanasaki et al., 2008)	GHM	tas, Pr, Snowfall, wind, rhs, rsds, rlds, ps	0.5°; daily	1-layer leaky bucket soil. Its runoff properties vary with climate zones.	Natural use: Globally uniform. No-specific land type is assigned, as known as Manabe's bucket.	Bulk formula	Energy balance	Yes/Yes	No	No
LPJmL (Bondeau et al., 2007)	GHM	Pr, tas, rsds	0.5°; daily	Five hydrologically active soil layers, coupled to carbon and thermal balance	Dynamic simulation of growth and productivity (with prescribed spatial distribution of crops and pasture); daily	Priestley-Taylor (modified for transpiration)	Degree-day method with precipitation factor	Irrigation only/Yes	Yes	No
PCRGLOBWB (Wada et al., 2014)	GHM	tas, Pr, pet	0.5°; daily	Two soil layers and one underlying groundwater layer at the bottom	Prescribed vegetation, agriculture, and land use cover	Hamon	Degree-day method	Yes/Yes	No	No
WaterGAP2 (Müller Schmied et al., 2016)	GHM	tas, Pr, rsds, rlds	0.5°; daily	One soil layer with varying rooting depth, dependent on land cover	IGBP land cover classes based on MODIS, temperature and precipitation based LAI-model, fixed rooting depth	Priestley-Taylor with two alpha factors depending on the aridity of the grid cell	Degree-day method	Yes/Yes	No	Yes

Table S3. Characteristics of hydrobelts and number of catchments within each hydrobelt (BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML=southern mid-latitude, SDR=southern dry and SST=southern subtropical).

Hydrobelt	Area of hydrobelt (10^3 km^2) (Meybeck et al., 2013)	Hydrobelt runoff (mm/yr) (Meybeck et al., 2013)	No. of selected catchments in each hydrobelt	Catchments area in each hydrobelt (10^3 km2)	Catchments area to hydrobelt area (%)
BOR	25995	223	14	13298	51.2
NML	24199	343	12	7940.8	32.8
NDR	30234	36	2	418.5	1.4
NST	10579	386	1	2074.2	19.6
EQT	16826	960	3	8561.9	50.9
SST	10599	233	4	1127.8	10.6
SDR	8677	31	2	1080.5	12.5
SML	4008	292	2	625.2	15.6
Sum/Ave.	131119*	277*	40	4390.9	24.3

*Total of non-glaciated land

Table S4. The GEP function set, constants and gene configuration. For detail explanations about control parameters see (Ferreira, 2001).

Settings	Control parameter	Value/Symbol
General settings	Number of chromosome	25
	Number of genes (equation components)	3
	Gene head size	10
	Linking function	addition
Functions, Fitness and model generation	Maximum generations	100,000
	Fitness function	Customised IPE
	Max fitness	1000
	Functions	+, -, *, / (2)*; Sqrt, Exp, x^2, x^3, Natural Log, Sine, Cosine (1)*;
Genetic operators	Mutation rate	0.044
	Inversion	0.1
	Insertion sequence (IS) transposition	0.1
	Root insertion sequence (RIS) transposition	0.1
	One-point recombination rate	0.3
	Two-point recombination rate	0.3
	Gene recombination	0.1
	Gene transposition	0.1
Numerical constants	Constants per gene	2
	Lower/upper bound	-10/+10
	Random numerical constants (RNC) mutation	0.01
	DC mutation	0.044
	DC inversion	0.1
	DC IS transposition	0.1

*Weight of the functions for calculating the "equation size"

Table S5. The training and validation subsets used to inform MMC development in each of the study catchments.

No	River	Total data length (months)	Period of Training Data	Period of Validation Data
1	LENA	384	1/1971-12/1990	1/1991-12/2002
2	AMUR	312	1/1971-12/2000	1/2001-12/2006
3	YENISEI	384	1/1975-12/1998	1/1999-12/2010
4	OB	468	1/1971-12/1990	1/1991-12/2010
5	KOLYMA	336	1/1978-12/1997	1/1998-12/2008
6	OLENEK	468	1/1971-12/1990	1/1992-12/2010
7	MACKENZIE RIVER	360	1/1971-12/1992	1/1993-12/2010
8	SASKATCHEWAN	480	1/1971-12/1990	1/1991-12/2010
9	ASSINIBOINE	480	1/1971-12/1990	1/1991-12/2010
10	RED RIVER	480	1/1971-12/1990	1/1991-12/2010
11	WINNIPEG RIVER	456	1/1971-12/1990	1/1991-12/2010
12	CHURCHILL RIVER	432	1/1971-12/1994	1/1995-12/2010
13	ALBANY RIVER	372	1/1973-12/1992	1/1993-12/2010
14	NORTHERN DVINA	372	1/1971-12/1992	1/1993-12/2005
15	YELLOW	480	1/1971-12/1990	1/1991-12/2010
16	COLUMBIA	480	1/1971-12/1990	1/1991-12/2010
17	MISSISSIPPI	444	1/1971-12/1990	1/1991-12/2010
18	ST.LAWRENCE	480	1/1971-12/1990	1/1991-12/2010
19	FRASER RIVER	480	1/1971-12/1990	1/1991-12/2010
20	LABE	480	1/1971-12/1990	1/1991-12/2010
21	RHINE RIVER	480	1/1971-12/1990	1/1991-12/2010
22	DANUBE	348	1/1971-12/1990	1/1991-12/1999
23	NEVA	480	1/1971-12/1990	1/1991-12/2010
24	VOLGA	468	1/1971-12/1990	1/1992-12/2010
25	DON	456	1/1971-12/1990	1/1993-12/2010
26	YANGTZE	372	1/1971-12/1990	1/1991-12/2001
27	COLORADO	480	1/1971-12/1990	1/1991-12/2010
28	SANTIAGO	372	1/1971-12/1990	1/1991-12/2003
29	NIGER	300	1/1971-12/1990	1/1991-12/2005
30	ZAIRE	480	1/1971-12/1990	1/1991-12/2010
31	AMAZONAS	324	1/1971-12/1990	1/1991-12/1997
32	XINGU	420	1/1971-12/1991	1/1992-12/2008
33	RIO PARNAIBA	300	1/1982-12/2001	1/2001-12/2006
34	SAO FRANCISCO	444	1/1971-12/1990	1/1991-12/2008
35	PARAGUAI	444	1/1971-12/1990	1/1991-12/2007
36	BURDEKIN	480	1/1971-12/1990	1/1991-12/2010
37	ORANJE	456	1/1971-12/1992	1/1993-12/2010
38	COOPER CREEK	396	1/1971-12/1993	1/1991-12/2006
39	FITZROY	480	1/1971-12/1990	1/1991-12/2010
40	DARLING RIVER	312	1/1971-12/2001	1/2002-12/2007

Table S6. Cross-correlation of the model residuals for the six catchments where their trade-off-based MMC solutions fail.

OLENEK					
	DBH	H08	LPJmL	PCRGLOBWB	WaterGAP2
DBH	1				
HO8	0.953627	1			
LPJmL	0.398167	0.461829	1		
PCRGLOBWB	0.946451	0.965436	0.519173	1	
WaterGAP2	0.81903	0.872276	0.375587	0.817426	1
MACKENZIE					
	DBH	H08	LPJmL	PCRGLOBWB	WaterGAP2
DBH	1				
HO8	0.491979	1			
LPJmL	-0.62203	-0.55728	1		
PCRGLOBWB	0.133255	0.455914	0.135442	1	
WaterGAP2	-0.42352	0.191301	0.397815	0.780964	1
WINNIPEG					
	DBH	H08	LPJmL	PCRGLOBWB	WaterGAP2
DBH	1				
HO8	-0.00723	1			
LPJmL	-0.24339	0.676216	1		
PCRGLOBWB	0.739637	0.183786	0.057652	1	
WaterGAP2	-0.19368	0.240907	0.343589	0.32813	1
COLUMBIA					
	DBH	H08	LPJmL	PCRGLOBWB	WaterGAP2
DBH	1				
HO8	-0.52298	1			
LPJmL	-0.77634	0.695721	1		
PCRGLOBWB	0.000928	0.183148	0.283918	1	
WaterGAP2	-0.10828	0.565067	0.58462	0.509789	1
LABE					
	DBH	H08	LPJmL	PCRGLOBWB	WaterGAP2
DBH	1				
HO8	-0.05791	1			
LPJmL	-0.08669	0.744627	1		
PCRGLOBWB	0.555102	0.097086	0.350186	1	
WaterGAP2	0.095586	0.548779	0.510525	0.128655	1
PARAGUAI					
	DBH	H08	LPJmL	PCRGLOBWB	WaterGAP2
DBH	1				
HO8	0.817754	1			
LPJmL	0.96891	0.819339	1		
PCRGLOBWB	0.938688	0.925835	0.956784	1	
WaterGAP2	0.702871	0.657761	0.70128	0.724402	1

Table S7. Changes in IPE score and performance gain of the best performing MMS solutions for the six catchments where their trade-off-based MMC solutions fail.

	IPE				PG relative to best GHM (%)		PG relative to EM (%)	
	Best GHM	**EM**	**original MMC**	**Best MMC**	**original MMC**	**Best MMC**	**original MMC**	**Best MMC**
Mackenzie	-1.30	-1.39	-1.33	-1.88	-3	-58	5	-49
Columbia	-1.11	-1.28	-1.21	-1.22	-10	-12	7	5
Olenek	-1.47	8.12	-1.15	-1.36	33	11	-1027	-1048
Paraguai	8.00	78.53	8.19	8.00	19	-0.4	-7034	-7053
Winnipeg	1.55	2.29	1.62	1.60	7	5	-67	-69
Labe	-1.47	3.10	-1.44	-1.65	2	-19	-554	-575

Section 2

Table S8. IPE and its components values for mean annual runoff for GHMs, EM and MMC

Catchment No.	Metrics	Calibration							Validation						
		DBH	H08	LPJmL	PCR_GLOBWB	WaterGap2	EM	MMC	DBH	H08	LPJmL	PCR_GLOBWB	WaterGap2	EM	MMC
1	RMSE	26.6	21.7	20.7	15.0	14.1	16.0	11.1	24.4	20.2	18.9	15.0	11.9	14.9	9.4
	MARE	1.8	2.0	1.8	1.7	1.0	1.4	0.6	1.4	2.1	1.4	1.6	0.8	1.2	0.5
	CE	-0.7	-0.1	0.0	0.5	0.5	0.4	0.7	-0.4	0.0	0.2	0.5	0.7	0.5	0.8
	IPE	1.1	-1.3	-1.3	-1.7	-2.0	-1.7	-3.3	1.6	2.0	1.4	1.5	-1.3	1.1	-2.0
2	RMSE	16.5	11.4	9.8	7.5	6.9	6.2	6.1	14.6	8.8	8.0	6.1	5.8	5.6	4.8
	MARE	0.7	1.1	0.7	0.9	0.7	0.6	0.3	0.8	1.2	0.7	0.9	0.8	0.7	0.4
	CE	-1.0	0.0	0.3	0.6	0.6	0.7	0.7	-1.5	0.1	0.3	0.6	0.6	0.6	0.7
	IPE	1.2	-1.3	-1.7	-1.7	-2.0	-2.5	-3.3	3.1	1.9	1.3	1.3	1.2	1.1	-1.4
3	RMSE	23.4	14.1	25.8	19.0	14.6	16.6	13.4	20.1	12.3	23.4	16.1	12.2	14.0	11.4
	MARE	0.8	0.5	0.8	0.4	0.4	0.4	0.2	0.7	0.4	0.7	0.3	0.3	0.3	0.2
	CE	0.0	0.6	-0.3	0.3	0.6	0.5	0.7	0.0	0.6	-0.4	0.3	0.6	0.5	0.7
	IPE	-1.4	-2.5	-1.3	-2.0	-2.5	-2.5	-3.3	1.2	-1.7	1.2	-1.4	-1.7	-1.7	-2.5
4	RMSE	21.3	12.8	32.1	10.2	6.0	11.0	4.4	20.9	14.0	27.0	8.9	6.7	10.7	3.9
	MARE	1.6	1.3	3.6	1.5	1.1	1.6	0.3	1.6	1.7	2.8	1.2	1.3	1.5	0.3
	CE	-5.2	-1.2	-13.0	-0.4	0.5	-0.7	0.7	-6.8	-2.5	-12.0	-0.4	0.2	-1.1	0.7
	IPE	2.4	1.2	5.2	1.0	-1.7	1.1	-3.3	8.4	4.7	13.9	2.6	2.5	3.5	-1.3
5	RMSE	21.1	17.7	20.3	14.8	18.7	13.4	6.3	17.6	14.1	20.2	13.7	15.8	10.8	7.4
	MARE	0.7	1.9	1.2	2.1	3.5	1.5	0.4	0.7	1.1	1.1	1.4	2.5	0.9	0.4
	CE	0.3	0.5	0.3	0.6	0.4	0.7	0.9	0.4	0.6	0.2	0.6	0.5	0.8	0.9
	IPE	-1.4	-1.4	-1.4	-1.7	1.0	-2.0	-5.0	-1.3	1.1	1.2	1.3	2.3	-1.3	-2.5
6	RMSE	26.1	20.9	24.1	25.2	7.5	16.7	7.3	30.5	26.5	23.9	30.1	13.2	21.4	13.2
	MARE	1.9	15.8	33.6	40.0	9.7	20.0	3.8	1.4	19.7	38.9	55.3	12.3	25.3	2.7
	CE	0.2	0.5	0.3	0.2	0.9	0.7	0.9	0.2	0.4	0.5	0.2	0.8	0.6	0.8
	IPE	-1.4	-1.4	1.1	1.3	-3.3	-1.4	-5.0	-1.4	6.3	12.5	17.7	3.9	8.1	-1.1
7	RMSE	18.5	8.3	17.6	5.6	6.2	4.1	4.5	17.0	8.7	14.3	5.4	5.0	4.0	4.9
	MARE	1.0	0.6	0.9	0.3	0.5	0.3	0.2	1.0	0.7	0.8	0.2	0.4	0.3	0.3
	CE	-2.5	0.3	-2.2	0.7	0.6	0.8	0.8	-2.7	0.0	-1.6	0.6	0.7	0.8	0.7
	IPE	1.5	-1.7	1.4	-3.3	-2.5	-3.3	-3.3	4.5	1.9	3.4	-1.3	1.1	-1.4	-1.4
8	RMSE	16.2	4.8	15.0	5.9	2.1	6.5	1.8	16.3	4.6	10.8	5.8	2.0	6.0	1.6
	MARE	3.0	1.2	2.2	1.1	0.4	1.4	0.3	2.7	1.1	1.8	0.9	0.3	1.2	0.3
	CE	-51.2	-3.7	-43.6	-6.0	0.1	-7.4	0.4	-51.8	-3.1	-21.9	-5.6	0.2	-6.1	0.5
	IPE	19.7	2.4	16.8	3.1	-1.4	3.7	-2.0	61.4	5.7	27.0	8.2	1.4	9.0	1.0
9	RMSE	14.3	6.9	28.6	4.1	1.0	8.9	0.8	18.1	6.0	20.9	4.8	0.8	8.5	0.7
	MARE	22.6	9.6	24.5	6.1	0.9	12.7	0.5	17.8	7.0	17.9	5.1	0.8	9.7	0.5
	CE	-134.0	-30.5	-537.0	-10.3	0.3	-51.3	0.6	-387.0	-42.4	-516.0	-26.9	0.3	-84.3	0.5
	IPE	68.1	16.4	267.8	6.4	-1.4	26.8	-2.5	384.8	44.5	512.3	28.9	1.6	85.8	1.1
10	RMSE	17.7	6.7	16.0	6.0	4.1	7.9	2.8	19.9	5.9	16.4	7.4	6.7	8.3	5.3
	MARE	11.3	4.2	9.8	3.7	2.0	6.2	0.9	5.7	1.8	4.4	2.3	1.6	3.1	0.7
	CE	-14.5	-1.2	-11.7	-0.8	0.2	-2.1	0.6	-6.9	0.3	-4.3	-0.1	0.1	-0.4	0.4
	IPE	8.1	1.5	6.7	1.3	-1.4	2.1	-2.5	6.6	1.6	4.8	2.1	1.5	2.8	-1.3

Table S8 (Cont.). IPE and its components values for mean annual runoff for GHMs, EM and MMC

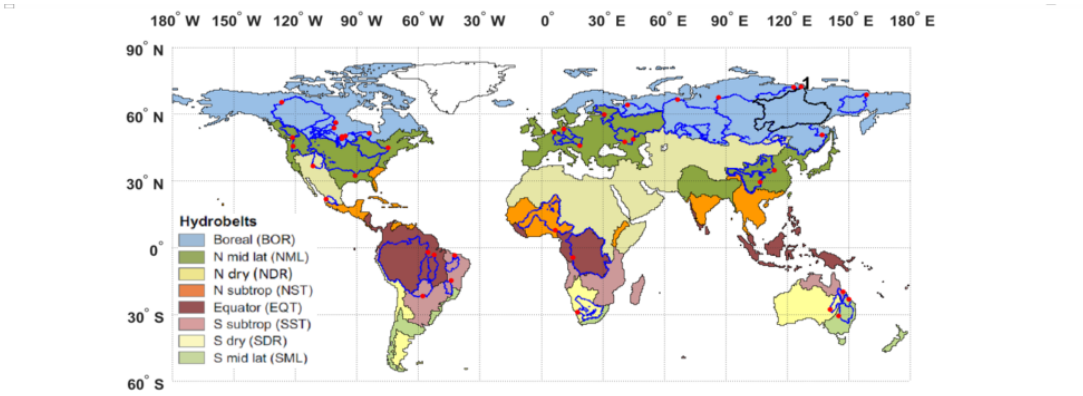
Catchment No.	Metrics	Calibration							Validation						
		DBH	H08	LPJmL	PCR_GLOBWB	WaterGap2	EM	MMC	DBH	H08	LPJmL	PCR_GLOBWB	Water-ap2	EM	MMC
11	RMSE	37.2	15.9	17.9	8.4	7.0	11.5	7.0	37.1	16.0	16.3	7.5	8.0	9.5	7.6
	MARE	2.4	0.7	0.7	0.4	0.4	0.7	0.3	1.8	0.5	0.6	0.3	0.3	0.5	0.4
	CE	-16.9	-2.3	-3.2	0.1	0.4	-0.7	0.4	-15.8	-2.1	-2.2	0.3	0.2	-0.1	0.3
	IPE	6.1	1.4	1.6	-1.7	-2.0	1.0	-2.0	24.2	4.8	5.0	1.6	1.7	2.3	1.6
12	RMSE	35.8	12.9	9.7	9.5	2.4	7.8	2.2	33.4	13.6	10.9	9.6	3.3	7.8	3.0
	MARE	3.0	1.2	0.7	1.0	0.2	0.7	0.2	2.8	1.3	0.6	1.0	0.3	0.7	0.3
	CE	-187.8	-23.6	-12.8	-12.4	0.2	-7.8	0.3	-89.9	-14.1	-8.7	-6.5	0.1	-4.0	0.3
	IPE	52.0	7.3	4.2	4.2	-2.0	2.9	-2.0	297.5	50.1	32.2	25.6	3.6	17.1	3.1
13	RMSE	47.4	22.6	50.9	10.6	22.0	13.6	9.6	40.8	15.6	39.6	14.5	21.2	11.1	13.4
	MARE	1.5	1.2	2.4	0.5	1.6	1.1	0.4	1.2	0.8	1.4	0.6	1.6	0.8	0.4
	CE	-4.0	-0.1	-4.7	0.8	-0.1	0.6	0.8	-3.0	0.4	-2.8	0.5	-0.1	0.7	0.6
	IPE	2.0	-1.4	2.3	-3.3	-1.4	-2.5	-5.0	2.8	1.0	2.8	-1.4	1.7	-1.3	-1.7
14	RMSE	34.9	17.4	51.1	29.1	17.7	18.4	13.4	36.7	14.1	48.6	30.4	18.3	18.7	13.7
	MARE	1.1	1.1	2.3	0.6	0.6	0.7	0.4	1.2	0.9	1.7	0.6	0.6	0.5	0.4
	CE	-0.5	0.6	-2.1	0.0	0.6	0.6	0.8	-0.4	0.8	-1.4	0.1	0.7	0.6	0.8
	IPE	1.0	-1.7	1.9	-1.4	-2.5	-2.0	-3.3	1.5	1.0	2.1	-1.1	-1.4	-1.7	-2.5
15	RMSE	11.6	5.5	5.9	15.5	3.7	6.9	2.7	10.7	4.8	5.8	14.9	2.1	6.7	1.6
	MARE	2.4	1.4	1.0	3.6	0.8	1.6	0.3	4.0	2.1	1.8	6.0	0.7	2.7	0.7
	CE	-11.3	-1.7	-2.2	-21.2	-0.2	-3.3	0.4	-43.2	-7.8	-12.2	-85.0	-0.7	-16.3	0.0
	IPE	5.4	1.7	1.7	9.3	-1.1	2.3	-2.0	23.4	5.5	7.4	44.9	1.5	9.7	1.2
16	RMSE	18.7	11.4	14.0	10.2	5.1	5.3	4.9	15.2	9.8	12.5	8.4	5.2	4.6	5.1
	MARE	0.6	0.4	0.5	0.4	0.2	0.2	0.1	0.5	0.4	0.5	0.4	0.2	0.2	0.2
	CE	-3.1	-0.5	-1.3	-0.2	0.7	0.7	0.7	-2.1	-0.3	-1.1	0.0	0.6	0.7	0.7
	IPE	2.1	1.0	1.4	-1.1	-2.5	-2.5	-2.5	4.3	2.1	3.1	1.7	-1.1	-1.3	-1.3
17	RMSE	15.6	4.5	5.6	7.8	3.1	5.7	3.3	16.6	4.4	5.6	8.6	3.4	6.0	3.7
	MARE	1.1	0.2	0.4	0.5	0.2	0.4	0.1	1.2	0.2	0.4	0.5	0.2	0.4	0.1
	CE	-2.5	0.7	0.5	0.1	0.9	0.5	0.8	-3.6	0.7	0.5	-0.3	0.8	0.4	0.8
	IPE	1.9	-2.5	-2.0	-1.4	-3.3	-2.0	-5.0	5.0	-1.7	1.1	1.7	-2.0	1.2	-2.0
18	RMSE	47.6	22.7	17.8	8.8	4.6	13.6	2.2	42.6	19.0	16.5	8.0	5.6	12.3	3.0
	MARE	1.3	0.6	0.5	0.3	0.1	0.4	0.1	1.3	0.6	0.6	0.3	0.2	0.3	0.1
	CE	-276.0	-61.9	-37.7	-8.5	-1.5	-21.7	0.4	-221.0	-43.4	-32.4	-6.9	-2.9	-17.4	-0.1
	IPE	86.3	20.0	12.6	3.5	1.2	7.5	-2.0	375.2	75.4	56.9	14.0	7.1	31.6	2.5
19	RMSE	20.4	30.4	43.8	20.4	13.8	18.4	11.3	19.8	29.1	41.0	19.6	15.2	18.3	11.8
	MARE	0.4	0.9	1.4	0.4	0.4	0.5	0.2	0.4	1.1	1.3	0.4	0.6	0.6	0.3
	CE	0.3	-0.6	-2.3	0.3	0.7	0.4	0.8	0.3	-0.5	-1.9	0.3	0.6	0.4	0.8
	IPE	-2.0	-1.1	1.5	-2.0	-2.5	-2.0	-5.0	1.2	2.5	4.1	1.1	1.2	1.3	-1.7
20	RMSE	19.4	18.4	12.6	22.3	3.8	12.1	3.6	21.4	17.3	13.7	23.0	4.4	13.4	4.2
	MARE	1.3	0.7	0.6	1.5	0.2	0.7	0.2	1.8	0.7	0.8	2.0	0.3	1.0	0.3
	CE	-6.2	-5.5	-2.0	-8.5	0.7	-1.8	0.8	-6.2	-3.7	-2.0	-7.4	0.7	-1.8	0.7
	IPE	3.5	3.0	1.6	4.5	-2.5	1.6	-3.3	6.7	4.1	3.0	7.7	-1.4	3.1	-1.4

Table S8 (Cont.). IPE and its components values for mean annual runoff for GHMs, EM and MMC

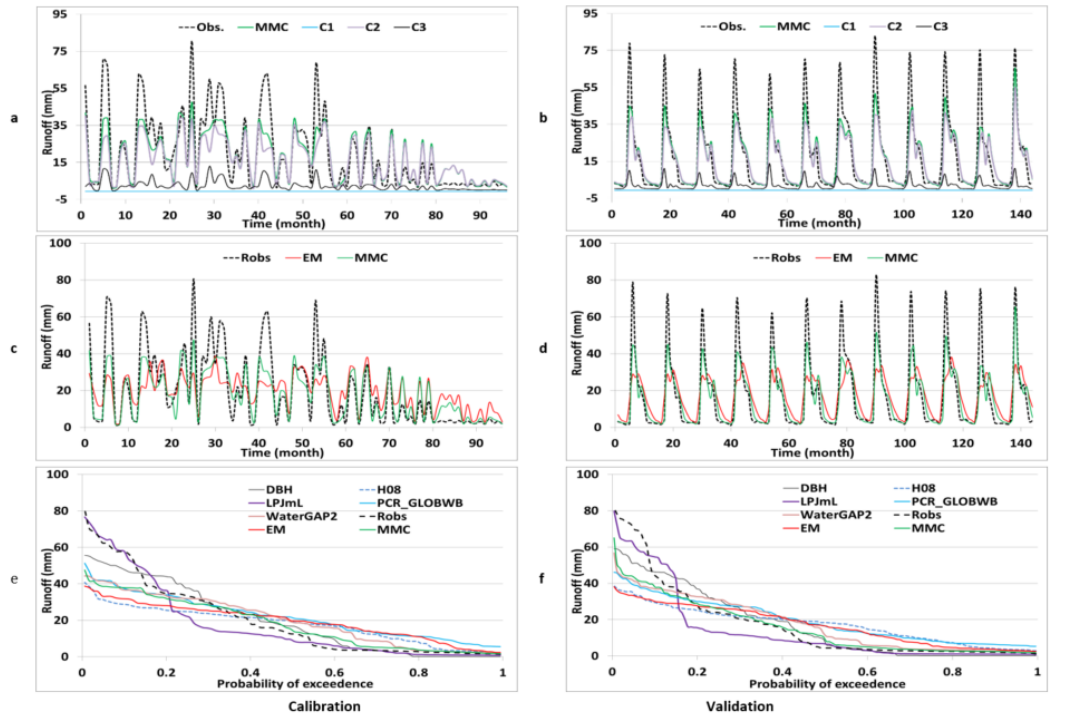
		Calibration							Validation						
Catchment No.	Metrics	DBH	H08	LPJmL	PCR_GLOBWB	WaterGap2	EM	MMC	DBH	H08	LPJmL	PCR_GLOBWB	Water-ap2	EM	MMC
21	RMSE	24.3	31.2	20.3	17.4	7.5	15.6	6.7	23.8	28.2	17.1	15.2	6.8	14.4	5.7
	MARE	0.6	0.6	0.4	0.4	0.2	0.3	0.1	0.6	0.5	0.3	0.4	0.2	0.3	0.1
	CE	-1.4	-2.9	-0.6	-0.2	0.8	0.0	0.8	-1.7	-2.8	-0.4	-0.1	0.8	0.0	0.8
	IPE	1.5	2.0	1.1	-1.1	-3.3	-1.4	-3.3	2.6	3.3	1.5	1.4	-2.0	1.2	-2.5
22	RMSE	23.4	18.7	13.2	16.8	6.3	9.1	5.4	23.1	18.5	11.0	15.3	5.8	8.8	4.9
	MARE	0.6	0.5	0.4	0.5	0.2	0.3	0.1	0.7	0.5	0.3	0.5	0.1	0.2	0.1
	CE	-5.1	-2.9	-0.9	-2.1	0.6	0.1	0.7	-4.7	-2.7	-0.3	-1.5	0.6	0.2	0.7
	IPE	2.5	1.7	1.0	1.5	-2.5	-1.7	-3.3	4.0	2.7	1.3	2.1	-2.0	-1.1	-2.0
23	RMSE	39.0	24.2	15.9	13.1	6.4	8.7	3.6	41.4	22.7	15.4	13.1	6.1	9.3	3.6
	MARE	1.1	0.9	0.6	0.5	0.3	0.3	0.1	1.1	0.8	0.6	0.5	0.3	0.3	0.1
	CE	-34.1	-12.5	-4.8	-3.0	0.0	-0.7	0.7	-48.9	-14.0	-5.9	-4.0	-0.1	-1.5	0.6
	IPE	10.8	4.5	2.2	1.6	-1.4	-1.1	-3.3	83.4	25.6	12.2	8.9	2.4	4.7	1.1
24	RMSE	28.7	16.7	17.0	7.2	4.2	10.4	4.1	28.4	16.5	12.9	5.8	4.6	9.9	4.5
	MARE	1.8	0.9	0.7	0.4	0.2	0.7	0.2	1.6	0.8	0.6	0.3	0.2	0.6	0.2
	CE	-10.8	-3.0	-3.2	0.2	0.7	-0.5	0.8	-9.4	-2.5	-1.2	0.6	0.7	-0.3	0.7
	IPE	5.3	2.2	2.1	-1.4	-2.5	1.2	-2.5	6.8	2.8	1.9	-1.4	-1.7	1.5	-2.0
25	RMSE	20.6	13.4	18.9	23.7	2.4	13.9	2.3	19.4	13.3	16.2	21.2	2.1	12.8	1.8
	MARE	4.2	3.0	4.1	5.0	0.5	3.3	0.4	3.5	2.3	3.3	3.7	0.3	2.6	0.2
	CE	-49.6	-20.4	-41.8	-66.3	0.3	-21.9	0.4	-94.4	-43.9	-65.4	-113.7	-0.1	-40.3	0.1
	IPE	20.2	9.1	17.2	26.6	-1.4	9.7	-1.7	83.5	39.9	58.8	100.1	1.5	37.1	1.2
26	RMSE	8.7	21.7	18.0	43.5	9.3	18.7	7.3	8.8	19.7	17.4	41.3	8.6	18.1	6.5
	MARE	0.2	0.4	0.4	0.8	0.1	0.3	0.1	0.2	0.4	0.5	0.9	0.1	0.4	0.1
	CE	0.9	0.4	0.6	-1.4	0.9	0.6	0.9	0.9	0.5	0.6	-1.0	0.9	0.6	1.0
	IPE	-5.0	-1.7	-2.0	1.5	-5.0	-2.0	-5.0	-2.5	-1.1	1.0	2.8	-3.3	-1.1	-5.0
27	RMSE	7.1	3.0	4.2	3.2	3.3	3.0	2.6	7.1	1.4	3.3	2.8	2.0	2.4	1.3
	MARE	1.4	0.4	0.9	0.6	0.5	0.6	0.4	1.4	0.3	0.8	0.5	0.4	0.5	0.3
	CE	-5.3	-0.1	-1.2	-0.2	-0.3	-0.1	0.1	-43.0	-0.7	-8.6	-5.7	-2.4	-3.9	-0.4
	IPE	3.0	-1.3	1.4	-1.1	-1.1	-1.1	-1.4	52.9	2.5	12.1	8.5	4.6	6.4	2.2
28	RMSE	21.3	17.4	10.9	18.4	2.9	12.9	2.9	21.1	15.3	8.9	19.2	2.8	12.1	2.9
	MARE	7.4	1.8	1.4	9.3	0.6	3.9	0.4	10.5	5.8	3.2	12.0	1.3	6.4	1.1
	CE	-11.2	-7.1	-2.2	-8.1	0.8	-3.5	0.8	-24.0	-12.2	-3.5	-19.8	0.5	-7.3	0.5
	IPE	7.9	5.2	2.2	6.2	-3.3	3.1	-3.3	15.1	8.3	3.8	15.0	1.4	7.3	1.2
29	RMSE	17.6	18.9	19.5	11.2	4.7	13.5	1.3	14.5	15.1	14.9	7.7	2.7	9.9	2.1
	MARE	2.9	3.5	2.3	2.9	0.7	2.4	0.2	1.8	2.3	1.3	1.6	0.4	1.3	0.3
	CE	-12.0	-13.9	-14.9	-4.3	0.1	-6.6	0.9	-7.5	-8.3	-8.0	-1.4	0.7	-2.9	0.8
	IPE	8.7	10.0	10.5	4.2	1.0	5.4	-5.0	9.7	10.7	10.0	3.6	-1.4	4.9	-1.7
30	RMSE	22.1	17.5	14.6	9.9	5.0	11.4	4.4	18.3	15.3	12.1	9.2	7.6	9.1	6.5
	MARE	0.7	0.5	0.4	0.3	0.1	0.3	0.1	0.5	0.4	0.3	0.3	0.2	0.2	0.2
	CE	-9.2	-5.4	-3.5	-1.1	0.5	-1.7	0.6	-4.9	-3.1	-1.6	-0.5	0.0	-0.5	0.2
	IPE	3.3	2.2	1.7	1.0	-2.5	1.2	-2.5	8.3	5.9	3.9	2.5	1.8	2.4	1.4

Table S8 (Cont.). IPE and its components values for mean annual runoff for GHMs, EM and MMC

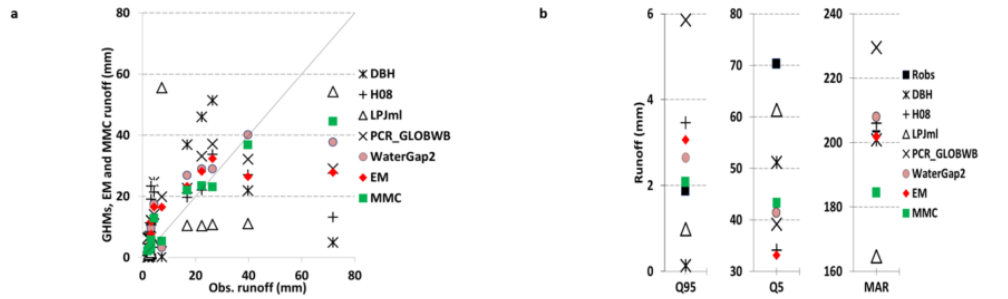
Catchment No.	Metrics	Calibration							Validation						
		DBH	H08	LPJmL	PCR_GLOBWB	WaterGap2	EM	MMC	DBH	H08	LPJmL	PCR_GLOBWB	Water-ap2	EM	MMC
31	RMSE	24.7	21.1	29.6	32.2	16.5	19.4	10.2	23.2	18.8	27.6	32.5	13.9	17.6	9.3
	MARE	0.2	0.2	0.3	0.2	0.1	0.2	0.1	0.3	0.2	0.2	0.3	0.1	0.1	0.1
	CE	0.1	0.3	-0.4	-0.6	0.6	0.4	0.8	0.3	0.5	0.0	-0.4	0.8	0.6	0.9
	IPE	-1.3	-1.7	-1.1	1.0	-2.5	-2.0	-3.3	2.0	1.5	2.6	3.4	-1.1	1.3	-2.0
32	RMSE	69.2	63.6	67.0	27.6	22.4	42.3	21.9	61.9	67.1	64.5	22.2	22.1	40.2	21.2
	MARE	2.7	1.0	1.5	0.6	0.4	1.1	0.4	5.0	1.9	3.1	1.0	1.1	2.2	0.9
	CE	-1.4	-1.1	-1.3	0.6	0.7	0.1	0.8	-1.3	-1.6	-1.4	0.7	0.7	0.1	0.7
	IPE	1.8	1.3	1.5	-2.5	-3.3	-1.3	-3.3	5.9	4.7	4.9	1.1	1.2	2.6	1.0
33	RMSE	32.7	32.7	33.2	10.8	4.6	21.0	1.7	32.7	39.7	37.5	13.3	4.2	23.6	1.7
	MARE	3.8	2.2	3.1	0.9	0.6	1.9	0.2	4.2	2.2	3.1	1.0	0.6	2.0	0.2
	CE	-52.5	-52.6	-54.2	-4.9	0.0	-21.1	0.9	-64.9	-95.9	-85.5	-9.9	-0.1	-33.2	0.8
	IPE	35.6	35.3	36.5	4.3	1.3	14.9	-3.3	48.8	70.8	63.4	8.4	1.5	25.4	-2.5
34	RMSE	41.0	35.9	28.6	31.7	12.4	26.5	9.7	38.4	32.5	21.8	24.5	8.4	22.5	6.7
	MARE	0.9	0.5	0.5	0.6	0.3	0.4	0.2	1.0	0.6	0.6	0.7	0.3	0.5	0.3
	CE	-3.2	-2.2	-1.0	-1.5	0.6	-0.7	0.8	-3.6	-2.3	-0.5	-0.9	0.8	-0.6	0.9
	IPE	2.9	2.2	1.6	1.9	-2.0	1.4	-2.5	4.8	3.5	1.9	2.2	-1.7	1.9	-2.0
35	RMSE	42.5	44.5	35.9	37.0	6.7	31.7	5.8	35.0	37.2	31.1	29.6	6.1	26.4	5.4
	MARE	2.7	2.2	2.1	2.2	0.4	1.8	0.3	5.8	4.4	4.6	5.1	2.3	4.4	2.6
	CE	-34.7	-38.2	-24.4	-26.1	0.1	-18.9	0.3	-39.2	-44.3	-30.7	-27.7	-0.2	-21.9	0.1
	IPE	12.1	13.1	8.8	9.3	-1.7	7.0	-1.7	136.9	153.0	108.1	98.4	8.0	78.5	8.2
36	RMSE	21.9	18.0	17.0	9.8	7.3	9.2	10.1	19.7	11.9	14.6	13.2	11.5	9.8	10.2
	MARE	63.2	2.1	38.8	17.0	8.4	25.7	1.8	17.9	3.7	8.1	5.2	4.2	7.6	1.8
	CE	-0.5	0.0	0.1	0.7	0.8	0.7	0.7	0.2	0.7	0.5	0.6	0.7	0.8	0.8
	IPE	1.2	-1.1	-1.3	-2.5	-3.3	-2.5	-2.5	6.9	1.4	3.1	2.0	1.6	2.9	-1.4
37	RMSE	11.5	1.7	11.1	5.8	1.0	5.8	0.8	7.7	1.6	8.0	5.0	0.8	4.1	1.0
	MARE	25.2	1.8	15.0	17.0	1.0	11.9	0.7	68.1	9.9	47.6	55.1	3.1	36.7	2.3
	CE	-154.9	-2.4	-144.0	-39.5	-0.2	-38.4	0.3	-102.8	-3.3	-109.3	-42.9	-0.2	-28.6	-0.7
	IPE	89.6	2.3	83.1	24.1	1.0	23.1	-1.4	83.1	7.1	81.1	46.4	2.3	31.2	2.0
38	RMSE	23.1	6.9	16.8	3.4	2.7	9.4	2.4	18.5	2.9	12.0	2.8	1.0	7.0	0.9
	MARE	64901	9658	307693	71119	7966	209090	557	1030724	21936	379899	92248	15753	308112	2954
	CE	-57.8	-4.3	-29.8	-0.3	0.2	-8.6	0.4	-521.6	-11.6	-221.3	-11.5	-0.4	-73.2	-0.3
	IPE	63.6	2.4	30.6	6.6	-1.1	19.5	-2.0	6992	149	2578	625	106	2088	20.1
39	RMSE	25.1	6.0	19.4	7.3	5.2	10.2	5.1	19.2	7.4	16.0	7.3	5.5	8.5	5.3
	MARE	2194	150	1839	856	71	1022	46	57166	4690	39895	24101	3430	25856	2731
	CE	-5.0	0.7	-2.6	0.5	0.7	0.0	0.7	-2.1	0.5	-1.1	0.6	0.7	0.4	0.8
	IPE	6.3	-2.0	4.9	2.1	-2.5	2.6	-3.3	641.2	52.6	447.5	270.3	38.5	290.0	30.6
40	RMSE	25.1	4.7	17.1	9.6	1.0	10.8	1.1	15.9	2.0	10.4	5.6	0.1	6.5	0.4
	MARE	69399	1132	27319	23162	457	24293	76	657576	19041	305242	193641	2223	235545	1992
	CE	-494.5	-16.2	-229.9	-72.1	0.2	-91.4	0.0	-21005	-325	-9036	-2597	0	-3552	-10.1
	IPE	200.6	7.0	92.3	35.2	-1.4	41.9	-1.7	39448	1126	18262	11432	131	13919	117



1. Lena catchment at Stob station.



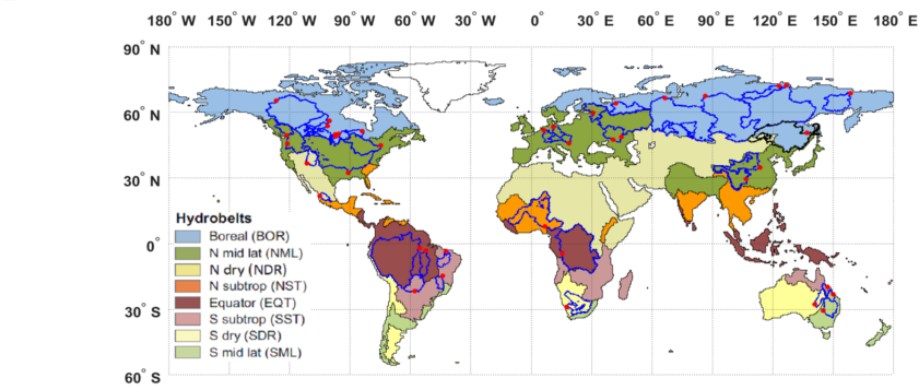
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



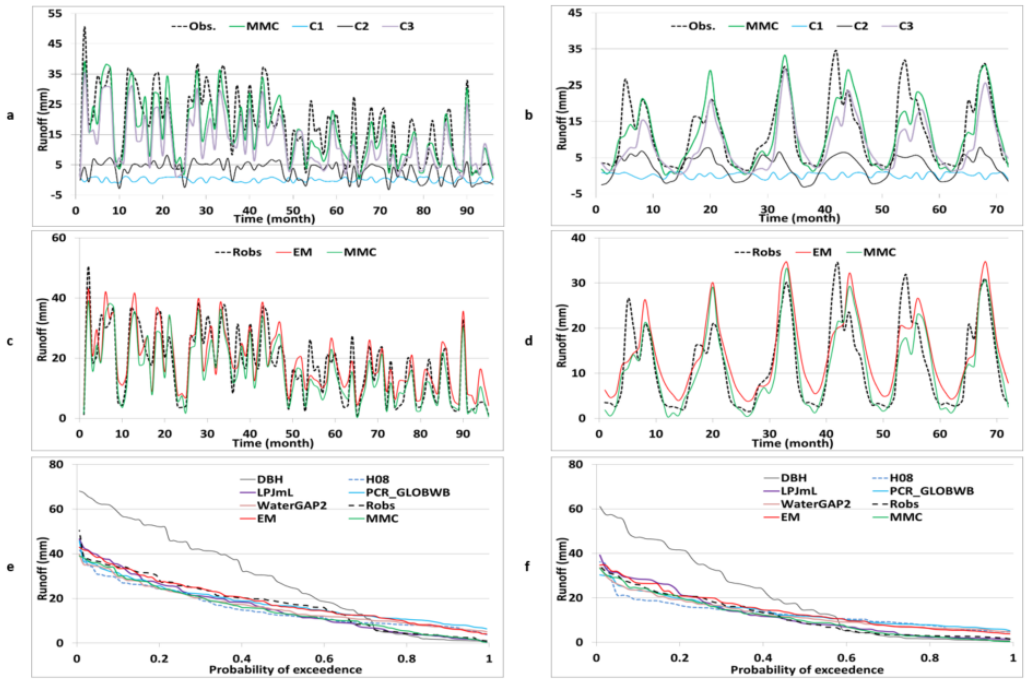
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM.

