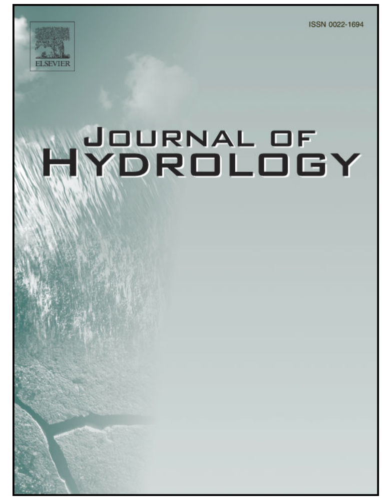
Journal Pre-proofs



Research papers

Improving performance of bucket-type hydrological models in high latitudes with multi-model combination methods: Can we wring water from a stone?

A. Todorović, T. Grabs, C. Teutschbein

PII: S0022-1694(24)00223-3
DOI: https://doi.org/10.1016/j.jhydrol.2024.130829
Reference: HYDROL 130829

To appear in: *Journal of Hydrology*

Please cite this article as: Todorović, A., Grabs, T., Teutschbein, C., Improving performance of bucket-type hydrological models in high latitudes with multi-model combination methods: Can we wring water from a stone?, *Journal of Hydrology* (2024), doi: https://doi.org/10.1016/j.jhydrol.2024.130829

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; Fatehifar 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" (MMCMs) rather than "model averaging". This recommendation is adopted in this study.

There are numerous MMCMs 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 MMCMs 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 MMCMs 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 MMCMs 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).

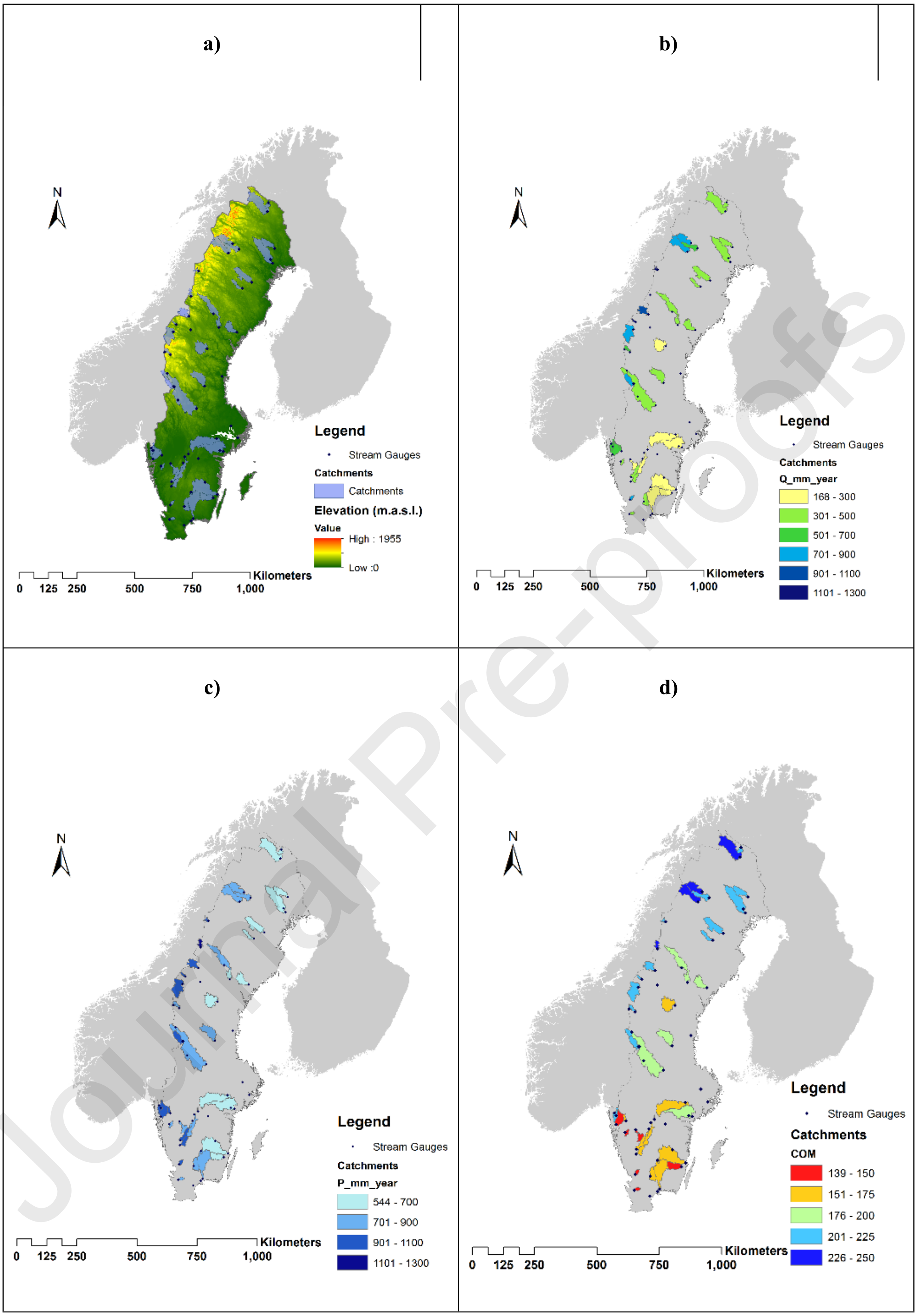


Figure 1. The selected catchments in Sweden: a) topography of the selected catchments, b) mean annual runoff (mm/year), c) mean annual precipitation (mm/year), d) timing of the centre of mass of annual flow (COM, in days of a water year), e) percentage of total precipitation as snowfall, and f) mean annual temperature (°C).

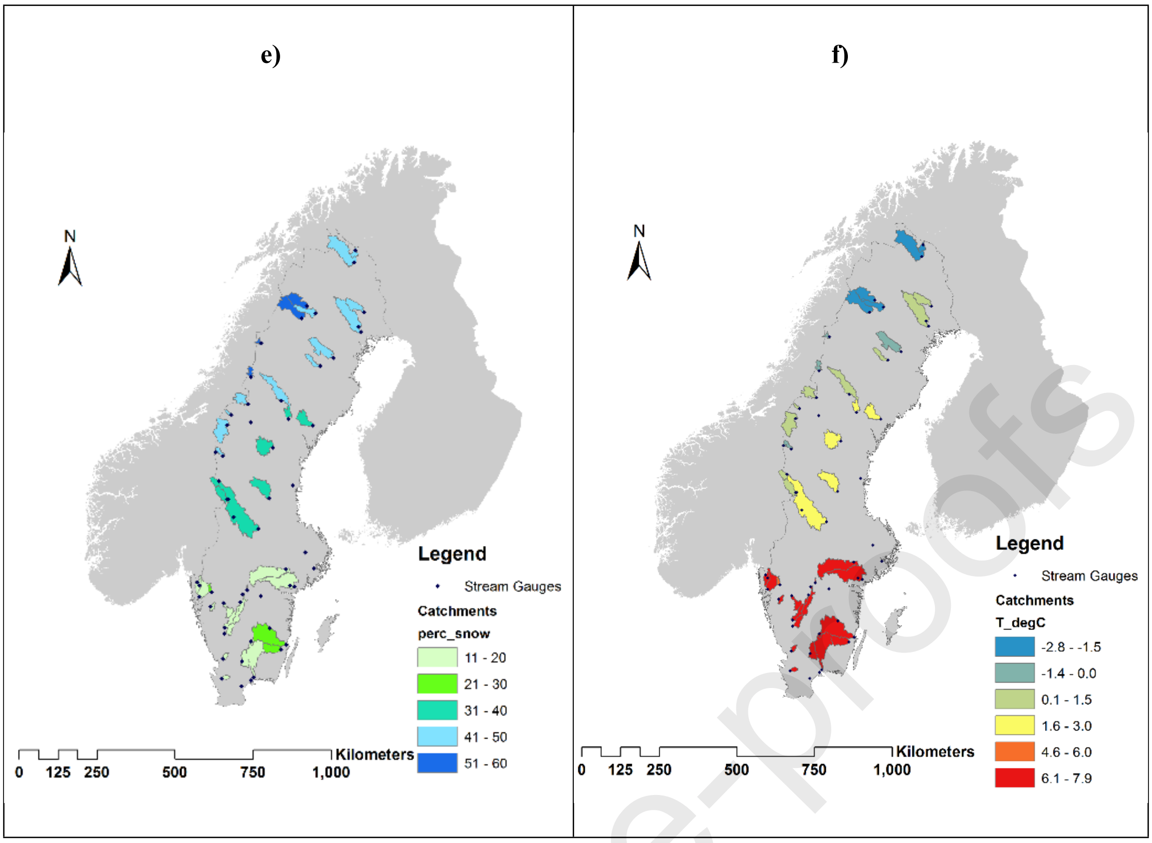


Figure 1. (continued)

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-liner 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). The 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.

ID	Model	Reference
1	3DNet-Catch	Todorović et al., 2019
2	ALPINE-2	Knoben et al., 2019
3	COSERO	Kling et al., 2015
4	ECHO	Knoben et al., 2019; Schaefli et al., 2014
5	FLEX-IS	Fenicia et al., 2011, 2008; Knoben et al., 2019b; Nijzink et al., 2016
6	GR4J	Perrin et al., 2003
7	GR5J	Pushpalatha et al., 2011

8	GR6J	Pushpalatha et al., 2011
9	GSM-SOCONT	Knoben et al., 2019b
10	HBV-light – basic version	Seibert and Vis, 2012; "HBV-light," 2020
11	HBV-light – standard version	Seibert and Vis, 2012; "HBV-light," 2020
12	HBV-light – one GW box	Seibert and Vis, 2012; "HBV-light," 2020
13	HBV-light – three GW boxes	Seibert and Vis, 2012; "HBV-light," 2020
14	HMETS	Martel et al., 2017; Francois, 2021
15	HYMOD	Knoben et al., 2019; Perra et al., 2018
16	IHACRES	Croke and Jakeman, 2004; DHI, 2017
17	MOPEX 2	Knoben et al., 2019b
18	MOPEX 3	Knoben et al., 2019b
19	MOPEX 4	Knoben et al., 2019b
20	MOPEX 5	Knoben et al., 2019b
21	MORDOR	Andreassian et al., 2006; Garavaglia et al., 2017
22	NAM	DHI, 2017
23	PDM	HOUGHTON-CARR, 1999; Moore, 2007; Moore and Bell, 2002; DHI, 2017
24	PRMS	Knoben et al., 2019b
25	SAC-SMA	Maurer et al., 2010; Newman et al., 2015; Agnihotri and Coulibaly, 2020
26	SIMHYD	Chiew et al., 2009; Chiew et al., 2010
27	TOPMODEL	Knoben et al., 2019b; Clark et al., 2008

28	VIC/ARNO	Clark et al., 2008; Knoben et al., 2019b
29	XINANJIANG	Xingnan, 1994

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 X are computed as a linear combination of flows simulated by all the models in the ensemble X_{m}, multiplied by corresponding model weights ω_{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 X, which represents a series of simulated flows (referred to as "model combination"), and reads as follows (Diks and Vrugt, 2010):

$$X = \sum_{m=1}^{M} \omega_{\mathrm{m}} X_{\mathrm{m}}^{T} \tag{1}$$

where X_{m}^{T} denotes transposed matrix of the simulated flows.

Table 2. Multi-model combination methods used in this study.

