Improving Performance of Bucket-Type Hydrological Models in High Latitudes with Multi-Model Combination Methods: Can We Wring Water from a Stone?

Todorović A.1, Grabs T.2, Teutschbein C.2*

1 University of Belgrade, Faculty of Civil Engineering, Institute of Hydraulic and Environmental Engineering, Bulevar kralja Aleksandra 73, 11000 Belgrade, Republic of Serbia
2 Uppsala University, Department of Earth Sciences, Program for Air, Water and Landscape Sciences, Villavägen 16, 752 36 Uppsala, Sweden

*Corresponding author: claudia.teutschbein@geo.uu.se

Key words: conceptual hydrological models; extreme flows; high-latitude catchments; hydrological signatures; information theory; multi-model averaging

Abstract

Multi-model combination (averaging) methods (MMCMs) are used to improve the accuracy of hydrological (precipitation-runoff) outputs in simulation or forecasting/prediction modes. In this paper, we examined if the application of MMCMs can improve model performance in reproducing distributions of hydrological signatures, such as annual maxima or minima of varying durations. To this end, ten MMCMs were applied to 29 bucket-type models to simulate runoff in 50 high-latitude catchments. The MMCMs were evaluated by comparing the resulting simulated flows to the reference (i.e., best-performing) individual model, considering various commonly used performance indicators, as well as model performance in reproducing the distributions of signatures. Additionally, we analysed whether (1) the selection of the candidate models, or (2) targeting specific signatures, such as annual maxima or minima, can improve performance of the model combinations. The results suggest that the application of MMCMs can improve accuracy of runoff simulations in terms of traditional performance indicators, but fails to improve performance in reproducing the distributions of signatures. Neither excluding poor-performing models nor applying the MMCMs with the targeted signatures, improves this aspect of model performance. These findings clearly reveal the need for further research aiming at enhancing model performance in reproducing the distributions of hydrological signatures, which is essential for climate-change impact studies.

1 Introduction

Simulated flows are expected to represent different features of observed flow series as closely as possible. This is a prerequisite for making accurate hydrological forecasting and predictions, such as those under climate change (e.g., Booij and Krol, 2010), and, consequently, for effective water resources management (Pechlivanidis et al., 2013). Accuracy of simulated flows is quantified in terms of performance indicators, which are usually computed by comparing entire series of simulated and observed flows (e.g., Crochemore et al., 2015; Kiraz et al., 2023). However, models should also reproduce statistical properties of hydrological signatures, such as mean flows, annual maxima or minima (Coffey et al., 2004; Willems, 2009; Todorović et al., 2022). This is important for climate-
change impact assessments, which predict changes in those signatures and their distributions under future climate (Lehner et al., 2006; Ludwig et al., 2009; Gain et al., 2013; Velazquez et al., 2013; Vansteenkiste et al., 2014; Karlsson et al., 2016; Gosling et al., 2017; Krysanova et al., 2017; Seiller et al., 2017; Mishra et al., 2020; Fatchir et al., 2021). To improve this aspect of model performance, bias-correction of simulated flows was proposed (González-Zeas et al., 2012; Bum Kim et al., 2021; Daraio, 2020; Farmer et al., 2018; Hales et al., 2023). Alternatively, parameters of hydrological model alone (Ricard et al., 2019), or together with the parameters of a bias-correction method (Ricard et al., 2020) can be optimised to reproduce statistical properties of flows as closely as possible. Nevertheless, it should be emphasised that these approaches are still in their infancy.

Another way to improve model performance is multi-modelling, which implies applications of numerous, different models that simulate the same variable(s) (Höge et al., 2019). The objective of multi-modelling is to combine outputs of individual models in an optimal manner to obtain so-called model combinations that outperform any individual model (“team-of-rivals”, Liang et al., 2011; Darbandsari and Coulibaly, 2020; Wan et al., 2021), since errors of individual candidate models are expected to partly cancel out (Tebaldi and Knutti, 2007). Combining (or averaging) the outputs of several candidate models is generally preferred over selection of a single “best performing” model (“winner-takes-all” approach, Höge et al., 2019), as it reduces bias towards selecting a single model (Raftery, 1995; Diks and Vrugt, 2010; Zhang and Liang, 2011). Broderick et al. (2016) highlighted that a single model cannot outperform all the other candidates according to different criteria. Therefore, different modelling objective(s) can lead to the selection of different models (Diks and Vrugt, 2010; Claeskens, 2016), which can yield quite divergent simulation results (Raftery, 1995). Additionally, small changes in the dataset can lead to a selection of a different model, which makes selection of single models less “stable” than multi-model averaging (Zhang and Liang, 2011). The latter is particularly convenient in cases where several models yield similar performances (Claeskens, 2016), which is quite common in hydrological modelling accompanied by parameter- (Beven and Binley, 1992) and structural equifinality (e.g., Knoben et al., 2020).

Many multi-model averaging methods act at the level of simulated outputs, i.e., not the models are being averaged, but their outputs are combined to obtain weighted simulated variable of interest (Claeskens, 2016). Therefore, Höge et al. (2019) proposed that the approaches focusing on the simulated outputs should be referred to as “multi-model combination methods” (MMCMMs) rather than “model averaging”. This recommendation is adopted in this study.

There are numerous MMCMMs that differ according to their theoretical foundations (Claeskens, 2016; Höge et al., 2019). Some of them, such as Bayesian model averaging, work with and result in probability density functions (Hjort and Claeskens, 2003; Wang et al., 2009; Mitra et al., 2019). Bayesian model averaging is frequently used in hydrological modelling (Ajami et al., 2007; Diks and Vrugt, 2010; Najafi et al., 2011). Alternatively, there are numerous MMCMMs that result in point estimates of simulated variables (Diks and Vrugt, 2010). Point estimates of the simulated variables have various practical applications, and decision-makers and stakeholders generally prefer easily graspable point estimates over probability distributions (Krysanova et al., 2018). Consequently, the use of these MMCMMs in hydrology have increased over the years (Liang et al., 2011; Najafi and Moradkhani, 2015; Arsenault et al., 2017).

Point estimates of a simulated variable are obtained by applying a weighting scheme over outputs of the candidate models (Spiegelhalter et al., 2002; Claeskens, 2016). Model weights are estimated over a training (calibration) period, and are further applied in an independent period, such as an evaluation period (in which the observations exist) or a future period that forecasts/predictions are made over (Tebaldi and Knutti, 2007; Diks and Vrugt, 2010; Arsenault et al., 2017). There are numerous MMCMMs that differ according to the way in which the model weights are estimated (Claeskens, 2016; Wang et al., 2009).

One approach to multi-modelling implies construction a regression-based statistical model (e.g., linear regression) to combine the outputs of candidate models (Dormann et al., 2018). The simplest method of
this kind is equal weighting (Block et al., 2009; Dormann et al., 2018). Although simple, it can often
outperform the individual candidates (Tebaldi and Knutti, 2007; Krinner and Flanner, 2018; Sun and
Trevor, 2018). Another method of this kind was proposed by Granger and Ramanathan (1984), who
considered the predictions by candidate models to be the predictors in a linear regression model. Model
weights are calculated as ordinary least square estimates of the regression model coefficients (Diks and
Vrugt, 2010). Model weights are not restricted to sum up to one (i.e., they are not simplex weights),
since Granger and Ramanathan (1984) demonstrated advantages of such weighting schemes.

Many MMCMs are based on information criteria (“bewildering alphabet of information criteria”,
Spiegelhalter et al., 2014). Information criteria encompass a model likelihood term and a penalty term,
and they are aimed at identifying a balance between flexibility and overfitting, both of which can be
related to the number of model parameters (Spiegelhalter et al., 2002; Diks and Vrugt, 2010; Moral-
Benito, 2015; Höge et al., 2019). Some of the methods stemming from the information theory include:
Akaike- (AIC, Akaike, 1970), deviance- (DIC, Spiegelhalter et al., 2014, 2002), focused- (FIC,
Claeskens and Hjort, 2003) or “widely applicable” information criteria (WAIC, Watanabe, 2013).
Alternatively, Bayesian- (BIC, Schwarz, 1978) and Kashyap information criteria (KIC) are grounded in
Bayesian theory (Höge et al., 2019). In these MMCMs model weights typically sum up to one (simplex
weights) (Claeskens and Hjort, 2001).

In some methods model weights are estimated based on the candidate model performance (Kiesel et al.,
2020; Wang et al., 2019). These weights heavily depend on the performance indicator considered, and
are, thus, highly subjective (Tebaldi and Knutti, 2007). Alternatively, the MMCM proposed by Bates
and Granger (1969) estimates model weights based on past prediction errors, which are commonly
approximated by sample variances of residual series (Diks and Vrugt, 2010).

In some MMCMs, model weights are optimised to minimise discrepancies between the linear
combination of outputs of the candidate models and the corresponding observations (Lee and Song,
2021). Dormann et al. (2018) referred to these methods as “tactical approach to estimating model
weights”. In many MMCMs, such as cross-validation, jackknife, stacking and extensions thereof (Yang,
2001), model weights are optimised to minimise predictive error over hold-out (validation) datasets
(Claeskens, 2016; Dormann et al., 2018). Model weights can also be obtained by minimising Mallows
criterion $C_p$ in the training period (Diks and Vrugt, 2010; Moral-Benito, 2015; Claeskens, 2016). These
methods, especially those that imply repetitive optimisation runs, can be computationally demanding.
Additionally, optimisation of the weights can introduce uncertainties in the final projections (Tebaldi
and Knutti, 2007).

