DOI: https://doi.org/10.1016/j.envsoft.2019.01.003

Optimising runoff simulations from an ensemble of global-scale hydrological models through multi-model combination weighting

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

Rest of co-authors in alphabetic order:

Dieter Gerten [b, c], Hannes Müller Schmied [d, e], Qiuhong Tang [f], Rutger Dankers [g], Stephanie Eisner [h], Xingcai Liu [f], Yoshihide Wada [i], Yoshimitsu Masaki [j]

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

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

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

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

[e] Senckenberg Biodiversity and Climate Research Centre (BiK-F), Frankfurt, 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] Met Office, FitzRoy Road, Exeter, EX1 3PB, United Kingdom,

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

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

[j] Hirosaki University, Hirosaki, Japan,

*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

Due to uncertainties associated with ensembles of global hydrological models (GHMs), their arithmetic mean (ensemble mean, EM) is normally used to report the projected hydrological impacts of different climate scenarios. This study presents a novel application of machine learning to deliver a more 'intelligent' multi-model combination (MMC) method that employs an evolutionary algorithm and symbolic regression to optimise how individual model outputs are combined. We exemplify the approach using runoff simulations from five GHMs. MMC solutions are developed for forty large global catchments and assessed against observed data. The median performance gain of the MMC solutions is 45% over the best performing GHM and exceeds 100% when compared to the EM. In light of the significantly improved performance offered by MMC, we recommend that future multi-model applications consider reporting MMCs, alongside the EM and intermodal range, to provide end-users of GHM ensembles with a better informed estimate of runoff.

Keywords:

Machine Learning; Gene Expression Programming; Global Hydrological Model; Optimization; Ensemble

Highlights:

- First use of multi-model combination (MMC) in a global hydrological model ensemble
- Model, ensemble mean and MMC performance assessed with a novel integrated metric
- MMC performs better than any individual model from the ensemble, overall
- MMC also performs better than the ensemble mean, overall

Conflict of interest

None

1. Introduction

Global Hydrological Models (GHMs) is a collective term that describes 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 [1]. GHMs include stand-alone global hydrology models [2, 3] as well as land surface models [4, 5] and dynamic vegetation models [6, 7] featuring the global water cycle among other processes. Considerable diversity exists amongst the numerous GHMs that have been developed, meaning that the "jungle" of models predicted by Kundzewicz [8], more than three decades ago, is now a reality. This multiplicity and diversity of different GHMs raises questions about whether one, or a group of GHMs, might be better suited than others to application in different regions of the world.

In practice, aleatory (random) and epistemic (resulting from deficits of knowledge) uncertainties mean that it is inappropriate to assume a single GHM can provide an adequate basis for a simulation [9], even though the acquisition of higher quality input data or efforts to improve individual model structures can enhance a model's ability to replicate observed data [10]. Instead, there is increasing recognition of the opportunities provided by combining outputs from a multitude of diverse models – a process known as multi-model combination (MMC). In practice, this is usually achieved by reporting the mean output of an ensemble of GHMs [11-16]; known as the ensemble mean (EM), and also sometimes the ensemble median [17, 18]. However, 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 [19-21]. Alternatively, the idea of combining an ensemble of models 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, represents a novel approach that has the potential to both enhance model outputs *and* gain valuable heuristic insights into the relative capabilities of different models under different contextual demands [9, 22]. The fact that this approach has yet to be adopted in GHM studies means that there is a unique opportunity to evaluate whether the optimised combination of multiple GHM outputs can provide both improved simulations and heuristic insight; which could be particularly valuable for climate change impact assessment applications that project the impacts of climate change on global hydrology. This is the objective of this paper.

To achieve this objective, we adopt a novel MMC approach by employing Gene Expression Programming (GEP) [23, 24]– a machine learning algorithm that has demonstrable success in delivering optimal solutions for hydrological [25, 26] and climatological problems [22, 26-30]. MMC encompasses a plethora of techniques designed to integrate the outputs of multiple models into a single integrated solution. It was first applied in economics but has been used in various forms across multiple disciplines including hydrological modelling [9, 13, 27, 31-50], but never at global scales. The 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 [11-15].

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 [9]. To overcome this problem, the development of MMC solutions based upon the identification and application of variable weights has been explored (see [49] for various weighting methods employed in hydrological modelling, and [51-54] for regional climate 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 (e.g. [6-9] and [17-19]), but the fundamental challenge remains that of finding computational methods that can optimise both the weighting strategy and the specific weights in a meaningful way.

To these ends, 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 [55] 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. They 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 [22].

Whilst the numerical solutions generated by some MLAs (e.g. artificial neural networks [9, 34, 56]) are black box and implicit [22], others deliver explicit equations. Being explicit enables 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 this way, the relative merits of different MMC solutions in different settings can be revealed in a more nuanced manner than is possible through the comparison of fit metrics alone.

In this paper, we use the Gene Expression Programming MLA to develop an optimised combination strategy for the runoff simulations of an ensemble of five global-scale hydrological models, spanning 40 large catchments across the globe (Figure 1). The model outputs are provided by teams engaged in the second phase of the Inter-sectoral Impact Model Inter-comparison Project (ISIMIP2a) [57] – one of the largest MIPs [1]. The catchments that are simulated are distributed across all of the world's hydrobelt types [58], thus promoting a robust spatial analysis. We assess the performance of our optimised model combination strategies by comparing our MMC outputs against those of the individual models and the EM respectively. Multiple characteristics of fit performance are assessed simultaneously through the use of the recently-devised, integrated Ideal Point Error metric (IPE) [59], alongside standard analyses of observed-vs-simulated plots. We also examine the spatial performance of our MMC strategies across different hydro-climatological regimes and disaggregate the MMC equations into their component parts. We suggest that this may help facilitate a new route towards diagnostic appraisal of the relative performance of different GHMs in different hydro-climatic settings.

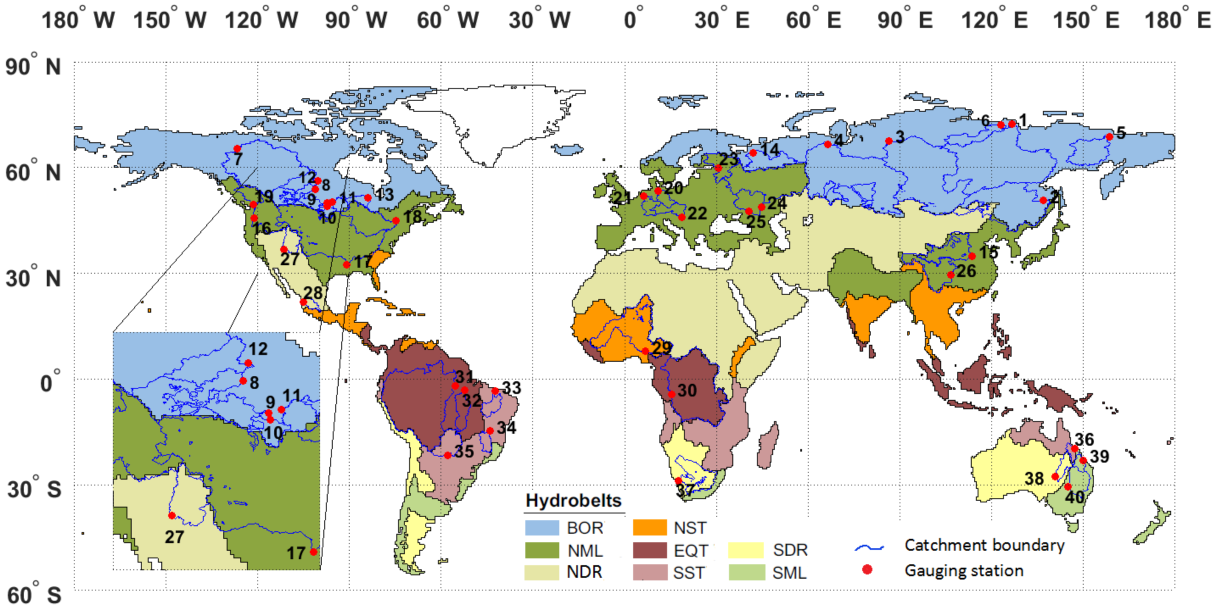

Figure 1. Locations of the 40 selected catchments (details in Table 1 and Table S1) across the hydrobelt system defined in [58]. BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial belts and SML, SDR, SST their southern analogues.

2. Methods

2.1 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 applied at different data ranges because of non-stationarities in the data. To these ends, GEP offers particular potential. GEP is a highly-adaptive, symbolic regression (SR) algorithm that searches for and learns optimal regression models 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 [22]. 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. It applies Darwinian principles to ‘evolve’ a set of optimised model solutions (in the form of symbolic regression 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 [23, 24]. GEP models are encoded in one or more linear structures (known as ‘chromosomes’) of fixed length where each structural element encodes a user-defined input, a function, a numerical constant or an arithmetic operator [47]. These are the building blocks of the SR model that GEP evolves. Chromosomes can be expressed as parse trees and combined (usually by addition) to deliver a final SR solution (often referred to as a program). Chromosome elements can be switched on and off; enabling SR solutions of different size and complexity 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 entirely from its solutions.

In the GEP learning process (Figure 2), 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 their ability to deliver a predefined objective function – usually a minimisation of the statistical fit between the model and a set of observed training data [27]. 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 [23] resulting in the evolution of a set that comprises a new generation of potential solutions. The best performing solution is then added back into this set and each solution is assessed for fitness before the best performing solution is isolated and the remaining solutions undergo adjustment once again. The iterations continue until a stopping point (usually a pre-defined number of iterations) is met and the set of candidate solutions is output.

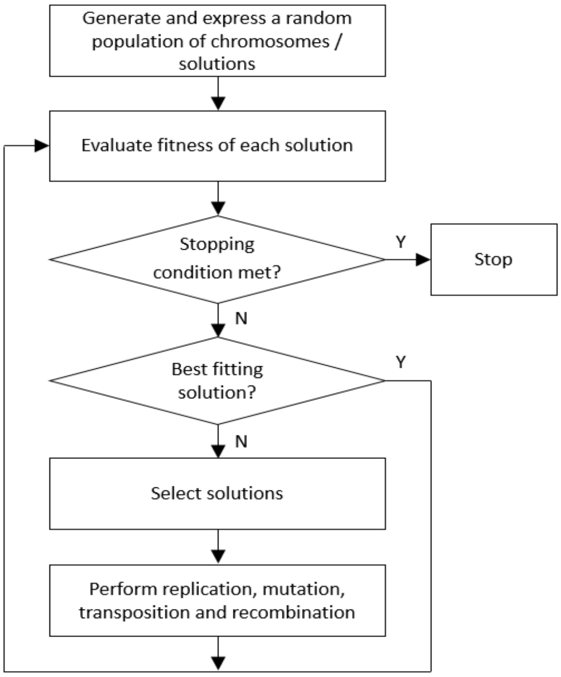
Generate and express a random population of chromosomes / solutions
Evaluate fitness of each solution
Stopping condition met?
Y
Stop
N
Best fitting solution?
Y
N
Select solutions
Perform replication, mutation, transposition and recombination

Figure 2: The GEP algorithm flowchart.

One important advantage of GEP over other MLAs is the way it structures and expresses it solutions to a modelling problem as component genes (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 to 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.

