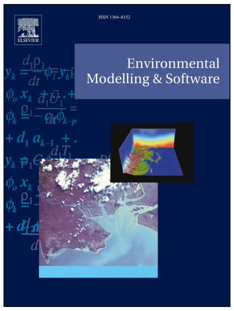


# Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models

Zaherpour Jamal, Mount Nick, N. Gosling Simon, Dankers Rutger, Eisner Stephanie, Gerten Dieter, Liu Xingcai, Masaki Yoshimitsu, Müller Schmied Hannes, Tang Qiuhong, Wada Yoshihide

PII: S1364-8152(17)30981-7

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

Reference: ENSO 4367

To appear in: *Environmental Modelling and Software*

Received Date: 18 September 2017

Revised Date: 11 December 2018

Accepted Date: 12 January 2019

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

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

Rest of co-authors in alphabetic order:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## Abstract

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

**Keywords:**

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

## Highlights:

- We present the first use of machine learning-based multi-model combination (MMC) applied to a global hydrological model ensemble.
- MMC performs better than any individual input model and the ensemble mean.
- MMC is not always able to out-perform model combination based on multiple linear regression.
- The physical interpretation of the MMC solutions is limited by the complexity of their non-linear weighting schemes.

## Software and/or data availability:

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

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

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

## Conflict of interest

None

## 1. Introduction

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

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

The global scope of GHMs, limited availability and quality of observed discharge data across the global domain and their use of spatially generalised parameters make them more difficult to calibrate than catchment hydrological models. Whilst examples of calibrated GHMs do exist (Müller Schmied et al., 2016), the majority of GHMs are uncalibrated (Gosling et al., 2016; Hattermann et al., 2017). This lack of calibration, coupled with the diversity of simplifications employed in the hydrological process representations, means that there can be large inconsistency in the skill, bias and uncertainty of an individual GHM at different locations, as well as large inconsistencies between different GHMs at any given location (van Huijgevoort et al., 2013; Zaherpour et al., 2018b). This spatial inconsistency means that GHMs risk becoming a "jungle of models" (Kundzewicz, 1986) in which it can be difficult to determine where a particular GHM output is likely to be capable of delivering optimal hydrological simulations. It also makes it dangerous to assume that any individual GHM will be an adequate basis for making projections at any given location, even if the model's ability to replicate observed data in particular catchments is enhanced through the acquisition of higher quality input data or efforts to improve process representations (Liu et al., 2007). To an extent, these arguments are also applicable to catchment hydrological models because whilst they have been shown to generally perform better than GHMs in model evaluation studies, ensembles of such models still result in an uncertainty range when the models are run with identical inputs (Hattermann et al., 2017; Hattermann et al., 2018).

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

A wide range of techniques can be used to generate an MMC solution. The simplest example is the calculation of the arithmetic mean of the input models (commonly referred to as an Ensemble Mean (EM)). More sophisticated techniques employ weighted schemes (Arsenault et al., 2015), with the differential weightings applied to each input model reflecting their relative strengths or weaknesses. The mathematical approach taken to determining the weights depends on the objective of the MMC. Where the primary objective is to minimise the difference between the MMC solution and observed data (i.e. maximise the predictive performance), without explicitly accounting for model or parameter uncertainty, the use of multiple linear regression (Doblas-Reyes et al., 2005) or machine learning algorithms (Lima et al., 2015; Worland et al., 2018) to 'learn' the optimal set weights to apply to each MMC input model is a popular approach (Marshall et al., 2007). The use of algorithms such as artificial neural networks (ANNs) (Shamseldin et al., 1997; Xiong et al., 2001) or gene expression programming (GEP) (Barbulescu and Bautu, 2010; Bărbulescu and Băutu, 2009; Fernando et al., 2012) to define non-linear weighting schemes have proven to be particularly effective. This is down to their ability to generate optimised, non-linear schemes rapidly, without the need for any prior knowledge of the model parameters.

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

An alternative approach is to use model combination methods that combine spatially co-incident variables in a dynamic manner. Such methods have included mechanistic approaches (Marshall et al., 2006) that adjust the weights as a conditional response to changes in one or more dynamic state variables (e.g. antecedent moisture) and statistical methods that maximise the temporal correlation of individual models through best linear unbiased estimation (Kim et al., 2015). However, dynamic approaches assume that is it possible to isolate, quantify and model the temporal relations contained within the suite of model outputs to be combined. It is unclear whether this will be possible for GHMs operating at the global-scale over multi-decadal periods because these relations, and the processes responsible for them are likely to be highly variable in space and time.

In this study we explore the potential of MMC for addressing the challenge of spatial inconsistency in simulations by GHMs, by combining outputs from a diverse set of five GHMs using GEP (Ferreira, 2001; Ferreira, 2006). 40 optimised MMC solutions of monthly mean runoff are generated for the period 1971 – 2010, one for each of 40 large catchments that are distributed throughout the world's eight hydrobelts (Meybeck et al., 2013) (Figure 1). In each catchment, the MMC's ability to replicate the observed monthly runoff is compared against that of the EM and each of the five GHMs from which the MMC is derived, as well as, the best-performing individual GHM from the ensemble. We also compare the MMC results against ordinary least squares multiple linear regression methods (Arsenault et al., 2015; Granger and Ramanathan, 1984) in order to assess the additional benefit gained by applying complex, machine learning methods rather than their simpler, linear counterparts (Arsenault et al., 2015; Mount and Abrahart, 2011).

The objectives of the paper are, therefore, twofold: 1) to assess the levels of performance gain that GEP-based MMC solutions can deliver to GHMs in different hydro-climatic settings and; 2) to critique the extent to which interpretation of GEP expressions can provide useful insights about the relative strengths and weaknesses of the different input models. Our experiments provide a clear demonstration that optimised MMCs of GHMs can deliver substantial performance gains in all hydrobelts when compared to the EM or individual GHMs, but that they do not always deliver benefits when compared to simpler, multiple linear regression approaches. They also highlight the challenges associated with delivering GEP-based MMCs that can be usefully and meaningfully interpreted.

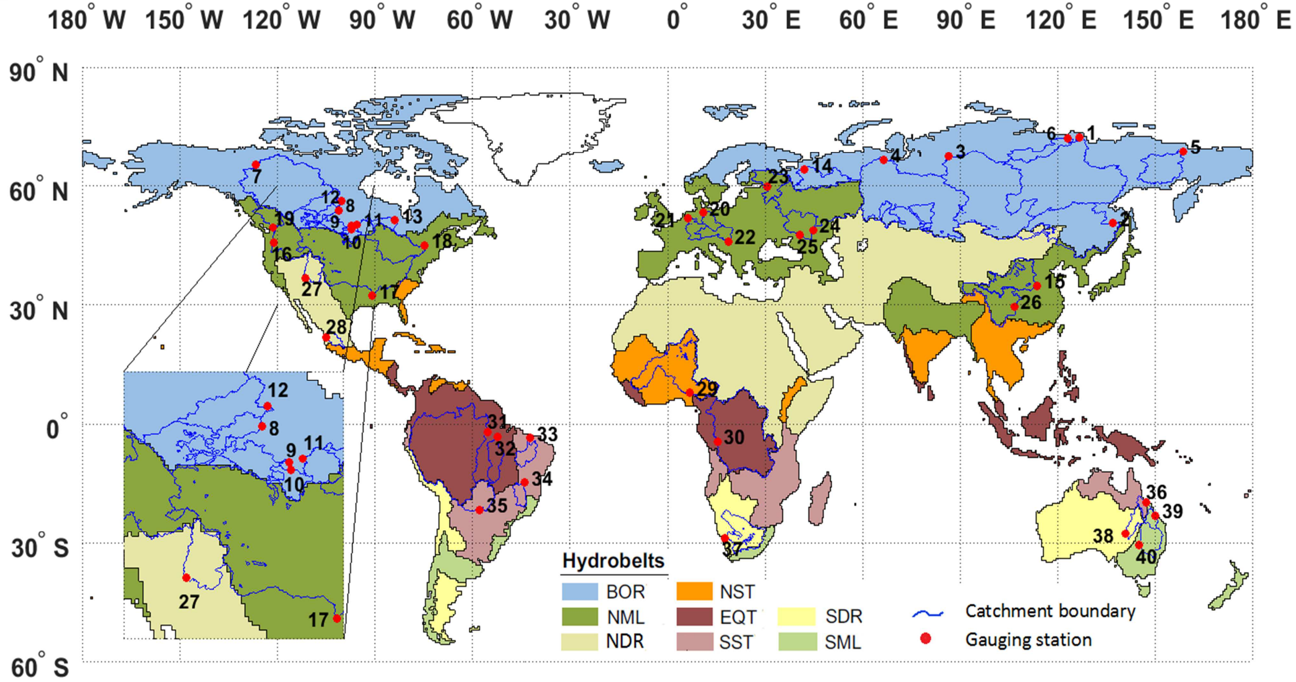


Figure 1. Locations of the 40 catchments (details in Table 1 and Table S1 in Supplementary Information) across the hydrobelt system defined in Meybeck et al. (2013). The hydrobelts are BOR= boreal, NML= northern mid-latitude, NDR= northern dry, NST = northern subtropical, EQT = equatorial, SML=southern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

## 2. MMC model inputs and study catchments

### 2.1. The GHMs

The study capitalises on the recent release of historical GHM simulations through the second phase of the Inter Sectoral Impacts Model Intercomparison Project (ISIMIP2a) (http://www.isimip.org; (Gosling et al., 2017)). ISIMIP2a provides a consistent modelling framework that ensures any inconsistencies between model outputs are a result of differences in the GHMs' structures or parameters. However, the GHMs providing ISIMIP2a simulation products are not generally calibrated and are not accompanied by detailed information about the aleatory or epistemic uncertainties associated with each simulation, or the PDFs of model parameters from which it was generated. Consequently, this study is focused on the use of MMC to maximise predictive performance gain and not to minimise model or parameter uncertainty.