In hydrological modelling, multi-modelling can be performed either with different model structures or
parameter sets obtained with different calibration strategies (Arsenault et al., 2017; Wan et al., 2021).
Such model combinations (i.e., series of flows obtained after applying a MMCM) were shown to have
better transferability and performance than individual models (Seiller et al., 2012; Gudmundsson et al.,
2012; Arsenault et al., 2017), and to improve the performance even in flow-ranges that are not
specifically targeted in the calibration (Arsenault et al., 2015). Dusa et al. (2023) demonstrated that
application of MMCMs improved not only the accuracy of simulated flows at the catchment outlet, but
also at the outlets of the nested subcatchments that were not considered in the estimation of the MMCM
weights. MMCMs were shown to improve accuracy of flow forecasts, especially with short lead times,
and in projections of flood flows (Darbandsari and Coulibaly, 2020). MMCMs have been extensively
used for climate-change impact assessments (Bastola et al., 2011), mainly in the context of climate
model outputs (Min et al., 2007; Simonis et al., 2007; Tebaldi and Knutti, 2007; Bohn et al., 2010; Bhat
et al., 2011; Fischer et al., 2012; Zhang et al., 2015). Arsenault et al. (2017) showed that application of
MMCMs to the inputs for hydrological models can also be advantageous. Applications of MMCMs to
different distribution functions for estimation of design floods were also reported (Di Baldassarre et al.,
2009; Okoli et al., 2018).

Notwithstanding the merits of MMCMs, their applications are accompanied by numerous uncertainties
related to the multi-modelling process. For example, different MMCMs may result in quite divergent
simulated outputs (Mitra et al., 2019), so there are great uncertainties associated with the selection of a
MMCM (Najafi et al., 2011; Najafi and Moradkhani, 2015). Many studies singled out the Granger-
Ramanathan MMCM (sometimes followed by a bias-correction) and optimisation of the model weights
as the most robust MMCMs (Diks and Vrugt, 2010; Broderick et al., 2016; Arsenault et al., 2017; Wan
et al., 2021). However, there is no straightforward guidance on MMCM selection according to specific
modelling objectives (Höge et al., 2019; Lee and Song, 2021), and selection of a MMCM has remained
a subjective decision (Kiesel et al., 2020; Tebaldi and Knutti, 2007).

The selection of candidate models is another important step in multi-modelling, which is also lacking
specific recommendations (Gosling et al., 2017). Some authors (e.g., Gosling et al., 2017) advocated
that the ensemble should include as many candidate models as possible, even parsimonious ones, as the
increasing number of models might mitigate the uncertainty (Tebaldi and Knutti, 2007). On the other
hand, some authors argued that a greater number of candidates can be computationally intractable, and
showed that inclusion of poor performing models increase uncertainties, and recommended that such
models should be omitted from the ensemble (Najafi et al., 2011; Huang et al., 2020). Some studies
demonstrated that even a small number of candidate models can yield satisfactory results, provided that
robust candidates are selected (Lee and Song, 2021). Wan et al. (2021) showed that an increase in the
number of candidate models considerably improves multi-model performance, but this improvement
becomes limited if ensemble includes more than nine candidate models. An informed selection of
candidate models is important not only to provide a more manageable ensemble, but also to reduce
redundancy (Kiesel et al., 2020). To minimise the redundancy in the ensemble, Darbandsari and
Coulibaly (2020) applied the Entropy-based selection algorithm to obtain a subset of the candidate
models that resulted in minimum correlation among its members prior to the application of MMCMs.
However, Tebaldi and Knutti (2007) argued that some degree of redundancy in the ensemble is
inevitable, because all models are grounded in same basic principles and methods, even though they
were developed by different research groups worldwide.

Providing a specific guidance on the selection of MMCMs or the candidate models in the ensemble is
rather challenging. Specifically, it is difficult to compare model combinations to the best performing
individual model, since these comparisons heavily depend on the performance indicators and simulated
variables analysed (Broderick et al., 2016; Seiller et al., 2015). In other words, a single performance
indicator might not fully reveal robustness of the multi-model combination compared to the best
performing individual model. Thus, different aspects of model performance should be considered
(Tebaldi and Knutti, 2007). Another challenge in evaluating model combinations is the fact that
MMCMs are not grounded in physical laws (e.g., Zaherpour et al., 2018). For example, mass
conservation does not necessarily hold for hydrographs obtained with a MMCM (Höge et al., 2019),
i.e., application of MMCMs can distort the water balance. Therefore, setting evaluation frameworks for
model combinations, especially for simulations under future hydroclimatic conditions, has remained an
open research question (Tebaldi and Knutti, 2007; Blöschl et al., 2019).

In view of all these challenges, the objective of this study is to provide novel insights into performance
of model combinations, and thereby facilitate making more informative decisions about applications of
MMCMs for hydrological modelling. This study focuses not only on the overall performance of
MMCMs quantified in commonly used indicators, but also on the MMCM ability to reproduce
distributions of hydrological signatures relevant for climate change impact assessment, especially
extreme flows, such as annual maxima and minima of various durations. The reason behind emphasising
performance in extreme flows is twofold. Firstly, accurate estimation of these flows is a prerequisite for
effective and sustainable water resources management and is, consequently, of great interest to decision-
makers and stakeholders (Broderick et al., 2016; Pechlivanidis et al., 2016). Secondly, accurate
predictions of extreme flows still represent a great challenge to hydrological modelling, and most models
fail to perform satisfactorily in extreme-flow range (Seibert, 2003; Oudin et al., 2006; Vaze et al., 2010;
Kim et al., 2011; Lane et al., 2019; Mizukami et al., 2019; Topalović et al., 2020; Brunner et al., 2021).
To advance previous research in the area of the MMCMs application in hydrological modelling, the
following research questions are addressed:
1) Can MMCMs improve different aspects of model performance, including the reproduction of the distributions of various hydrological signatures relevant for climate change impact studies, such as mean- or extreme flows of various durations?

2) Can preselection of candidate models (i.e., ensemble members) based on their performance, improve efficiency of MMCMs, including efficiency in reproducing distributions of various hydrological signatures?

3) Can performance of MMCMs in reproducing the distributions of signatures be improved if the MMCMs are applied over the series of these signatures?

To address these research questions, ten commonly used MMCMs are applied to 29 spatially-lumped, bucket-type modes of various complexity, in 50 high-latitude catchments that cover a broad range of hydroclimatic conditions. Such multi-catchment, multi-model setup makes this study one of few in the area of multi-modelling in hydrology that involves a large set of catchments and models (e.g., Arsenault et al., 2017, 2015; Seiller et al., 2012; Wan et al., 2021; Dusa et al., 2023).

2 Methodology

2.1 Data and Catchments

Evaluation of MMCMs in this study is conducted in 50 high-latitude catchments in Sweden (Figure 1), with long continuous series of observed daily precipitation, temperature and flows (1961-2020). These catchments are characterised by a relatively low variation in elevation and mild slopes (Figure 1a); however, they vary considerably in areas and latitudes (Table S1 of the Supplementary material). Specifically, the selected catchments cover a longitudinal range between 56°N to 68°N, and all three major climate zones in Sweden: namely, the polar tundra climate zone in the Scandinavian Mountains in north-western Sweden, the subarctic boreal climate in central and northern Sweden, and the warm-summer hemiboreal climate zone (Dfb) in southern Sweden (Teutschbein et al., 2022; Tootoonchi et al., 2023). As for land cover, the selected catchments are mainly covered by forests, and only few of the catchments are extensively cultivated. Fractions of glaciers and urbanised areas are negligible (up to 1.6% and 3.1% of catchment area, respectively), while the share of lakes and wetlands is rather low in most catchments (median area of 12.1%; Table S1). Approximately one third of the study catchments are regulated, but the degree of regulation (i.e., impact of the reservoirs on flows) is quite low (Todorović et al., 2022; Tootoonchi et al., 2023), which is important since accommodation of reservoirs in hydrological models can be quite challenging, especially in a spatially-lumped setup (Todorović et al., 2019; Oliveira et al., 2023). The selected catchments are predominately humid, with the wettest catchments being located in western Sweden (in terms of both precipitation and runoff; Figure 1b,c). Snow-dominated and transitional catchments prevail over the rain-dominated ones, as indicated by high values (above 150) of the centre of timing of the centre of mass of annual runoff (COM) in Figure 1d (Kormos et al., 2016; see section 2.4.1 and Table S1). In half of the catchments, snowfall represents more than one third of total annual precipitation (Figure 1e). A distinct north-south gradient in temperatures is apparent across the catchments (Figure 1f). The main physiographic and hydroclimatic characteristics of the selected catchments are outlined in Table S1 of the Supplementary material.

Daily precipitation, temperature and runoff series over period 1961-2020 are obtained from a publicly accessible database (http://vattenwebb.smhi.se/), hosted by the Swedish Meteorological and Hydrological Institute (SMHI). The daily temperature and precipitation series are obtained from the SMHI’s national precipitation-temperature grid with 4 km x 4 km spatial resolution (SMHI, 2005; Johansson, 2000). The catchment-averaged values are calculated as an area-weighted average of all grid
cells partly or fully lying within a catchment. Locations of the stream gauges are obtained from SMHI’s SVAR database (Eklund, 2011; Henestål et al., 2012).
Figure 1. The selected catchments in Sweden: a) topography of the selected catchments, b) mean annual runoff (mm/year), c) mean annual precipitation (mm/year), d) timing of the centre of mass of annual flow (COM, in days of a water year), e) percentage of total precipitation as snowfall, and f) mean annual temperature (°C).
2.2 Hydrological Models

This research builds on the modelling exercise in which 29 spatially-lumped bucket-type hydrological models of varying complexity (Table 1) are calibrated to simulate runoff in 50 high-latitude catchments (section 2.1), and evaluated by applying an approach proposed by Todorović et al. (2022). This approach to model evaluation takes into account model ability to reproduce distributions of series of hydrological signatures, in addition to commonly considered performance indicators, such as Nash-Sutcliffe efficiency coefficient (Nash and Sutcliffe, 1970).