№	Method	Description and Equations	Simplex Weights	Reference
1	Equal weights ("democracy"), *EW*	$\omega_{\mathrm{m}} = \frac{1}{M}$	yes	Kiesel et al., 2020
2	Akaike information criterion, *AIC*	$C_{\mathrm{AIC,m}} = \frac{\exp\left(0.5\Delta_{\mathrm{AIC,m}}\right)}{\sum_{i=1}^{M}\exp\left(0.5\Delta_{\mathrm{AIC},i}\right)}$ $\Delta_{\mathrm{AIC,m}} = \mathrm{AIC_m} - \min_i \mathrm{AIC}_i$ $AIC_{\mathrm{m}} = -2\ln L + 2p_{\mathrm{m}}$ $-2\ln L = N\log S_{\mathrm{m}}^2 + N$	yes	Posada and Buckley, 2004; Diks and Vrugt, 2010; Liang et al., 2011; Symonds and Moussalli, 2011; Schöniger et al., 2014; (Claeskens, 2016
3	Corrected Akaike information criterion, *AICc*	*AICc* differs from AIC according to the penalty term, which is modified to account for size of the dataset (Höge et al., 2019). $\omega_{\mathrm{AICc,m}} = \frac{\exp\left(0.5\Delta_{\mathrm{AICc,m}}\right)}{\sum_{i=1}^{M}\exp\left(0.5\Delta_{\mathrm{AICc},i}\right)}$ $\Delta_{\mathrm{cAIC,m}} = AIC_{\mathrm{c,m}} - \min_i AIC_{\mathrm{c},i}$ $AIC_{\mathrm{c,m}} = AIC_{\mathrm{m}} + \frac{2p_{\mathrm{m}}\left(p_{\mathrm{m}}+1\right)}{N - p_{\mathrm{m}} - 1}$ $-2\ln L = N\log S_{\mathrm{m}}^2 + N$	yes	Schöniger et al., 2014; Lute and Luce, 2017; Okoli et al., 2018
4	Bayesian information criterion, *BIC*	$\omega_{\mathrm{BICm}} = \frac{\exp\left(0.5\Delta_{\mathrm{BICm}}\right)}{\sum_{i=1}^{M}\exp\left(0.5\Delta_{\mathrm{BIC}i}\right)}$ $\Delta_{\mathrm{BIC,m}} = BIC_{\mathrm{m}} - \min_i BIC_i$ $BIC_{\mathrm{m}} = -2\ln L + p_{\mathrm{m}}\ln N$ $-2\ln L = N\log S_{\mathrm{m}}^2 + N$	yes	Diks and Vrugt, 2010; Schöniger et al., 2014

№	Method	Description and Equations		Simplex Weights	Reference
5	Hannan-Quinn information criterion, *HQIC*	$\omega_{\mathrm{HQICm}} = \dfrac{\exp\left(0.5\Delta_{\mathrm{HQICm}}\right)}{\sum_{i=1}^{M}\exp\left(0.5\Delta_{\mathrm{HQIC}i}\right)}$ $HQIC_{\mathrm{m}} = -2\ln L + p_{\mathrm{m}}\ln\left(\ln N\right)$	$\Delta_{\mathrm{HQIC,m}} = HQIC_{\mathrm{m}} - \min_i HQIC_i$ $-2\ln L = N\log S_{\mathrm{m}}^2 + N$	yes	Claeskens, 2016; Ye et al., 2004
6	Kashyap information criterion, *KIC*	$\omega_{\mathrm{KICm}} = \dfrac{\exp\left(0.5\Delta_{\mathrm{KICm}}\right)}{\sum_{i=1}^{M}\exp\left(0.5\Delta_{\mathrm{KIC}i}\right)}$ $KIC_{\mathrm{m}} = -2\ln L + 2p_{\mathrm{m}}\ln\left(\dfrac{N}{2\pi}\right) + \ln FI$	$\Delta_{\mathrm{KIC,m}} = KIC_{\mathrm{m}} - \min_i KIC_i$ $-2\ln L = N\log S_{\mathrm{m}}^2 + N$	yes	Ye et al., 2004
7	Bates-Granger method, *BG*	$\omega_{\mathrm{m}} = \dfrac{1/S_{\mathrm{m}}^2}{\sum_{i=1}^{M}1/S_i^2}$ S_{m} is the sample variance of residual series ε_{m} of the m$^{\mathrm{th}}$ model in the calibration period: $\varepsilon_{\mathrm{m}} = X_{\mathrm{m}} - Y$ The denominator value is obtained from the residual series of all models within the ensemble.		yes	Diks and Vrugt, 2010
8	Granger-Ramanathan method, *GR*	This method yields a column-vector of the set of weights Ω: $\Omega = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}Y$		no	Diks and Vrugt, 2010

№	Method	Description and Equations	Simplex Weights	Reference
9	Mallows method, *MM*	Model weight vector Ω_m is obtained by minimising the Mallows criterion, which penalises model complexity, i.e., number of parameters of the m[th] model, p_m: $C(\Omega)=\sum_{i=1}^{N}\left(Y_{i,1}-\Omega X_{i,m}\right)^{2}+2\sum_{m=1}^{M}\Omega_m\, p_m\, S_m^2$ 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.	no	Diks and Vrugt, 2010
10	Mallows method with simplex weights, $MM_{simplex}$	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).	yes	Diks and Vrugt, 2010

Notation: ω_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 ε_m; FI – discriminant of the observed Fisher information matrix; Ω – 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.

2.4 Evaluation of the Multi-Model Combination Methods

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 is 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 (E_0) 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 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).

Performance Indicator	**Description, Equation and References**
KGE $KGE_{1/\sqrt{Q}}$ KGE_{wy}	Kling-Gupta efficiency (KGE) coefficient is computed as follows (Gupta et al., 2009): $KGE = 1-\sqrt{(r-1)^2+(\alpha-1)^2+(\beta-1)^2}$ $r=\frac{\sum\left(Q_{\mathrm{obs},i}-\bar{Q}_{\mathrm{obs}}\right)\left(Q_{\mathrm{sim},i}-\bar{Q}_{\mathrm{sim}}\right)}{\sqrt{\sum\left(Q_{\mathrm{obs},i}-\bar{Q}_{\mathrm{obs}}\right)^2\sum\left(Q_{\mathrm{sim},i}-\bar{Q}_{\mathrm{sim}}\right)^2}}$ $\alpha=\frac{\hat{S}_{Q_{\mathrm{sim}}}}{\hat{S}_{Q_{\mathrm{obs}}}}$ $\beta=\frac{\bar{Q}_{\mathrm{sim}}}{\bar{Q}_{\mathrm{obs}}}$ KGE is computed from: ○ daily flows, KGE; ○ reciprocal of root-transformed daily flows ($KGE_{1/\sqrt{Q}}$) to put more emphasis on low flows (Santos et al., 2018); ○ daily flows in a representative year, obtained by averaging daily flows on a specific calendar day over the entire simulation period, KGE_{wy} (Schaefli et al., 2014).
$NPKGE$	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: $NPKGE=1-\sqrt{\left(r_{\mathrm{Spearman}}-1\right)^2+\left(\alpha_{\mathrm{NP}}-1\right)^2+(\beta-1)^2}$, $\alpha_{\mathrm{NP}}=1-\frac{1}{2}\sum_{i=1}^{N}\left\lvert\frac{FDC_{\mathrm{sim},i}}{N\bar{Q}_{\mathrm{sim}}}-\frac{FDC_{\mathrm{obs},i}}{N\bar{Q}_{\mathrm{obs}}}\right\rvert$
NSE	Nash-Sutcliffe efficiency coefficient (Nash and Sutcliffe, 1970) is computed from flows (NSE) and log-transformed flows ($NSE_{\log Q}$) by applying the following equation: $NSE=\frac{\sum_{i=1}^{N}\left(Q_{\mathrm{obs},i}-Q_{\mathrm{sim},i}\right)^2}{\sum_{i=1}^{N}\left(Q_{\mathrm{obs},i}-\bar{Q}_{\mathrm{obs}}\right)^2}$
LME	Liu-Mean Efficiency represents a modification of KGE computed from daily flow series (Liu, 2020): $LME=1-\sqrt{\left[\left(k_1-1\right)^2+(\beta-1)^2\right]}$, $k_1=r\frac{\hat{S}_{Q_{\mathrm{sim}}}}{\hat{S}_{Q_{\mathrm{obs}}}}=\alpha r$

Performance Indicator	**Description, Equation and References**
R^2	Coefficient of determination (e.g., Krause et al., 2005): $R^2 = \frac{\sum_{i=1}^{N}\left(Q_{\mathrm{obs},i}-\bar{Q}_{\mathrm{obs}}\right)\left(Q_{\mathrm{sim},i}-\bar{Q}_{\mathrm{sim}}\right)}{\sqrt{\sum_{i=1}^{N}\left(Q_{\mathrm{obs},i}-\bar{Q}_{\mathrm{obs}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Q_{\mathrm{sim},i}-\bar{Q}_{\mathrm{sim}}\right)^2}}$
VE	Volumetric efficiency (Criss and Winston, 2008): $VE = 1-\frac{\sum_{i=1}^{N}\left\|Q_{\mathrm{sim},i}-Q_{\mathrm{obs},i}\right\|}{\sum_{i=1}^{N}Q_{\mathrm{obs},i}}$
LE	Lindström efficiency coefficient represents a modified version of *NSE* coefficient (Seibert and Vis, 2010): $LE = NSE - 0.1\frac{\sum_{i=1}^{N}\left\|Q_{\mathrm{obs},i}-Q_{\mathrm{sim},i}\right\|}{\sum_{i=1}^{N}Q_{\mathrm{obs},i}}$
KGE_{fdc}	*KGE* obtained from entire flow duration curves (FDC; KGE_{fdc}) and from different FDC segments obtained from flows that are exceeded given % of time of the simulation period: o extremely high flows: exceeded 5% of time; o high flows: exceeded 5-20% of time; o mean flows, exceeded 20-70% of time; o low flows, exceeded 70-95% of time; o extremely low flows, exceeded 95-100% of time (Pfannerstill et al., 2014); o overall low flow segment, exceeded 70-100% of time.
$AB_{Q\max}$ $AB_{Q\min}$	These performance indices are computed from the mean values of the extremes, obtained from daily series of simulated $\left(\mu_{Q\max,\mathrm{sim}}, \mu_{Q\min,\mathrm{sim}}\right)$ and observed flows $\left(\mu_{Q\max,\mathrm{obs}}, \mu_{Q\min,\mathrm{obs}}\right)$ (following Mizukami et al., 2019): $AB_{Q\max} = \sqrt{\left(\frac{\mu_{Q\max,\mathrm{sim}}}{\mu_{Q\max,\mathrm{obs}}}-1\right)^2}$ $AB_{Q\min} = \sqrt{\left(\frac{\mu_{Q\min,\mathrm{sim}}}{\mu_{Q\min,\mathrm{obs}}}-1\right)^2}$ 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 Todorovic et al., 2022).

Hydrological Signature	**Description, Equation and References**
Mean annual flow, Q_{mean}	Mean flows in a water year (from 1st October to 30th September).
Mean spring flow, Q_{spring}	Series of mean flows in the spring (1st March through 31st May) over the simulation period (Chen et al., 2017).
1-, 5- and 30-day maximum annual flows, $Q_{\mathrm{max},d}$ for d=1, 5 and 30	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).
1-,3-, 7-, 10-, 20-, 30- and 90 day minimum flows, $Q_{\mathrm{min},d}$ for d=1, 3, 7, 10, 20 and 30	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).
10th and 90th flow percentiles in wet seasons, $Q_{\mathrm{wet,10p}}$ and $Q_{\mathrm{wet,90p}}$	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).
10th and 90th flow percentiles in dry seasons, $Q_{\mathrm{dry,10p}}$ and $Q_{\mathrm{dry,90p}}$	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).
Timing of the centre of mass of annual flow, *COM*	Timing is computed from daily flows Q_{i} and for each year in a simulation period (Mendoza et al., 2015; Kormos et al., 2016): $COM=\frac{\sum_i Q_{\mathrm{i}} t_{\mathrm{i}}}{\sum_i Q_{\mathrm{i}}}$ 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, $T_{Q\mathrm{max}}$	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, $T_{Q\mathrm{min}}$	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).

