Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size

Yongjing Wan [a], Jie Chen [a,b*], Chong-Yu Xu[c], Ping Xie [a], Wenyan Qi [a], Daiyuan Li[d], Shaobo Zhang [a]

[a] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China

[b] Hubei Key Laboratory of Water System Science for Sponge City Construction, Wuhan University, Wuhan 430072, China

[c] Department of Geosciences, University of Oslo, P.O. Box 1047 Blindern, N-0316 Oslo, Norway

[d] Hydrology and Water Resources Department, Nanjing Hydraulic Research Institute, Nanjing 210029, China

* Corresponding author: Jie Chen (jiechen@whu.edu.cn)

Abstract：

The multi-model combination is a technique to improve the performances of hydrological streamflow simulations. An area that has not been investigated much is the performance dependence of combination techniques on the hydrological model calibration strategy and ensemble size. This study aims at investigating the joint effect of the hydrological models, calibration strategies and ensemble sizes on combination abilities for selecting the most appropriate multi-model combination method. The ensemble members were constructed by applying four hydrological models and four objective functions over 383 catchments in China. The ensemble members were combined by using nine commonly used methods, which are Equal Weights (EWA), Akaike Information Criterion (AICA), Bayes Information Criterion (BICA), Bates and Granger (BGA), Granger Ramanathan A, B, and C (GRA, GRB, and GRC), Bayesian Model Averaging (BMA) and Multi-model Super

Ensemble (MMSE). The GRC is found as the best multi-model combination method for hydrological simulations. Adding ensemble members by either multiple hydrological models or calibration strategies could help to improve the simulation abilities. Specifically, the increase of ensemble members can obviously enhance the performance of multi-model combinations when the ensemble size is less than six, while only limited improvement is achieved when the ensemble size is more than nine. The combination of ensemble members with various calibration strategies is hard to compensate for the weakness of hydrological model structures. As well, the application of a single calibration strategy in ensemble members only emphasizes single discharge periods and neglects other important discharge periods. This study found that various models with different objective functions are more robust and efficient. The combination performs better than any individual model in terms of Nash–Sutcliffe efficiency (NSE) for approximately 70% catchments, but the multi-model combination is less efficient in terms of low-flow simulations.

Keywords: multi-model combination method; ensemble modeling; calibration strategy; ensemble size; joint effect

1. Introduction

Hydrological models are essential tools for addressing a wide spectrum of hydrological and water resources problems, including water resources planning, drought and flood control, simulation at ungauged locations, and impact studies for climate or land-use changes (Kotsuki et al., 2014; Kudo et al., 2017; Lane et al., 2019; Wang et al., 2020; Zhang et al., 2014). Over the last decades, a large number of hydrological models have been developed, ranging from lumped conceptual models to physically-based distributed models (Arnold et al., 1998; Chiew et al., 2002; Edijatno et al., 1999; Liang et al., 1994; Xu, 2021; Zhao et al., 1980). The performance of those models varies for diverse catchments characterized by different climate, land use and topography, according to the strengths and weaknesses of the modeling (Mendoza et al., 2016; Pechlivanidis et al., 2011; Vansteenkiste et al., 2014a; Zhang et al., 2020). It is hard to determine a priori which model is most appropriate for a given application over widely differing characteristics of catchments. A single model is not able to consistently outperform the others for all catchment characteristics and heterogeneous climatology (Arsenault et al., 2015; Kumar et al., 2015; Velazquez et al., 2010). Several studies (Arsenault et al., 2015; Velázquez et al., 2011; Zhang et al., 2020) found that multi-model combinations are more robust and efficient than their individual members with the concept of using ensemble to reducing errors with an optimal bias and variance trade-off.

A wide range of methods can be used to generate a multi-model combination solution. The simplest example is the calculation of the arithmetic mean of the input models (commonly referred to as the Equal Weights Averaging (EWA)). More sophisticated techniques employ weighted schemes, with differential weightings applied to each input model reflecting their relative

advantages or limitations. There are some popular techniques used to obtain the optimal set weights for multi-model combinations like multiple linear regression (Arsenault et al., 2015; Granger and Ramanathan, 1984; Kumar et al., 2015), machine learning algorithms (Jeong and Kim, 2009; Shamseldin et al., 1997; Zaherpour et al., 2019), Bayesian model averaging (Neuman, 2003) and Information Criterion Averaging (Akaike, 1974; Schwarz, 1978). The challenge in ensemble modeling is to determine the ensemble size and to identify the best averaging method (Arsenault et al., 2015; Buizza and Palmer, 1998; Kumar et al., 2015).

Many studies (Jeong and Kim, 2009; Shamseldin et al., 1997; Sun and Trevor, 2018; Zaherpour et al., 2019) have attempted to identify the best multi-model combination method for hydrological simulations. Shamseldin et al. (1997) applied three methods (simple arithmetic mean, constrained ordinary least-squares weighting, and neural network) to combine four hydrological models for 11 catchments and found the constrained ordinary least-squares weighting and neural network are more robust and efficient than the simple arithmetic mean in terms of the Nash–Sutcliffe efficiency (NSE). Broderick et al. (2016) analyzed the performance of four ensemble averaging techniques using four hydrological models for 37 Irish catchments. They concluded that GRA is the best ensemble averaging technique, and the averaging methods performed better for the NSE as opposed to bias metrics. In addition, some studies investigated the effect of ensemble size on the performance of multi-model ensemble simulations. For example, Arsenault et al. (2015) compared nine multi-model averaging approaches using 12 hydrographs (4 models $\times$ 3 metrics) over 429 catchments. They found that GRC performs better than other averaging methods and no catchment requiring more than seven ensemble members to maximize the NSE with this method. Kumar et al. (2015) compared ten different multi-model ensemble methods using eight hydrological models to select

the best multi-model ensemble method for the discharge estimation over a catchment of the Mahanadi river basin in India. They showed that the constrained multiple linear regression is the most suitable multi-model ensemble method in terms of NSE, root mean square error (RMSE) and Pearson's correlation coefficient (R), and five ensembles show the best performance for the study area.