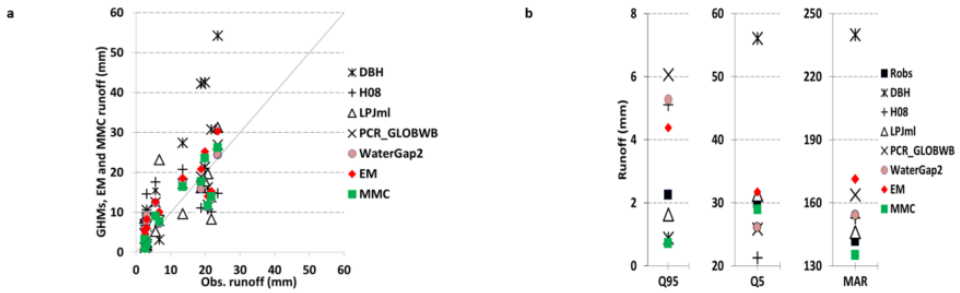
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	-1.1	1.0	-1.4	-1.3	-2.4	-1.7	-3.6	-183.0	-613.2	-121.7	-152.4
Calibration	1.1	-1.2	-1.3	-1.7	-2.2	-1.8	-3.0	-128.9	-705.5	-88.9	-105.9
Validation	1.6	2.0	1.4	1.5	-1.2	1.1	-2.0	-415.2	-41341.1	-78.4	-386.6



2. Amur catchment at Komsomolsk station.



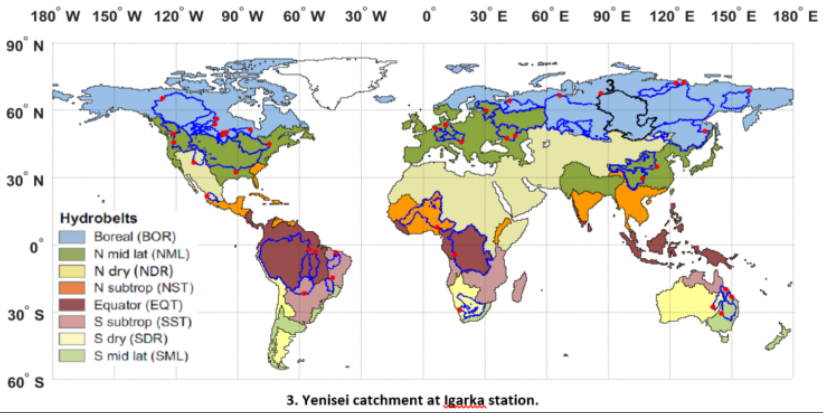
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



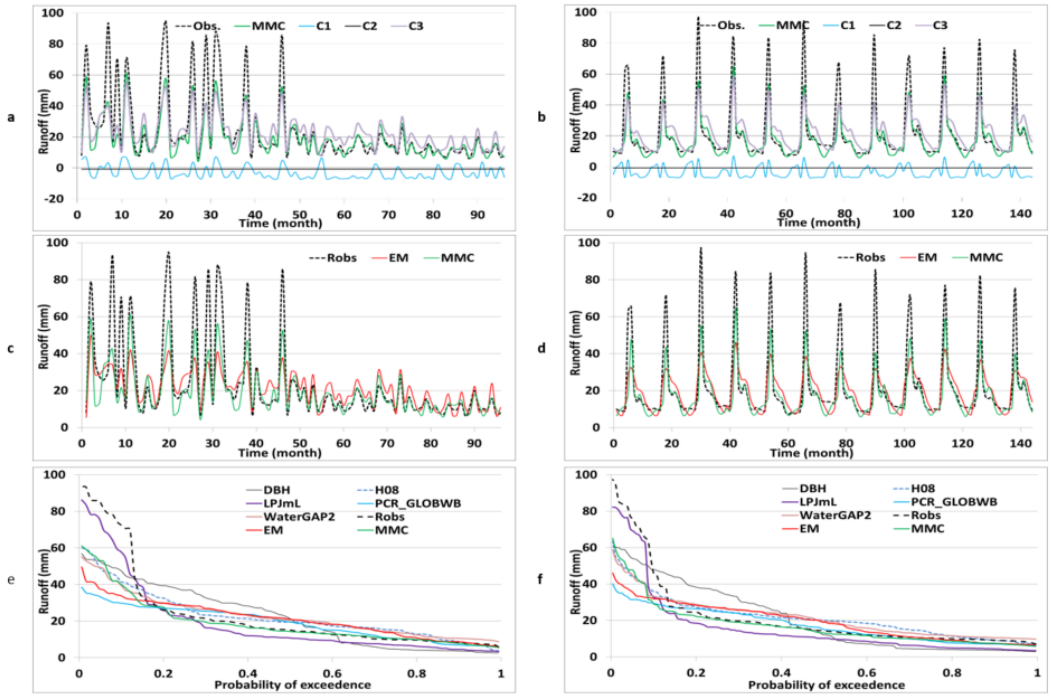
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

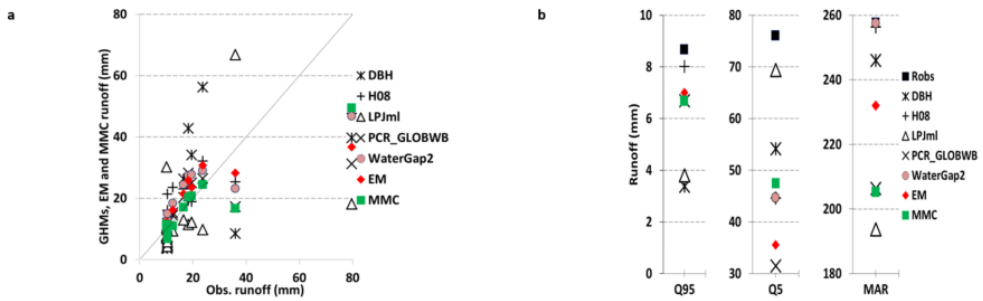
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	-1.2	-1.1	-1.7	-1.4	-1.7	-2.1	-3.4	-134.7	-613.2	-165.8	-152.4
Calibration	1.2	-1.2	-1.6	-1.8	-2.1	-2.4	-3.0	-66.1	-705.5	-91.1	-105.9
Validation	3.1	1.9	1.3	1.3	1.2	1.1	-1.5	-356.3	-41341.1	-466.4	-386.6



3. Yenisei catchment at Igarka station.



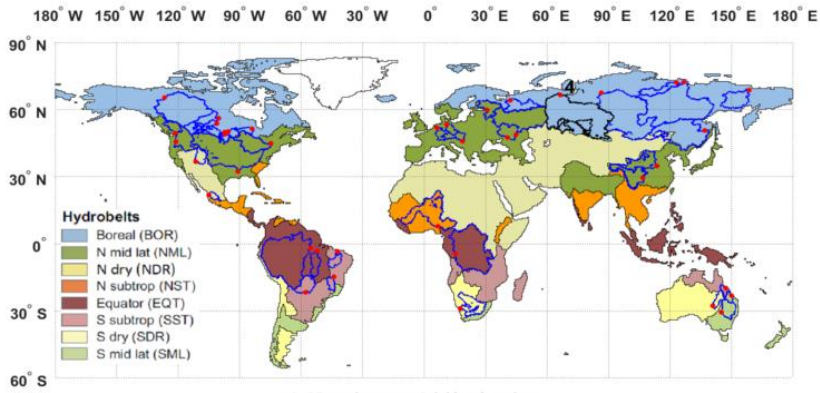
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



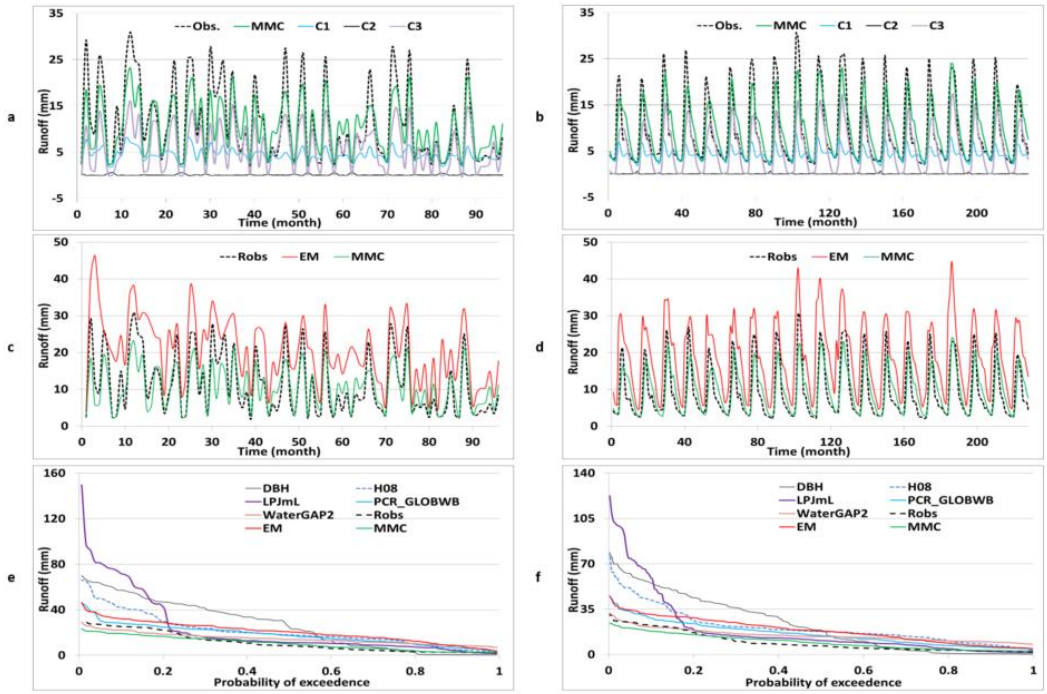
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

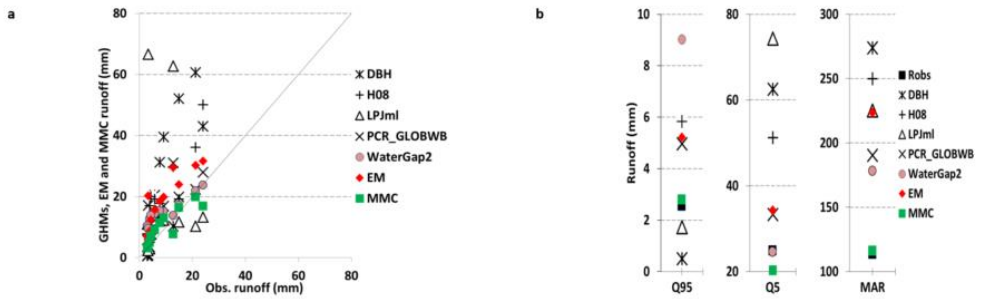
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	-1.2	-2.0	-1.1	-1.9	-2.1	-2.2	-3.0	-81.4	-613.2	-89.3	-152.4
Calibration	-1.4	-2.5	-1.3	-2.1	-2.6	-2.4	-3.2	-79.1	-705.5	-63.1	-105.9
Validation	1.2	-1.5	1.2	-1.5	-1.7	-1.7	-2.3	-60.8	-41341.1	-57.7	-386.6



4. OB catchment at Salekhard station.



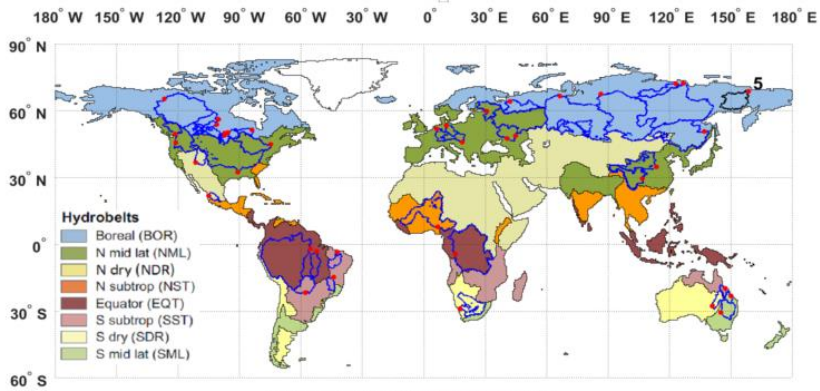
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



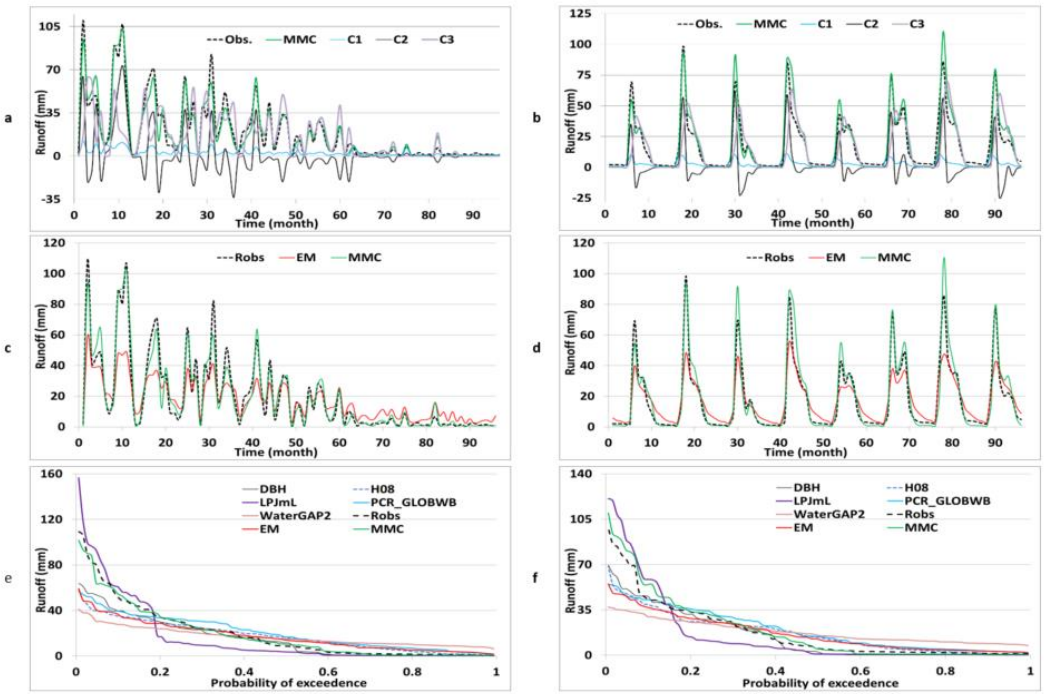
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

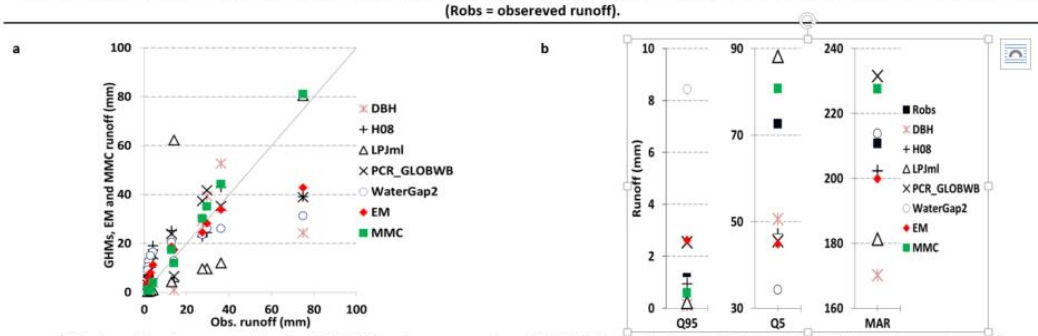
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	2.5	1.8	4.9	1.1	1.1	1.3	-2.6	-488.4	-613.2	-566.6	-152.4
Calibration	2.4	1.2	5.2	1.0	-1.6	1.1	-3.4	-551.6	-705.5	-178.1	-105.9
Validation	8.4	4.7	13.9	2.6	2.5	3.5	-1.3	-583.5	-41341.1	-581.0	-386.6



5. Kolyma catchment at Kolymskaya station.



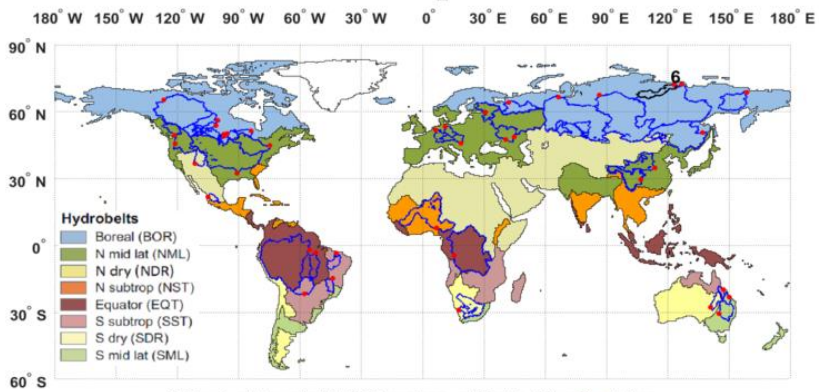
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



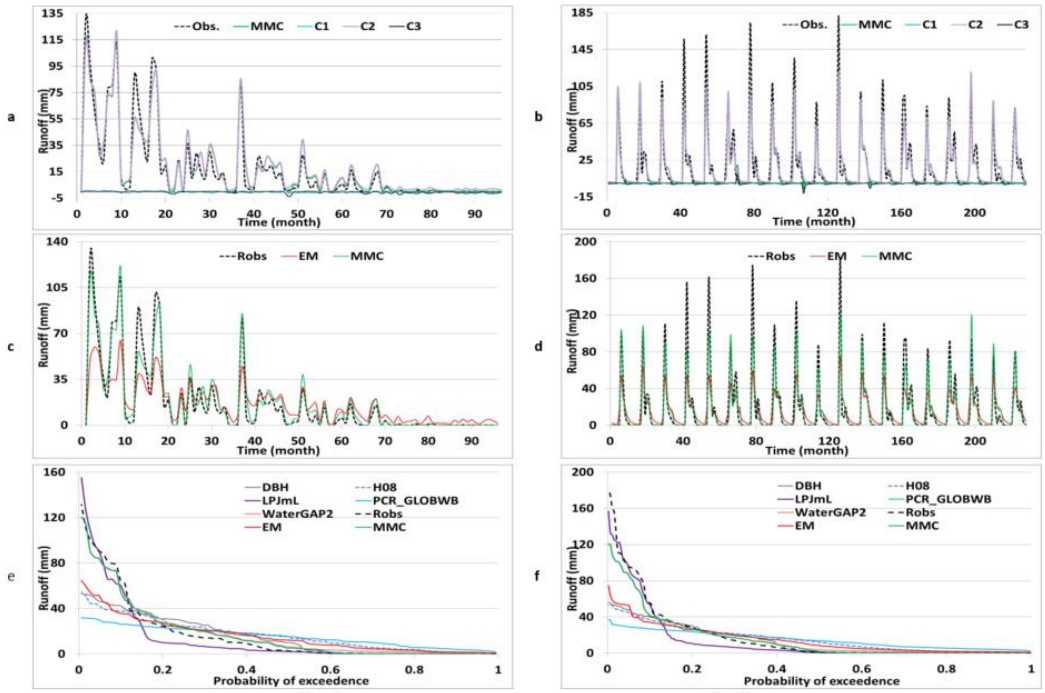
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

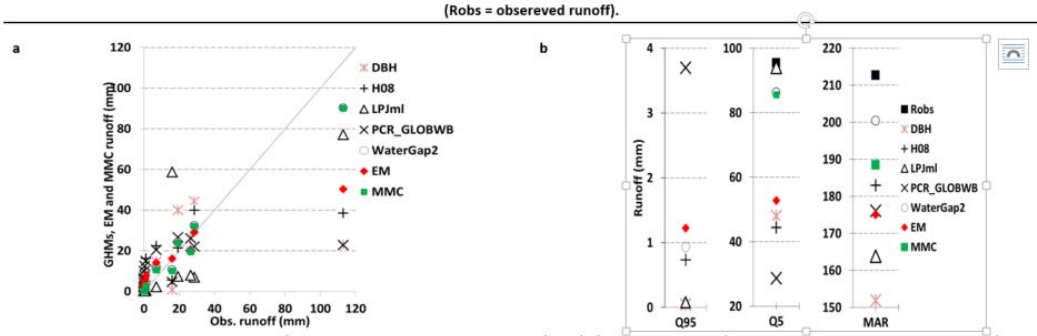
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	-1.7	-1.4	-1.7	-1.1	1.8	-1.5	-5.2	-369.7	-613.2	-350.4	-152.4
Calibration	-1.5	-1.5	-1.5	-1.6	-1.0	-2.0	-5.6	-357.6	-705.5	-400.1	-105.9
Validation	-1.2	1.1	1.2	1.3	2.3	-1.2	-2.4	-119.4	-41341.1	-113.9	-386.6



6. Olenek catchment at 7.5KM Downstream of Mouth of River Pur station.



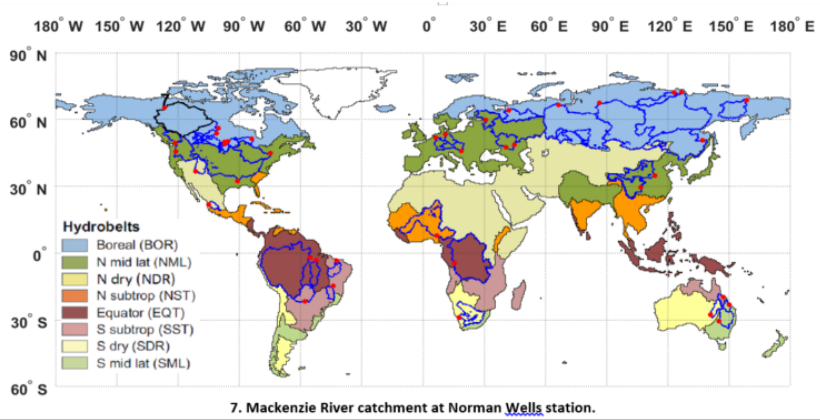
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



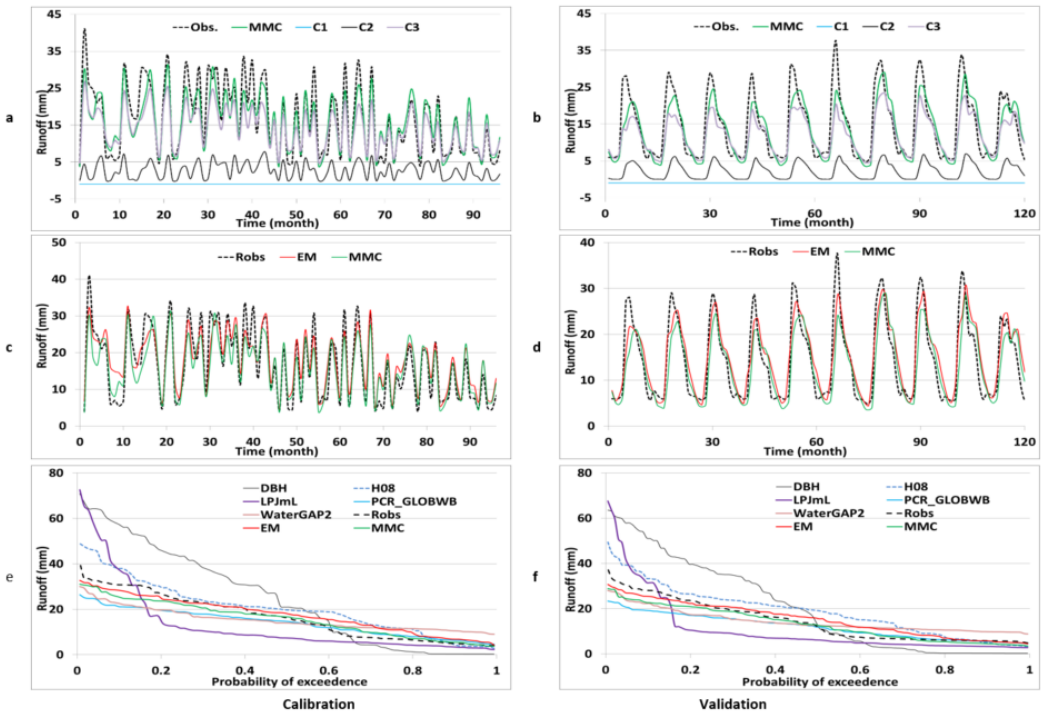
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

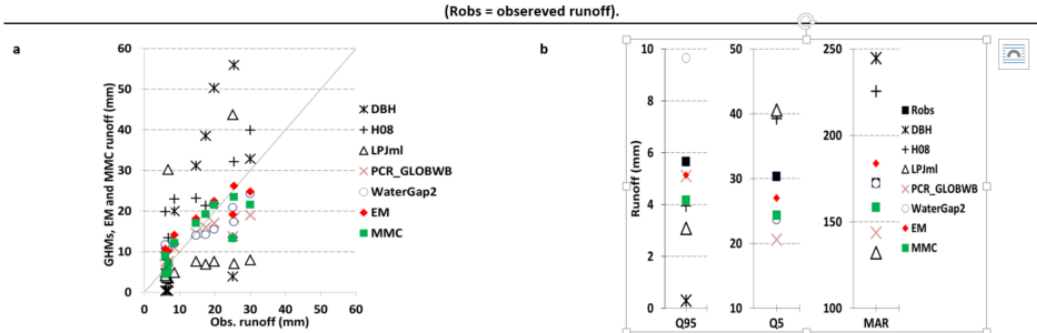
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	-1.5	-1.7	-1.6	1.2	-3.2	-2.0	-6.1	-410.0	-613.2	-290.9	-152.4
Calibration	-1.4	-1.5	1.1	1.3	-3.3	-1.5	-6.0	-448.6	-705.5	-269.7	-105.9
Validation	-1.5	6.3	12.5	17.7	3.9	8.1	-1.2	-1026.9	-41341.1	32.7	-386.6



7. Mackenzie River catchment at Norman Wells station.



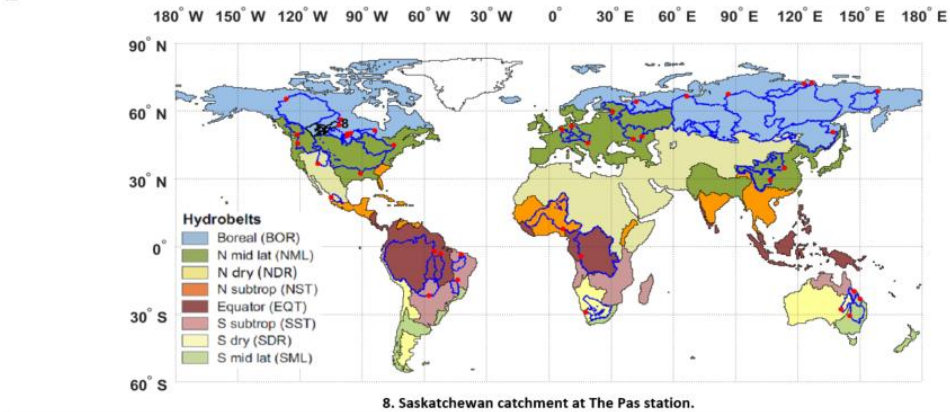
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



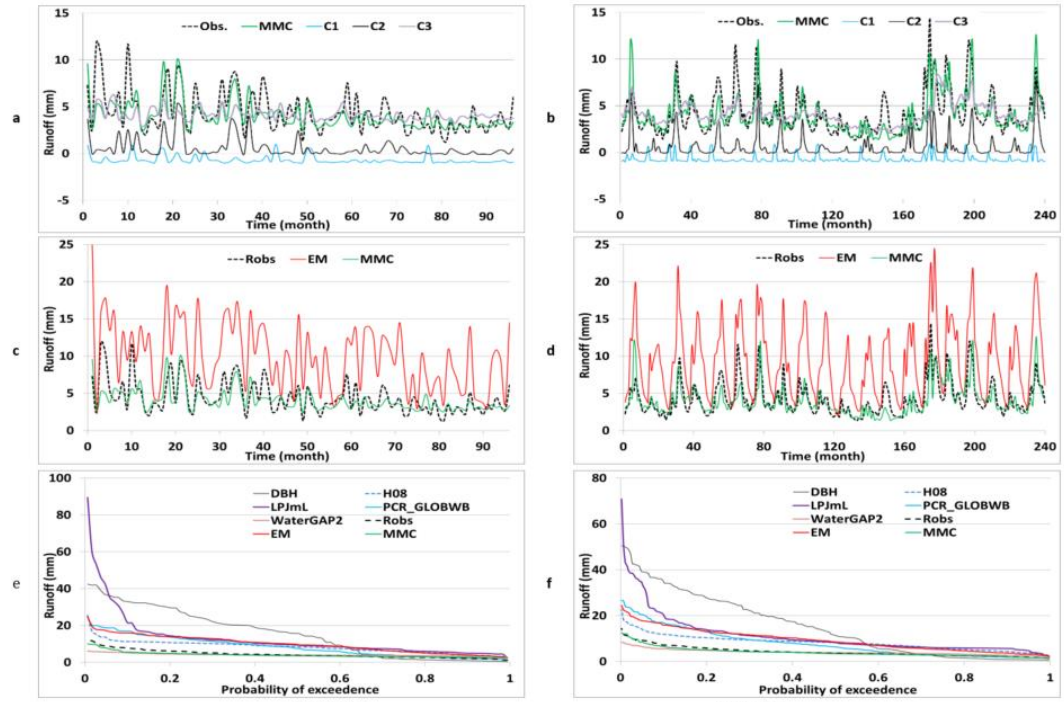
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

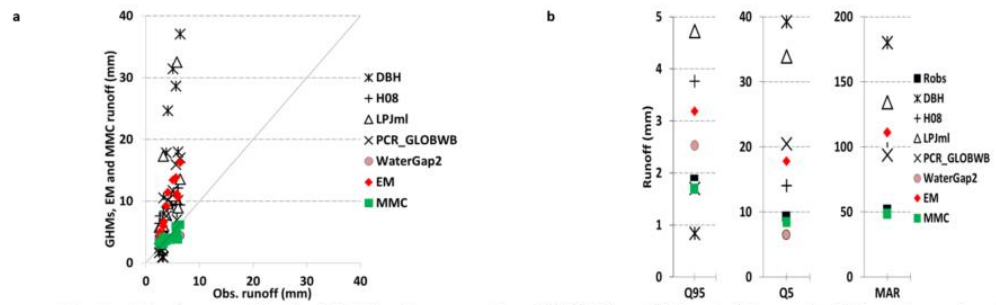
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	1.9	-1.1	1.7	-2.7	-1.4	-2.8	-3.3	-48.8	-613.2	-62.4	-152.4
Calibration	1.5	-1.7	1.4	-3.2	-2.3	-3.7	-3.9	-23.7	-705.5	-74.3	-105.9
Validation	4.5	1.9	3.4	-1.3	1.1	-1.4	-1.3	5.4	-41341.1	-3.5	-386.6



8. Saskatchewan catchment at The Pas station.



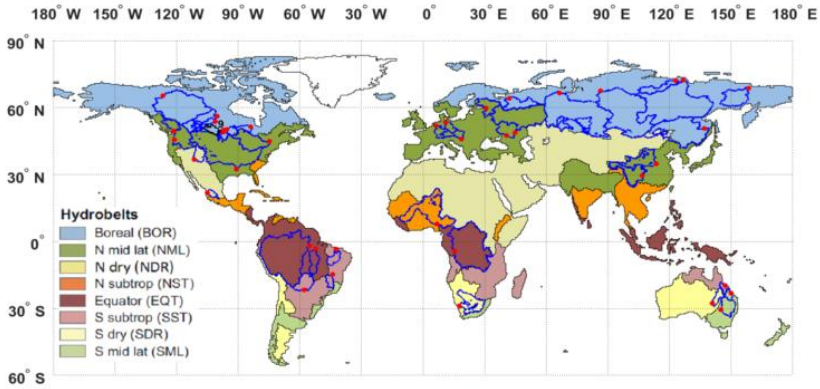
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



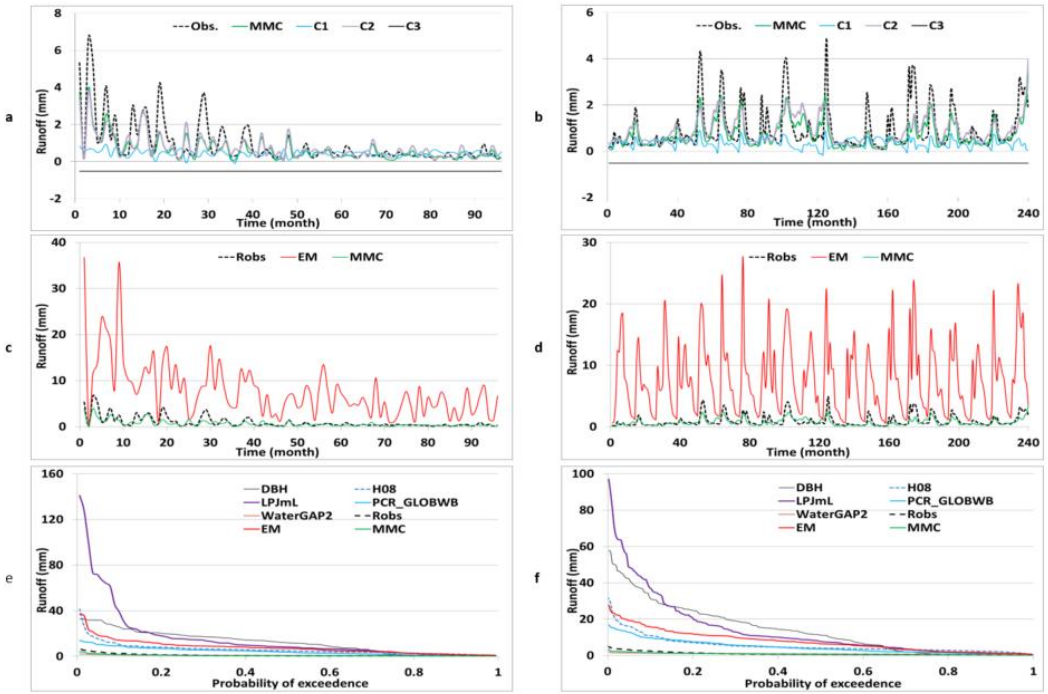
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

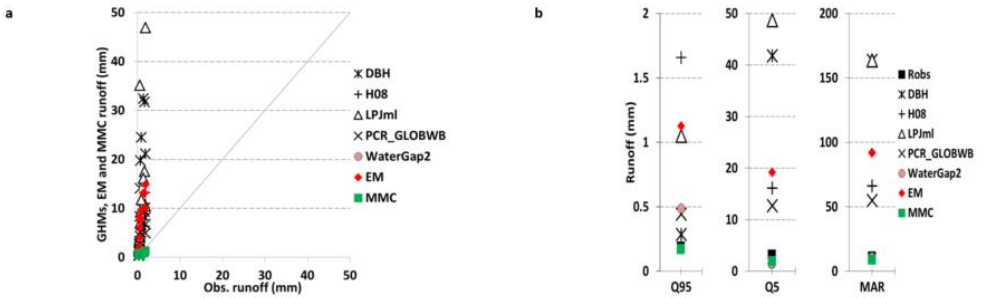
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	15.9	2.6	7.4	2.4	-1.5	2.9	-2.5	-642.6	-613.2	-101.4	-152.4
Calibration	19.7	2.4	16.8	3.1	-1.5	3.7	-1.9	-661.4	-705.5	-46.2	-105.9
Validation	61.4	5.7	27.0	8.2	1.4	9.0	1.0	-793.3	-41341.1	-39.9	-386.6



9. Assiniboine catchment at Headingley station.



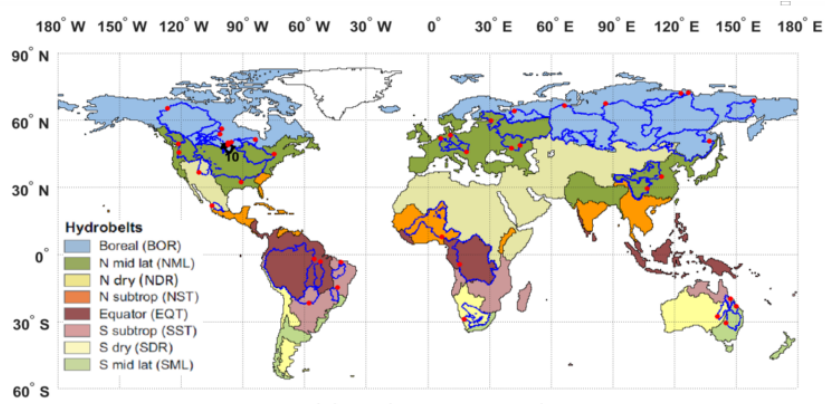
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



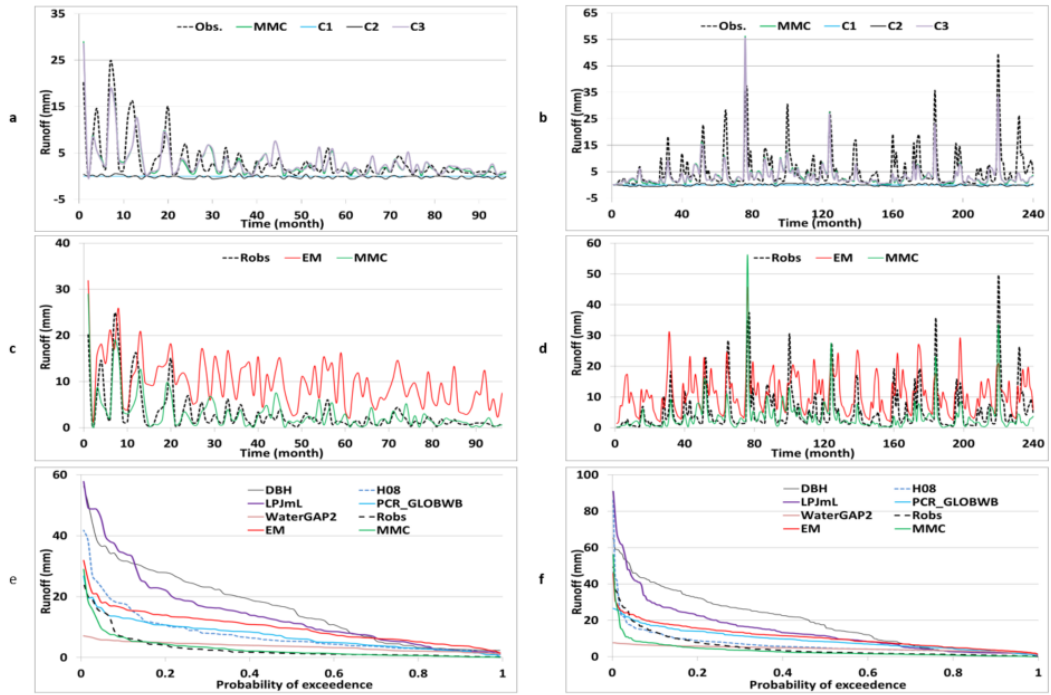
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

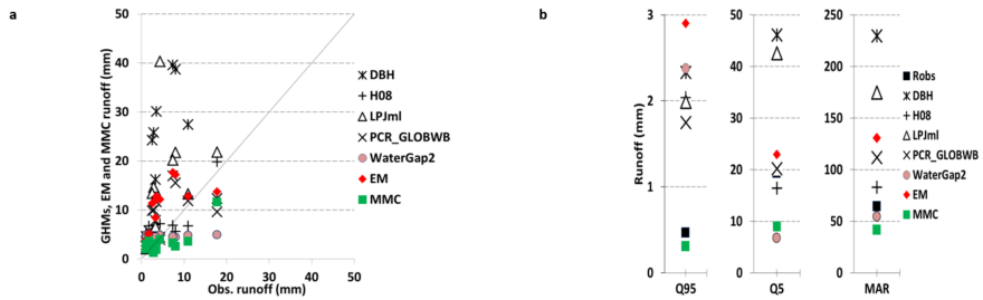
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	113.9	31.2	195.1	9.7	-1.4	31.3	-2.0	-3435.5	-613.2	-61.7	-152.4
Calibration	68.1	16.4	267.8	6.4	-1.5	26.8	-2.2	-3005.5	-705.5	-75.5	-105.9
Validation	384.8	44.5	512.3	28.9	1.6	85.8	1.1	-8473.8	-41341.1	-51.4	-386.6



10. Red River catchment at Emerson station.



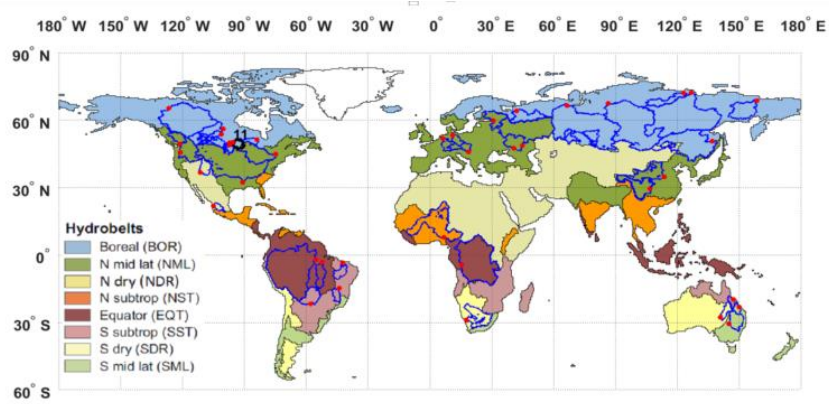
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



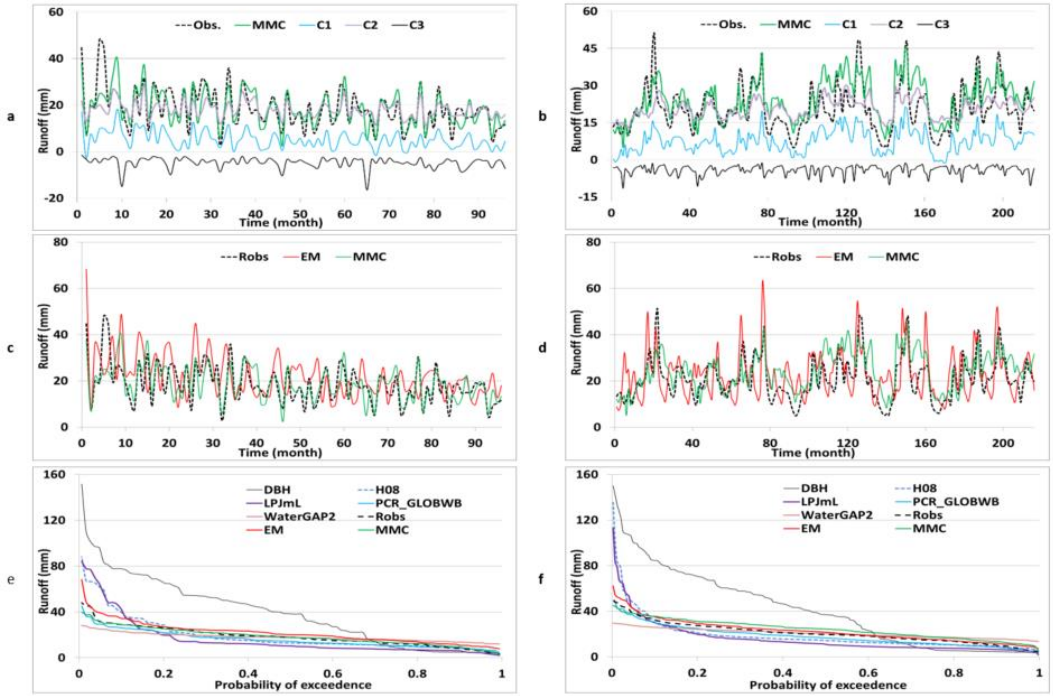
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

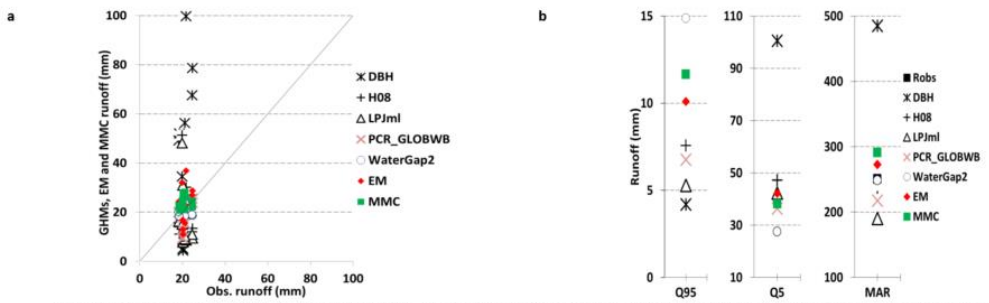
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	6.3	1.5	7.1	1.3	-1.0	2.4	-2.8	-618.9	-613.2	-178.6	-152.4
Calibration	8.1	1.5	6.7	1.3	-1.4	2.1	-2.4	-550.1	-705.5	-105.7	-105.9
Validation	6.6	1.6	4.8	2.1	1.5	2.8	-1.3	-502.2	-41341.1	-477.7	-386.6



11. Winnipeg River catchment at Slave Falls station.



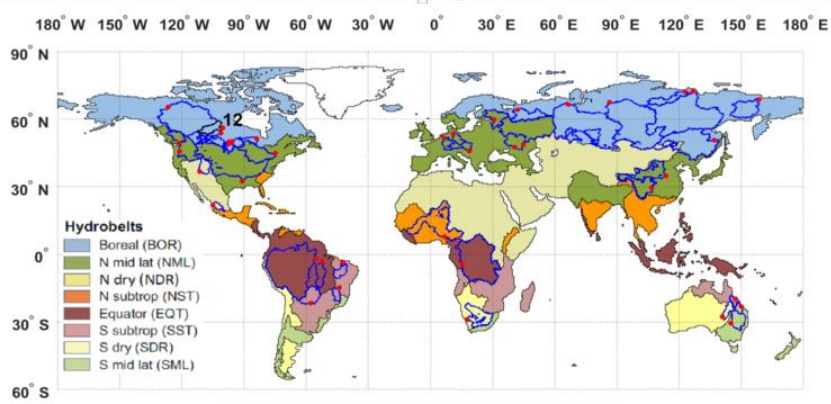
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



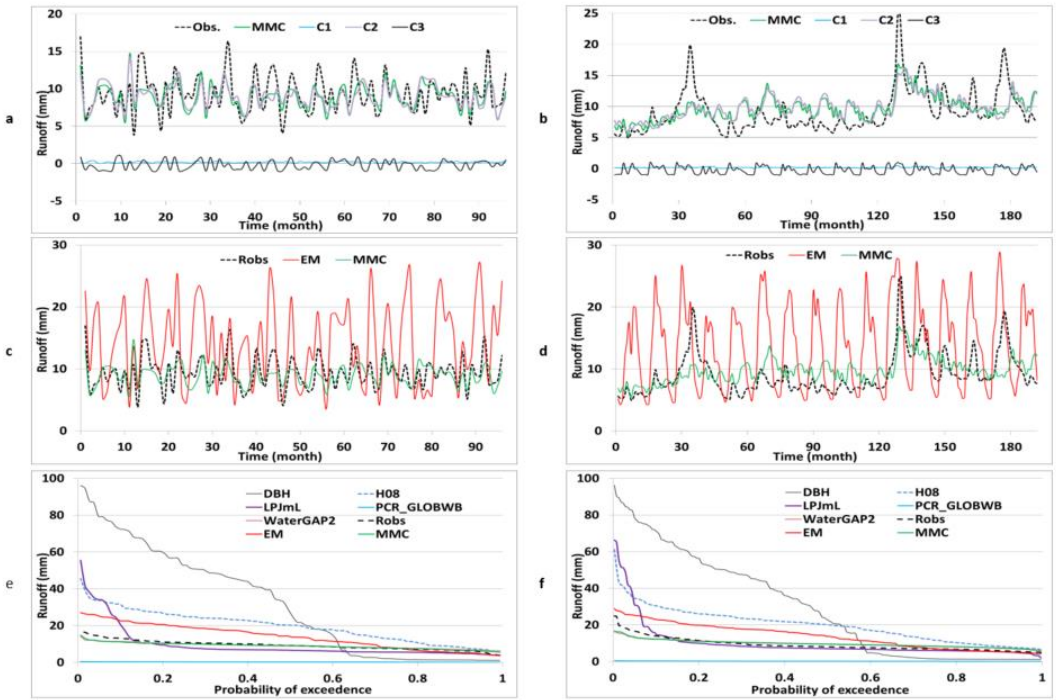
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

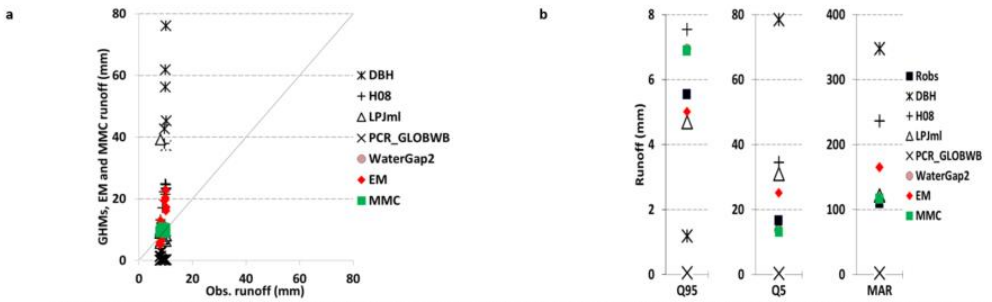
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	6.2	1.1	1.3	-1.6	-2.0	-1.1	-3.1	-197.3	-613.2	-112.3	-152.4
Calibration	6.1	1.4	1.6	-1.7	-2.1	-1.0	-2.6	-152.8	-705.5	-48.9	-105.9
Validation	24.2	4.8	5.0	1.6	1.7	2.3	1.6	-67.1	-41341.1	6.7	-386.6



12. Churchill River catchment at Above Granville Falls station.



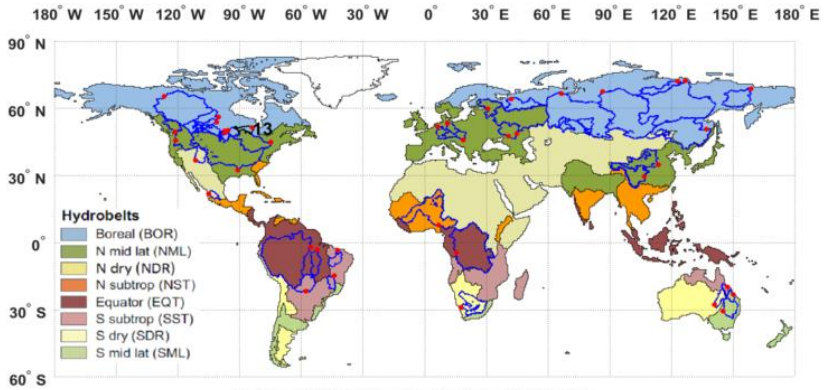
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