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

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

### 2.2. Study catchments and observed data

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

1- Catchments had to be larger than 100,000 km$^2$ to conform with the World Meteorological Organisation's definition of 'major' catchments (WMO, 2006). This ensured that the catchments were of sufficient size to accommodate the output resolution of the models (0.5° x 0.5°).

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

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

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

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

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

| No | GRDC Reference | River | Gauging Station | Total data length (years) | Catchment Area ($km^2$) | Hydro-belt |
|---|---|---|---|---|---|---|
| 1 | 2903430 | LENA | STOLB | 32 | 2,460,000 | BOR |
| 2 | 2906900 | AMUR | KOMSOMOLSK | 26 | 1,730,000 | BOR |
| 3 | 2909150 | YENISEI | IGARKA | 32 | 2,440,000 | BOR |
| 4 | 2912600 | OB | SALEKHARD | 39 | 2,949,998 | BOR |
| 5 | 2998510 | KOLYMA | KOLYMSKAYA | 28 | 526,000 | BOR |
| 6 | 2999910 | OLENEK | 7.5KM DOWNSTREAM OF MOUTH OF RIVER PUR | 39 | 198,000 | BOR |
| 7 | 4208150 | MACKENZIE RIVER | NORMAN WELLS | 30 | 1,570,000 | BOR |
| 8 | 4213550 | SASKATCHEWAN | THE PAS | 40 | 347,000 | BOR |
| 9 | 4213650 | ASSINIBOINE | HEADINGLEY | 40 | 153,000 | BOR |
| 10 | 4213680 | RED RIVER | EMERSON | 40 | 104,000 | BOR |
| 11 | 4213800 | WINNIPEG RIVER | SLAVE FALLS | 38 | 126,000 | BOR |
| 12 | 4214260 | CHURCHILL RIVER | ABOVE GRANVILLE FALLS | 36 | 228,000 | BOR |
| 13 | 4214520 | ALBANY RIVER | NEAR HAT ISLAND | 31 | 118,000 | BOR |
| 14 | 6970250 | NORTHERN DVINA | UST-PINEGA | 31 | 348,000 | BOR |
| 15 | 2180800 | YELLOW | HUAYUANKOU | 40 | 730,036 | NML |
| 16 | 4115200 | COLUMBIA | THE DALLES, OREG. | 40 | 613,830 | NML |
| 17 | 4127800 | MISSISSIPPI | VICKSBURG, MISS. | 37 | 2,964,252 | NML |
| 18 | 4143550 | ST.LAWRENCE | CORNWALL(ONTARIO), NEAR MASSENA, N.Y. | 40 | 773,892 | NML |
| 19 | 4207900 | FRASER RIVER | HOPE | 40 | 217,000 | NML |
| 20 | 6340110 | LABE | NEU-DARCHAU | 40 | 131,950 | NML |
| 21 | 6435060 | RHINE RIVER | LOBITH | 40 | 160,800 | NML |
| 22 | 6442600 | DANUBE | MOHACS | 29 | 209,064 | NML |
| 23 | 6972430 | NEVA | NOVOSARATOVKA | 40 | 281,000 | NML |

| | | | | | | |
|---|---|---|---|---|---|---|
| 24 | 6977100 | VOLGA | VOLGOGRAD POWER PLANT | 39 | 1,360,000 | NML |
| 25 | 6978250 | DON | RAZDORSKAYA | 38 | 378,000 | NML |
| 26* | 7222222 | YANGTZE | CUNTAN | 31 | 804,859 | NML |
| 27 | 4152450 | COLORADO | LEES FERRY, ARIZ. | 40 | 289,562 | NDR |
| 28 | 4356100 | SANTIAGO | EL CAPOMAL | 31 | 128,943 | NDR |
| 29 | 1834101 | NIGER | LOKOJA | 25 | 2,074,171 | NST |
| 30 | 1147010 | ZAIRE | KINSHASA | 40 | 3,475,000 | EQT |
| 31 | 3629000 | AMAZONAS | OBIDOS | 27 | 4,640,300 | EQT |
| 32 | 3630050 | XINGU | ALTAMIRA | 35 | 446,570 | EQT |
| 33 | 3650481 | RIO PARNAIBA | LUZILANDIA | 26 | 322,823 | SST |
| 34 | 3651805 | SAO FRANCISCO | MANGA | 37 | 200,789 | SST |
| 35 | 3667060 | PARAGUAI | PORTO MURTINHO (FB/DNOS) | 37 | 474,500 | SST |
| 36 | 5101200 | BURDEKIN | CLARE | 40 | 129,660 | SST |
| 37 | 1159100 | ORANJE | VIOOLSDRIF | 38 | 850,530 | SDR |
| 38 | 5410100 | COOPER CREEK | CALLAMURRA | 33 | 230,000 | SDR |
| 39 | 5101301 | FITZROY | THE GAP | 40 | 135,860 | SML |
| 40 | 5204250 | DARLING RIVER | LOUTH | 26 | 489,300 | SML |

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

## 3. Developing MMC solutions via Gene Expression Programming

### 3.1. GEP

GEP, which is detailed fully in Ferreira (2001, 2006), is an automated, machine learning algorithm that searches for optimal symbolic regression expressions to relate one or more series of input data to an independent, observed series. Unlike standard linear regression, where the expression structure is limited to the input and output variables, numerical constants (the regression coefficients) and addition and multiplication operators; GEP expressions can incorporate the full range of arithmetic operators, as well as, mathematical functions (which are selected by the modeller). This makes it possible for GEP to relate input and observed data series via non-linear expressions. GEP expressions are modular; they are comprised of component trees (hereafter simply termed components) which are themselves made up of *bases* - the individual inputs, functions, constants and operators that comprise the component. Components are aggregated together using mathematical operators (usually addition) to form more complex expressions that can be readily translated into standard algebraic equations (Figure 2).

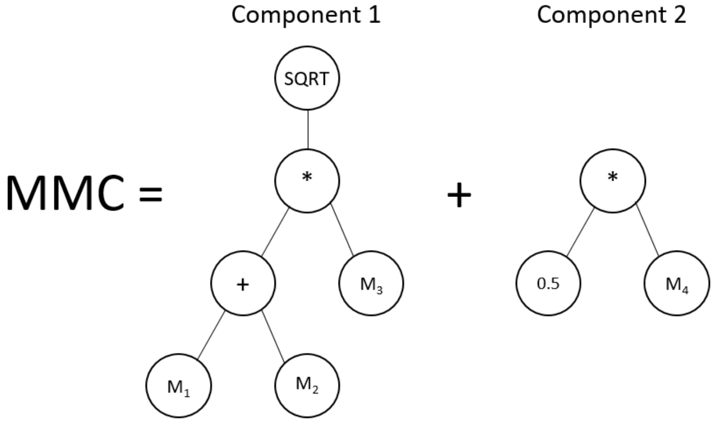


Figure 2. A GEP-based MMC solution (MMC) expressed as two components. The first component is made up of six bases and the second is made up of three. The MMC solution combines the four input models ($M_1$ to $M_4$) into an expression that includes a constant (0.5), operators (+ and *) and a non-linear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

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

The GEP algorithm is an example of an iterative evolutionary algorithm that evolves a set of expressions to relate the input data series to the observed series (Figure 3). The algorithm begins by creating a random set of expressions which are then evolved in subsequent iterations. The set of expressions that GEP develops in each iteration are analogous to the genetic codes of biological 'organisms'. Each organism's likelihood of survival to the next iteration of the algorithm is dependent upon the extent to which its genetic code (i.e. the GEP expression) optimises the fit between the input data series and the observed data according to a pre-determined metric (a process known as 'training'). In this study we use the ideal point error metric (Dawson et al., 2012) to determine fitness, (see Section 3.4), due to its incorporation of multiple error metrics into a single fitness measure. Each expression is then applied to an independent set of model inputs and the fit is validated to ensure that the expression can be generalised beyond the specific data from which it was learnt. If, at the end of an iteration, the best fitting expression is new, it is added to the candidate solution set which is output at the end of the GEP run. It is also preserved in the expression set (known as replication) whilst the remaining expressions are modified through adjustments to the bases in each component. These modifications can include mutation (where bases are randomly replaced with an alternative function, operator, input or constant) or transposition (where the arrangement of bases in the component is changed). In addition, entire components can be recombined by pairing them and exchanging their locations in the overall expression. The degree of modification allowed by each in any iteration is controlled by a rate set by the user. The number of iterations of the algorithm is also determined by a stopping point that is controlled by the user. This is usually a fixed number of iterations that is a large multiple of the number of data points in the observed series (i.e. to ensure adequate sampling of input data during training). Similarly, the user controls the complexity (equation size) of the expression by setting how many components it should include and the set of operators, functions and number of constants that can be included in the GEP expressions. The user settings applied in this study are provided in Table 2 and more detailed in Table S4.