The method of multi-model combination is usually used to extract as much information as possible from a group of existing models, which may produce a better overall simulation, as each simulation of the group provides specific information. In addition to the selection of the hydrological models, the process of parameter identification is also a crucial step in streamflow modeling. The calibration strategies (which here mean choice of objective functions) reflect the goodness of fitting between hydrological model simulations and observations, which can substantially influence the model parameters and the streamflow projections (Krysanova et al., 2018; Lane et al., 2019; Mizukami et al., 2019; Seiller et al., 2017). The hydrological models with various structures and calibration strategies have certain capacities to predict the streamflow. In general, the hydrological simulations achieve maximum accuracy in terms of specific hydrological properties using a particular metric, but that might limit the modeling skill in other aspects (Arsenault et al., 2015; Mizukami et al., 2019; Seiller et al., 2017). For example, the most widely used calibration strategies, such as Nash–Sutcliffe efficiency (NSE) and Kling–Gupta efficiency (KGE), emphasize high flow events and their timing (Gupta et al., 2009; Mizukami et al., 2019; Nash and Sutcliffe, 1970). Those metrics calculated on the natural logarithm of the flow values put more weight on low flows (Pushpalatha et al., 2012; Seiller et al., 2017). The use of multi-model combination scheme is expected to benefit from the variation of the parameter sets derived from objective functions targeted

at different hydrological processes to produce a better overall simulation. Although the influence of calibration strategies exists in hydrological modeling, the impacts of hydrological model calibration strategies on the performance of multi-model combinations and their joint effects with ensemble sizes are not clear. Moreover, there is no consensus in the hydrological community in terms of the selection of particular multi-model ensemble sizes to ensure good model performances.

This study aims to investigate the joint effect of ensemble sizes and hydrological model calibration strategies on combination abilities for selecting the most appropriate multi-model combination method. Specifically, nine commonly used multi-model combination techniques are compared over 383 catchments in China using ensemble members derived from 4 hydrological models calibrated with 4 objective functions. The rest of the paper is organized as follows. Section 2 presents a brief introduction of the study data and the methodology, including the hydrological models, the calibration strategies, the multi-model combination techniques, and the details of the evaluation technique used in this study. The results are presented in Section 3, followed by the discussion and conclusion in Section 4.

2. Data and Methodology

2.1 Study region and data

This study used a gridded meteorological dataset (0.5° × 0.5°) over China for the period of 1961–2016, which contains four climate variables, including daily precipitation, daily maximum, minimum and mean air temperatures to represent observed data. This dataset was generated from 2,472 in-situ gauge stations across China by thin-plate spline interpolation method and GTOPO30 (Global 30 Arc-Second Elevation) data sampling and is considered as the latest gridded

meteorological data with the highest spatial resolution in China. This dataset has been commonly used in many hydro-climatological studies in China (Gu et al., 2020; Li et al., 2019; Yin et al., 2020), and downloaded from the China Meteorological Data Sharing Service System (http://www.cma.gov.cn).

The daily streamflow series over 383 catchments in China were used (Figure 1). These catchments with a wide range of climatic conditions and hydrological regimes span over all the nine major river basins in China. Based on climate type and physical geography, this study region was divided into four major climate regions: continental climate zone of Northwest (NW), the highland climate zone of Southwest (Tibetan Plateau, SW), the temperate monsoon region of Northeast (NE), and the tropical and subtropical monsoon region of Southeast (SE) (Ding, 2013; Wu et al., 2016). NE is the driest region according to the average aridity index value (Figure 2). The size of the catchments ranges from 612 km^2 to 995,343 km^2. The streamflow dataset covers the 1961–2016 period with a maximum length of 52 years and a minimum length of 22 years. The average annual precipitation of the catchment varies greatly with clear gradients depending on the region. The mean annual precipitation is more than 1400 mm in the southern region, while it is less than 600 mm in the northern region.

<Figure 1>

<Figure 2>

2.2 Hydrological models

A wide range of hydrological models is used for different application purposes. Some studies compared the performance of lumped and distributed models for outlet streamflow simulation and

found that two types of models may lead to similar accuracy (Kumar et al., 2015; Lobligeois et al., 2014; Vansteenkiste et al., 2014b), even for quite large catchments (Merz et al., 2009). Considering the scope of this study, the lumped model (the most common model type used by hydrologists for water resources assessment, flood forecasting, and impact of climate change studies) was chosen, and the distributed model with expensive computations was excluded. Four lumped models with different complexity were used, i.e., modèle du Génie Rural à 4 paramètres Journalier (GR4J) (Edijatno et al., 1999; Perrin et al., 2003), hydrological model of école de technologie supérieure (HMETS) (Martel et al., 2017), simple HYDROLOG (SIMHYD) (Chiew et al., 2002), and Xinanjiang (XAJ) (Zhao, 1992; Zhao et al., 1980). Those models have been widely used in streamflow simulation and have been shown to be relatively efficient (Arsenault et al., 2015; Broderick et al., 2016; Jones et al., 2006; Liang et al., 2013; Mathevet et al., 2020). Table 1 briefly summarized the basic information for these hydrological models.

The four models have different numbers of parameters and are different in model structures and underlying mechanisms. For example, the physical process is described in more detailed and complex mechanisms in HMETS, SIMHYD and XAJ than in the most parsimonious structure GR4J with only four free parameters. The main feature of the runoff generation of HMETS and XAJ is using the saturation excess flow mechanism based on the soil moisture content of the aeration zone reaching its field capacity. While SIMHYD considers both infiltration excess runoff and saturation excess runoff in streamflow production calculated by an interception store, a nonlinear soil moisture store. For the simulation of evaporation, XAJ uses a three-layer evaporation model, while HMETS and SIMHYD use a one-layer model. Additionally, GR4J and HMETS consider the incorporation of groundwater exchange by surface water–groundwater interaction functions, but XAJ and

SIMHYD do not have this consideration.

Since GR4J, XAJ and SIMHYD do not simulate snow accumulation or melt processes, a snow module (CEMANEIGE) with 2 free parameters (Valéry et al., 2014) was combined with the original model to make it applicable in seasonally snow-covered catchments in northern China. The basic inputs of these four models are catchment-averaged precipitation and temperature/potential evapotranspiration over the entire basin for the computation of discharge. The potential evapotranspiration was estimated using a temperature-based method proposed by Oudin et al. (2006a).