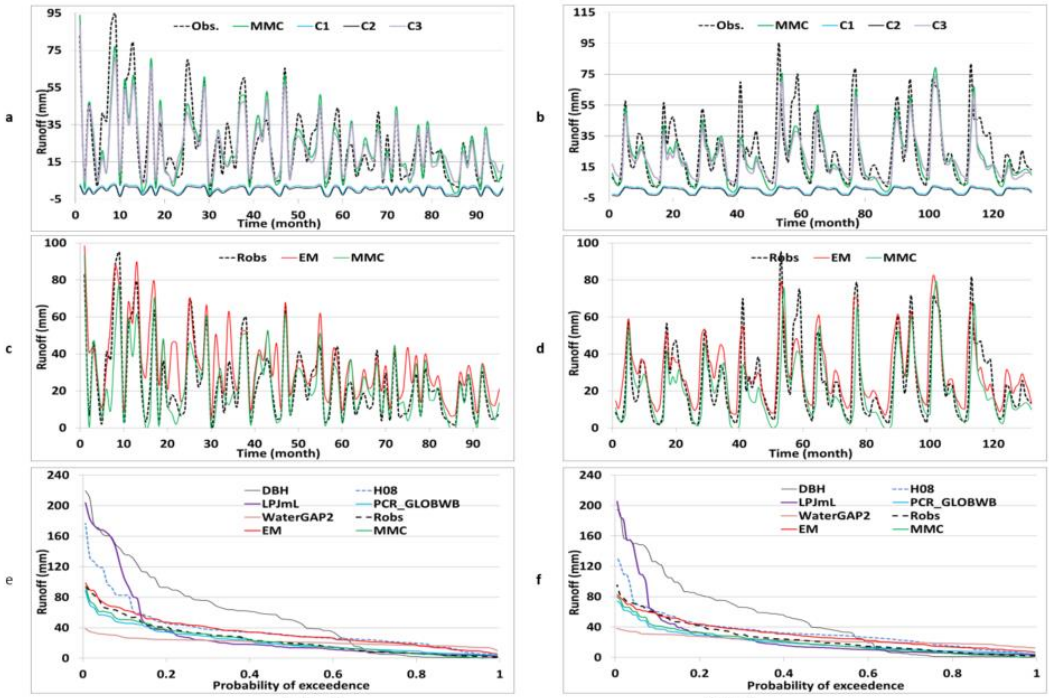
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

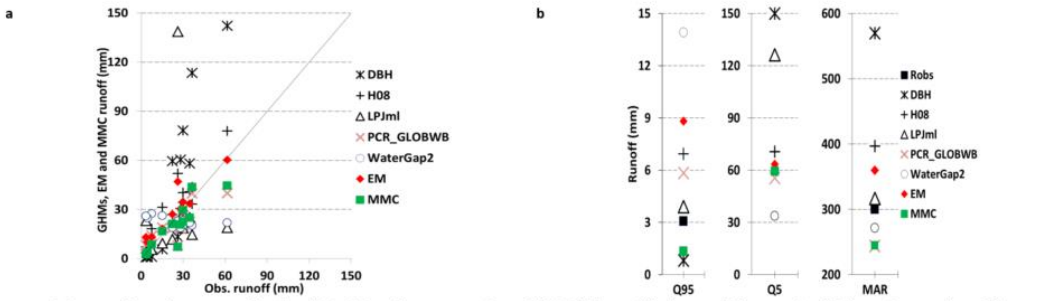
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	36.8	6.7	4.0	4.0	-2.2	2.4	-2.4	-583.8	-613.2	-26.0	-152.4
Calibration	52.0	7.3	4.2	4.2	-1.8	2.9	-2.1	-598.8	-705.5	-22.5	-105.9
Validation	297.5	50.1	32.2	25.6	3.6	17.1	3.1	-1397.9	-41341.1	-50.3	-386.6



13. Albany River catchment at Near Hat Island station.



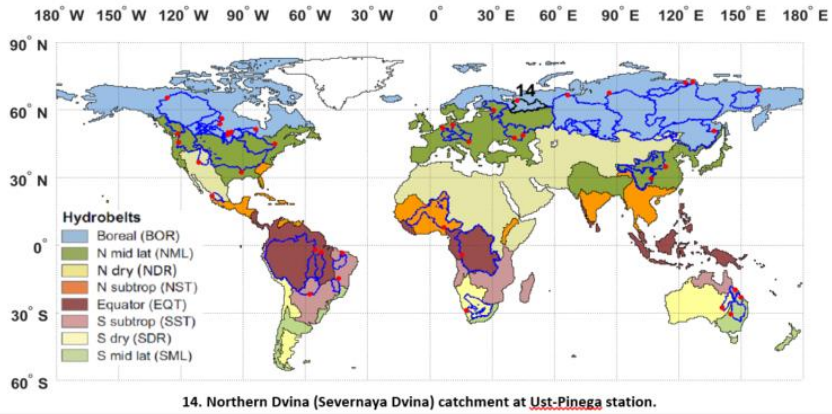
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



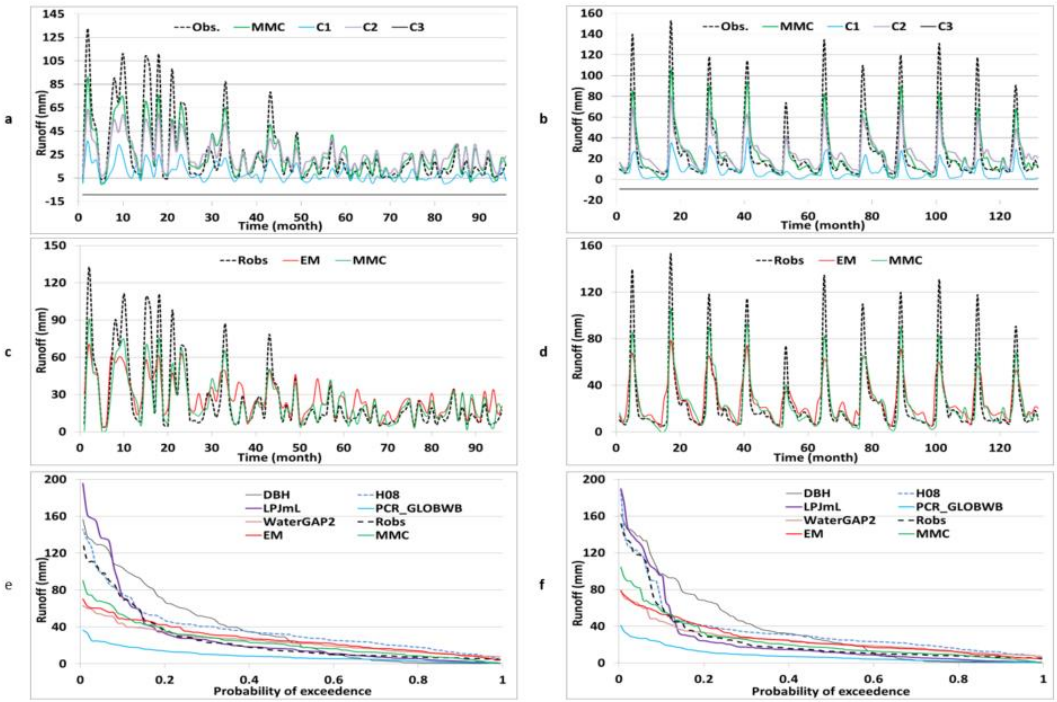
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

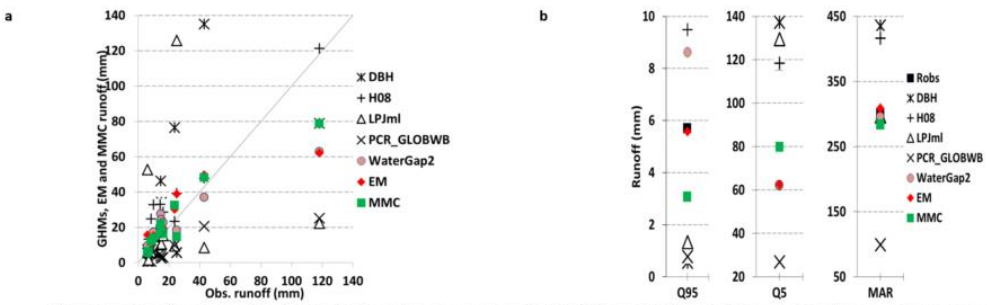
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	2.1	-1.1	2.4	-2.3	1.3	-1.6	-2.9	-131.9	-613.2	-56.6	-152.4
Calibration	2.0	-1.5	2.3	-3.7	-1.4	-2.4	-4.3	-183.4	-705.5	-51.8	-105.9
Validation	2.8	-1.0	2.8	-1.3	1.7	-1.2	-1.7	-44.1	-41341.1	-32.9	-386.6



14. Northern Dvina (Severnaya Dvina) catchment at Ust-Pinega station.



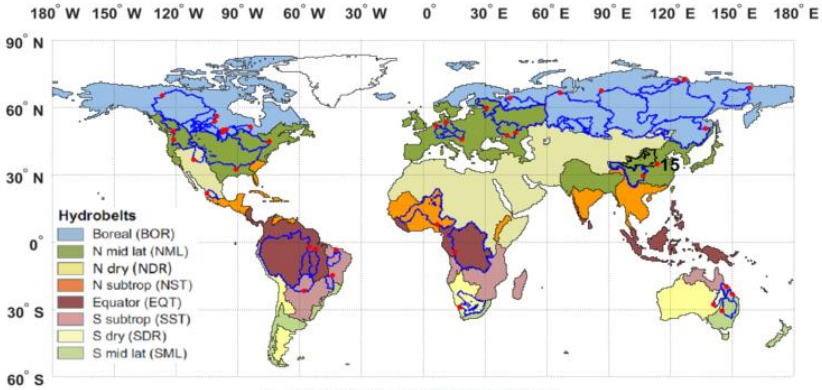
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



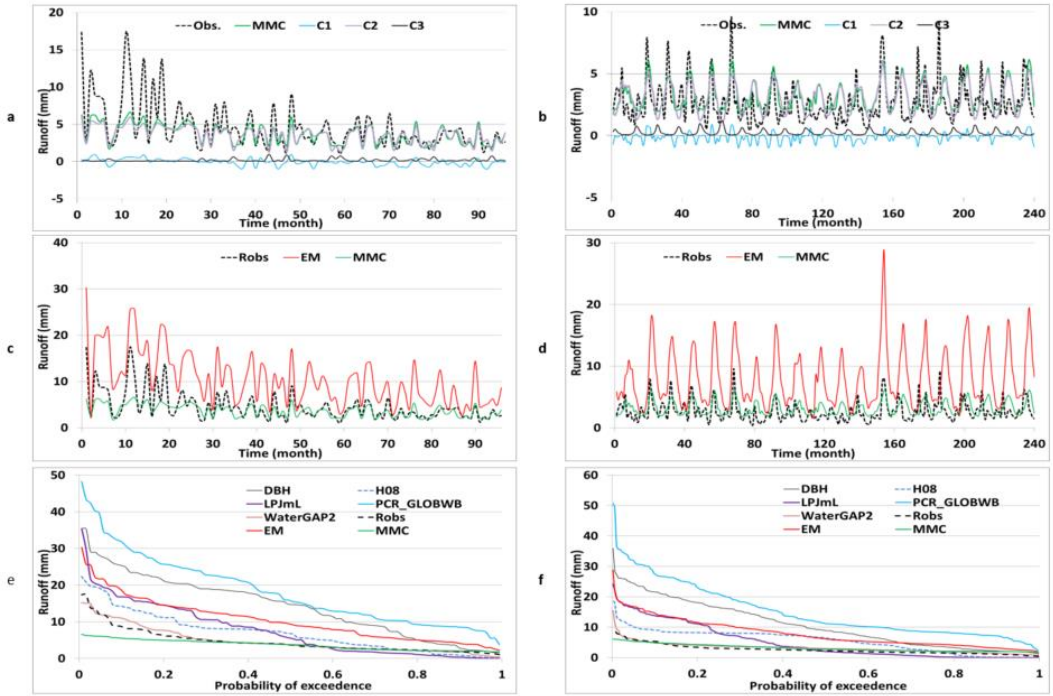
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

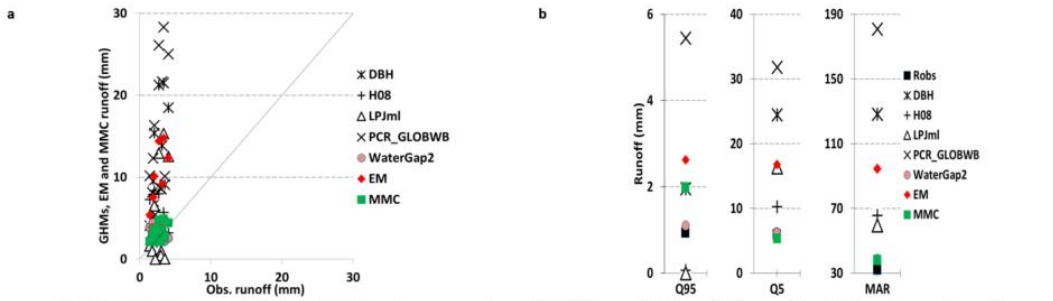
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	1.3	-1.6	1.5	-1.4	-2.1	-2.1	-3.3	-120.8	-613.2	-120.4	-152.4
Calibration	-1.0	-1.7	1.9	-1.4	-2.3	-2.0	-3.2	-118.0	-705.5	-87.5	-105.9
Validation	1.5	-1.0	2.1	-1.2	-1.5	-1.5	-2.2	-70.1	-41341.1	-73.6	-386.6



15. Yellow catchment at Huayuankou station.



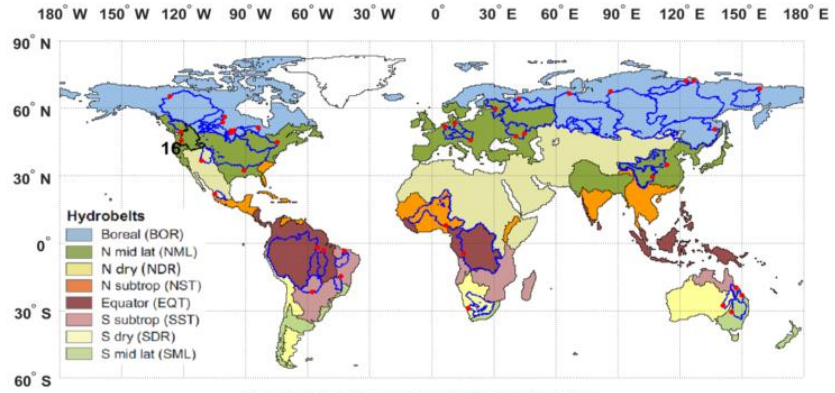
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



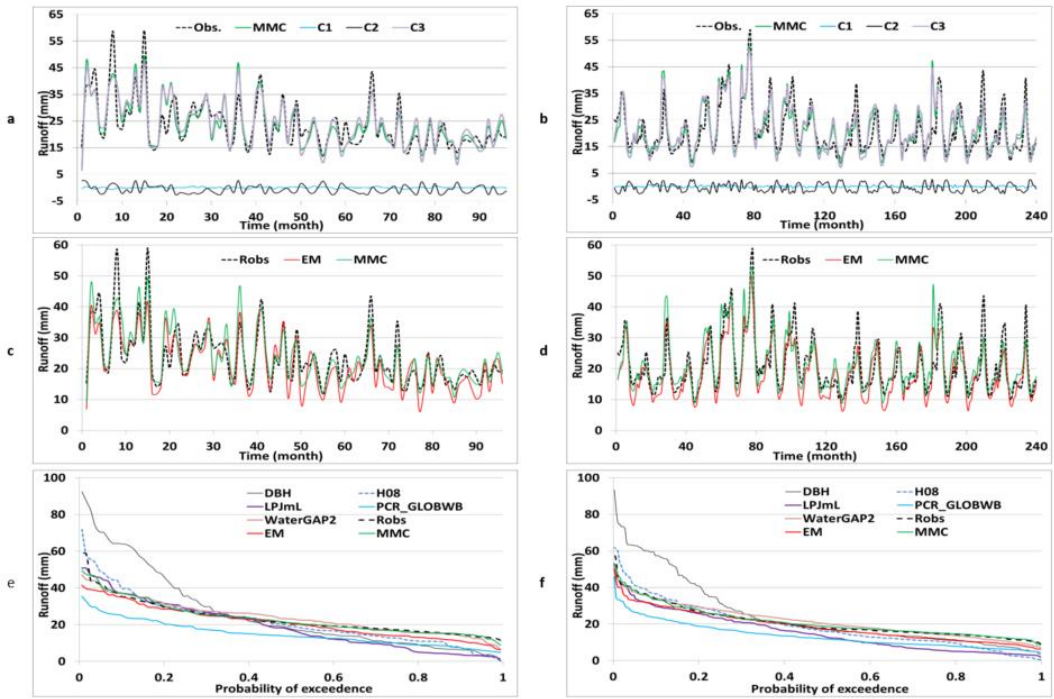
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

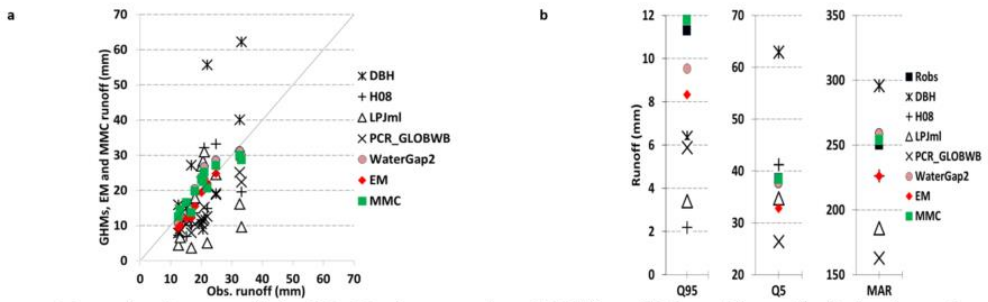
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJml	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	3.8	1.6	1.2	8.0	-1.3	1.7	-1.9	-460.5	-613.2	-63.6	-152.4
Calibration	5.4	1.7	1.7	9.3	-1.1	2.3	-1.9	-522.4	-705.5	-85.9	-105.9
Validation	23.4	5.5	7.4	44.9	1.5	9.7	1.2	-859.2	-41341.1	-33.0	-386.6



16. Columbia catchment at The Dalles, OREG. Station.



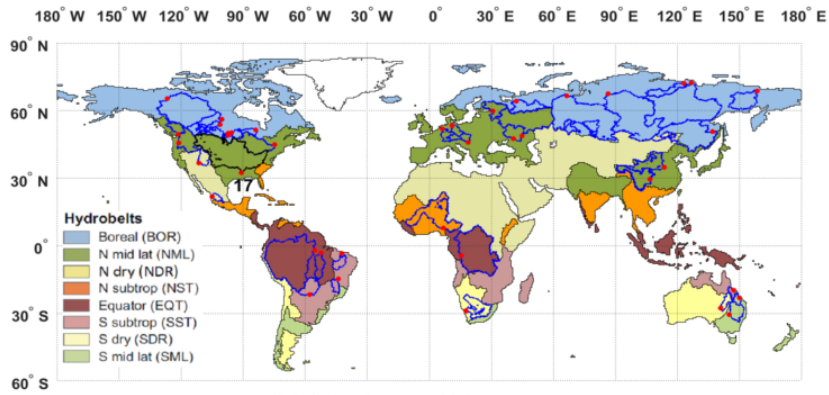
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



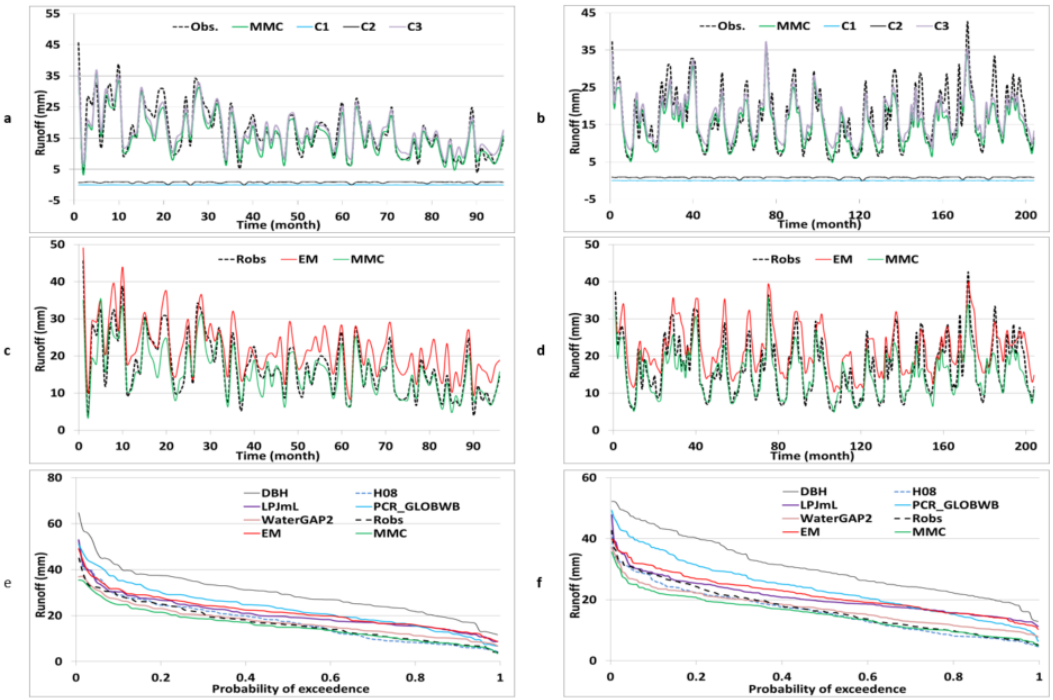
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

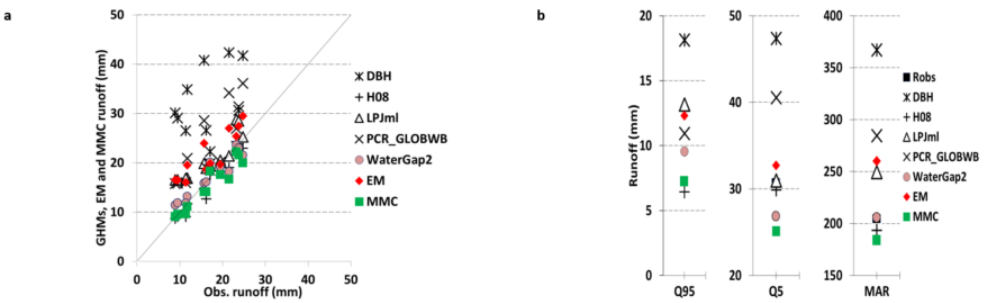
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	1.7	-1.0	1.4	-1.1	-2.6	-2.1	-3.1	-97.5	-613.2	-47.5	-152.4
Calibration	2.1	1.0	1.4	-1.1	-2.5	-2.4	-2.8	-39.1	-705.5	-28.1	-105.9
Validation	4.3	2.1	3.1	1.7	-1.1	-1.3	-1.2	6.5	-41341.1	-10.3	-386.6



17. Mississippi catchment at Vicksburg, MISS. Station.



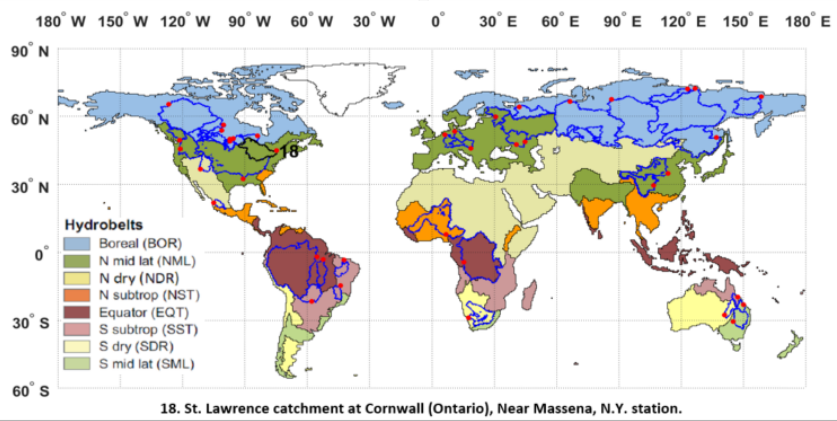
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



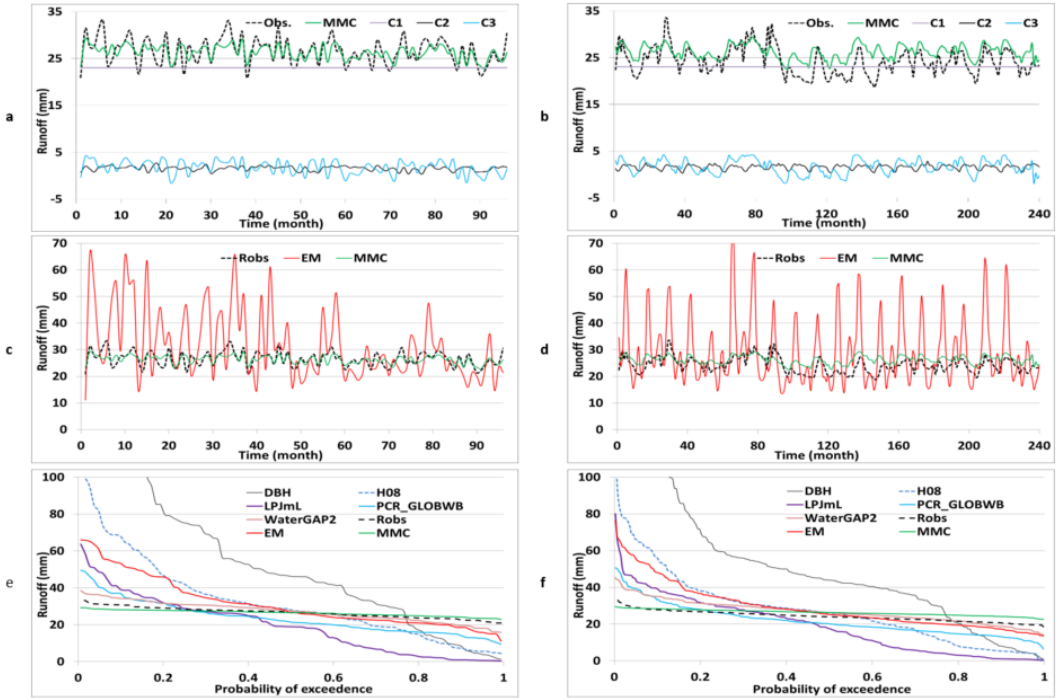
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	2.5	-2.8	-1.6	-1.2	-3.7	-1.5	-4.2	-262.4	-613.2	-50.9	-152.4
Calibration	1.9	-2.8	-2.0	-1.4	-3.9	-1.9	-4.2	-227.3	-705.5	-24.6	-105.9
Validation	5.0	-1.6	1.1	1.7	-1.9	1.2	-2.0	-418.0	-41341.1	-14.2	-386.6

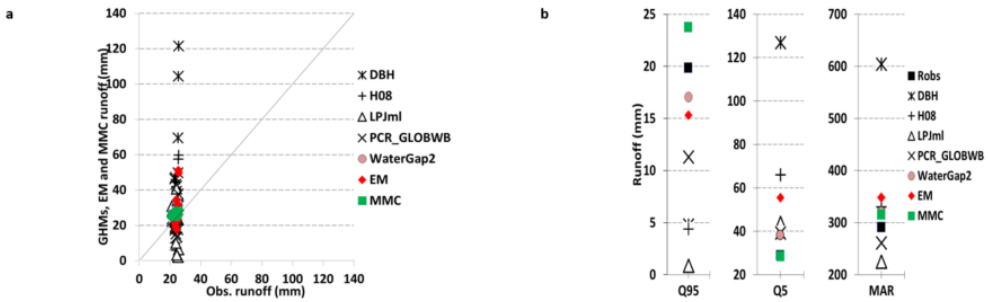


18. St. Lawrence catchment at Cornwall (Ontario), Near Massena, N.Y. station.



Calibration Validation

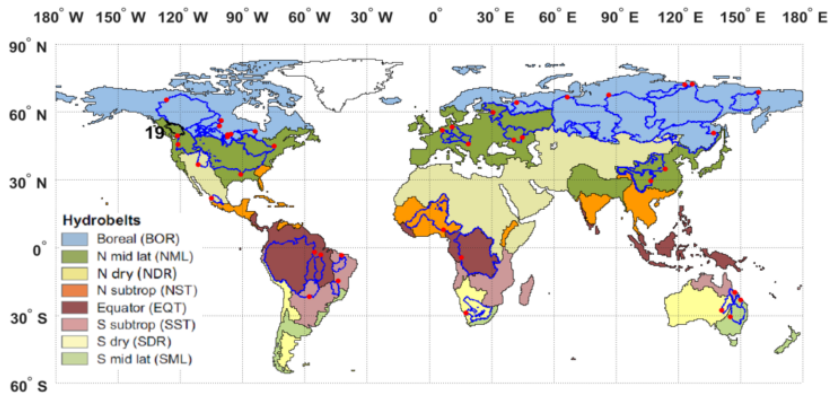
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



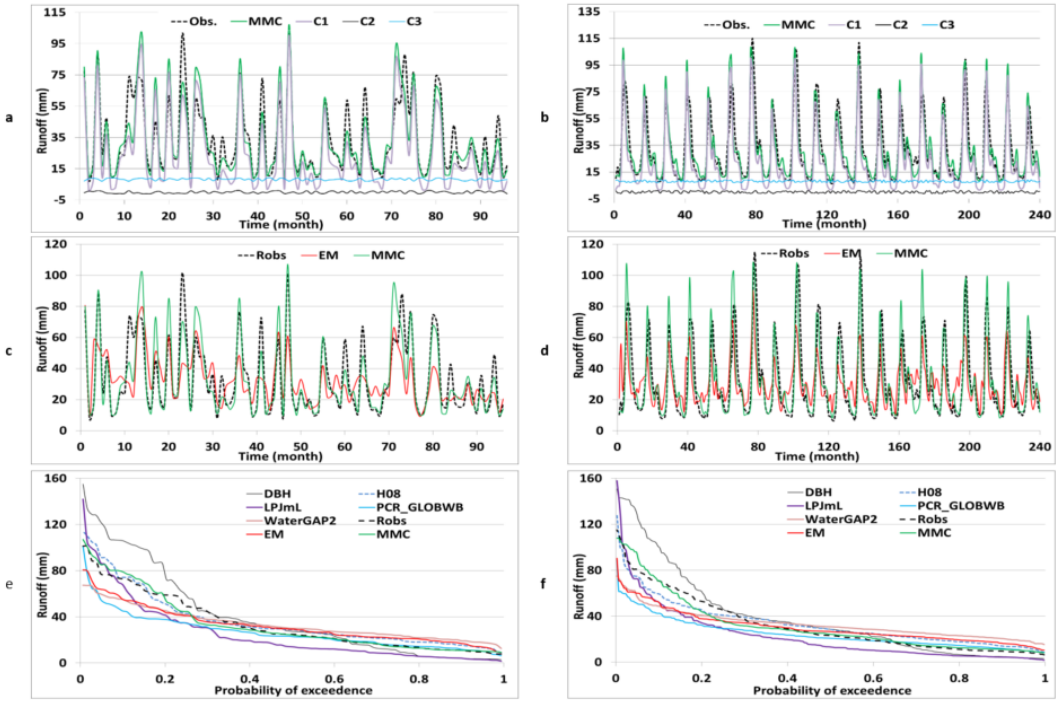
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

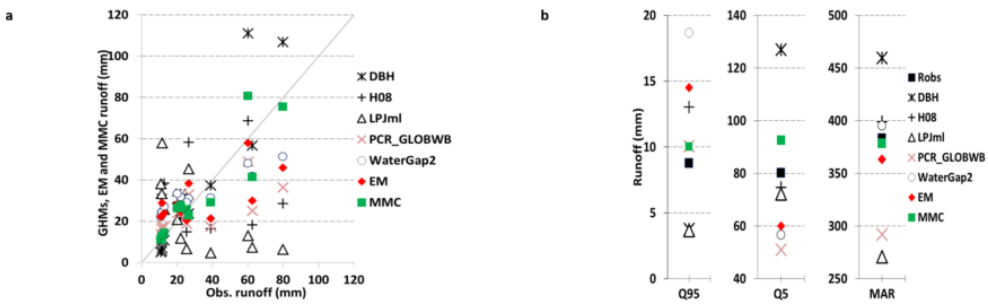
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	61.8	15.9	13.6	3.6	1.4	5.7	-2.0	-871.4	-613.2	-547.8	-152.4
Calibration	86.3	20.0	12.6	3.5	1.2	7.5	-2.1	-1061.4	-705.5	-533.7	-105.9
Validation	375.2	75.4	56.9	14.0	7.1	31.6	2.5	-2914.3	-41341.1	-461.9	-386.6



19. Fraser River catchment at Hope station.



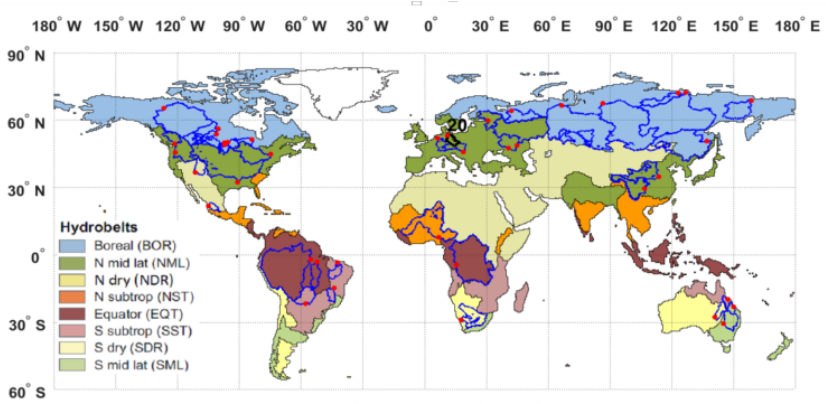
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



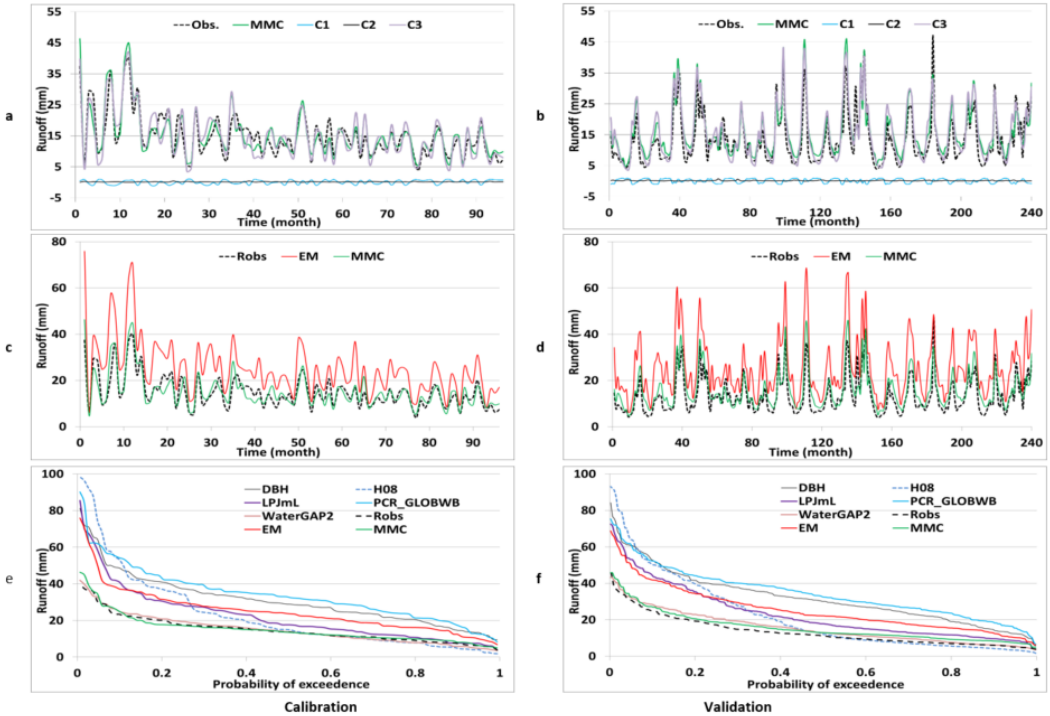
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

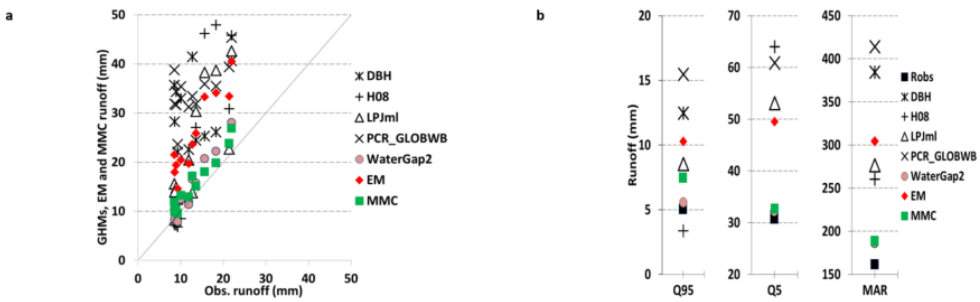
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	-2.4	-1.2	1.5	-2.1	-2.4	-2.0	-4.1	-211.9	-613.2	-167.4	-152.4
Calibration	-2.1	-1.1	1.5	-2.1	-2.7	-2.1	-4.2	-208.5	-705.5	-140.4	-105.9
Validation	1.2	2.5	4.1	1.1	1.2	1.3	-1.6	-392.4	-41341.1	-477.3	-386.6



20. Labe catchment at Neu-Darchau station.



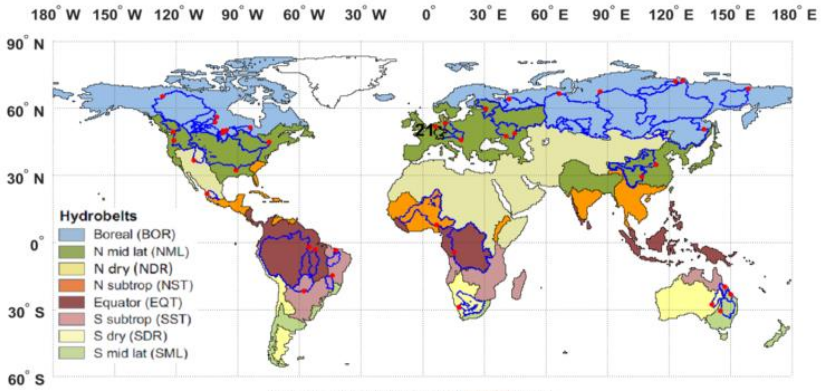
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



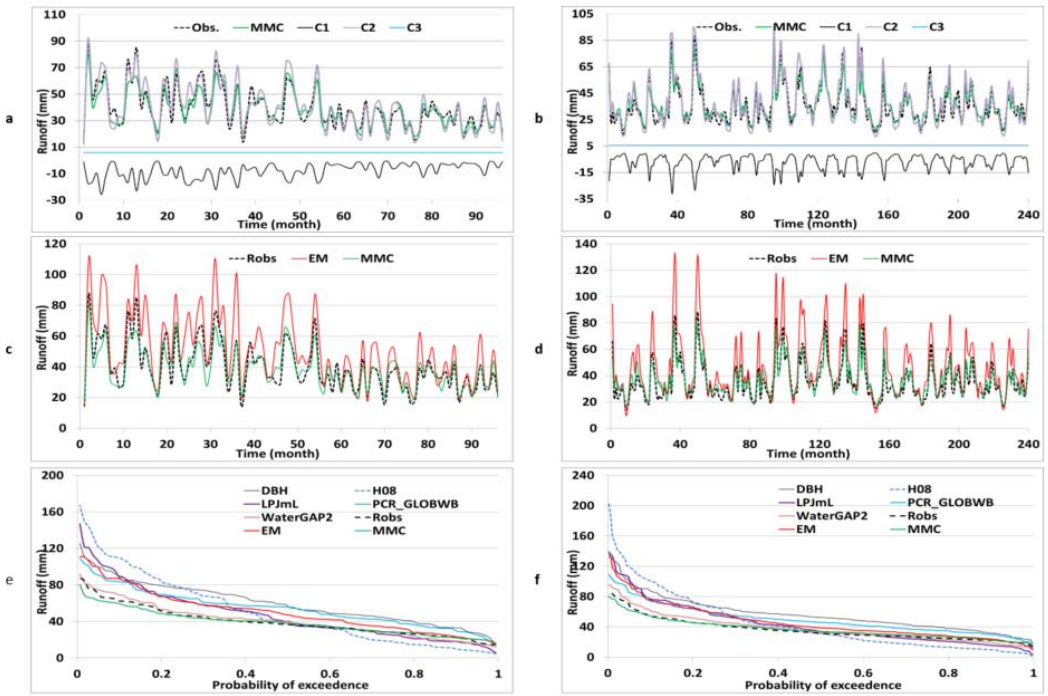
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

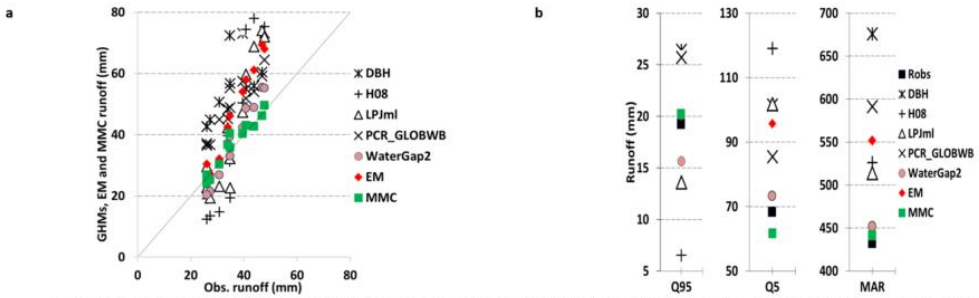
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	4.6	3.2	2.1	5.3	-2.2	1.9	-3.3	-622.3	-613.2	-118.5	-152.4
Calibration	3.5	3.0	1.6	4.5	-2.7	1.6	-3.0	-557.7	-705.5	-27.8	-105.9
Validation	6.7	4.1	3.0	7.7	-1.5	3.1	-1.4	-554.0	-41341.1	2.0	-386.6



21. Rhine River catchment at Lobith station.



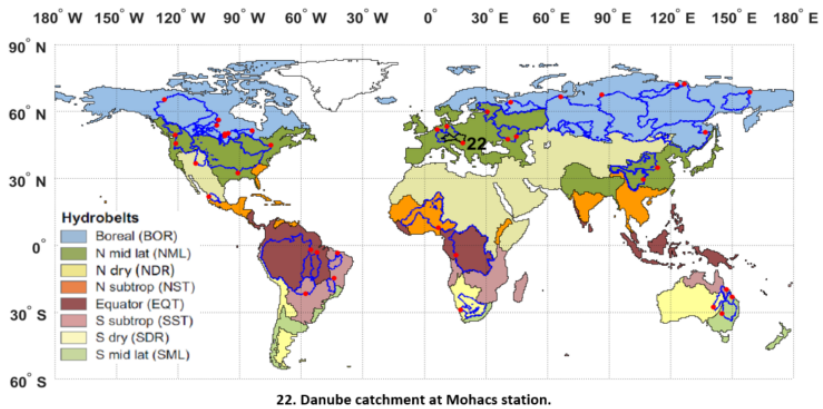
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



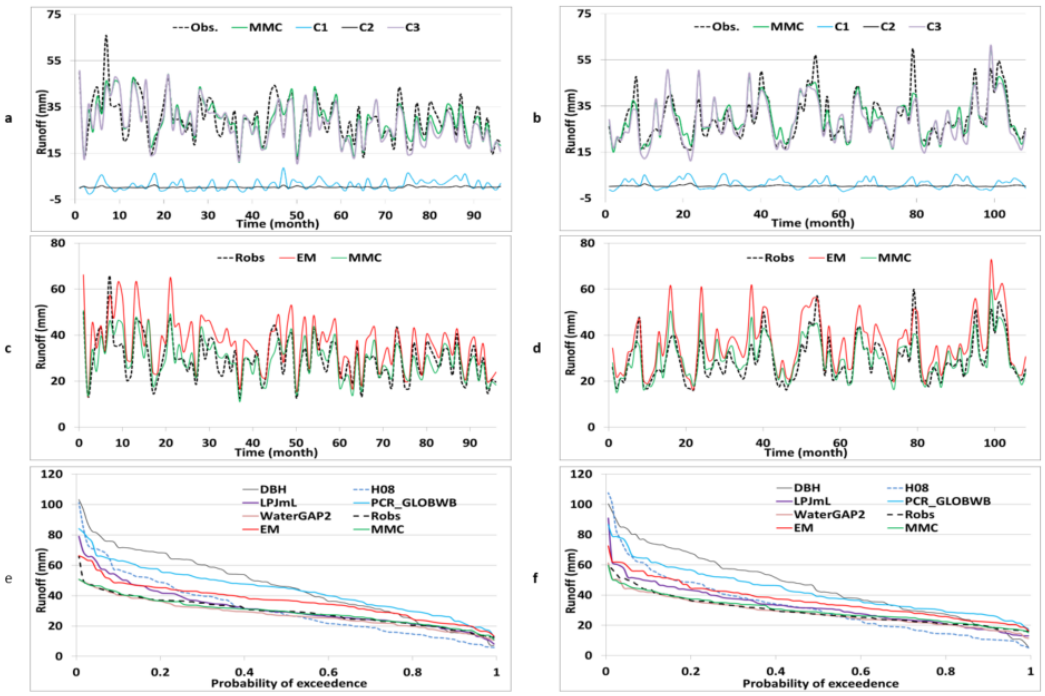
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

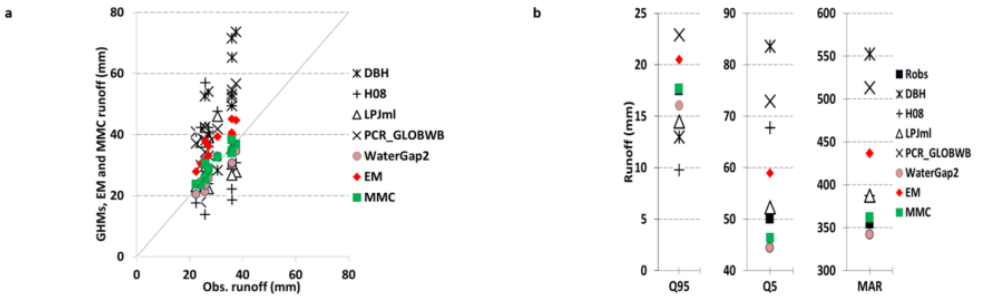
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	1.9	2.4	1.2	1.1	-2.3	-1.2	-3.2	-201.5	-613.2	-85.0	-152.4
Calibration	1.5	2.0	1.1	-1.1	-2.9	-1.3	-3.5	-214.3	-705.5	-56.3	-105.9
Validation	2.6	3.3	1.5	1.4	-2.0	1.2	-2.5	-462.8	-41341.1	-50.7	-386.6



22. Danube catchment at Mohacs station.



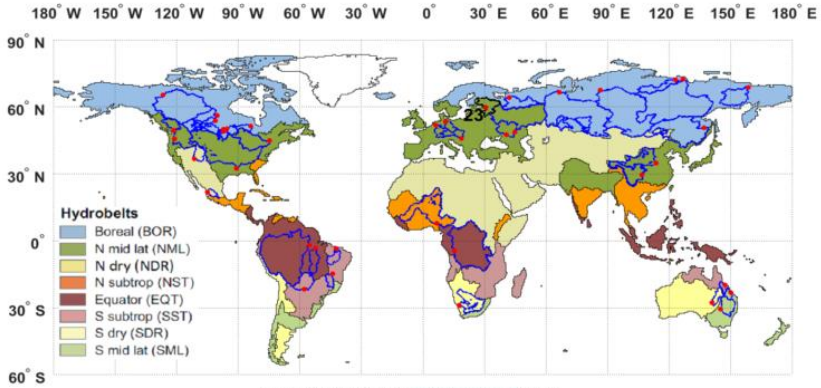
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