The selected models vary in complexity: namely, the number of free model parameters range from six (ALPINE-2) to 21 (3DNet-Catch, HMETS), while the number of storages takes values between two and seven. All the models include a snow routine, which is based on the degree-day method in most models, whereas only few models contain a canopy routine. Complexity and conceptualisation of the soil and groundwater routines varies considerably across the models. Linear and non-linear outflow equations are applied for runoff routing in most models, while unit hydrographs are implemented in few models, such as HMETS, MORDOR and the models of the GR-group. To run all of these models in 50 high-latitude study catchments, some adjustments to the original model formulations have to be made. For example, snow routines are added to those models that do not have this feature in their original formulations (the GR-group of models, HYMOD, SAC-SMA, TOPMODEL, XINANJIANG). Conversely, snow routines of some models (HMETS, MORDOR) are simplified to be applicable with the available data that do not include minimum or maximum daily temperatures. Since the selected catchments are not covered with glaciers (section 2.1), glacier routines are omitted from those models that encompass such routine (COSERO, GSM-SOCONT). The details on the models are elaborated in Table S2 of the Supplementary material.

The models are run with catchment-averaged daily precipitation depths and mean daily temperatures (section 2.1), and potential evapotranspiration computed with the Hamon method (Hamon, 1961).
models are calibrated by using the Genetic algorithm (Vrugt, 2015) with the non-parametric version (Pool et al., 2018) of Kling-Gupta efficiency (Gupta et al., 2009) as the objective function. This performance indicator is selected as the objective function since it provides balanced sensitivity to model performance in both high- and low-flow range (Pool et al., 2018). The models are calibrated over the first half of the record period (water years 1962-1991), and evaluated in the second half of the record (water years 1991-2020), with one year for model warm-up that precedes both simulation periods. Such calibration/evaluation scheme is selected to enable assessment of model performance over climatically different periods, i.e., a differential-split sample test is applied (Klemeš, 1986). This calibration/evaluation scheme is considered robust (Seibert, 2003), and is generally recommended in cases of model application for climate change impact studies (Fowler et al., 2018). Additionally, the adopted calibration/evaluation scheme in this paper implies simulations over long periods that correspond to those in climate change impact studies (Todorović et al., 2022).

The comparison between the two simulation periods shows an increase in temperature (especially in the winter, with the median value of 2.35°C), and a general increase in precipitation (except in the autumn in 64% of catchments). The median annual increase in temperature between the two periods amounts to +1.1°C, and closely corresponds to the projected increase in the future- (2070-2100) relative to the baseline period (1971-2010) under the RCP 2.6 scenario (Gutiérrez, J.M. et al., 2021; SMHI, 2022). The median observed increase in annual precipitation amounts to +9.5%, and it lies in-between the projected values obtained with the RCP 2.6 and RCP 8.5 scenarios (SMHI, 2022). Increase in temperature and in precipitation depths over the record period (1961-2020) is also confirmed by the results of the Mann-Kendall test (Kendall, 1938; Mann, 1945). The test detects statistically significant upward trend in the mean annual- and winter temperatures in all catchments, and in spring-, summer-, and autumn temperatures in vast majority of catchments (in 96%, 86% and 92% of catchments, respectively). The Mann-Kendall test also indicates significant increasing trends in annual- and winter precipitation depths in most catchments (78% and 76%, respectively). Detected increase in temperature and in precipitation over the record period suggest that the adopted calibration/evaluation scheme provides a solid ground for proper assessment of performance under changing climate in the selected catchments.

Table 1. The hydrological models used for multi-modelling in this study.

<table>
<thead>
<tr>
<th>ID</th>
<th>Model</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3DNet-Catch</td>
<td>Todorović et al., 2019</td>
</tr>
<tr>
<td>2</td>
<td>ALPINE-2</td>
<td>Knoben et al., 2019</td>
</tr>
<tr>
<td>3</td>
<td>COSERO</td>
<td>Kling et al., 2015</td>
</tr>
<tr>
<td>4</td>
<td>ECHO</td>
<td>Knoben et al., 2019; Schaefl et al., 2014</td>
</tr>
<tr>
<td>5</td>
<td>FLEX-IS</td>
<td>Fenicia et al., 2011, 2008; Knoben et al., 2019b; Nijzink et al., 2016</td>
</tr>
<tr>
<td>6</td>
<td>GR4J</td>
<td>Perrin et al., 2003</td>
</tr>
<tr>
<td>7</td>
<td>GR5J</td>
<td>Pushpalatha et al., 2011</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>Reference</td>
</tr>
<tr>
<td>---</td>
<td>----------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>8</td>
<td>GR6J</td>
<td>Pushpalatha et al., 2011</td>
</tr>
<tr>
<td>9</td>
<td>GSM-SOCNT</td>
<td>Knoben et al., 2019b</td>
</tr>
<tr>
<td>10</td>
<td>HBV-light – basic version</td>
<td>Seibert and Vis, 2012; “HBV-light,” 2020</td>
</tr>
<tr>
<td>12</td>
<td>HBV-light – one GW box</td>
<td>Seibert and Vis, 2012; “HBV-light,” 2020</td>
</tr>
<tr>
<td>13</td>
<td>HBV-light – three GW boxes</td>
<td>Seibert and Vis, 2012; “HBV-light,” 2020</td>
</tr>
<tr>
<td>14</td>
<td>HMETS</td>
<td>Martel et al., 2017; Francois, 2021</td>
</tr>
<tr>
<td>15</td>
<td>HYMOD</td>
<td>Knoben et al., 2019; Perra et al., 2018</td>
</tr>
<tr>
<td>16</td>
<td>IHACRES</td>
<td>Croke and Jakeman, 2004; DHI, 2017</td>
</tr>
<tr>
<td>17</td>
<td>MOPEX 2</td>
<td>Knoben et al., 2019b</td>
</tr>
<tr>
<td>18</td>
<td>MOPEX 3</td>
<td>Knoben et al., 2019b</td>
</tr>
<tr>
<td>19</td>
<td>MOPEX 4</td>
<td>Knoben et al., 2019b</td>
</tr>
<tr>
<td>20</td>
<td>MOPEX 5</td>
<td>Knoben et al., 2019b</td>
</tr>
<tr>
<td>21</td>
<td>MORDOR</td>
<td>Andreassian et al., 2006; Garavaglia et al., 2017</td>
</tr>
<tr>
<td>22</td>
<td>NAM</td>
<td>DHI, 2017</td>
</tr>
<tr>
<td>23</td>
<td>PDM</td>
<td>HOUGHTON-CARR, 1999; Moore, 2007; Moore and Bell, 2002; DHI, 2017</td>
</tr>
<tr>
<td>24</td>
<td>PRMS</td>
<td>Knoben et al., 2019b</td>
</tr>
<tr>
<td>25</td>
<td>SAC-SMA</td>
<td>Maurer et al., 2010; Newman et al., 2015; Agnihotri and Coulibaly, 2020</td>
</tr>
<tr>
<td>26</td>
<td>SIMHYD</td>
<td>Chiew et al., 2009; Chiew et al., 2010</td>
</tr>
<tr>
<td>27</td>
<td>TOPMODEL</td>
<td>Knoben et al., 2019b; Clark et al., 2008</td>
</tr>
</tbody>
</table>
2.3 Selection and Application of Multi-Model Combination Methods

Ten multi-model combination methods (MMCMs) used for point estimation are selected for this study. The method selection is primarily led by the ease of their implementation and computational requirements, i.e., by their practical applicability. Therefore, many computationally intensive methods are omitted from the analysis, such as stacking or jackknife methods (Dormann et al., 2018). Most of the selected methods are based on some information criterion (see Table 2). Key features of the selected methods are outlined in Table 2, while their detailed descriptions can be found in the cited references, primarily in the seminal paper by Diks and Vrugt (2010). Although bias correction is generally expected to improve the performance of model combinations (Arsenault et al., 2015), the results presented by Diks and Vrugt (2010) suggest negligible effects of such a procedure. Therefore, bias-correction is not applied in this study.

The implementation of MMCMs in this study closely follows the application of the hydrological models. Specifically, model weights are estimated over the calibration period (water years 1962-1991), and then applied in the evaluation period (water years 1991-2020) to assess the robustness of each MMCM (section 2.4). The combined simulated flows $\mathbf{X}$ are computed as a linear combination of flows simulated by all the models in the ensemble $\mathbf{X}_m$, multiplied by corresponding model weights $\omega_m$, which may or may not be simplex (i.e., the weights may or may not add up to 1). The resulting multi-model combination $\mathbf{X}$, which represents a series of simulated flows (referred to as ”model combination”), and reads as follows (Diks and Vrugt, 2010):

$$
\mathbf{X} = \sum_{m=1}^{M} \omega_m \mathbf{X}_m^T
$$

where $\mathbf{X}_m^T$ denotes transposed matrix of the simulated flows.
Table 2. Multi-model combination methods used in this study.