<Table 1>

2.3 Calibration and evaluation metrics

This study used four objective functions to calibrate the four hydrological models over 383 catchments. The four calibration strategies are the widely-used Nash–Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970), the NSE computed on natural logarithm and square root of the flow values (NSE(ln) and NSE(sqrt)), and percent bias (PBias). NSE, NSE(ln) and NSE(sqrt) all range from negative infinity to 1, with a value of 1 indicating perfect fitting. The PBias' value being closer to zero indicates the better simulating performances.

Different calibration strategies were included in the combination since they emphasize different aspects of hydrological streamflow properties. The original NSE without discharge transformation puts great emphasis on high flows (Li et al., 2019; Mizukami et al., 2019). The natural logarithm of discharge transformation (NSE(ln)) is to optimize the performance for low flow segments (Seiller et al., 2017). The analysis of NSE(sqrt) well balances simulated streamflow

without too much emphasis on low or high flow (Oudin et al., 2006b), and the PBias emphasizes the total water balance. Regardless of the objective function, the hydrological models optimized the parameter using the Shuffled Complex Evolution-University of Arizona (SCE-UA) algorithm (Duan et al., 1992). A cross-validation method that divided the complete record of each catchment (Arsenault et al., 2015; Yang et al., 2020) into odd and even years for model calibration and validation was used, to reduce the influence from the non-stationarity of hydro-climatological conditions. In this process, the first year in the calibration period was used for model warm-up.

In addition to using NSE, NSE(ln), NSE(sqrt) and PBias as evaluation metrics to represent the overall performance of simulations, the high and low flows were also analyzed based on the discharge segments of flow duration curves (FDC) to identify the influence of ensemble method on the performance of various flow components. Following previous studies (Laaha and Blöschl, 2006; Pfannerstill et al., 2014; Yilmaz et al., 2008), the flow exceedance probability of 70% was used to represent the low flow, and the mid-flow segment was shifted from 20% to 70%. The very high-flow range was defined between 0% and 5%, and the high-flow range between 5% and 20%. The performance of the model simulations within these FDC segments was analyzed using PBias (PBiasFSV), noted by PBiasFSV-5, PBiasFSV-20, PBiasFSV-mid, and PBiasFSV-low. The basic information of those metrics which were used for evaluating the performance of the different hydrograph phases was shown in Table 2.

<Table 2>

2.4 Multi-model averaging methods

There are several methods available in the literature for developing the multi-model

211 combination. Here, we compared the performance of nine commonly used deterministic ensemble
212 techniques for creating multi-model ensembles. The selected methods include Equal Weights
213 Averaging (EWA), Akaike Information Criterion Averaging (AICA), Bayes Information Criterion
214 Averaging (BICA), Bates and Granger Averaging (BGA), Granger Ramanathan A Averaging (GRA),
215 Granger Ramanathan B Averaging (GRB), Granger Ramanathan C Averaging (GRC), Bayesian
216 Model Averaging (BMA) and Multi-model Super Ensemble (MMSE). The general model for
217 averaging methods can be expressed as：

218 $$Q_{ens} = \sum_{i=1}^{n} W_i \cdot Q_{simi}$$

219 where Q_{ens} is the ensemble simulation, Q_{simi} is the individual simulation, and W_i is the weight
220 of the ensemble member.

221 A description of the selected averaging methods was given in the Appendix. The basic
222 characteristics of the nine multi-model averaging techniques were summarized in Table 3.

223 <Table 3>

224 # 3. Results

225 ## 3.1 Performance of the individual members

226 The performances of the 16 ensemble members over 383 catchments in the calibration and
227 validation were analyzed using the evaluation of NSE, NSE(ln), NSE(sqrt), and PBias representing
228 the average simulated abilities and using the PBias of four different FDC segments (PBiasFSV-5,
229 PBiasFSV-20, PBiasFSV-mid and PBiasFSV-low) representing the high and low flows, as presented
230 in Figure 3. The result shows that the evaluation metrics in the validation period are consistent with

those in the calibration period, which indicates the hydrological simulations are robust and transferable.

The ensemble members with calibration strategies closely related to the evaluation metrics work best, as expected. Models calibrated with NSE(sqrt) maintain good performances in terms of NSE and NSE(ln) evaluation with median values around 0.7 and 0.8. All ensemble members show the absolute value of PBias being less than 10%, but the models calibrated with PBias yield inferior performance in terms of NSE-based evaluation metrics. According to the FDC segment evaluation values (Figures 3e-3h), the models calibrated with NSE show good performance for high flow segments that obtain the best score for 38% and 42% of catchments in terms of the PBiasFSV-5 and PBiasFSV-20, respectively (Table 4). The models calibrated with NSE(ln) put high weight on low flows and obtain the best score at a frequency of 42% catchments. Models calibrated with NSE(sqrt) emphasize middle flows, which obtain the best score at a frequency of 38% catchments. Those results confirmed the specialization of the objective functions for specific parts of the hydrographs.

This figure also demonstrates that hydrologic models perform differently with respect to different calibration strategies and evaluation metrics. For example, calibrated with NSE, both HMETS and GR4J with median NSE(ln) value of 0.6 perform better than SIMHYD and XAJ with median NSE(ln) value of 0.3 and 0.5. In addition, HMETS and GR4J tend to overestimate the midle and low flow segments, while the other two models (i.e., SIMHYD and XAJ) tend to underestimate those flows. Comparing the hydrological models, XAJ generally performs best for the evaluation metrics, followed by HMETS, GR4J and SIMHYD.

When looking at four sub-regions based on the median value of evaluation metrics (Figure 4),

different models show similar spatial patterns in terms of NSE-based metrics with better performances in the wetter southern catchments (SW and SE) than in drier northern catchments (NW and NE). However, the performance of the four hydrological models is not the same. GR4J and SIMHYD perform better than XAJ and HMETS in terms of NSE-based metrics over SW when using NSE(ln) as the objective function.

<Figure 3>

<Figure 4>

<Table 4>

3.2 Performance of the multi-model averaging methods

The nine multi-model combination methods were applied to calculate the optimal weights based on the observed and the 16 simulated streamflow series. The performance of those methods over 383 catchments in the validation period is presented in Figure 5. The result shows that the differences of the ensemble simulations are evident not just in how well the averaging methods but also in the evaluation metrics used.