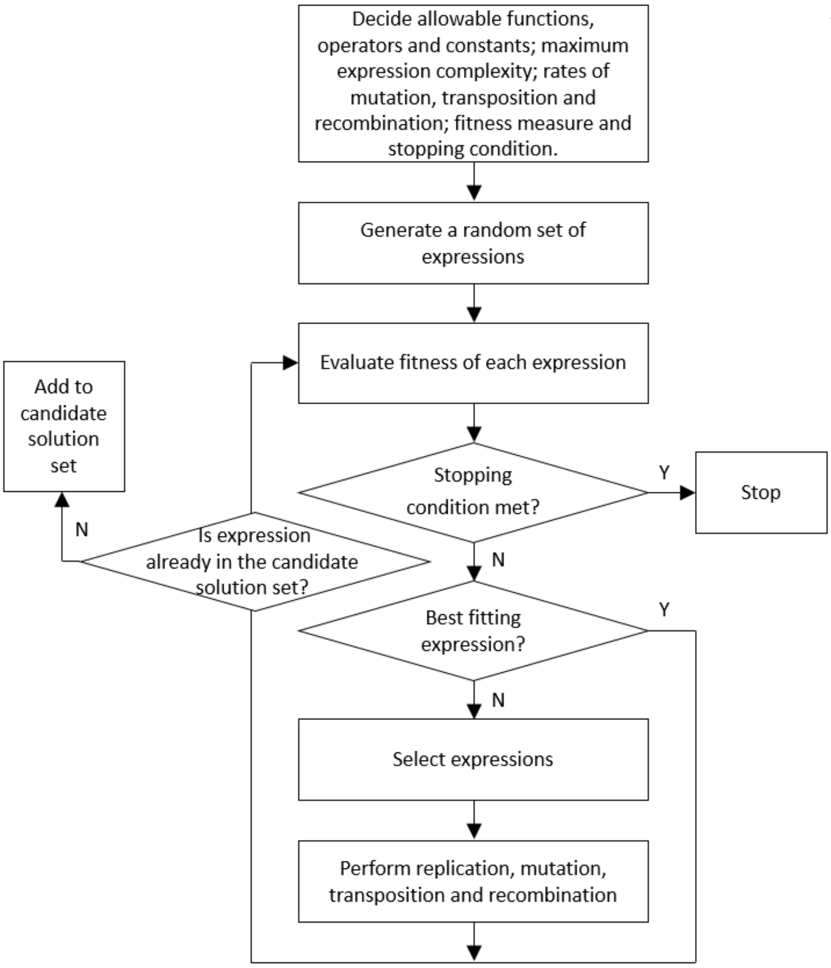

Decide allowable functions, operators and constants; maximum expression complexity; rates of mutation, transposition and recombination; fitness measure and stopping condition.
Generate a random set of expressions
Evaluate fitness of each expression
Stopping condition met?
Y
Stop
N
Best fitting expression?
Y
N
Select expressions
Perform replication, mutation, transposition and recombination
Is expression already in the candidate solution set?
N
Add to candidate solution set

Figure 3. The GEP algorithm.

Table 2. User settings for the GEP.

| Control | Setting used |
|---|---|
| Number of components | 3 |
| Allowable operators | +, -, *, / |
| Allowable functions | Sqrt, Exp, $x^2$, $x^3$, Natural Log, Sine, Cosine |
| Number of constants allowed per component | 2 |
| Mutation rate | 0.044 |

| | |
|---|---|
| Transposition rate | 0.1 |
| Recombination rate | 0.7 |
| Stopping condition | 100,000 iterations |
| Fit measure | IPE (see Section 3.4 below) |

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

Insights into the extent to which complex non-linear MMC methods offer benefits over simpler, linear MMC counterparts are gained by comparing the performance gains of GEP-MMC to that of a simpler, multiple linear regression (MLR) method. We use the bias corrected, ordinary least square (OLS) algorithm of Granger and Ramanathan (1984) which is unconstrained (the sum of the weights can exceed unity) as tests indicate improved performance when compared to non-bias-corrected and/or constrained alternatives (Arsenault et al., 2015).

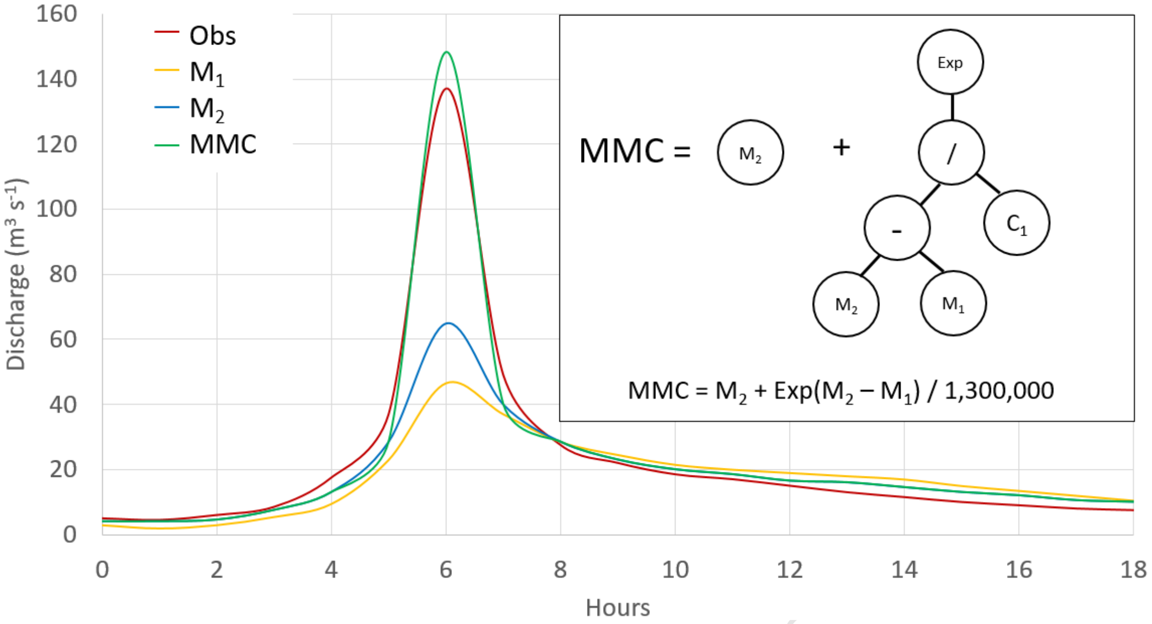


Figure 4. An example of a non-linear, GEP-based MMC solution in which the difference between two poorly performing models ($M_1$ and $M_2$) is used to correct the underestimation of peak discharge. $C_1$ in the second MMC component is a constant equal to 1,300,000.

### 3.2. Data splitting for GEP expression development

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

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

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

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

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

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

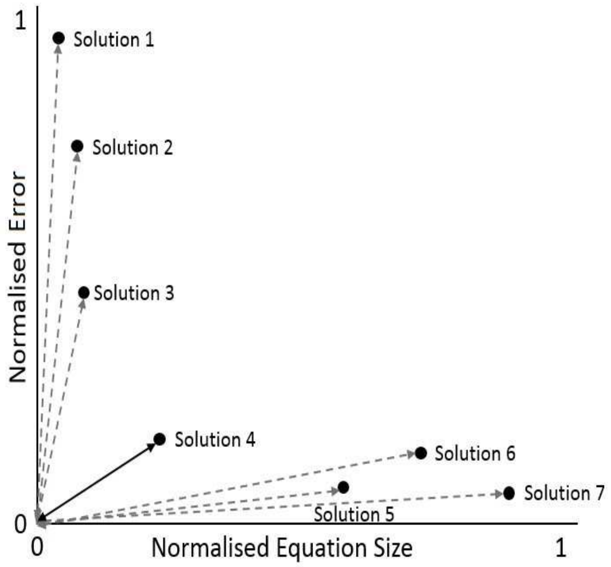


Figure 5. Selecting the GEP solution from a normalised fitness-equation space.
Solution 4 is selected because it has the smallest Euclidean distance from the origin.

### 3.4. Fit metrics

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

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

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

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

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

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

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

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

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

indicates that there is a 1.2 times reduction in performance of the MMC solution relative to its counterpart.