<table>
<thead>
<tr>
<th>№</th>
<th>Method</th>
<th>Description and Equations</th>
<th>Simplex Weights</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Equal weights (&quot;democracy&quot;), <em>EW</em></td>
<td>$\omega_m = \frac{1}{M}$</td>
<td>yes</td>
<td>Kiesel et al., 2020</td>
</tr>
<tr>
<td>2</td>
<td>Akaike information criterion, <em>AIC</em></td>
<td>$C_{AIC,m} = \frac{\exp\left(0.5\Delta_{AIC,m}\right)}{\sum_{i=1}^{M} \exp\left(0.5\Delta_{AIC,i}\right)}$</td>
<td>yes</td>
<td>Posada and Buckley, 2004; Diks and Vrugt, 2010; Liang et al., 2011; Symonds and Moussalli, 2011; Schöniger et al., 2014; (Claeskens, 2016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\Delta_{AIC,m} = AIC_m - \min_{i} AIC_i$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$AIC_m = -2\ln L + 2p_m$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$-2\ln L = N \log S_m^2 + N$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Corrected Akaike information criterion, <em>AICc</em></td>
<td>$\omega_{AICc,m} = \frac{\exp\left(0.5\Delta_{AICc,m}\right)}{\sum_{i=1}^{M} \exp\left(0.5\Delta_{AICc,i}\right)}$</td>
<td>yes</td>
<td>Schöniger et al., 2014; Lute and Luce, 2017; Okoli et al., 2018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\Delta_{AICc,m} = AIC_{c,m} - \min_{i} AIC_{c,i}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$AIC_{c,m} = AIC_m + \frac{2p_m(p_m+1)}{N - p_m - 1}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$-2\ln L = N \log S_m^2 + N$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Bayesian information criterion, <em>BIC</em></td>
<td>$\omega_{BIC,m} = \frac{\exp\left(0.5\Delta_{BIC,m}\right)}{\sum_{i=1}^{M} \exp\left(0.5\Delta_{BIC,i}\right)}$</td>
<td>yes</td>
<td>Diks and Vrugt, 2010; Schöniger et al., 2014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\Delta_{BIC,m} = BIC_m - \min_{i} BIC_i$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$BIC_m = -2\ln L + p_m \ln N$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$-2\ln L = N \log S_m^2 + N$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>№</td>
<td>Method</td>
<td>Description and Equations</td>
<td>Simplex Weights</td>
<td>Reference</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>5</td>
<td>Hannan-Quinn information criterion, HQIC</td>
<td>( \omega_{HQIC} = \frac{\exp(0.5\Delta_{HQIC})}{\sum_{i=1}^{M} \exp(0.5\Delta_{HQIC})} ) ( \Delta_{HQIC} = HQIC_m - \min_{j} HQIC_j )</td>
<td>yes</td>
<td>Claeskens, 2016; Ye et al., 2004</td>
</tr>
<tr>
<td>6</td>
<td>Kashyap information criterion, KIC</td>
<td>( \omega_{KIC} = \frac{\exp(0.5\Delta_{KIC})}{\sum_{i=1}^{M} \exp(0.5\Delta_{KIC})} ) ( \Delta_{KIC} = KIC_m - \min_{i} KIC_i )</td>
<td>yes</td>
<td>Ye et al., 2004</td>
</tr>
<tr>
<td>7</td>
<td>Bates-Granger method, BG</td>
<td>( \omega_{BG} = \frac{1/\sigma^2_m}{\sum_{i=1}^{M} 1/\sigma^2_i} )</td>
<td>yes</td>
<td>Diks and Vrugt, 2010</td>
</tr>
<tr>
<td>8</td>
<td>Granger-Ramanathan method, GR</td>
<td>This method yields a column-vector of the set of weights ( \Omega ):</td>
<td>no</td>
<td>Diks and Vrugt, 2010</td>
</tr>
</tbody>
</table>

\( HQIC_m = -2 \ln L + p_m \ln (\ln N) \) \(-2 \ln L = N \log S^2_m + N\)

\( KIC_m = -2 \ln L + 2p_m \ln \left( \frac{N}{2\pi} \right) + \ln FI \) \(-2 \ln L = N \log S^2_m + N\)

\( S_m \) is the sample variance of residual series \( \varepsilon_m \) of the \( m \)th model in the calibration period:

\( \varepsilon_m = X_m - Y \)

The denominator value is obtained from the residual series of all models within the ensemble.
Mallows method, MM

Model weight vector $\Omega_m$ is obtained by minimising the Mallows criterion, which penalises model complexity, i.e., number of parameters of the $m$th model, $p_m$:

$$ C(\Omega) = \sum_{i=1}^{N} (Y_{i,\Omega} - \Omega X_{i,m})^2 + 2 \sum_{m=1}^{M} \Omega_m p_m S_m^2 $$

$S_m$ is an estimate of the variance of the residual series. In this study, $S_m$ is obtained from the model that yielded minimum RMSE, averaged over all catchments in the calibration period (following Diks and Vrugt, 2010). Optimisation is performed with the AMALGAM algorithm (Vrugt et al., 2009; Vrugt and Robinson, 2007) and from the prior distributions of the model weights.

Mallows method with simplex weights, $\text{MM}_{\text{simplex}}$

Non-simplex model weights obtained by applying the Mallows method are rescaled to have non-negative values that sum up to one. In case of negative weights obtained by applying the Mallows method, their value is set to 0 (following recommendations by Lee and Song, 2021).

Notation:

- $\omega_m$ – weight of the $m$th model;
- $N$ – length of the data series;
- $M$ – the number of the models of the ensemble;
- $p_m$ – number of free parameters of $m$th model;
- $L$ – model likelihood;
- $S_m$ – estimated error of series simulated by $m$th model, which is approximated by the sample variance of the residuals $\varepsilon_m$;
- $FI$ – discriminant of the observed Fisher information matrix;
- $\Omega$ – vector of the model weights;
- $X_m$ – matrix of all simulated series by the $M$ models, the matrix size is $N$ by $M$;
- $Y$ – column vector of observations (here: series of observed daily flows); Simplex weights imply non-negative weights that sum up to 1.
Three research questions are addressed in this study: (1) can MMCMs improve model performance, including reproduction of distributions of the hydrological signatures, and (2) can preselection of the candidate models, or (3) targeting particular signatures, enhance performance in reproducing distribution of the signatures? To address the first research question, performance of different model combinations (i.e., series of daily flows obtained with a MMCM) is evaluated in terms of various performance indicators, such as Nash Sutcliffe (Nash and Sutcliffe, 1970) or Kling-Gupta coefficients (Gupta et al., 2009), and with respect to how well the combinations can reproduce distributions of series of selected hydrological signatures (i.e., annual maxima and minima of various durations), following the approach presented by Todorović et al. (2022). Performance of the model combinations is compared to the (on average) best performing individual model, which is considered a reference model in this study (section 2.4.1). Six alternative subsets of models (i.e., ensembles of the candidate models) are created and fed into the MMCMs to address the second research question (section 2.4.2). These alternative model combinations are evaluated in the same way as the combinations obtained with the complete model ensemble. In order to address the third research question, the MMCMs are applied with the series of annual maxima or minima of various length, as opposed to their application with entire series of simulated daily flows used to address the previous research questions (section 2.4.3).

2.4.1 Performance of Model Combinations

Model weights of ten selected multi-model combination methods (MMCMs, Table 2) are obtained from the complete flow series in the calibration period (water years 1962-1991). This results in a set of model weights for the complete model ensemble (E0) for each MMCM, i.e., in ten different sets comprising 29 weight values each. These weights are employed to obtain model combinations in the evaluation period (water years 1991-2020). This procedure is looped over the fifty catchments.

Evaluation of the model combinations builds on the approach proposed by (Todorović et al., 2022), i.e., performance is quantified in terms of (1) numerous performance indicators (Table 3), and (2) the percentage of the catchments in which the distributions of hydrological signatures are well reproduced (Table 4). A large number of performance indicators is selected to facilitate a rigorous evaluation of MMCMs, and is in line with the recommendations in the literature (Tebaldi and Knutti, 2007). Performance in reproducing distributions of the signatures is evaluated by applying the Wilcoxon sign-rank test with annual series of the signatures obtained from observed and simulated flows (Todorović et al., 2022). This test is based on the locations of the distributions, with an underlying assumption that the shapes of the distributions are similar (Kvam and Vidakovic, 2007; Montgomery and Runger, 2003). Although not all the properties of the distributions are taken into consideration through the testing procedure (e.g., variance, skewness), it is considered here that a model combination properly reproduces the distribution of a signature if the null hypothesis of the test is not rejected at 5% level of significance (following Todorović et al., 2022).

To evaluate the effects of multi-modelling, the model combinations are compared to the individual, on average best performing model (Table 2, Figures S1 and S2 of the Supplementary material). To identify the model that is on average best performing one, all the candidate models (Table 1) are ranked (1) according to their performance in terms of various indicators, and (2) in reproducing distributions of the signatures. These ranks are obtained from (1) the median values of all indicators in all catchments, and (2) as the median percentage of catchments with properly reproduced distributions of the signatures. The model ranks obtained in the calibration and evaluation periods are averaged, as presented in Table 5. In case that two models share the same rank value, the lower value (i.e., higher rank) is assigned to the model that yields better performance in the evaluation period. This procedure indicates the MORDOR model (Andreassian et al., 2006; Garavaglia et al., 2017) as on average best performing one across the selected 50 catchments, according to various aspects of performance. Thus, this model is assigned the overall rank of 1 in Table 5, and is considered a reference in this study (hereafter referred to as the reference model). A single reference model for all catchments is preferred over a selection of the reference model for each catchment individually because a multi-catchment approach is employed in this study. Hence, it is deemed that one “multi-catchment reference” model can provide a
A reliable benchmark that can assure consistent assessment of the MMCMs performance. Additionally, this approach is commonly adopted for evaluation of MMCMs in hydrological literature (e.g., Seiller et al., 2012).

To evaluate MMCMs in this study, a performance indicator obtained by a model combination is compared to the corresponding value by the reference model in that catchment. This procedure is repeated for all performance indicators, and in all catchments, resulting in the percentage of catchments in which the MMCM outperformed the reference model according to a specific indicator. In other words, this study focuses on the frequency of outperformance, rather than on the values of the indicators per se. High frequency (greater than 50%) indicates a robust MMCMs. Concerning the distributions of signatures, the percentage of catchments with well reproduced distributions by the reference model is subtracted from the percentage of obtained by the model combination. Differences are preferred over ratios between the two percentage values to avoid potential division by zero, in case that the reference model reproduces distributions properly in none of the catchments. High positive values of these differences suggest that the multi-model combination outperforms the reference model.

Table 3. Performance indicators used for evaluation of the multi-model combination methods (adapted from Todorović et al., 2022).