For NSE metric, almost all multi-model combination methods show similar performances with median values around 0.8 except for EWA, BGA, and BMA with median values less than 0.75. For NSE(ln) metric, EWA, BGA, and BMA perform better than others with median values around 0.8, followed by GRA, GRB and GRC with median values around 0.75. Almost all averaging methods show similar performances in terms of NSE(sqrt) metric with median values exceeding 0.81. For PBias and PBiasFSV-5, GRA, GRB, GRC and MMSE show similar performances and outperform

others, followed by AICA and BICA. For PBiasFSV-20, PBiasFSV-mid and PBiasFSV-low, GRC performs the best, followed by GRA, GRB and MMSE. In general, all simulations tend to underestimate the high flow but overestimating the mid and low flows.

Among various averaging methods, GRA, GRB, GRC, and MMSE show similar performance, since they derived from the same optimal weighting group. The AICA and BICA perform poorly in terms of NSE(ln) and PBiasFSV-low metrics compared with other averaging methods, as they put almost all weights on the individual member with minimum RMSE. It can also be seen that the BMA, EWA, and BGA without bias-correction show worse performances in terms of PBias-based measures, especially for PBias, PBiasFSV-5 and PBiasFSV-low.

Overall, the GRA, GRB, GRC, and MMSE consistently outperformed other combination methods in terms of the eight evaluation metrics, and GRC provided the best performances, especially in terms of PBias and PBiasFSV. Therefore, GRC is considered as the best multi-model averaging method for hydrological simulations in this study.

<Figure 5>

3.3 Impact of the hydrological model, calibration strategy, and ensemble size on multi-model combinations

This section investigated the joint influence of the hydrological models, calibration strategies and ensemble size on the performance of GRC. The selection of ensemble members is based on multiple hydrological models calibrated with a single objective function, a single hydrological model calibrated with multiple objective functions, and multiple hydrological models calibrated with multiple objective functions. Figure 6 shows the median performance value over 383

catchments in the validation period to represent the overall ability of those combinations.

The performance of multiple hydrological models calibrated with a single objective function varies depending on the selection of objective function and is improved along with the increase of hydrological model numbers (areas numbered 2 in Figure 6). Areas numbered 3 in Figure 6 show the performances of a hydrological model with multiple objective functions. Though using multiple objective functions could improve the combination abilities, the performances still depend on the selection of hydrological models, as each hydrological model has its own advantages and limitations. For example, the combinations based on SIMHYD generally perform worse than others in terms of NSE and NSE(sqrt) but perform better in terms of NSE(ln), PBias and PBiasFSV-low. In addition, a single model is not able to consistently outperform others for all catchments with various hydro-climatic regimes. The combinations of multiple hydrological models calibrated with various objective functions taking advantage of all ensemble members generally perform more robust and efficient than combinations only using a single calibration or a single hydrological model in terms of all NSE-based metrics (areas numbered 4 in Figure 6).

Overall, either using multiple hydrological models or multiple calibration strategies could improve the multi-model combination performances in terms of NSE-based metrics but is less efficient in PBias-based evaluation metrics. The above results also indicate that the influence of increasing hydrological models on combination is larger than increasing model calibration strategies, and the effect of the individual simulation is decreasing along with the increase of ensemble members. The qualitative comparison of different combinations and individual simulations would be analyzed later (Section 3.4).

<Figure 6>

The combinations with ensemble sizes ranging from 2 to 16 ensemble members are generated by re-sampling the ensemble members 100 times from all combination members to investigate the effect of ensemble size on simulating abilities. Figure 7 shows the relationship between the ensemble size and the combination performances. The 0.05 and 0.25 quantiles of NSE, NSE(ln) and NSE(sqrt) values correspond to the poor efficiency modeling, while 0.75 and 0.95 quantiles correspond to simulations with good performance. The median performance is represented by quantile 0.50 (red lines). The absolute values of PBias and four PBiasFSV being infinite to zero indicating better performances. This result indicates that the performance is improved along with the increase in ensemble sizes, and the effect of ensemble size on the low quantile values is larger than that on the high quantile. In general, the performance of the multi-model combinations is sensitive to the ensemble size and the selection of ensemble members when the member is less than six. The influence of enlarging ensemble numbers on simulation performances could be ignored when the ensemble member is more than nine. In other words, when the ensemble member is more than nine, neither enriching the hydrological model nor increasing calibration strategies show limited improvement in simulating abilities, as larger ensemble sizes mean each individual member with smaller weights and limited influence on the combination.

<Figure 7>

In order to investigate whether all ensemble members contribute to the performance of the combination, the frequency of the individual member being selected in the best-performed combinations of all catchments is calculated and plotted in Figure 8. Here, the ensemble sizes of the

335 combinations are nine since the results from Figure 7 concluded that the effect of increasing
336 ensemble numbers more than nine could be ignored. Figure 8 shows that all individual members
337 have similar chancies of being selected in the best-performed combinations in terms of different
338 evaluation metrics, indicating all ensemble members contribute to enhancing the combination
339 simulation.

340 <Figure 8>

341 ## 3.4 Comparison of the multi-model combination and ensemble members

342 The performances of GRC and the ensemble members were compared using the cumulative
343 distributions of NSE, NSE(ln), NSE(sqrt), absolute values of PBias and PBiasFSV over the 383
344 catchments in the validation period (Figure 9). Here, five combinations are represented, including
345 four hydrological models calibrated with a specified objective function (i.e., H4-NSE, H4-NSE(ln),
346 H4-NSE(sqrt) and H4-PBias) and four hydrological models calibrated with four calibration
347 strategies (i.e., H4C4).

348 It is apparent that the multi-model combinations, either using multiple hydrological models or
349 using multiple calibration strategies, outperform the individual member in terms of those
350 performance metrics, except for NSE(ln) and PBiasFSV-low. The combinations of four hydrological
351 models calibrated with four calibration strategies perform the best and show 85% catchments have
352 NSE values exceeding 0.6, and 75% catchments have NSE values exceeding 0.7 (Figure 9a). In
353 addition, 91% catchments have NSE(sqrt) values exceeding 0.6, and 73% catchments have NSE(sqrt)
354 values exceeding 0.7 (Figure 9b). The absolute value of PBias is less than 10 for more than 95%
355 catchments, and that is less than 5 for 78% catchments. In addition, the absolute values of PBiasFSV-

5, PBiasFSV-20, and PBiasFSV-mid are less than 10 for 55%, 80% and 80% of catchments, respectively. However, the combinations are less efficient in terms of NSE(ln) and PBiasFSV-low evaluation than the individual members using NSE(ln) calibration, since the averaging methods based on minimum RMSE are more sensitive to flood peaks than low-flow period.