2.4.2 Assessment of the Impact of Preselection of Candidate Models on Performance of Model Combinations

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.

ID	Model	Model Ranks			Ensemble with preselected models						
		Preform. Indicators	Distributions	Overall	E_0 (29)	E_1 (25)	E_2 (20)	E_3 (15)	E_4 (10)	E_5 (5)	E_{uni} (21)
1	3DNet-Catch	16	17	16	1	1	0	0	0	1	1
2	ALPINE-2	26	29	28	0	0	0	0	0	1	0
3	COSERO	19	21	21	1	0	0	0	0	1	1
4	ECHO	22	19	22	1	0	0	0	0	1	1
5	FLEX-IS	25	26	26	0	0	0	0	0	1	0
6	GR4J	1	5	2	1	1	1	1	1	1	1
7	GR5J	5	9	6	1	1	1	1	0	0	1

8	GR6J	8	11	8	1	1	1	1	0	0	1
9	GSM-SOCONT	13	7	11	1	1	1	0	0	1	1
10	HBV-light – basic version	6	3	5	1	1	1	1	1	0	1
11	HBV-light – standard version	4	2	3	1	1	1	1	1	1	1
12	HBV-light – one GW box	9	15	12	1	1	1	0	0	0	1
13	HBV-light – three GW boxes	3	4	4	1	1	1	1	1	0	1
14	HMETS	24	25	24	1	0	0	0	0	1	1
15	HYMOD	10	22	15	1	1	1	0	0	1	1
16	IHACRES	11	8	9	1	1	1	1	0	1	1
17	MOPEX 2	17	20	18	1	1	0	0	0	0	1
18	MOPEX 3	20	18	19	1	1	0	0	0	0	1
19	MOPEX 4	14	6	10	1	1	1	1	0	1	1
20	MOPEX 5	23	16	20	1	1	0	0	0	0	1
21	**MORDOR**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
22	NAM	21	14	17	1	1	0	0	0	1	1
23	PDM	18	24	23	1	0	0	0	0	1	1
24	PRMS	29	28	29	0	0	0	0	0	1	0
25	SAC-SMA	15	13	14	1	1	1	0	0	1	1
26	SIMHYD	28	27	27	0	0	0	0	0	1	0
27	TOPMODEL	12	12	13	1	1	1	0	0	1	1
28	VIC/ARNO	27	23	25	1	0	0	0	0	1	1

29	XINANJIANG	7	10	7	1	1	1	1	0	1	1

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 wights 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 (E_0). 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 (E_0).

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 ($NSE_{\log Q}$) and *KGE* computed for a representative year (KGE_{wy}) 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 ($KGE_{95\text{-}100}$), and the indicators computed from annual maxima ($AB_{Q\max}$) and minima ($AB_{Q\min}$). 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.

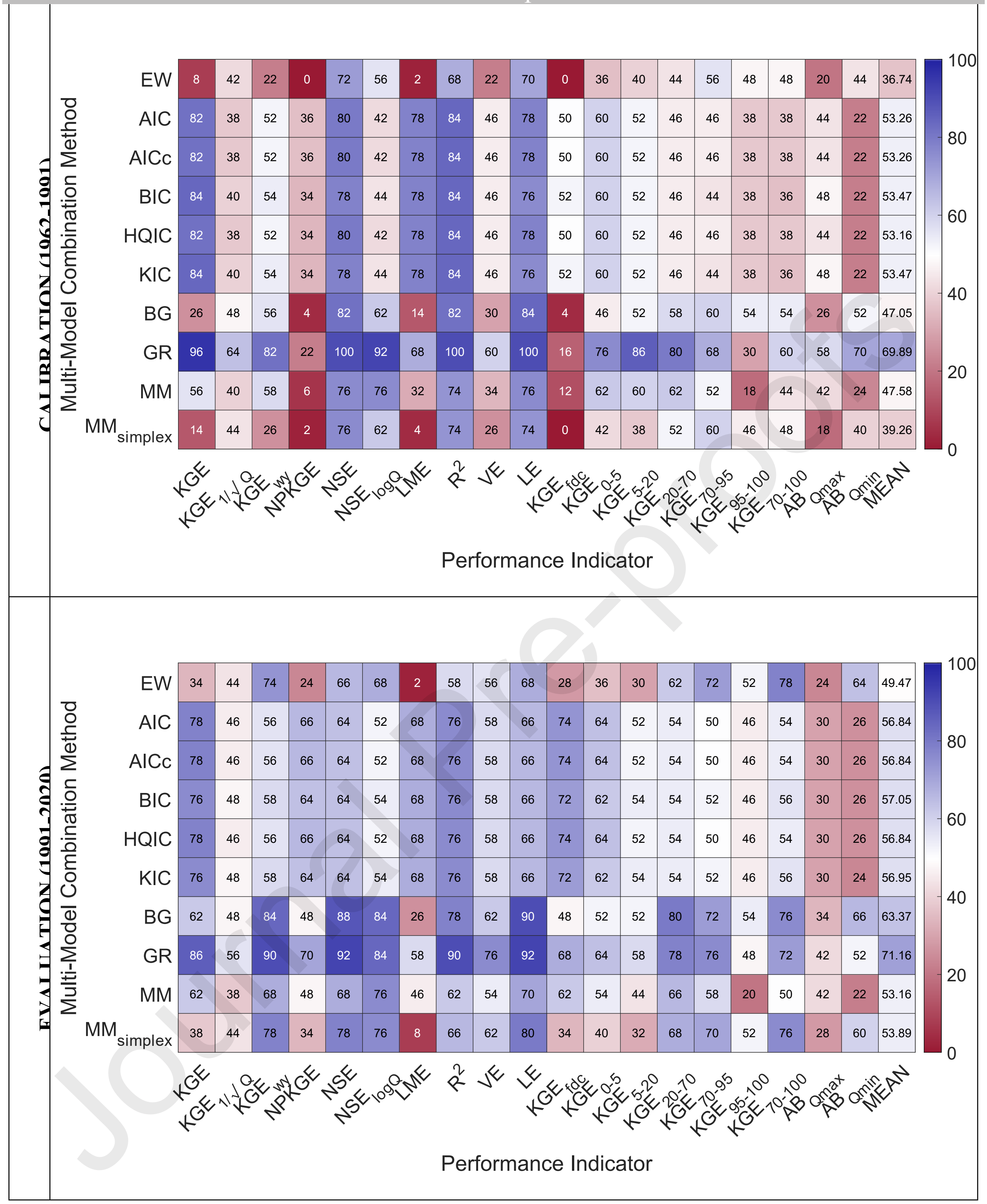


Figure 2. The percentage of catchments in which a multi-model combination method outperforms the reference model according to a specific indicator (Table 3) in the calibration (top) and evaluation periods (bottom panel).

As opposed to the performance indicators, the application of the MMCMs does not improve performance in reproducing distributions of the hydrological signatures (Figure 3). Specifically, the percentage of catchments in which the distribution of a signatures is properly reproduced by the model combinations is consistently 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.

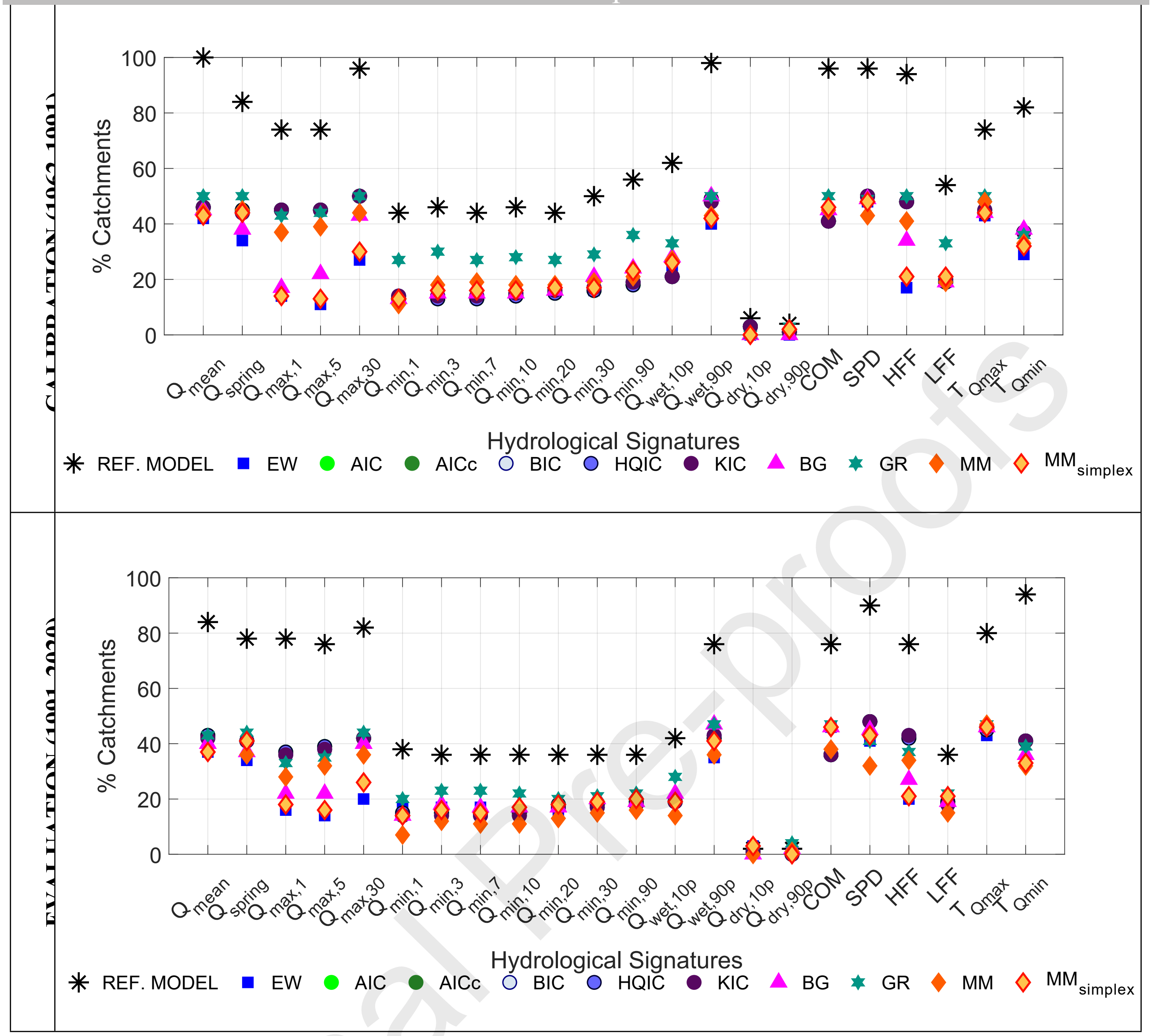


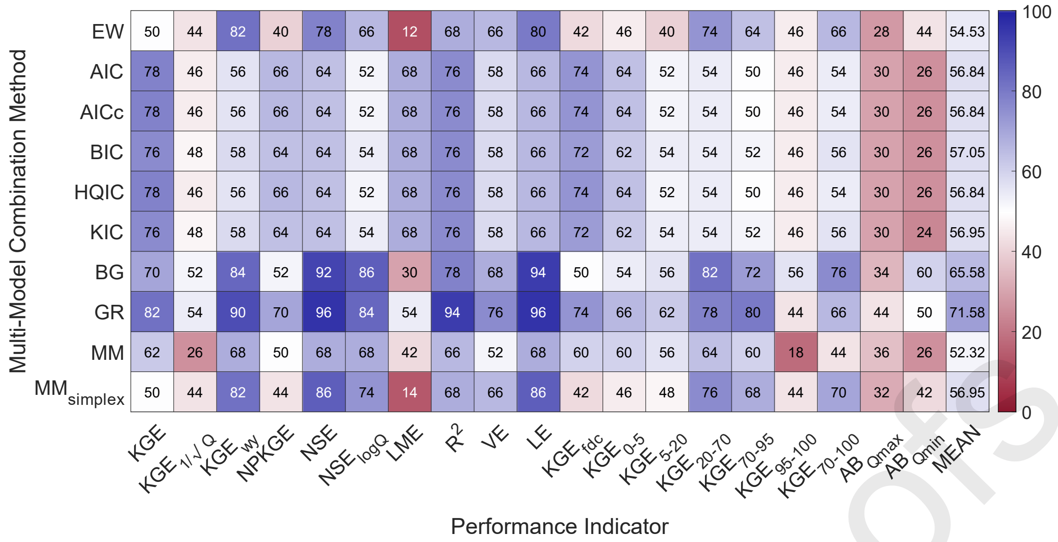
Figure 3. The percentage of catchments in which a multi-model combination method outperforms the reference model in reproducing distributions of the hydrological signatures (Table 4) according to the results of the Wilcoxon rank sum test in the calibration (top) and evaluation periods (bottom panel).