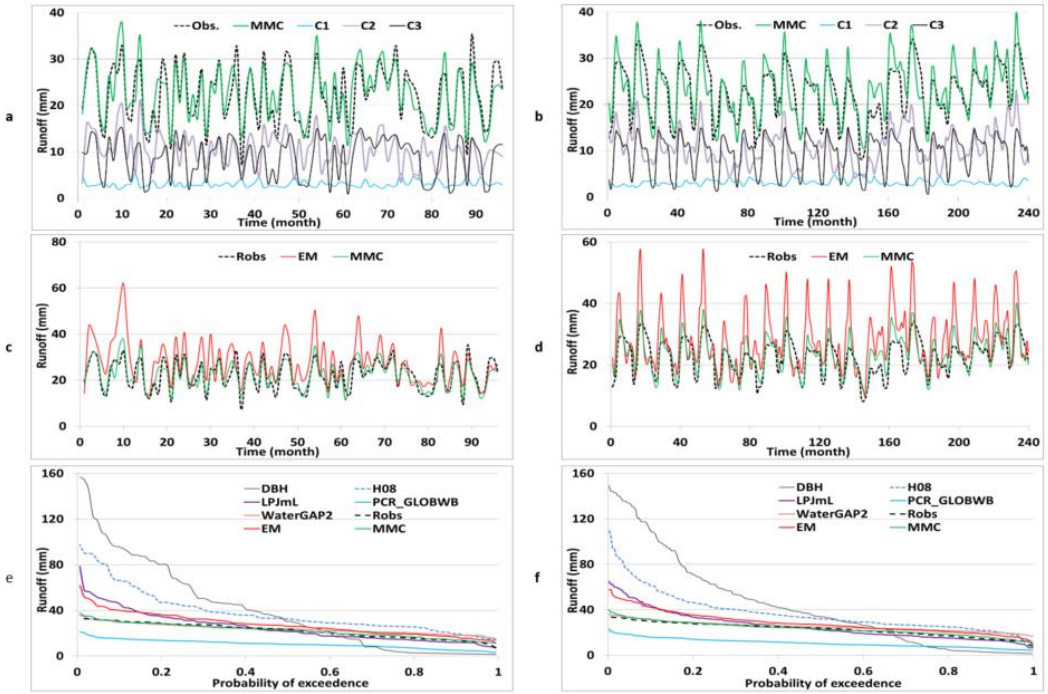
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

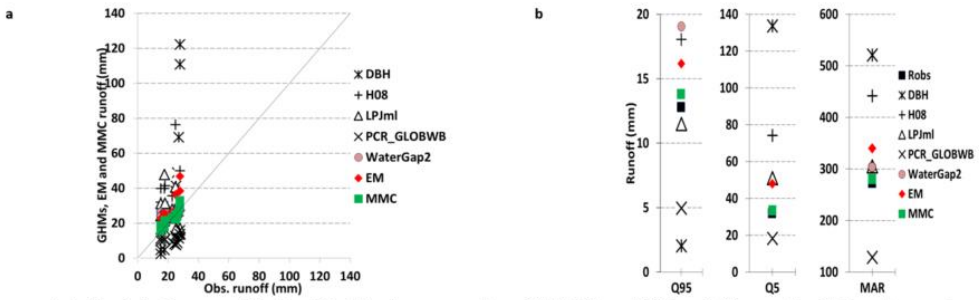
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	2.8	2.0	1.0	1.4	-2.6	-1.6	-3.3	-171.7	-613.2	-75.6	-152.4
Calibration	2.5	1.7	1.0	1.5	-2.5	-1.6	-2.9	-138.8	-705.5	-45.6	-105.9
Validation	4.0	2.7	1.3	2.1	-1.9	-1.1	-2.2	-114.9	-41341.1	-31.9	-386.6



23. Neva catchment at Novosaratovka station.



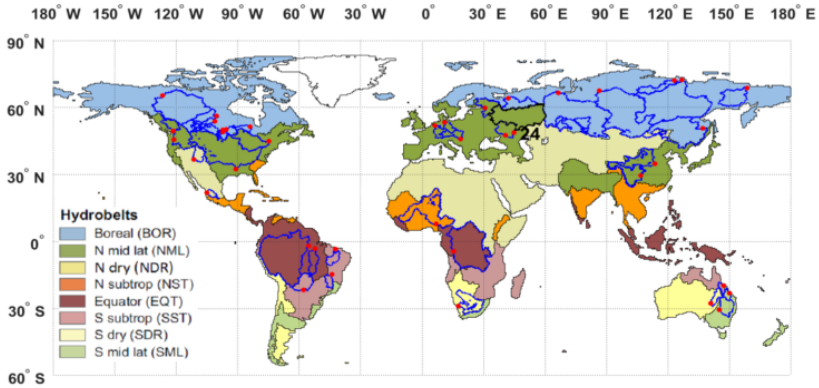
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



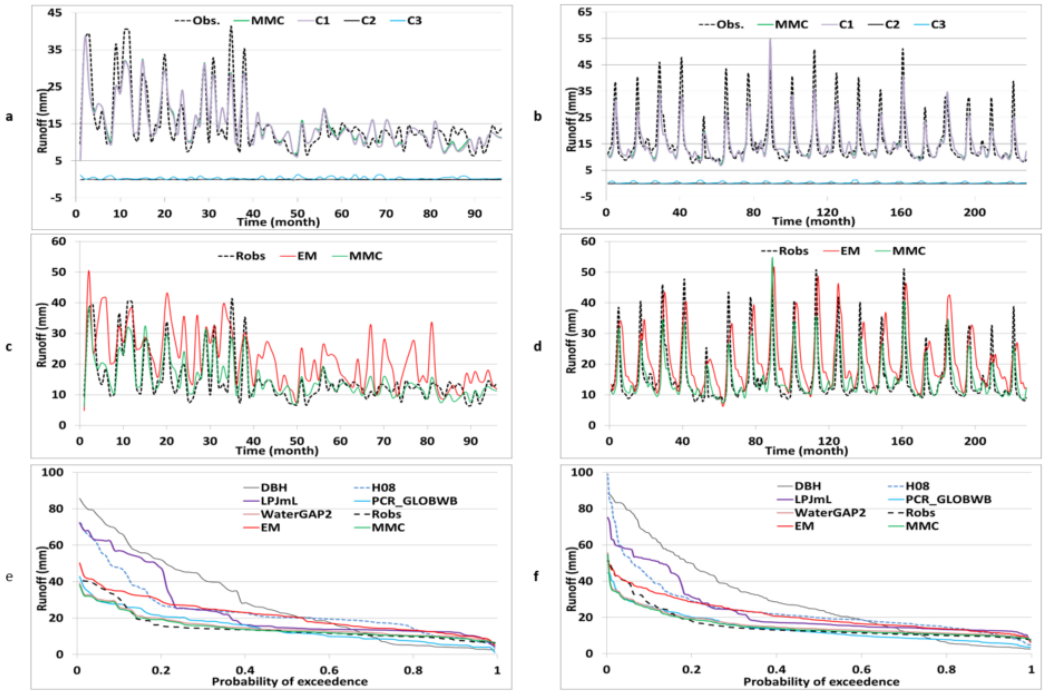
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

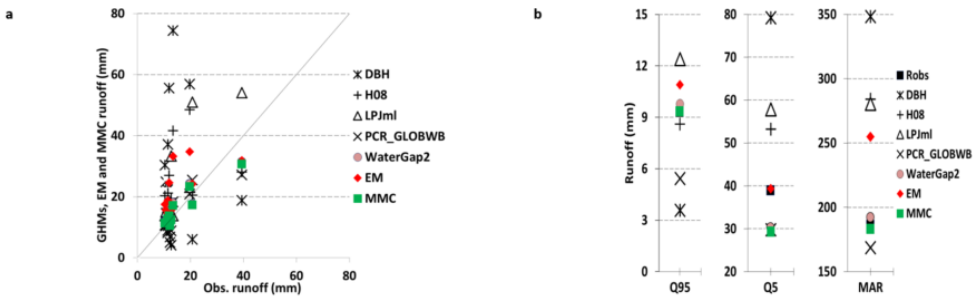
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	10.9	3.8	2.4	2.0	-1.5	-1.2	-3.0	-175.4	-613.2	-144.4	-152.4
Calibration	10.8	4.5	2.2	1.6	-1.5	-1.1	-3.1	-200.1	-705.5	-162.4	-105.9
Validation	83.4	25.6	12.2	8.9	2.4	4.7	1.1	-364.7	-41341.1	-132.7	-386.6



24. Volga catchment at Volgograd Power Plant station.



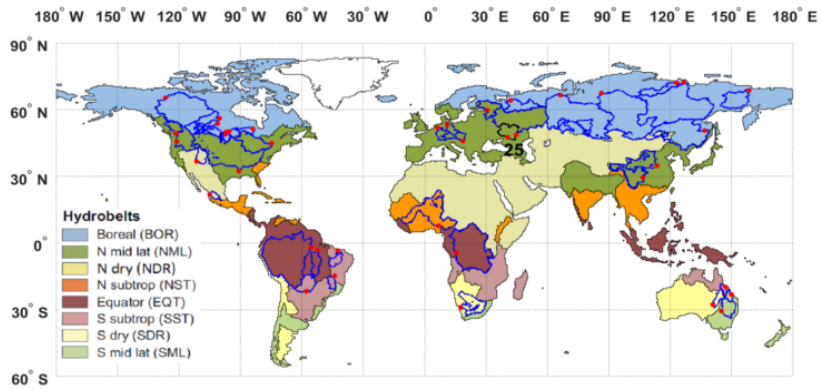
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



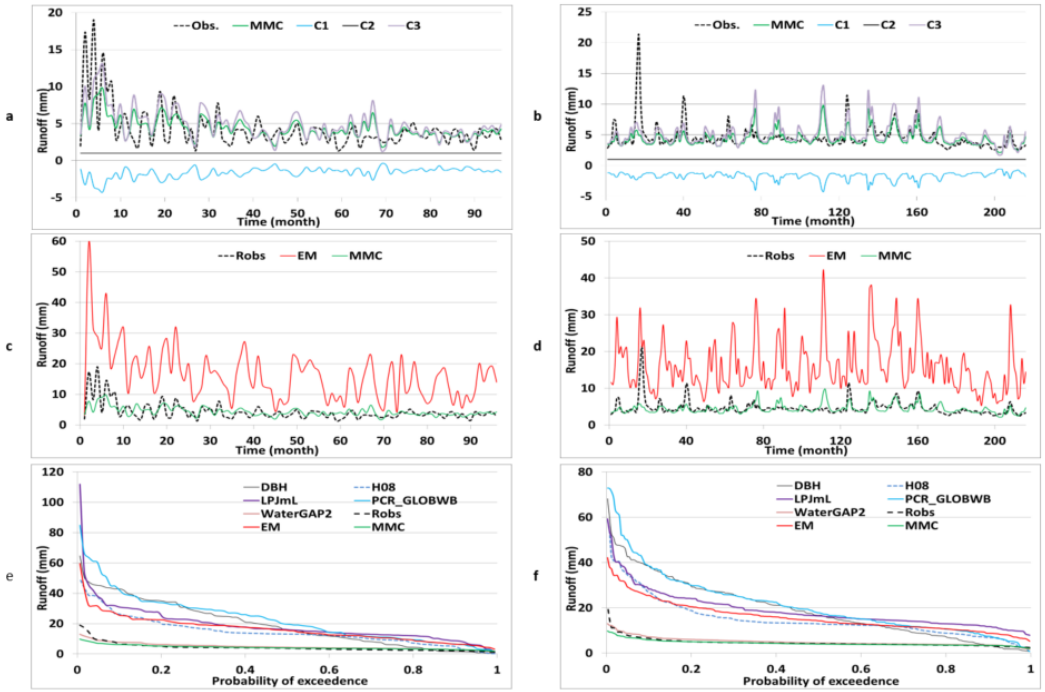
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

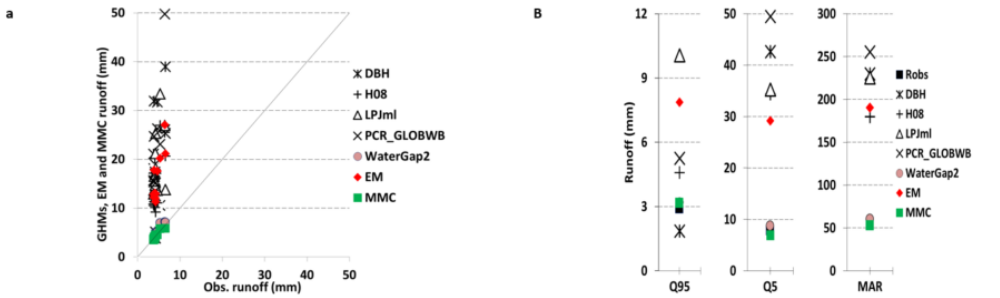
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	5.8	2.4	2.0	-1.2	-2.4	1.3	-2.6	-489.3	-613.2	-23.6	-152.4
Calibration	5.3	2.2	2.1	-1.4	-2.4	1.2	-2.7	-487.0	-705.5	-24.2	-105.9
Validation	6.8	2.8	1.9	-1.3	-1.8	1.5	-2.0	-450.0	-41341.1	-23.1	-386.6



25. Don catchment at Razdorskaya station.



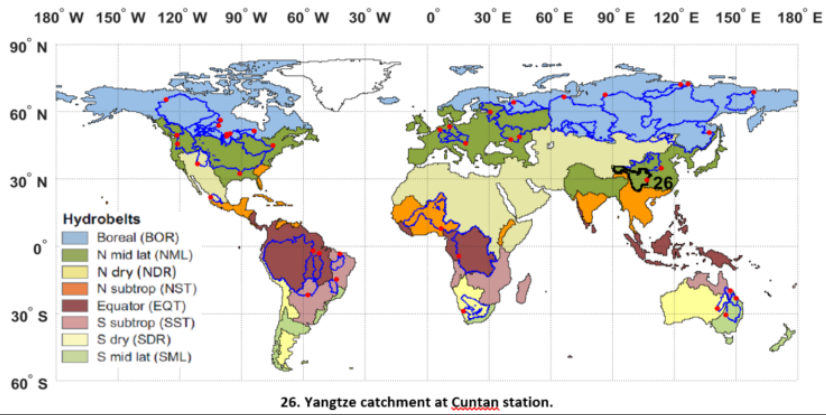
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



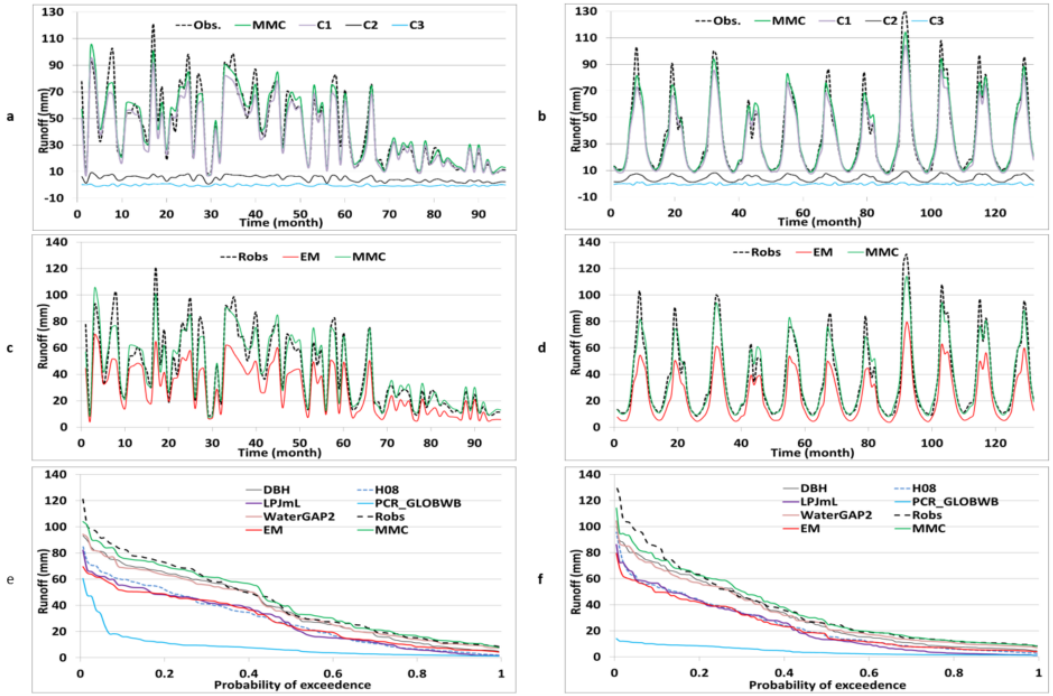
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

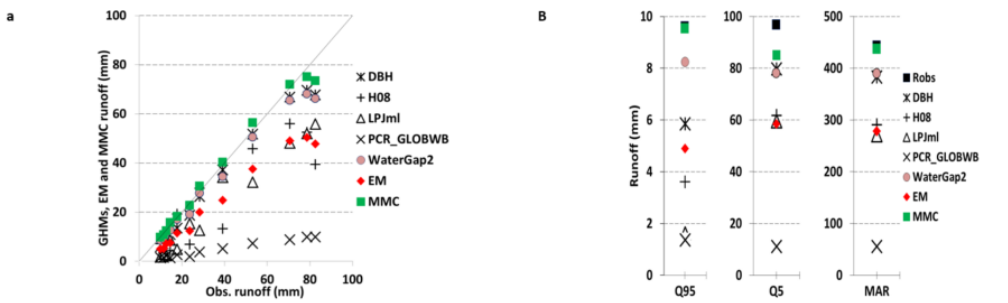
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	36.2	18.2	27.0	41.7	-1.3	15.9	-1.6	-1850.7	-613.2	-24.8	-152.4
Calibration	20.2	9.1	17.2	26.6	-1.3	9.7	-1.5	-1227.9	-705.5	-19.9	-105.9
Validation	83.5	39.9	58.8	100.1	1.5	37.1	1.2	-3591.7	-41341.1	-31.3	-386.6



26. Yangtze catchment at Cuntan station.



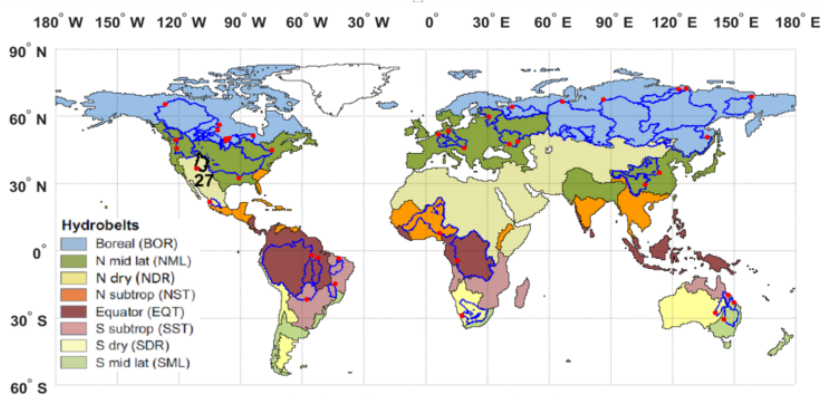
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



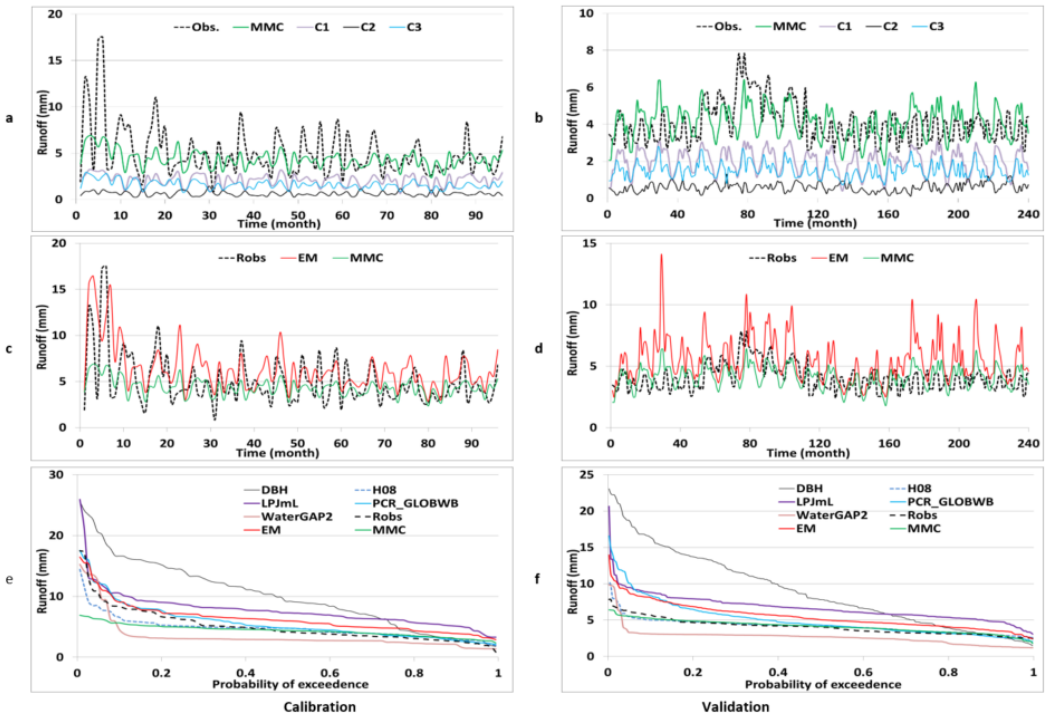
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

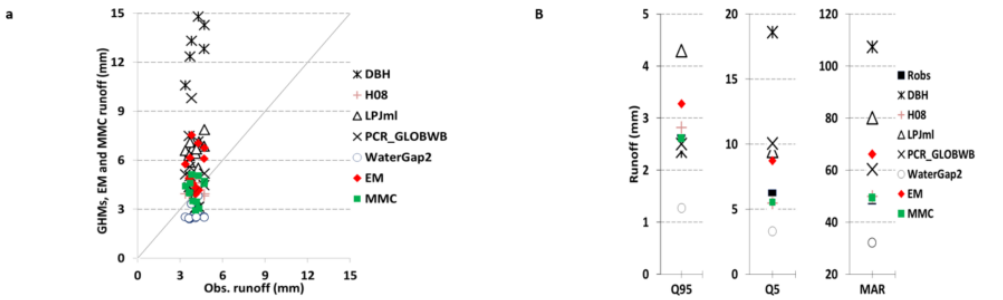
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJml	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	-3.4	-1.7	-1.6	1.5	-4.7	-1.9	-5.9	-404.7	-613.2	-127.3	-152.4
Calibration	-4.7	-1.8	-2.0	1.5	-4.9	-2.1	-5.9	-384.9	-705.5	-102.0	-105.9
Validation	-2.4	-1.1	-1.1	2.8	-3.1	-1.1	-4.2	-301.0	-41341.1	-107.7	-386.6



27. Colorado catchment at Lees Ferry Ariz station.



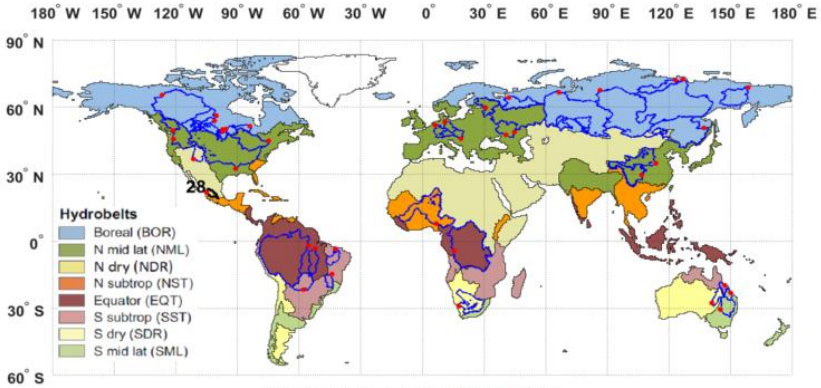
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



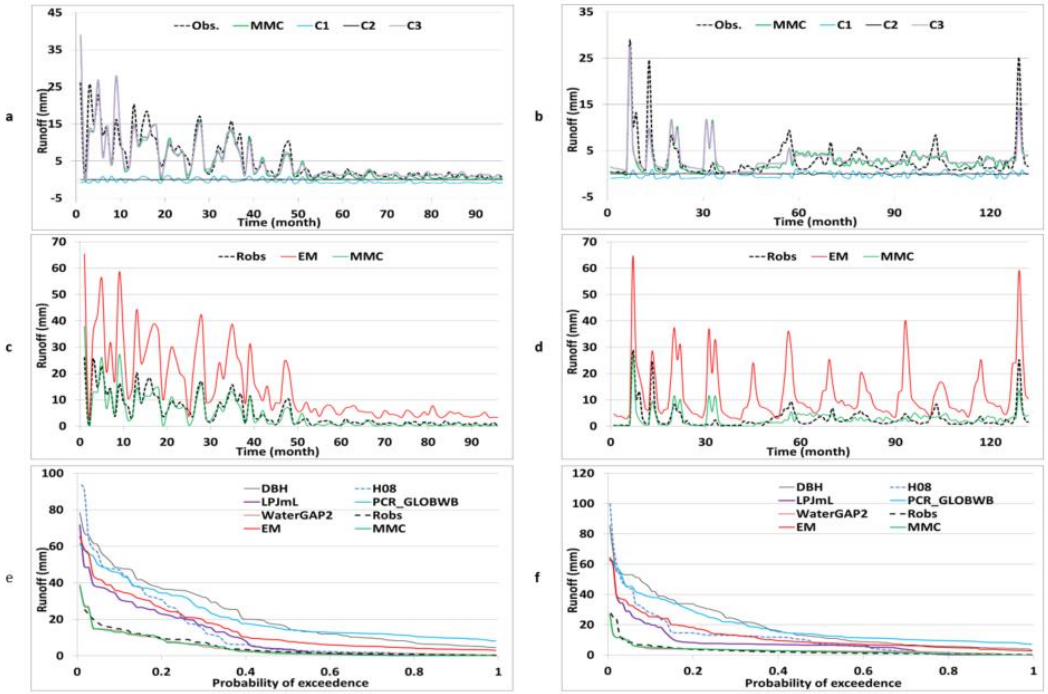
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

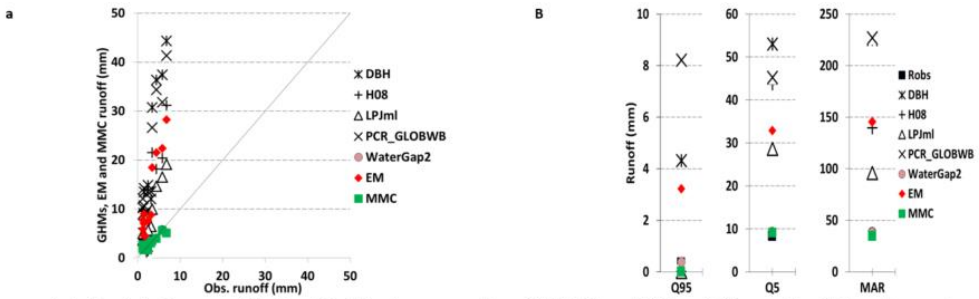
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJml	PCR_GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	3.2	-1.4	2.1	1.1	-1.0	-1.0	-1.7	-64.4	-613.2	-26.5	-152.4
Calibration	3.0	-1.3	1.4	-1.1	-1.1	-1.2	-1.5	-35.1	-705.5	-24.3	-105.9
Validation	52.9	2.5	12.1	8.5	4.6	6.4	2.2	-422.1	-41341.1	-28.6	-386.6



28. Santiago catchment at El Capomal station.



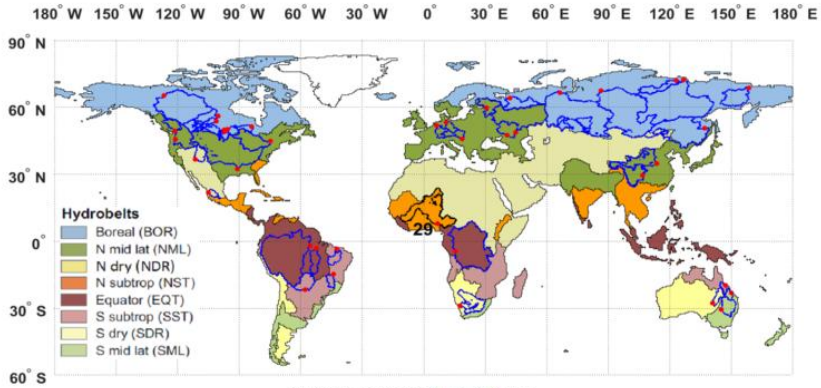
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



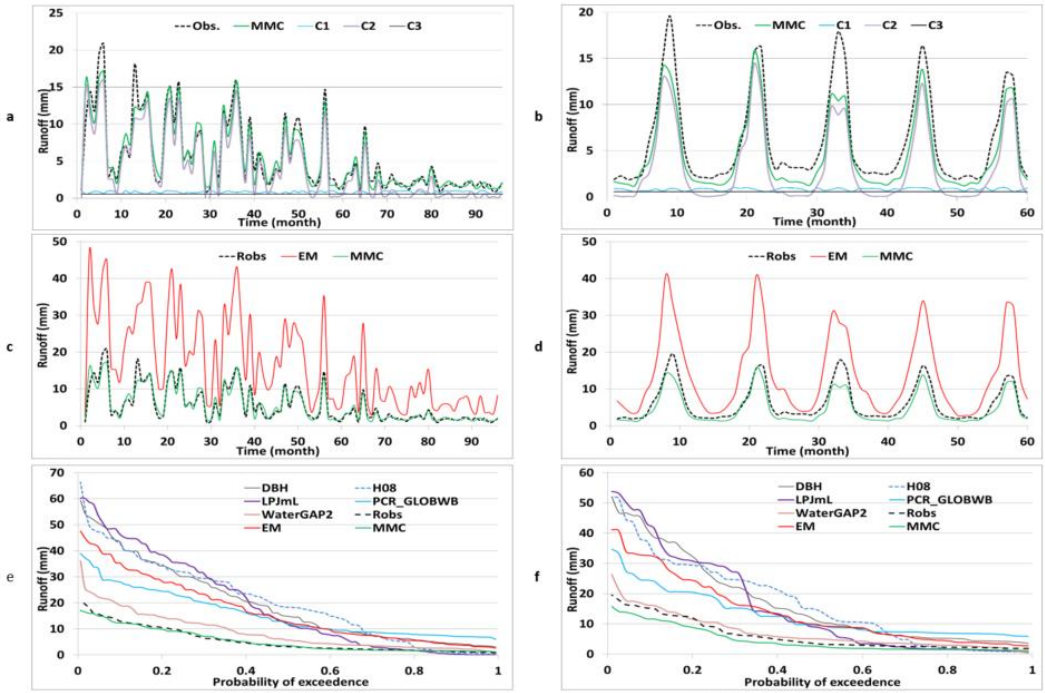
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

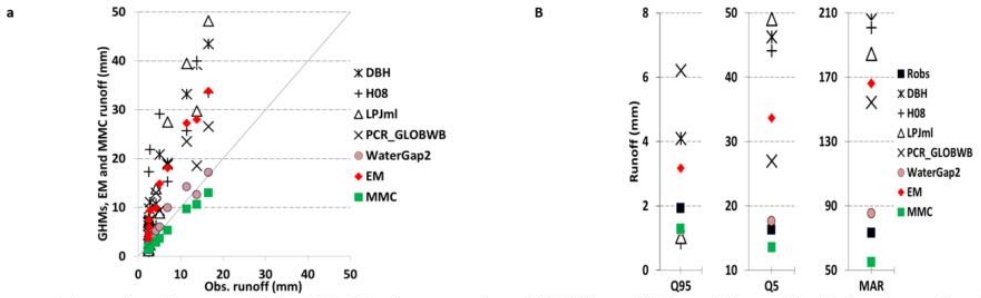
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	5.7	4.0	1.6	5.2	-3.5	2.4	-3.6	-691.7	-613.2	-5.1	-152.4
Calibration	7.9	5.2	2.2	6.2	-2.9	3.1	-3.0	-717.2	-705.5	-12.0	-105.9
Validation	15.1	8.3	3.8	15.0	1.4	7.3	1.2	-617.4	-41341.1	-19.3	-386.6



29. Niger catchment at Lokoja station.



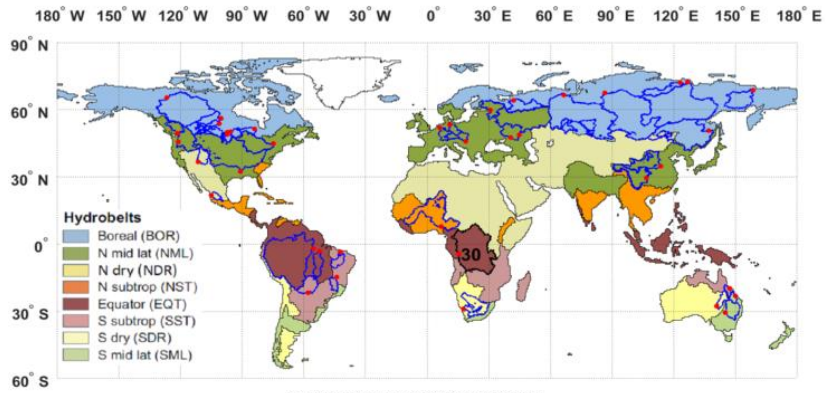
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



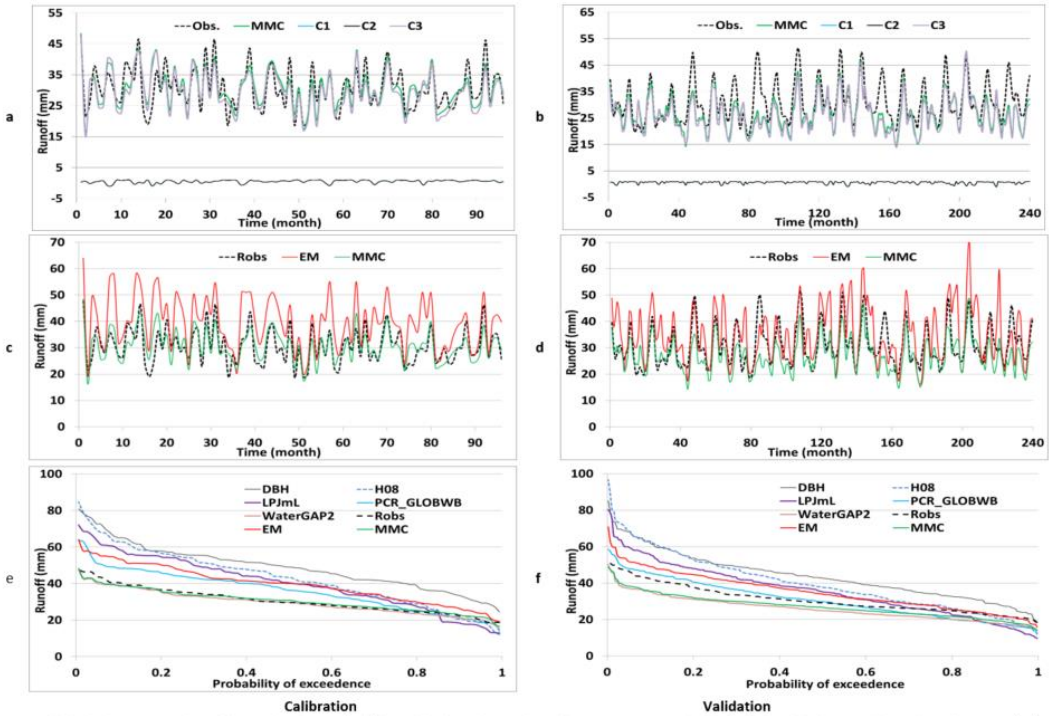
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

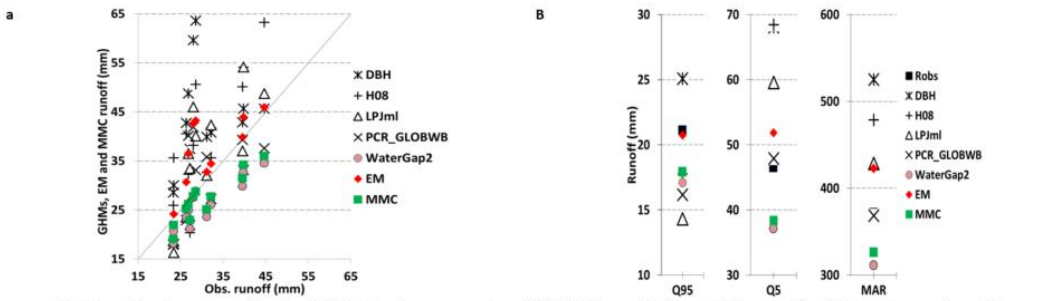
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	7.1	8.2	7.0	3.7	-1.4	4.0	-5.3	-1031.4	-613.2	-393.7	-152.4
Calibration	8.7	10.0	10.5	4.2	-1.0	5.4	-4.4	-1071.1	-705.5	-333.4	-105.9
Validation	9.7	10.7	10.0	3.6	-1.4	4.9	-1.8	-763.7	-41341.1	-41.3	-386.6



30. Zaire catchment at Kinshasa station.



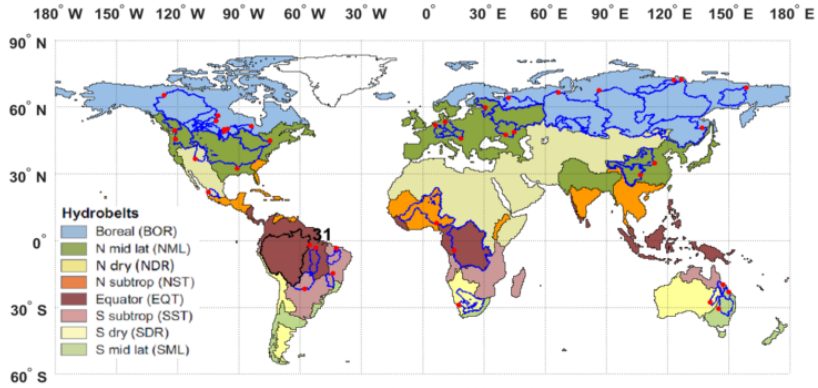
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



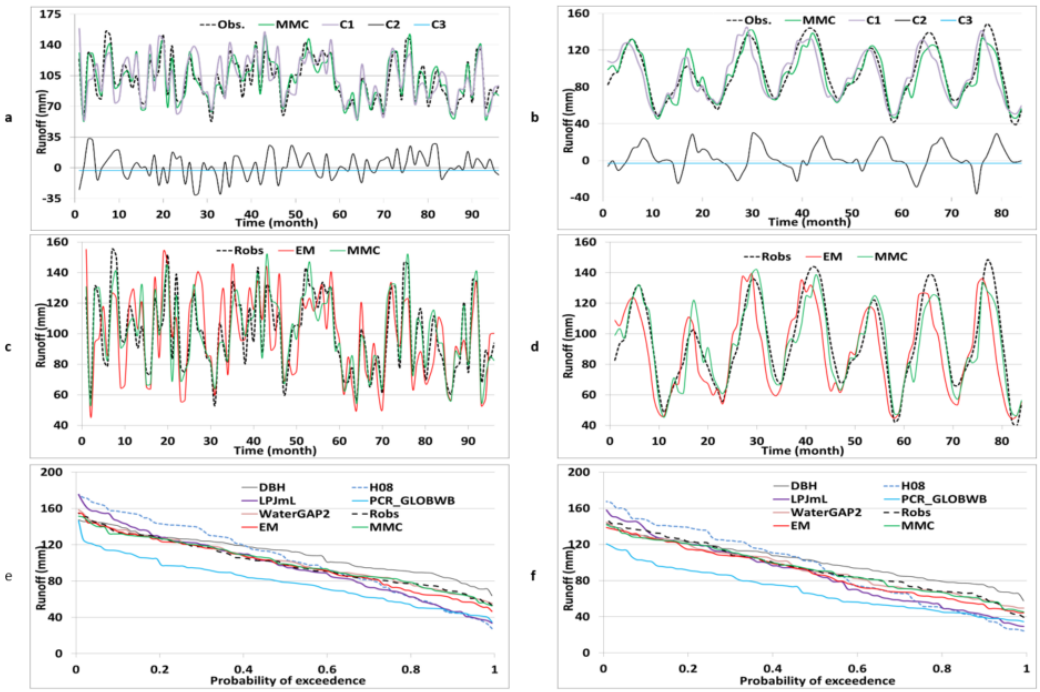
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

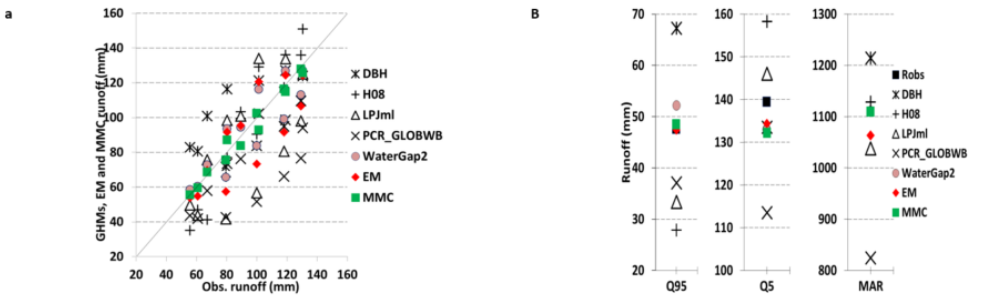
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	3.6	2.2	1.8	1.3	-2.3	1.4	-2.5	-486.7	-613.2	-24.6	-152.4
Calibration	3.3	2.2	1.7	-1.0	-2.4	1.2	-2.7	-485.0	-705.5	-30.2	-105.9
Validation	8.3	5.9	3.9	2.5	1.8	2.4	1.4	-98.4	-41341.1	-36.3	-386.6



31. Amazonas catchment at Obidos station.



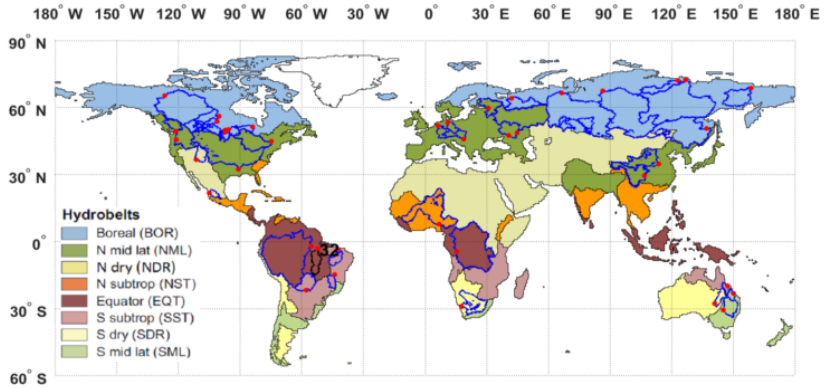
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



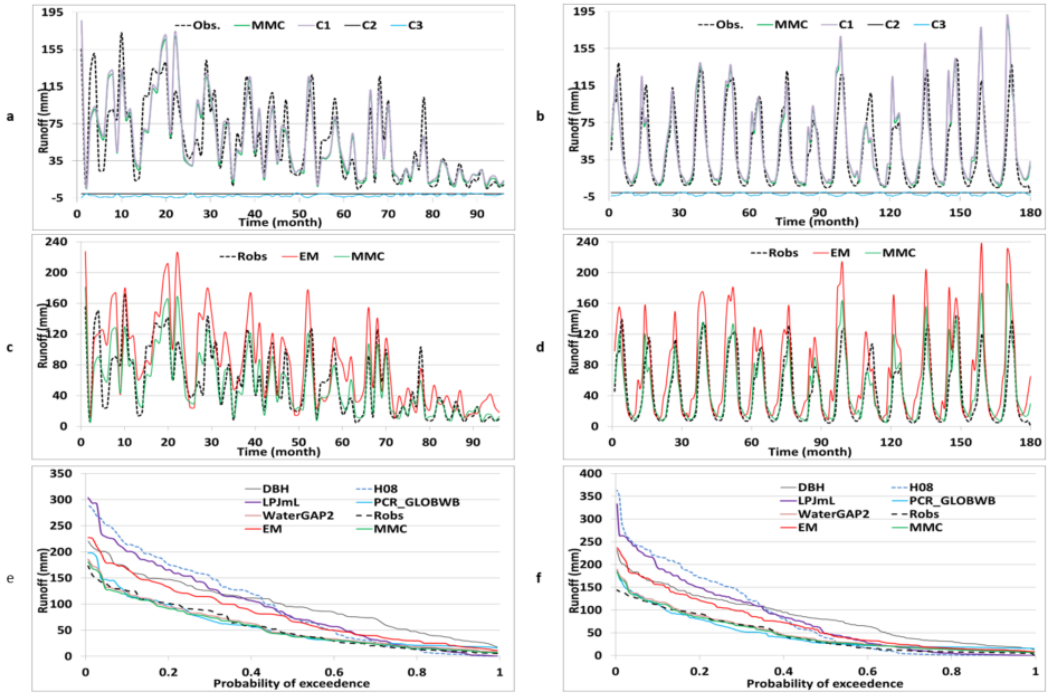
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

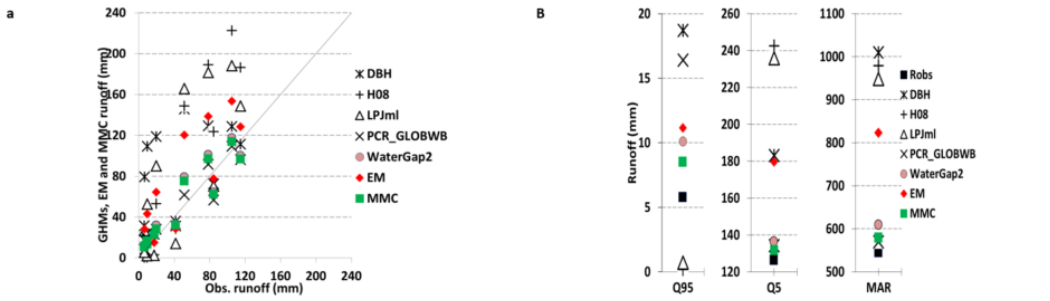
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	-1.4	-1.6	-1.2	-1.0	-2.7	-2.1	-4.7	-267.0	-613.2	-206.1	-152.4
Calibration	-1.3	-1.7	-1.1	-1.0	-2.3	-1.8	-3.8	-196.1	-705.5	-153.0	-105.9
Validation	2.0	1.5	2.6	3.4	-1.1	1.3	-1.8	-410.2	-41341.1	-75.0	-386.6



32. Xingu catchment at Altamira station.



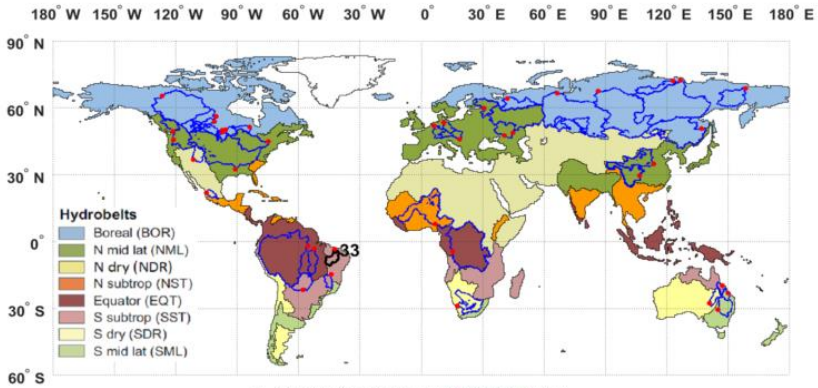
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