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

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

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

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

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

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

## 4. GHM, EM, MMC and MLR Performance

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

### 4.1. GHM performance

To assess the performance of the different GHMs, the fit of the monthly simulated and observed runoff time series was computed against the validation data for each model as well as the EM and the MMC solution in each of the 40 catchments. The IPE metrics for each catchment are reported in Table 3 and the spatial distribution of the best individual GHM and the best overall model is mapped in Figure 6. This reveals that WaterGAP2 is the GHM most able to improve upon the naïve model benchmark. It outperforms the other GHMs in

32 catchments, and also performs better than the EM for the majority of catchments (34). This finding is perhaps unsurprising given that this is the only calibrated model in the ensemble. However, it is noteworthy that the dominant performance of WaterGAP2 is considerably less evident in the boreal hydrobelt compared to the other hydrobelts. Here both PCRGLOBWB and DBH are the best performing individual models in 5 of the 14 catchments. Across the remaining hydrobelts, calibrated WaterGAP2 is out-performed by its uncalibrated counterparts in only 3 out of 26 catchments and these are spread across south sub-tropical, north dry belt and north mid-latitude without any apparent spatial pattern.

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

Table 3. IPE scores for individual GHMs, EM, MLR 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** | **MLR** | **MMC** |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | LENA | BOR | 1.58 | 2.04 | 1.42 | 1.51 | -1.22 | 1.15 | -1.56 | **-2.00** |
| **2** | AMUR | BOR | 3.06 | 1.91 | 1.33 | 1.34 | 1.17 | 1.07 | -1.34 | **-1.49** |
| **3** | YENISEI | BOR | 1.18 | -1.54 | 1.25 | -1.54 | -1.72 | -1.69 | -2.03 | **-2.33** |
| **4** | OB | BOR | 8.42 | 4.75 | 13.92 | 2.61 | 2.50 | 3.53 | -1.30 | **-1.32** |
| **5** | KOLYMA | BOR | -1.23 | 1.10 | 1.18 | 1.27 | 2.30 | -1.19 | -1.21 | **-2.38** |
| **6** | OLENEK | BOR | **-1.47** | 6.32 | 12.45 | 17.70 | 3.94 | 8.12 | 4.05 | -1.15 |
| **7** | MACKENZIE RIVER | BOR | 4.50 | 1.85 | 3.37 | -1.30 | 1.07 | -1.39 | **-2.19** | -1.33 |
| **8** | SASKATCHEWAN | BOR | 61.42 | 5.75 | 27.03 | 8.16 | 1.43 | 8.97 | **-1.22** | 1.03 |
| **9** | ASSINIBOINE | BOR | 384.84 | 44.46 | 512.25 | 28.94 | 1.57 | 85.79 | **-1.01** | 1.06 |
| **10** | RED RIVER | BOR | 6.56 | 1.62 | 4.83 | 2.12 | 1.52 | 2.77 | -1.20 | **-1.25** |
| **11** | WINNIPEG RIVER | BOR | 24.16 | 4.85 | 5.05 | **1.55** | 1.67 | 2.29 | 1.71 | 1.63 |
| **12** | CHURCHILL RIVER | BOR | 297.53 | 50.12 | 32.22 | 25.65 | 3.60 | 17.08 | 3.94 | **3.10** |
| **13** | ALBANY RIVER | BOR | 2.82 | -1.03 | 2.76 | -1.33 | 1.73 | -1.22 | **-2.50** | -1.67 |
| **14** | NORTHERN DVINA | BOR | 1.48 | -1.04 | 2.14 | -1.15 | -1.52 | -1.54 | -2.25 | **-2.27** |
| **15** | YELLOW | NML | 23.41 | 5.50 | 7.42 | 44.87 | 1.49 | 9.75 | 2.04 | **1.16** |
| **16** | COLUMBIA | NML | 4.25 | 2.12 | 3.11 | 1.75 | -1.11 | -1.28 | **-1.58** | -1.20 |
| **17** | MISSISSIPPI | NML | 4.98 | -1.56 | 1.07 | 1.70 | -1.89 | 1.16 | **-2.50** | -2.04 |
| **18** | ST.LAWRENCE | NML | 375.18 | 75.36 | 56.89 | 13.97 | 7.09 | 31.61 | 2.74 | **2.47** |
| **19** | FRASER RIVER | NML | 1.18 | 2.53 | 4.06 | 1.15 | 1.16 | 1.30 | **-1.78** | -1.61 |
| **20** | LABE | NML | 6.70 | 4.11 | 2.98 | 7.67 | -1.47 | 3.10 | **-1.58** | -1.45 |
| **21** | RHINE RIVER | NML | 2.63 | 3.29 | 1.50 | 1.39 | -1.96 | 1.15 | **-3.20** | -2.50 |
| **22** | DANUBE | NML | 4.02 | 2.72 | 1.25 | 2.07 | -1.89 | -1.08 | **-3.12** | -2.22 |
| **23** | NEVA | NML | 83.42 | 25.58 | 12.19 | 8.94 | 2.42 | 4.74 | 1.40 | **1.09** |
| **24** | VOLGA | NML | 6.80 | 2.79 | 1.89 | -1.35 | -1.75 | 1.52 | **-2.17** | -2.00 |
| **25** | DON | NML | 83.47 | 39.91 | 58.79 | 100.12 | 1.54 | 37.14 | 1.28 | **1.23** |
| **26** | YANGTZE | NML | -2.44 | -1.10 | -1.05 | 2.81 | -3.03 | -1.15 | -3.71 | **-4.17** |
| **27** | COLORADO | NDR | 52.90 | 2.50 | 12.10 | 8.50 | 4.59 | 6.44 | 2.51 | **2.22** |
| **28** | SANTIAGO | NDR | 15.13 | 8.26 | 3.84 | 14.97 | 1.35 | 7.33 | 1.60 | **1.16** |
| **29** | NIGER | NST | 9.67 | 10.65 | 10.04 | 3.61 | -1.37 | 4.86 | **-1.99** | -1.79 |
| **30** | ZAIRE | EQT | 8.28 | 5.92 | 3.89 | 2.47 | 1.78 | 2.40 | **-1.05** | 1.42 |
| **31** | AMAZONAS | EQT | 2.05 | 1.46 | 2.60 | 3.44 | -1.09 | 1.27 | -1.75 | **-1.85** |
| **32** | XINGU | EQT | 5.89 | 4.65 | 4.89 | 1.12 | 1.16 | 2.65 | **-1.16** | 1.04 |
| **33** | RIO PARNAIBA | SST | 48.77 | 70.84 | 63.41 | 8.39 | 1.46 | 25.41 | **-2.52** | -2.27 |
| **34** | SAO FRANCISCO | SST | 4.81 | 3.48 | 1.89 | 2.25 | -1.64 | 1.94 | -1.65 | **-1.92** |
| **35** | PARAGUAI | SST | 136.88 | 153.69 | 108.09 | 98.44 | **8.00** | 78.53 | 8.78 | 8.51 |
| **36** | BURDEKIN | SST | 6.87 | 1.44 | 3.13 | 2.03 | 1.65 | 2.92 | -1.19 | **-1.35** |
| **37** | ORANJE | SDR | 83.15 | 7.09 | 81.10 | 46.42 | 2.26 | 31.15 | 3.58 | **2.04** |
| **38** | COOPER CREEK | SDR | 6993.0 | 149.00 | 2578.0 | 625.00 | 107.00 | 2089.0 | 124.58 | **20.05** |
| **39** | FITZROY | SML | 641.17 | 52.61 | 447.46 | 270.32 | 38.47 | 290.00 | 86.85 | **30.64** |
| **40** | DARLING RIVER | SML | 200.58 | 6.95 | 92.30 | 35.20 | -1.54 | 41.93 | 591.22 | **-1.64** |