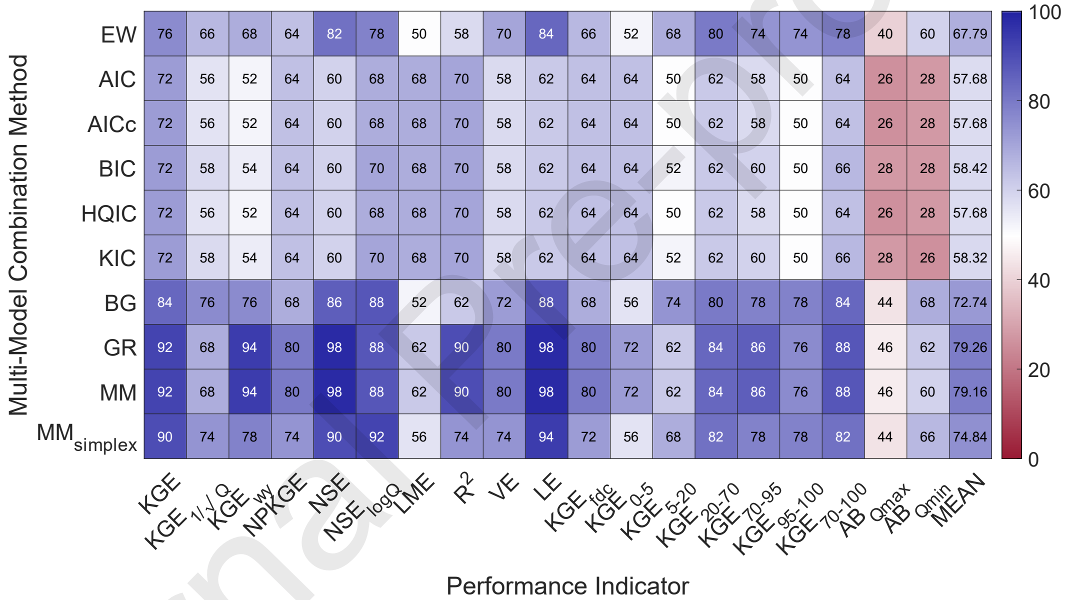
3.2 Impact of Preselection of Candidate Models on Performance of Model Combinations

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 ($AB_{Q\mathrm{max}}$, $AB_{Q\mathrm{min}}$), 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 $MM_{simplex}$, and in GR and MM in the evaluation period. Conversely, the information criteria-related MMCMs exhibit fairly consistent performance across all ensembles (including E_{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_{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_{uni} in the calibration; Figure 5), even though size of these ensembles varies considerably from five (E_5) 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.


Multi-Model Combination Method
EW 50 44 82 40 78 66 12 68 66 80 42 46 40 74 64 46 66 28 44 54.53
AIC 78 46 56 66 64 52 68 76 58 66 74 64 52 54 50 46 54 30 26 56.84
AICc 78 46 56 66 64 52 68 76 58 66 74 64 52 54 50 46 54 30 26 56.84
BIC 76 48 58 64 64 54 68 76 58 66 72 62 54 54 52 46 56 30 26 57.05
HQIC 78 46 56 66 64 52 68 76 58 66 74 64 52 54 50 46 54 30 26 56.84
KIC 76 48 58 64 64 54 68 76 58 66 72 62 54 54 52 46 56 30 24 56.95
BG 70 52 84 52 92 86 30 78 68 94 50 54 56 82 72 56 76 34 60 65.58
GR 82 54 90 70 96 84 54 94 76 96 74 66 62 78 80 44 66 44 50 71.58
MM 62 26 68 50 68 68 42 66 52 68 60 60 56 64 60 18 44 36 26 52.32
MMsimplex 50 44 82 44 86 74 14 68 66 86 42 46 48 76 68 44 70 32 42 56.95
100 80 60 40 20 0
KGE
KGE 1/√Q
KGE wy
NPKGE
NSE
NSE logQ
LME
R2
VE
LE
KGE fdc
KGE 0-5
KGE 5-20
KGE 20-70
KGE 70-95
KGE 95-100
KGE 70-100
AB Qmax
AB Qmin
MEAN
Performance Indicator


Multi-Model Combination Method
EW 76 66 68 64 82 78 50 58 70 84 66 52 68 80 74 74 78 40 60 67.79
AIC 72 56 52 64 60 68 68 70 58 62 64 64 50 62 58 50 64 26 28 57.68
AICc 72 56 52 64 60 68 68 70 58 62 64 64 50 62 58 50 64 26 28 57.68
BIC 72 58 54 64 60 70 68 70 58 62 64 64 52 62 60 50 66 28 28 58.42
HQIC 72 56 52 64 60 68 68 70 58 62 64 64 50 62 58 50 64 26 28 57.68
KIC 72 58 54 64 60 70 68 70 58 62 64 64 52 62 60 50 66 28 26 58.32
BG 84 76 76 68 86 88 52 62 72 88 68 56 74 80 78 78 84 44 68 72.74
GR 92 68 94 80 98 88 62 90 80 98 80 72 62 84 86 76 88 46 62 79.26
MM 92 68 94 80 98 88 62 90 80 98 80 72 62 84 86 76 88 46 60 79.16
MMsimplex 90 74 78 74 90 92 56 74 74 94 72 56 68 82 78 78 82 44 66 74.84
100 80 60 40 20 0
KGE
KGE 1/√Q
KGE wy
NPKGE
NSE
NSE logQ
LME
R2
VE
LE
KGE fdc
KGE 0-5
KGE 5-20
KGE 20-70
KGE 70-95
KGE 95-100
KGE 70-100
AB Qmax
AB Qmin
MEAN
Performance Indicator

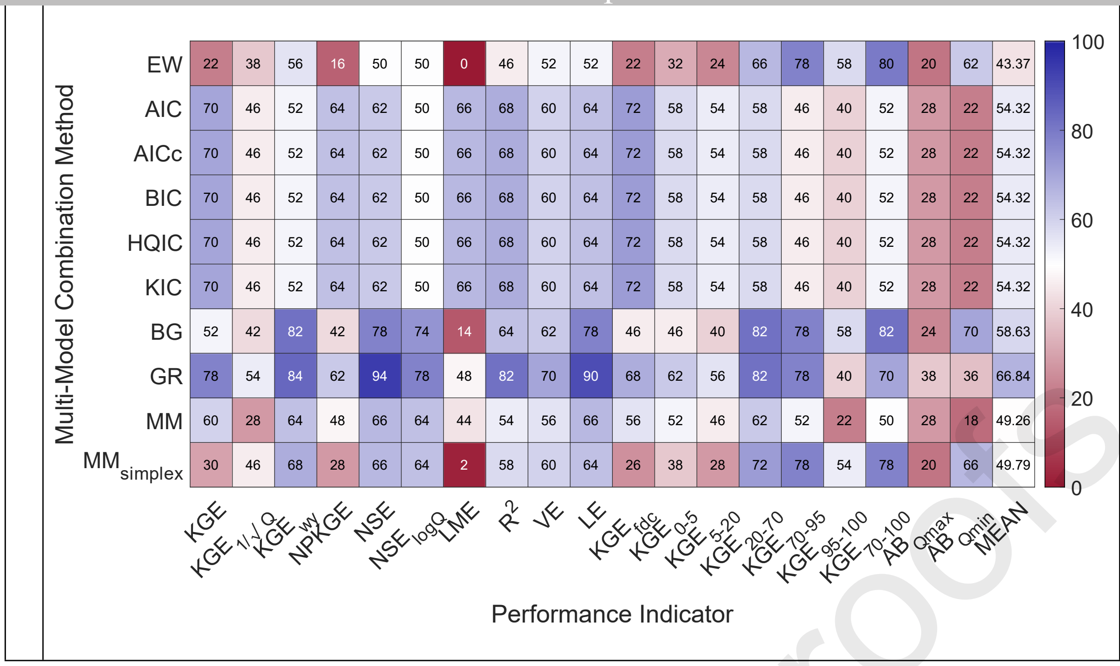


Figure 4. The percentage of the catchments in which a model combination outperforms the reference model according to a specific performance indicator (Table 3) in the evaluation period. The model combinations are obtained with the daily flow series and with three different ensembles (Table 5) that include 25 (E_1) and 5 (E_5) candidate models, and 21 candidate models from different groups (E_{uni}).

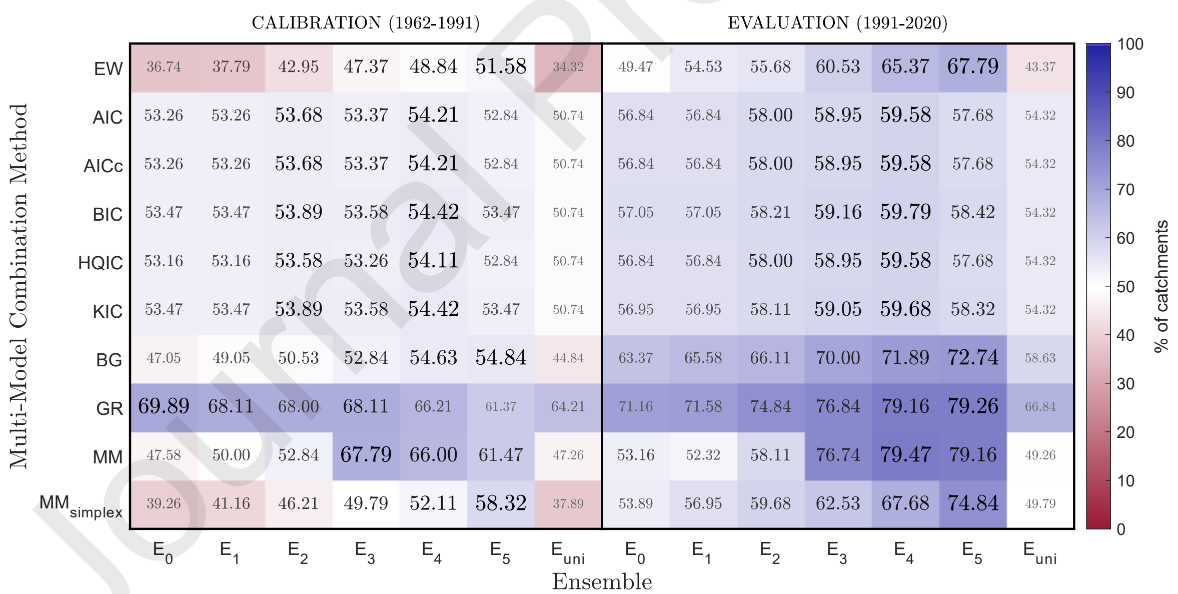


Figure 5. The average performance of the model combinations obtained with all ensembles (Table 5). The cell values show the percentage of catchments in which the model combination outperforms the reference model, averaged over all performance indicators (Table 3). The font size is adjusted to indicate 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 (E_0) yield a higher percentage of catchments with properly reproduced distributions in many instances, especially in the evaluation period (Figure S8). Specifically, E_0 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).