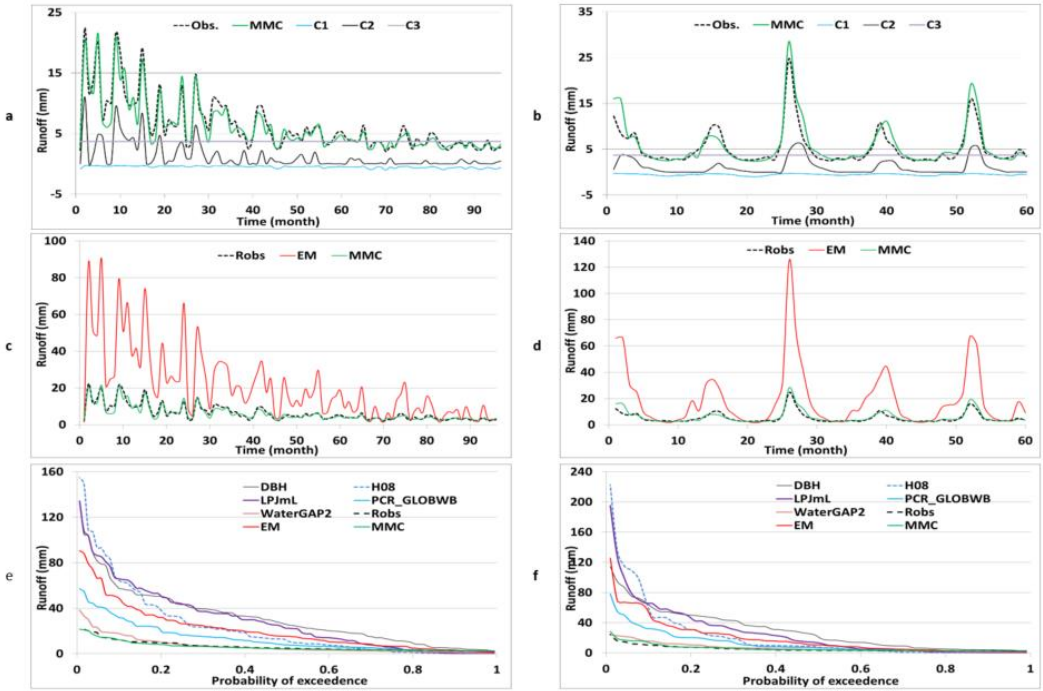
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

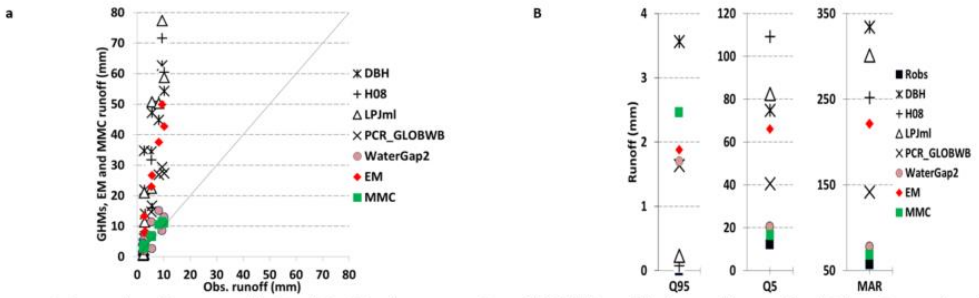
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	1.6	1.4	1.3	-2.0	-3.0	-1.4	-3.3	-189.3	-613.2	-24.8	-152.4
Calibration	1.8	1.3	1.5	-2.3	-3.0	-1.2	-3.2	-191.9	-705.5	-15.9	-105.9
Validation	5.9	4.7	4.9	1.1	1.2	2.6	1.0	-161.0	-41341.1	-8.6	-386.6



33. Rio Parnaiba catchment at Luzilandia station.



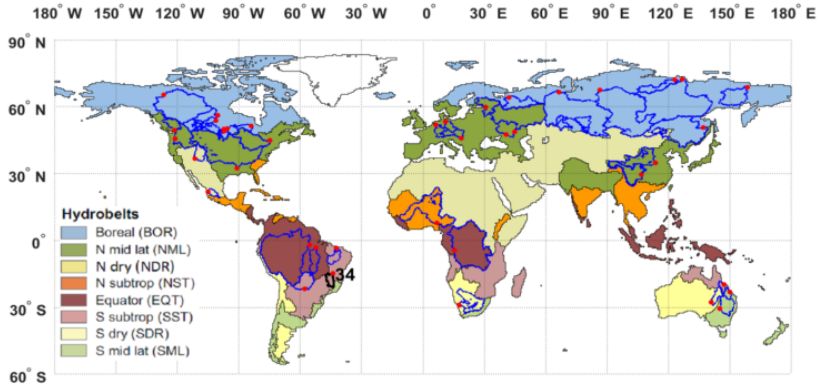
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



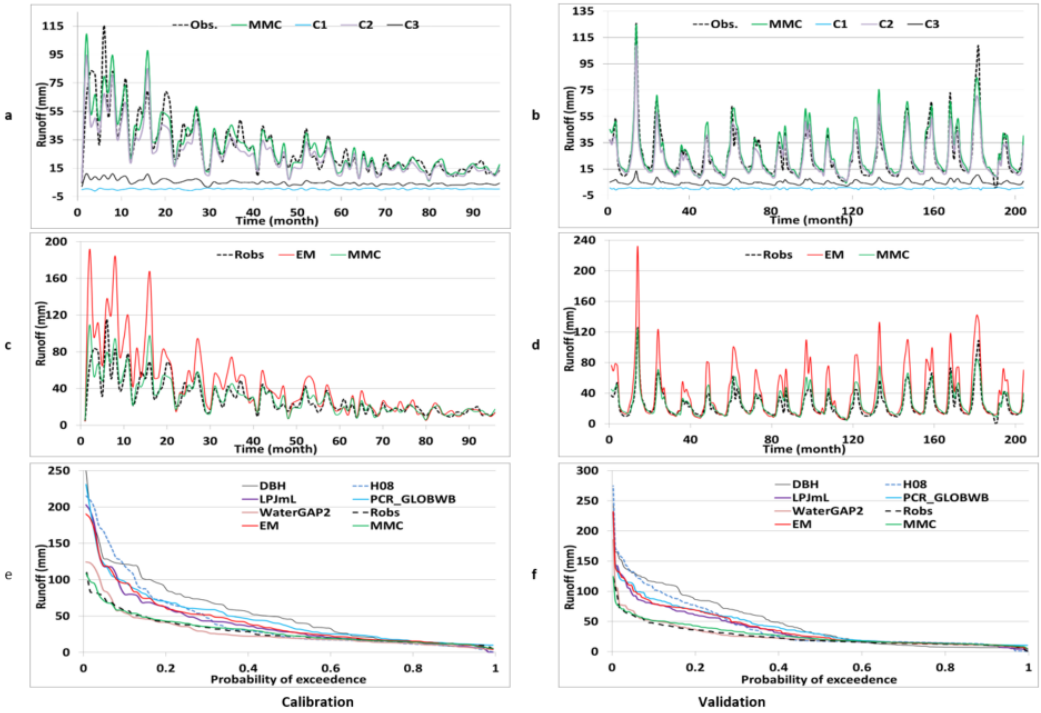
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

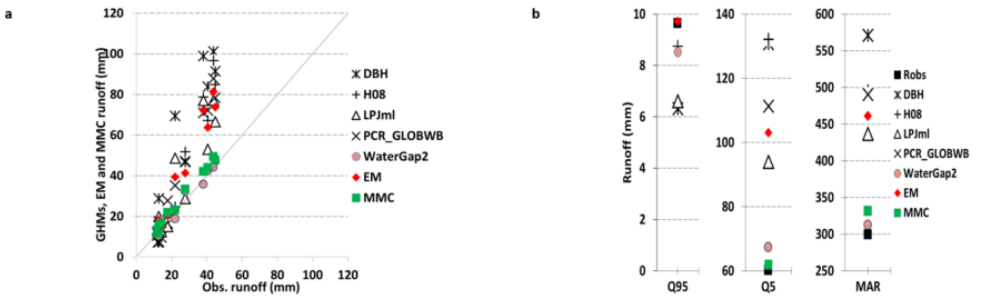
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	26.3	28.1	26.2	3.2	1.2	11.5	-3.2	-1569.6	-613.2	-637.5	-152.4
Calibration	35.6	35.3	36.5	4.3	1.3	14.9	-2.9	-1880.7	-705.5	-614.4	-105.9
Validation	48.8	70.8	63.4	8.4	1.5	25.4	-2.3	-2868.6	-41341.1	-574.0	-386.6



34. SAO Francisco catchment at MANGA station.



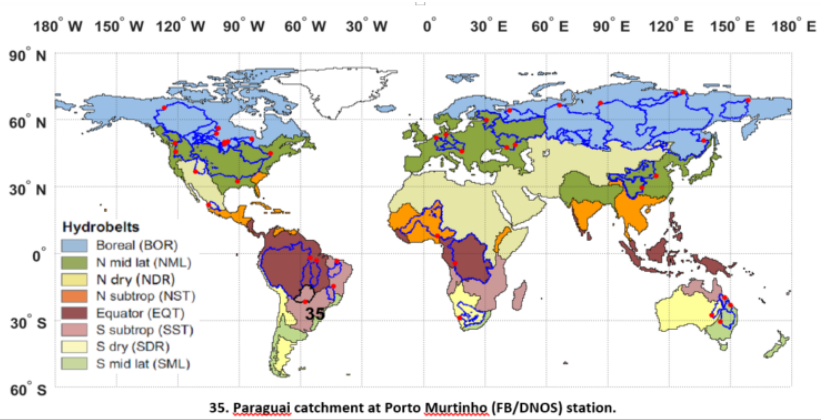
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



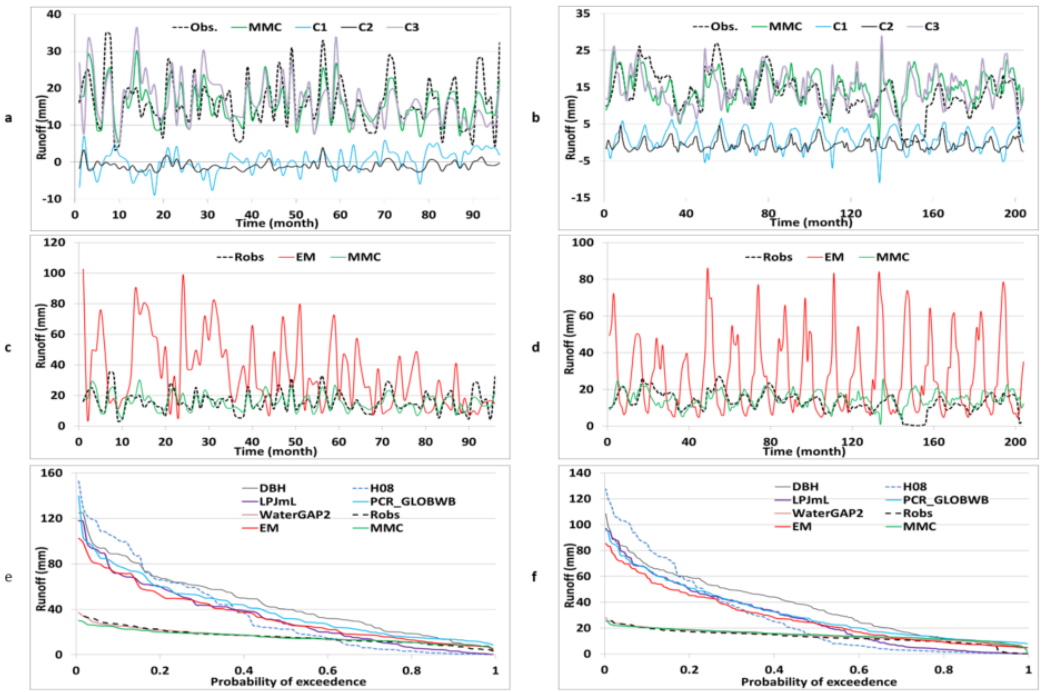
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

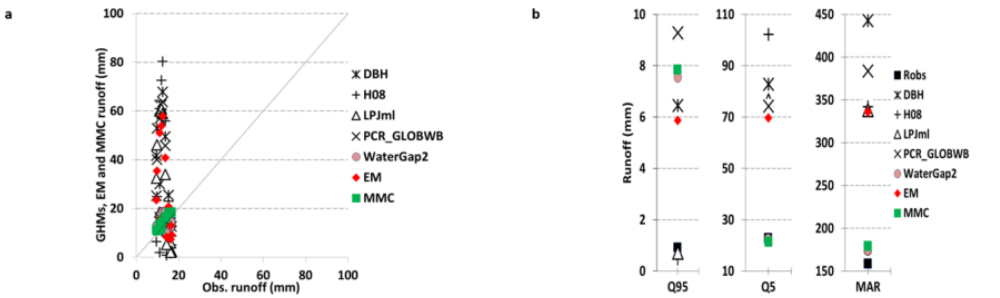
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	2.6	2.2	1.5	1.3	-1.7	1.2	-2.8	-503.5	-613.2	-109.7	-152.4
Calibration	2.9	2.2	1.6	1.9	-1.8	1.4	-2.5	-487.5	-705.5	-63.7	-105.9
Validation	4.8	3.5	1.9	2.2	-1.6	1.9	-1.9	-485.5	-41341.1	-29.2	-386.6



35. Paraguai catchment at Porto Murtinho (FB/DNOS) station.



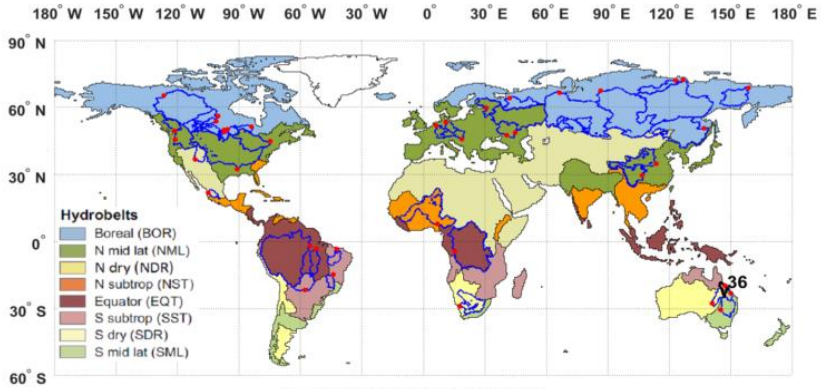
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



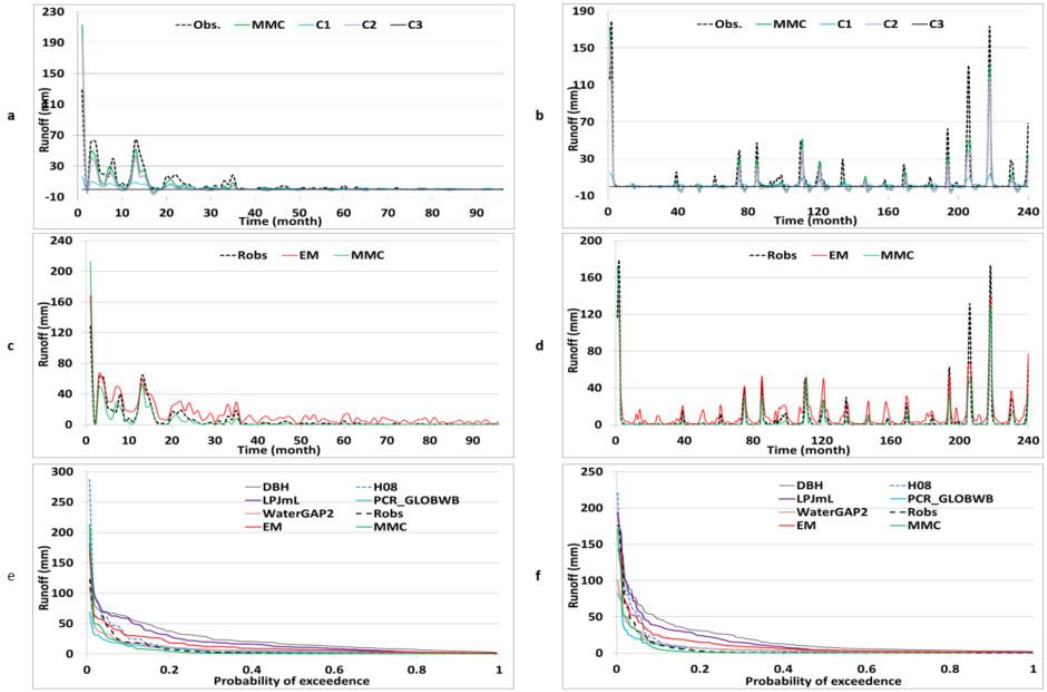
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

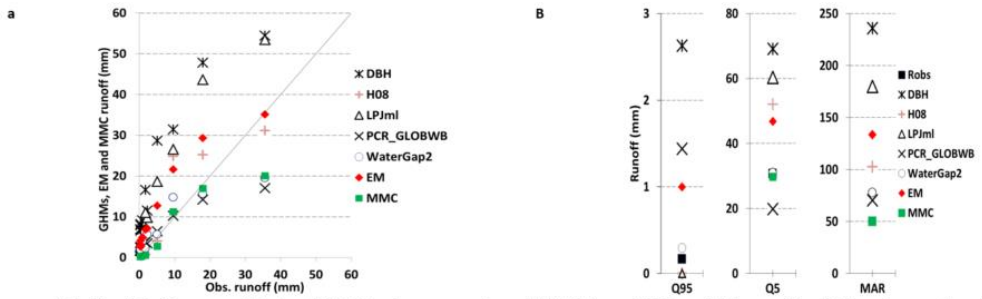
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	9.5	9.8	7.4	7.6	-1.7	5.5	-2.5	-905.0	-613.2	-88.0	-152.4
Calibration	12.1	13.1	8.8	9.3	-1.6	7.0	-2.2	-1016.5	-705.5	-59.3	-105.9
Validation	136.9	153.7	108.1	98.4	8.0	78.5	8.2	-7033.6	-41341.1	18.6	-386.6



36. Burdekin catchment at Clare station.



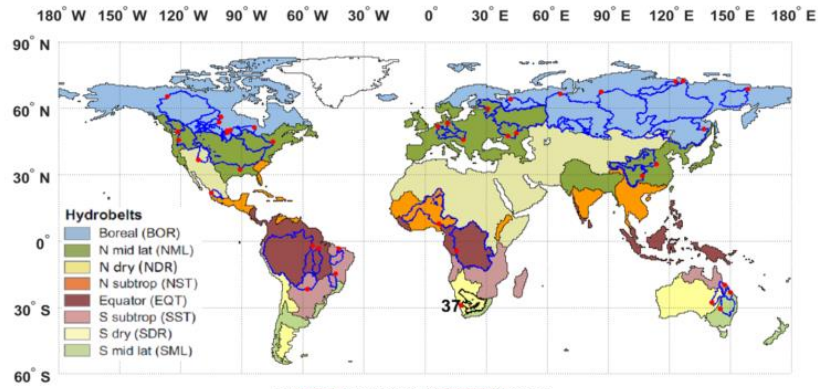
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



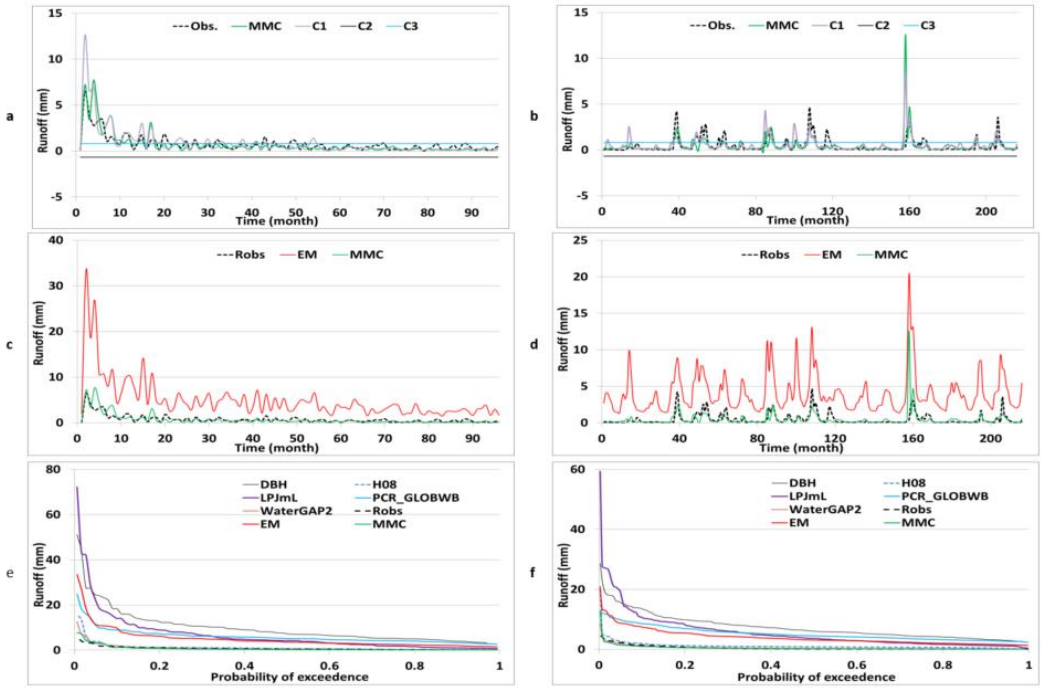
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

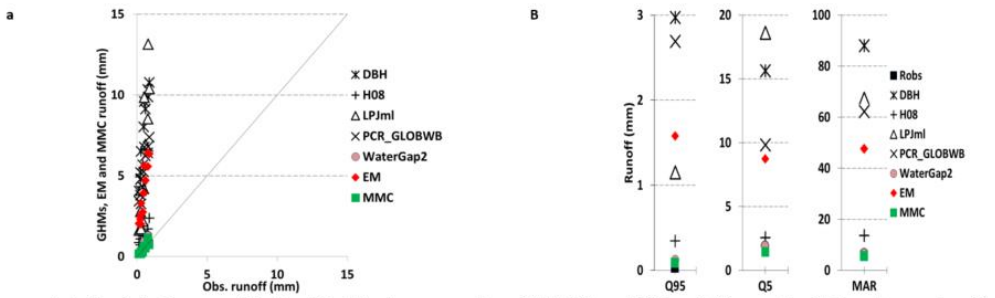
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	-1.3	-3.0	-2.2	-3.3	-2.7	-3.0	-4.5	-142.7	-613.2	-120.3	-152.4
Calibration	1.2	-1.1	-1.2	-2.6	-3.7	-2.7	-2.5	19.8	-705.5	120.4	-105.9
Validation	6.9	1.4	3.1	2.0	1.6	2.9	-1.4	-527.9	-41341.1	-479.2	-386.6



37. Oranje catchment at Vioolsdrif station.



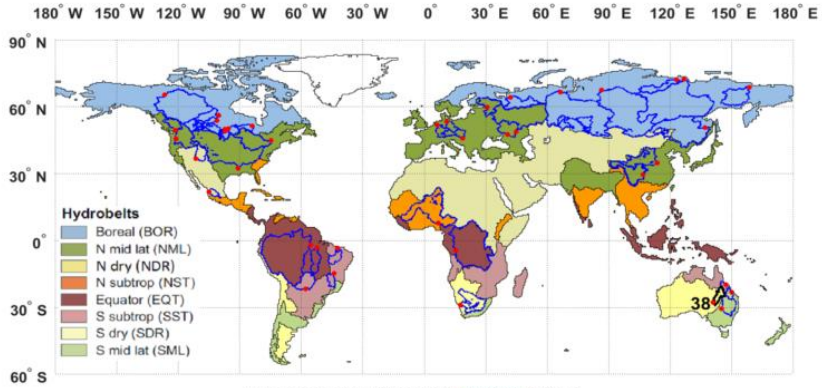
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



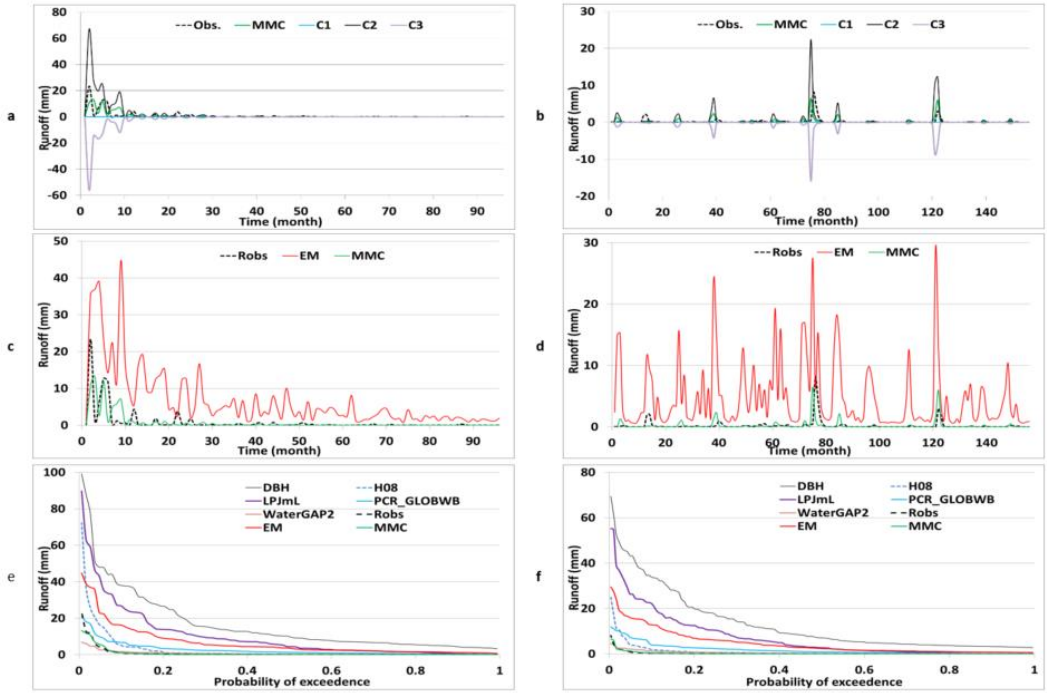
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

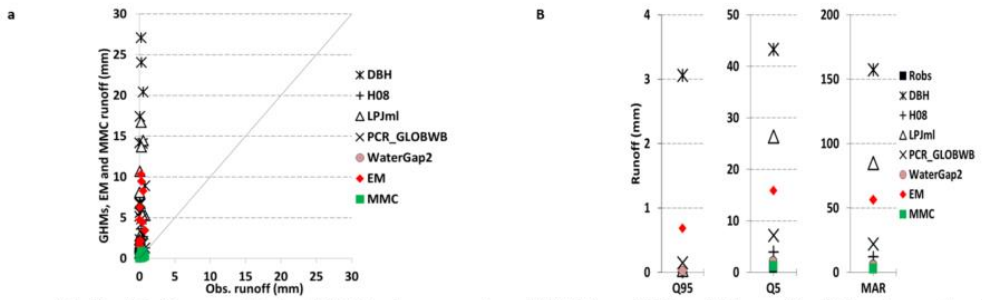
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	16.4	2.1	13.8	9.3	-2.0	6.2	-3.3	-1056.3	-613.2	-131.6	-152.4
Calibration	89.6	2.3	83.1	24.1	-1.0	23.1	-1.5	-2567.6	-705.5	-50.0	-105.9
Validation	83.1	7.1	81.1	46.4	2.3	31.2	2.0	-2911.3	-41341.1	-22.3	-386.6



38. Cooper Creek catchment at Callamurra station.



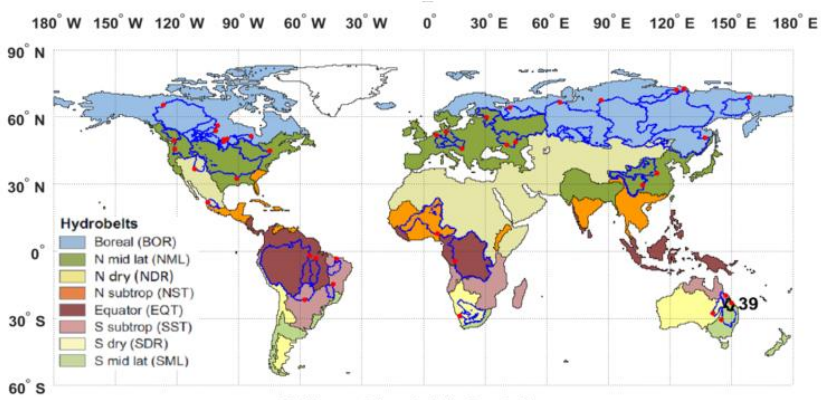
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



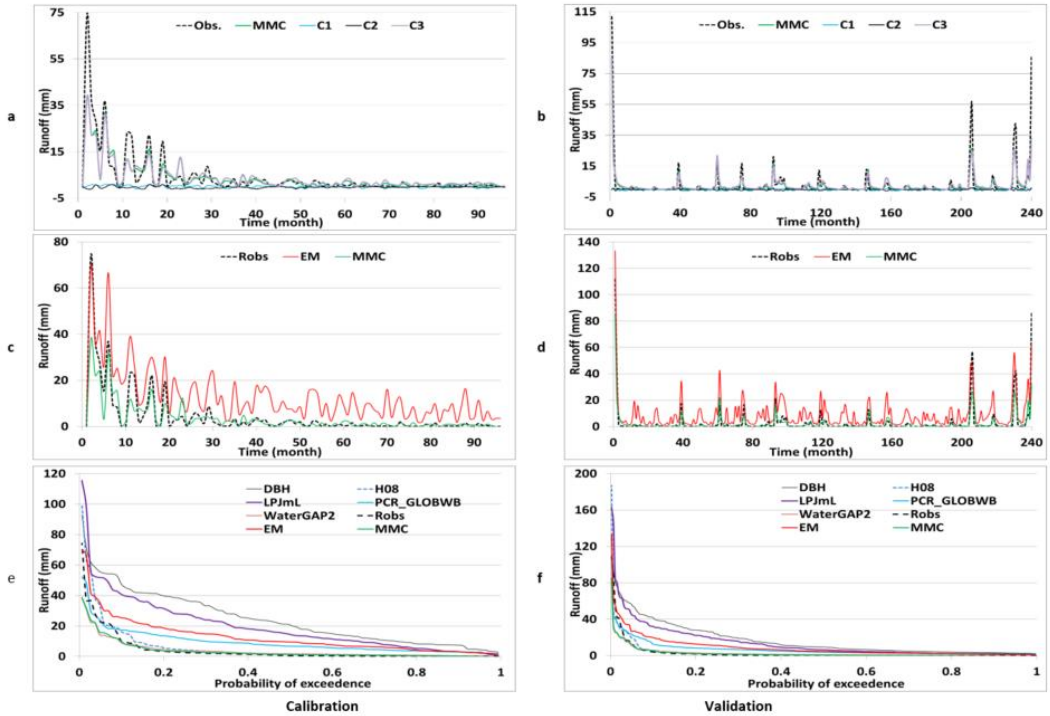
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

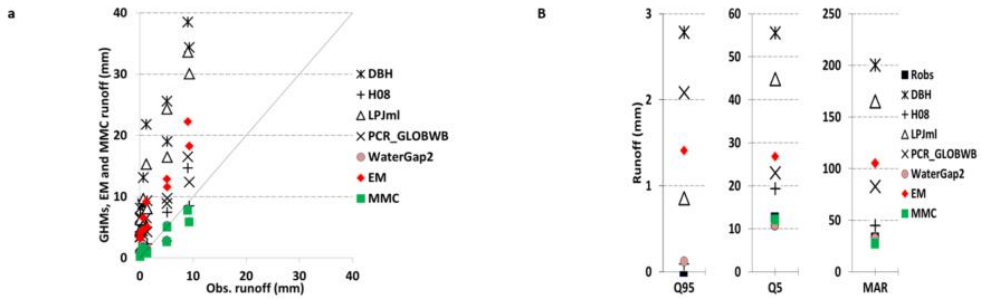
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	16.9	1.2	10.8	-1.4	-1.8	3.2	-6.4	-1063.8	-613.2	-463.9	-152.4
Calibration	63.6	2.4	30.6	6.6	-1.1	19.5	-2.2	-2276.7	-705.5	-109.8	-105.9
Validation	6992.9	148.9	2578.2	625.3	106.8	2088.9	20.1	-206888.9	-41341.1	-8673.7	-386.6



39. Fitzroy catchment at The Gap station.



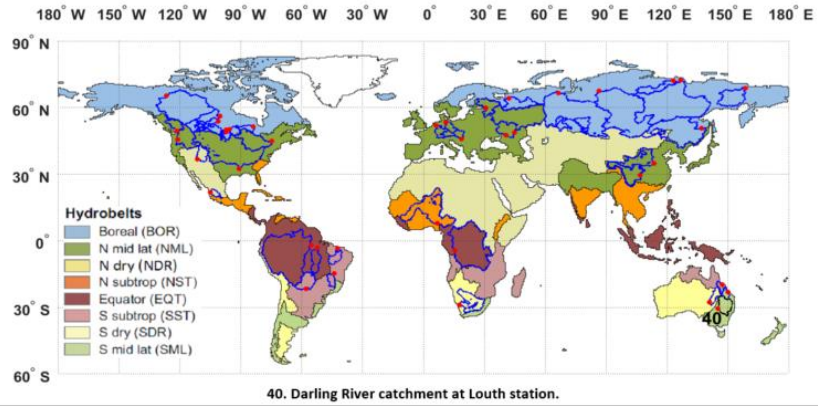
2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



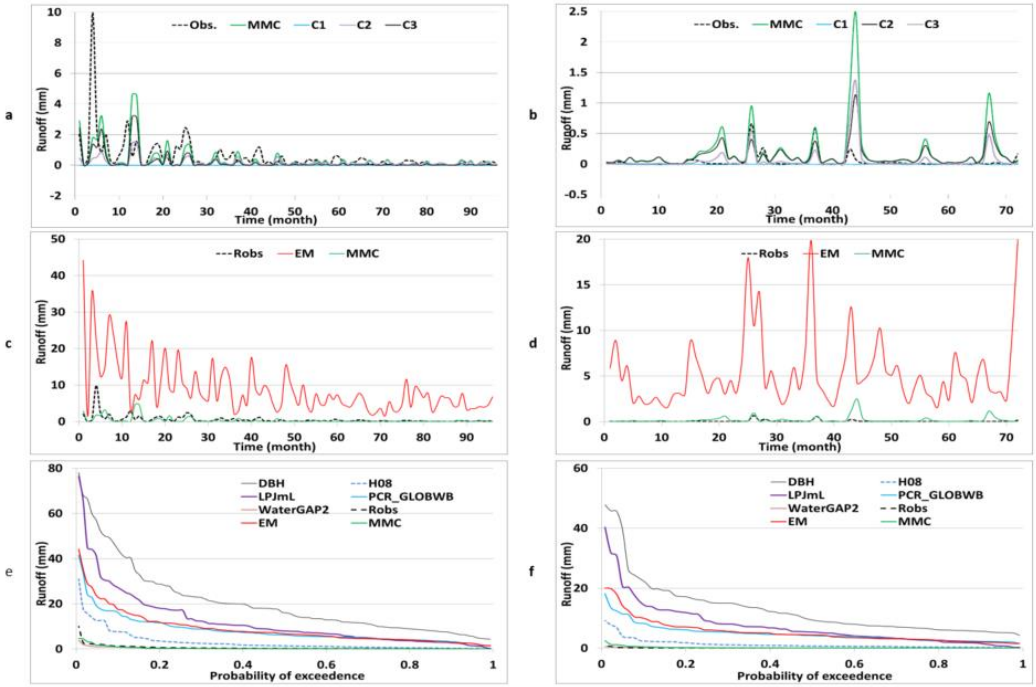
3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

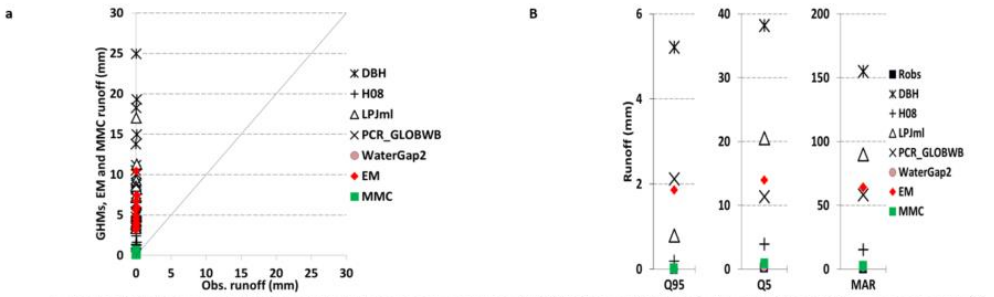
	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	2.9	-1.3	2.0	-2.1	-3.5	-1.2	-3.8	-256.2	-613.2	-32.6	-152.4
Calibration	6.3	-1.9	4.9	2.1	-2.8	2.6	-3.1	-663.2	-705.5	-25.6	-105.9
Validation	641.2	52.6	447.5	270.3	38.5	290.0	30.6	-25936.7	-41341.1	-783.5	-386.6



40. Darling River catchment at Louth station.



2. Partial solution spaces plot of the optimal MMC (a,b) for split data, time series of the EM, MMC and Robs (c,d) for split data, and flow duration curve (e,f), (Robs = obsereved runoff).



3. Scatter plots of mean monthly a) and Q5, Q95 and mean annual runoff (MAR) b) on validation set, (Obs. runoff and Robs = obsereved runoff).

4. IPE values for the GHMs (best GHM underlined), EM and optimal MMC and improvement in IPE achieved by MMC relative to the best GHM and the EM

	IPE							MMC_{PG} relative to EM		MMC_{PG} relative to best GHM	
	DBH	H08	LPJmL	PCR_ GLOBWB	Water-GAP2	EM	MMC	Catchment	All catchments (Average)	Catchment	All catchments (Average)
Training	141.5	8.0	63.5	20.6	-1.5	27.3	-2.0	-3032.4	-613.2	-49.2	-152.4
Calibration	200.6	7.0	92.3	35.2	-1.5	41.9	-1.6	-4456.9	-705.5	-10.8	-105.9
Validation	39448.4	1126.7	18262.3	11432.5	130.8	13919.3	117.3	-1380198.3	-41341.1	-1350.2	-386.6

References

Bondeau, A., Smith, P.C., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., Gerten, D., Lotze-Campen, H., MÜLler, C., Reichstein, M., Smith, B., 2007. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. Global Change Biology 13(3) 679-706.

Ferreira, C., 2001. Gene Expression Programming: A New AdaptiveAlgorithm for Solving Problems. 13(2) 87-129.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., Tanaka, K., 2008. An integrated model for the assessment of global water resources Part 1: Model description and input meteorological forcing. Hydrology and Earth System Sciences 12(4) 1007-1025.

Meybeck, M., Kummu, M., Dürr, H.H., 2013. Global hydrobelts and hydroregions: improved reporting scale for water-related issues? Hydrology and Earth System Sciences 17(3) 1093-1111.

Müller Schmied, H., Adam, L., Eisner, S., Fink, G., Flörke, M., Kim, H., Oki, T., Portmann, F.T., Reinecke, R., Riedel, C., Song, Q., Zhang, J., Döll, P., 2016. Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use. Hydrology and Earth System Sciences 20(7) 2877-2898.

Tang, Q.H., Oki, T., Kanae, S., Hu, H.P., 2007. The influence of precipitation variability and partial irrigation within grid cells on a hydrological simulation. Journal of Hydrometeorology 8(3) 499-512.

Wada, Y., Wisser, D., Bierkens, M.F.P., 2014. Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. Earth System Dynamics 5(1) 15-40.

Appendix C (Supplementary information for Paper III)

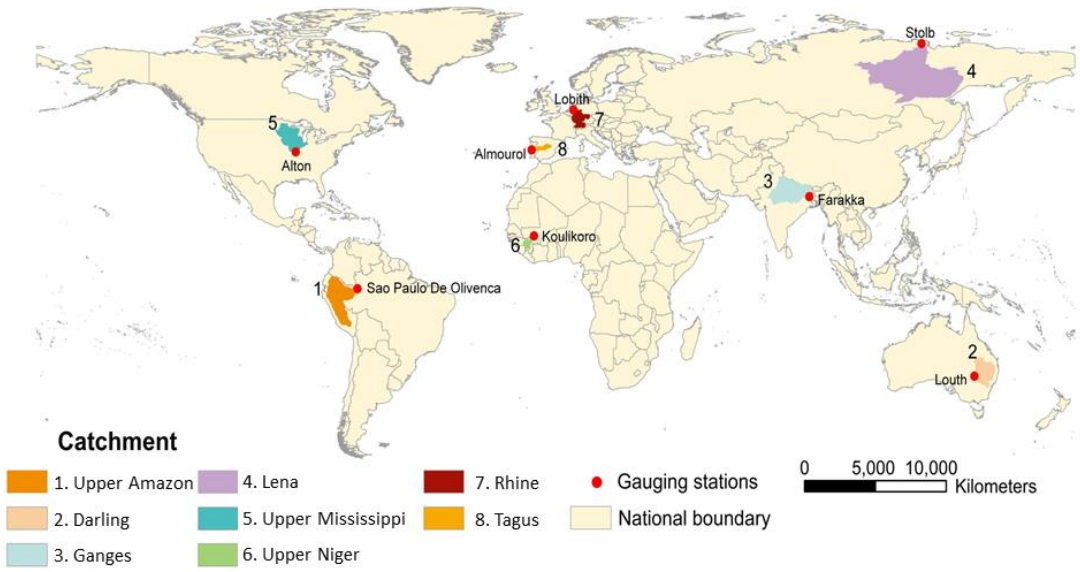
Electronic Supplementary Materials for "A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1°C, 2°C and 3°C."

Climatic Change

Simon N. Gosling[1*], Jamal Zaherpour[1], Nick Mount[1], Fred F. Hattermann[2], Rutger Dankers[3], Berit Arheimer[4], Lutz Breuer[5,6], Jie Ding[7,8], Ingjerd Haddeland[9], Rohini Kumar[10], Dipangkar Kundu[11], Junguo Liu[7,12], Ann van Griensven[13,14], Ted I.E. Veldkamp[15], Tobias Vetter[2], Xiaoyan Wang[16], Xinxin Zhang[12]

[1] School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom.
[2] Potsdam Institute for Climate Impact Research, Telegrafenberg A62, D-14473 Potsdam, Germany.
[3] Met Office, Exeter, United Kingdom.
[4] Swedish Meteorological and Hydrological Institute (SMHI), 60176 Norrköping, Sweden.
[5] Institute for Landscape Ecology and Resources Management (ILR), Research Centre for BioSystems, Land Use and Nutrition (IFZ), Justus Liebig University Giessen, Germany
[6] Centre for International Development and Environmental Research, Justus Liebig University, Germany.
[7] School of Environmental Science and Engineering, South University of Science and Technology of China, Shenzhen, 518055, China.
[8] Institute of Water Resources Management, Hydrology and Agricultural Hydraulic Engineering, Leibniz Hannover University, Hannover 30167, Germany.
[9] Norwegian Water Resources and Energy Directorate, Oslo, Norway.
[10] UFZ-Helmholtz Centre for Environmental Research, Leipzig, Germany.
[11] Department of Environmental Sciences, Faculty of Agriculture and Environment, The University of Sydney.
[12] School of Nature Conservation, Beijing Forestry University, Beijing, 10083, China.
[13] Department of Hydrology and Hydraulic Engineering, Vrije Universiteit Brussel, Belgium.
[14] UNESCO-IHE Institute for Water Education, the Netherlands.
[15] Vrije Universiteit, Institute for Environmental Studies (IVM), Amsterdam, the Netherlands.
[16] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Center for Global Change and Water Cycle, Hohai University, Nanjing 210098, China.
* All correspondence to:
Email: simon.gosling@nottingham.ac.uk
Telephone: +44(0)115 951 5437
Fax: +44 (0)115 951 5249

Online Resource 1: The eight study catchments (shaded) and their gauges (red filled circles).



Online Resource 2: Key model characteristics of the Glob-HMs and Cat-HMs employed in this study. Tmin = minimum temperature, Tmax = maximum temperature, P = precipitation, AH= air humidity, RAD = solar radiation, WS = wind speed, LAI = leaf area index, LW = longwave radiation, PET = potential evapotranspiration, SP = surface pressure.

Model	***Spatial disaggregation***	***Representation of soils***	***Representation of vegetation***	***Input climate parameters***	***Method: potential evapotranspiration***	***Method: snow melt***	***Method: runoff routing***
VIC-Cat-HM (Cat-HM)	Grid cells with sub-grid heterogeneity accounting method	Three soil layers, 19 parameters	Fixed monthly plant characteristics	Tmin, Tmax, P, AH, WS	Penman-Monteith	Two-layer energy Balance at the snow surface	Linearized St. Venant's equations
SWIM (Cat-HM)	Subbasins and hydrotopes	Up to 10 soil layers, 11 soil parameters	A simplified EPIC approach	Tmin, Tmean, Tmax, P, AH, RAD	Priestley-Taylor or Turc-Ivanov	An extended degree-day method	Muskingum method + reservoirs and irrigation
WaterGAP3 (Cat-HM)	Grid cells (5 arc min.) with elevation subgrid (1 arc min.)	One soil layer, 2 soil parameters	Temperature-dependent LAI, fixed rooting depth	Tmean, P, RAD, LW	Priestley-Taylor	Degree-day method	Linear reservoir, flow velocity based on Manning-Strickler
mHM (Cat-HM)	Grid cells with sub-grid heterogeneity accounting method	N soil layers (here two soil layers were used)	Fixed monthly plant characteristics	Tmax, Tmin, Tmean and P	Hargreaves and Samani method + aspect correction	Enhanced Degree-day method	Muskingum method
HBV (Cat-HM)	Subbasins, 10 elevation zones & land use classes	One soil layers, 2 soil parameters	Fixed monthly plant characteristics	Tmean and P	Blaney–Criddle	Degree-day method	A simple time lag method
HYMOD (Cat-HM)	Lumped model, a single basin structure	One soil layer, 7 soil parameters	NA	P, Tmean and PET	Hargreaves and Samani method	Degree-day method	Quick and slow flows tanks and linear reservoirs
SWAT (Cat-HM)	Subbasins and hydrologic response units (HRU)	Up to 10 soil layers, 11 soil parameters	A simplified EPIC approach	Tmin, Tmax, P, AH, RAD, WS	Penman-Monteith, Hargreaves, Priestely-Talvor	Degree-day method	Muskingum, Variable Storage + Reservoirs, ponds and wetlands method
HYPE (Cat-HM)	Subbasins and hydrological response units (HRU)	Three soil layers and up to 10 soil types with individual parameters	Fixed plant characteristics	Mandatory: Tmean & P, optional: Tmin, Tmax, AH, RAD	Priestley-Taylor, modified Hargreaves-Semani and other	Degree-day method or simplified energy approaches	Reservoir cascade from ground and surface discharge
DBH (Glob-HM)	Grid cells with sub-grid heterogeneity accounting method	Three soil layers and one underlying groundwater layer at the bottom	Prescribed spatial distribution of natural vegetation and agricultural land cover	P, Tmean, WS,AH, LW, RAD, SP	Energy balance	Energy balance	Kinematics wave method to compute flow routing in a river network

Online Resource 2 (Cont.): Key model characteristics of the Glob-HMs and Cat-HMs employed in this study. Tmin = minimum temperature, Tmax = maximum temperature, P = precipitation, AH= air humidity, RAD = solar radiation, WS = wind speed, LAI = leaf area index, LW = longwave radiation, PET = potential evapotranspiration, SP = surface pressure.