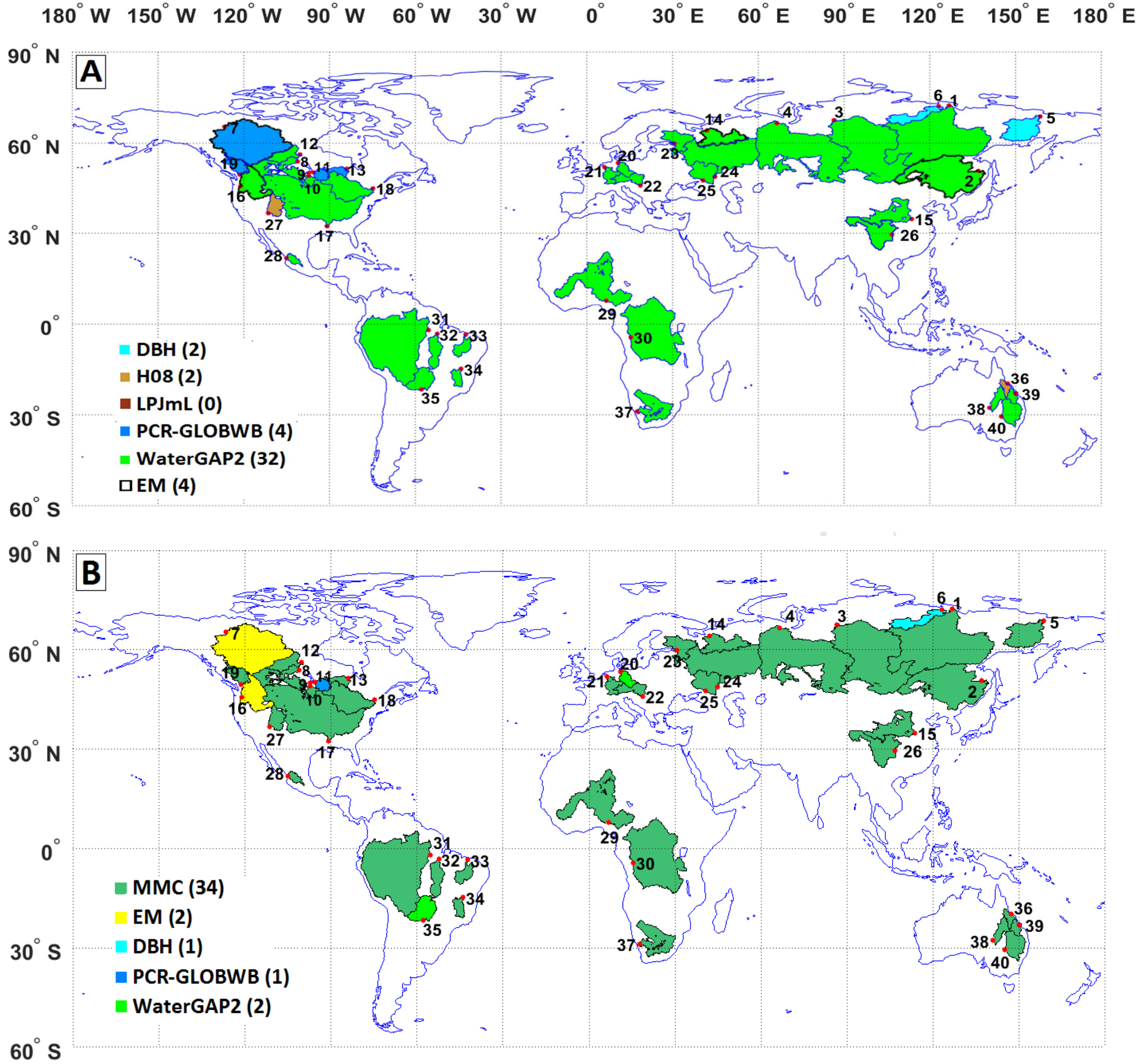


Figure 6. The best performing individual GHM (A); four catchments (2, 7, 14 and 16) where the EM outperforms the individual models have borders in bold black lines (in these cases the catchment is still shaded according to the best performing individual GHM). The best performing overall model/MMC (B);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

### 4.2. EM Performance

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

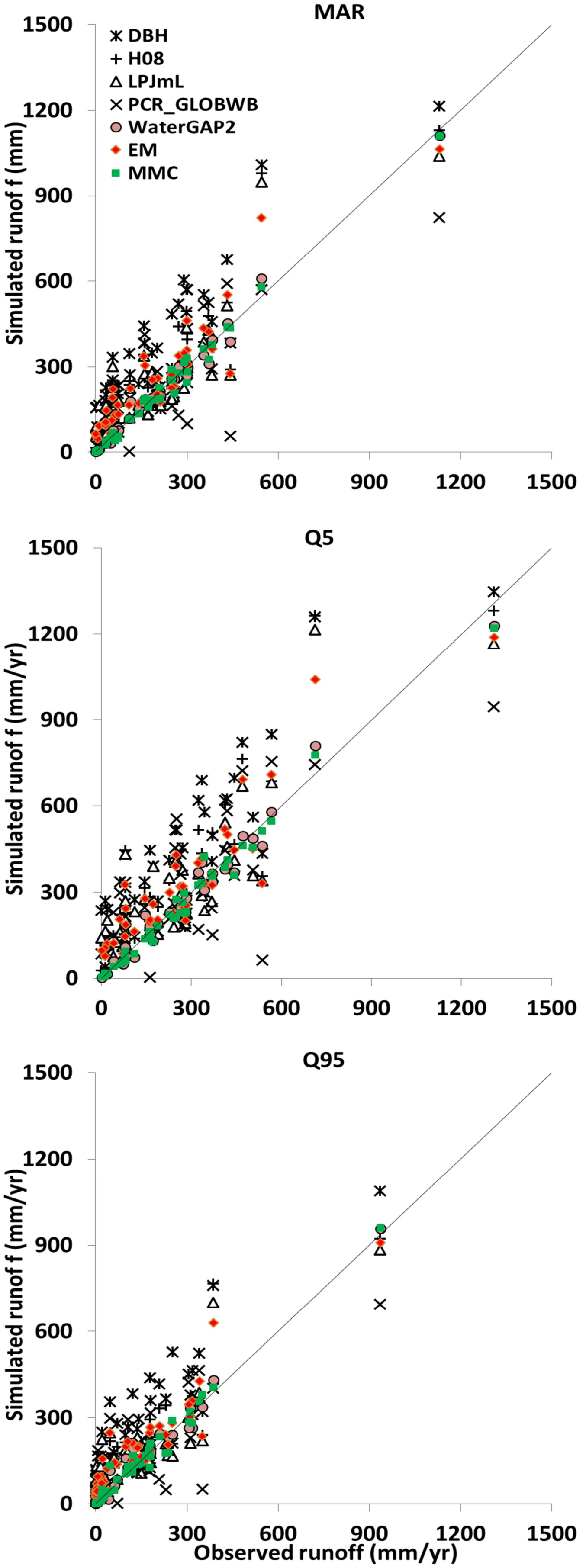


Figure 7. Plots of observed versus simulated runoff for each GHM, the EM and the MMC for mean annual runoff, Q5 and Q95.

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

**4.3. MMC and MLR Performance**

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

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

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

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

When summarised by hydrobelt (Table 5), it is evident from the median $\mathrm{MMC_{PG}}$ score that MMC solutions generally deliver substantial improvements over their EM and GHM counterparts in all hydrobelts. The MMC performance gain is largest against the EM than the best-performing GHM in all hydrobelts. It is always several orders of magnitude greater and reflects the limiting impact that positive biases in GHM outputs have on the performance of the EM. When compared against the best-performing GHM, the median MMC performance gain is lowest in the northern dry hydrobelt (-24%) and highest in southern sub-tropical (-254%) and the boreal (-55%) hydrobelts. Northern mid-latitude catchments see performance gains of -32%. However, it is important to acknowledge that whilst IPE facilitates comparison of MMCs across hydrobelts, the robustness of the comparison is limited by the lower proportion of the total hydrobelt area represented by catchments in NDR, SST, SDR and SML hydrobelts. Addressing this will require data from a greater number of study catchments to be made available, with the temporally-extensive runoff records needed to support robust application of the machine learning algorithms that underpin MMC development. This highlights the importance of improving data collection systems in these hydrobelts in particular.

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

Perhaps surprisingly, MLR outperforms GEP-based MMC in approximately one third (n = 15) of the catchments and, whilst the magnitude of the additional performance achieved by MLR is generally small, occasionally MLR does outperform GEP-based MMC by a substantial margin (e.g. the Mackenzie River). The number of catchments in which MLR achieves a large performance gain ($MLR_{PG}$ >50%) over the best GHM or the EM (Table 4) is almost the same as that of GEP-based MMC (21 catchments and 20 respectively). However, MLR fails to perform as well as either in 12 catchments – double the number of catchments in which this occurs with GEP-based MMC. Moreover, where performance loss occurs, its average magnitude is greater for MLR than GEP-based MMC (median loss of 77% compared to 7%). It is noteworthy the three catchments in which GEP-based MMC delivers the greatest performance gain (Cooper Creek, Darling River and Fitzroy river) are the three in which MLR performs worst. This indicates that linear MMC methods may be poorly suited to the non-linear challenge of MMC in arid and semi-arid hydrobelts, although the small number of catchments in these hydrobelts requires caution in drawing general conclusions (Table 5).

Aggregated across hydrobelts, inconsistency in the relative performance gain of GEP-based MMC versus MLR remains. The Boreal (BOR, n=14) and Northern Mid Latitude (NML, n=12) hydrobelts are the only ones with a sufficiently large number of catchments to support general interpretations but it is nonetheless difficult to generalise (Figure 8). Whilst in both of these hydrobelts MLR has a small, mean performance gain over GEP-based MMC, the number of catchments in which either method outperforms the other is similar and the magnitude of the relative performance gain varies substantially from catchment to catchment – with each method achieving order-of-magnitude relative performance gains over the other in certain catchments.