<table>
<thead>
<tr>
<th>Performance Indicator</th>
<th>Description, Equation and References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KGE</strong></td>
<td>Kling-Gupta efficiency (KGE) coefficient is computed as follows (Gupta et al., 2009):</td>
</tr>
<tr>
<td></td>
<td>$KGE = 1 - \sqrt{\frac{(r-1)^2 + \alpha^2 + (\beta-1)^2}{\sum(Q_{obs,i} - \bar{Q}<em>{obs})^2 \sum(Q</em>{sim,i} - \bar{Q}_{sim})^2}}$</td>
</tr>
<tr>
<td></td>
<td>$\alpha = \frac{\hat{Q}<em>{obs}}{\bar{Q}</em>{obs}}, \quad \beta = \frac{\hat{Q}<em>{sim}}{\bar{Q}</em>{sim}}$</td>
</tr>
<tr>
<td><strong>KGE_{1/2}</strong></td>
<td>KGE is computed from:</td>
</tr>
<tr>
<td></td>
<td>o daily flows, KGE;</td>
</tr>
<tr>
<td></td>
<td>o reciprocal of root-transformed daily flows (KGE_{1/2}) to put more emphasis on low flows (Santos et al., 2018);</td>
</tr>
<tr>
<td></td>
<td>o daily flows in a representative year, obtained by averaging daily flows on a specific calendar day over the entire simulation period, KGE_{wy} (Schaeffli et al., 2014).</td>
</tr>
<tr>
<td><strong>NPKGE</strong></td>
<td>Non-parametric formulation of KGE indicator (Pool et al., 2018) is computed as KGE, with Spearman instead of Pearson correlation coefficient, and with the ratio of standard deviations estimated from FDCs:</td>
</tr>
<tr>
<td></td>
<td>$NPKGE = 1 - \sqrt{\frac{(\text{spearman} - 1)^2 + (\alpha_{NPKGE} - 1)^2 + (\beta_{NPKGE} - 1)^2}{\sum(Q_{obs,i} - \bar{Q}<em>{obs})^2 \sum(Q</em>{sim,i} - \bar{Q}_{sim})^2}}$</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{NPKGE} = 1 - \frac{\sum_{i=1}^{N} FDC_{sim,i}}{NQ_{sim} - FDC_{obs}}$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{NPKGE} = 1 - \frac{\sum_{i=1}^{N} FDC_{obs,i}}{NQ_{obs} - FDC_{sim}}$</td>
</tr>
<tr>
<td><strong>NSE</strong></td>
<td>Nash-Sutcliffe efficiency coefficient (Nash and Sutcliffe, 1970) is computed from flows (NSE) and log-transformed flows (NSE_{log}) by applying the following equation:</td>
</tr>
<tr>
<td></td>
<td>$NSE = 1 - \frac{\sum_{i=1}^{N} (Q_{obs,i} - Q_{sim,i})^2}{\sum_{i=1}^{N} (Q_{obs,i} - \bar{Q}_{obs})^2}$</td>
</tr>
<tr>
<td><strong>LME</strong></td>
<td>Liu-Mean Efficiency represents a modification of KGE computed from daily flow series (Liu, 2020):</td>
</tr>
</tbody>
</table>
| | $LME = 1 - \sqrt{(k - 1)^2 + (\beta - 1)^2}$, \quad k = \frac{\hat{Q}_{obs}}{\bar{Q}_{obs}} = ar$
<table>
<thead>
<tr>
<th>Performance Indicator</th>
<th>Description, Equation and References</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R^2 ) Coefficient of determination (e.g., Krause et al., 2005):</td>
<td></td>
</tr>
</tbody>
</table>
| \[
R^2 = \frac{\sum_{i=1}^{N} (Q_{i,\text{sim}} - \bar{Q}_{\text{sim}})(Q_{i,\text{obs}} - \bar{Q}_{\text{obs}})}{\sqrt{\sum_{i=1}^{N} (Q_{i,\text{sim}} - \bar{Q}_{\text{sim}})^2 \sum_{i=1}^{N} (Q_{i,\text{obs}} - \bar{Q}_{\text{obs}})^2}}
\] |
| Volumetric efficiency (Criss and Winston, 2008): |
| \[
VE = 1 - \frac{\sum_{i=1}^{N} |Q_{i,\text{obs}} - Q_{i,\text{sim}}|}{\sum_{i=1}^{N} Q_{i,\text{sim}}}
\] |
| Lindström efficiency coefficient represents a modified version of NSE coefficient (Seibert and Vis, 2010): |
| \[
LE = NSE - 0.1 \frac{\sum_{i=1}^{N} |Q_{i,\text{obs}} - Q_{i,\text{sim}}|}{\sum_{i=1}^{N} Q_{i,\text{sim}}}
\] |

\( KGE_{\text{fdc}} \) obtained from entire flow duration curves (FDC; \( KGE_{\text{fdc}} \)) and from different FDC segments obtained from flows that are exceeded given % of time of the simulation period:
- extremely high flows: exceeded 5% of time;
- high flows: exceeded 5-20% of time;
- mean flows, exceeded 20-70% of time;
- low flows, exceeded 70-95% of time;
- extremely low flows, exceeded 95-100% of time (Pfannerstill et al., 2014);
- overall low flow segment, exceeded 70-100% of time.

These performance indices are computed from the mean values of the extremes, obtained from daily series of simulated \( (\mu_{\text{sim,obs}}, \mu_{\text{sim,obs}}) \) and observed flows \( (\mu_{\text{sim,obs}}, \mu_{\text{sim,obs}}) \) (following Mizukami et al., 2019):

\[
AB_{Q_{\text{max}}} = \left( \frac{\mu_{\text{sim,obs}}}{\mu_{\text{sim,obs}}} - 1 \right)^2 \quad AB_{Q_{\text{min}}} = \left( \frac{\mu_{\text{sim,obs}}}{\mu_{\text{sim,obs}}} - 1 \right)^2
\]

Series of annual maxima and minima obtained from daily flows are considered in this study.
Table 4. Hydrological signatures used for evaluation of the multi-model combination methods (adapted from Todorović et al., 2022).

<table>
<thead>
<tr>
<th>Hydrological Signature</th>
<th>Description, Equation and References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean annual flow, $Q_{\text{mean}}$</td>
<td>Mean flows in a water year (from 1st October to 30th September).</td>
</tr>
<tr>
<td>Mean spring flow, $Q_{\text{spring}}$</td>
<td>Series of mean flows in the spring (1st March through 31st May) over the simulation period (Chen et al., 2017).</td>
</tr>
<tr>
<td>1-, 5- and 30-day maximum annual flows, $Q_{\text{max, d}}$ for $d=1$, 5 and 30</td>
<td>Series of annual maxima obtained from daily flows averaged over 5 and 30-days in each water year of the simulation period (Dankers et al., 2014; Vis et al., 2015).</td>
</tr>
<tr>
<td>1-, 3-, 7-, 10-, 20-, 30- and 90 day minimum flows, $Q_{\text{min, d}}$ for $d=1$, 3, 7, 10, 20 and 30</td>
<td>Series on minimum flows averaged over a given number of days obtained in each water year of the simulation period (Richter et al., 1996; Olden and Poff, 2003; Garcia et al., 2017).</td>
</tr>
<tr>
<td>10th and 90th flow percentiles in wet seasons, $Q_{\text{wet,10p}}$ and $Q_{\text{wet,90p}}$</td>
<td>Series of specific flow percentiles obtained in each water year of the simulation period. Wet season is defined as period from 1st April through 30th September (Yarnell et al., 2020).</td>
</tr>
<tr>
<td>10th and 90th flow percentiles in dry seasons, $Q_{\text{dry,10p}}$ and $Q_{\text{dry,90p}}$</td>
<td>Series of specific flow percentiles obtained in each water year of the simulation period. Dry season is defined as period from 1st October through 31st March (Yarnell et al., 2020).</td>
</tr>
</tbody>
</table>

Timing of the centre of mass of annual flow, $COM$:

$$COM = \frac{\sum_{i} Q_{t_i}}{\sum_{i} t_i}$$

where $t_i$ represents the $i$th ordinal day of a water year.

Spring onset (spring “pulse day”), $SPD$: Spring onset is the ordinal number of the day in which the negative difference between the streamflow mass curve and the mean streamflow mass curve is the greatest (Cunderlik and Ouarda, 2009). Spring onset series is obtained from values in each water year of a simulation period.

High flow frequency, $HFF$: Series of mean number of days in a water year with flows greater than 5 times the mean observed flow in the simulation period. In the literature, flows greater than 9 times the mean observed flow are used for high flow frequency computations (Westerberg and McMillan, 2015; Krysanova et al., 2017). Since the considered catchments in this paper exhibit relatively low flow variability, this threshold is reduced to 5.

Low-flow frequency, $LFF$: Series of mean number of days in a water year with flows smaller than 20% of the mean observed flow in the simulation period (Nicolle et al., 2014; Westerberg and McMillan, 2015; Krysanova et al., 2017).

Timing of the maximum annual flow, $T_{Q_{\text{max}}}$: Ordinal number of a day in which maximum annual flow occurred, obtained in each water year of the simulation period (Richter et al., 1996).

Timing of the minimum annual flow, $T_{Q_{\text{min}}}$: Ordinal number of a day in which minimum annual flow occurred, obtained in each water year of the simulation period. If there are several consecutive days with the same minimum flows, the mean timing of these days in a water year is adopted (Vis et al., 2015; Parajka et al., 2016).
To assess the effects of preselecting candidate models on the performance of multi-model combination (the second research question), six alternative ensembles are created (denoted by E$_1$ through E$_5$ and E$_{uni}$, Table 5). Five of these ensembles (E$_1$ – E$_5$) are obtained by successively reducing the number of candidate models (approximately in steps of five) by omitting the poor performing models, down to the smallest ensemble with five best performing (i.e., most elite) candidates. The model selection is guided by the overall model ranks (Table 5), which are obtained from their performance quantified in terms of various indicators, and performance in reproducing the distributions of signatures (Figures S1-S2 of the Supplementary material). Ensembles with fewer than five members are not considered here (following recommendations by Seiller et al., 2012). Another alternative ensemble (denoted by E$_{uni}$) is created by keeping only one best-performing model from each group, and discarding remaining models from the group (namely, GR-, MOPEX- and HBV-light groups), which results in the ensemble of 21 models. The objective of creating such an ensemble is to reduce potential redundancy in the pool of candidate models (following Kiesel et al., 2020).