In this study, GEP was implemented using GeneXpro Tools 4.0 (GXPT4) (http://www.gepsoft.com/). The user-defined function set, the allowable constants and the number of genes (Table S2) were chosen to maximise the parsimony of the symbolic regression solutions that GEP delivered, whilst supporting a limited set of non-linear functions, so that the potential for meaningful interpretation of the regression equations was maximised.

2.2. Selecting the GEP-based MMC solution

The candidate solutions generated by GEP vary in terms of their complexity and level of fit. As a general rule, the higher the number of iterations completed by the GEP algorithm during training, the more complex the candidate solutions become. Similarly, the more complex the solution (incorporating larger chromosomes and thus having a larger equation size), the greater the degree of fit between modelled and observed data. However, the more complex the solution the greater the chance that it is overfitted and the harder it is to interpret the SR 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 so that the final solution has both a good degree of fit and is parsimonious with respect to its numerical complexity so that overfitting is prevented.

In the absence of a generally accepted method for doing this [49, 51, 52], we devised a simple trade-off between solution fitness and equation size (a proxy for model complexity evaluated according to the number of inputs, constants, operators and functions in an equation) to identify a final solution from a candidate set produced from a training sequence of 100,000 iterations (Figure 3). Firstly, the fitness and equation size of each candidate solution was normalised to a 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.

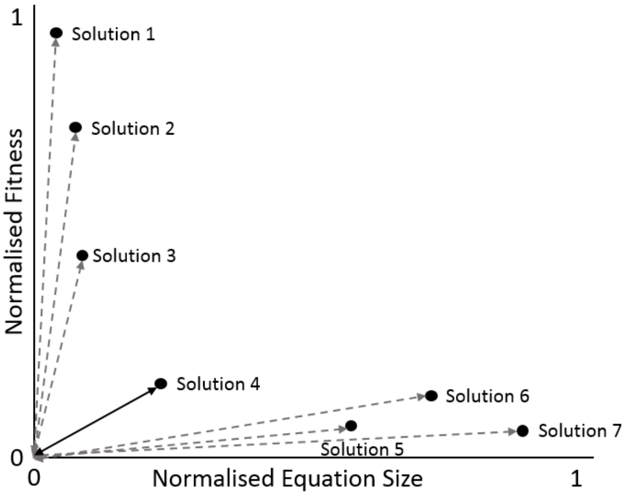

Figure 3. Selecting the GEP solution from a normalised fitness-equation space. Solution 4 is selected because it has the smallest Euclidean distance from the origin.

2.3. GEP validation metrics

It is standard practice to evaluate the performance of models based on summary metrics that are designed to quantify the degree of fit of each model to observed data. However, this approach is problematic when performance comparisons across multiple catchments of different scales and variable hydro-climatic regimes are sought. This is because it is difficult to assess relative model performance in a meaningful way without the adoption of a transferrable benchmark against which model performance in different catchments can be consistently compared and understood. In addition, different metrics are more or less suited to assessing individual characteristics of a model's fit. For models developed using GEP this is a significant issue because the objective function that directs the learning process is usually a minimisation of one specific fit metric and this can bias the solutions that are generated to an individual fit characteristic if a diversity of metrics is not somehow incorporated into the objective function [59, 60].

To overcome these two challenges, we adopted an integrated performance metric called the Ideal Point Error (IPE) [59]. This was used to inform the learning process during the development of GEP solutions (i.e. the minimisation of IPE is the objective function used to train the GEP algorithm). It is also used to assess the relative ability of individual GHMs, the EM and the MMC solutions to replicate out-of-sample, observed data. IPE expresses the ratio of performance gain / loss of a model compared to a model benchmark (in our case the naïve t-1 model, where runoff in time *t* is predicted by runoff in *t-1* following [61]); with the ratio assessed using a suite of metrics that are integrated into the single IPE value. The basic IPE equation is presented in (1) and is adapted from the original formula in [59]. 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. 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) as well as their orthogonality. 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.

IPE scores can range between minus one and minus infinity (performance gain over benchmark model) and one and infinity (performance loss over benchmark model). Where a model exactly equals the performance of the benchmark against which it is assessed, its IPE score will be one. 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.

$$\varepsilon = \left\{[0.333 * ((\mathrm{RMSE_m}/\mathrm{RMSE_b})^2 + (\mathrm{MARE_m}/\mathrm{MARE_b})^2 + ((\mathrm{CE_m} - 1)/(\mathrm{CE_b} - 1))^2)]^{\frac{1}{2}}\right\}$$

$$\mathrm{IPE} = \varepsilon \qquad \mathrm{IF}\ \varepsilon > 1 \tag{1}$$

$$\mathrm{IPE} = -1/\varepsilon \qquad \mathrm{IF}\ \varepsilon < 1 \tag{2}$$

Where:

RMSE = root mean squared error

MARE = mean absolute relative error

CE = Coefficient of Efficiency

m = modelled data

b = benchmark data from the naïve (t-1) model)

The IPE performance gain (PG) of one model output (i.e. model *A*) relative to another (i.e. model *B*) can be expressed in percentage terms. The way that this is computed is dependent on the respective signs of the IPE scores for the models being compared (equations 3-5). PG values are 0% where there is no difference in the performance gain / loss relative to the benchmark delivered by model *A* over model *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 model *A* than model *B*. Similarly, a PG value of 120% would indicate that the 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:

$$\mathrm{PG}\ = 0 - (\mathrm{IPE}_{\mathrm{modelA}} - \mathrm{IPE}_{\mathrm{modelB}}) \times 100 \tag{3}$$

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

$$\mathrm{PG} = 0 - \big((\mathrm{IPE}_{\mathrm{modelA}} - 1) - (\mathrm{IPE}_{\mathrm{modelB}} + 1)\big) \times 100 \tag{4}$$

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

$$\mathrm{MMC}_{\mathrm{PG}} = 0 - \big((\mathrm{IPE}_{\mathrm{modelA}} + 1) - (\mathrm{IPE}_{\mathrm{modelB}} - 1)\big) \times 100 \tag{5}$$

3. Data Sets and MMC Development

3.1. Study catchments and observed data

Catchments were selected for inclusion in the study according to two criteria designed to ensure the spatial resolution of the GHMs (0.5° x 0.5°) was accommodated and the availability of observed data series of sufficient length to support robust GEP training, testing and out-of-sample validation procedures:

1) Catchment size >100,000 km^2 (conforming to the World Meteorological Organisation's definition of 'major' catchments [62]).
2) Availability of an observed mean monthly runoff record with length > 25 years between 1971 and 2010.

This resulted in the selection of 40 catchments for use in the study. Observed runoff data were derived from river discharge observations held in the Global Runoff Data Centre (GRDC) reference database (http://grdc.bafg.de). 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.

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	2460000	BOR
2	2906900	AMUR	KOMSOMOLSK	26	1730000	BOR
3	2909150	YENISEI	IGARKA	32	2440000	BOR
4	2912600	OB	SALEKHARD	39	2949998	BOR
5	2998510	KOLYMA	KOLYMSKAYA	28	526000	BOR
6	2999910	OLENEK	7.5KM DOWNSTREAM OF MOUTH OF RIVER PUR	39	198000	BOR
7	4208150	MACKENZIE RIVER	NORMAN WELLS	30	1570000	BOR
8	4213550	SASKATCHEWAN	THE PAS	40	347000	BOR
9	4213650	ASSINIBOINE	HEADINGLEY	40	153000	BOR
10	4213680	RED RIVER	EMERSON	40	104000	BOR
11	4213800	WINNIPEG RIVER	SLAVE FALLS	38	126000	BOR
12	4214260	CHURCHILL RIVER	ABOVE GRANVILLE FALLS	36	228000	BOR
13	4214520	ALBANY RIVER	NEAR HAT ISLAND	31	118000	BOR
14	6970250	NORTHERN DVINA	UST-PINEGA	31	348000	BOR
15	2180800	YELLOW	HUAYUANKOU	40	730036	NML
16	4115200	COLUMBIA	THE DALLES, OREG.	40	613830	NML
17	4127800	MISSISSIPPI	VICKSBURG, MISS.	37	2964252	NML
18	4143550	ST.LAWRENCE	CORNWALL(ONTARIO), NEAR MASSENA, N.Y.	40	773892	NML
19	4207900	FRASER RIVER	HOPE	40	217000	NML
20	6340110	LABE	NEU-DARCHAU	40	131950	NML
21	6435060	RHINE RIVER	LOBITH	40	160800	NML
22	6442600	DANUBE	MOHACS	29	209064	NML
23	6972430	NEVA	NOVOSARATOVKA	40	281000	NML
24	6977100	VOLGA	VOLGOGRAD POWER PLANT	39	1360000	NML
25	6978250	DON	RAZDORSKAYA	38	378000	NML
26*	7222222	YANGTZE	Cuntan	31	121000	NML
27	4152450	COLORADO	LEES FERRY, ARIZ.	40	289562	NDR
28	4356100	SANTIAGO	EL CAPOMAL	31	128943	NDR
29	1834101	NIGER	LOKOJA	25	2074171	NST
30	1147010	ZAIRE	KINSHASA	40	3475000	EQT
31	3629000	AMAZONAS	OBIDOS	27	4640300	EQT
32	3630050	XINGU	ALTAMIRA	35	446570	EQT
33	3650481	RIO PARNAIBA	LUZILANDIA	26	322823	SST
34	3651805	SAO FRANCISCO	MANGA	37	200789	SST
35	3667060	PARAGUAI	PORTO MURTINHO (FB/DNOS)	37	474500	SST
36	5101200	BURDEKIN	CLARE	40	129660	SST
37	1159100	ORANJE	VIOOLSDRIF	38	850530	SDR
38	5410100	COOPER CREEK	CALLAMURRA	33	230000	SDR
39	5101301	FITZROY	THE GAP	40	135860	SML
40	5204250	DARLING RIVER	LOUTH	26	489300	SML

*not included in GRDC database, obtained from local authorities (the code invented by the authors)

3.2. Input models to the MMC

This study has been made possible by ongoing efforts to inter-compare GHMs through ISIMIP [54]. ISIMIP modelling groups use 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). 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.

We herein capitalise on the opportunity provided by the latest simulations from the water sector of the current phase of ISIMIP (ISIMIP2a). Our MMC integrates simulation outputs from an ensemble of five GHMs: DBH, H08, LPJmL, PCR-GLOBWB (hereafter called PCRGLOBWB in the main text in order to avoid confusion by '-' in MMC equations) and WaterGAP2 (Table 2), which were selected because they have made available simulated catchment runoff using a protocol that accounts for time-varying human management such as water usage, water withdrawals, and dams operation (referred to as "varsoc" in the ISIMIP2a protocol). 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/). In all cases, the discharge 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 [63]. It is important to note that, of the five models, only WaterGAP2 has been calibrated against long-term mean annual runoff for a selection of catchments [64] for the ISIMIP2a simulations. The inclusion of calibrated WaterGAP2 helps to highlight the benefits (or otherwise) of calibrating global scale models – an activity that remains relatively uncommon with GHMs compared with catchment-scale hydrological models [65].