H08 (Glob-HM)	Globally covered 0.5x0.5 degree resolution	1-layer leaky bucket soil. Its runoff properties vary with climate zones.	Natural use: Globally uniform. No-specific land type is assigned, as known as Manabe's bucket.	Tmean, Rainfall, Snowfall, WS, AH, RAD, LW, surface pressure	Bulk formula	Energy balance	Saturation-excess, with a non-linear function of soil moisture.
LPJmL (Glob-HM)	Grid cells (30 arc min resolution) with land cover and land use classes	Five hydrologically active soil layers, coupled to carbon and thermal balance	Dynamic simulation of growth and productivity (with prescribed spatial distribution of crops and pasture); daily	P, Tmean, RAD	Priestley-Taylor (modified for transpiration)	Degree-day method with precipitation factor	Continuity equation derived from linear reservoir model
Mac-PDM.09 (Glob-HM)	0.5x0.5 degree grid across the land-surface of the globe	One soil layer, 7 soil texture types	Fixed plant charactertistcs, 13 types of natural land-cover	P, Tmean, WS, AH, LW, RAD	Penman-Monteith	Degree-day method	Two linear reservoirs for direct runoff
MATSIRO (Glob-HM)	Grid cells (0.5x0.5 degree)	19 layers with dynamical ground water schime. 13 soil type with 7 parameters.	Monthly LAI, and fixed vegetation characteristics for 12 types of natural land cover	Tmean, Rainfall, Snowfall, AH, LW, RAD, WS, SP	Penman-Monteith	Energy balance	Linear reservoir, fixed flow velocity at 0.5 m/s globally
MPI-HM (Glob-HM)	Grid cells (0.5 deg)	1 soil layer (bucket scheme), 6 parameters	Fixed climatology of vegetation fraction	Tmean, P, PET	Penman-Monteith	Degree-day method	Linear reservoir
PCR-GLOBWB (Glob-HM)	Grid cells (30 arc min.) with elevation subgrid (30 arc sec.)	Two soil layers and one underlying groundwater layer at the bottom	Prescribed vegetation, agriculture, and land use cover	Tmean, P, PET	Hamon	Degree-day method	Characteristic distance approach based on channel Characteristics.
VIC-Glob-HM (Glob-HM)	Grid cells with sub-grid heterogeneity accounting method	Three soil layers, 19 parameters	Fixed monthly plant characteristics	Tmin, Tmax, P, AH, WS, SP, RAD, LW	Penman-Monteith	Two-layer energy Balance at the snow surface	Linearized St. Venant's equations
WBM (Glob-HM)	Grid cell, resolution flexible	1 soil layer, bucket model, one parameter	Fixed vegetation, constant over time	P, Tmean	Hamon	Empirical Tmean and P based formula	Muskingum method

Online Resource 3: Central years of 31-year periods corresponding to the different levels of global-mean warming (compared to pre-industrial) in the individual GCMs under RCP8.5. Other RCPs are available (van Vuuren et al., 2011) but we used RCP8.5 in line with previous work (Schewe et al., 2014) because it enable investigation of global-mean warming levels up to 3°C for all five GCMs, whereas not all other RCPs reach this level of warming.

Warming level	HadGEM2-ES	IPSL-CM5A-LR	MIROC-ESM-CHEM	GFDL-ESM2M	NorESM1-M
1°C	2006	2008	2008	2015	2013
2°C	2031	2035	2032	2053	2047
3°C	2051	2055	2051	2082	2072

Online Resource 4: Further details on the calculation of the magnitude of annual maximum daily runoff and annual minimum 7-day runoff associated with the 2, 5, 10, 20, 25, and 50-year return periods.

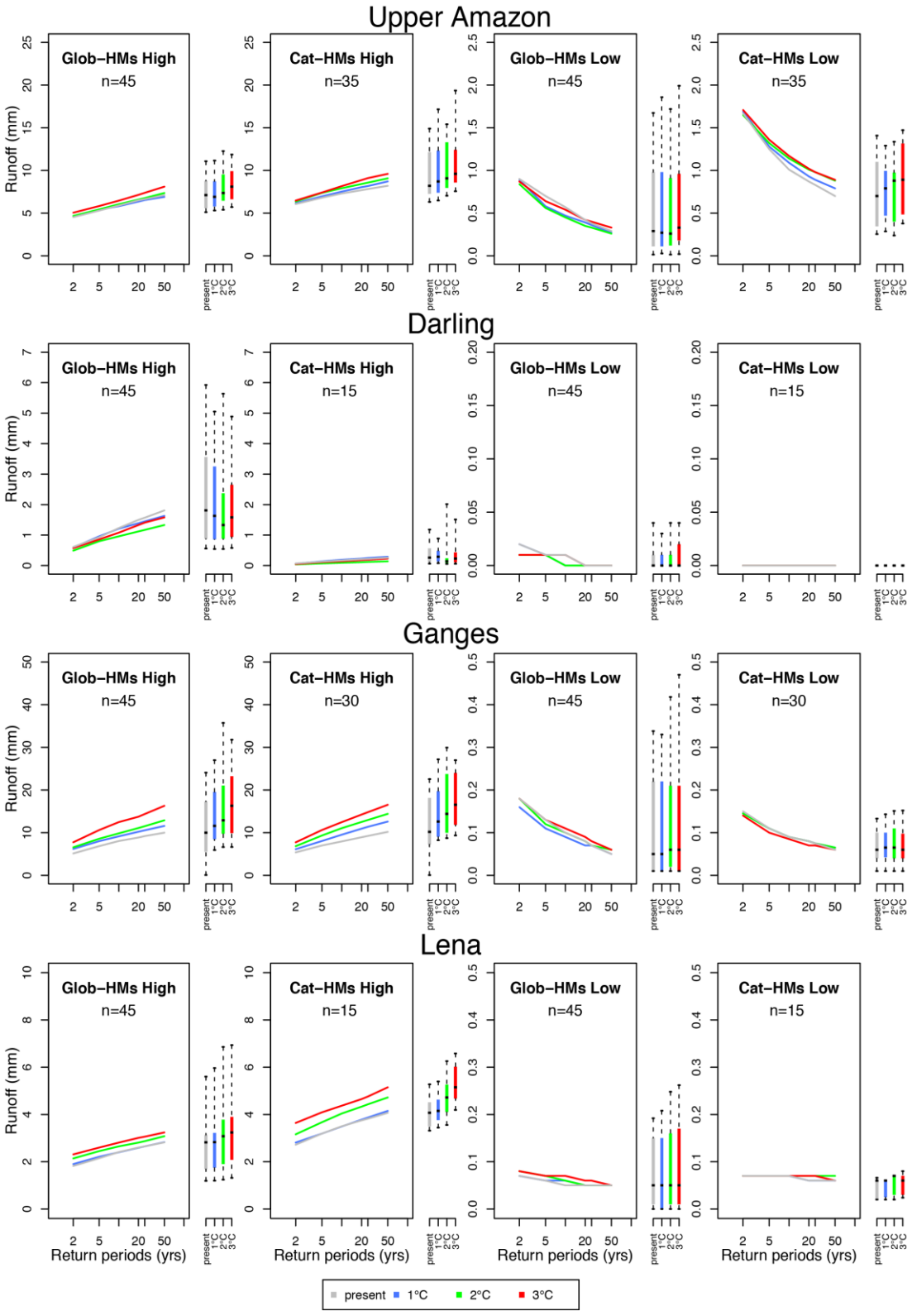
Maximum and minimum daily runoff values corresponding to different occurrence intervals (2, 5, 10, 20, 25, and 50 year return periods) were estimated in this study using extreme value analysis techniques.

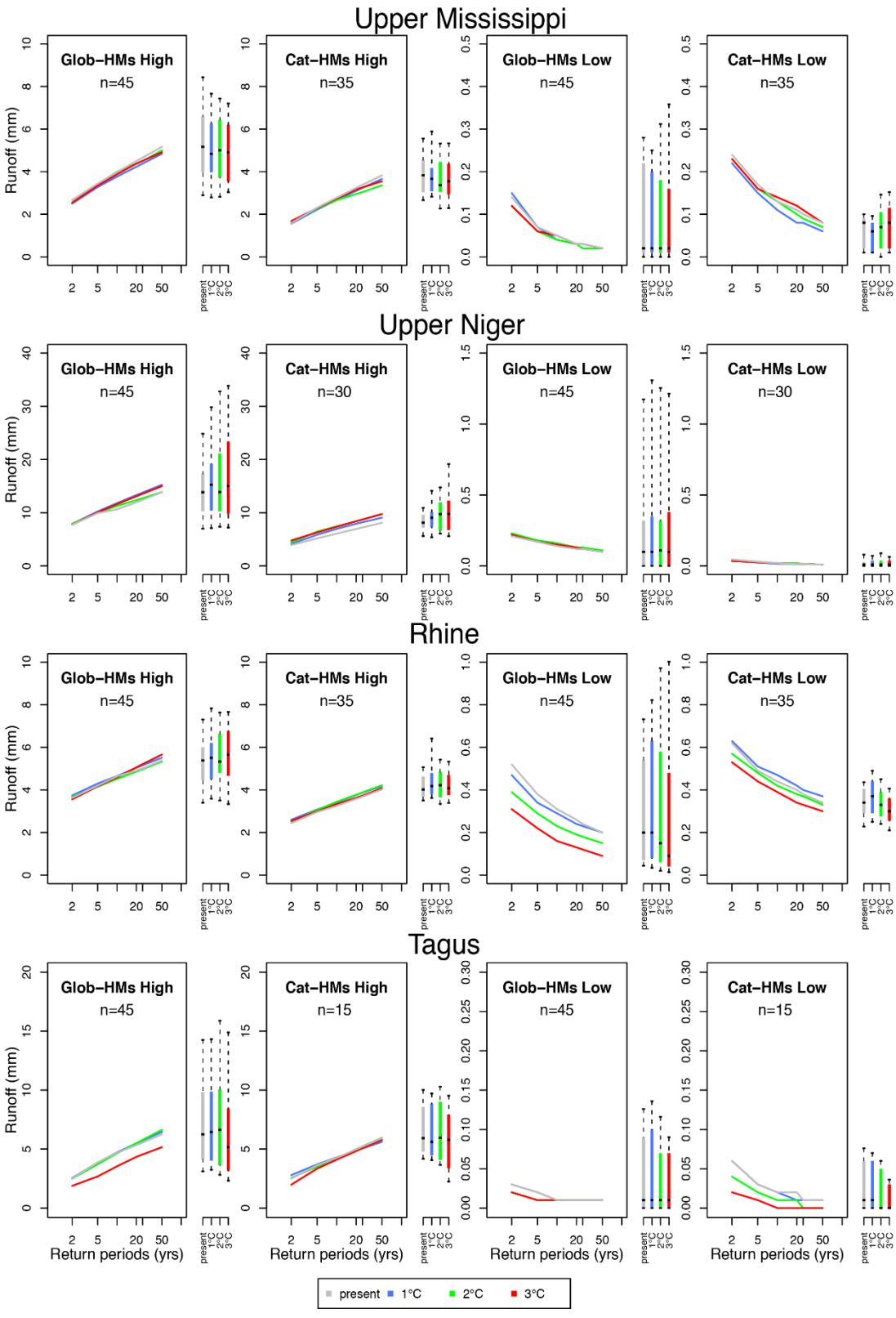
To cover maximum runoff values, a Gumbel distribution (Gumbel, 1941) was fit through the 31-year long time series of annual maximum daily runoff, for each GCM-Glob-HM/Cat-HM combination, for each climate change scenario, and for each catchment. Subsequently, we applied the estimated Gumbel parameters to calculate the maximum daily runoff values for return periods varying from 2 up to 50 years. We thereby accounted continuously for the probability of zero runoff (Stedinger et al., 1993). To ensure that the runoff time-series were statistically independent, a pre-requisite for such an extreme value analysis (Katz et al., 2002; Leadbetter et al., 1983), we used hydrological years rather than calendar years to find the maximum daily runoff values per year (Dettinger and Diaz, 2000; Ward et al., 2010). For each catchment and GCM-Glob-HM/Cat-HM combination, we assessed for the historical time-period the timing of long-term mean maximum daily runoff values. If this maximum daily long-term mean runoff value of a GCM-Glob-HM/Cat-HM combination in a specific catchment, fell in the months September, October, or November, we used the period July-June to characterise the hydrological year. In all other cases we used the standard hydrological year convention for peak flows that runs from October-September (Ward et al., 2014).

We covered extreme minimum runoff with the use of annual minimum 7-day flows (WMO, 2009), using a moving average of 7-days runoff. Subsequently, we fitted a Weibull distribution (Weibull, 1951) – often used in low flow studies due to its lower bound and flexibility – through each of the 31-year long time-series of annual minimum 7-day flow. We used the estimated Weibull parameters to return minimum 7-day runoff values for the different return periods. In doing so, we accounted for the occurrence of zero flows by the use of a conditional probability model (Stedinger et al., 1993). The use of calendar years might not be always the proper choice to select the annual minimum 7-day flows for input of the extreme value analysis (Katz et al., 2002; Leadbetter et al., 1983; WMO, 2009). We

therefore assessed for each GCM-Glob-HM/Cat-HM combination and for each catchment whether the use of calendar years would possibly split the long-term mean low-flow season in the historical time-period (WMO, 2009). If the timing of the long-term mean low flow period fell in the months December, January, or February, we applied a hydrological year that runs from August-July (WMO, 2009). In all other cases, we used the standard hydrological year distinction for low flows, which runs from January-December (WMO, 2009).

Online Resource 5: The magnitude of maximum daily runoff ("High") and minimum 7-day moving average of runoff ("Low") associated with 2, 5, 10, 20, 25, and 50-year return periods for the Glob-HM ensemble and Cat-HM ensemble. The lines show the median for each ensemble and the box-whiskers show the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of changes for the 50-year return period only.




Upper Mississippi
Glob–HMs High
n=45
Cat–HMs High
n=35
Glob–HMs Low
n=45
Cat–HMs Low
n=35
Upper Niger
Glob–HMs High
n=45
Cat–HMs High
n=30
Glob–HMs Low
n=45
Cat–HMs Low
n=30
Rhine
Glob–HMs High
n=45
Cat–HMs High
n=35
Glob–HMs Low
n=45
Cat–HMs Low
n=35
Tagus
Glob–HMs High
n=45
Cat–HMs High
n=15
Glob–HMs Low
n=45
Cat–HMs Low
n=15
Runoff (mm)
Return periods (yrs)
present
1°C
2°C
3°C

References for the Electronic Supplementary Materials

Dettinger MD, Diaz HF (2000) Global Characteristics of Stream Flow Seasonality and Variability. Journal of Hydrometeorology 1:289-310.

Gumbel EJ (1941) The Return Period of Flood Flows. 12:163-190.

Katz RW, Parlange MB, Naveau P (2002) Statistics of extremes in hydrology. Advances in Water Resources 25:1287-1304.

Leadbetter MR, Lindren G, Rootzen H (1983) Extremes and related properties of random sequences and processes. Springer-Verlag, New York, USA.

Schewe J, Heinke J, Gerten D, Haddeland I, Arnell NW, Clark DB, Dankers R, Eisner S, Fekete BM, Colón-González FJ, Gosling SN, Kim H, Liu X, Masaki Y, Portmann FT, Satoh Y, Stacke T, Tang Q, Wada Y, Wisser D, Albrecht T, Frieler K, Piontek F, Warszawski L, Kabat P (2014) Multimodel assessment of water scarcity under climate change. Proceedings of the National Academy of Sciences 111:3245-3250.

Stedinger JR, Vogel RM, Foufoula-Georgiou E (1993) Frequency analysis of extreme events. in Maidment DR (ed.) Handbook of Hydrology. McGraw-Hill, New York, pp. 18.11- 18.66.

van Vuuren D, Edmonds J, Kainuma M, Riahi K, Thomson A, Hibbard K, Hurtt G, Kram T, Krey V, Lamarque J-F, Masui T, Meinshausen M, Nakicenovic N, Smith S, Rose S (2011) The representative concentration pathways: an overview. Climatic Change 109:5-31.

Ward PJ, Beets W, Bouwer LM, Aerts JCJH, Renssen H (2010) Sensitivity of river discharge to ENSO. Geophysical Research Letters 37:L12402.

Ward PJ, Eisner S, Flörke M, Dettinger MD, Kummu M (2014) Annual flood sensitivities to El Niño–Southern Oscillation at the global scale. Hydrol. Earth Syst. Sci. 18:47-66.

Weibull W (1951) Wide applicability. Journal of Applied Mechanics 13:293-297.

WMO (2009) Manual on low-flow estimation and prediction (http://www.wmo.int/pages/prog/hwrp/publications/low-flow_estimation_prediction/WMO%201029%20en.pdf).

Appendix D (Supplementary information for Paper IV)

Multi-model combination with a super-ensemble of catchment-scale and global-scale hydrological models

Jamal Zaherpour[1*], Simon N Gosling[1], Nick Mount[1], Fred Hattermann[2]

[1]School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom.
[2]Potsdam Institute for Climate Impact Research, Telegrafenberg, 14473 Potsdam, Germany

*Corresponding author:

Tel.: +44 115 951 5428, Fax: +44 (0)115 951 5249
E-mail addresses: zaherpour@gmail.com , lgxjz1@nottingham.ac.uk
Postal address: School of Geography, University of Nottingham, Nottingham NG7 2RD, United Kingdom

Table S1: Key model characteristics of the RHMs and GHMs employed in this study. Tmin = minimum temperature, Tmax = maximum temperature, P = precipitation, AH= air humidity, RAD = solar radiation, WS = wind speed, LAI = leaf area index, LW = longwave radiation, PET = potential evapotranspiration, SP = surface pressure.

Model	***Spatial disaggregation***	***Representation of soils***	***Representation of vegetation***	***Input climate parameters***	***Method: potential evapotranspiration***	***Method: snow melt***	***Method: runoff routing***
VIC (RHM)	Grid cells with sub-grid heterogeneity accounting method	Three soil layers, 19 parameters	Fixed monthly plant characteristics	Tmin, Tmax, P, AH, WS	Penman-Monteith	Two-layer energy Balance at the snow surface	Linearized St. Venant's equations
SWIM (RHM)	Subbasins and hydrotopes	Up to 10 soil layers, 11 soil parameters	A simplified EPIC approach	Tmin, Tmean, Tmax, P, AH, RAD	Priestley-Taylor or Turc-Ivanov	An extended degree-day method	Muskingum method + reservoirs and irrigation
WaterGAP3 (RHM)	Grid cells (5 arc min.) with elevation subgrid (1 arc min.)	One soil layer, 2 soil parameters	Temperature-dependent LAI, fixed rooting depth	Tmean, P, RAD, LW	Priestley-Taylor	Degree-day method	Linear reservoir, flow velocity based on Manning-Strickler
mHM (RHM)	Grid cells with sub-grid heterogeneity accounting method	N soil layers (here two soil layers were used)	Fixed monthly plant characteristics	Tmax, Tmin, Tmean and P	Hargreaves and Samani method + aspect correction	Enhanced Degree-day method	Muskingum method
HBV (RHM)	Subbasins, 10 elevation zones & land use classes	One soil layers, 2 soil parameters	Fixed monthly plant characteristics	Tmean and P	Blaney–Criddle	Degree-day method	A simple time lag method
HYMOD (RHM)	Lumped model, a single basin structure	One soil layer, 7 soil parameters	NA	P, Tmean and PET	Hargreaves and Samani method	Degree-day method	Quick and slow flows tanks and linear reservoirs
SWAT (RHM)	Subbasins and hydrologic response units (HRU)	Up to 10 soil layers, 11 soil parameters	A simplified EPIC approach	Tmin, Tmax, P, AH, RAD, WS	Penman-Monteith, Hargreaves, Priestely-Talyor	Degree-day method	Muskingum, Variable Storage + Reservoirs, ponds and wetlands method
HYPE (RHM)	Subbasins and hydrological response units (HRU)	Three soil layers and up to 10 soil types with individual parameters	Fixed plant characteristics	Mandatory: Tmean & P, optional: Tmin, Tmax, AH, RAD	Priestley-Taylor, modified Hargreaves-Semani and other	Degree-day method or simplified energy approaches	Reservoir cascade from ground and surface discharge
ECOMAG	Subbasins, soil and land-use classes within them	Three soil layers, 5 parameters	Fixed plant characteristics	3 parameters: Tmean, P, AH	Dalton formulae	Degree-day method	Kinematic wave equations
CLM (GHM)	Grid cells with sub-grid heterogeneity accounting	15 soil layers (soil moisture is simulated for 10 layers)	Fixed monthly plant characteristics	6 parameters: Tmean, P, AH, RAD, WS, SAP	Modified Penman-Monteith	Degree-day method	Linear reservoir, constant flow velocity

Table S1 (Cont.): Key model characteristics of the RHMs and GHMs employed in this study. Tmin = minimum temperature, Tmax = maximum temperature, P = precipitation, AH= air humidity, RAD = solar radiation, WS = wind speed, LAI = leaf area index, LW = longwave radiation, PET = potential evapotranspiration, SP = surface pressure.

DBH (GHM)	Grid cells with sub-grid heterogeneity accounting method	Three soil layers and one underlying groundwater layer at the bottom	Prescribed spatial distribution of natural vegetation and agricultural land cover	P, Tmean, WS,AH, LW, RAD, SP	Energy balance	Energy balance	Kinematics wave method to compute flow routing in a river network
H08 (GHM)	Globally covered 0.5x0.5 degree resolution	1-layer leaky bucket soil. Its runoff properties vary with climate zones.	Natural use: Globally uniform. No-specific land type is assigned, as known as Manabe's bucket.	Tmean, Rainfall, Snowfall, WS, AH, RAD, LW, surface pressure	Bulk formula	Energy balance	Saturation-excess, with a non-linear function of soil moisture.
LPJmL (GHM)	Grid cells (30 arc min resolution) with land cover and land use classes	Five hydrologically active soil layers, coupled to carbon and thermal balance	Dynamic simulation of growth and productivity (with prescribed spatial distribution of crops and pasture); daily	P, Tmean, RAD	Priestley-Taylor (modified for transpiration)	Degree-day method with precipitation factor	Continuity equation derived from linear reservoir model
MATSIRO (GHM)	Grid cells (0.5x0.5 degree)	19 layers with dynamical ground water schime. 13 soil type with 7 parameters.	Monthly LAI, and fixed vegetation characteristics for 12 types of natural land cover	Tmean, Rainfall, Snowfall, AH, LW, RAD, WS, SP	Penman-Monteith	Energy balance	Linear reservoir, fixed flow velocity at 0.5 m/s globally
MPI-HM (GHM)	Grid cells (0.5 deg)	1 soil layer (bucket scheme), 6 parameters	Fixed climatology of vegetation fraction	Tmean, P, PET	Penman-Monteith	Degree-day method	Linear reservoir
PCR-GLOBWB (GHM)	Grid cells (30 arc min.) with elevation subgrid (30 arc sec.)	Two soil layers and one underlying groundwater layer at the bottom	Prescribed vegetation, agriculture, and land use cover	Tmean, P, PET	Hamon	Degree-day method	Characteristic distance approach based on channel Characteristics.
WaterGAP2 (GHM)	Grid cells (0.5 degree) with elevation subgrid (1 arc-min)	One soil layer, 2 soil parameters	Temperature-dependent LAI, fixed rooting depth	4 parameters: Tmean, P, RAD (short- & longwave)	Priestley-Taylor	Degree-day method	Linear reservoir, flow velocity based on Manning-Strickler

Table S2. The gene expression programming function set, constants and gene configuration. For detail explanations about control parameters see (Ferreira, 2001).

Settings	**Control parameter**	**Value/Symbol**	
		3-component settings	**2-component settings**
General settings	Number of genes	25	20
	Number of chromosomes (equation components)	3	2
	Gene head size	10	8
	Linking function	addition	addition
Functions, Fitness and model generation	Maximum generations	100,000	The same
	Fitness function	Customised IPE	
	Max fitness	1000	
	Functions	+, -, *, / (2) †; Sqrt, Exp, x^2, x^3, Natural Log, Sine, Cosine (1)†;	
Genetic operators	Mutation rate	0.044	The same
	Inversion	0.1	
	Insertion sequence (IS) transposition	0.1	
	Root insertion sequence (RIS) transposition	0.1	
	One-point recombination rate	0.3	
	Two-point recombination rate	0.3	
	Gene recombination	0.1	
	Gene transposition	0.1	
Numerical constants	Constants per gene	2	The same
	Lower/upper bound	-10/+10	
	Random numerical constants (RNC) mutation	0.01	
	DC mutation	0.044	
	DC inversion	0.1	
	DC IS transposition	0.1	

† *(1) and (2) w*eights of the operators and functions

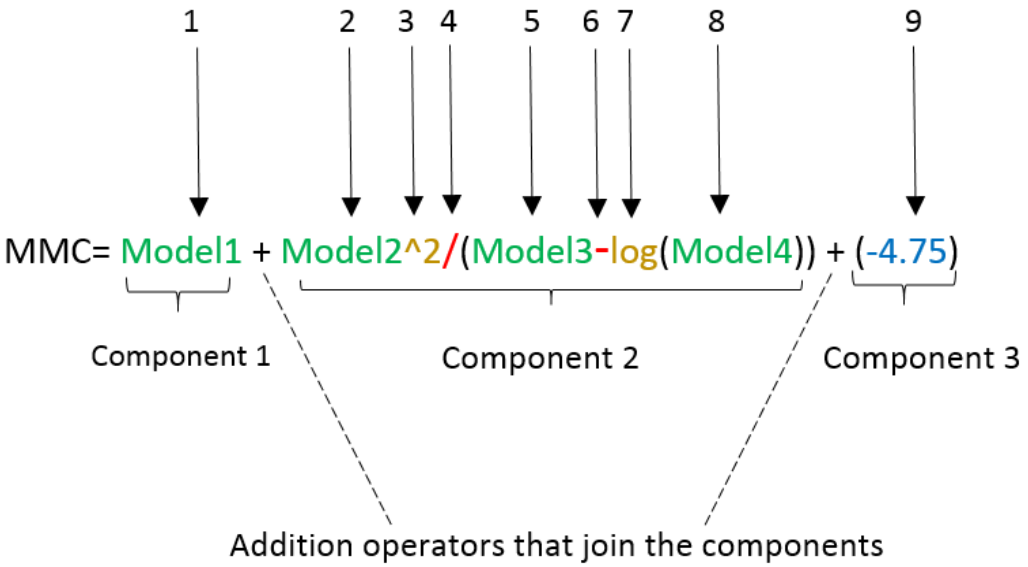


Figure S1. A hypothetical MMC equation with complexity of 9 formed by three components (C1+C2+C3). Elements 1 to 9 are model inputs, arithmetic operators, functions or numerical constants counted towards the calculation of the overall MMC's complexity.

Table S3. Single metrics applied in this study

Metric Name	Formula	Range	Ideal value
Root mean squared error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{n}}$	[0-∞)	0
Mean absolute relative error (MARE)	$\mathrm{MARE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Q_i - \hat{Q}_i\rvert}{Q_i}$	[0-∞)	0
Coefficient of determination (RSqr)	$\mathrm{RSqr} = \left[\frac{\sum_{i=1}^{n}(Q_i - \overline{Q})\left(\hat{Q} - \tilde{Q}\right)}{\sqrt{\sum_{i=1}^{n}(Q_i - \overline{Q})^2 \sum_{i=1}^{n}\left(\hat{Q}_i - \tilde{Q}\right)^2}}\right]^2$	[0-1]	1
Nash-Sutcliffe (NSE) or Coefficient of efficiency (CE)	$\mathrm{CE} = 1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \overline{Q})^2}$	(-∞, 1]	1
Percentage bias (PBIAS)	$PBIAS = \frac{\sum_{i=1}^{n}(Q - \hat{Q}) * 100}{\sum_{i=1}^{n} Q}$	(-∞,∞)	0*

*Result of zero does not necessarily indicate low error due to cancellation.

In the above equations, Q is the observed value, $\hat{Q}$ denotes modelled value (where *i* = *1* to *n* data points), $\overline{Q}$ is the mean of the observed data, and $\tilde{Q}$ is the mean of the modelled values.

Table S4. MMCs and equations ranked by MMC performance gain (MMC_{PG}) in the out-of-sample validation data set for the 3- and 2-component MMC for the GHM ensemble. MMC_{PG} is measured against *either* the best performing GHM *or* the EM, whichever of the two performs better.

		3-component					2-component				
River	**Best performing GHM or EM and IPE score**	**MMC IPE score**	**MMC_{PG} (%)**	**Rank**	**MMC equation separated into its components. MMC = C1 + C2 + C3. Components (C1…C3) are ordered according to their explanatory power as assessed by IPE. The IPE value for each component is provided in square brackets.**	**Eqn. size[1]**	**MMC IPE score**	**MMC_{PG} (%)**	**Rank**	**MMC equation separated into its components. MMC = C1 + C2. Components (C1…C3) are ordered according to their explanatory power as assessed by IPE. The IPE value for each component is provided in square brackets.**	**Eqn. size[1]**
Upper Amazon	WaterGAP2 IPE = -1.43	-1.78	35.4	5	C1 [-1.44]: WaterGAP2 C2 [11.56]: + sin(-237.393+75.343*MATSIRO-sin DBH) C3 [12.93]: + (WaterGAP2-CLM)/ -3.467	15	-1.44	0.62	5	C1 [-1.43]: WaterGAP2 C2 [12.14]: + -0.16787	2
Lena	MATSIRO IPE = -1.67	-2.14	46.9	2	C1 [-1.66]: sqrt(LPJmL*MATSIRO) C2 [-1.45]: + MATSIRO-sqrt(sqrt DBH) C3 [1.60]: + sin(PCRGLOBWB)-log(LPJmL)	14	-1.59†	-7.58	7	C1 [-1.67]: MATSIRO C2 [1.12]: + log(log(WaterGAP2)^2)^2	6
Upper Mississippi	WaterGAP2 IPE = -1.40	-1.42	2.4	6	C1 [-1.39]: sin(DBH)+WaterGAP2 C2 [4.32]: + log(sqrt(LPJmL)) C3 [5.41]: + (6.180-MATSIRO-MPI)/MATSIRO	14	-1.48	8.10	4	C1 [-1.39]: WaterGAP2 C2 [4.86]: + log(MATSIRO)-3.619	5
Upper Niger	WaterGAP2 IPE = -1.30	-1.66	35.7	4	C1 [-1.30]: WaterGAP2 C2 [1.79]: + (MATSIRO/(PCRGLOBWB-log(LPJmL)))^3 C3 [2.88]: + (-4.772)	9	-1.62	31.80	2	C1 [-1.30]: WaterGAP2 C2 [2.73]: + cos(sqrt(MPI))- 3.539	6
Rhine	WaterGAP2 IPE = -2.51	-2.53	2.1	7	C1 [-2.51]: WaterGAP2 C2 [5.86]: + sin(sqrt(log CLM)+sin MPI) C3 [6.07]: + LPJmL/(MPI+MATSIRO-LPJmL^3+(WaterGAP2+MATSIRO)^2)	21	-2.50	-0.57	6	C1 [-2.49]: WaterGAP2-((CLM-MPI)/MATSIRO) C2 [6.10]: + sin(PCRGLOBWB*8.551*exp(sin(8.551)/MPI))	17
Tagus	MATSIRO IPE = -1.26	-1.68	42.3	3	C1 [-1.39]: 0.834*MATSIRO C2 [1.28]: + sin(MATSIRO)^3 C3 [1.30]: + sin(sqrt(LPJmL))	9	1.47†	-472.84	8	C1 [1.17]: exp(CLM/DBH) C2 [1.21]: + WaterGAP2	5
Yangtze	EM IPE = -1.79	-3.27	148.5	1	C1 [-1.75]: LPJmL+(1/MPI) C2 [3.08]: + log(MATSIRO) C3 [3.29]: + CLM/LPJmL	10	-2.48	69.79	1	C1 [-1.73]: LPJmL C2 [2.87]: + sqrt(CLM)	3
Yellow	EM IPE =-1.18	-1.15	-3.4	8	C1 [1.50]: MATSIRO C2 [1.50]: + MATSIRO C3 [2.82]: + Log (H08)	4	-1.26	8.52	3	C1 [-1.15]: MATSIRO+log(CLM) C2 [1.50]: + MATSIRO	5

1-As defined in Section 2.4, equation size is calculated according to the number of model inputs, constants, operators and functions in an equation.

Table S5. MMCs and equations ranked by MMC performance gain (MMC_{PG}) in the out-of-sample validation data set for the 3- and 2-component MMC for the CHM ensemble. MMC_{PG} is measured against *either* the best performing GHM *or* the EM, whichever of the two performs better.

		3-component					2-component				
River	**Best performing GHM or EM and IPE score**	**MMC IPE score**	**MMC_{PG} (%)**	**Rank**	**MMC equation separated into its components. MMC = C1 + C2 + C3. Components (C1…C3) are ordered according to their explanatory power as assessed by IPE. The IPE value for each component is provided in square brackets.**	**Eqn. size[1]**	**MMC IPE score**	**MMC_{PG} (%)**	**Rank**	**MMC equation separated into its components. MMC = C1 + C2. Components (C1…C3) are ordered according to their explanatory power as assessed by IPE. The IPE value for each component is provided in square brackets.**	**Eqn. size[1]**
Upper Amazon	SWIM IPE = -1.99	1.98	0.8	2	C1 [-1.98]: cos(HBV/((VIC- 6.766)/HYMOD))+SWIM C2 [11.94]: + 1.0 C3 [12.33]: + (-6.962-HBV)/mHM	16	-2.00	1.07	4	C1 [-1.99]: SWIM C2 [12.11]: + sin(log(sin(0.119/WATERGAP3^3)^3))	9
Lena	SWIM IPE= -3.40	2.17	123.2	8	C1 [-2.56]: HYPE-cos(cos(HYPE)) C2 [1.20]: + 1.813 C3 [1.45]: + (-7.250)/(sin SWIM*(WATERGAP3-381.141))	14	-4.04	64.16	2	C1 [-2.24]: HYPE+(cos(exp(sin(ECOMAG)-WATERGAP3)^2)^2) C2 [1.87]: + 1/sin(-6.326/WATERGAP3)	16
Upper Mississippi	EM IPE = -2.47	1.73	74.0	6	C1 [-1.70]: mHM+sqrt(WATERGAP3) C2 [4.78]: + 0.0 C3 [4.81]: + (-1.593)/(sin(2*VIC-(mHM-SWAT))+(1.593+WATERGAP3-log(VIC)))	22	-2.48	0.49	5	C1 [-2.18]: mHM C2 [4.38]: + 0.0553*VIC-(8.385/(3.851+HYMOD))	10
Upper Niger	HYNOD IPE = -2.22	1.58	63.6	5	C1 [-1.56]: VIC+sin(SWIM) C2 [2.30]: + cos(VIC+ 5.784) C3 [2.32]: + cos(log(WATERGAP3))	11	-1.51	-70.91	7	C1 [-1.91]: sqrt(VIC*mHM) C2 [2.31]: + sin(WATERGAP3)	5
Rhine	EM IPE= -5.58	4.36	121.6	7	C1 [-1.86]: HBV*mHM/(5.080+HBV) C2 [5.19]: + log(SWAT) C3 [5.21]: + log(WATERGAP3)	11	-4.25	-132.22	8	C1 [-2.57]: SWAT C2 [7.19]: + log(0.220*VIC*HBV/(WATERGAP3*SWAT))^3	12
Tagus	VIC IPE = -1.50	1.61	-10.7	1	C1 [-1.68]: HBV-sqrt(HBV) C2 [1.30]: + sin(HBV*cos(HYPE/(VIC^4-HBV*VIC^2))) C3 [1.33]: + sin(VIC+ exp(HYPE/((-6.979-WATERGAP3)/HBV)) -9.210)	30	-1.64	14.31	3	C1 [-2.06]: 0.671*HBV-1.76 C2 [1.31]: + sin(HBV)	7
Yangtze	EM IPE = -3.86	3.85	1.4	3	C1 [-3.48]: VIC C2 [3.35]: + 0.982 C3 [3.44]: + sin(cos(VIC*((-0.129*cos(HBV)/SWIM)*(VIC-8.252))))	15	-4.56	69.17	1	C1 [-3.48]: HBV-sqrt(HBV*VIC)-(5.246/HBV) C2 [3.41]: + VIC	11
Yellow	mHM IPE = -1.72	1.69	2.6	4	C1 [-1.84]: mHM- 1.277 C2 [3.26]: + HBV/SWIM C3 [1.45]: + (-7.250)/(sin SWIM*(WATERGAP3-381.141))	9	-1.12	-59.79	6	C1 [-1.13]: HBV C2 [3.66]: + sin((HYMOD*sin(VIC))^3+log(SWIM/WATERGAP3))	12

1-As defined in Section 2.4, equation size is calculated according to the number of model inputs, constants, operators and functions in an equation.

Table S6. MMCs and equations ranked by MMC performance gain (MMC_{PG}) in the out-of-sample validation data set for the 3-component MMC for the super-ensemble experiment. MMC_{PG} is measured against *either* the best performing GHM *or* the EM, whichever of the two performs better.

No	River	MMC IPE score	Global			Regional			MMC equation separated into its components. MMC = C1 + C2 + C3. Components (C1…C3) are ordered according to their explanatory power as assessed by IPE. The IPE value for each component is provided in square brackets.	Eqn. size[1]
			Best performing GHM or EM and IPE score	MMC_{PG} (%)	Rank	Best performing GHM or EM and IPE score	MMC_{PG} (%)	Rank		
1	Upper Amazon	-1.95	WaterGAP2 IPE = -1.43	51.5	8	SWIM IPE = -1.98	-4.5	5	C1 [-1.99]: SWIM+ 0.158538 C2 [12.08]: + cos PCRGLOBWB C3 [12.11]: + cos(PCRGLOBWB ^2*HYMOD)	10
2	Lena	-3.36	MATSIRO IPE = -1.67	168.7	1	SWIM IPE= -3.40	-2.6	4	C1 [-3.40]: SWIM C2 [1.44]: + cos((SWIM/LPJmL)^3) C3 [1.49]: + cos MATSIRO	8
3	Upper Mississippi	-2.60	WaterGAP2 IPE = -1.40	120.6	3	EM‡ IPE = -2.47	42.1	2	C1 [-2.18]: mHM C2 [3.00]: + sqrt(WATERGAP3+log(exp(HYMOD))) C3 [6.22]: + -4.618866	8
4	Upper Niger	-2.06	WaterGAP2 IPE = -1.30	2.4	6	HYMOD IPE = -2.22	-15.9	6	C1 [-1.62]: SWIM C2 [2.31]: + sin(sin(sin SWIM)) C3 [2.76]: + -sqrt(LPJmL/log PCRGLOBWB)	10
5	Rhine	-3.66	WaterGAP2 IPE = -2.51	115.5	4	EM‡ IPE= -5.58	-3.8	3	C1 [-3.65]: SWIM C2 [5.85]: + VIC/SWAT C3 [6.08]: + cos LPJmL^3	7
6	Tagus	-2.13	MATSIRO IPE = -1.26	86.8	5	VIC IPE = -1.50	62.7	1	C1 [1.03]: HBV C2 [1.23]: + sqrt(VIC/MATSIRO^2) C3 [1.99]: + -2.263672-log(MPI-1.335)	12
7	Yangtze	-3.05	EM IPE = -1.79	132.1	2	EM‡ IPE = -3.86	-53.9	8	C1 [-3.48]: VIC C2 [3.10]: + log H08 C3 [3.49]: + sin SWIM	5
8	Yellow	-1.53	EM IPE =-1.18	67.0	7	mHM IPE = -1.72	-19.0	7	C1 [-1.72]: mHM C2 [3.22]: + log(sqrt(LPJmL)) C3 [3.68]: + sin HYMOD	9

‡*EM of CHMs across the catchment*

1-As defined in Section 2.4, equation size is calculated according to the number of model inputs, constants, operators and functions in an equation.

Table 7. MMCs and equations ranked by MMC performance gain (MMC_{PG}) in the out-of-sample validation data set for the 2-component MMC for the super-ensemble experiment. MMC_{PG} is measured against *either* the best performing GHM *or* the EM, whichever of the two performs better.

			Global			**Regional**				
No	**River**	**MMC IPE score**	**Best performing GHM or EM and IPE score**	**MMC_{PG} (%)**	**Rank**	**Best performing GHM or EM and IPE score**	**MMC_{PG} (%)**	**Rank**	**MMC equation separated into its components. MMC = C1 + C2. Components (C1…C3) are ordered according to their explanatory power as assessed by IPE. The IPE value for each component is provided in square brackets.**	**Eqn. size[1]**
1	Upper Amazon	-2.04	WaterGAP2 IPE = -1.43	61.04	6	SWIM IPE = -1.98	5.06	3	C1 [-1.92]: SWIM-9.555 C2 [10.56]: + sqrt(MATSIRO)	5
2	Lena	-4.07	MATSIRO IPE = -1.67	240.07	1	SWIM IPE= -3.40	67.10	1	C1 [-2.89]: HYPE C2 [1.71]: + (cos(PCRGLOBWB)-LPJmL)/sqrt(PCRGLOBWB+ECOMAG)	10
3	Upper Mississippi	-2.29	WaterGAP2 IPE = -1.40	89.70	4	EM‡ IPE = -2.47	-18.04	4	C1 [-2.18]: mHM C2 [34.55]: + YMOD*(SWAT/(((mHM/MATSIRO)*MPI)*WaterGAP2))	12
4	Upper Niger	-2.00	WaterGAP2 IPE = -1.30	69.93	5	HYMOD IPE = -2.22	-21.80	5	C1 [-1.54]: VIC C2 [2.83]: + -0.109*VIC	4
5	Rhine	-3.86	WaterGAP2 IPE = -2.51	135.45	3	EM‡ IPE= -5.58	-171.40	8	C1 [-3.86]: SWAT-4.210 C2 [6.04]: + 4.210/WaterGAP2	6
6	Tagus	-1.80	MATSIRO IPE = -1.26	54.27	7	VIC IPE = -1.50	-30.19	7	C1 [-1.48]: HBV-2.351 C2 [151]: + -0.103*CLM	6
7	Yangtze	-3.99	EM IPE = -1.79	220.07	2	EM‡ IPE = -3.86	11.54	2	C1 [-3.48]: VIC C2 [3.49]: + sin(log(SWIM))	4
8	Yellow	-1.49	EM IPE =-1.18	31.36	8	mHM IPE = -1.72	-22.73	6	C1 [-1.79]: 0.983*mHM C2 [3.07]: + log(MATSIRO)	5

‡*EM of CHMs across the catchment*

1-As defined in Section 2.4, equation size is calculated according to the number of model inputs, constants, operators and functions in an equation.

Table S8. IPE scores for individual GHMs and CHMs, their EM and MMC solutions (all computed for the out-of-sample validation period). Grey shading denotes where an individual model, the EM, or MMC solution outperforms the benchmark model. The best performing of all individual models (i.e. GHMs and CHMs), the EM and all MMC solutions, for each catchment is in bold. Red text denotes where an MMC solution does not perform better than the best individual model or EM, while green denotes where an MMC does.

GHMs

Catchment	CLM	DBH	H08	LPJmL	MATSIRO	MPI	PCR-GLOBWB	WaterGAP2		EM	3-component MMC	2-component MMC	Super-ensemble 3-component MMC	Super-ensemble 2-component MMC
Upper Amazon	1.44	1.07	1.95	1.77	2.44	1.45	2.99	-1.43	–	1.24	-1.78	-1.44	-1.95	**-2.04**
Lena	1.16	1.92	1.86	1.04	-1.67	-1.18	1.16	-1.10	–	-1.15	-2.14	-1.59	-3.36	**-4.07**
Upper Mississippi	2.19	6.57	1.26	1.58	-1.09	2.00	1.84	-1.40	–	1.20	-1.42	-1.48	**-2.60**	-2.29
Upper Niger	5.33	8.70	10.05	12.98	3.92	4.31	3.95	-1.30	–	4.42	-1.66	-1.62	**-2.06**	-2.00
Rhine	5.34	3.62	2.14	2.09	1.26	1.95	1.11	-2.51	–	1.53	-2.53	-2.50	-3.66	**-3.86**
Tagus	1.46	4.43	2.35	2.96	-1.26	1.67	1.38	1.21	–	1.70	-1.68	1.47	**-2.13**	-1.80
Yangtze	1.44	-1.60	1.15	-1.73	-1.13	-1.00	-1.68	-1.63	–	-1.79	-3.27	-2.48	-3.05	**-3.99**
Yellow	2.34	3.48	1.92	1.31	1.50	1.39	4.41	1.14	–	-1.18	-1.15	-1.26	**-1.53**	-1.49
Mean	2.59	3.52	2.84	2.75	0.50	1.32	1.90	-0.88	–	0.75	-1.95	-1.36	-2.54	**-2.69**

CHMs

Catchment	ECOMAG	HBV	HYMOD	HYPE	mHM	SWAT	SWIM	VIC	WATERGAP3	EM	3-component MMC	2-component MMC	Super-ensemble 3-component MMC	Super-ensemble 2-component MMC
Upper Amazon		-1.62	-1.17		-1.09	-1.39	-1.99	-1.61	1.46	-1.48	-1.98	-2.00	-1.95	**-2.04**
Lena	-2.76			-2.89			-3.40	-1.41	1.29	-1.99	-2.17	-4.04	-3.36	**-4.07**
Upper Mississippi		-1.58	-1.44		-2.18	-1.79	-1.70	-1.40	-1.14	-2.47	-1.73	-2.48	**-2.60**	-2.29
Upper Niger		-1.28	**-2.22**		-2.01	1.17	-1.62	-1.54	1.46	-1.68	-1.58	-1.51	-2.06	-2.00
Rhine		-3.70	-2.41	-2.52	-3.77	-2.57	-3.65	-3.12	-2.15	**-5.58**	-4.36	-4.25	-3.66	-3.86
Tagus		1.03		-1.22				-1.50	1.11	-1.32	-1.61	-1.64	**-2.13**	-1.80
Yangtze		-3.59				-2.31	-2.34	-3.48		-3.86	-3.86	**-4.56**	-3.05	-3.99
Yellow		-1.13	-1.26		**-1.72**		-1.20	1.11	-1.61	-1.44	-1.69	-1.12	-1.53	-1.49
Mean	-2.76	-1.70	-1.70	-2.21	-2.15	-1.38	-2.27	-1.62	0.06	-2.48	-2.37	**-2.70**	-2.54	-2.69

Table S9. Performance gain (PG) achieved by the EM and MMCs (both relative to the best individual model), and the PG achieved by MMC relative to the EM. Grey shading denotes performance gain. The best performing model combination, i.e. the EM or one of the MMCs, is shown in bold for each catchment.