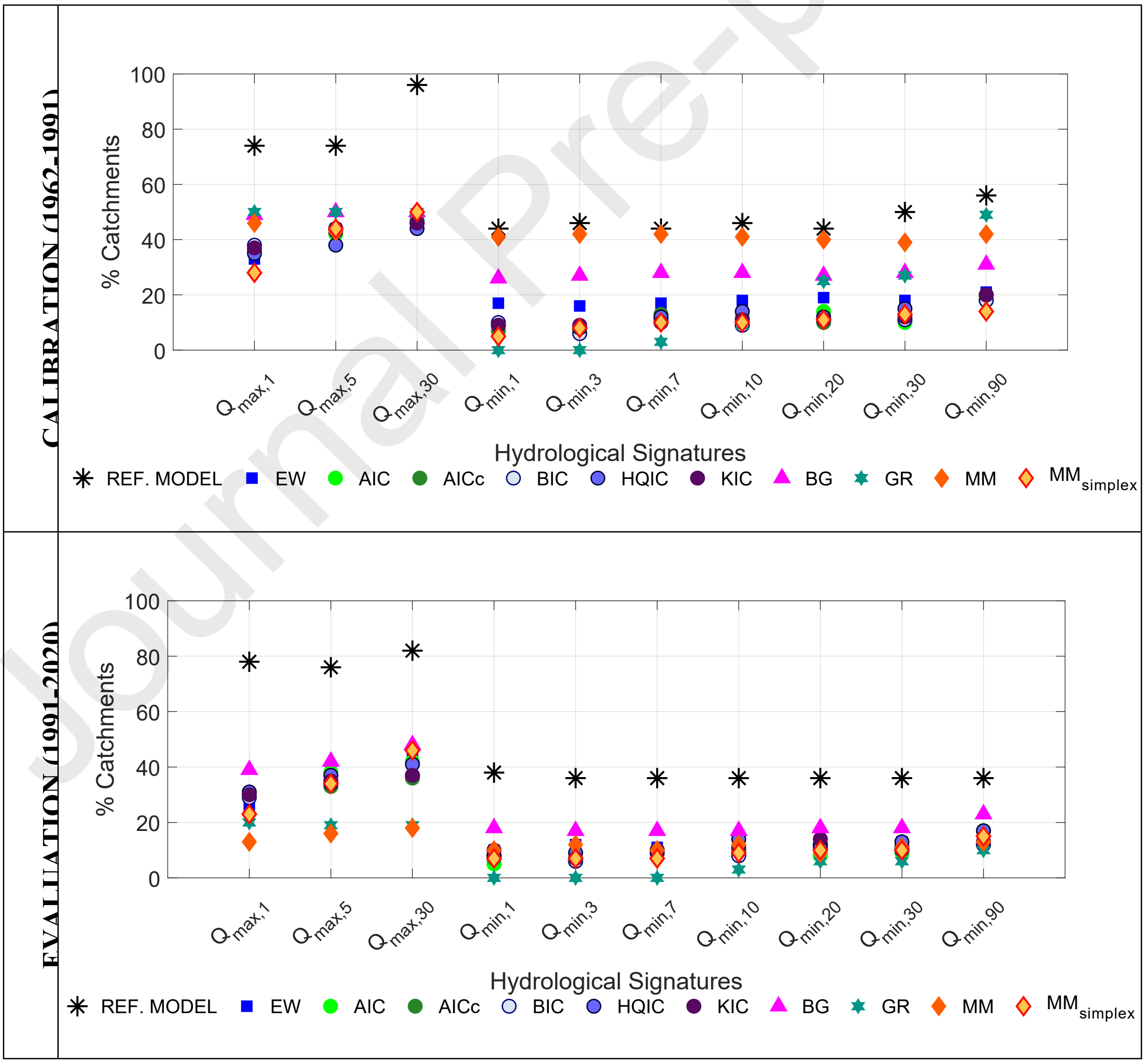


Figure 6. The percentage of catchments in which model combinations obtained with the series of annual maxima and minima of various durations, and the reference model properly reproduce distributions of these signatures in the calibration (top) and evaluation periods (bottom panels).

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 $AB_{Q\mathrm{min}}$), 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 $KGE_{70\text{-}100}$, 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 suggest 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.

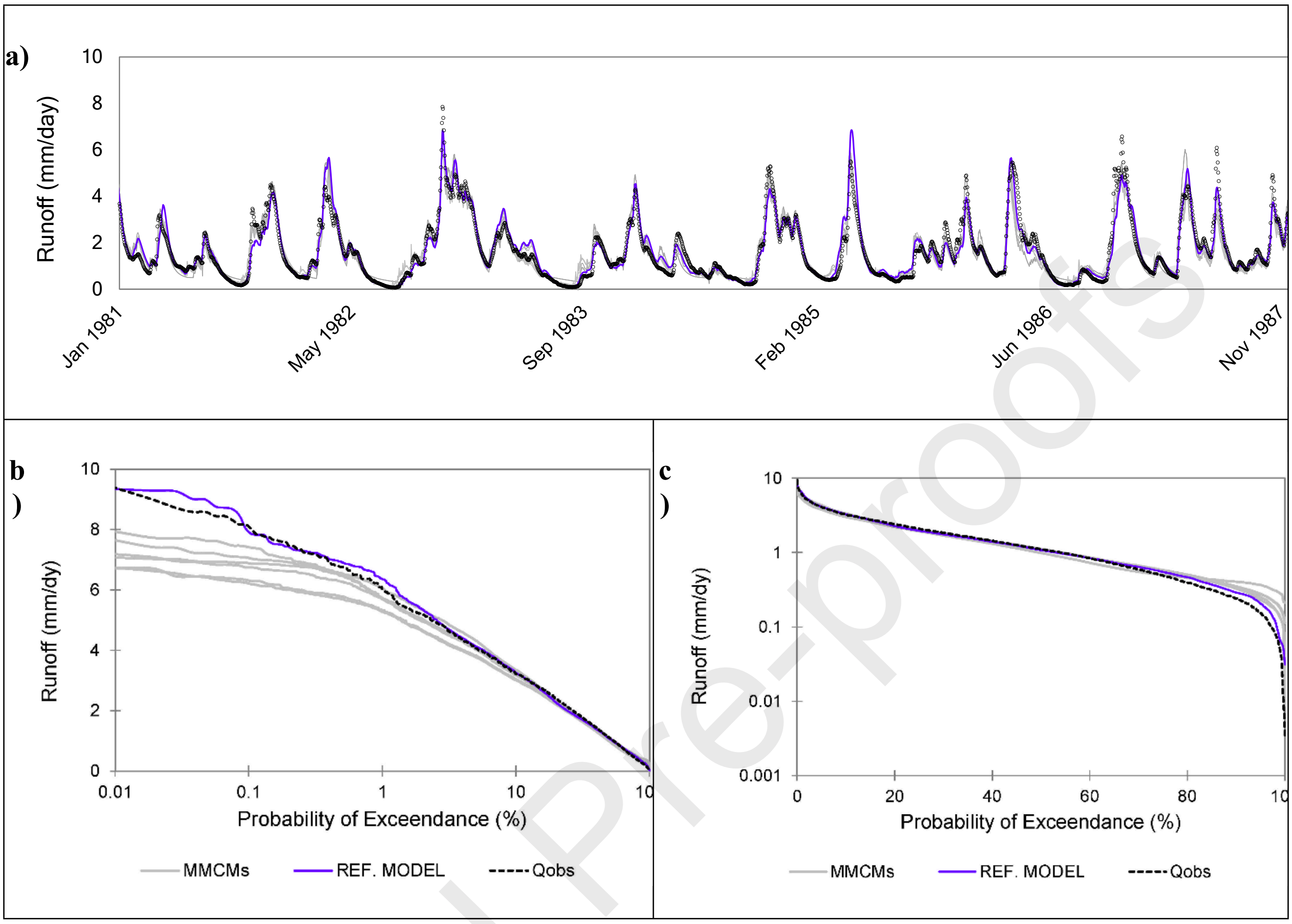


Figure 7. Runoff in the Vassbotten catchment in the calibration period: a) observed (Q_{obs}, black line) and simulated runoff with the best performing individual model (purple line), and with ten multi-model combination methods (grey lines), and flow duration curves with b) logarithmically scaled abscissa to emphasise high flows and c) logarithmically scaled ordinate to emphasise low flows.

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_{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., Vukmirović 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 al., 2020), 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 extend 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, $MM_{simplex}$. Further research is needed to analyses 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 irresponsive 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

Agnihotri, J., Coulibaly, P., 2020. Evaluation of Snowmelt Estimation Techniques for Enhanced Spring Peak Flow Prediction. Water 12, 1290. https://doi.org/10.3390/w12051290

Ajami, N.K., Duan, Q., Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resour. Res. 43, 1–19. https://doi.org/10.1029/2005WR004745

Andreassian, V., Bergström, S., Chahinian, N., Duan, Q., Gusev, Y.M., Littlewood, I., Mathevet, T., Michel, C., Montanari, A., Moretti, G., Moussa, R., Nasonova, O.N., O'Connor, K., Paquet, E., Perrin, C., Rousseau, A., Schaake, J., Wagener, T., Xie, Z., Garç, R., Gailhard, J., Croke, B., 2006. Catalogue of the models used in MOPEX 2004/2005. IAHS-AISH Publ. 41–93.

Arnold, J.G., Allen, P.M., Muttiah, R., Bernhardt, G., 1995. Automated Base Flow Separation and Recession Analysis Techniques. Ground Water 33, 1010–1018. https://doi.org/10.1111/j.1745-6584.1995.tb00046.x

Arsenault, R., Essou, G.R.C., Brissette, F.P., 2017. Improving Hydrological Model Simulations with Combined Multi-Input and Multimodel Averaging Frameworks. J. Hydrol. Eng. 22, 04016066. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001489

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. J. Hydrol. 529, 754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001

Bastola, S., Murphy, C., Sweeney, J., 2011. The role of hydrological modelling uncertainties in climate change impact assessments of Irish river catchments. Adv. Water Resour. 34, 562–576. https://doi.org/10.1016/j.advwatres.2011.01.008

Bates, J.M., Granger, C., 1969. Combination of Forecasts. Oper. Res. Q. 20, 451–468. https://doi.org/10.2307/3008764

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Process. 6, 279–298. https://doi.org/10.1002/hyp.3360060305

Bhat, K., Haran, M., Terando, A., Keller, K., 2011. Climate projections using Bayesian model averaging and space-time dependence. J. Agric. Biol. Environ. Stat. 16, 606–628.