Table 2. The five 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 [66]	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 [3]	GHM	tas, Pr, Snowfall, wind, rhs, rsds, rlds, ps	0.5º; 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 [67]	GHM	Pr, tas, rsds	0.5º; 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 [68]	GHM	tas, Pr, pet	0.5º; 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 [64]	GHM	tas, Pr, rsds, rlds	0.5º; 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

3.3. Data splitting for MMC development and testing

In line with standard practices [69-71], each of the candidate solutions produced was evaluated against an 'in sample' training/testing data subset (effectively a calibration subset) and an independent, 'out-of-sample' validation data subset of the observed data. Consequently, the nature of a GEP-based MMC solution will be inexorably linked to the statistical characteristics of the in-sample, training/testing data upon which it was trained and its reported validity will be dependent on the statistical characteristics of the out-of-sample validation data subsets. It is, therefore, important to ensure that data subsets (especially training and testing) are representative of the observed data set and of each other.

Arbitrary data splitting approaches (e.g. taking the first 50% of an observed record for training/testing and second for validation) cannot be guaranteed to achieve this. Therefore, a range of splitting methods have been developed [71-73] that are based on variations of cluster-based sampling or data proximity considerations. Tests of the effectiveness of alternative splitting techniques [71] have shown the DUPLEX method [72] to be particularly well suited to delivering representative data splits for use in model development by MLAs. 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 [71]. We were consistent across all 40 catchments in the size of the training/testing data, which comprised 144 months (12 years) and 96 months (8 years) for the training and testing data subsets respectively, 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. However, it was never less than 60 months (5 years) and extended up to 240 months (20 years) in some catchments (Table 3).

Table 3. The training, testing and validation subsets used to inform MMC development in each of the study catchments.

No	River	Total data length (months)	Period of Training and Testing 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

4. Model Performance

In the following section we summarise the performance of individual GHMs and the EM, and quantify the performance gain achieved over them by MMC solutions. We pay particular attention to differences in performance gain across hydrobelts to explore spatial homogeneity (or otherwise) in any performance gains achieved by MMC. Detailed results are provided on a catchment-by-catchment basis in the Supplementary Information. The performance metrics for all models are detailed for both calibration and validation data subsets (Table S3). In addition, observed versus simulated plots for mean annual runoff and the exceedance probability curves are provided for all models. Plots for each SR equation component (i.e. the output of the gene expressed by each GEP chromosome) are presented in Section S2.

4.1. GHM performance

To assess the performance of the different models, the fit of the monthly simulated and observed runoff time series was computed against the validation data for each GHM 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 4 and the spatial distribution of the best individual GHM and the best overall model is mapped in Figure 4. This reveals that WaterGAP2 is the GHM most able to improve upon the naïve model benchmark. It outperforms other GHMs in 32 catchments, and also performs better than the EM for the majority of catchments. 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; Darling) the IPE scores of one or more GHMs exceeds 10, indicating a failure to deliver a performance anywhere close to that of the simpler model benchmark. In the ephemeral catchments of Cooper Creek (No. 38) and Fitzroy (No. 39) 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 because a high ratio between the mean absolute relative error of the individual GHMs ($MARE_m$) and those of the naïve model benchmark ($MARE_b$) 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.

Table 4. 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**

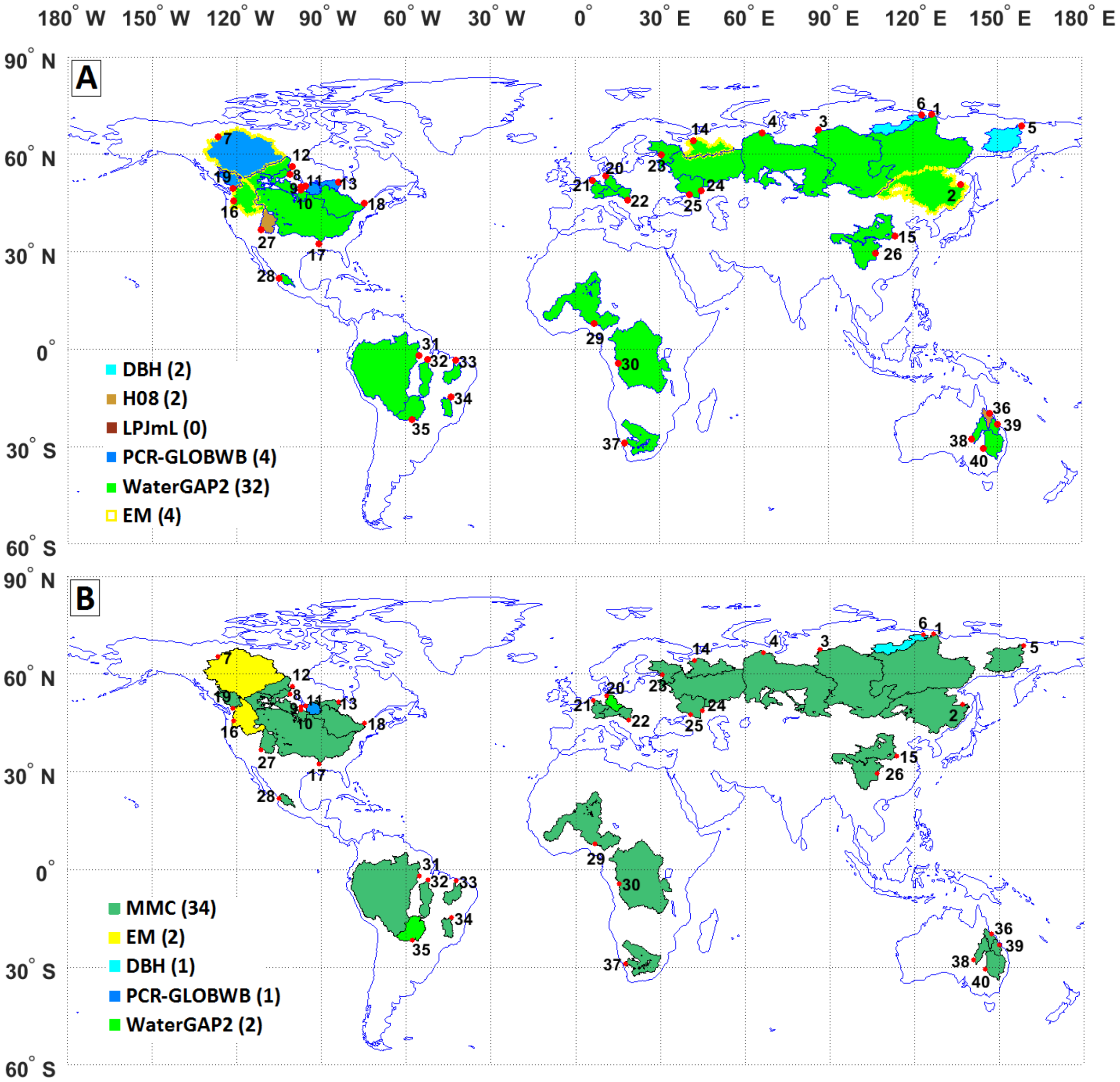

Figure 4. The best performing individual GHM (A); four catchments where the EM outperforms the individual models have borders in yellow. The best performing overall model/MMC (B); the two catchments where the EM is the best are shaded in yellow. Numbers in parentheses denote number of catchments where each model performs best.

4.2. EM Performance

Table 4 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 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 5).

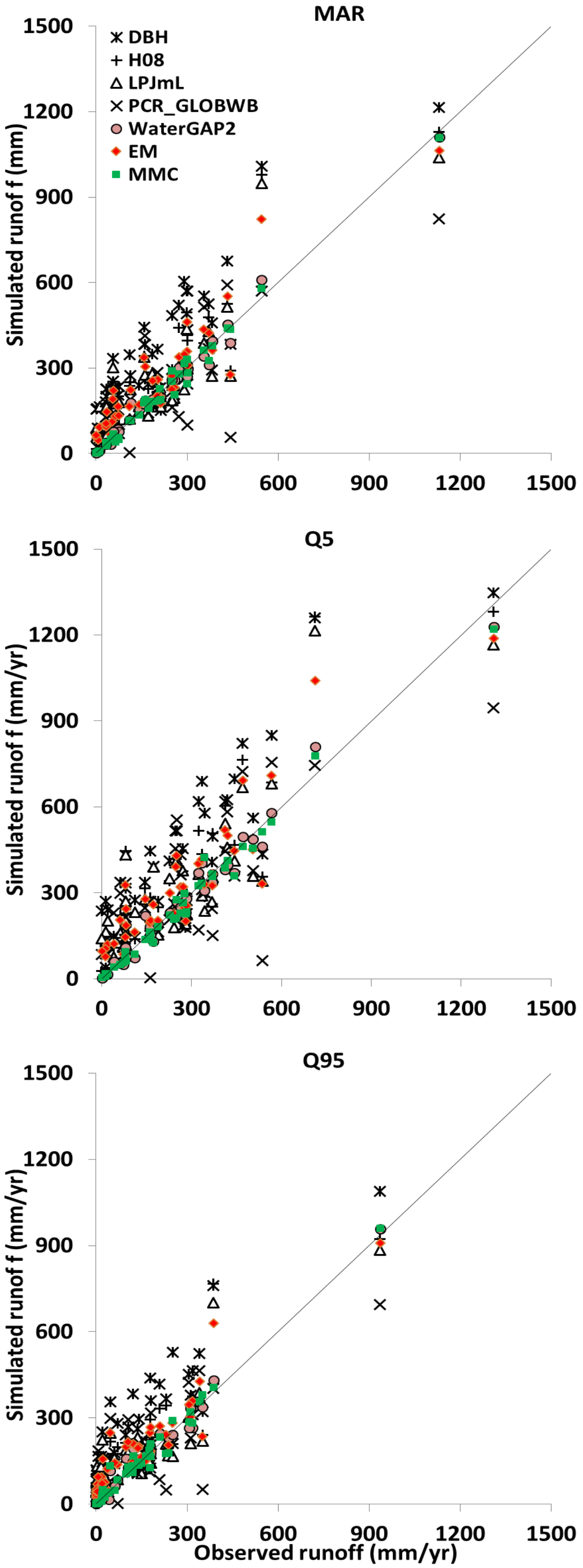

Figure 5. Plots of observed versus simulated runoff for each GHM, the EM and the MMC for mean annual runoff, Q5 and Q95.

This positive biases amongst the GHMs from which the EM is calculated, precludes a better performance by the EM relative to the best performing individual model, calibrated WaterGAP2. This is because the frequent overestimations of runoff by the other non-calibrated models prevents any beneficial compensation of under- and over-estimation of runoff required to generate a gain in EM performance. 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 4): 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 benefits 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 individual GHMs, EM and MMC for the validation period in each catchment are presented in Table 4. The MMC solutions, and their symbolic regression equations for each catchment are detailed in Table 5 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 5) 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 45.80% across all 40 catchments. If the outliers of Cooper Creek, Darling and Fitzroy River are omitted, the median MMC_{PG} is 39.88% 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 15.01% greater than for the MMC on average. Similarly, in 2 catchments (Mackenzie and Columbia) the EM delivers performance gains over the MMC equal to 5.42% and 6.53% respectively. 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 -29.24% and -217.2% respectively. This is considerably smaller than their northern hemisphere equivalents; -41.28% and -118.91%.

When summarised by hydrobelt (Table 6), 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 bias in GHM outputs has 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 (-23.97%) and highest in southern sub-tropical (-254.24%) and the boreal (-51.35%) hydrobelts. Northern mid-latitude catchments see performance gains of -32.44%.