GHMs

Catchment	IPE						PG over best individual GHM (%)					PG over GHM EM (%)			
	Best GHM	EM	3-component MMC	2-component MMC	Super-ensemble 3-component MMC	Super-ensemble 2-component MMC	EM	3-component MMC	2-component MMC	Super-ensemble 3-component MMC	Super-ensemble 2-component MMC	3-component MMC	2-component MMC	Super-ensemble 3-component MMC	Super-ensemble 2-component MMC
Upper Amazon	-1.43	1.24	-1.78	-1.44	-1.95	-2.04	-466.9	35.4	0.6	51.5	**61.0**	402.3	367.5	418.4	**427.9**
Lena	-1.67	-1.15	-2.14	-1.59	-3.36	-4.07	-51.9	46.9	-7.9	168.8	**240.1**	98.8	44.0	220.6	**291.9**
Upper Mississippi	-1.40	1.20	-1.42	-1.48	-2.60	-2.29	-459.2	2.4	8.1	**120.6**	89.7	361.7	367.3	**479.8**	448.9
Upper Niger	-1.30	4.42	-1.66	-1.62	-2.06	-2.00	-772.4	35.7	31.8	**75.9**	69.9	708.1	704.2	**748.2**	742.3
Rhine	-2.51	1.53	-2.53	-2.50	-3.66	-3.86	-603.2	2.1	-0.6	115.5	**135.5**	505.3	502.6	618.7	**638.7**
Tagus	-1.26	1.70	-1.68	1.47	-2.13	-1.80	-495.7	42.3	-472.8	**86.8**	54.3	438.1	22.9	**482.5**	450.0
Yangtze	-1.73	-1.79	-3.27	-2.48	-3.05	-3.99	5.7	154.2	75.5	132.1	**225.8**	148.5	69.8	126.4	**220.1**
Yellow	1.14	-1.18	-1.15	-1.26	-1.53	-1.49	331.9	328.5	340.4	**367.0**	363.2	-3.4	8.5	**35.1**	31.4
Mean		0.75	-1.95	-1.36	-2.54	-2.69	-314.0	81.0	-3.1	139.8	**154.9**	332.4	260.8	391.2	**406.4**

CHMs

Catchment	IPE						PG over best individual CHM (%)					PG over CHM EM (%)			
	Best CHM	EM	3-component MMC	2-component MMC	Super-ensemble 3-component MMC	Super-ensemble 2-component MMC	EM	3-component MMC	2-component MMC	Super-ensemble 3-component MMC	Super-ensemble 2-component MMC	3-component MMC	2-component MMC	Super-ensemble 3-component MMC	Super-ensemble 2-component MMC
Upper Amazon	-1.99	-1.48	-1.98	-2.00	-1.95	-2.04	-51.3	-0.8	1.1	-4.5	**5.1**	50.6	52.4	46.9	**56.4**
Lena	-3.40	-1.99	-2.17	-4.04	-3.36	-4.07	-141.1	-123.2	64.2	-4.2	**67.1**	17.9	205.2	136.8	**208.2**
Upper Mississippi	-2.18	-2.47	-1.73	-2.48	-2.60	-2.29	29.3	-44.7	29.8	**42.1**	11.2	-74.0	0.5	**12.8**	-18.0
Upper Niger	-2.22	-1.68	-1.58	-1.51	-2.06	-2.00	-54.2	-63.6	-70.9	**-15.9**	-21.8	-9.4	-16.7	**38.3**	32.4
Rhine	-3.70	-5.58	-4.36	-4.25	-3.66	-3.86	**187.5**	66.0	55.3	-3.8	16.2	**121.6**	-132.2	-191.4	-171.4
Tagus	-1.50	-1.32	-1.61	-1.64	-2.13	-1.80	-18.4	10.7	14.3	**62.7**	30.2	29.1	32.7	**81.1**	48.6
Yangtze	-3.59	-3.86	-3.86	-4.56	-3.05	-3.99	-28.2	26.8	**97.4**	-54.0	39.8	-1.4	**69.2**	-82.2	11.5
Yellow	-1.72	-1.44	-1.69	-1.12	-1.53	-1.49	-27.6	**-2.6**	-59.8	-19.0	-22.7	**25.0**	-32.2	8.6	4.9
Mean		-2.48	-2.37	-2.70	-2.54	-2.69	-5.9	-16.4	**16.4**	0.4	15.6	-10.5	**22.4**	6.4	21.6

Table S10. The contribution of individual GHMs and CHMs to MMCs derived for the super-ensemble. Crosses in underlined bold (and totals in brackets) indicate the inclusion of the individual models in the first component of the symbolic regression equation.

Catchment	CLM	DBH	H08	LPJmL	MATSIRO	MPI	PCR-GLOBWB	WaterGAP2	ECOMAG	HBV	HYMOD	HYPE	mHM	SAWT	SWIM	VIC	WATERGAP3	No of models
Upper Amazon							x				x				**x**			3(1)
Lena				x	x										**x**			3(1)
Upper Mississippi											x		**x**				x	3(1)
Upper Niger				x			x								**x**			3(1)
Rhine				x										x	**x**	x		4(1)
Tagus					x	x				**x**						x		4(1)
Yangtze			x												x	**x**		3(1)
Yellow				x							x		**x**					3(1)
Totals	0 (0)	0 (0)	1 (0)	4 (0)	2 (0)	1 (0)	2 (0)	0 (0)	0 (0)	1 (1)	3 (0)	0 (0)	2 (2)	1 (0)	5 (4)	3 (1)	1 (0)	

.

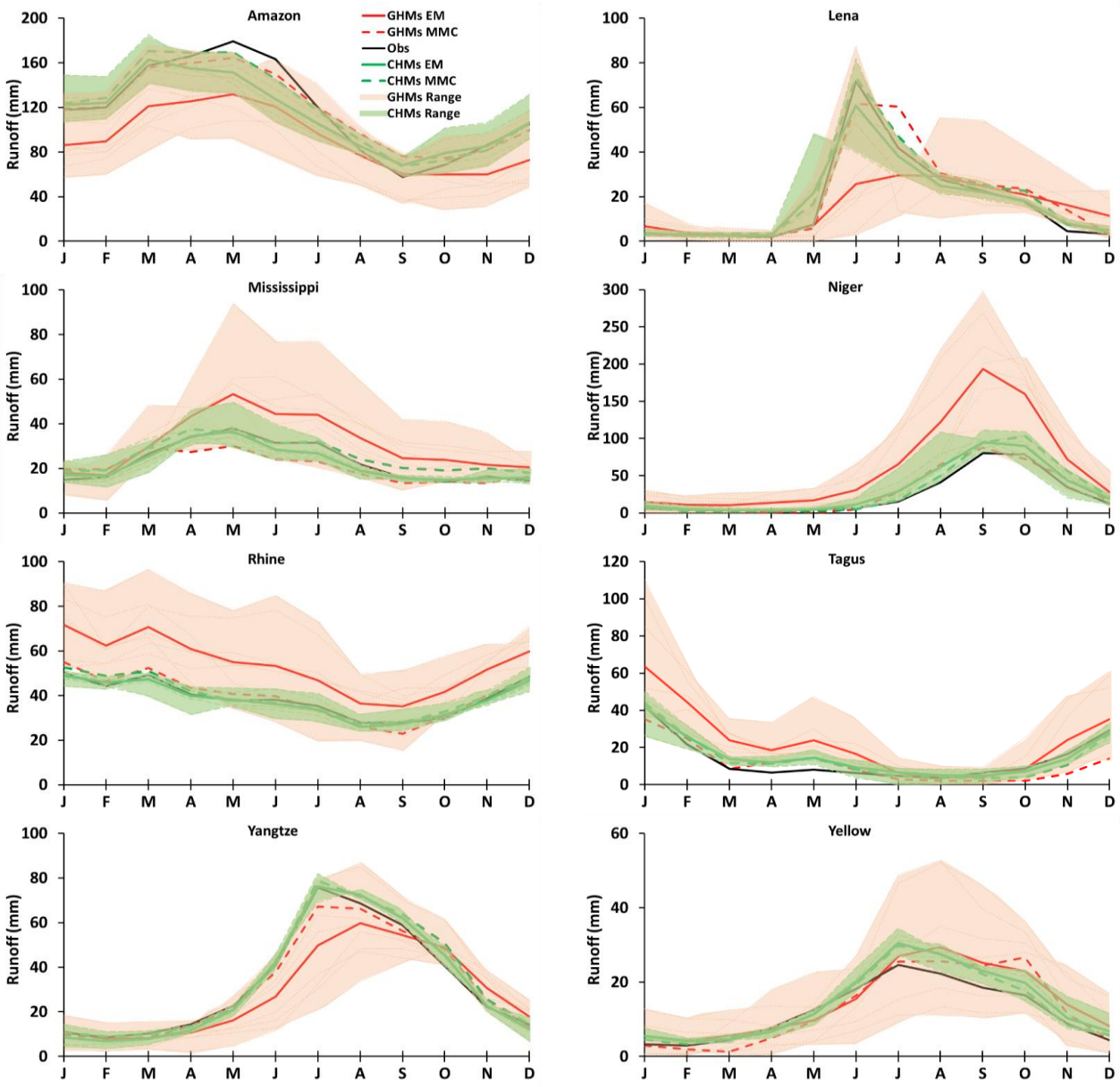
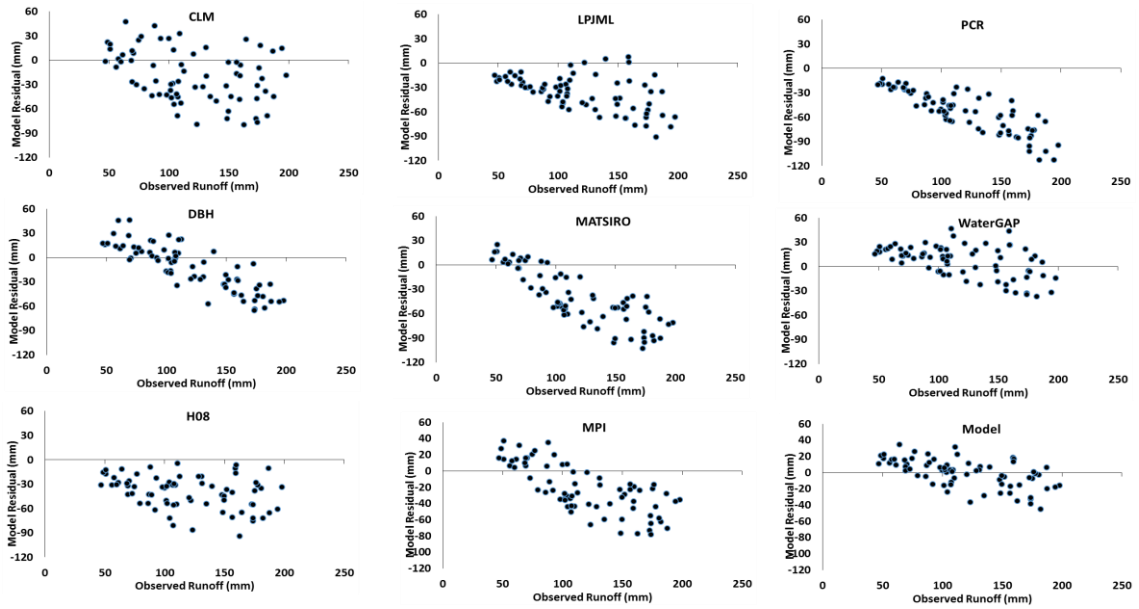
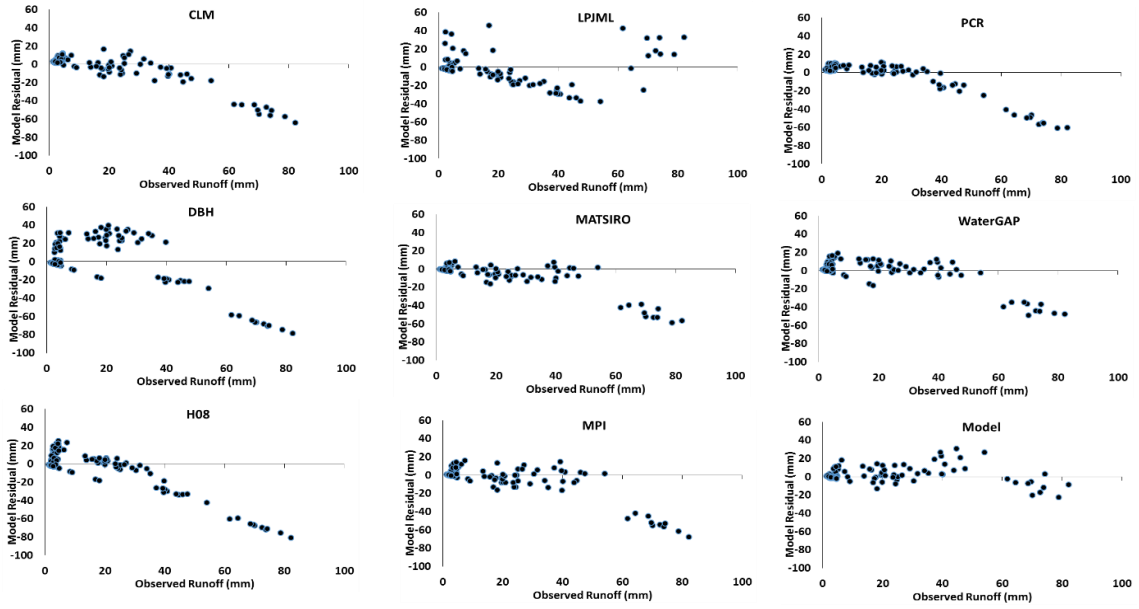


Figure S2. Seasonal cycle (long-term mean monthly runoff, MMR). The shaded region denotes the models' performance range.

.

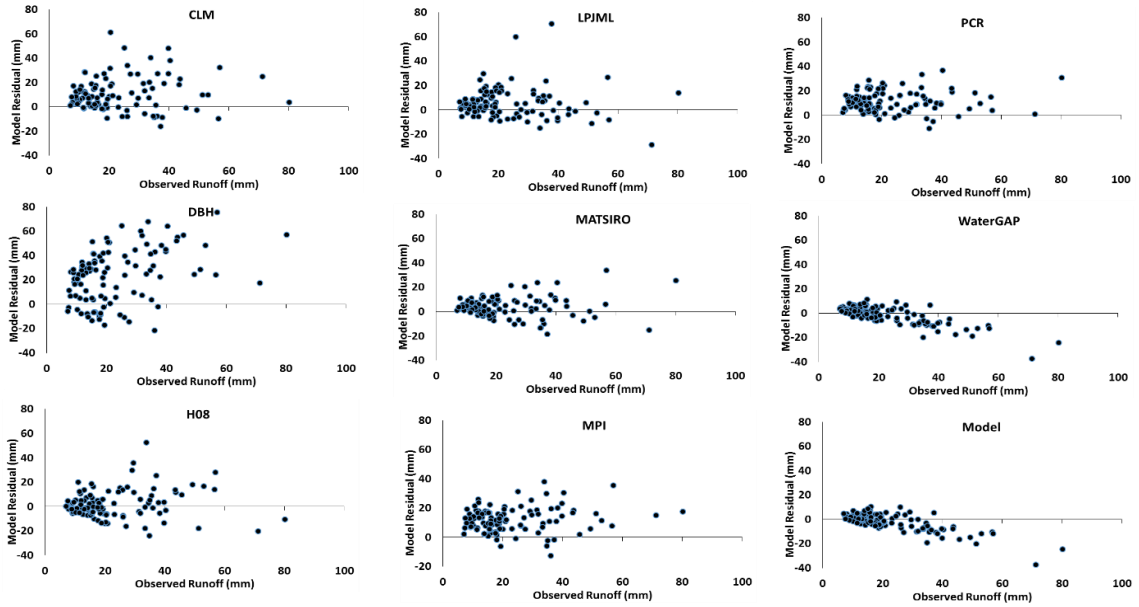


Upper Amazon

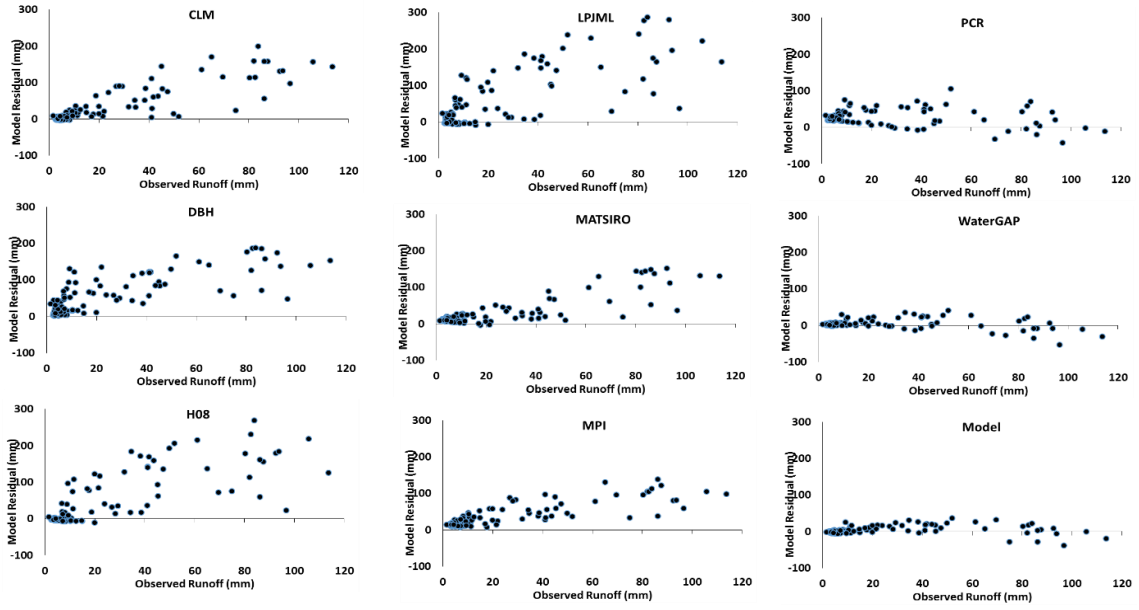


Lena

Figure S3. Residual plots for the GHMs and 3-component MMC (Model).



Upper Mississippi



Upper Niger

Figure S3 (continued). Residual plots for the GHMs and 3-component MMC (Model).

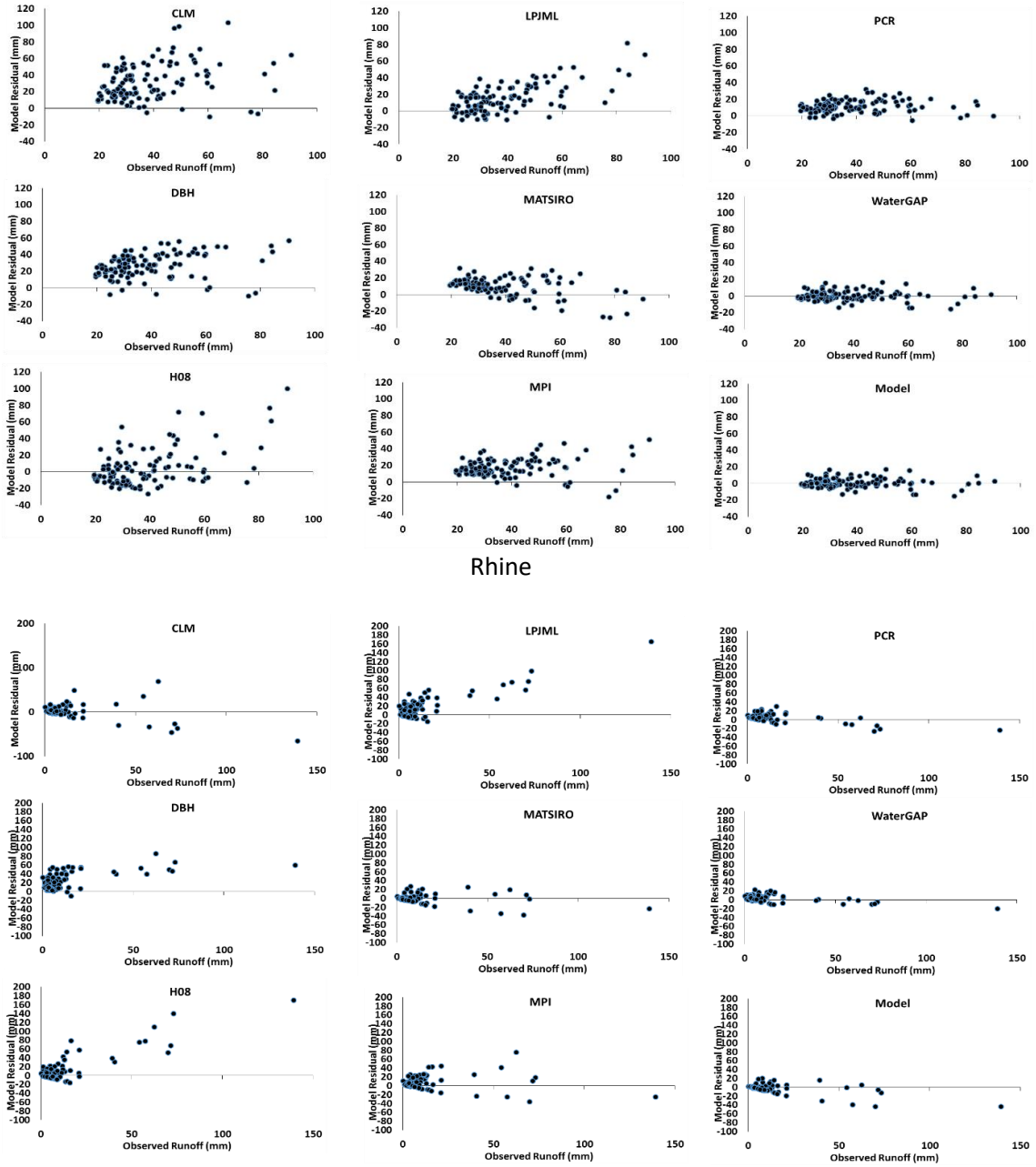
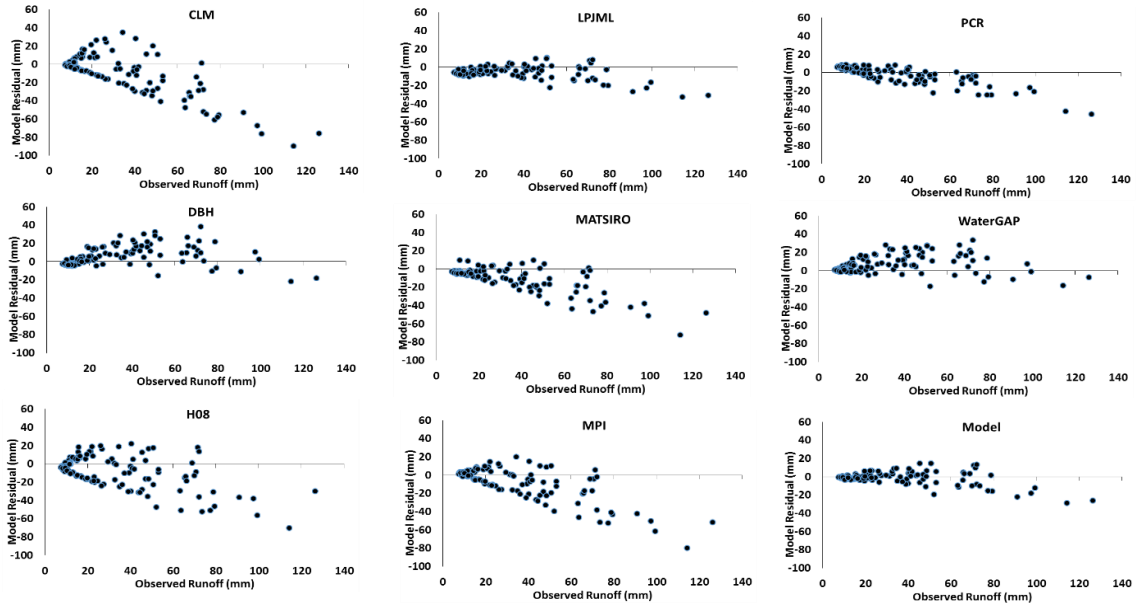
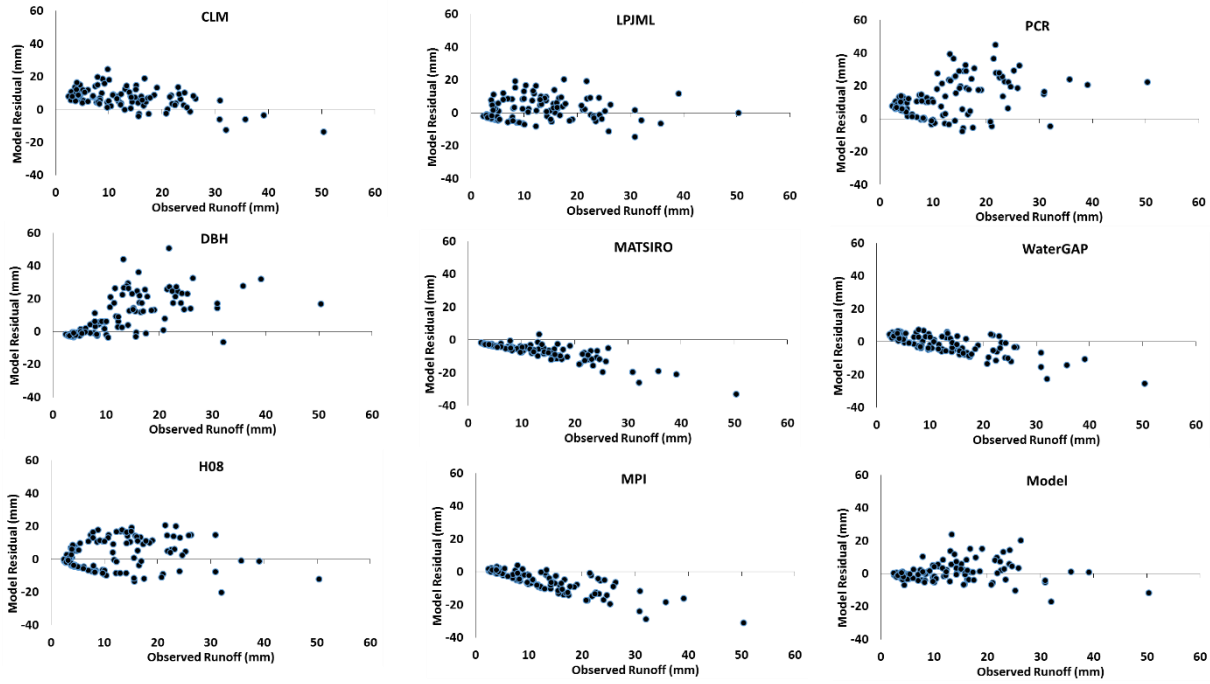


Figure S3 (continued). Residual plots for the GHMs and 3-component MMC (Model).



Yangtze



Yellow

Figure S3 (continued). Residual plots for the GHMs and 3-component MMC (Model).

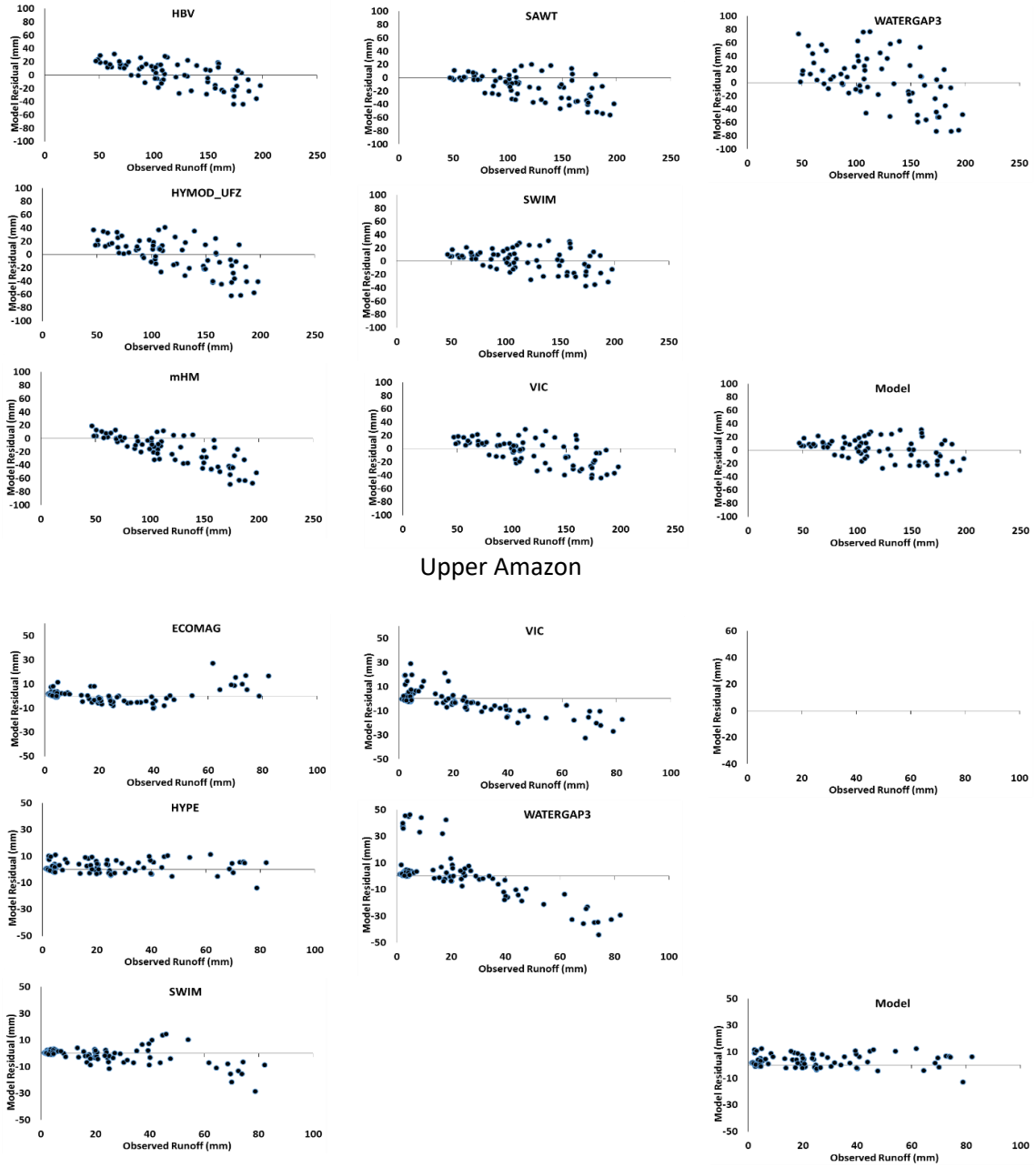
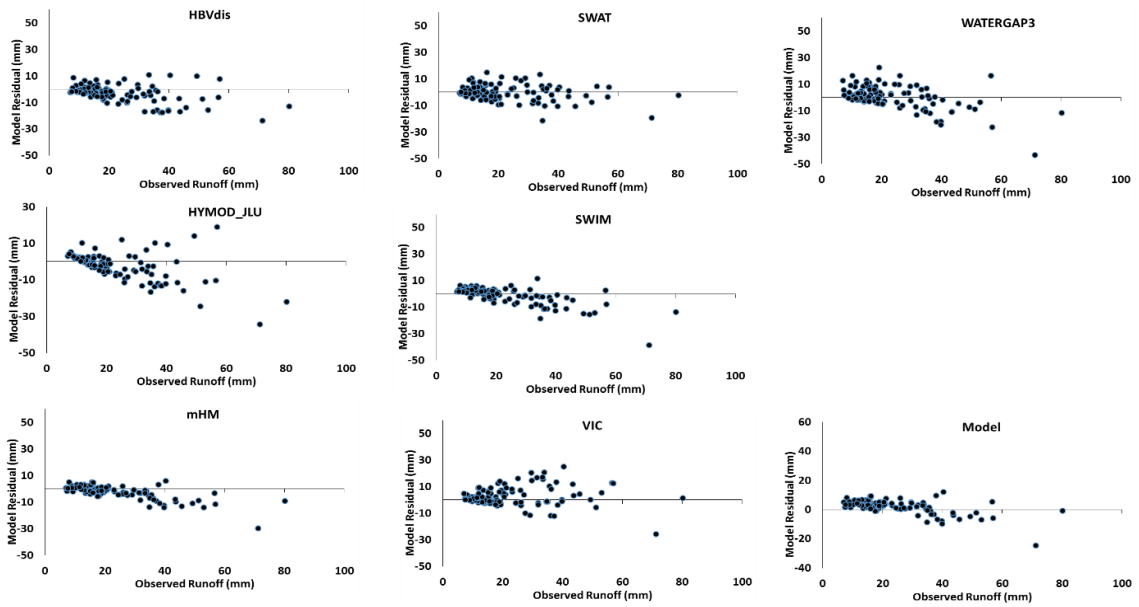
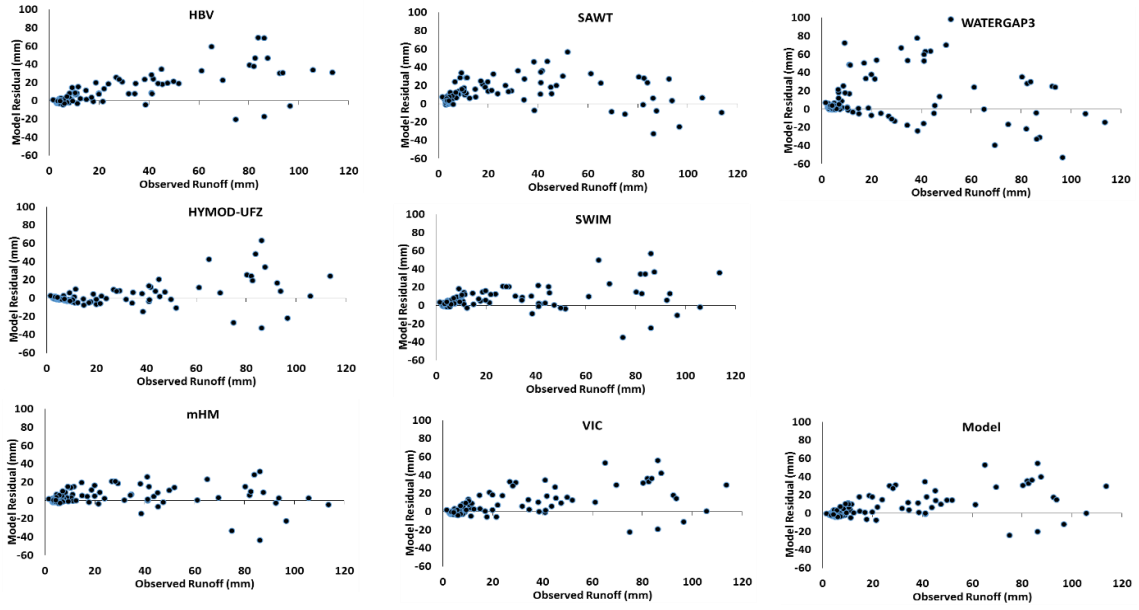


Figure S4. Residual plots for the CHMs and 3-component MMC (Model).

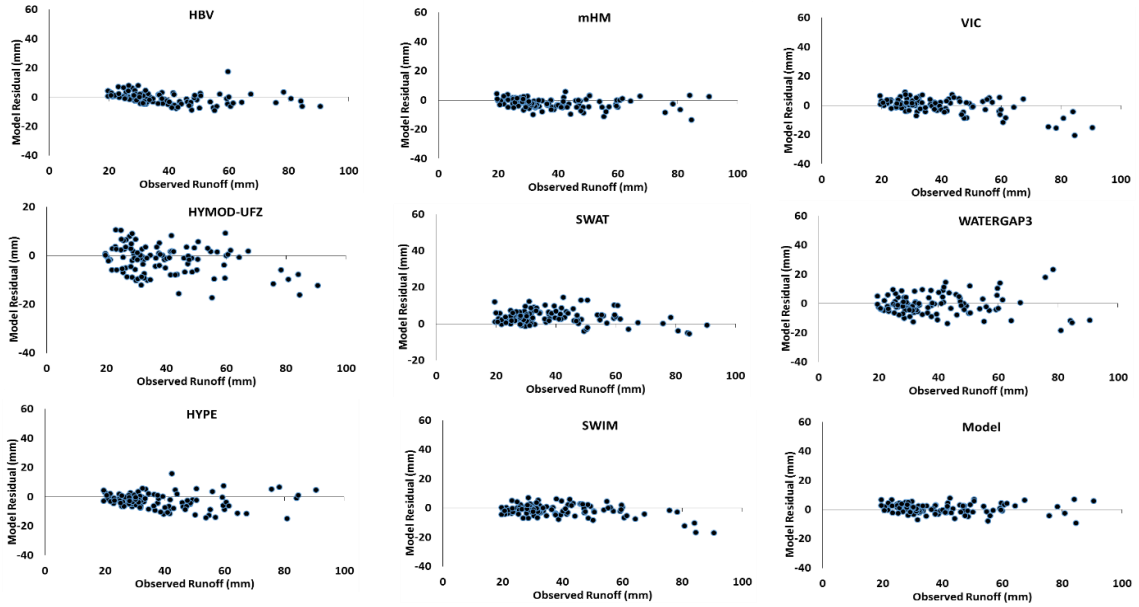


Upper Mississippi

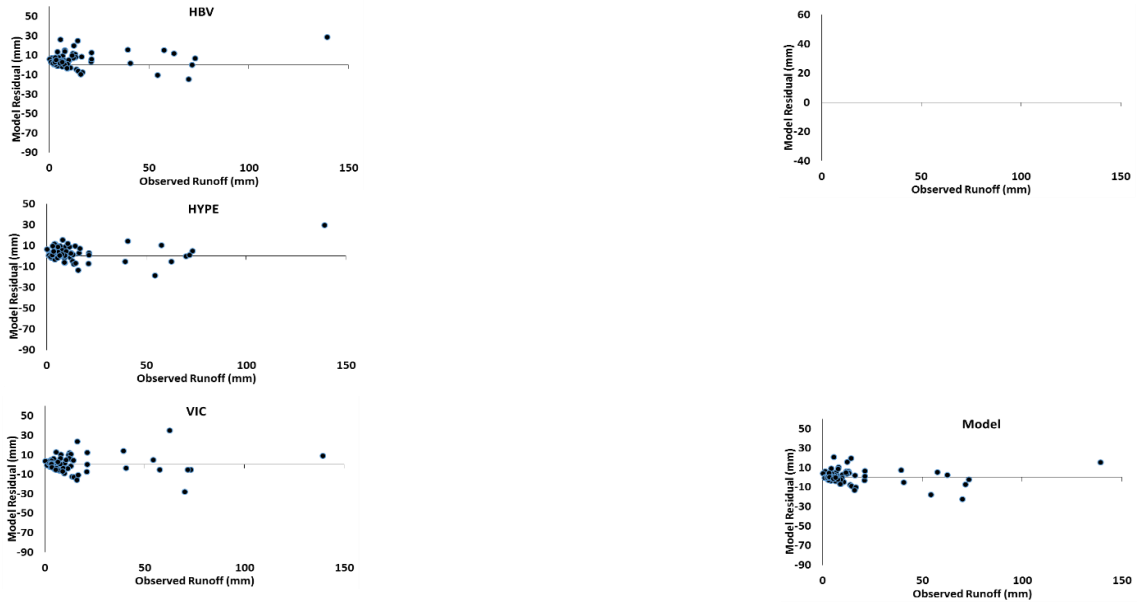


Upper Niger

Figure S4 (continued). Residual plots for the CHMs and 3-component MMC (Model).



Rhine



Tagus

Figure S4 (continued). Residual plots for the CHMs and 3-component MMC (Model).

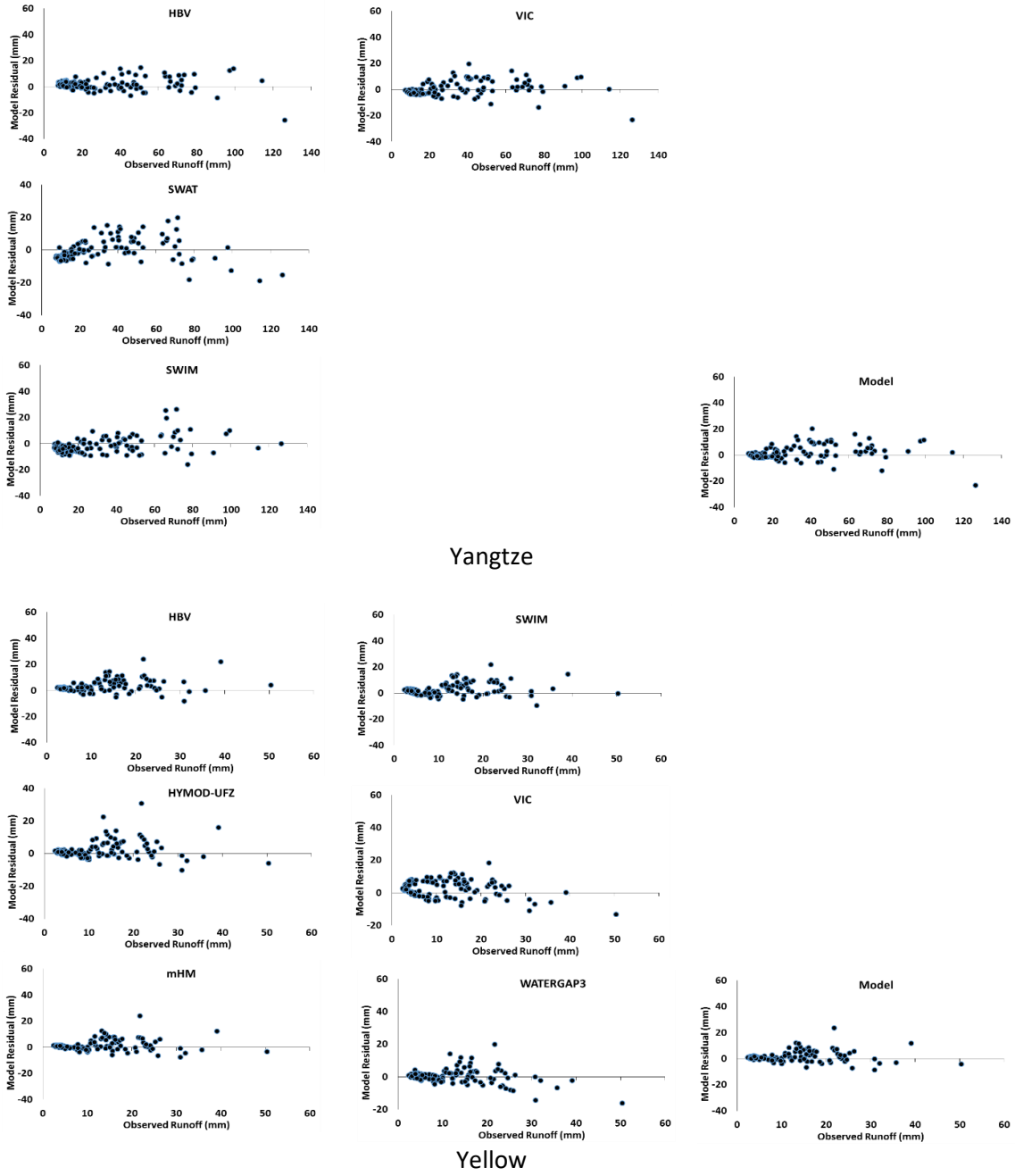


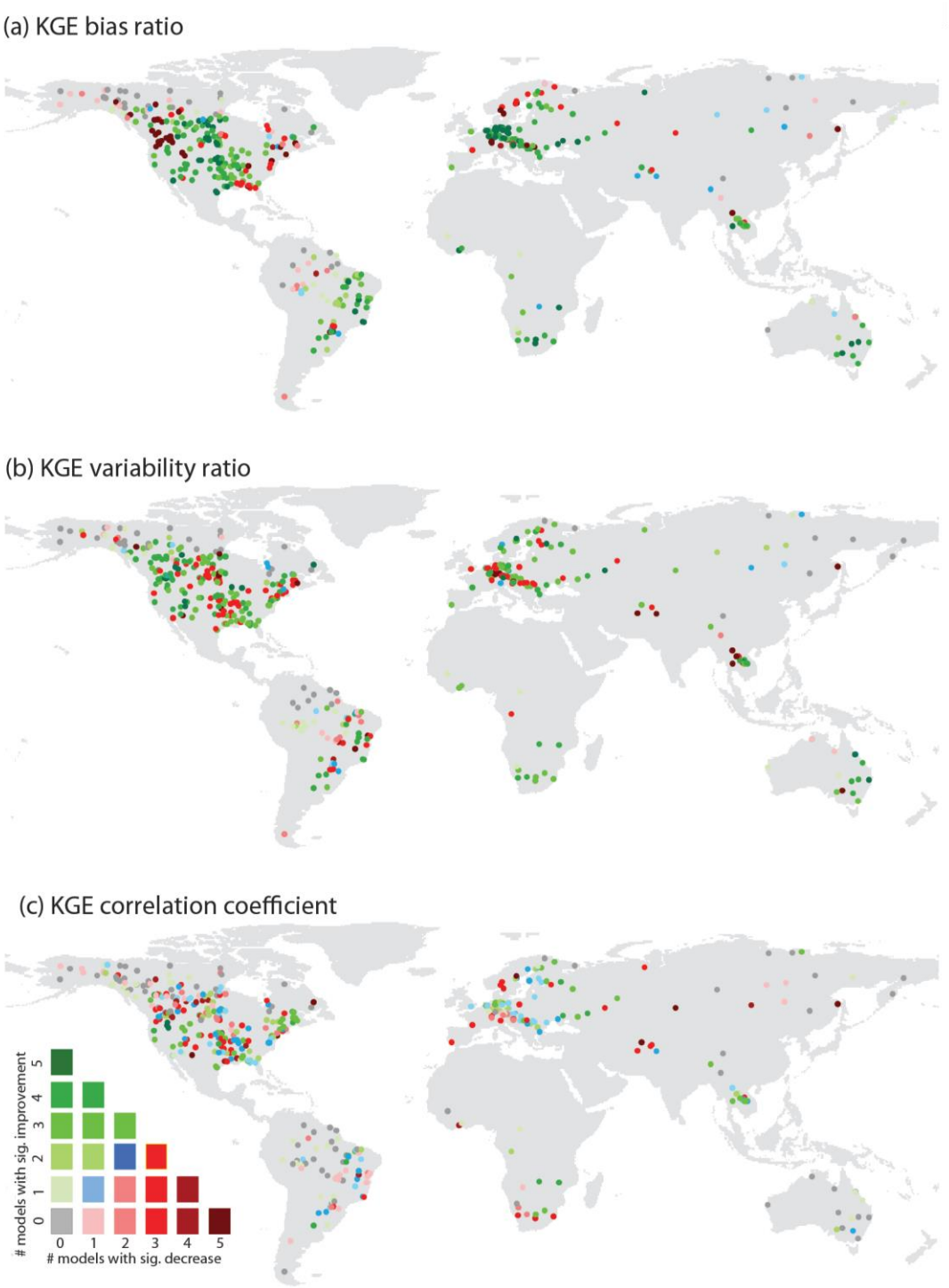
Figure S4 (continued). Residual plots for the CHMs and 3-component MMC (Model).