Block, P.J., Souza Filho, F.A., Sun, L., Kwon, H.-H., 2009. A Streamflow Forecasting Framework using Multiple Climate and Hydrological Models. JAWRA J. Am. Water Resour. Assoc. 45, 828–843. https://doi.org/10.1111/j.1752-1688.2009.00327.x

Blöschl, G., Bierkens, M.F.P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J.W., McDonnell, J.J., Savenije, H.H.G., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S.T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S.K., Baker, V., Bardsley, E., Barendrecht, M.H., Bartosova, A., Batelaan, O., Berghuijs, W.R., Beven, K., Blume, T., Bogaard, T., Borges de Amorim, P., Böttcher, M.E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Yangbo, Chen, Yuanfang, Chifflard, P., Claps, P., Clark, M.P., Collins, A.L., Croke, B., Dathe, A., David, P.C., de Barros, F.P.J., de Rooij, G., Di Baldassarre, G., Driscoll, J.M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W.H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Gonzalez Bevacqua, A., González-Dugo, M.P., Grimaldi, S., Gupta, A.B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T.H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., Kalantari, Z., Kalvans, A., Khanal, S., Khatami, S., Kiesel, J., Kirkby, M., Knoben, W., Kochanek, K., Kohnová, S., Kolechkina, A., Krause, S., Kreamer, D., Kreibich, H., Kunstmann, H., Lange, H., Liberato, M.L.R., Lindquist, E., Link, T., Liu, J., Loucks, D.P., Luce, C., Mahé, G., Makarieva, O., Malard, J., Mashtayeva, S., Maskey, S., Mas-Pla, J., Mavrova-Guirguinova, M., Mazzoleni, M., Mernild, S., Misstear, B.D., Montanari, A., Müller-Thomy, H., Nabizadeh, A., Nardi, F., Neale, C., Nesterova, N., Nurtaev, B., Odongo, V.O., Panda, S., Pande, S., Pang, Z., Papacharalampous, G., Perrin, C., Pfister, L., Pimentel, R., Polo, M.J., Post, D., Prieto Sierra, C., Ramos, M.-H., Renner, M., Reynolds, J.E., Ridolfi, E., Rigon, R., Riva, M., Robertson, D.E., Rosso, R., Roy, T., Sá, J.H.M., Salvadori, G., Sandells, M., Schaefli, B., Schumann, A., Scolobig, A., Seibert, J., Servat, E., Shafiei, M., Sharma, A., Sidibe, M., Sidle, R.C., Skaugen, T., Smith, H., Spiessl, S.M., Stein, L., Steinsland, I., Strasser, U., Su, B., Szolgay, J., Tarboton, D., Tauro, F., Thirel, G., Tian, F., Tong, R., Tussupova, K., Tyralis, H., Uijlenhoet, R., van Beek, R., van der Ent, R.J., van der Ploeg, M., Van Loon, A.F., van Meerveld, I., van Nooijen, R., van Oel, P.R., Vidal, J.-P., von Freyberg, J., Vorogushyn, S., Wachniew, P., Wade, A.J., Ward, P., Westerberg, I.K., White, C., Wood, E.F., Woods, R., Xu, Z., Yilmaz, K.K., Zhang, Y., 2019. Twenty-three unsolved problems in hydrology (UPH) – a community perspective. Hydrological Sciences Journal 64, 1141–1158. https://doi.org/10.1080/02626667.2019.1620507

Booij, M.J., Krol, M.S., 2010. Balance between calibration objectives in a conceptual hydrological model. Hydrol. Sci. J. 55, 1017–1032. https://doi.org/10.1080/02626667.2010.505892

Broderick, C., Matthews, T., Wilby, R.L., Bastola, S., Murphy, C., 2016. Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. Water Resour. Res. 52, 8343–8373. https://doi.org/10.1002/2016WR018850

Brunner, M.I., Slater, L., Tallaksen, L.M., Clark, M., 2021. Challenges in modeling and predicting floods and droughts: A review. WIREs Water 1–32. https://doi.org/10.1002/wat2.1520

Bum Kim, K., Kwon, H.-H., Han, D., 2021. Bias-correction schemes for calibrated flow in a conceptual hydrological model. Hydrol. Res. 196–211. https://doi.org/10.2166/nh.2021.043

Chen, J., Brissette, F.P., Lucas-Picher, P., Caya, D., 2017. Impacts of weighting climate models for hydro-meteorological climate change studies. J. Hydrol. 549, 534–546. https://doi.org/10.1016/j.jhydrol.2017.04.025

Chiew, F.H.S., Kirono, D.G.C., Kent, D.M., Frost, A.J., Charles, S.P., Timbal, B., Nguyen, K.C., Fu, G., 2010. Comparison of runoff modelled using rainfall from different downscaling methods for historical and future climates. J. Hydrol. 387, 10–23. https://doi.org/10.1016/j.jhydrol.2010.03.025

Chiew, F.H.S., Teng, J., Vaze, J., Post, D.A., Perraud, J.M., Kirono, D.G.C., Viney, N.R., 2009. Estimating climate change impact on runoff across southeast Australia: Method, results, and implications of the modeling method. Water Resour. Res. 45, 1–17. https://doi.org/10.1029/2008WR007338

Claeskens, G., 2016. Statistical Model Choice. Annu. Rev. Stat. Its Appl. 3, 233–256. https://doi.org/10.1146/annurev-statistics-041715-033413

Claeskens, G., Cunen, C., Hjort, N.L., 2019. Model Selection via Focused Information Criteria for Complex Data in Ecology and Evolution. Front. Ecol. Evol. 7, 1–13. https://doi.org/10.3389/fevo.2019.00415

Claeskens, G., Hjort, N.L., 2003. The Focused Information Criterion. J. Am. Stat. Assoc. 98, 900–916. https://doi.org/10.1198/016214503000000819

Claeskens, G., Hjort, N.L., 2001. Model Selection and Model Averaging. Cambridge University Press. https://doi.org/10.1017/CBO9780511790485

Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R.A., Vrugt, J.A., Gupta, H. V., Wagener, T., Hay, L.E., 2008. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resour. Res. 44, W0B02, 1-14. https://doi.org/10.1029/2007WR006735

CORINE Land Cover [WWW Document], 2012. URL https://land.copernicus.eu/pan-european/corine-land-cover/clc-2012?tab=mapview

Criss, R.E., Winston, W.E., 2008. Do Nash values have value ? Discussion and alternate proposals. Hydrol. Process. 22, 2723–2725. https://doi.org/10.1002/hyp

Crochemore, L., Perrin, C., Andréassian, V., Ehret, U., Seibert, S.P., Grimaldi, S., Gupta, H., Paturel, J.-E., 2015. Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrol. Sci. J. https://doi.org/10.1080/02626667.2014.903331

Croke, B., Jakeman, A., 2004. A catchment moisture deficit module for the IHACRES rainfall-runoff model. Environ. Model. Softw. 19, 1–5. https://doi.org/10.1016/j.envsoft.2003.09.001

Cunderlik, J.M., Ouarda, T.B.M.J., 2009. Trends in the timing and magnitude of floods in Canada. J. Hydrol. 375, 471–480. https://doi.org/10.1016/j.jhydrol.2009.06.050

D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, T. L. Veith, 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Trans. ASABE 50, 885–900. https://doi.org/10.13031/2013.23153

Dankers, R., Arnell, N.W., Clark, D.B., Falloon, P.D., Fekete, B.M., Gosling, S.N., Heinke, J., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., Wisser, D., 2014. First look at changes in flood hazard in the Inter-Sectoral Impact Model Intercomparison Project ensemble. Proc. Natl. Acad. Sci. 111, 3257–3261. https://doi.org/10.1073/pnas.1302078110

Daraio, J.A., 2020. Hydrologic Model Evaluation and Assessment of Projected Climate Change Impacts Using Bias-Corrected Stream Flows. Water 12, 2312. https://doi.org/10.3390/w12082312

Darbandsari, P., Coulibaly, P., 2020. Introducing entropy-based Bayesian model averaging for streamflow forecast. J. Hydrol. 591, 125577. https://doi.org/10.1016/j.jhydrol.2020.125577

DHI, 2017. A Modelling System for Rivers and Channels Reference Manual.

Di Baldassarre, G., Laio, F., Montanari, A., 2009. Design flood estimation using model selection criteria. Phys. Chem. Earth, Parts A/B/C 34, 606–611. https://doi.org/10.1016/j.pce.2008.10.066

Diks, C.G.H., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Environ. Res. Risk Assess. 24, 809–820. https://doi.org/10.1007/s00477-010-0378-z

Dormann, C.F., Calabrese, J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J.J., Pollock, L.J., Reineking, B., Roberts, D.R., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Wood, S.N., Wüest, R.O., Hartig, F., 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecol. Monogr. 88, 485–504. https://doi.org/10.1002/ecm.1309

Dusa, S., Manikanta, V., Das, J., Umamahesh, N.V., 2023. Does the performance enhancement through multi-model averaging at the catchment outlet gets translated to the interior ungauged points? Journal of Hydrology 627, 130389. https://doi.org/10.1016/j.jhydrol.2023.130389

Eklund, A., 2011. SVAR, Svenskt vattenarkiv (No. 53). Norrköping, Sweden.

Farmer, W.H., Over, T.M., Kiang, J.E., 2018. Bias correction of simulated historical daily streamflow at ungauged locations by using independently estimated flow duration curves. Hydrol. Earth Syst. Sci. 22, 5741–5758. https://doi.org/10.5194/hess-22-5741-2018

Fatehifar, A., Goodarzi, M.R., Montazeri Hedesh, S.S., Siahvashi Dastjerdi, P., 2021. Assessing watershed hydrological response to climate change based on signature indices. Journal of Water and Climate Change 12, 2579–2593. https://doi.org/10.2166/wcc.2021.293

Fenicia, F., Kavetski, D., Savenije, H.H.G., 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resour. Res. 47, 1–13. https://doi.org/10.1029/2010WR010174

Fenicia, F., Savenije, H.H.G., Matgen, P., Pfister, L., 2008. Understanding catchment behavior through stepwise model concept improvement. Water Resour. Res. 44, W01402, 1–13. https://doi.org/10.1029/2006WR005563

Fischer, A.M., Weigel, A.P., Buser, C.M., Knutti, R., Künsch, H.R., Liniger, M.A., Schär, C., Appenzeller, C., 2012. Climate change projections for Switzerland based on a Bayesian multi-model approach. Int. J. Climatol. 32, 2348–2371. https://doi.org/10.1002/joc.3396

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., Woods, R., Zhang, L., 2018. Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement. Water Resour. Res. 54, 9812–9832. https://doi.org/10.1029/2018WR023989

Francois, D., 2021. HMETS hydrological model.

Garavaglia, F., Le Lay, M., Gottardi, F., Garçon, R., Gailhard, J., Paquet, E., Mathevet, T., 2017. Impact of model structure on flow simulation and hydrological realism: from a lumped to a semi-distributed approach. Hydrol. Earth Syst. Sci. 21, 3937–3952. https://doi.org/10.5194/hess-21-3937-2017

Garcia, F., Folton, N., Oudin, L., 2017. Which objective function to calibrate rainfall–runoff models for low-flow index simulations? Hydrol. Sci. J. 62, 1149–1166. https://doi.org/10.1080/02626667.2017.1308511

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., 2017. A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C. Clim. Change 141, 577–595. https://doi.org/10.1007/s10584-016-1773-3

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. J. Forecast. 3, 197–204. https://doi.org/10.1002/for.3980030207

Gupta, H. V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gutiérrez, J.M., R.G. Jones, G.T. Narisma, L.M. Alves, M. Amjad, I.V. Gorodetskaya, M. Grose, N.A.B. Klutse, S., Krakovska, J. Li, D. Martínez-Castro, L.O. Mearns, S.H. Mernild, T. Ngo-Duc, B. van den H., Yoon, J.-H., Masson-Delmotte, V., P. Zhai, A. Pirani, S.L., Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R., Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B.Z., 2021. Atlas. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (In Press). WMO; UNEP.