<Figure 9>

Table 5 shows the rate at which the averaging simulation surpasses the best individual member in terms of four average performance metrics and four FDC segment metrics over the 383 catchments in the validation period. Here, the ‘‘best individual member” value is selected from the 16 members independently for each catchment. The results show that four hydrological models calibrated with four calibrations (H4C4) perform more robustly than other combinations in terms of various performance metrics. The combination of four hydrological models calibrated with a specified objective only performs efficiently for the specific parts of the hydrographs. The combinations perform better than the best individual member for about 70% catchments in terms of NSE and for about 55% catchments in terms of NSE(sqrt). On the contrary, combinations show less efficiency for performance values in terms of NSE(ln) and PBias.

Figures 10(a) and 10(b) show the geographic distribution of the NSE value in the validation period from the best individual member and combinations. The spatial patterns are consistent with better performances in the wetter southern catchments than in the drier northern catchments. For catchments with annual precipitation larger than 600 mm, 82% catchments obtain NSE value exceeding 0.7, and the combination outperforms the best individual in 82% cases. However, for catchments with annual precipitation less than 600 mm, only 51% catchments obtain NSE value

exceeding 0.7, and the combinations outperform the best individual only in 45% cases (Figures 10c and 10d). Since hydrological simulations are big challenges for the arid catchments, especially when using lumped models, which cannot represent the rainfall and loss variability that tends to be higher. The multi-model combinations have less benefit when the ensemble members cannot accurately simulate the streamflow.

<Table 5>

<Figure 10>

4. Discussion and conclusion

The multi-model combination is a technique widely used to improve the performance of hydrological streamflow simulations, as it extracts potentially useful information from a group of existing models. The commonly used multi-model combination usually takes information from different hydrological models while neglects the additional information from calibration strategies. This study investigated the joint effect of hydrological models, calibration strategies and ensemble sizes on combination abilities for selecting the appropriate multi-model combination method. This study compared different multi-model ensemble methods by combining the simulated discharge with four hydrological models and four different calibration strategies.

Generally, the hydrological model performances were similar in terms of geographic distribution, with better performances in wetter catchments than drier catchments. Nevertheless, there were some obvious differences between those models. GR4J and HMETS with surface water–groundwater interaction functions performed better than XAJ and SIMHYD in terms of NSE(ln) value. This is consistent with Pushpalatha et al. (2011) and Fleckenstein et al. (2006), who found

that the lumped model with the incorporation of groundwater exchange functions improves the predictions of low flows.

One of the other goals was to compare the nine multi-model averaging methods. Overall, GRA, GRB, GRC, and MMSE with similar methods for optimal weight showed consistent performances and performed better than other combination methods. GRC with bias-correction showed better performances in terms of PBias and PBiasFSV metrics than other combination methods and was considered as the best multi-model averaging method for hydrological simulations in this study. As AICA and BICA put all weights on the individual model with a minim root-mean-square error, they performed similarly to the best-performing individual model. BGA, without the consideration of the biases from ensemble members, only performed slightly better than EWA, which is consistent with the previous study (Arsenault et al., 2015). This study also found that BMA, which was widely used in previous ensemble studies (Li et al., 2019; Zhang et al., 2020), performs similarly to BGA in terms of PBias-baed evaluation metrics. BMA is recommended to apply on ensemble simulations, whereas it is sometimes difficult to effectively remove bias from the predictions of complex streamflow simulation (Madadgar and Moradkhani, 2014), which is related to the poor performance in inters of PBias evaluations.

In addition, this study found that adding ensemble members by either increasing hydrological models or increasing calibration strategies could improve the simulation abilities, but the combinations of taking advantage of different hydrological models and objective functions were more robust and efficient in terms of different hydrological properties and hydro-climatic regimes. In addition, this study found that the increase of ensemble members could obviously improve the

multi-model combination performance when the ensemble size is less than six, but has limited effects when the ensemble size is more than nine. This is consistent with Kumar et al. (2015) and Arsenault et al. (2015), who found that four and seven ensemble members are efficient for multi-model combinations.

Comparing the combinations and individual simulations, the combinations of different hydrological models with various objective functions performed better than any individual model for around 70% of 383 catchments in terms of NSE scores. In contrast, the combinations were less efficient than the best individual members in terms of low-flow simulations. This study also found that the multi-model combinations perform better for wetter catchments than for drier catchments. The frequencies of ensemble simulation outperforming the best individual simulation in terms of NSE were 82% catchments in wetter regions with precipitation more than 600 mm/year and 45% catchments in drier regions with precipitation less than 600 mm/year.

Acknowledgments

This work was partially supported by the National Key Research and Development Program of China (No. 2017YFA0603704), the National Natural Science Foundation of China (Grant No. 52079093), the Hubei Provincial Natural Science Foundation of China (Grant No. 2020CFA100), and the Overseas Expertise Introduction Project for Discipline Innovation (111 Project) funded by Ministry of Education and State Administration of Foreign Experts Affairs P.R. China (Grant No. B18037). The authors wish to thank the China Meteorological Data Sharing Service System for providing gauged precipitation for China.

Appendix：Multi-model ensemble methods

A description of the multi-model averaging methods is presented here.

1. Equal Weights Averaging (EWA)

In the EWA (Shamseldin et al., 1997), the equal weight is simply assigned to each of the ensemble members. This is expressed mathematically as:

$$\mathrm{w} = \frac{1}{N} \tag{A1}$$

where N is the number of ensemble models.

2. Akaike and Bayes information criteria averaging (AICA and BICA)

AICA (Akaike, 1974) and BICA (Schwarz, 1978) methods combine ensemble members based on both performance and model complexity. Weight represents a trade-off between reducing the simulated error while tending toward less complex, which is calculated as:

$$\mathrm{w} = \frac{\exp\left(-\frac{1}{2}I\right)}{\sum_{i=1}^{N}\exp\left(-\frac{1}{2}\mathrm{I}_i\right)} \tag{A2}$$

where I is the information criterion estimated based on the mean of the logarithm of the ensemble member variances and the number of calibration parameters, and is calculated as:

$$\mathrm{I} = -2\log(L) + q(p) \tag{A3}$$

where L and $q(p)$ are the maximum likelihood of ensemble member and the penalty term, respectively.