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

| No | River | Hydro-belt | MMC IPE score | Best performing model (GHM or EM) and IPE score | MMC$_{PG}$ (%) | Rank | MMC solution separated into its GEP-expression components. MMC = C1 + C2 + C3. Components are ordered according to their explanatory power as assessed by their IPE. | Eqn. size[1] | MLR IPE score | MLR$_{PG}$ (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| 38 | COOPER CREEK | SDR | 20.05 | WaterGAP2<br>IPE = 107.00 | -8674 | 1 | C1: 0<br>C2: + (-0.143) * H08 * (WaterGAP2 +1) * cos(cos(WaterGAP2))<br>C3: + 0.436*H08*sqrt WaterGAP2 | 18 | 124.58 | 438 |
| 40 | DARLING RIVER | SML | -1.64 | WaterGAP2<br>IPE = -1.54 | -1350 | 2 | C1: 0.174*H08^2/DBH<br>C2: + (-0.06/DBH)<br>C3: + H08/DBH | 11 | 591.22 | 46041 |
| 39 | FITZROY | SML | 30.64 | WaterGAP2<br>IPE = 38.47 | -784 | 3 | C1: sin(H08/-4.91)<br>C2: + WaterGAP2<br>C3: + sin((LPJmL - sqrt DBH-8.45)*(WaterGAP2+H08)/( DBH *PCRGLOBWB)) | 20 | 86.58 | 4837 |
| 4 | OB | BOR | -1.32 | WaterGAP2<br>IPE = 2.50 | -581 | 4 | C1: 2*DBH/(log(sin H08)+6247.9)<br>C2: +  sqrt H08<br>C3: + WaterGAP2/H08^2 | 15 | -1.30 | -580 |
| 33 | RIO PARNAIBA | SST | -2.27 | WaterGAP2<br>IPE = 1.46 | -574 | 5 | C1: 3.695<br>C2: + 0.625*((cos(0.227/H08))^6*(log(WaterGAP2))^4)<br>C3: + 1.472 / (log(1/PCRGLOBWB) – 1.08396) | 20 | -2.52 | -597 |
| 36 | BURDEKIN | SST | -1.35 | H08<br>IPE = 1.44 | -479 | 6 | C1: 0<br>C2: + sqrt H08<br>C3: + H08 * sin(log(log(PCRGLOBWB/2))) | 10 | -1.19 | -462 |
| 10 | RED RIVER | BOR | -1.25 | WaterGAP2<br>IPE = 1.52 | -478 | 7 | C1: H08*WaterGAP2/10.045<br>C2: + sin PCRGLOBWB^3/(DBH^3*H08+H08-LPJmL-5.44)<br>C3: + sin(cos(WaterGAP2))^3 | 23 | -1.20 | -472 |
| 19 | FRASER RIVER | NML | -1.61 | PCRGLOBWB<br>IPE = 1.15 | -477 | 8 | C1: 0.33*DBH*sqrt(log(PCRGLOBWB))<br>C2: + cos((H08+1.63)/LPJmL)+8.12<br>C3: + cos H08 | 17 | -1.78 | -493 |
| 18 | ST. LAWRENCE | NML | 2.47 | WaterGAP2<br>IPE = 7.09 | -462 | 9 | C1: 23.04<br>C2: + 0.67*sqrt WaterGAP2 * cos(sqrt WaterGAP2+ 1.42/H08)<br>C3: + 1.1*sqrt(DBH/PCRGLOBWB) | 19 | 2.74 | -435 |
| 2 | AMUR | BOR | -1.49 | EM<br>IPE = 1.07 | -356 | 10 | C1: 2.534*(DBH-H08-LPJmL-LPJmL/H08)/PCRGLOBWB<br>C2: + WaterGAP2-4.33<br>C3: + sin DBH | 18 | -1.34 | -450 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 23 | NEVA | NML | 1.09 | WaterGAP2<br>IPE =2.42 | -133 | 11 | C1: PCRGLOBWB<br>C2: + log(DBH^3)<br>C3: + WaterGAP2/PCRGLOBWB + 0.5*log(log(WaterGAP2)) | 13 | 1.40 | -102 |
| 5 | KOLYMA | BOR | -2.38 | DBH<br>IPE = -1.23 | -<br>-114 | 12 | C1: DBH<br>C2: + sqrt LPJmL<br>C3: + DBH*(-2.74*DBH+LPJmL-3.133)/WaterGAP2 | 14 | -1.21 | 2 |
| 26 | YANGTZE | NML | -4.17 | WaterGAP2<br>IPE = -3.03 | -108 | 13 | C1: WaterGAP2<br>C2: + sqrt LPJmL<br>C3: + cos(PCRGLOBWB +0.039*H08*PCRGLOBWB/DBH) | 13 | -3.71 | -63 |
| 1 | LENA | BOR | -2.00 | WaterGAP2<br>IPE = -1.22 | -78 | 14 | C1: WaterGAP2-sqrt DBH<br>C2: + LPJmL/(2*LPJmL/WaterGAP2^2+5.575)<br>C3: + (-0.626) | 15 | -1.56 | -34 |
| 31 | AMAZONAS | EQT | -1.85 | WaterGAP2<br>IPE = -1.09 | -75 | 15 | C1: WaterGAP2<br>C2: + (H08-DBH+LPJmL+0.77)* (WaterGAP2-LPJmL- 0.77)/(PCRGLOBWB+24.9)<br>C3: + (-2.98) | 19 | -1.75 | -66 |
| 14 | NORTHERN DVINA | BOR | -2.27 | EM<br>IPE= -1.54 | -70 | 16 | C1: WaterGAP2<br>C2: + PCRGLOBWB<br>C3: + (-9.29) | 3 | -2.25 | -73 |
| 3 | YENISEI | BOR | -2.32 | WaterGAP2<br>IPE = -1.72 | -58 | 17 | C1: WaterGAP2<br>C2: + (-0.742)<br>C3: + 7.0*sin(sqrt H08) | 7 | -2.3 | -31 |
| 9 | ASSINIBOINE | BOR | 1.06 | WaterGAP2<br>IPE = 1.57 | -51 | 18 | C1: WaterGAP2^2<br>C2: + sin(0.5*log(0.268*H08+cosWaterGAP2/WaterGAP2+0.003))<br>C3: + 0.064 | 17 | -1.01 | -458 |
| 21 | RHINE RIVER | NML | -2.50 | WaterGAP2<br>IPE = -1.96 | -51 | 19 | C1: WaterGAP2<br>C2: + 5.813<br>C3: + (-0.153)*H08 | 5 | -3.20 | -123 |
| 12 | CHURCHILL RIVER | BOR | 3.10 | WaterGAP2<br>IPE = 3.60 | -50 | 20 | C1: WaterGAP2<br>C2: + sin PCRGLOBWB<br>C3: + cos(sqrt H08) | 6 | 3.94 | -66 |
| 29 | NIGER | NST | -1.79 | WaterGAP2<br>IPE = -1.37 | -41 | 21 | C1: 0.062* log(DBH)^4*(cos(4.647/PCRGLOBWB))^6<br>C2: + cos(sin LPJmL/WaterGAP2)<br>C3: + 0.556 | 17 | -1.99 | -62 |
| 8 | SASKATCHEWAN | BOR | 1.03 | WaterGAP2<br>IPE = 1.43 | -40 | 22 | C1: WaterGAP2<br>C2: + (cos(cos(DBH + log WaterGAP2 + 0.31))-sin(sqrt PCRGLOBWB^3))^3<br>C3: + -sin((log LPJmL^3)/8-sin(cos(0.401*LPJmL)+1.723) | 29 | -1.22 | -464 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | ZAIRE | EQT | 1.42 | WaterGAP2<br>IPE = 1.78 | -36 | 23 | C1: WaterGAP2<br>C2: + cos(sqrt DBH)<br>C3: + cos(sqrt DBH) | 7 | -1.05 | -483 |
| 15 | YELLOW | NML | 1.16 | WaterGAP2<br>IPE = 1.49 | -33 | 24 | C1: sqrt(DBH)<br>C2: + DBH*WaterGAP2^5/4/(DBH^2*WaterGAP2-0.043*PCRGLOBWB)<br>C3: + (sin WaterGAP2)^2*sin(sqrt(PCRGLOBWB+DBH)) | 26 | 2.04 | 55 |
| 13 | ALBANY RIVER | BOR | -1.66 | PCRGLOBWB<br>IPE = -1.33 | -33 | 25 | C1: PCRGLOBWB<br>C2: + log(0.106*DBH)<br>C3: + log(0.041*DBH) | 9 | -2.50 | -116 |
| 22 | DANUBE | NML | -2.22 | WaterGAP2<br>IPE = -1.89 | -32 | 26 | C1: WaterGAP2<br>C2: + DBH/H08- H08/(PCRGLOBWB-1)<br>C3: + 7.93/H08 | 13 | -3.12 | -122 |
| 25 | DON | NML | 1.23 | WaterGAP2<br>IPE = 1.54 | -32 | 27 | C1: WaterGAP2<br>C2: + 1<br>C3: + (-0.325)*WaterGAP2 | 5 | 1.28 | -26 |
| 34 | SAO FRANCISCO | SST | -1.92 | WaterGAP2<br>IPE = -1.64 | -29 | 28 | C1: sqrt(WaterGAP2)<br>C2: + 1.46*(PCRGLOBWB+WaterGAP2-5.75)/log(PCRGLOBWB)<br>C3: + cos(H08/LPJmL) | 16 | -1.65 | -2 |
| 27 | COLORADO | NDR | 2.22 | H08<br>IPE = 2.50 | -29 | 29 | C1: log(DBH)<br>C2: + log(PCRGLOBWB)<br>C3: + WaterGAP2/PCRGLOBWB | 7 | 2.51 | 1 |
| 24 | VOLGA | NML | -2.00 | WaterGAP2<br>IPE = -1.75 | -23 | 30 | C1: WaterGAP2-0.978<br>C2: + 3.35/DBH<br>C3: + 0.999/LPJmL | 9 | -2.17 | -41 |
| 37 | ORANJE | SDR | 2.04 | WaterGAP2<br>IPE = 2.26 | -22 | 31 | C1: WaterGAP2<br>C2: + 0.808<br>C3: + (-0.672) | 3 | 3.58 | 131 |
| 28 | SANTIAGO | NDR | 1.16 | WaterGAP2<br>IPE = 1.35 | -19 | 32 | C1: sin(LPJmL^2*(0.319-LPJmL/DBH))/DBH<br>C2: + WaterGAP2<br>C3: + sin((sin(((sin((LPJmL))-(((LPJmL)/(WaterGAP2))^3))^2))-(WaterGAP2))) | 24 | 1.60 | 25 |
| 17 | MISSISSIPPI | NML | -2.04 | WaterGAP2<br>IPE = -1.89 | -14 | 33 | C1: WaterGAP2<br>C2: + (log(WaterGAP2^3)-WaterGAP2)/PCRGLOBWB<br>C3: + (-1.70-DBH)/PCRGLOBWB | 13 | -2.50 | -62 |
| 32 | XINGU | EQT | 1.04 | WaterGAP2<br>IPE = 1.16 | -9 | 34 | C1: WaterGAP2<br>C2: + (-0.494)<br>C3: + (-0.204)*LPJmL/sqrt WaterGAP2 | 8 | -1.16 | -428 |