Hales, R.C., Williams, G.P., James Nelson, E., Sowby, R.B., Ames, D.P., Lozano, J.L.S., 2023. Bias correcting discharge simulations from the GEOGloWS global hydrologic model. Journal of Hydrology 626, 130279. https://doi.org/10.1016/j.jhydrol.2023.130279

Hamon, W.R., 1961. Estimating potential evaporation, in: Proceedings of the American Society of Civil Engineers, Division, J.o.H. pp. 107–120.

HBV-light, 2020.

Henestål, J., Ranung, J., Gyllander, G., Johnson, Å., Olsson, H., Pettersson, O., Westman, Y., Wingqvist, E.-M., 2012. Arbete med SVAR version 2012_1 och 2012_2, Svenskt Vattenarkiv, en databas vid SMHI. DM #154591.

Hoang, L.P., Lauri, H., Kummu, M., Koponen, J., Vliet, M.T.H. Van, Supit, I., Leemans, R., Kabat, P., Ludwig, F., 2016. Mekong River flow and hydrological extremes under climate change. Hydrol. Earth Syst. Sci. 20, 3027–3041. https://doi.org/10.5194/hess-20-3027-2016

Höge, M., Guthke, A., Nowak, W., 2019. The hydrologist's guide to Bayesian model selection, averaging and combination. J. Hydrol. 572, 96–107. https://doi.org/10.1016/j.jhydrol.2019.01.072

Houghton-Carr, H.A., 1999. Assessment criteria for simple conceptual daily rainfall-runoff models. Hydrol. Sci. J. 44, 237–261. https://doi.org/10.1080/02626669909492220

Huang, S., Shah, H., Naz, B.S., Shrestha, N., Mishra, V., Daggupati, P., Ghimire, U., Vetter, T., 2020. Impacts of hydrological model calibration on projected hydrological changes under climate change—a multi-model assessment in three large river basins. Clim. Change 163, 1143–1164. https://doi.org/10.1007/s10584-020-02872-6

Johansson, B., 2000. Areal Precipitation and Temperature in the Swedish Mountains. Hydrol. Res. 31, 207–228. https://doi.org/10.2166/nh.2000.0013

Kendall, M.G., 1938. A New Measure of Rank Correlation. Biometrika 30, 81. https://doi.org/10.2307/2332226

Kiesel, J., Stanzel, P., Kling, H., Fohrer, N., Jähnig, S.C., Pechlivanidis, I., 2020. Streamflow-based evaluation of climate model sub-selection methods. Clim. Change 163, 1267–1285. https://doi.org/10.1007/s10584-020-02854-8

Kim, H.S., Croke, B.F.W., Jakeman, A.J., Chiew, F.H.S., 2011. An assessment of modelling capacity to identify the impacts of climate variability on catchment hydrology. Mathematics and Computers in Simulation 81, 1419–1429. https://doi.org/10.1016/j.matcom.2010.05.007

Kiraz, M., Coxon, G., Wagener, T., 2023. A Signature-Based Hydrologic Efficiency Metric for Model Calibration and Evaluation in Gauged and Ungauged Catchments. Water Resources Research 59, e2023WR035321. https://doi.org/10.1029/2023WR035321

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. 31, 13–24. https://doi.org/10.1080/02626668609491024

Kling, H., Stanzel, P., Fuchs, M., Nachtnebel, H.-P., 2015. Performance of the COSERO precipitation–runoff model under non-stationary conditions in basins with different climates. Hydrol. Sci. J. 60, 1374–1393. https://doi.org/10.1080/02626667.2014.959956

Knoben, W.J.M., Freer, J.E., Fowler, K.J.A., Peel, M.C., Woods, R.A., 2019a. Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. Geosci. Model Dev. 12, 2463–2480. https://doi.org/10.5194/gmd-12-2463-2019

Knoben, W.J.M., Freer, J.E., Fowler, K.J.A., Peel, M.C., Woods, R.A., 2019b. Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations - Supplement. Geosci. Model Dev. 12, 2463–2480. https://doi.org/10.5194/gmd-12-2463-2019

Knoben, W.J.M., Freer, J.E., Peel, M.C., Fowler, K.J.A., Woods, R.A., 2020. A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. Water Resour. Res. 56, 1–24. https://doi.org/10.1029/2019WR025975

Kormos, P.R., Luce, C.H., Wenger, S.J., Berghuijs, W.R., 2016. Trends and sensitivities of low streamflow extremes to discharge timing and magnitude in Pacific Northwest mountain streams. Water Resour. Res. 52, 4990–5007. https://doi.org/10.1002/2015WR018125

Krause, P., Boyle, D.P., Base, F., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97. https://doi.org/10.5194/adgeo-5-89-2005

Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., Kundzewicz, Z.W., 2018. How the performance of hydrological models relates to credibility of projections under climate change. Hydrol. Sci. J. 63, 696–720. https://doi.org/10.1080/02626667.2018.1446214

Krysanova, V., Vetter, T., Eisner, S., Huang, S., Pechlivanidis, I., Strauch, M., Gelfan, A., Kumar, R., Aich, V., Arheimer, B., Chamorro, A., van Griensven, A., Kundu, D., Lobanova, A., Mishra, V., Plötner, S., Reinhardt, J., Seidou, O., Wang, X., Wortmann, M., Zeng, X., Hattermann, F.F., 2017. Intercomparison of regional-scale hydrological models and climate change impacts projected for 12 large river basins worldwide—a synthesis. Environ. Res. Lett. 12, 105002. https://doi.org/10.1088/1748-9326/aa8359

Kvam, P.H., Vidakovic, B., 2007. Nonparametric Statistics with Applications to Science and Engineering. Wiley-interscience, New Jersey.

Lane, R.A., Coxon, G., Freer, J.E., Wagener, T., Johnes, P.J., Bloomfield, J.P., Greene, S., Macleod, C.J.A., Reaney, S.M., 2019. Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain. Hydrology and Earth System Sciences 23, 4011–4032. https://doi.org/10.5194/hess-23-4011-2019

Lee, Y., Song, J., 2021. Robustness of model averaging methods for the violation of standard linear regression assumptions. Commun. Stat. Appl. Methods 28, 189–204. https://doi.org/10.29220/CSAM.2021.28.2.189

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241. https://doi.org/10.1029/1998WR900018

Liang, H., Zou, G., Wan, A.T.K., Zhang, X., 2011. Optimal Weight Choice for Frequentist Model Average Estimators. J. Am. Stat. Assoc. 106, 1053–1066. https://doi.org/10.1198/jasa.2011.tm09478

Liu, D., 2020. A rational performance criterion for hydrological model. J. Hydrol. 590, 125488. https://doi.org/10.1016/j.jhydrol.2020.125488

Lute, A.C., Luce, C.H., 2017. Are Model Transferability and Complexity Antithetical? Insights From Validation of a Variable-Complexity Empirical Snow Model in Space and Time. Water Resour. Res. 53, 8825–8850. https://doi.org/10.1002/2017WR020752

Mann, H.B., 1945. Nonparametric Tests Against Trend. Econometrica 13, 245. https://doi.org/10.2307/1907187

Martel, J.L., Demeester, K., Brissette, F., Poulin, A., Arsenault, R., 2017. HMETS—A Simple and Efficient Hydrology Model forTeaching Hydrological Modelling, Flow Forecasting andClimate Change Impacts. Int. J. Eng. Educ. 33, 1307–1316.

Mathevet, T., Gupta, H., Perrin, C., Andréassian, V., Le Moine, N., 2020. Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds. J. Hydrol. 585, 124698. https://doi.org/10.1016/j.jhydrol.2020.124698

Maurer, E.P., Brekke, L.D., Pruitt, T., 2010. Contrasting Lumped and Distributed Hydrology Models for Estimating Climate Change Impacts on California Watersheds1. JAWRA J. Am. Water Resour. Assoc. 46, 1024–1035. https://doi.org/10.1111/j.1752-1688.2010.00473.x

Mendoza, P.A., Clark, M.P., Mizukami, N., Newman, A.J., Barlage, M., Gutmann, E.D., Rasmussen, R.M., Rajagopalan, B., Brekke, L.D., Arnold, J.R., 2015. Effects of Hydrologic Model Choice and Calibration on the Portrayal of Climate Change Impacts. J. Hydrometeorol. 16, 762–780. https://doi.org/10.1175/JHM-D-14-0104.1

Mitra, P., Lian, H., Mitra, R., Liang, H., Xie, M., 2019. A general framework for frequentist model averaging. Sci. China Math. 62, 205–226. https://doi.org/10.1007/s11425-018-9403-x

Mizukami, N., Rakovec, O., Newman, A.J., Clark, M.P., Wood, A.W., Gupta, H. V., Kumar, R., 2019. On the choice of calibration metrics for “high-flow” estimation using hydrologic models. Hydrol. Earth Syst. Sci. 23, 2601–2614. https://doi.org/10.5194/hess-23-2601-2019

Montgomery, D.C., Runger, G.C., 2003. Applied Statistics and Probability for Engineers, Third Edit. ed. John Wiley & Sons.

Moore, R.J., 2007. The PDM rainfall-runoff model. Hydrol. Earth Syst. Sci. 11, 483–499. https://doi.org/10.5194/hess-11-483-2007

Moore, R.J., Bell, V.A., 2002. Incorporation of groundwater losses and well level data in rainfall-runoff models illustrated using the PDM. Hydrol. Earth Syst. Sci. 6, 25–38. https://doi.org/10.5194/hess-6-25-2002

Najafi, M.R., Moradkhani, H., 2015. Multi-model ensemble analysis of runoff extremes for climate change impact assessments. J. Hydrol. 525, 352–361. https://doi.org/10.1016/j.jhydrol.2015.03.045

Najafi, M.R., Moradkhani, H., Jung, I.W., 2011. Assessing the uncertainties of hydrologic model selection in climate change impact studies. Hydrol. Process. 25, 2814–2826. https://doi.org/10.1002/hyp.8043

Nash, J.E., Sutcliffe, J. V., 1970. River flow forecasting through conceptual models, Part I - A discussion of principles. J. Hydrol. 10, 282–290.