The model weights are obtained separately for each ensemble, yielding thereby six sets of weights for each of ten MMCM, i.e., sixty sets of weights in every catchment. The model weights are estimated from the entire flow series in the calibration period (water years 1962-1991), and further applied in the evaluation period (water years 1991-2020). The model combinations are evaluated by applying the same approach as in case of the complete ensemble (E$_0$).

Table 5. The hydrological models and their ranks according to performance indicators and efficiency in reproducing distributions of hydrological signatures, and the overall ranks, and different model ensembles created based on the overall model ranks. The size of the ensemble is gradually reduced from E$_0$ (includes all models) down to E$_5$ (includes five most elite models), while E$_{uni}$ contains only one model from each group (GR-, MOPEX-, and HBV-light groups). Values in parentheses by the ensemble labels indicate the number of models included, which are indicated by the values of 1 in the shaded cells. The reference model with the overall rank of 1 is shown in bold.

<table>
<thead>
<tr>
<th>ID</th>
<th>Model</th>
<th>Preform. Indicators</th>
<th>Distributions</th>
<th>Overall</th>
<th>E$_0$ (29)</th>
<th>E$_1$ (25)</th>
<th>E$_2$ (20)</th>
<th>E$_3$ (15)</th>
<th>E$_4$ (10)</th>
<th>E$_5$ (5)</th>
<th>E$_{uni}$ (21)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3DNet-Catch</td>
<td>16</td>
<td>17</td>
<td>16</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>ALPINE-2</td>
<td>26</td>
<td>29</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>COSEREO</td>
<td>19</td>
<td>21</td>
<td>21</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>ECHO</td>
<td>22</td>
<td>19</td>
<td>22</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>FLEX-IS</td>
<td>25</td>
<td>26</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>GR4J</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>GR5J</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>---</td>
<td>------</td>
<td>------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td>--------------------------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>1 1 1 1 1 0 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>13</td>
<td>7</td>
<td>11</td>
<td>1 1 1 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>1 1 1 1 1 1 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>3</td>
<td>11</td>
<td>1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>15</td>
<td>12</td>
<td>1 1 1 0 0 0 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>4</td>
<td>11</td>
<td>1 1 1 1 1 1 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>24</td>
<td>25</td>
<td>24</td>
<td>1 0 0 0 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>22</td>
<td>15</td>
<td>1 1 1 0 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>8</td>
<td>9</td>
<td>1 1 1 1 1 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>20</td>
<td>18</td>
<td>1 1 0 0 0 0 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>20</td>
<td>18</td>
<td>19</td>
<td>1 1 0 0 0 0 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>14</td>
<td>6</td>
<td>10</td>
<td>1 1 1 1 1 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>23</td>
<td>16</td>
<td>20</td>
<td>1 1 0 0 0 0 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>21</td>
<td>14</td>
<td>17</td>
<td>1 1 0 0 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>18</td>
<td>24</td>
<td>23</td>
<td>1 0 0 0 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>29</td>
<td>28</td>
<td>29</td>
<td>0 0 0 0 0 0 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>15</td>
<td>13</td>
<td>14</td>
<td>1 1 1 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>28</td>
<td>27</td>
<td>27</td>
<td>0 0 0 0 0 0 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>1 1 1 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>27</td>
<td>23</td>
<td>25</td>
<td>1 0 0 0 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.4.3 Assessment of the Impact of Selection of a Target Hydrological Signature on Performance of Model Combinations

The MMCMs can be applied both over the entire flow series, and over the series of targeted hydrological signatures, such as series of annual maxima. To address the third research question, model weights are obtained from the series of selected signatures in the calibration period. In this study, this analysis is performed only with the series of extreme flows, i.e., annual maxima and minima of various durations (Table 4). Specifically, series of 1-, 5-, and 30-day annual maxima, and 1-, 3-, 7-, 10-, 20-, 30-, and 90-day annual minima obtained in water years (Table 4), are used to estimate model weights, yielding ten sets of weights for each MMCM, i.e., one hundred weights in total. Series of particular extreme flows (e.g., 30-day annual minima) in the evaluation period are simulated by using the corresponding set of the weights. These analyses are conducted with the complete ensemble of 29 candidates (E₀). Application of MMCMs with the series of a targeted signature results only in the series of that particular signature over the simulation period. Therefore, performance of the MMCMs can be quantified only as the percentage of catchments in which the distribution of a particular signature (extreme flows) is well reproduced, and these values can be compared to the corresponding results of the reference model or the MMCMs applied over the complete series of daily flows (E₀).

3 Results

3.1 Performance of Model Combinations

Application of MMCMs generally improves model efficiency, i.e., the model combinations outperform the reference individual model in over half of the catchments in most performance indicators, particularly over the evaluation period (Figure 2). On the contrary, variability in the performance of MMCMs across the indicators is more pronounced in the calibration period (indicated by dark-shaded colours in the top panel of Figure 2). The MMCMs improve values of Kling-Gupta coefficient (KGE), Nash-Sutcliffe coefficient computed from daily flows (NSE), $R^2$ and Lindström efficiency coefficient (LE), as indicated by dark-blue cells in Figure 2. These indicators reflect model performance in reproducing runoff dynamics, and are generally sensitive to high flows (Moriasi et al., 2007; Krause et al., 2005; Legates and McCabe, 1999). Subtle improvement in Nash-Sutcliffe coefficient computed from log-transformed flows ($NSE_{log}$) and KGE computed for a representative year ($KGE_{200}$) is also obtained with the multi-modelling. As for the other indicators, the effects of multi-modelling are either negligible (e.g., Liu-mean efficiency coefficient, $LME$), or model combinations are largely outperformed by the reference model, especially $KGE$ obtained from extremely low flows ($KGE_{95\text{th}}$). The objective function in the calibration of the candidate models, NPKGE (section 2.2), which already takes high values (Figure S1), is improved in only few cases in the evaluation period.

The reference model is most often outperformed by the MMCMs based on the information criteria and especially by the Granger-Ramanathan method (GR, Table 2), which outperforms the reference model in terms of NSE, $R^2$ and LE in all catchments in the calibration period (Figure 2). The information criteria-related MMCMs exhibit rather uniform (almost identical) performance, regardless of the indicator. Overall performance across ten MMCM is largely consistent over two simulation periods, with exception of the Bates-Granger method (BG), which performance improves in the evaluation period, and the Mallows methods (MM and MM$_{simplex}$), but to a lesser extent.
468 reproducing distributions of the hydrological signatures (Figure 3). Specifically, the percentage of catchments
469 As opposed to the performance indicators, the application of the MMCMs does not improve performance in
470 Figure 2. The percentage of catchments in which a multi-model combination method outperforms the reference model
471 according to a specific indicator (Table 3) in the calibration (top) and evaluation periods (bottom panel).
472
473 As opposed to the performance indicators, the application of the MMCMs does not improve performance in
474 reproducing distributions of the hydrological signatures (Figure 3). Specifically, the percentage of catchments
475 in which the distribution of a signatures is properly reproduced by the model combinations is consistently
476 lower than the values obtained by the reference model (Figure 3). The MMCMs neither improve performance
477 lower than the values obtained by the reference model (Figure 3). The MMCMs neither improve performance
In the signatures that are well reproduced by the candidate models (e.g., 30-day annual maxima or spring onset, SPD), nor in the signatures that are poorly reproduced by the candidate models, such as annual minima (Figure S2 of the Supplementary material). The smallest differences between the model combinations and the reference model are obtained in the dry season flow percentiles, which are well reproduced in almost none of the catchments in both simulation periods. Performance of the MMCMs in reproducing distributions of signatures is somewhat higher in the signatures related to mean-, spring-, or high-flows, or runoff timings (e.g., SPD, COM or timing of annual maxima, T_{Q_{max}}; Table 4), than in signatures related to low-flows (Figure 3). Differences in the percentage of catchments with properly reproduced distributions across MMCMs are minor in most signatures, with few exceptions, such as annual maxima of various durations or high-flow frequency (HFF). This suggests that none of the MMCMs is clearly superior over the others. Nevertheless, the GR method slightly outperforms other MMCM in reproducing distributions of some signatures, such as mean and spring flows, 30-day annual maxima or minima or wet season flow percentiles, particularly in the calibration period (Figure 3). The GR method is closely followed (and even outperformed in few instances) by the information criteria-based MMCMs, and this pattern persists in both simulation periods.

The results presented thus far refer to the entire set of catchments, meaning that considerable variation in their hydroclimatic and physiographic properties (section 2.1) is overlooked. To examine whether catchment properties reflect on the performance of MMCMs, the performance of five selected methods in each study catchment is mapped (Figure S3 of the Supplementary material). To facilitate the presentation of the results, only percentages of the performance indicators and the hydrological signatures according to which a MMCM outperformed the reference model in each catchment, are shown. These maps reveal large variability in MMCM performance across the catchments, however, no apparent relationship between the performance level of a MMCM, and catchment area, latitude, or climate zone can be found.
3.2 Impact of Preselection of Candidate Models on Performance of Model Combinations

To assess the impact of preselection of the candidate models on efficiency of the MMCMs, various performance indicators (Table 3) are computed for six alternative ensembles $E_1$ – $E_5$, and $E_{uni}$ (Table 5), and compared to the reference model. This results in the percentage of catchments in which the reference model is outperformed by a MMCM according to a specific indicator. These results for three selected ensembles are shown in Figure 4, while the complete results are presented in Figure S4 of the Supplementary material.