The difference between the AICA and BICA methods lies in the penalty term calculation. The penalty terms of AICA and BICA are estimated by equations (A4) and (A5):

$$q = 2p \tag{A4}$$

$$q = p\log(k) \tag{A5}$$

where p donates the number of calibrated parameters in the members, and k donates the sample size (here is the number of time steps).

3. Bates and Granger Averaging (BGA)

The BGA (Bates and Granger, 1969) method aims to produce a combined ensemble by minimizing the Root Mean Square Error (RMSE) between the observations and simulations. The model weighting vector is estimated according to:

$$\mathrm{W} = \frac{{}^{1}/_{RMSE^2}}{\sum_{i=1}^{N} {}^{1}/_{RMSE_i{}^2}} \tag{A6}$$

4. Granger Ramanathan A, B and C (GRA, GRB and GRC)

The GRA (Granger and Ramanathan, 1984) approach minimizes the RMSE setting weights based on the ordinary least squares (OLS) algorithm. The weights are estimated by:

$$W = \left(Q_{sim}{}^{T} Q_{sim}\right)^{-1} Q_{sim}{}^{T} Q_{obs} \tag{A7}$$

where Q_{obs} is the observation. The GRB variant is similar to the GRA method, but the OLS algorithm is constrained such that the sum of the weights to unity. The GRC variant is unconstrained, but the averaged streamflow is bias-corrected through the use of a constant term.

5. Bayesian Model Averaging (BMA)

BMA (Neuman, 2003) determines the weights of each member through the use of the ensemble members' probability distribution functions (PDFs). The combined distribution is bias-corrected, and the difference between the distributions is minimized. According to BMA, the posterior

probability of the predictand (y) is described as:

$$p(y \mid Q_{\text{obs}}) = \sum_{i=1}^{N} P(Q_{simi} \mid Q_{\text{obs}}) p(y \mid Q_{simi}, Q_{\text{obs}}) \quad \text{(A8)}$$

where $p(y \mid Q_{simi}, Q_{\text{obs}})$ is the posterior predictive distribution of y on the condition of the given sample Q_{obs} and each individual model Q_{simi}; $P(Q_{simi} \mid Q_{\text{obs}})$ is the optimal model on the condition of the given sample Q_{obs} denoted the weight (w_i) of each ensemble member. The mean and variance of y are given:

$$E[y \mid Q_{\text{obs}}] = \sum_{i=1}^{N} P(Q_{simi} \mid Q_{\text{obs}}) \int_{-\infty}^{+\infty} y p(y \mid Q_{simi}, Q_{\text{obs}}) dy = \sum_{i=1}^{N} w_i \eta_i \quad \text{(A9)}$$

$$\text{Var}[y \mid Q_{\text{obs}}] = \sum_{i=1}^{N} w_i \left(\eta_i - \sum_{i=1}^{N} w_i \eta_i\right)^2 + \sum_{i=1}^{N} w_i \sigma_i^2 \quad \text{(A10)}$$

$$w_i = P(Q_{simi} \mid Q_{\text{obs}}) \quad \text{(A11)}$$

where w_i is the weight of the ensemble member and sum to unite; η and σ are the expectation and variance of y, respectively, on the condition of the given sample Q_{obs}.

6. Multi-model Super Ensemble (MMSE)

MMSE (Krishnamurti et al., 2000) uses the logic of bias reduction along with variance reduction through using the mean of observational value and the combination of ensemble member, respectively. According to this method, the ensemble discharge is estimated as:

$$Q_{MMSE,j} = \bar{Q}_{obs} + \sum_{i=1}^{N} w_i \left[Q_{sim,i,j} - \bar{Q}_{sim,i}\right] \quad \text{(A12)}$$

where w_i is the weight of ith model, which is estimated by the unconstrained least square technique (Eq. (A1)).

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6): 716-723. https://doi.org/10.1109/TAC.1974.1100705

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. LARGE AREA HYDROLOGIC MODELING AND ASSESSMENT PART I: MODEL DEVELOPMENT1. JAWRA Journal of the American Water Resources Association, 34(1): 73-89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. Journal of Hydrology, 529: 754-767. http://dx.doi.org/10.1016/j.jhydrol.2015.09.001

Bates, J.M., Granger, C.W.J., 1969. The Combination of Forecasts. Journal of the Operational Research Society, 20(4): 451-468. https://doi.org/10.1057/jors.1969.103

Broderick, C., Matthews, T.K.R., Wilby, R.L., Bastola, S., Murphy, C., 2016. Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. Water Resources Research, 52(10): 8343-8373. https://doi.org/10.1002/2016WR018850.

Buizza, R., Palmer, T.N., 1998. Impact of Ensemble Size on Ensemble Prediction. Monthly Weather Review, 126(9): 2503-2518. https://doi.org/10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2

Chiew, F.H.S., Peel, M.C., Western, A.W., Singh, V.P., Frevert, D., 2002. Application and testing of the simple rainfall-runoff model SIMHYD. Water Resources Publications, Colorado, USA.

Ding, Y., 2013. China Climate. Science Press, Beijing, China.

Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resources Research, 28(4): 1015-1031. https://doi.org/10.1029/91WR02985

Edijatno, De Oliveira Nascimento, N., Yang, X., Makhlouf, Z., Michel, C., 1999. GR3J: a daily watershed model with three free parameters. Hydrological Sciences Journal, 44(2): 263-277. https://doi.org/10.1080/02626669909492221

Fleckenstein, J.H., Niswonger, R.G., Fogg, G.E., 2006. River-aquifer interactions, geologic heterogeneity, and low-flow management. Ground Water, 44(6): 837-852. https://doi.org/10.1111/j.1745-6584.2006.00190.x

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting, 3(2): 197-204. https://doi.org/10.1002/for.3980030207

Gu, L. et al., 2020. On future flood magnitudes and estimation uncertainty across 151 catchments in mainland China. International Journal of Climatology, n/a(n/a). https://doi.org/10.1002/joc.6725

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377(1): 80-91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Jeong, D.I., Kim, Y.-O., 2009. Combining single-value streamflow forecasts – A review and guidelines for selecting techniques. Journal of Hydrology, 377(3): 284-299. https://doi.org/10.1016/j.jhydrol.2009.08.028

Jones, R.N., Chiew, F., Boughton, W.C., Lu, Z.J.A.i.W.R., 2006. Estimating the sensitivity of mean annual runoff to climate change using selected hydrological models. Advances in Water Resources, 29(10): 1419-1429. https://doi.org/10.1016/j.advwatres.2005.11.001

Kotsuki, S., Tanaka, K., Watanabe, S., 2014. Projected hydrological changes and their consistency under future climate in the Chao Phraya River Basin using multi-model and multi-scenario of CMIP5

dataset. Hydrological Research Letters, 8(1): 27-32. https://doi.org/10.3178/hrl.8.27

Krishnamurti, T.N. et al., 2000. Multimodel Ensemble Forecasts for Weather and Seasonal Climate. Journal of Climate, 13(23): 4196-4216. https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2

Krysanova, V. et al., 2018. How the performance of hydrological models relates to credibility of projections under climate change. Hydrological Sciences Journal, 63(5): 696-720. https://doi.org/10.1080/02626667.2018.1446214

Kudo, R., Yoshida, T., Masumoto, T., 2017. Nationwide assessment of the impact of climate change on agricultural water resources in Japan using multiple emission scenarios in CMIP5. Hydrological Research Letters, 11(1): 31-36. https://doi.org/10.3178/hrl.11.31

Kumar, A., Singh, R., Jena, P.P., Chatterjee, C., Mishra, A., 2015. Identification of the best multi-model combination for simulating river discharge. Journal of Hydrology, 525(525): 313-325. http://dx.doi.org/10.1016/j.jhydrol.2015.03.060

Laaha, G., Blöschl, G., 2006. Seasonality indices for regionalizing low flows. Hydrological Processes, 20(18): 3851-3878. https://doi.org/10.1002/hyp.6161

Lane, R. et al., 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(10): 4011-4032. https://doi.org/10.5194/hess-23-4011-2019

Li, X., Chen, J., Xu, C., Li, L., Chen, H., 2019. Performance of Post-Processed Methods in Hydrological Predictions Evaluated by Deterministic and Probabilistic Criteria. Water Resources Management, 33(9): 3289-3302. https://doi.org/10.1007/s11269-019-02302-y

Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. Journal of Geophysical Research: Atmospheres, 99(D7): 14415-14428. https://doi.org/10.1029/94JD00483

Liang, Z., Wang, D., Guo, Y., Zhang, Y., Dai, R., 2013. Application of Bayesian Model Averaging Approach to Multimodel Ensemble Hydrologic Forecasting. Journal of Hydrologic Engineering, 18(11): 1426-1436. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000493

Lobligeois, F., Andréassian, V., Perrin, C., Tabary, P., Loumagne, C., 2014. When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events. Hydrol. Earth Syst. Sci., 18(2): 575-594. https://hess.copernicus.org/articles/18/575/2014/

Madadgar, S., Moradkhani, H., 2014. Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. Water Resources Research, 50(12): 9586-9603. https://doi.org/10.1002/2014WR015965

Martel, J.-L., Demeester, K., Brissette, F.P., Arsenaul, R., Poulin, A., 2017. HMET: a simple and efficient hydrology model for teaching hydrological modelling, flow forecasting and climate change impacts. The International journal of engineering education, 33(4): 1307–1316. https://dialnet.unirioja.es/servlet/articulo?codigo=6897050

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. Journal of Hydrology, 585: 124698. https://doi.org/10.1016/j.jhydrol.2020.124698

Mendoza, P.A. et al., 2016. How do hydrologic modeling decisions affect the portrayal of climate change impacts? Hydrological Processes, 30(7): 1071-1095. https://doi.org/10.1002/hyp.10684

Merz, R., Parajka, J., Blöschl, G., 2009. Scale effects in conceptual hydrological modeling. Water Resources Research, 45(9): W09405. https://doi.org/10.1029/2009WR007872

Mizukami, N. et al., 2019. On the choice of calibration metrics for “high-flow” estimation using hydrologic models. Hydrology and Earth System Sciences, 23(6): 2601-2614. https://doi.org/10.5194/hess-23-2601-2019

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology, 10(3): 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stochastic Environmental Research and Risk Assessment, 17(5): 291-305. https://doi.org/10.1007/s00477-003-0151-7

Oudin, L., Andreassian, V., Mathevet, T., Perrin, C., Michel, C., 2006a. Dynamic averaging of rainfall - runoff model simulations from complementary model parameterizations. Water Resources Research, 42(7): W07410. https://doi.org/10.1029/2005WR004636

Oudin, L., Perrin, C., Mathevet, T., Andréassian, V., Michel, C., 2006b. Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. Journal of Hydrology, 320(1): 62-83. https://doi.org/10.1016/j.jhydrol.2005.07.016

Pechlivanidis, I.G., Jackson, B., Mcintyre, N., Wheater, H.S., 2011. 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. Global Nest Journal, 13(3): 193-214.

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology, 279(1): 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. Journal of Hydrology, 510: 447-458. http://dx.doi.org/10.1016/j.jhydrol.2013.12.044

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. Journal of Hydrology, 411(1): 66-76. https://doi.org/10.1016/j.jhydrol.2011.09.034

Pushpalatha, R., Perrin, C., Moine, N.L., Andréassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. Journal of Hydrology, 420-421: 171-182. https://doi.org/10.1016/j.jhydrol.2011.11.055

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist., 6(2): 461-464.