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 5). 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 4). 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 5. 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 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]
38	COOPER CREE	SDR	20.05†	WaterGAP2 IPE = 107.00	-8673.69	1	C1 [-1.17]: 0	18
							C2 [79.32]: + (-0.143) * H08 * (WaterGAP2 +1) * cos(cos(WaterGAP2))	
							C3 [99.23]: + 0.436*H08*sqrt WaterGAP2	
40	DARLING RIVER	SML	-1.64	WaterGAP2 IPE = -1.54	-1350.25	2	C1 [10.09]: 0.174*H08^2/DBH	11
							C2 [23.87]: + (-0.06/DBH)	
							C3 [131.00]: + H08/DBH	
39	FITZRO	SML	30.64†	WaterGAP2 IPE = 38.47	-783.52	3	C1 [10.65]: sin(H08/-4.91)	20
							C2 [38.47]: + WaterGAP2	
							C3 [45.72]: + sin((LPJmL - sqrt DBH-8.45)*(WaterGAP2+H08)/(DBH *PCRGLOBWB))	
4	Ob	BOR	-1.32	WaterGAP2 IPE = 2.50	-580.95	4	C1 [1.33]: 2*DBH/(log(sin H08)+6247.9)	15
							C2 [1.59]: + sqrt H08	
							C3 [3.29]: + WaterGAP2/H08^2	
33	RIO PARNAIBA	SST	-2.27	WaterGAP2 IPE = 1.46	-573.98	5	C1 [1.26]: 3.695	20
							C2 [2.15]: + 0.625*((cos(0.227/H08))^6*(log(WaterGAP2))^4)	
							C3 [3.26]: + 1.472 / (log(1/PCRGLOBWB) – 1.08396)	
36	BURDEKIN	SST	-1.35	H08 IPE = 1.44	-479.24	6	C1 [-1.12]: 0	10
							C2 [-1.01]: + sqrt H08	
							C3 [1.48]: + H08 * sin(log(log(PCRGLOBWB/2)))	
10	RED RIVER	BOR	-1.25†	WaterGAP2 IPE = 1.52	-477.75	7	C1 [-1.27]: H08*WaterGAP2/10.045	23
							C2 [1.37]: + sin PCRGLOBWB^3/(DBH^3*H08+H08-LPJmL-5.44)	
							C3 [1.42]: + sin(cos(WaterGAP2))^3	
19	FRASER RIVER	NML	-1.61	PCRGLOBWB IPE = 1.15	-477.34	8	C1 [-1.10]: 0.33*DBH*sqrt(log(PCRGLOBWB))	17
							C2 [2.66]: + cos((H08+1.63)/LPJmL)+8.12	
							C3 [3.70]: + cos H08	
18	ST. LAWRENCE	NML	2.47	WaterGAP2 IPE = 7.09	-461.94	9	C1 [2.51]: 23.04	19
							C2 [107.37]: + 0.67*sqrt WaterGAP2 * cos(sqrt WaterGAP2+ 1.42/H08)	
							C3 [108.82]: + 1.1*sqrt(DBH/PCRGLOBWB)	
2	AMUR	BOR	-1.49	EM IPE = 1.07	-356.32	10	C1 [-1.10]: 2.534*(DBH-H08-LPJmL-LPJmL/H08)/PCRGLOBWB	18
							C2 [2.27]: + WaterGAP2-4.33	
							C3 [3.23]: + sin DBH	

23	NEVA	NML	1.09	WaterGAP2 IPE =2.42	-132.70	11	C1 [8.94]: PCRGLOBWB	13
							C2 [9.91]: + log(DBH^3)	
							C3 [21.09]: + WaterGAP2/PCRGLOBWB + 0.5*log(log(WaterGAP2))	
5	KOLYMA	BOR	-2.38	DBH IPE = -1.23	- -113.88	12	C1 [-1.23]: DBH	14
							C2 [1.12]: + sqrt LPJmL	
							C3 [1.22]: + DBH*(-2.74*DBH+LPJmL-3.133)/WaterGAP2	
26	YANGTZE	NML	-4.17	WaterGAP2 IPE = -3.03	-107.67	13	C1 [-3.03]: WaterGAP2	13
							C2 [2.90]: + sqrt LPJmL	
							C3 [3.49]: + cos(PCRGLOBWB +0.039*H08*PCRGLOBWB/DBH)	
1	LENA	BOR	-2.00	WaterGAP2 IPE = -1.22	-78.43	14	C1 [-1.72]: WaterGAP2-sqrt DBH	15
							C2 [1.20]: + LPJmL/(2*LPJmL/WaterGAP2^2+5.575)	
							C3 [1.53]: + 0.997	
31	AMAZONAS	EQT	-1.85	WaterGAP2 IPE = -1.09	-74.97	15	C1 [-1.09]: WaterGAP2	19
							C2 [26.65]: + (H08-DBH+LPJmL+0.77)* (WaterGAP2-LPJmL- 0.77)/(PCRGLOBWB+24.9)	
							C3 [29.16]: + (-2.98)	
14	NORTHERN DVINA	BOR	-2.27	EM IPE= -1.54	-70.10	16	C1 [-1.51]: WaterGAP2	3
							C2 [-1.15]: + PCRGLOBWB	
							C3 [2.17]: + (-9.29)	
3	YENISEI	BOR	-2.32	WaterGAP2 IPE = -1.72	-57.68	17	C1 [-1.72]: WaterGAP2	7
							C2 [1.90]: + (-0.742)	
							C3 [2.18]: + 7.0*sin(sqrt H08)	
9	ASSINIBOINE	BOR	1.06	WaterGAP2 IPE = 1.57	-51.35	18	C1 [1.37]: WaterGAP2^2	17
							C2 [1.81]: + sin(0.5*log(0.268*H08+cosWaterGAP2/WaterGAP2+0.003))	
							C3 [4.93]: + 0.064	
21	RHINE RIVER	NML	-2.50	WaterGAP2 IPE = -1.96	-50.72	19	C1 [-1.96]: WaterGAP2	5
							C2 [4.59]: + 5.813	
							C3 [8.41]: + (-0.153)*H08	
12	CHURCHILL RIVER	BOR	3.10	WaterGAP2 IPE = 3.60	-50.33	20	C1 [3.60]: WaterGAP2	6
							C2 [25.66]: + sin PCRGLOBWB	
							C3 [27.48]: + cos(sqrt H08)	
29	NIGER	NST	-1.79	WaterGAP2 IPE = -1.37	-41.28	21	C1 [1.16]: 0.062* log(DBH)^4*(cos(4.647/PCRGLOBWB))^6	17
							C2 [2.77]: + cos(sin LPJmL/WaterGAP2)	
							C3 [2.92]: + 0.556	
8	SASKATCHEWAN	BOR	1.03	WaterGAP2 IPE = 1.43	-39.88	22	C1 [1.43]: WaterGAP2	29
							C2 [4.60]: + (cos(cos(DBH + log WaterGAP2 + 0.31))-sin(sqrt PCRGLOBWB^3))^3	
							C3 [7.59]: + -sin((log LPJmL^3)/8-sin(cos(0.401*LPJmL)+1.723)	

30	ZAIRE	EQT	1.42	WaterGAP2 IPE = 1.78	-36.35	23	C1 [1.78]: WaterGAP2 C2 [23.20]: + cos(sqrt DBH) C3 [23.20]: + cos(sqrt DBH)	7
15	YELLOW	NML	1.16	WaterGAP2 IPE = 1.49	-32.99	24	C1 [1.23]: sqrt(DBH) C2 [2.28]: + DBH*WaterGAP2^5/4/(DBH^2*WaterGAP2-0.043*PCRGLOBWB) C3 [2.67]: + (sin WaterGAP2)^2*sin(sqrt(PCRGLOBWB+DBH))	26
13	ALBANY RIVER	BOR	-1.66	PCRGLOBWB IPE = -1.33	-32.88	25	C1 [-1.33]: PCRGLOBWB C2 [1.02]: + log(0.106*DBH) C3 [1.06]: + log(0.041*DBH)	9
22	DANUBE	NML	-2.22	WaterGAP2 IPE = -1.89	-31.90	26	C1 [-1.89]: WaterGAP2 C2 [6.43]: + DBH/H08- H08/(PCRGLOBWB-1) C3 [6.77]: + 7.93/H08	13
25	DON	NML	1.23	WaterGAP2 IPE = 1.54	-31.33	27	C1 [1.54]: WaterGAP2 C2 [4.91]: + 1 C3 [11.26]: + (-0.325)*WaterGAP2	5
34	SAO FRANCISCO	SST	-1.92†	WaterGAP2 IPE = -1.64	-29.24	28	C1 [-2.17]: sqrt(WaterGAP2) C2 [2.53]: + 1.46*(PCRGLOBWB+WaterGAP2-5.75)/log(PCRGLOBWB) C3 [3.33]: + cos(H08/LPJmL)	16
27	COLORADO	NDR	2.22	H08 IPE = 2.50	-28.63	29	C1 [6.42]: log(DBH) C2 [8.58]: + log(PCRGLOBWB) C3 [14.37]: + WaterGAP2/PCRGLOBWB	7
24	VOLGA	NML	-2.00	WaterGAP2 IPE = -1.75	-23.08	30	C1 [-1.92]: WaterGAP2-0.978 C2 [3.21]: + 3.35/DBH C3 [3.30]: + 0.999/LPJmL	9
37	ORANJE	SDR	2.04	WaterGAP2 IPE = 2.26	-22.35	31	C1 [2.26]: WaterGAP2 C2 [6.89]: + 0.808 C3 [7.19]: + (-0.672)	3
28	SANTIAGO	NDR	1.16	WaterGAP2 IPE = 1.35	-19.30	32	C1 [1.33]: sin(LPJmL^2*(0.319-LPJmL/DBH))/DBH C2 [1.35]: + WaterGAP2 C3 [1.66]: + sin((sin(((sin((LPJmL))-(((LPJmL)/(WaterGAP2))^3))^2))-(WaterGAP2)))	24
17	MISSISSIPPI	NML	-2.04	WaterGAP2 IPE = -1.89	-14.18	33	C1 [-1.89]: WaterGAP2 C2 [5.43]: + (log(WaterGAP2^3)-WaterGAP2)/PCRGLOBWB C3 [5.88]: + (-1.70-DBH)/PCRGLOBWB	13
32	XINGU	EQT	1.04	WaterGAP2 IPE = 1.16	-8.63	34	C1 [1.16]: WaterGAP2 C2 [3.81]: + (-0.494) C3 [4.01]: + (-0.204)*LPJmL/sqrt WaterGAP2	8

20	LABE	NML	-1.45†	WaterGAP2 IPE = -1.47	2.00	35	C1 [-1.47]: WaterGAP2 C2 [3.68]: + 4.32/DBH C3 [3.71]: + 0.962*sin((DBH-H08)/PCRGLOBWB+ cos(0.15*WaterGAP2))	17
7	MACKENZIE RIVER	BOR	-1.33	EM IPE = -1.39	5.42	36	C1 [-1.30]: PCRGLOBWB C2 [3.46]: + 0.107*DBH C3 [4.92]: + (-0.978)	5
16	COLUMBIA	NML	-1.20	EM IPE = -1.28	6.53	37	C1 [-1.11]: WaterGAP2 C2 [8.96]: + sin(cos(LPJmL)^3)^2*sin(PCRGLOBWB*cos(3.78*PCRGLOBWB)) C3 [9.35]:+ exp(cos(cos(LPJmL)*sin(WaterGAP2)))* sin(0.479+0.166*WaterGAP2)	28
11	WINNIPEG RIVER	BOR	1.63	PCRGLOBWB IPE = 1.55	6.72	38	C1 [1.67]: WaterGAP2 C2 [8.64]: + H08/DBH C3 [11.05]: + (-4.91+log(PCRGLOBWB))	8
35	PARAGUAI	SST	8.51	WaterGAP2 IPE = 8.00	18.61	39	C1 [8.00]: WaterGAP2 C2 [18.71]: log(9.84/LPJmL) C3 [23.74]: 0.99- (LPJmL/(PCRGLOBWB-((LPJmL+WaterGAP2)/945.48)))	16
6	OLENEK	BOR	-1.15	DBH IPE = -1.47	32.72	40	C1 [2.14]: -sin(0.004* LPJmL^2*PCRGLOBWB-LPJmL+9.04) C2 [2.71]: + PCRGLOBWB/(-0.31*DBH^2*cosec(PCRGLOBWB) -7.71) C3 [3.94]: + WaterGAP2	22

1-As defined in Section 2.2, equation size is calculated according to the number of inputs (GHMs), constants, operators and functions in an equation.