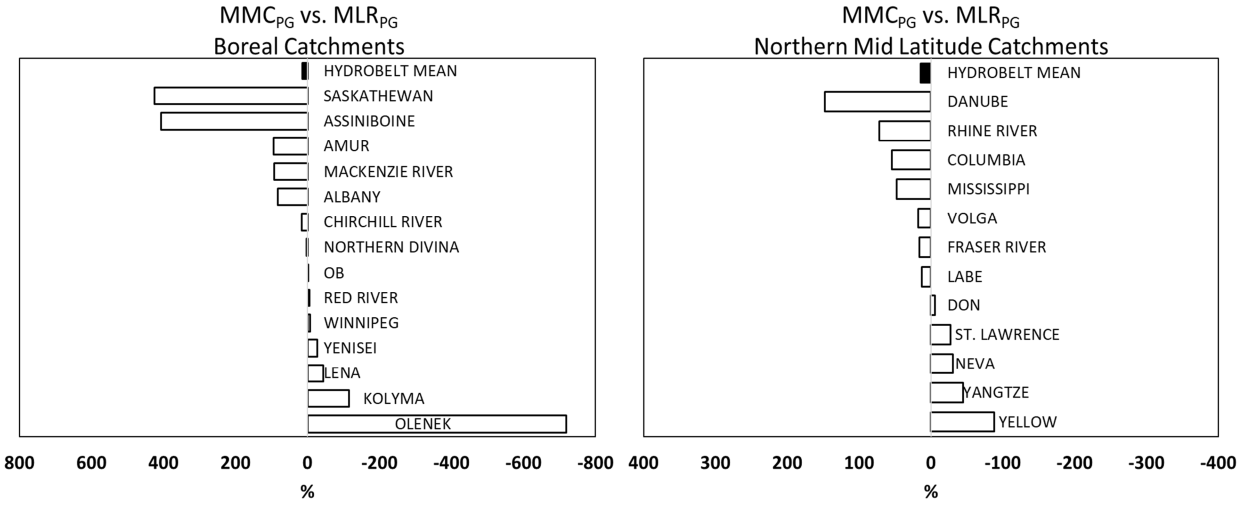
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | LABE | NML | -1.45 | WaterGAP2<br>IPE = -1.47 | 2 | 35 | C1: WaterGAP2<br>C2: + 4.32/DBH<br>C3: + 0.962*sin((DBH-H08)/PCRGLOBWB+ cos(0.15*WaterGAP2)) | 17 | -1.58 | -11 |
| 7 | MACKENZIE RIVER | BOR | -1.33 | EM<br>IPE = -1.39 | 5 | 36 | C1: PCRGLOBWB<br>C2: + 0.107*DBH<br>C3: + (-0.978) | 5 | -2.19 | -88 |
| 16 | COLUMBIA | NML | -1.20 | EM<br>IPE = -1.28 | 7 | 37 | C1: WaterGAP2<br>C2: + sin(cos(LPJmL)^3)^2*sin(PCRGLOBWB*cos(3.78*PCRGLOBWB))<br>C3:+ exp(cos(cos(LPJmL)*sin(WaterGAP2)))* sin(0.479+0.166*WaterGAP2) | 28 | -1.58 | -47 |
| 11 | WINNIPEG RIVER | BOR | 1.63 | PCRGLOBWB<br>IPE = 1.55 | 8 | 38 | C1: WaterGAP2<br>C2: + H08/DBH<br>C3: + (-4.91+log(PCRGLOBWB)) | 8 | 1.71 | 16 |
| 35 | PARAGUAI | SST | 8.51 | WaterGAP2<br>IPE = 8.00 | 19 | 39 | C1: WaterGAP2<br>C2: log(9.84/LPJmL)<br>C3: 0.99- (LPJmL/(PCRGLOBWB-((LPJmL+WaterGAP2)/945.48))) | 16 | 8.78 | 77 |
| 6 | OLENEK | BOR | -1.15 | DBH<br>IPE = -1.47 | 33 | 40 | C1: -sin(0.004* LPJmL^2*PCRGLOBWB-LPJmL+9.04)<br>C2: + PCRGLOBWB/(-0.31*DBH^2*cosec(PCRGLOBWB) -7.71)<br>C3: + WaterGAP2 | 22 | 4.05 | 752 |

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

Table 5. Median MMC performance gain ($MMC_{PG}$) for each hydrobelt, for the validation data set. Figures in bold highlight where each of the methods performs best.

| Hydrobelt | No. of catchments | Median PG over best-performing GHM (%) | | Median PG over EM (%) | |
|---|---|---|---|---|---|
| | | MMC | MLR | MMC | MLR |
| BOR | 14 | -55 | -80 | **-415** | -355 |
| NML | 12 | -32 | -62 | -434 | -467 |
| NDR | 2 | **-24** | 13 | **-520** | -483 |
| NST | 1 | -41 | -62 | -764 | -785 |
| EQT | 3 | -36 | -428 | -161 | -445 |
| SST | 4 | **-254** | -232 | -1698 | -1701 |
| SDR | 2 | **-4348*** | 955 | **-104900*** | -99596 |
| SML | 2 | **-1067*** | 25439 | **-703068*** | -676561 |

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


$MMC_{PG}$ vs. $MLR_{PG}$ Boreal Catchments

$MMC_{PG}$ vs. $MLR_{PG}$ Northern Mid Latitude Catchments

Figure 8. Relative performance gain of GEP-based MMC versus MLR for BOR and NML catchments. A negative % value indicates the MLR is out-performed by GEP-based MMC and a positive value indicates the opposite.

# 5. Discussion

## 5.1. Interpretability of MMC solutions

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

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

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

Thirdly, we acknowledge that any attempt to weight models may be viewed by some as futile so long as the current generation of GHMs (or any model) are far from being empirically adequate for purpose (Stainforth et al., 2007). Other work has shown that the GHMs applied here are imperfect (Zaherpour et al., 2018b) and in this sense it can be argued that applying weights to any type of model that is known to contain errors is counter-intuitive because the errors in even well performing models will be weighted inherently in the approach. Where weights are applied in a simple manner (e.g. each GHM output is multiplied by a single coefficient), this is certainly the case. However, a key advantage of GEP is that it develops more complex schemes in which the products of more than one model can be weighted (e.g. the difference in performance between two or more models at different hydrological response ranges - see Figure 4). Intuitively, this gives it an advantage over MMC methods that have a fixed structure, such as MLR, because it offers the potential to exploit the characteristic differences in the capabilities and/or failings of the models that are combined: allowing GEP-based MMC solutions to deliver performance gains based on non-linear adjustments made to the characteristic differences between each model input. Where GEP is concerned, it can be argued that it is its counter-intuitive ability to exploit model failings in the MMC solutions that provides a strong argument for using it rather than simple weighting – especially where the objective is to combine models known to be lacking with respect to their empirical fitness-for-purpose.