These results reveal a strong resemblance among ensembles $E_1$ – $E_5$, which is indicated by the overall patterns in the heatmaps, and by the average performance across the MMCMs (shown in the rightmost column of each heatmap). Specifically, the ranks of ten MMCMs according to the average performance remain consistent across the ensembles (including $E_0$), and in both simulation periods. The results also show similarities between $E_1$ – $E_5$ and $E_0$, which reflect in the indicators that are most improved with multi-modelling ($KGE$, $NSE$, $R^2$ and $LE$), and in the higher frequency of outperformance of the reference model in the evaluation period (section 3.1). Despite the overall resemblance, there are differences among the ensembles $E_1$ – $E_5$. The most apparent
is a general increase in the mean performance of MMCMs with reducing the ensemble size down to the most elite candidate models (Figure 5). In other words, most MMCMs yield best average performance either with $E_4$ or $E_5$, with exception of GR and MM in the calibration that yield the highest performance with $E_0$ and $E_3$, respectively (Figure 5). Additionally, model combinations obtained with $E_0$ are on average outperformed by ensembles $E_1 - E_5$ in most cases (indicated by the prevalence of blue cells in Figure S5), except for the GR method in both periods, and MM in the evaluation (Figure 5). The indicator values generally increase with reducing the ensemble down to most elite candidates (Figures S4 and S5 of the Supplementary material), but such behaviour is not exhibited by indicators computed from annual maxima and minima ($AB_{Q_{\text{max}}}$, $AB_{Q_{\text{min}}}$), and some other indicators with GR and especially $E_5$ ($KGE$, $LME$, $VE$; Figure S5). An increase in performance with selection of more elite candidates is primarily noted in EW, BG and MM$_{\text{simplex}}$, and in GR and MM in the evaluation period. Conversely, the information criteria-related MMCMs exhibit fairly consistent performance across all ensembles (including $E_{\text{uni}}$) in both simulation periods, i.e., they can be considered least “sensitive” to the selection of the candidate models. The performance of ensemble $E_{\text{uni}}$ is notably lower in comparison to the other ensembles, with an exception of the GR method in the calibration (Figure 5). However, it should be emphasised that the differences in the overall performance across the ensembles can be minor in some cases (e.g., differences across information criteria-based MMCMs, in both periods, or between $E_0$ and $E_{\text{uni}}$ in the calibration; Figure 5), even though size of these ensembles varies considerably from five ($E_3$) to 29 ($E_0$) candidate models.

Preselection of the candidate models does not improve the performance of MMCMs in reproducing distributions of the signatures, and the MMCMs clearly remain outperformed by the reference model in this respect, regardless of the ensemble used (Figure S6 of the Supplementary material). Furthermore, performance of different ensembles in reproducing distributions of signatures largely corresponds the performance by $E_0$, and it remains fairly constant across the ensembles (pale-shaded cells in Figure S7), with marginally higher performance by $E_3$, $E_4$ and $E_5$, mainly with the EW and both MM methods. In some instances, a decrease in this aspect of performance relative to $E_0$ is obtained (e.g., with the GR method with $E_5$ in the calibration period). The differences between $E_0$ and other alternative ensembles are mainly detected in annual maxima of various durations.
### Performance Indicators

#### Multi-Model Combination Method

<table>
<thead>
<tr>
<th>Method</th>
<th>KGE</th>
<th>NSE</th>
<th>LE</th>
<th>VE</th>
<th>r²</th>
<th>GR</th>
<th>AB</th>
<th>VE</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>EW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>78</td>
<td>48</td>
<td>58</td>
<td>68</td>
<td>76</td>
<td>54</td>
<td>60</td>
<td>46</td>
<td>50</td>
</tr>
<tr>
<td>AICc</td>
<td>78</td>
<td>48</td>
<td>58</td>
<td>68</td>
<td>76</td>
<td>54</td>
<td>56</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>BIC</td>
<td>76</td>
<td>48</td>
<td>58</td>
<td>68</td>
<td>76</td>
<td>54</td>
<td>54</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>HQIC</td>
<td>76</td>
<td>48</td>
<td>58</td>
<td>68</td>
<td>76</td>
<td>54</td>
<td>55</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>KIC</td>
<td>76</td>
<td>48</td>
<td>58</td>
<td>68</td>
<td>76</td>
<td>54</td>
<td>54</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>BG</td>
<td>70</td>
<td>52</td>
<td>52</td>
<td>62</td>
<td>68</td>
<td>54</td>
<td>54</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>GR</td>
<td>82</td>
<td>54</td>
<td>54</td>
<td>64</td>
<td>76</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>30</td>
</tr>
<tr>
<td>MM</td>
<td>62</td>
<td>26</td>
<td>26</td>
<td>34</td>
<td>34</td>
<td>34</td>
<td>34</td>
<td>34</td>
<td>20</td>
</tr>
</tbody>
</table>

#### Performance Indicator

![Performance Indicator Chart](chart.png)
Figure 4. The percentage of the catchments in which a model combination outperforms the reference model according to a specific performance indicator (Table 3) in the evaluation period. The model combinations are obtained with the daily flow series and with three different ensembles (Table 5) that include 25 ($E_{i}$) and 5 ($E_{s}$) candidate models, and 21 candidate models from different groups ($E_{uni}$).

Figure 5. The average performance of the model combinations obtained with all ensembles (Table 5). The cell values show the percentage of catchments in which the model combination outperforms the reference model, averaged over all performance indicators (Table 3). The font size is adjusted to indicate best performing ensemble for each multi-model combination method.

### 3.3 Impact of Selection of a Target Hydrological Signature on Performance of Model Combinations

Estimation of model combinations obtained from the series of targeted signatures (i.e., annual maxima and minima of various durations), instead of entire flow series, results in the model combinations that can be used
for simulation of those series alone. Therefore, these model combinations can only be evaluated in terms of performance in reproducing distributions of the series of targeted signatures. Figure 6 shows the percentage of catchments in which distributions of annual maxima and minima of various durations are properly reproduced by the model combinations obtained with the targeted signatures, and by the reference model. The reference model outperforms all model combinations according to the targeted signatures both simulation periods, although it is closely followed by the MM method in reproducing annual minima of short durations in calibration. Performance in reproducing distributions of annual maxima is considerably lower in comparison to the reference model in both periods (Figure 6 and Figure S8 of the Supplementary Material). The BG and MM methods slightly outperform other MMCMs in the calibration period, but performance of MM noticeably drops in evaluation, and the greatest difference in average performance between two simulation periods is shown by this exact MMCM. The BG method slightly outperforms other MMCMs in all targeted signatures in the evaluation period (Figure S8).

Marginal impact of application of MMCMs with the targeted signatures is also indicated by the fact that the model combinations obtained with the complete daily flow records (E₀) yield a higher percentage of catchments with properly reproduced distributions in many instances, especially in the evaluation period (Figure S8). Specifically, E₀ outperforms the model combinations created with the targeted signatures with all information criteria-based MMCMs in both simulation periods, and with GR and MM methods in the evaluation period. This pattern is particularly pronounced in annual minima of various durations. The exceptions in this regard are the EW, BG and the MM methods in the calibration period (Figure S8).
Figure 6. The percentage of catchments in which model combinations obtained with the series of annual maxima and minima of various durations, and the reference model properly reproduce distributions of these signatures in the calibration (top) and evaluation periods (bottom panels).

4 Discussion

4.1 Performance of Model Combinations

Application of the MMCMs improves model performance in terms of commonly used indicators in many instances, which corroborates the results of previous studies (Diks and Vrugt, 2010; Seiller et al., 2012; Dusa et al., 2023). The greatest improvements are obtained for the indicators that reflect performance in runoff dynamics and in high flows, such as $KGE$, $NSE$, $R^2$ or $LE$ (Table 3), which are improved in all catchments by the Granger-Ramanathan (GR) method in the calibration period. Application of MMCMs has a minor impact on the performance indicators related to extreme flows, especially to low flows (e.g., $KGE$ computed from the extremely low-flow FDC segment or $AB_{Q_{min}}$), which are not satisfactorily reproduced by the candidate models in this study (Figure S2). Slight improvements are also obtained for many indicators that are already well reproduced by the model ensemble, such as the objective function used for model calibration ($NPKGE$) or $VE$ or $KGE_{FDC}$ in calibration (Figure S1). This pattern can suggest that multi-modelling cannot lead to a noteworthy improvement of the indicators that already take rather low- or high values, which could explain a higher effect of multi-modelling in the evaluation period than in calibration. However, this should not be considered a strict rule, since opposing examples can be found. For example, $KGE$ or $KGE_{wy}$, which are well reproduced by the model candidates in the calibration period, are improved in many catchments by applying MMCMs, while $KGE_{70-100}$, which takes quite low values in majority of the candidate models is improved in some catchments in evaluation. In other words, a relationship between the indicator value obtained by the candidate models, and the level of improvement by multi-modelling is not a straightforward one.

Application of equal weighting (EW) improves values of the indicators, but to a lesser degree than other methods. The greatest improvement in performance indicators is obtained with the MMCMs based on the information criteria and (especially) the GR method. The former yields fairly consistent levels of improvement across the simulation periods, which can be explained by the fact that all these methods result in rather high weights assigned to a single model (here: parsimonious GR4J model, Table 1). Further, in this study a large dataset is used, which counteract the effects of the penalty terms in the information criteria-based methods. Good performance of the GR method was also demonstrated in many studies (e.g., Diks and Vrugt, 2010; Arsenault et al., 2015; Broderick et al., 2017; Arsenault et al., 2017; Wan et al., 2021). The Mallow’s method generally outperforms its simplex version, which is consistent with the results presented by Diks and Vrugt (2010).

Although the MMCMs improve values of many performance indicators, no improvements are obtained in terms of reproduction of distributions of signatures in comparison to the reference model. In other words, the null hypothesis of the Wilcoxon rank sum test, stating that the distributions of series of signatures obtained from observed flow series and from model combinations have same properties, are not rejected at 5% level of significance in fewer catchments. This is in line with the conclusions presented by Todorović et al. (2022), who argued that good performance in terms of the commonly used indicators does not assure that the distributions of the series of hydrological signatures are well reproduced. Poor performance is rather pronounced in the signatures related to low-flows, which corroborates the results by Wan et al. (2021).