Note: † is used to identify catchments where the IPE score of an individual MMC component is lower than the IPE score for the overall MMC equation.

Table 6. 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	-51.35	-415.19
NML	12	-32.44	-434.02
NDR	2	-23.97	-519.78
NST	1	-41.28	-763.67
EQT	3	-36.35	-160.98
SST	4	-254.24	-1698.24
SDR	2	-4348.02*	-104900.11*
SML	2	-1066.88*	-703067.50*

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

5. Using MMC to Gain Heuristic Insights

In this section we use results from selected catchments to exemplify ways in which the SR equations that define the MMC solutions may, or not, be usefully interpreted, and the insights that may be gained from their successes and failures. In so doing, we also suggest areas for fruitful future research.

5.1. Potential reasons for MMC failure

The basic premise of GEP is that it can formulate optimised SR equations for combining and adjusting individual GHM inputs so that an improved MMC solution is delivered. However, to be effective, the structural complexity of the SR equation should reflect the complexity of the combinatorial problem, and there should be sufficient diversity, coherently structured, in the GHM outputs to facilitate learning of optimal combination mechanisms from their inherent numerical patterns. If either, or both, of these factors is lacking, the likelihood that GEP will be able to improve upon an individual GHM will decrease. Examples of both excessive structural complexity and a lack of diversity in residual structures are evident in the six catchments where GEP failed to deliver performance gains over the best-performing GHM or EM.

5.1.1. Structural complexity.

The Labe catchment provides a good example of how an inappropriate level of structural complexity in the SR equation can limit GEP's capacity to deliver MMC performance gain. In this catchment, the best performing GHM is WaterGAP2 and the MMC solution does not quite match its performance [Labe IPE_{BPM} = -1.47 and IPE_{MMC} = -1.45 (Table 4)]. It is evident from the

SR equation (Equation 6) that the GEP algorithm has correctly learnt the efficacy of WaterGAP2 in the Labe catchment because it is the sole term in the first equation component:

$$\begin{gathered} MMC_{Labe} = WaterGAP2 + 4.32/DBH + \\ 0.962*\sin((DBH\text{-}H08)/PCRGLOBWB+\cos(0.15*WaterGAP2)) \end{gathered} \tag{6}$$

However, the default structural complexity of the equation (which is fixed at three components - see Table S2) dictates that the GEP learning process delivers terms to the remaining two equation components. In doing so, the performance of the complete SR equation is reduced so that it is worse than its first component and the best-performing GHM counterpart. In other words, the default level of complexity used in the GEP algorithm has reduced the MMC performance. An additional complexity problem is evident in the SR equation for the Columbia catchment MMC (Equation 7) which fails to improve upon the performance of the EM:

$$\begin{gathered} MMC_{Columbia} = WaterGAP2 + \\ \sin(\cos(LPJmL)^\wedge 3)^\wedge 2*\sin(PCRGLOBWB*\cos(3.78*PCRGLOBWB)) + \\ \exp(\cos(\cos(LPJmL)*\sin(WaterGAP2)))*\sin(0.479+0.166*WaterGAP2) \end{gathered} \tag{7}$$

Here, WaterGAP2 is the best performing model (IPE_{BPM} = -1.11) and is once again the sole term in the first SR equation component. Whilst the additional two components of the equation do deliver a moderate performance gain over WaterGAP2 (IPE_{MMC} = -1.20; MMC_{PG} = 9%) their complexity is excessive relative to the performance gain achieved and the complexity precludes any meaningful mechanistic interpretation of the equation. This raises questions about whether the MMC solution is sufficiently parsimonious or over fitted. Similar questions surround the MMC solutions for the Sao Francisco, Red River, Fitzroy River and Cooper Creek catchments.

These examples highlight the fact that a 'one-size-fits-all' setting for equation complexity can both reduce the efficacy of GEP-based MMC and limit the opportunities for gaining useful insights from the equations themselves. It also indicates that, in some cases, the fitness / size trade off we devised to avoid excessive MMC solution complexity performs poorly (Figure 3). Consequently, we suggest that more flexible, data-driven, *a priori* methods are needed that can estimate the most appropriate complexity from the GHM outputs prior to the GEP algorithm being employed – with approaches based on information content [74, 75] offering a potential way forward.

5.1.2. Residual structures.

At its most fundamental, MMC-GEP employs numerical combination methods to exploit complementary patterns in model outputs and deliver performance gains. The potential for these gains will, therefore, be limited in situations where:

i. The residuals of individual models are all similarly structured (i.e. all models perform similarly poorly and in much the same way which limits the GEP algorithm's ability to exploit complementary residual structures as a mechanism for correcting GHM output errors) and / or;

ii. There is little or no structure in residuals of all of the individual GHMs (i.e. the randomness of the GHM error structures prevents the GEP algorithm from learning numerical mechanisms for correcting GHM outputs.

The Olenek and Winnipeg basins exemplify the importance of these. In the Olenek (Figure 6), the residual structure is very similar for all individual GHMs and is characterised by an approximately linear increase in runoff under-estimation with increasing observed runoff. The lack of diversity in the residual structures offers the GEP algorithm little in the way of complementarity in the GHM residuals to exploit during the learning process. This is likely to be a reason why the MMC failed to improve upon the best performing GHM in this catchment. In the Winnipeg (Figure 7) the residuals are more diverse – especially for DBH, H08 and LPJmL. Nonetheless, a linear trend of under-estimations with increasing observed runoff is again evident in all GHM outputs (see the data points inside the red circles); again limiting the opportunity for GEP learning through complementary residual structures. The residuals that are not part of this trend have no evident structure and thus provide the GEP algorithm little additional opportunity for learning effective correction mechanisms.

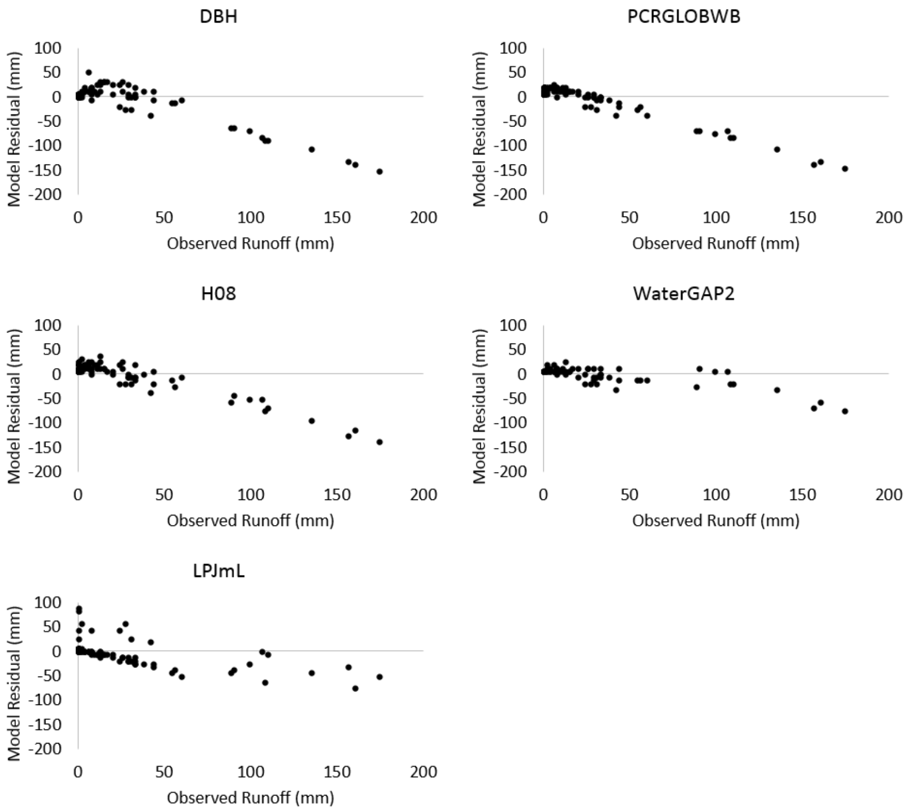

Figure 6. Residuals of the individual GHMs in the Olenek catchment.

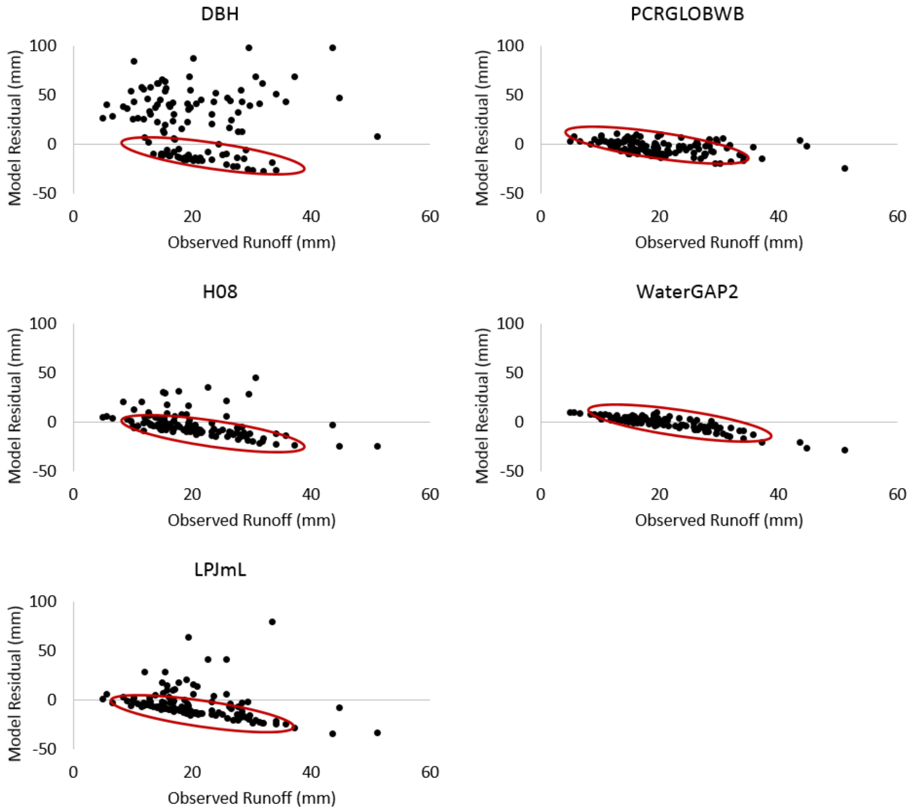

Figure 7. Residuals of the individual GHMs in the Winnipeg catchment.