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci. 19, 209–223. https://doi.org/10.5194/hess-19-209-2015

Nijzink, R., Hutton, C., Pechlivanidis, I., Capell, R., Arheimer, B., Freer, J., Han, D., Wagener, W., McGuire, K., Savenije, H., Hrachowitz, M., 2016. The evolution of root zone moisture capacities after land use change: a step towards predictions under change? Hydrol. Earth Syst. Sci. 20, 4775–4799. https://doi.org/10.5194/hess-2016-427

Okoli, K., Breinl, K., Brandimarte, L., Botto, A., Volpi, E., Di Baldassarre, G., 2018. Model averaging versus model selection: estimating design floods with uncertain river flow data. Hydrol. Sci. J. 63, 1913–1926. https://doi.org/10.1080/02626667.2018.1546389

Olden, J.D., Poff, N.L., 2003. Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River Res. Appl. 19, 101–121. https://doi.org/10.1002/rra.700

Oliveira, A.R., Ramos, T.B., Pinto, L., Neves, R., 2023. Direct integration of reservoirs’ operations in a hydrological model for streamflow estimation: coupling a CLSTM model with MOHID-Land. Hydrol. Earth Syst. Sci. 27, 3875–3893. https://doi.org/10.5194/hess-27-3875-2023

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., Michel, C., 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resources Research 42, W07410, 1–10. https://doi.org/10.1029/2005WR004636

Parajka, J., Blaschke, A.P., Blöschl, G., Haslinger, K., Hepp, G., Laaha, G., Schöner, W., Trautvetter, H., Viglione, A., Zessner, M., 2016. Uncertainty contributions to low flow projections in Austria. Hydrol. Earth Syst. Sci. 20, 2085–2101. https://doi.org/10.5194/hess-20-2085-2016

Pechlivanidis, I.G.G., Arheimer, B., Donnelly, C., Hundecha, Y., Huang, S., Aich, V., Samaniego, L., Eisner, S., Shi, P., 2016. Analysis of hydrological extremes at different hydro-climatic regimes under present and future conditions. Clim. Change 141, 467–481. https://doi.org/10.1007/s10584-016-1723-0

Pechlivanidis, I.G.G., Jackson, B.M., Mcintyre, N.R., Wheater, H.S., 2013. Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. Glob. NEST J. 13, 193–214. https://doi.org/10.30955/gnj.000778

Perra, E., Piras, M., Deidda, R., Paniconi, C., Mascaro, G., Vivoni, E.R., Cau, P., Marras, P.A., Ludwig, R., Meyer, S., 2018. Multimodel assessment of climate change-induced hydrologic impacts for a Mediterranean catchment. Hydrol. Earth Syst. Sci. 22, 4125–4143. https://doi.org/10.5194/hess-22-4125-2018

Perrin, C., Michel, C., Andre, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279, 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Pfannerstill, M., Guse, B., Fohrer, N., 2014. Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. J. Hydrol. 510, 447–458. https://doi.org/10.1016/j.jhydrol.2013.12.044

Pool, S., Vis, M., Seibert, J., 2018. Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. Hydrol. Sci. J. 63, 1941–1953. https://doi.org/10.1080/02626667.2018.1552002

Posada, D., Buckley, T.R., 2004. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. Syst. Biol. 53, 793–808. https://doi.org/10.1080/10635150490522304

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., Andréassian, V., 2011. A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. J. Hydrol. 411, 66–76. https://doi.org/10.1016/j.jhydrol.2011.09.034

Qiao, L., 2021. Baseflow filter using the recursive digital filter technique.

Raftery, A.E., 1995. Bayesian Model Selection in Social Research. Sociol. Methodol. 25, 111–163.

Ricard, S., Sylvain, J.-D., Anctil, F., 2020. Asynchronous Hydroclimatic Modeling for the Construction of Physically Based Streamflow Projections in a Context of Observation Scarcity. Front. Earth Sci. 8, 1–16. https://doi.org/10.3389/feart.2020.556781

Ricard, S., Sylvain, J.D., Anctil, F., 2019. Exploring an alternative configuration of the hydroclimatic modeling Chain, based on the notion of asynchronous objective functions. Water (Switzerland) 11. https://doi.org/10.3390/w11102012

Richter, B.D., Baumgartner, J. V., Powell, J., Braun, D.P., 1996. A Method for Assessing Hydrologic Alteration within Ecosystems. Conserv. Biol. 10, 1163–1174.

Santos, L., Thirel, G., Perrin, C., 2018. Technical note: Pitfalls in using log-transformed flows within the KGE criterion. Hydrol. Earth Syst. Sci. Discuss. 1–14. https://doi.org/10.5194/hess-2018-298

Schaefli, B., Nicótina, L., Imfeld, C., Da Ronco, P., Bertuzzo, E., Rinaldo, A., 2014. SEHR-ECHO v1.0: a Spatially Explicit Hydrologic Response model for ecohydrologic applications. Geosci. Model Dev. 7, 2733–2746. https://doi.org/10.5194/gmd-7-2733-2014

Schöniger, A., Wöhling, T., Samaniego, L., Nowak, W., 2014. Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. Water Resour. Res. 50, 9484–9513. https://doi.org/10.1002/2014WR016062

Seibert, J., 2003. Reliability of Model Predictions Outside Calibration Conditions. Hydrology Research 34, 477–492. https://doi.org/10.2166/nh.2003.0019

Seibert, J., Vis, M.J.P., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. Hydrol. Earth Syst. Sci. 16, 3315–3325. https://doi.org/10.5194/hess-16-3315-2012

Seibert, J., Vis, M.J.P., 2010. HBV-light HELP.

Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrol. Earth Syst. Sci. 16, 1171–1189. https://doi.org/10.5194/hess-16-1171-2012

Seiller, G., Hajji, I., Anctil, F., 2015. Improving the temporal transposability of lumped hydrological models on twenty diversified U.S. watersheds. J. Hydrol. Reg. Stud. 3, 379–399. https://doi.org/10.1016/j.ejrh.2015.02.012

SMHI, 2005. PTHBV klimatdatabas för hydrologiska beräkningar [PTHBV Climate Database for Hydrological Calculations] (Produktblad). Norrköping, Sweden.

SMHI, 2022. Basic Climate Change Scenario Service [WWW Document]. Future Climate. URL https://www.smhi.se/en/climate/future-climate/basic-climate-change-scenario-service/sverige/ (accessed 2.1.22)

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2014. The deviance information criterion: 12 years on. J. R. Stat. Soc. Ser. B (Statistical Methodol. 76, 485–493. https://doi.org/10.1111/rssb.12062

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 583–616. https://doi.org/10.1111/1467-9868.00353

Sun, W., Trevor, B., 2018. Multiple model combination methods for annual maximum water level prediction during river ice breakup. Hydrol. Process. 32, 421–435. https://doi.org/10.1002/hyp.11429

Tebaldi, C., Knutti, R., 2007. The use of the multi-model ensemble in probabilistic climate projections. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 365, 2053–2075. https://doi.org/10.1098/rsta.2007.2076

Teutschbein, C., Quesada Montano, B., Todorović, A., Grabs, T., 2022. Streamflow droughts in Sweden: Spatiotemporal patterns emerging from six decades of observations. J. Hydrol. Reg. Stud. 42, 101171. https://doi.org/10.1016/J.EJRH.2022.101171

Todorović, A., Grabs, T., Teutschbein, C., 2022. Advancing Traditional Strategies for Testing Hydrological Model Fitness in a Changing Climate. Hydrol. Sci. J. https://doi.org/10.1080/02626667.2022.2104646

Todorović, A., Stanić, M., Vasilić, Ž., Plavšić, J., 2019. The 3DNet-Catch hydrologic model: Development and evaluation. J. Hydrol. 568, 26–45. https://doi.org/10.1016/j.jhydrol.2018.10.040

Tootoonchi, F., Todorović, A., Grabs, T., Teutschbein, C., 2023. Uni- and multivariate bias adjustment of climate model simulations in Nordic catchments: Effects on hydrological signatures relevant for water resources management in a changing climate. Journal of Hydrology 623, 129807. https://doi.org/10.1016/j.jhydrol.2023.129807

Topalović, Ž., Todorović, A., Plavšić, J., 2020. Evaluating the transferability of monthly water balance models under changing climate conditions. Hydrol. Sci. J. 65, 1–23. https://doi.org/10.1080/02626667.2020.1725238

Vaze, J., Post, D. a., Chiew, F.H.S., Perraud, J.-M., Viney, N.R., Teng, J., 2010. Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies. J. Hydrol. 394, 447–457. https://doi.org/10.1016/j.jhydrol.2010.09.018

Vis, M., Knight, R., Pool, S., Wolfe, W., Seibert, J., 2015. Model Calibration Criteria for Estimating Ecological Flow Characteristics. Water 7, 2358–2381. https://doi.org/10.3390/w7052358

Vrugt, J.A., 2015. Multi-criteria Optimization Using the AMALGAM Software Package: Theory, Concepts, and MATLAB Implementation.

Vrugt, J.A., Robinson, B.A., 2007. Improved evolutionary optimization from genetically adaptive multimethod search, in: Proceedings of the National Academy of Sciences of the United States of America. pp. 708–11. https://doi.org/10.1073/pnas.0610471104

Vrugt, J.A., Robinson, B.A., Hyman, J.M., 2009. Self-Adaptive Multimethod Search for Global Optimization in Real-Parameter Spaces. IEEE Trans. Evol. Comput. 13, 243–259.

Vukmirović, V., Plavšić, J. (1997) Flood flow analysis using renewal processes, in: UNESCO-IHP V Tech. Documents in Hydrology No. 11 (Annual FRIEND-AMHY Meeting, Thessaloniki, 1995). pp. 159–169.

Wan, Y., Chen, J., Xu, C.-Y., Xie, P., Qi, W., Li, D., Zhang, S., 2021. Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size. J. Hydrol. 603, 127065. https://doi.org/10.1016/j.jhydrol.2021.127065

Wang, H.-M., Chen, J., Xu, C., Chen, H., Guo, S., Xie, P., Li, X., 2019. Does the weighting of climate simulations result in a better quantification of hydrological impacts? Hydrol. Earth Syst. Sci. 23, 4033–4050. https://doi.org/10.5194/hess-23-4033-2019

Wang, H., Zhang, X., Zou, G., 2009. Frequentist model averaging estimation: a review. J. Syst. Sci. Complex. 22, 732–748. https://doi.org/10.1007/s11424-009-9198-y

Wang, X., Jiang, D., Lang, X., 2021. Future changes in Aridity Index at two and four degrees of global warming above preindustrial levels. Int. J. Climatol. 41, 278–294. https://doi.org/10.1002/joc.6620

Watanabe, S., 2013. A widely applicable bayesian information criterion. J. Mach. Learn. Res. 14, 867–897.

Westerberg, I.K., McMillan, H.K., 2015. Uncertainty in hydrological signatures. Hydrol. Earth Syst. Sci. 19, 3951–3968. https://doi.org/10.5194/hess-19-3951-2015

Xingnan, Z., 1994. A comparative study of the hbv model and development of an automatic calibration scheme. Norrköping, Sweden.

Yang, Y., 2001. Adaptive Regression by Mixing. J. Am. Stat. Assoc. 96, 574–588. https://doi.org/10.1198/016214501753168262

Yarnell, S.M., Stein, E.D., Webb, J.A., Grantham, T., Lusardi, R.A., Zimmerman, J., Peek, R.A., Lane, B.A., Howard, J., Sandoval-Solis, S., 2020. A functional flows approach to selecting ecologically relevant flow metrics for environmental flow applications. River

Res. Appl. 36, 318–324. https://doi.org/10.1002/rra.3373

Ye, M., Neuman, S.P., Meyer, P.D., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour. Res. 40, 1–17. https://doi.org/10.1029/2003WR002557

Zaherpour, J., Gosling, S.N., Mount, N., Schmied, H.M., Veldkamp, T.I.E., Dankers, R., Eisner, S., Gerten, D., Gudmundsson, L., Haddeland, I., Hanasaki, N., Kim, H., Leng, G., Liu, J., Masaki, Y., Oki, T., Pokhrel, Y., Satoh, Y., Schewe, J., Wada, Y., 2018. Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts. Environ. Res. Lett. 13. https://doi.org/10.1088/1748-9326/aac547

Zhang, X., Liang, H., 2011. Focused information criterion and model averaging for generalized additive partial linear models. Ann. Stat. 39, 174–200. https://doi.org/10.1214/10-AOS832

Zhang, Y., Su, F., Hao, Z., Xu, C., Yu, Z., Wang, L., Tong, K., 2015. Impact of projected climate change on the hydrology in the headwaters of the Yellow River basin. Hydrol. Process. 29, 4379–4397. https://doi.org/10.1002/hyp.10497

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.

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

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