References

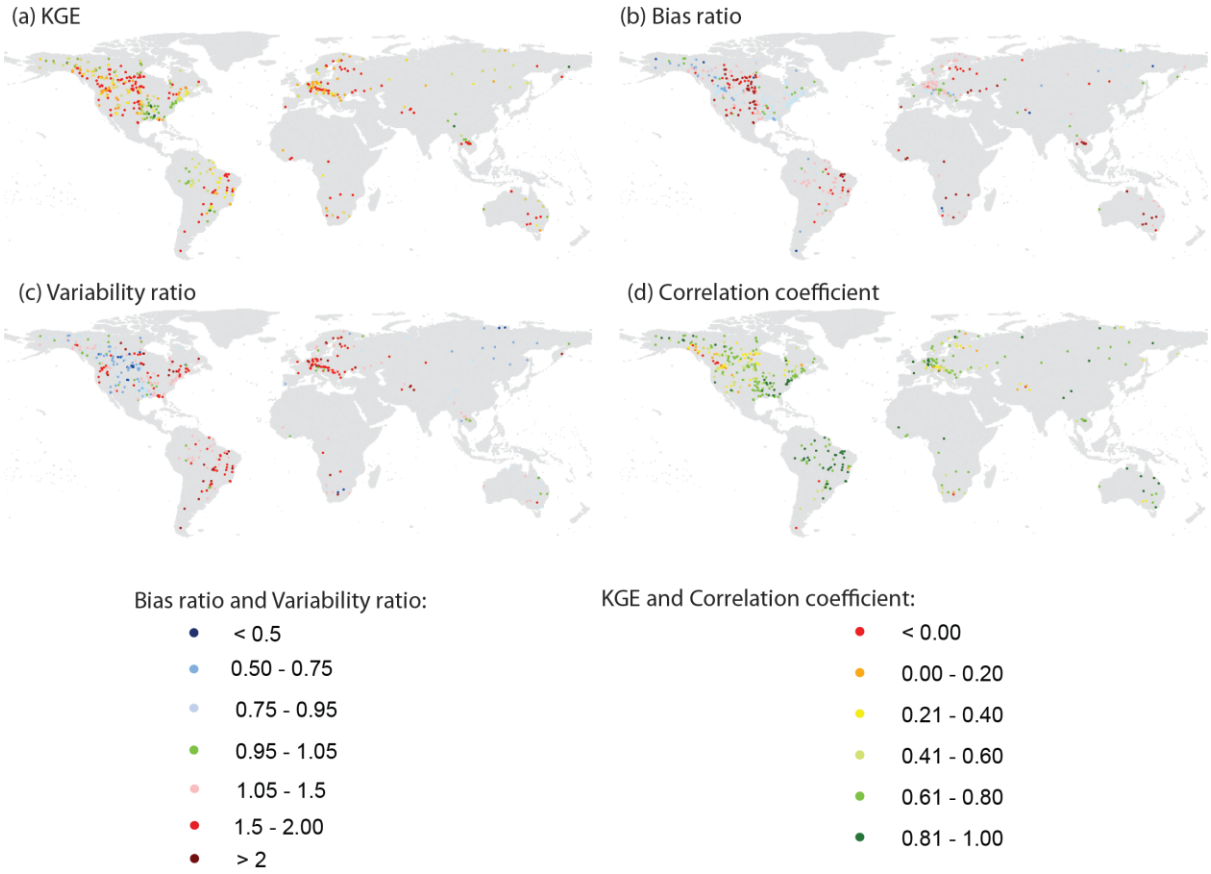
FERREIRA, C. 2001. Gene Expression Programming: A New AdaptiveAlgorithm for Solving Problems. 13**,** 87-129.

Appendix D (Supplementary information for Paper V)

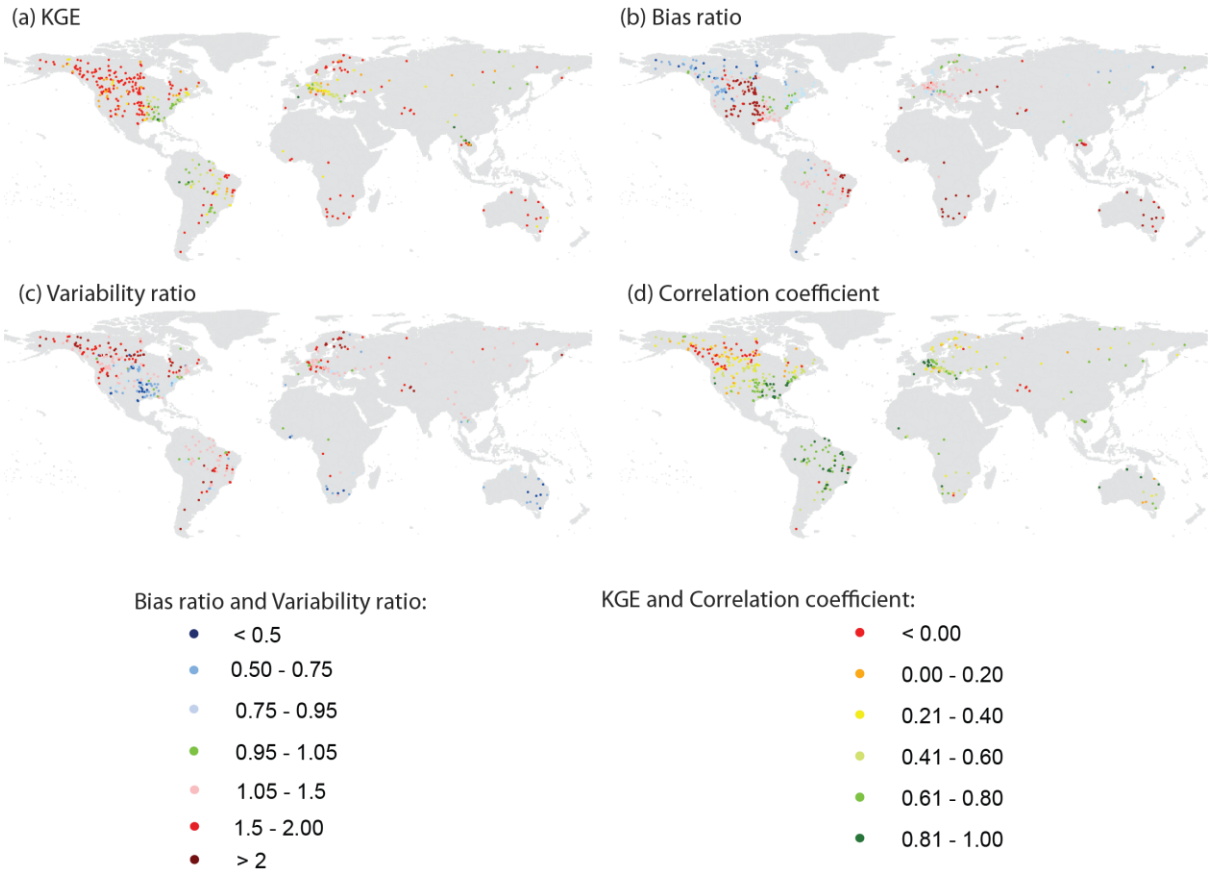
Supplementary Figures



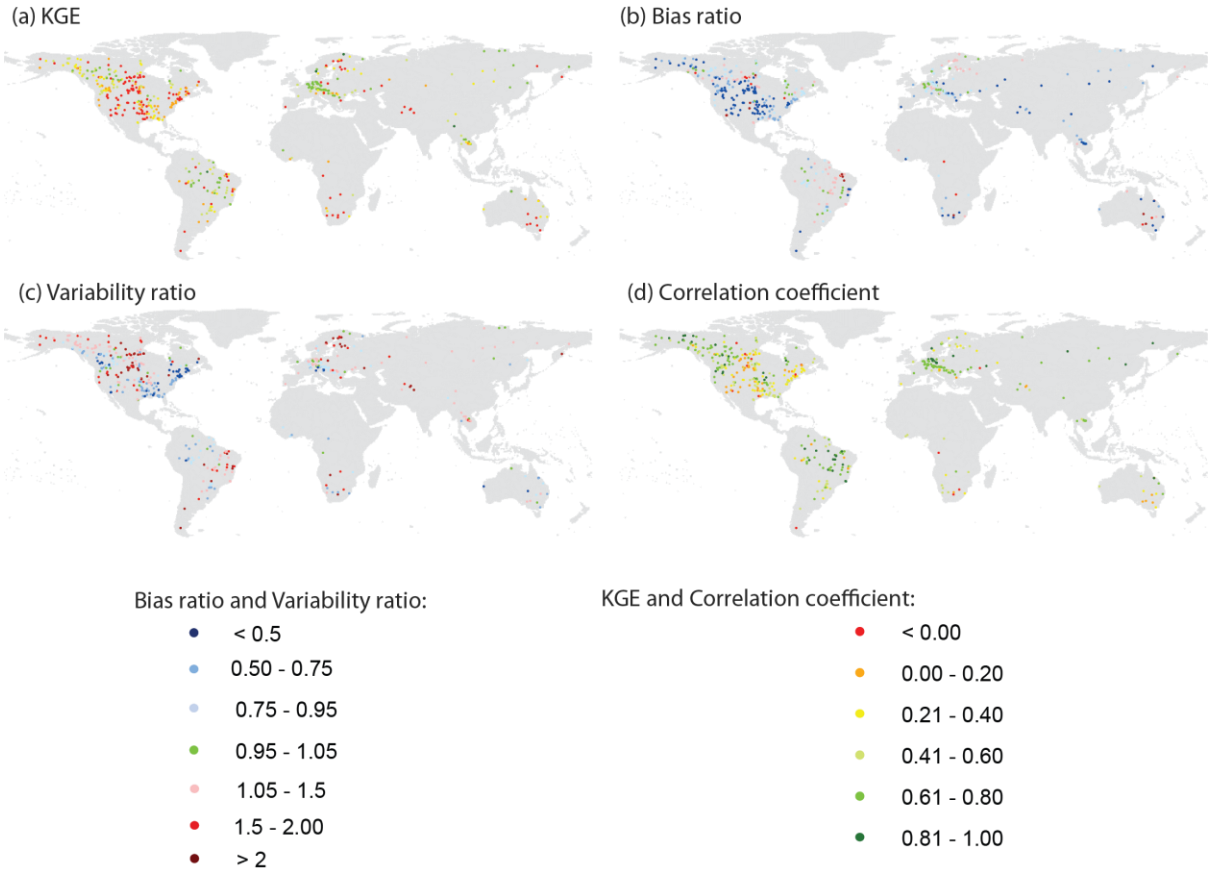
Supplementary figure 1: Number of GHMs with a significant increase or decrease in hydrological performance (KGE) due to HIP. Figures (a), (b), and (c) show for each of the underlying KGE sub-parameters (bias ratio, variability ratio, correlation coefficient) the number of models with a significant increase or decrease in performance.



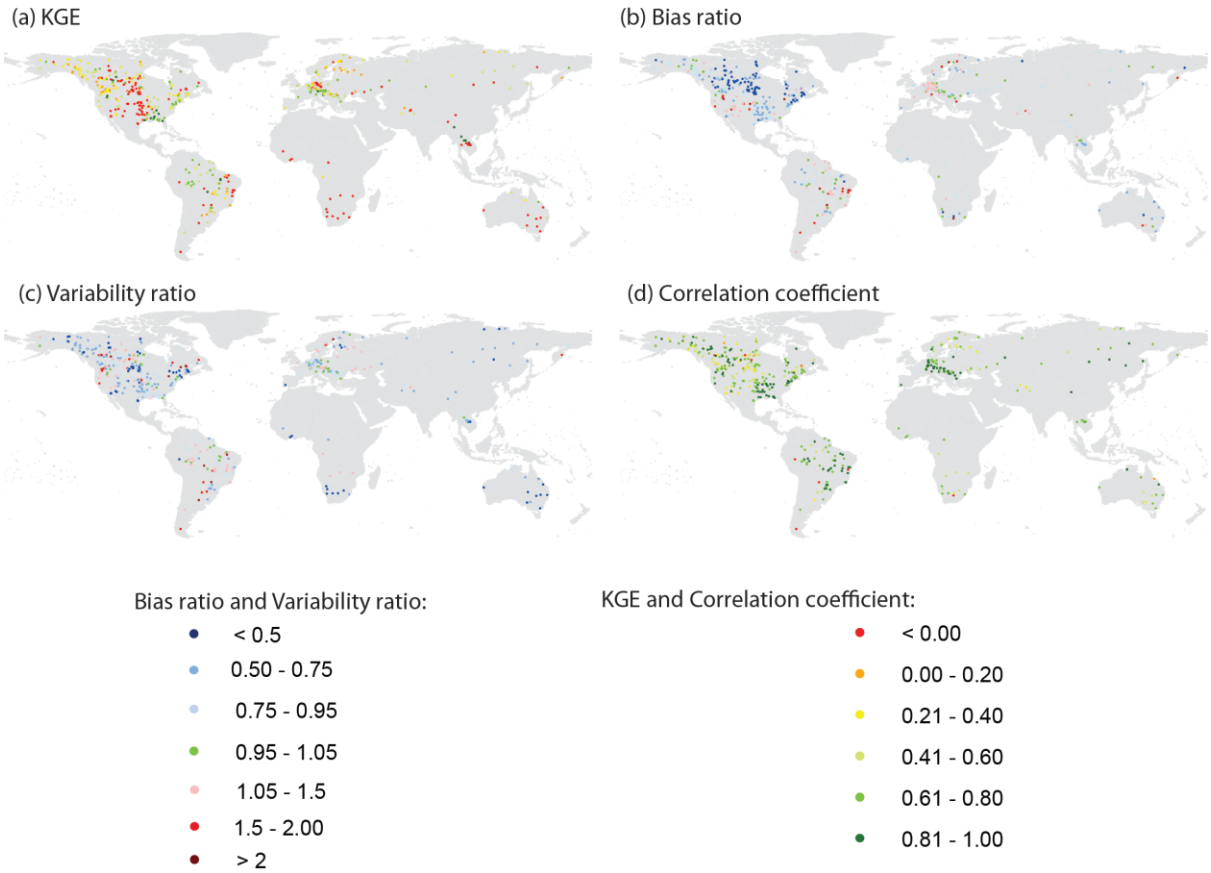
Supplementary figure 2a: KGE performance of H08 under HIP conditions.



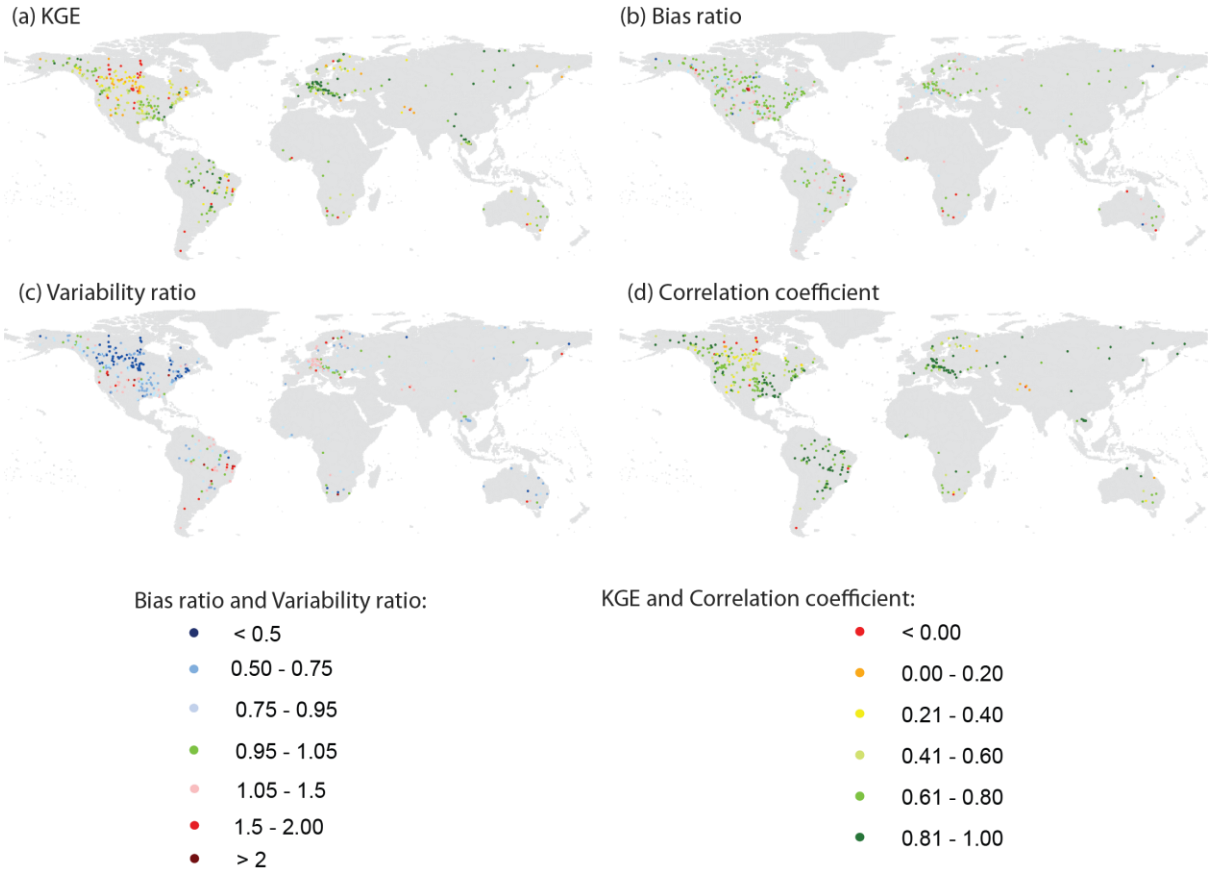
Supplementary figure 2b: KGE performance of LPJmL under HIP conditions.



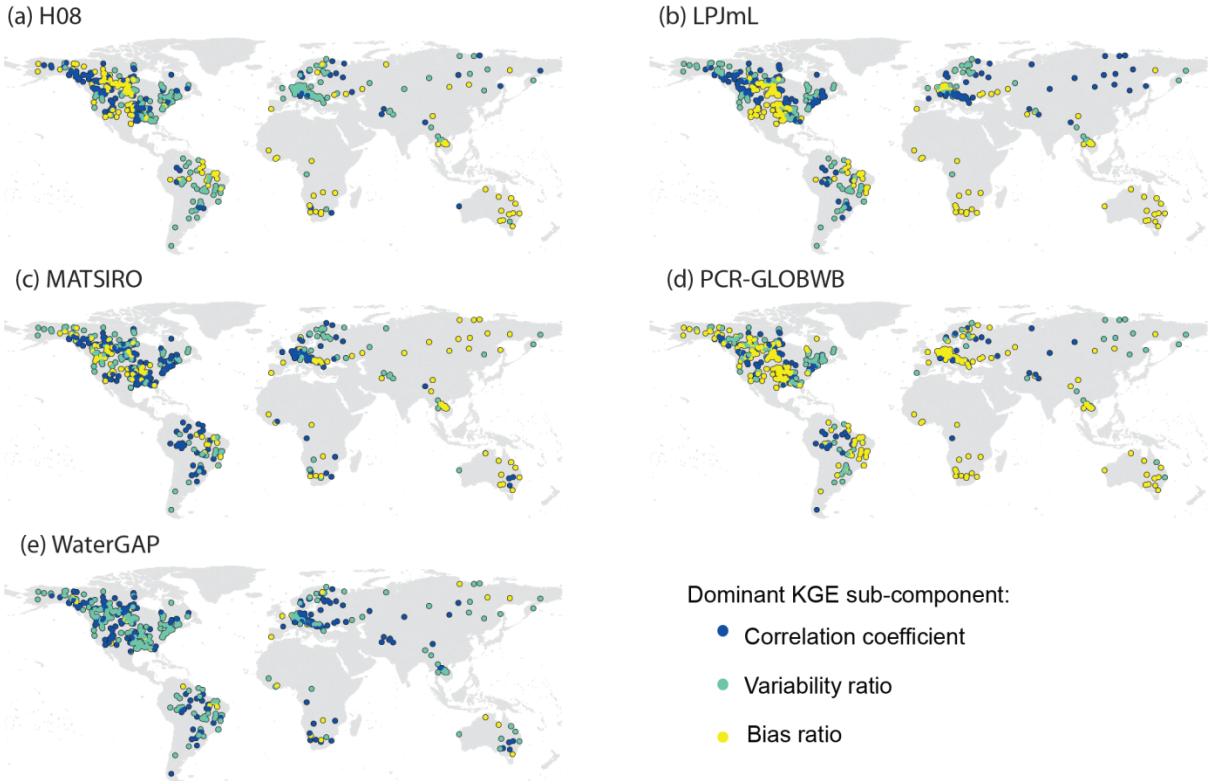
Supplementary figure 2c: KGE performance of MATSIRO under HIP conditions.



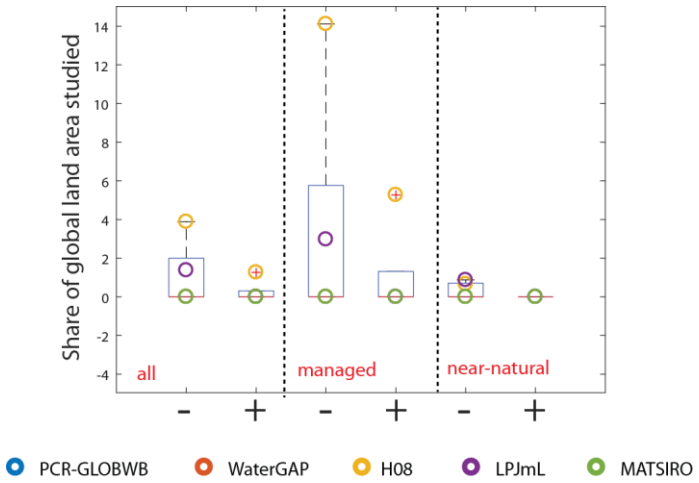
Supplementary figure 2d: KGE performance of PCR-GLOBWB under HIP conditions.



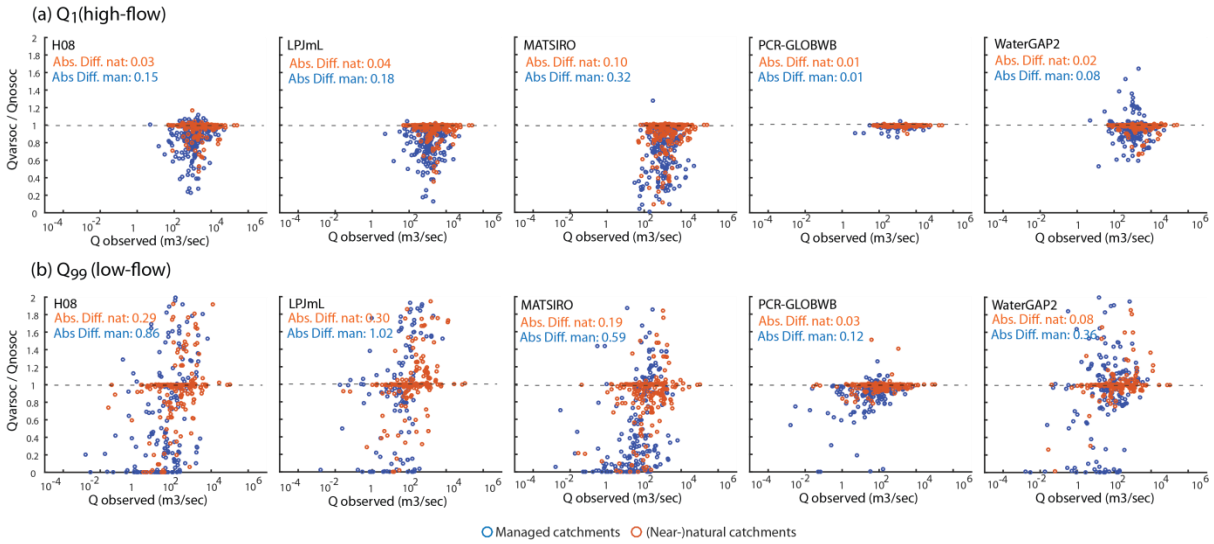
Supplementary figure 2e: KGE performance of WaterGAP2 under HIP conditions.



Supplementary figure 3: Spatial distribution of dominant KGE sub-component limiting optimal hydrological performance.



Supplementary figure 4: Share of land area with a significant change in the representation of the exceedance probability curve, as tested with the Kolmogorov-Smirnov test. WaterGAP2 and PCR-GLOBWB are located underneath the results of MATSIRO.



Supplementary figure 5: Scatterplot showing for the individual global hydrological models (GHMs) the difference in hydrological extremes due to HIP.

The two sets of panels show the difference (factor) in modelled (a) Q_1 high-flow and (b) Q_{99} low-flow discharges under the HIP and NOHIP conditions for the managed (blue) and near-natural catchments (orange). Values >1 imply that discharges are larger under HIP conditions compared to NOHIP conditions, and vice versa. To improve visibility, results are capped at a value of 2. The absolute differences given in each of the panels represent the average absolute deviation from unity for the managed and near-natural catchments. The larger the average absolute deviation, the larger the difference in modelled Q_1 high-flow and Q_{99} low-flow discharges between HIP and NOHIP conditions.

Supplementary Tables

Supplementary table 1: Description of the modelling framework and parameterizations of the models. Table 1.A presents the representation of the hydrological processes in the models, after Zhao et al., 2017. Table 1.B-I present the representation of the human dimensions and its interactions with the hydrological processes for each of the models.

1.A Representation of hydrological processes

Model name	Energy balance	Evaporation scheme	Runoff scheme	Snow scheme	Routing scheme	Flow velocity	Soil Water Layer Depth	References
H08	Yes	Bulk formula	Saturation excess, non-linear	Energy balance	TRIP Linear reservoir (Oki and Sud 1998)	0.5m/s	One soil layer with a depth of 1m	Hanasaki et al., 2008a,b
LPJmL	No	Priestley-Taylor	Saturation excess	Degree-day	Continuity equation derived from linear reservoir model	1 m/s	Five hydrologically active layers of 20, 30, 50, 100 and 100 cm thickness, respectively	Bondeau et al., 2007; Schaphoff et al., 2013
MATSIRO	Yes	Bulk formula	Overland flow, infiltration excess, saturation excess, groundwater.	Energy balance	TRIP Linear reservoir (Oki and Sud 1998)	0.5m/s	12 fully resolved layers (5cm, 20cm, 75cm, and nine next layers of 1m) and a 90m groundwater layer.	Takata et al., 2003; Pokhrel et al., 2012; 2015
PCR-GLOBWB	No	Hamon	Saturation Excess Beta Function	Degree Day	Travel time routing (characteristic distance)	Variable based on Manning's equation	Variable up to 1.5 m soil layers and 50 m groundwater layer	van Beek et al., 2011; Wada et al., 2011
WaterGAP2	No	Priestley Taylor with two alpha factors depending on the aridity of the grid cell	Beta function, saturation excess	Degree Day	Linear reservoir	Variable, based on Manning-Strickler	One soil layer, varying depth in dependence on land cover type (0.1 to 4 m)	Müller Schmied et al., 2014,2016; Verzano et al., 2012

1.B Representation of livestock water use

Model name	Model	Drivers	Parameters	References
H08	-	-	-	
LPJmL	-	-	-	
MATSIRO	-	-	-	
PCR-GLOBWB	Time-series regression by individual countries and regions	Cattle stock	Set from literature review and time-series data	Wada et al., 2014b
WaterGAP2	Time-series regression by individual countries and regions	Cattle stock	Set from literature review and time-series data	Flörke et al., 2013

1.C Representation of domestic water use

Model name	Model	Drivers	Parameters	References
H08	Time-series regression by individual countries and regions	Population	Set from literature review and time-series data	Hanasaki et al., 2008a, b; Yoshikawa et al., 2014
LPJmL	-	-	-	
MATSIRO	Time-series regression by individual countries and regions	Population	Set from literature review and time-series data	Takata et al., 2003; Pokhrel et al., 2012; 2015
PCR-GLOBWB	Time-series regression by individual countries and regions	Population, GDP per capita	Set from literature review and time-series data	Wada et al., 2014b
WaterGAP2	Time-series regression by individual countries and regions	Population, GDP per capita	Calibrated from time-series data	Flörke et al., 2013

1.D Representation of industrial water use

Model name	Model	Sector	Drivers	Parameters	References
H08	Time-series regression by individual countries and regions	Industry	Infrastructure area	Set from literature review and time-series data	Hanasaki et al., 2008a,b, Yoshikawa et al., 2014
LPJmL	-	-	-	-	
MATSIRO	Time-series regression by individual countries and regions	Industry	electricity production	Set from literature review and time-series data	Takata et al., 2003; Pokhrel et al., 2012; 2015
PCR-GLOBWB	Time-series regression by individual countries and regions	Industry	GDP, electricity production, energy consumption, household consumption	Set from literature review and time-series data	Wada et al., 2014b
WaterGAP2	Time-series regression by individual countries and regions	Manufacturing	Manufacturing structural water intensity, Manufacturing gross value added (GVA), Technological change rate for manufacturing	Calibrated from time-series data	Flörke et al., 2013
		Electricity production	Annual Electricity Production, Water use intensity of thermal power plants (by type of plant and cooling system), Technological change rate		

1043 **1.E Representation of irrigation water use**

Model name	Calculation of irrigation water requirements	Crop Types	Crop Calendar	Additional components	References
H08	Model-based	MIRCA 2000 (Portmann et al., 2010)	Planting date was determined to obtain maximum yield under meteorological conditions for 1960-1999. The planting date was fixed throughout the simulation period. Harvesting date was calculated in the model and changed with years according to meteorological conditions.	Country/Region specific irrigation efficiency	Hanasaki et al., 2008a,b
LPJmL	Model-based	MIRCA 2000 (Portmann et al., 2010)	n.a. (simulated growing season length)	Country-specific irrigation efficiencies	Bondeau et al., 2007; Rost et al., 2008
MATSIRO	Model-based	MIRCA 2000 (Portmann et al., 2010)	Model-based	Country-specific irrigation efficiencies	Takata et al., 2003; Pokhrel et al., 2012; 2015
PCR-GLOBWB	Model-based	MIRCA 2000 (Portmann et al., 2010) with rice/non-rice distinction	Fixed calendar	Dynamically calculated irrigation efficiency	Wada et al., 2014b
WaterGAP2	Model-based	Rice/Non-rice	Model-based	Country specific irrigation efficiency	Döll and Siebert 2002; Döll et al., 2012, 2014; Portmann et al., 2010; Müller Schmied et al., 2014, 2016

1044

1045 **1.F Water allocation**

Model name	Sources of water included in water use framework	Parameterizations	References
H08	Surface water (including reservoirs)		Hanasaki et al., 2008a,b
LPJmL	River discharge (incl. renewable groundwater), green water in soils		Bondeau et al., 2007; Rost et al., 2008
MATSIRO	Surface water, Groundwater	Pokhrel et al. 2012; 2015	Takata et al., 2003; Pokhrel et al., 2012; 2015
PCR-GLOBWB	Desalinated, Groundwater, Surface water	Wada et al., 2014a,b	Wada et al., 2014b
WaterGAP2	Groundwater, surface water	Sectoral groundwater use fractions	Döll et al., 2012

1046

1047 **1.G Representation of return flows**

Model name	Irrigation	Industry	Domestic	Livestock	References
H08	Soil via infiltration; Groundwater: via additional recharge.	-	-	-	Hanasaki et al., 2008b
LPJmL	River; 50% returns	-	-	-	Bondeau et al., 2007; Rost et al., 2008
MATSIRO	River	River	River	-	Pokhrel et al., 2012; 2015
PCR-GLOBWB	Soil Layers via infiltration; Groundwater layer: via additional recharge. Amount determined by irrigation efficiency	River: same day. Amount determined by recycling ratios	River: same day. Amount determined by recycling ratios	No return flow	Wada et al., 2014b
WaterGAP2	Returned to groundwater, Fraction of irrigation water abstraction (from groundwater or surface water) returning to groundwater $f_{rgi} = 0.95 – 0.75f_{drain}$ (cell-specific artificial drainage fraction).	Returned to surface water. Difference of water abstraction and consumptive use.	Returned to surface water. Difference of water abstraction and consumptive use.	No return flow, as only consumptive use.	Döll et al., 2012, 2014b

1048

1049 **1.H Representation of Reservoirs (after Pokhrel et al., 2016)**

Model name	Purposes included	Parameterizations	References
H08	Non-irrigation	Hanasaki et al., 2006	Hanasaki et al., 2006
LPJmL	Irrigation/Non-irrigation	Biemans et al., 2011	Biemans et al., 2011
MATSIRO	Irrigation/Non-irrigation	Hanasaki et al., 2006	Pokhrel et al., 2012
PCR-GLOBWB	Water supply/Flood control/Hydropower/Navigation	Haddeland et al., 2006; Adam et al., 2007	van Beek et al., 2011
WaterGAP2	Irrigation/Non-irrigation	Hanasaki et al., 2006	Döll et al., 2009

1050

1051 **1.I Calibration of runoff and discharge**

Model name		References
H08	Bias in runoff corrected with modification of two parameters of subsurface flow for four climatic groups: daily maximum subsurface runoff, relation between subsurface runoff and soil moisture.	Hanasaki et al., 2008a,b
LPJmL	-	

MATSIRO	-	
PCR-GLOBWB	-	
WaterGAP2	Biases in long-term mean annual discharge corrected by adjusting exponent of function where part of effective precipitation becomes runoff from land depending on exponent of ratio of actual to maximum soil storage.	Müller Schmied et al., 2014

1052

Supplementary table 2: Abbreviations

Abbreviation	Full name
GRDC	Global Runoff Data Centre
KGE	Modified Kling-Gupta Efficiency
rKGE	KGE correlation coefficient (Pearson)
βKGE	KGE bias ratio
γKGE	KGE variability ratio
Q_{99}	Low-flow indicator
Q_1	High-flow indicator
NSE	Nash-Sutcliffe Model Efficiency
KS	Kolmogorov-Smirnov test
NOHIP conditions	'naturalized' model run without human influence
HIP conditions	Model run including time-varying human influence
DDM30	Drainage direction map
HIP	Human impact parameterizations
n	Number
P	Probability
s_i	Simulated discharge
o_i	Observed discharge
i	Station
μ_s	Simulated mean monthly discharge
μ_o	Observed mean monthly discharge
σ_s	Standard deviation of simulated discharge
σ_o	Standard deviation of observed discharge
Q_s	Simulated hydrological extreme
Q_o	Observed hydrological extreme
GVA	Gross value added

Supplementary results for selected river basins

Twelve GRDC stations (**supplementary table 3**) were selected to serve as illustrative focus cases in this study. These are: the Amazonas, Amur, Colorado, Congo, Guadiana, Mackenzie, Murray, Ob, Rhine, Tocantins, Volga, and the Zambezi. The stations are selected for their heterogeneous spatial distribution and belong to both the managed and near-natural catchment groups. Moreover, station-specific results for all stations are provided as a **separate supplement**.

Supplementary table 3: Characteristics of focus stations.

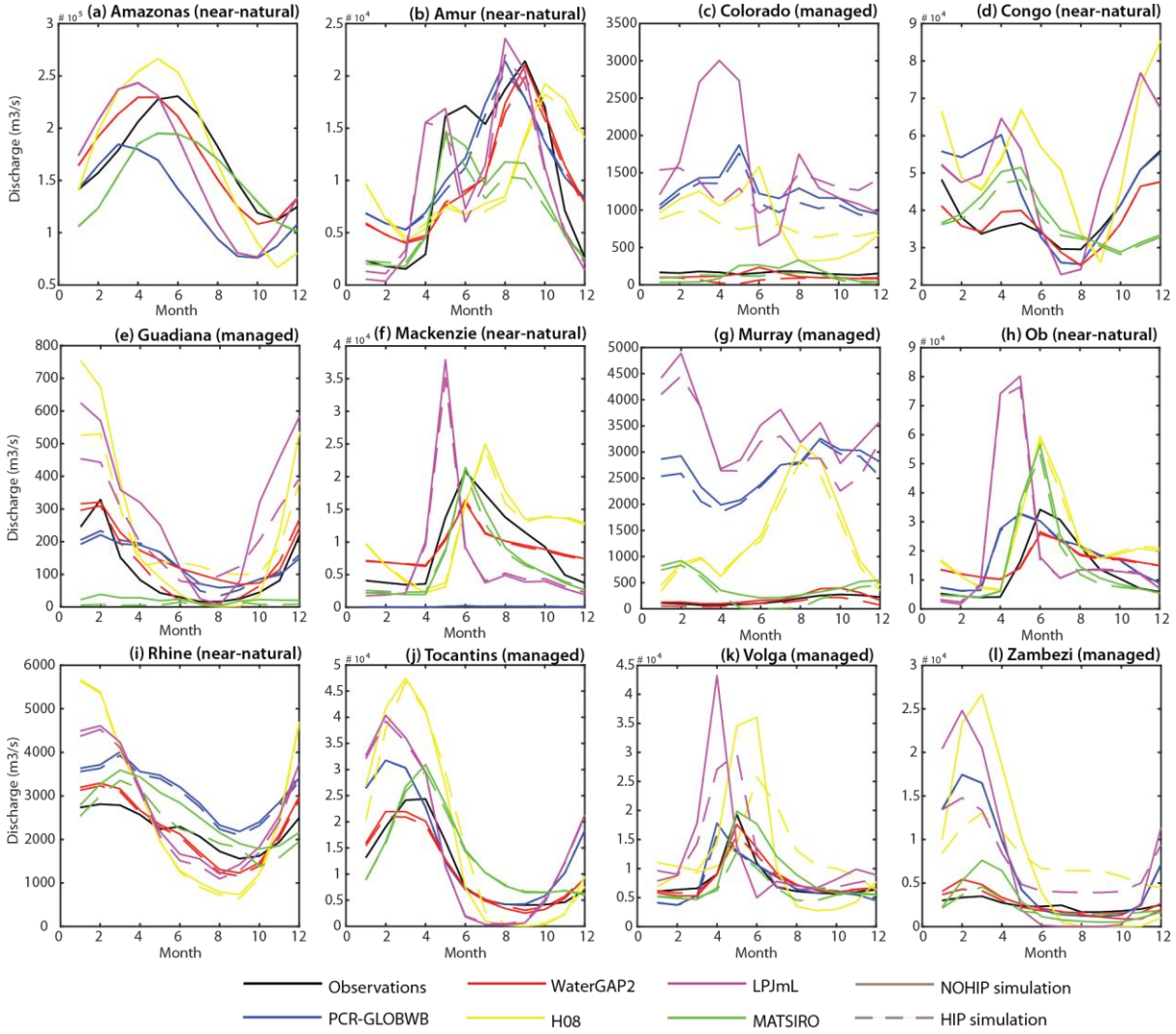
River (station name)	GRDC number	GRDC Catchment area (km^2)	Mean discharge (m^3/sec)	Start- end-year	data-points coverage (%)	Managed/ near-natural
Amazonas (Obidos-Porto)	3629000	4,680,000	170,356	1971-1998	97.9%	near-natural
Amur (Bogorodskoy)	2906901	1,790,000	10,386	1971-1987	99.5%	near-natural
Colorado (Lindero Internacional Norte)	4352100	631,960	160	1971-1995	92%*	managed
Congo (Kinshasa)	1147010	3,475,000	39,127	1971-2010	100%	near-natural
Guadiana (Pulo do Lobo)	6116200	60,883	107	1971-1990	98.8%	managed
Mackenzie (Arctic Red River)	4208025	1,660,000	9,225	1972-2010	98.1%	near-natural
Murray (Overland Corner)	5404270	1,000,001	164	1985-2010	98.1	managed
Ob (Salekhard)	2912600	2,949,998	13,132	1971-2010	97.5%	near-natural
Rhine (Lobith)	6435060	160,800	2,235	1971-2010	100%	near-natural
Tocantins (Tucurui)	3649950	742,300	11,186	1978-2010	94.95%*	managed
Volga (Volgograd Power Plant)	6977100	1,360,000	7,885	1971-2010	97.5%	managed
Zambezi (Matundo-Cais)	1891500	940,000	2,455	1971-1990	96.67%	managed

*stations do not comply to minimum requirement of data-point coverage in time-series (95%) set for full analysis but were included for their representative spatial distribution and representative human/(near-)natural characteristics.

1.Impact of HIP in overall hydrological performance

The change in hydrological performance for 12 managed and near-natural focus stations is shown in **supplementary figure 6**. Little to no change in performance (as calculated with the KGE over the full time-series) is found before and after HIP for the Amazonas, Congo, Mackenzie, Ob, Rhine, and the Tocantins. In contrast, HIP significantly influences the hydrographs of the Amur, Colorado, Guadiana, Murray, Volga, and the Zambezi river. Monthly discharge diminishes significantly for the latter catchments, either in an equal way throughout all months of the year (Murray, Amur) or predominantly during the peak-flow months (Colorado, Guadiana, Volga, Zambezi). For the latter four catchments we observe in most models a significant redistribution of monthly discharge between

the peak-flow and low-flow months due to HIP. These cases highlight the influence of dams and reservoirs (and their operation), being part of the HIP framework, on the volume and timing of estimated river discharge. The ability to represent the general hydrologic characteristics varies significantly from catchment to catchment and across models (**supplementary figure 6, supplementary table 4**). Results show that some catchments (Amazonas, Amur, Rhine, or the Tocantins) are generally better represented, whereas the models have relatively more difficulties representing the hydrologic characteristics of others (Colorado, Murray). On the other hand, different models outperform others looking at the various catchments. An overall best performing model does not exist, with WaterGAP2 producing the best results for the Colorado, Congo, Murray, Tocantins, Volga, and Zambezi, compared to LPJmL for the Amur, or MATSIRO for the Amazonas and Ob (**supplementary table 4**).



Supplementary figure 6: Hydrographs for the focus stations under HIP and NOHIP. Each panel shows the long-term mean monthly discharge for each model. Dashed lines indicate the long-term mean monthly discharges with HIP whilst the solid lines indicate the performance under NOHIP conditions.

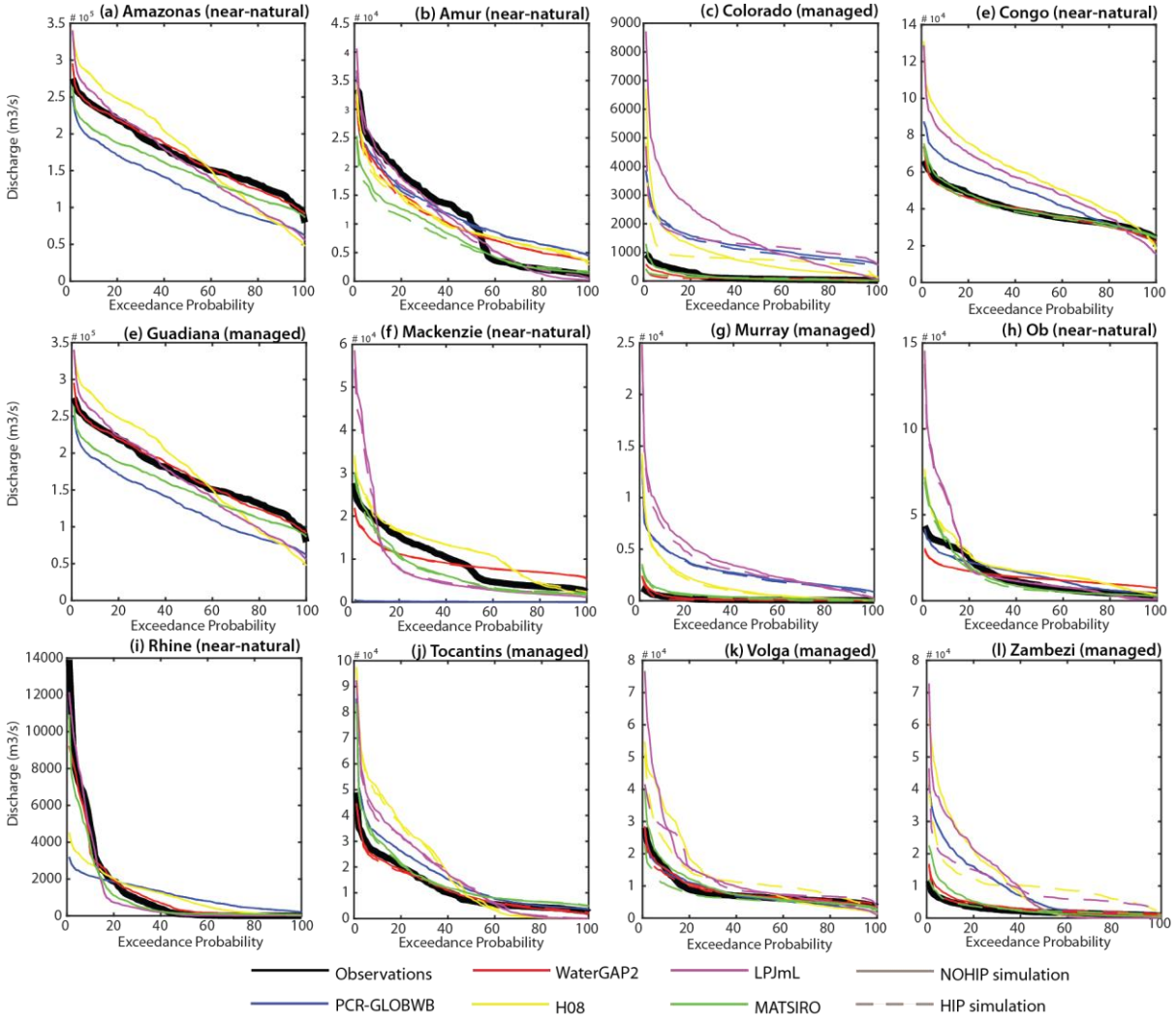
Supplementary table 4: Overall hydrological performance (KGE) for different models for the focus stations.

Performance values with inclusion of human impact parameterizations (HIP) are presented first, performance values without HIP are presented after the slash. Bold values show the best performing models per station. Underscored values show the best performing station per model.

	Amazonas	**Amur**	**Colorado**	**Congo**	**Guadiana**	**Mackenzie**	**Murray**	**Ob**	**Rhine**	**Tocantins**	**Volga**	**Zambezi**
PCR-GLOBWB	0.45/ 0.45	0.55/ 0.54	-4.21/ -4.68	0.36/ 0.36	0.32/**0.29**	**0.70**/ **0.70**	-13.19/ -14.15	0.51/ 0.50	0.54/0.50	0.58/ 0.58	0.63/ 0.58	-0.92/ -0.92
WaterGAP2	0.81/ 0.82	0.67/ 0.67	**0.13**/ **0.23**	**0.66**/ **0.66**	**0.49**/0.18	0.48/ 0.47	**-0.11**/ **0.64**	0.34/ 0.33	**0.78**/**0.81**	**0.88**/ **0.91**	**0.73**/ **0.85**	**0.59**/ **0.57**
H08	0.41/ 0.41	0.26/ 0.28	-3.36/ -3.88	0.24/ 0.23	-0.32/-0.41	0.48/ 0.49	-5.51/ -6.42	0.23/ 0.23	-0.20/-0.14	0.27/ 0.23	0.15/ 0.04	-1.15/ -1.43
LPJmL	0.41/ 0.40	**0.71**/ **0.73**	-7.13/ -8.56	0.23/ 0.23	-0.50/-1.07	-0.22/ -0.38	-16.96/ -19.44	-0.31/ -0.34	0.34/0.36	0.41/ 0.36	0.36/ -0.31	-1.10/ -1.37
MATSIRO	**0.84**/ **0.84**	0.64/ 0.70	-0.21/ 0.14	-0.07/ -0.11	-0.16/0.09	0.60/ 0.55	-0.33/ -0.95	**0.57**/ **0.58**	0.76/0.70	0.71/ 0.69	0.69/ 0.72	0.52/ 0.23

2 Impact of HIP in representation of hydrological extremes

Supplementary figure 7 shows for the focus stations the modelled probability exceedance curves for the different models for both NOHIP and HIP conditions, with as reference the probability exceedance curves based on observational data. Substantial changes in modelled discharge estimates that are the result of HIP can be found under various exceedance probability levels for the Amur, Colorado, Murray, Volga, and Zambezi. For some stations, the most substantial changes are being found at the tails of the exceedance probability curve (at high- and/or low-flows), e.g. the Zambezi or Volga. Other stations show changes for predominantly intermediate exceedance probability levels, e.g. the Amur or Murray. While all models tend to generally overestimate both high- and low-flows in the Murray, Colorado, Tocantins and Zambezi; the results are more varied for the Amazonas, Amur, Guadiana, Mackenzie, and the Ob. For the Congo and Volga most models overestimate high-flows whereas they underestimate low-flow discharges, even with HIP. For the Rhine, this is the other way around, with mostly underestimations for high-flow discharges but overestimations when moving towards low-flows. Both the Amur and Mackenzie see a sudden drop in observed discharges at intermediate levels of exceedance probability. This is likely caused by the strong seasonal characteristics that play an important role in the generation of discharge in these rivers, with a substantial period of the year being snow-dominated. None of the models seems to be able to reflect these particular hydrologic characteristics correctly, both without and with HIP.



Supplementary figure 7: Exceedance probability curves for the focus stations.
Each panel shows the modelled discharge at various levels of exceedance probability with HIP (dashed line) and without HIP (solid line). The thick black lines show the observed discharge at various levels of exceedance probability.