5.2. Interpretation of MMC solutions

The explicit nature of the output delivered by GEP-based MMC, and the disaggregation of MMC solutions into separate SR equation components, combined by addition, may provide opportunities for their meaningful interpretation. We here consider three potential dimensions for interpretation whilst recognising that formal methods to ensure robustness and correctness of the interpretations that are made (an area for future research) are both lacking and needed:

i. GHM contribution – may help to reveal which GHM, or combination of GHMs, is better suited / poorly suited to solving the characteristic prediction challenge presented by each catchment or hydrobelt.

ii. MMC equation size – may help to reveal the complexity of the challenge of improving GHM predictions in different catchments. Catchments where the MMC optimisation requires only simple numerical adjustment mechanisms should have lower equation sizes than their more challenging counterparts. Particularly large equation sizes may indicate overfitting.

iii. Equation component examination – may help qualify the numerical mechanisms by which MMC optimisation is achieved, especially for medium and large size equations. This may help reveal the degree of linearity / non-linearity in the adjustments that each component of the MMC solution encodes. Identifying which GHM is responsible for delivering the majority of the MMC fit (i.e. has the highest explanatory power in the solution), and examining how alternative GHMs are used by MMC to adjust the fit residuals offers a means of diagnosing how specific weakness in a given GHM might be addressed by incorporating modelling mechanisms from alternative models.

5.2.1. Interpreting GHM contributions

One of GEP's unusual properties is its ability to preferentially select or deselect GHMs entirely from its MMC solutions and the SR equation components that comprise them. On the assumption that this is optimised during the learning process, the extent of inclusion / exclusion of GHMs should offer diagnostic insights about their relative merits.

GHM contribution rates for all 40 catchments (Table 7) reveal the preferential selection of WaterGAP2 by the GEP algorithm – it contributed to 87.5 % of the MMC solutions. By contrast, the output from LPJml contributed in only 40 % of MMC solutions. All other models contributed in at least 50% of the MMC solutions: DBH (65%); PCRGLOBWB (65%); H08 (50%). As well as having the highest overall contribution rate, WaterGAP2 was also the most frequent contributor

(65%) to component 1 of the SR equation – i.e. the component that delivered the best, standalone IPE score (see Table 5). This compares to 20% (both DBH and H08), 15% (PCRGLOBWB), and only 5% (JPJmL). Given that WaterGAP2 is the only calibrated GHM in the ensemble, its frequency of occurrence, particularly within the component 1 of the SR equation, is unsurprising. However, the inclusion (or not) of a GHM in a GEP-based MMC should not be interpreted in terms of its 'stand-alone' performance. Indeed, GHMs that have relatively weak performance may be included in the MMC because their outputs have specific residual structures that can be exploited to deliver numerical adjustments to consistent weaknesses in better performing GHMs. For this reason, all 40 MMC solutions involve two or more GHMs. As reflected by their disproportionate contribution to equation components 2 and 3 rather than component 1, DBH, H08 and PCRGLOBWB appear to regularly be used in this way. The infrequent contribution of JPJmL and its lack of contribution to the component 1 across the MMC solutions is also noteworthy. This may reflect general weaknesses in JPJmL's stand-alone performance relative to the other models, and/or indicate patterns of fit residuals that are difficult to exploit as adjustment mechanisms.

When examined by hemisphere, an interesting contrast between DBH and H08 is evident. DBH is a contributor in 76% of the MMC solutions for northern hemisphere catchments; contributing to component 1 at a rate of 24%. In the southern hemisphere catchments it contributes to only 25% of the MMC solutions and to component 1 at a rate of 13%. By contrast, H08's contribution to northern hemisphere catchment MMC solutions is low (38% overall; 7% for component 1). However, it contributes to 75% of the MMC solutions for southern hemisphere catchments and is in component 1 for 50% of these catchments. The reasons for this are not clear, but is does raise the question of whether there are characteristic differences between the models, or the catchments, or both that make the former better suited to runoff simulation in the northern hemisphere, and the latter better suited to application in southern hemisphere catchments.

When examined at the level of hydrobelts, it is apparent that the pre-eminence of WaterGAP2 in the MMC solutions is not universal. In the SST hydrobelt, H08 is a more frequent contributor to the MMC solutions – contributing to all four SST MMC equations (Table 7). In all other hydrobelts WaterGAP2 has the greatest rate of contribution to the MMC equations, although in the BOR and NML hydrobelts both DBH and PCRGLOBWB have rates that are significantly higher than rates of H08 or JPJmL.

Table 7. The contribution of individual GHMs to the MMC solution. Larger crosses in bold (and totals in brackets) indicate the inclusion of the GHM in the component 1 of the SR equation.

Catchment No.	River	Hydrobelt	DBH	H08	LPJmL	PCRGLOBWB	WaterGAP2
1	LENA	BOR	**X**		x		**X**
2	AMUR	BOR	x	x	x	x	**X**
3	YENISEI	BOR		x			**X**
4	OB	BOR	**X**	**X**			x
5	KOLYMA	BOR	**X**		x		x
6	OLENEK	BOR	x		**X**	**X**	x
7	MACKENZIE RIVER	BOR	x			**X**	
8	SASKATCHEWAN	BOR	x		x	x	**X**
9	ASSINIBOINE	BOR		x			**X**
10	RED RIVER	BOR	x	**X**	x	x	**X**
11	WINNIPEG RIVER	BOR	x	x		x	**X**
12	CHURCHILL RIVER	BOR		x		x	**X**
13	ALBANY RIVER	BOR	x			**X**	
14	NORTHERN DVINA	BOR				x	**X**
BOREAL HYDROBELT TOTALS			10(3)	7(2)	6(1)	9(3)	12(9)
15	YELLOW	NML	**X**			x	x
16	COLUMBIA	NML			x	x	**X**
17	MISSISSIPPI	NML	x			x	**X**
18	ST.LAWRENCE	NML	x	x		x	x
19	FRASER RIVER	NML	**X**	x	x	**X**	
20	LABE	NML	x	x		x	**X**
21	RHINE RIVER	NML		x			**X**
22	DANUBE	NML	x	x		x	**X**
23	NEVA	NML	x			**X**	x
24	VOLGA	NML	x		**X**		x
25	DON	NML					**X**
26	YANGTZE	NML	x	x	x	x	**X**
NORTHERN MID LATITUDE HYDROBELT TOTALS			9(2)	6(0)	4(1)	9(2)	11(7)
27	COLORADO	NDR	**X**			x	x
28	SANTIAGO	NDR	x		x		**X**
NORTHERN DRY HYDROBELT TOTALS			2(1)	0(0)	1(0)	1(0)	2(1)
29	NIGER	NST	**X**		x	**X**	x
NORTHERN SUB TROPICAL HYDROBELT TOTALS			1(1)	0(0)	1(0)	1(1)	1(0)
30	ZAIRE	EQT	x				**X**
31	AMAZONAS	EQT	x	x	x	x	**X**
32	XINGU	EQT			x		**X**
EQUATORIAL HYDROBELT TOTALS			2(0)	1(0)	2(0)	1(0)	3(3)
33	RIO PARNAIBA	SST		**X**		x	**X**
34	SAO FRANCISCO	SST		x	x	**X**	**X**
35	PARAGUAI	SST			x	x	**X**
36	BURDEKIN	SST		**X**		**X**	
SOUTHERN SUB TROPICAL HYDROBELT TOTALS			0(0)	3(2)	1(0)	4(2)	3(3)
37	ORANJE	SDR					**X**
38	COOPER CREEK	SDR		**X**			**X**
SOUTHERN DRY HYDROBELT TOTALS			0(0)	1(1)	0(0)	0(0)	2(2)
39	FITZROY	SML	x	x	x	x	**X**
40	DARLING RIVER	SML	**X**	**X**			
SOUTHERN MID LATITIUDE TOTALS			2(1)	2(1)	1(0)	1(0)	1(1)
ALL HYDROBELTS TOTALS			**26(8)**	**20(6)**	**16(2)**	**26(8)**	**35(26)**

5.2.2. Interpreting MMC equation sizes

The MMC equation size is the total number of mathematical functions, operators, model inputs and constants that comprise it. The MMC equation size distribution by hydrobelt is presented in Table 8. The overall distribution of equation complexity approximates a right skewed, normal distribution (skewedness = 0.26) with a minimum of 3, maximum of 29 and a median of 14.5. The mean complexity of MMC solutions is lowest in the SDR and EQT hydrobelts – although the robustness of this observation is limited by the small number of catchments in these hydrobelts, negating meaningful interpretation. More meaningful is the difference between equation sizes in the BOR (mean = 13.6; StDev = 7.7) and NML (mean = 14.8; StDev = 7.2) hydrobelts which suggests that Boreal catchments present a slightly greater modelling challenge overall, and have a greater diversity in the scale of the MMC challenge (highlighted by the higher standard deviation) compared to their Northern Mid Latitude counterparts. This may be a result of the specific difficulties associated with modelling snowmelt processes in some Boreal catchments.

Table 8. MMC equation size by hydrobelt.

Hydrobelt	No. of catchments	Maximum MMC equation size	Minimum MMC equation size	Mean MMC equation size	Standard Deviation of Mean (where n =>4)
BOR	14	29	3	13.6	7.7
NML	12	28	5	14.8	7.2
NDR	2	24	7	15.5	N/A
NST	1	17	17	17	N/A
EQT	3	19	7	11.3	N/A
SST	4	20	10	15.5	4.1
SDR	2	18	3	10.5	N/A
SML	2	20	11	15.5	N/A

Two catchments (Northern Dvina and Oranje) had the minimum MMC solution complexity of 3 and, in both cases, the MMC solution that was developed was a simple, linear adjustment to a GHM. Simple, linear adjustments were also present in the Rhine, Don and Mackenzie catchments. In the case of the Oranje (Equation 8), the adjustment comprised the addition of a constant (0.136) to the output of WaterGAP2. This reduced the IPE score from 2.26 to 2.00 (IPE_{PG} = -22.35 %) and indicates that, despite WaterGAP2 being the best-performing model in the catchment, it delivers a systematic and consistent under-estimation of runoff which can be easily improved by a simple correction factor. It highlights the potential of MMC as a method for bias correction and informing improved model calibration procedures.

$$MMC_{Oranje} = WaterGAP2 + (-0.672) + 0.808 \qquad (8)$$

In the case of Northern Dvina, the MMC solution summed the output of PCRGLOBWB with that of WaterGAP2; adjusted by the subtraction of a constant value of 9.29 (Equation 9). This has the effect of amplifying the peak runoff in the MMC output (which the individual models significantly underestimate) and reducing the IPE score from -1.53 to -2.27 (MMC_{PG} = -73.59%). The solution indicates a lack of sensitivity in the runoff response in both PCRGLOBWB and WaterGAP2 in this catchment.

$$MMC_{NorthernDvina} = PCRGLOBWB + WaterGAP2 + (-9.29) \qquad (9)$$

With an increase in the equation size comes a general shift from relatively simple linear to more complex non-linear adjustments. For example, the MMC for the Yenisei catchment (equation size = 9) applies a non-linear adjustment to WaterGAP2 that is informed by the output of H08 (Equation 10) and that reduces the IPE from -1.72 to -2.32 (MMC_{PG} = -57.68 %). This adjustment alters the low flows in WaterGAP2 downwards whilst maintaining the high flows to correct for the over-estimation of low flows by WaterGAP2 – suggesting that, in this catchment, the mechanisms for modelling low flows in H08 may be preferable to those used in WaterGAP2.