These results could be explained by analysing high and low flows. Figure 7a presents observed and simulated hydrographs in the rainfall-dominated Vassboten catchment, as well as the flow duration curve (FDC), scaled to highlight agreement between simulated and observed runoff in high- (Figure 7b) and low flows (Figure 7c). The simulated hydrographs and FDCs are obtained by the reference model and with ten model combinations. The hydrographs and the FDCs clearly show that the model combinations tend to underestimate the highest peak flows, and overestimate low flows during prolonged dry periods, and these discrepancies are more pronounced than in the reference model. Numerous previous studies showed that high model performance in extreme flows represents a great challenge, since model calibration tends to move flow distribution tails towards the central value (“squeezing” of the flow distribution, Farmer et al., 2018). The results presented in
This study suggests that application of the MMCMs with long daily flow series “squeezes” the flow distribution even more, and, thereby, deteriorates model performance in reproduction of distributions of the signatures.

4.2 Impact of Preselection of Candidate Models on Performance of Model Combinations

Alternative ensembles are created to evaluate impact of preselection of candidate models on the performance of MMCMs: namely, five ensembles comprising 25 ($E_1$) through five most elite candidate models ($E_5$), and one ensemble with one model from each model group ($E_{uni}$; Table 5). These ensembles are specifically created to enable assessment of the effects of robustness of the candidate models ($E_1$ – $E_5$), and redundancy in the ensemble ($E_{uni}$) on MMCM performance.

Our results suggest a general increase in model performance with exclusion of poor performing models. In most cases, the best results are obtained either with $E_4$ (10 candidate models) or $E_5$, which suggests that the ensemble size is not decisive for MMCM performance, and that high performance can be achieved with a low number of robust candidates, which corroborates findings by Lee and Song (2021). On the other hand, these results to some extent contradict findings by Wan et al. (2021), who demonstrated that increasing ensemble size from five to nine noticeably improves performance of the combinations, while any further increase in the ensemble size yields only marginal improvements. Our results corroborate the latter; however, no “leap” in MMCM performance level between $E_4$ and $E_5$ is not obtained in this study.
Ensemble $E_{\text{uni}}$ encompasses 21 candidate models, but few well-performing models are left out of the ensemble to reduce potential “redundancy” (Table 5 and Figure S1). This leads to poorer performance than other ensembles in this study, which also suggests that performance (i.e., robustness) of the candidate models, rather than their diversity, dictates performance of MMCMs. As for the size of the ensemble, this study shows that this is not the key criterion for creation of an ensemble, provided that there are five or more candidate models.

The candidate models in this study are selected specifically according to their performance, including performance level, and consistency across different aspects and periods. Nevertheless, alternative approach to candidate selection could be adopted, such as application of the information criteria (Claeskens et al., 2019). Bearing in mind that information criteria, when applied with complete daily series, result in the highest weight assigned to a single model, whereas the weights assigned to the other candidates take values close to zero (even in the smallest ensembles of only five candidate models), it is deemed that such an approach could not significantly facilitate creation of model ensembles in this study. Application of other approaches to preselection of candidate models, such as a re-sampling procedure used by Wan et al. (2021), requires future research.

Although average performance does not vary substantially across the ensembles, there are variations across the indicators and across the MMCMs. Reduction in the size of the ensemble deteriorates performance in some indicators (mainly those that are most improved with the $E_0$ compared to the reference model), but improves in some other indicators. As for the variations among the MMCMs, the results indicate the methods based on the information criteria as least sensitive to the ensemble, which could be explained by the fact that these methods consistently assign the greatest weight to a single candidate model, as discussed in the previous section. On the other hand, the GR method is shown most sensitive to the selection of the candidate models, closely followed by EW, BG, and both MM methods.

Concerning performance in reproduction of the distributions, there is a strong similarity across the ensembles (including $E_0$), which remains consistent across the MMCMs, signatures and simulation periods. Eliminating poorly performing models leads to slight improvements, mostly in the signatures related to annual maxima and with the EW, and both MM methods. However, none of the ensembles outperforms the reference model in this regard, suggesting that the issue with “squeezing” of the flow distribution (Farmer et al., 2018) cannot be compensated by eliminating candidate models from the ensemble.

The conclusions on the impact of the candidate model preselection on MMCM performance presented here are drawn from the application of 29 models in 50 high-latitude catchments. Further analyses can be conducted in other catchments (e.g., from other climate zones), and with larger ensembles that can be created by employing some of the modular frameworks, such as FUSE (Clark et al., 2008) or MARRMoT (Knoben et al., 2022), or by implementing e.g., a semi-distributed model setup (following Dusa et al., 2023).

### 4.3 Impact of Selection of a Target Hydrological Signatures on Performance of Model Combinations

Application of MMCMs only with series of targeted signatures, such as annual maxima or minima of various durations, generally does not assure that the distributions of these series are better reproduced than by the reference model or by the model combinations obtained with the complete daily flow series. Our results suggest in fact a poorer performance, which is inconsistent with the general scientific perception that targeting signatures leads to improvement in performance of the model combinations (e.g., Tebaldi and Knutti, 2007). This could be attributed to the fact that the annual series (here: annual maxima and minima) are insufficiently long to enable proper estimation of model weights, as opposed to the entire flow series. Application of the peak-over-threshold method (e.g., Vukmirovíc and Plavšić, 1997) or the pooled station-year method over homogenous regions (such as those identified by Teutschbein et al. (2022) for the study catchments) can assure longer series of extreme flows. Estimation of MMCM weights over such series could improve performance in this regard, and testing of this hypothesis requires further research.

Poor model performance in reproducing distributions of extreme flows could be to some extent attributed to the fact that most of the candidate models do not give high performance in this regard, especially when it comes to low flows (Figure S2). This could partly be attributed to the calibration strategy (e.g., Topalović et
which is not considered in this study. Specifically, this study focuses on the benefits of MMCMs application itself, and it is deemed that the impacts of the calibration of the candidate models on reasoning on effects of the MMCMs in this study are marginal. However, model calibration with alternative objective function(s) that put emphases on extreme flows might improve this aspect of model performance (by the candidate models and MMCMs), but further research is needed to test this hypothesis.

Multi-modelling with the targeted series in this study reveals some features of the MMCMs that are not exhibited to that extent when applied with complete series of flows. For example, considerable drop in performance of the MM method in the evaluation period can suggest its proneness to overfitting, when applied with short series. Interestingly, this is not exhibited by its simplex version, $\text{MM}_{\text{simplex}}$. Further research is needed to analyse under which conditions overfitting occurs, and to which extent that can affect the transferability of the model combinations. The GR method is singled out as one of the best performing when applied with the complete flow series. However, such characterisation cannot be attributed in simulations with the series of targeted signatures, which can indicate sensitivity of this method to the input data.

5 Conclusions

This study provides novel insights in performance of model combinations obtained by applying ten multi-model combination methods (MMCMs) with daily flow series simulated by 29 models in 50 high-latitude catchments. Performance of the model combinations is quantified in terms of commonly used indicators, such as Nash-Sutcliffe and Kling-Gupta coefficients, and in terms of percentage of catchments in which the distributions of hydrological signatures are properly distributed. The model combinations are evaluated by comparing different aspects of their performance to the performance of the reference model.

The application of the MMCMs can improve model performance in terms of some indicators when compared to the reference model in both calibration and evaluation periods. The greatest improvement is obtained with the MMCMs based on the information criteria and the Granger-Ramanathan (GR) method. However, it should be emphasised that no MMCM is superior over the others: for example, GR can be sensitive to the input data, the Mallows method can be prone to overfitting, while information criteria-based methods can be irresponsible to selection of the candidate models, and application with elite ensembles may not improve their performance. As for the remaining MMCMs, their performance improves by omitting poor performing models from the ensemble. This study shows that neither the size of the ensemble (provided that it comprises more than five candidates), nor diversity within the ensemble are as crucial for the MMCM performance as the robustness of the candidate models.

No improvement is obtained in terms of reproducing the distributions of the signatures in this study. From the standpoint of annual maxima or minima, application of multi-model combination methods leads to further “squeezing” of the flow distribution, moving the distribution tails even closer to the mean values. Neither preselection of candidate models, i.e., excluding of poor-performing models from the ensemble, nor application of the combination methods specifically with a series of the targeted signatures, such as annual maxima or minima, improves performance of the model combinations in this regard. These results suggest that we cannot wring water from a stone, i.e., application of MMCMs cannot enhance this aspect of model performance of the original model ensemble. Proper reproduction of the distributions of signatures, particularly extreme flows, clearly remains a great challenge to hydrological models. Thus, we argue that further research is needed to improve this aspect of their performance, in particular because robust and reliable simulations of such distributions are crucial for climate-change impact studies and sustainable water resources management in general.

Acknowledgements

This research is conducted as part of project “Reducing uncertainties in hydrological climate change impact research to allow for robust streamflow simulations” supported by the Swedish Research Council (VR starting grant: 2017-04970).
The data used in this paper were obtained from Meteorological and Hydrological Institute. The authors thank the reviewers for their constructive feedback.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**


Francois, D., 2021. HMETS hydrological model.


Gosling, S.N., Zaherpour, J., Mount, N.J., Hattermann, F.F., Dankers, R., Arheimer, B., Breuer, L., Ding, J., Haddeland, I., Kumar, R.,


Improving Performance of Bucket-Type Hydrological Models in High Latitudes with Multi-Model Combination Methods: Can We Wring Water from a Stone?

Todorović A.¹, Grabs T.², Teutschbein C.²*

¹ University of Belgrade, Faculty of Civil Engineering, Institute of Hydraulic and Environmental Engineering, Bulevar kralja Aleksandra 73, 11000 Belgrade, Republic of Serbia

² Uppsala University, Department of Earth Sciences, Program for Air, Water and Landscape Sciences, Villavägen 16, 752 36 Uppsala, Sweden

*Corresponding author: claudia.teutschbein@geo.uu.se

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Highlights

- Ten multi-model combination methods (MMCMs) are created from 29 models in 50 basins
- MMCMs improve some performance indicators, especially the Granger-Ramanathan method
- MMCMs do not improve performance in reproducing distributions of hydrological signatures
- Application with series of targeted signatures does not improve MMCM performance
- MMCM performance is improved by selecting more robust candidate models for ensembles