Current model combination approaches in hydrological modelling include simple model averaging (Arsenault et al., 2015; Cloke and Pappenberger, 2009) and complex weighting approaches (Ajami et al., 2006; Arsenault et al., 2015; Shamseldin et al., 2007) comprising machine learning algorithms, as described here. The data we present, and the above critique, indicate that on a global scale MMC based on machine learning algorithms may offer little in the way of average performance gain over simpler, linear methods such as MLR. However, at the catchment level, and in certain hydrobelts, there can be significant differences in their relative performance. This suggests that the adoption of a stepwise approach to multi-model combination is prudent in which simple, linear methods are attempted first and, where they fail to deliver adequate performance gain, non-linear machine learning approaches are subsequently employed.

The evidence we present also indicates that the application of complex weighting schemes via machine learning algorithms can make it difficult to understand the reasons behind the relative performance of individual models. For example, it is difficult to understand the relative weightings of individual models (i.e. which models are weighted more/less than others, e.g. see the solution for the Columbia river in Table 4), let alone why those weights have been applied (e.g. are the weights applied due to a model's ability to simulate high flows well?) and why some models are excluded altogether. Therefore, whilst we have demonstrated that generally a complex MMC solution can perform better than the EM, the interpretability of the MMC can become limited. This suggests that a more interpretable, but still intelligent, approach to model combination is needed. An alternative approach would be to follow the framework described by Krysanova et al. (2018) for global- and catchment models. They recommend first evaluating model performance for several hydrological variables over various time periods, as in a classical model evaluation (Zaherpour et al., 2018b), and if performance is considered to be acceptable then the models can be weighted, otherwise they are excluded from the ensemble. Although there is value in the approach, no specific recommendations are provided on how to weight the models, other than weighting based upon model performance. In addition, identification of a threshold for "good performance" is not straightforward, and the approach rejects, *a priori*, poorly performing models. One of the arguable advantages of GEP is that it can exploit the characteristic error patterns of poorly performing models by using them as mechanisms to adjust other models through the MMC development, as we have demonstrated. Merging a more interpretable MMC approach with that of Krysanova et al. (2018) may be a pragmatic way forwards for future model combination and weighting studies.

**5.2. MMC does not always deliver optimal solutions**

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

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

Secondly, deficiencies in our error-complexity trade-off method to select the final MMC solution from the candidate set (Figure 5) could be a factor. Whilst the trade-off is necessary to limit the complexity of the final GEP-based MMC solution, it does mean that the best performing MMC solution in the candidate set can be overlooked in favour of a simpler, lower-performing counterpart. This means that although the GEP algorithm may have developed a candidate MMC solution that outperforms either the EM or best-performing GHM, if its complexity is high relative to other solutions, it will not be selected as the final MMC solution. To check whether this is a factor behind MMC's poor performance in the six catchments, the best performing solutions from GEP's candidate solution set, irrespective of their complexity, are compared to the EM and best performing GHM in each catchment (Table S7). In the Mackenzie and Labe catchments, the best-performing MMC solution from the candidate set does outperform both the EM and the best-performing GHM. In the Paraguai catchment it equals it. However, in the Columbia, Olenek and Winnipeg even the best-performing candidate MMC solution fails to outperform the best individual GHM and the reasons for MMC failure remain unclear – particularly in the Columbia and Winnipeg catchments.

Thirdly, GEP's user settings (Table 2) are fundamental controls of the complexity of the MMC solutions that will be produced. The number of components included sets a 'baseline' for the solution complexity, whilst the number of constants and allowable function set will strongly influence the nature and complexity of its inherent non-linearity. Where these user settings encourage solutions whose complexity is excessive for the nature of the combination problem at hand, 'redundancy' in the MMC solutions is likely. This may be achieved simply (i.e. the assignment of a constant of value zero to component 1 in the solution for Burdekin, Table 4), or through complex equations that deliver insignificant outputs. Applying different user settings for the algorithm may to some extent solve this problem – but it is impossible to know, *a priori* what the most suitable settings might be. As an alternative, allowing the algorithm more iterations (we applied 100,000 in this study) might provide the algorithm with the opportunity to find improved solutions based on the development of lower-complexity MMC equations. Ongoing research by the authors is exploring the impact of applying different settings (specifically a lower number of MMC components) on the performance of the MMC approach (Zaherpour et al., 2018a).

### 5.3. Accounting for and presenting uncertainty in MMC development

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

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

## 6. Conclusions

This study has, for the first time, applied a set of 'intelligently defined' weights to a state-of-the-art ensemble of global-scale hydrological models. The GEP-based MMC applied, is shown to employ a diverse array of linear and non-linear adjustments to exploit information in runoff estimates from the individual GHMs. The result is that in 34 catchments (85%) the MMC performs better than the best performing GHM and EM with the median performance gain over a naïve benchmark model being 45% across all 40 catchments. The EM performs better than individual GHMs in only 10% (4) of our catchments. However, is cannot be assumed that complex, machine-learning MMC methods will deliver performance gains over simpler approaches, such as MLR. Indeed, it this study we find the relative performance of GEP-based MMC versus simpler MLR varies hugely from catchment-to-catchment and hydrobelt-to-hydrobelt and that MLR out-performs GEP-based MMC in around a third of the study catchments.

Despite the good performance of MMC 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 GEP-based MMC solution, with GHMs' lack of insufficient 'raw material' for the GEP algorithm to exploit, or deficiency in our error-complexity trade-off method for selecting final MMC being potentially responsible for this.

More importantly, the GEP approach applied here includes weighting schemes whose complexity prevents meaningful physical interpretation of the MMCs solutions and realisation of the absolute and relative power and contribution of individual GHMs. More research is, therefore, needed to explore the effect of application of different levels of constraints on GEP-based algorithm performance in providing more interpretable MMC solutions.

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

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

## Acknowledgements

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

## References

Abrahart, R.J., See, L., 2002. Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments. Hydrology and Earth System Sciences 6(4) 655-670.
Ajami, N.K., Duan, Q.Y., Gao, X.G., Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: Application to Distributed Model Intercomparison Project results. Journal of Hydrometeorology 7(4) 755-768.
Ajami, N.K., Duan, Q.Y., Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resources Research 43(1).
Arnold, J.G., Allen, P.M., Bernhardt, G., 1993. A comprehensive surface-groundwater flow model. Journal of Hydrology 142(1–4) 47-69.
Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J.L., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. Journal of Hydrology 529 754-767.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gosling, S.N., Zaherpour, J., Mount, N.J., Hattermann, F.F., Dankers, R., Arheimer, B., Breuer, L., Ding, J., Haddeland, I., Kumar, R., Kundu, D., Liu, J., van Griensven, A., Veldkamp, T.I.E., Vetter, T., Wang, X., Zhang, X., 2016. A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C. Climatic change 141(3) 577-595.

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

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of forecasting 3(2) 197-204.

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

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

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

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

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

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

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

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

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

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

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

Kim, S., Parinussa, R.M., Liu, Y.Y., Johnson, F.M., Sharma, A., 2015. A framework for combining multiple soil moisture retrievals based on maximizing temporal correlation. Geographical Research Letters 42(16) 6662-6670.

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

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

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

Lima, A.R., Cannon, A.J., Hsieh, W.W., 2015. Nonlinear regression in environmental sciences using extreme learning machines: A comparative evaluation. Environmental Modelling & Software 73 175-188.

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

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

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

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

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

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

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

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

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

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

Mount, N.J., Abrahart, R.J., 2011. Discussion of “River flow estimation from upstream flow records by artificial intelligence methods” by M.E. Turan, M.A. Yurdusev [J. Hydrol. 369 (2009) 71–77]. Journal of Hydrology 396(1-2) 193-196.

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

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

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

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

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

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

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

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

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

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

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

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

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

Van Beek, L.P.H., Bierkens, M.F.P., 2008. The Global Hydrological Model PCR-GLOBWB: Conceptualization, Parameterization and Verification, Report
Department of Physical Geography, Utrecht University, Utrecht, The Netherlands,.

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

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

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

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

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

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

Worland, S.C., Farmer, W.H., Kiang, J.E., 2018. Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. Environmental Modelling & Software 101 169-182.

Wu, W., May, R., Dandy, G.C., Maier, H.R., 2012. A method for comparing data splitting approaches for developing hydrological ANN models, In: R. Seppelt, A.A.V., S. Lange, D. Bankamp (Eds.) (Ed.), International Environmental Modelling and Software Society (iEMSs), 2012 International Congress on Environmental Modelling and Software, Managing Resources of a Limited Planet, Sixth Biennial Meeting: Leipzig, Germany.

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

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

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

Zaherpour, J., Gosling, S.N., Mount, N., Hattermann, F., 2018a. Multi-model combination with a super-ensemble of catchment-scale and global-scale hydrological models, in review. Journal of Hydrology.

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

**Highlights:**

- We present the first use of machine learning-based multi-model combination (MMC) applied to a global hydrological model ensemble.
- MMC performs better than any individual input model and the ensemble mean.
- MMC is not always able to out-perform model combination based on multiple linear regression.
- The physical interpretation of the MMC solutions is limited by the complexity of their non-linear weighting schemes.