$$MMC_{Yenisei} = 7.0*\sin(\mathrm{sqrt}\ H08) + (-0.742) + WaterGAP2 \qquad (10)$$

MMC solutions with complexity values close to or above the median are considerably more complex and more difficult to interpret directly from the equations, because they tend to involve a greater range of GHMs and more complex non-linear adjustment mechanisms (although the dominant solution components in Olenek and Columbia do comprise just a single model). For example, the Ob catchment MMC equation has the median equation size of 15 (Equation 11) and applies a set of non-linear adjustments to the outputs from H08, WaterGAP2 and DBH to deliver its performance gain (MMC_{PG} = -580.95%):

$$MMC_{Ob} = \mathrm{sqrt}H08 + WaterGAP2/H08\hat{}2 + 2*DBH/(\log(\sin H08) + 6247.9) \qquad (11)$$

Whilst the effect of the first component of the Ob MMC equation is relatively easy to understand (applying a square root to H08 achieves a non-linear reduction to its output amplitude that corrects for error in the simulated runoff magnitude), the combined effect of the remainder of the solution components is more difficult to determine. This is particularly the case for MMC solutions with particularly large equation sizes. These are impossible to interpret directly from the equations and may require particular attention because they may be over fitted. For example, the MMC equation for the Olenek has a size of 22 (Equation 12) and incorporates a series of complex, non-linear adjustments to its four contributing models. Despite this, the MMC solution does not perform as well as the best performing individual model (DBH) – a result, that indicates a case of overfitting.

$$\begin{gathered}\text{MMC}_{\text{Olenek}} = -\sin(0.004 * \text{LPJmL}^{\wedge}2 * \text{PCRGLOBWB} - \text{LPJmL} + 9.04) \\ + \text{PCRGLOBWB}/(-0.31 * \text{DBH}^{\wedge}2 * \operatorname{cosec}(\text{PCRGLOBWB}) - 7.71) \\ + \text{WaterGAP2}\end{gathered} \tag{12}$$

5.2.3. Interpreting SR equation components

The SR equations that underpin the MMC solutions comprise three components, combined via simple addition (see Table 5). It has been suggested that, during the evolutionary learning process employed in GEP, each component is invoked and adjusted separately so that it is tailored towards solving a specific characteristic of the overall modelling problem [23, 24]. By plotting and examining the behaviour of the SR components, we suggest that for some catchments, it may be possible to gain useful, qualitative insights into the mechanisms by which performance gains have been achieved, and the roles of different GHMs in achieving this on a catchment-by-catchment basis. Plots of the MMC components for each catchment are included in the Supplementary Information.

A complete, catchment-by-catchment analysis and characterisation of the correction mechanisms used by the SR equation components is beyond the scope of this paper. Indeed, the complexity of many of the SR equations preclude a simplistic mechanistic interpretation and the development of effective methods for achieving this represent an important opportunity for further research. Nonetheless, we indicate how an MMC might support improved qualitative understanding of model deficiencies and correction mechanisms using the case of the Amazonas (Equation 13). This catchment provides a particularly clear exemplification of how different SR equation components can combine to address timing and magnitude errors (and associated MMC performance gains) in a GHM – in this case WaterGAP2.

$$\begin{gathered}\text{MMC}_{\text{Amazonas}} = \text{WaterGAP2} + \\ (\text{H08} - \text{DBH} + \text{JPJmL} + 0.77) * (\text{WaterGAP2} - \text{JPJmL} * 0.77) / (\text{PCRGLOBWB} + 24.9) + \\ (-2.98)\end{gathered} \tag{13}$$

WaterGAP2 is the best-performing GHM in the Amazonas catchment (IPE = -1.09) and is the sole term in component 1 of the SR equation. However, when plotted alongside the observed data (Figure 8 top), its output is seen to be out-of-phase and slightly ahead of the observed data. This timing error is relatively consistent in each of the seven years of validation data. By plotting component 1 alongside the sum of components 1 and 2 (Figure 8 bottom), it is evident that the GEP algorithm has learnt a way of using the outputs of the other GHMs to correct for this timing error by inducing a phase shift – effectively combining the outputs of the other GHMs to generate a 'switch' that delivers a negative adjustment to WaterGAP2 on its rising limb and a positive adjustment on its falling limb. In addition, the component 3 uses a constant to deliver a small, negative adjustment to runoff magnitude to correct for runoff over-estimation. The net effect of this is a 74.97% MMC performance gain over that delivered by WaterGAP2 and a characterisation of a numerical mechanism by which WaterGAP2's deficiencies in simulating runoff in the catchment may be reduced.

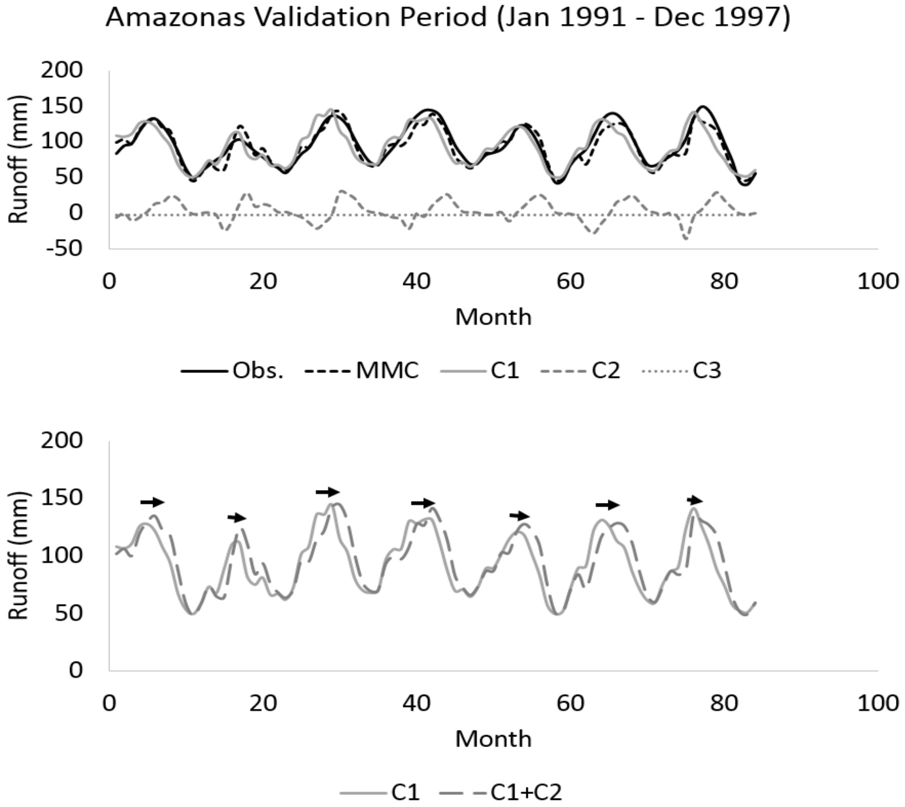

Figure 8. Equation components for the Amazons catchment, validation dataset.

6. Discussion

6. 1. To weight, or not to weight?

Our rationale for developing MMCs was in part a response to a question frequently asked by modellers, decision-makers, and the public, when presenting simulations from an ensemble of GHMs: *why not weight / adjust the models according to their performance*? We acknowledge that in other disciplines [51, 76, 77], including climate modelling [52, 78], weighting strategies have been shown to be highly effective in improving the performance of a model ensemble. However, our MMC approach goes beyond the application of constant weights and incorporates more complex non-linear adjustments alongside the application of weights to deliver better performing ensemble outputs – harnessing the potential of machine learning as a means of discovering optimal combination strategies. We argue that this may provide opportunities for gaining heuristic insights about the individual GHMs that are not possible using simple weighting strategies. Thus, the present study is a novel example of a more 'intelligent' approach to combining an ensemble of models and a first of its kind for GHMs. Nonetheless, a counter-argument remains that asserts any attempt to weight models will be futile as long as the current generation of models are far from being empirically adequate for purpose [79]. We agree with such concerns, but in this study have attempted to exemplify how the interpretation of MMC solutions may provide useful and important insights into the relative empirical strengths and weaknesses of different models; thereby offering an important means by which limitations in process representation may be better identified and understood.

Whilst our study highlights how MMC outputs will 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 4 and 5 and Table 4). Even though MMC techniques employed in other disciplines have been claimed to result in a "reduction of the uncertainty range" [51], 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 provides 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.

We acknowledge that our results reflect a fundamental error-complexity trade-off 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. Indeed, ensemble weighting exercises in other disciplines include both simple [78] and more complex [51] weighting approaches. However, with greater complexity comes a tendency towards overfitting of the MMC solutions and the delivery of MMC equations that are too complex to be meaningfully interpreted. Taken to its extreme, MMC could become nothing more than a meaningless curve fitting exercise. By applying an error-complexity trade-off selection method (Section 2.2; Figure 3) we have sought to minimise the risk of selecting uninterpretable, over-fitted MMC solutions. Whilst in Section 5.3 we exemplify how the SR equation components of the MMCs can be used to examine individual GHM performance and areas for GHM improvement, we also acknowledge that the simpler MMC equations are the easier to interpret. There is, therefore, an argument for 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. The increased interpretability of the solutions would, however, be at the expense of MMC performance. Identifying the 'sweet spot' where both performance gain and interpretation is maximise is an area for fruitful future research.

6.2. Future applications of MMC

We developed MMCs for individual catchments as opposed to developing a single MMC that operates across the globe. We made this decision *a priori* because we anticipated that whilst a single MMC would prove easy to apply by a user, it would be unlikely to transfer well across space. Indeed, weighting of global climate models has shown that individual model weights vary strongly with location, so that a model that receives nearly zero weight in some areas may still receive a large weight elsewhere [80]. The trade-off is that a more significant computing resource is required to compute MMCs for individual catchments but the benefit is significantly improved ensemble performance. In future work we plan to develop grid-cell level MMCs by making use of an observed gridded global runoff database [81].

Opportunities exist for gaining improved diagnostic insights from MMC equations where methods are available for identifying the characteristics adjustment mechanisms that are observed in the SR equations across the multiple catchments, and quantifying their importance globally and regionally. To this end, applying data-mining methods to the full set of SR equations may offer potential. Similarly, questions remain over the robustness with which MMCs that rely on historical data for their training, testing and validation (as in this study), and the extent to which they can be used for supporting future scenario projections; recognising that numerical adjustment mechanisms optimised for past data may be sub-optimal for future extrapolation. To this end, the ISIMIP2b project [82] will provide an important framework for applying MMCs to climate change scenarios of 1.5°C and 2.0°C global mean warming relative to pre-industrial levels.

7. Summary

Whilst model weighting has been applied in other disciplines, such as climate modelling, in this paper we have for the first time applied a set of intelligently defined weights to the individual global-scale hydrological models that comprise a state-of-the-art ensemble from the global hydrology modelling community.

The MMCs we have presented are shown to employ a diverse array of linear and non-linear adjustments to counteract the biases and fit residuals in the runoff estimates from 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 (or 40% with the outlying Cooper Creek, Darling and Fitzroy River catchments removed). The largest MMC performance gains are achieved in hydrobelts located in the southern hemisphere. However, the low number of catchments in the southern hemisphere hydrobelts preclude a more in-depth understanding of the potential and specific benefits of MMC therein. To address this will require the availability of data from a greater number of study catchments with the temporally-extensive runoff records that are needed to support the robust application of the MLAs that underpin MMC development. 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.

The EM performs better than individual GHMs in only 10% (4) of our catchments. This is 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[11, 15, 83-85]. The likely reason for our finding is the combination of a single, calibrated model (WaterGAP2) and a large number of uncalibrated counterparts, whose consistent tendency to over or underestimate runoff, in the end biases the EM. This supports earlier work that shows uncalibrated GHMs, and their ensemble mean, tend to show large positive biases relative to calibrated catchment-scale models [65].

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. The ISIMIP2a 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. JZ was supported by the Islamic Development Bank. YM was supported by the Environmental Research and Technology Fund (S-10) of the Ministry of the Environment, Japan.

References

[1] Bierkens M F P 2015 Global hydrology 2015: State, trends, and directions *Water Resources Research* **51(7)**: 4923-47

[2] Gosling S N and Arnell N W 2011 Simulating current global river runoff with a global hydrological model: model revisions, validation, and sensitivity analysis *Hydrological Processes* **25(7)**: 1129-45

[3] 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

[4] Liang X, Lettenmaier D P, Wood E F and 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(D7)**: 14415-28

[5] Koirala S, Yeh P J F, Hirabayashi Y, Kanae S and 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

[6] Prentice I C, Cramer W, Harrison S P, Leemans R, Monserud R A and Solomon A M 1992 Special Paper: A Global Biome Model Based on Plant Physiology and Dominance, Soil Properties and Climate *Journal of Biogeography* **19(2)**: 117

[7] Sitch S, et al. 2003 Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model *Global Change Biology* **9(2)**: 161-85

[8] Kundzewicz Z W 1986 The Hydrology of Tomorrow *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques* **31(2)**: 223-35

[9] Shamseldin A Y, OConnor K M and Liang G C 1997 Methods for combining the outputs of different rainfall-runoff models *Journal of Hydrology* **197(1-4)**: 203-29

[10] Liu Y, Guo H, Zhang Z, Wang L, Dai Y and Fan Y 2007 An optimization method based on scenario analysis for watershed management under uncertainty *Environ Manage* **39(5)**: 678-90

[11] Phillips T J and Gleckler P J 2006 Evaluation of continental precipitation in 20th century climate simulations: The utility of multimodel statistics *Water Resources Research* **42(3)**:

[12] Gosling S N, Bretherton D, Haines K and Arnell N W 2010 Global hydrology modelling and uncertainty: running multiple ensembles with a campus grid *Philos Trans A Math Phys Eng Sci* **368(1926)**: 4005-21

[13] Sanderson B M and Knutti R 2012 On the interpretation of constrained climate model ensembles *Geophysical Research Letters* **39**:

[14] Gudmundsson L, et al. 2012 Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe *Journal of Hydrometeorology* **13(2)**: 604-20

[15] 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 Resources Research* **48**:

[16] Masaki Y, Hanasaki N, Biemans H, Müller Schmied H, Tang Q, Wada Y, Gosling S, Takahashi K and 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*:

[17] 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

[18] van Huijgevoort M H J, et al. 2013 Global Multimodel Analysis of Drought in Runoff for the Second Half of the Twentieth Century *Journal of Hydrometeorology* **14(5)**: 1535-52

[19] Gao X and Dirmeyer P A 2006 A multimodel analysis, validation, and transferability study of global soil wetness products *Journal of Hydrometeorology* **7(6)**: 1218-36

[20] Guo Z C, Dirmeyer P A, Gao X and Zhao M 2007 Improving the quality of simulated soil moisture with a multi-model ensemble approach *Q J Roy Meteor Soc* **133(624)**: 731-47

[21] 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

[22] Beriro D. Gene Expression Programming Models of Bioaccessible Benzo[a]pyrene in Coking Works Soils: Nottingham; 2015.

[23] Ferreira C 2001 Gene Expression Programming: A New AdaptiveAlgorithm for Solving Problems **13(2)**: 87-129

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

[25] Bărbulescu A and 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

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

[27] Fernando A K, Shamseldin A Y and Abrahart R J 2012 Use of Gene Expression Programming for Multimodel Combination of Rainfall-Runoff Models *Journal of Hydrologic Engineering* **17(9)**: 975-85

[28] Aytek A, Asce M and Alp M 2008 An application of artificial intelligence for rainfall-runoff modeling *J Earth Syst Sci* **117(2)**: 145-55

[29] Beriro D J, Abrahart R J, Nathanail C P, Moreno J and 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)**: 29-41

[30] Traore S and 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)**: 1-10

[31] Shamseldin A Y and 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(6)**: 895-912

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

[33] Shamseldin A Y, Nasr A E and 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(4)**: 671-84

[34] Shamseldin A Y, O'Connor K M and 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

[35] See L and Abrahart R J 2001 Multi-model data fusion for hydrological forecasting *Computers & Geosciences* **27(8)**: 987-94

[36] Abrahart R J and 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)**: 655-70

[37] Georgakakos K P, Seo D J, Gupta H, Schaake J and Butts M B 2004 Towards the characterization of streamflow simulation uncertainty through multimodel ensembles *Journal of Hydrology* **298(1-4)**: 222-41

[38] Ajami N K, Duan Q Y, Gao X G and Sorooshian S 2006 Multimodel combination techniques for analysis of hydrological simulations: Application to Distributed Model Intercomparison Project results *Journal of Hydrometeorology* **7(4)**: 755-68

[39] Ajami N K, Duan Q Y and 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)**:

[40] Duan Q Y, Ajami N K, Gao X G and Sorooshian S 2007 Multi-model ensemble hydrologic prediction using Bayesian model averaging *Advances in Water Resources* **30(5)**: 1371-86

[41] Jeong D I and Kim Y-O 2009 Combining single-value streamflow forecasts – A review and guidelines for selecting techniques *Journal of Hydrology* **377(3-4)**: 284-99

[42] Viney N R, Vaze J, Chiew F H S, Perraud J, Post D A and Teng J 2009 Comparison of multi-model and multi-donor ensembles for regionalisation of runoff generation using five lumped rainfall-runoff models *18th World Imacs Congress and Modsim09 International Congress on Modelling and Simulation*: 3428-34

[43] Azmi M, Araghinejad S and 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

[44] Velázquez J A, Anctil F, Ramos M H and Perrin C 2011 Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures *Advances in Geosciences* **29**: 33-42

[45] Vel´azquez J A, Schmid J, Ricard S, Muerth M J, St-Denis B G, Minville M, Chaumont D, Caya D, Ludwig R and Turcotte R 2013 An ensemble approach to assess hydrological models' contribution to uncertainties in the analysis of climate change impact on water resources *Hydrol Earth Syst Sci* **17**: 565–78

[46] Nasseri M, Zahraie B, Ajami N K and Solomatine D P 2014 Monthly water balance modeling: Probabilistic, possibilistic and hybrid methods for model combination and ensemble simulation *Journal of Hydrology* **511**: 675-91

[47] Hibon M. E T 2005 To combine or not to combine: selecting among forecasts and their combinations *International Journal of Forecasting* **21** 15– 24

[48] Hibon M and Evgeniou T 2005 To combine or not to combine:selecting among forecasts and their combinations *International Journal of Forecasting* **21**: 15-24

[49] Arsenault R, Gatien P, Renaud B, Brissette F and Martel J L 2015 A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation *Journal of Hydrology* **529**: 754-67

[50] Li W, Sankarasubramanian A, Ranjithan R S and 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(8)**: 2255-69

[51] 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 *Journal of Climate* **15(10)**: 1141-58

[52] Fowler H J and Ekström M 2009 Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes *International Journal of Climatology* **29(3)**: 385-416

[53] H C J, E K, F G, G L and M R 2010 Weight assignment in regional climate models *CLIMATE RESEARCH* **44**: 179–94

[54] Chandler R E 2013 Exploiting strength, discounting weakness: combining information from multiple climate simulators *Philos Trans A Math Phys Eng Sci* **371(1991)**: 20120388

[55] Solomatine D P, See L M and 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.*

[56] Freitas P S A and Rodrigues A J L 2006 Model combination in neural-based forecasting *European Journal of Operational Research* **173(3)**: 801-14

[57] Warszawski L, Frieler K, Huber V, Piontek F, Serdeczny O and Schewe J 2014 The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): project framework *Proc Natl Acad Sci U S A* **111(9)**: 3228-32

[58] 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

[59] 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 *Hydrology and Earth System Sciences* **16(8)**: 3049-60

[60] Pushpalatha R, Perrin C, Le Moine N and Andreassian V 2012 A review of efficiency criteria suitable for evaluating low-flow simulations *Journal of Hydrology* **420**: 171-82

[61] Seibert J 2001 On the need for benchmarks in hydrological modelling *Hydrological Processes* **15(6)**: 1063-4

[62] WMO 2006 Technical Regulations, Volume III: Hydrology *Available online at: http://librarywmoint/pmb_ged/wmo_49-v3-2006_enpdf:*

[63] Haddeland I, et al. 2011 Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results *Journal of Hydrometeorology* **12(5)**: 869-84

[64] 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

[65] 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

[66] 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

[67] Bondeau A, et al. 2007 Modelling the role of agriculture for the 20th century global terrestrial carbon balance *Global Change Biology* **13(3)**: 679-706

[68] 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

[69] Wu W Y, Dandy G C and 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**: 108-27

[70] Phukoetphim P, Shamseldin A Y and 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

[71] Wu W, May R, Dandy G C and Maier H R, editors. A method for comparing data splitting approaches for developing hydrological ANN models. 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.

[72] Snee R D 1977 Validation of Regression-Models - Methods and Examples *Technometrics* **19(4)**: 415-28

[73] May R J, Maier H R and Dandy G C 2010 Data splitting for artificial neural networks using SOM-based stratified sampling *Neural Netw* **23(2)**: 283-94

[74] Liepe J, Filippi S, Komorowski M and Stumpf M P 2013 Maximizing the information content of experiments in systems biology *PLoS Comput Biol* **9(1)**: e1002888

[75] Beven K and Smith P 2015 Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models *Journal of Hydrologic Engineering* **20(1)**: A4014010

[76] Qi Y, Qian C and Yan Z 2017 An alternative multi-model ensemble mean approach for near-term projection *International Journal of Climatology* **37(1)**: 109-22

[77] Gillett N P 2015 Weighting climate model projections using observational constraints *Philos Trans A Math Phys Eng Sci* **373(2054)**:

[78] Christensen J H, Kjellstrom E, Giorgi F, Lenderink G and Rummukainen M 2010 Weight assignment in regional climate models *Climate Research* **44(2-3)**: 179-94

[79] Stainforth D A, Allen M R, Tredger E R and Smith L A 2007 Confidence, uncertainty and decision-support relevance in climate predictions *Philos Trans A Math Phys Eng Sci* **365(1857)**: 2145-61

[80] Räisänen J and Ylhäisi J S 2011 Can model weighting improve probabilistic projections of climate change? *Climate Dynamics* **39(7-8)**: 1981-98

[81] Koster R D, Fekete B M, Huffman G J and Stackhouse P W 2006 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(D22)**:

[82] Frieler K, et al. 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

[83] 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

[84] Hagemann S and Jacob D 2007 Gradient in the climate change signal of European discharge predicted by a multi-model ensemble *Climatic Change* **81**: 309-27

[85] 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 Resources Research* **46(6)**, W06507, doi:10.1029/2009WR007811