Seiller, G., Roy, R., Anctil, F., 2017. Influence of three common calibration metrics on the diagnosis of climate change impacts on water resources. Journal of Hydrology, 547: 280-295. https://doi.org/10.1016/j.jhydrol.2017.02.004

Shamseldin, A.Y., O'Connor, K.M., Liang, G.C., 1997. Methods for combining the outputs of different rainfall–runoff models. Journal of Hydrology, 197(1): 203-229. https://doi.org/10.1016/S0022-1694(96)03259-3

Sun, W., Trevor, B., 2018. Multiple model combination methods for annual maximum water level prediction during river ice breakup. Hydrological Processes, 32(3): 421-435. https://doi.org/10.1002/hyp.11429

Valéry, A., Andréassian, V., Perrin, C., 2014. ‘As simple as possible but not simpler’: What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments. Journal of Hydrology, 517: 1176-1187.

https://doi.org/10.1016/j.jhydrol.2014.04.058
Vansteenkiste, T. et al., 2014a. Intercomparison of hydrological model structures and calibration approaches in climate scenario impact projections. Journal of Hydrology, 519: 743-755. https://doi.org/10.1016/j.jhydrol.2014.07.062
Vansteenkiste, T. et al., 2014b. Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation. Journal of Hydrology, 511: 335-349. https://doi.org/10.1016/j.jhydrol.2014.01.050
Velazquez, J.A., Anctil, F., Perrin, C., 2010. Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. Hydrology and Earth System Sciences, 14(11): 2303-2317. https://doi.org/10.5194/hess-14-2303-2010
Velázquez, J.A., Anctil, F., Ramos, M.H., Perrin, C., 2011. Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Advances in Geosciences, 29: 33-42. https://doi.org/10.5194/adgeo-29-33-2011
Wang, H.-M., Chen, J., Xu, C.-Y., Zhang, J., Chen, H., 2020. A Framework to Quantify the Uncertainty Contribution of GCMs Over Multiple Sources in Hydrological Impacts of Climate Change. Earth's Future, 8(8): e2020EF001602. https://doi.org/10.1029/2020EF001602
Wu, Y., Wu, S.-Y., Wen, J., Xu, M., Tan, J., 2016. Changing characteristics of precipitation in China during 1960 – 2012. International Journal of Climatology, 36(3): 1387-1402. https://doi.org/10.1002/joc.4432
Xu, C.-y., 2021. Issues influencing accuracy of hydrological modeling in a changing environment. Water Science and Engineering. https://doi.org/10.1016/j.wse.2021.06.005
Yang, W. et al., 2020. Temporal and spatial transferabilities of hydrological models under different climates and underlying surface conditions. Journal of Hydrology, 591: 125276. https://doi.org/10.1016/j.jhydrol.2020.125276
Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. Water Resources Research, 44(9): W09417. https://doi.org/10.1029/2007WR006716
Yin, J. et al., 2020. Projected changes of bivariate flood quantiles and estimation uncertainty based on multi-model ensembles over China. Journal of Hydrology, 585: 124760. https://doi.org/10.1016/j.jhydrol.2020.124760
Zaherpour, J. et al., 2019. Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models. Environmental Modelling and Software, 114: 112-128. https://doi.org/10.1016/j.envsoft.2019.01.003
Zhang, J. et al., 2020. Combining Postprocessed Ensemble Weather Forecasts and Multiple Hydrological Models for Ensemble Streamflow Predictions. Journal of Hydrologic Engineering, 25(1): 04019060. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001871
Zhang, Y., Vaze, J., Chiew, F.H.S., Teng, J., Li, M., 2014. Predicting hydrological signatures in ungauged catchments using spatial interpolation, index model, and rainfall–runoff modelling. Journal of Hydrology, 517: 936-948. https://doi.org/10.1016/j.jhydrol.2014.06.032
Zhao, R.-J., 1992. The Xinanjiang model applied in China. Journal of Hydrology, 135(1): 371-381. https://doi.org/10.1016/0022-1694(92)90096-E
Zhao, R.J., Zuang, Y., Fang, L., Liu, X., Zhang, Q., 1980. The Xinanjiang Model.

Replies to Referee #1

Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size

Yongjing Wan, Jie Chen, Chong-Yu Xu, Ping Xie, Wenyan Qi, Daiyuan Li, Shaobo Zhang

We would like to thank the referee for constructive comments and suggestions. All comments are very helpful for improving this paper and beneficial to our research in general. We have provided detailed point-by-point responses to each comment below and have revised the manuscript accordingly. For clarity, comments are given in black, and our responses are given in blue. Please note that the page and line numbers mentioned in reviewers' comments refer to the original version, while in our reply, they refer to the revised version.

The authors have paid careful attention to the reviewers' comments and significantly improved the paper. I have a few minor suggestions in annotations in the attached file. The paper, especially the latest revisions, would benefit from a proof reading for English quality.

Reply: We sincerely thank the reviewer for the positive evaluation of our manuscript and for providing insightful comments. All comments in annotations in the attachment have been addressed as follows. The English has been carefully revised.

Line 103. Perhaps this is expected, if the metric on which the multi-model average is assessed is some generic metric like NSE. If the metric is more specific like high flows, then we do not expect the combination derived from using different objective functions to work better than simply using a high-flow objective. This is a critical point that needs to be made clear.

Reply: We agree with the reviewer. The use of multi-model combination scheme is expected to benefit from the variation of the parameter sets derived from objective functions targeted at different hydrological processes for producing a better overall simulation.

We have made a clear declaration in line 105 of the revised manuscript.

Line 195. This is poorly written. Needs a new sentence; and needs written more clearly and checked by an English speaker.

Reply: We have made a clear declaration in line 195 of the revised manuscript.

Line 277. Unclear what this means.

Reply: Sorry for the confusion. This sentence has been modified as: Among various averaging methods, GRA, GRB, GRC, and MMSE show similar performance, since they derived from the same optimal weighting group.

We have clarified this in line 276 of the revised manuscript.

Line 319. Random selection from all combinations?

Reply: Yes, and we have clarified this in line 316 of the revised manuscript.

Line 331. not the right word here? "significant", or "much"?

Reply: Thanks, and we have verified this sentence.

Line 335. does not make sense to me. Do you mean "whether all ensemble members contribute to the performance of the combination"

Reply: Yes, and we have made a clear declaration in line 332 of the revised manuscript.

Line 381. Especially when using lumped models, which cannot represent the rainfall and loss variability that tends to be higher in arid areas.

Reply: Thanks for the supplement explanation. We have added this in the revised manuscript in line 378.

Line 428. This gives limited insight into where and why the performances were better. At least, the effect of catchment area should be explored too.

Reply: Thanks for the suggestion. We have analyzed the relationship between the combination performance and catchment area, as shown in figure R1. The result indicates that there is no obvious relationship between two of them. Thus, this resut was not shown in the revised manuscript.

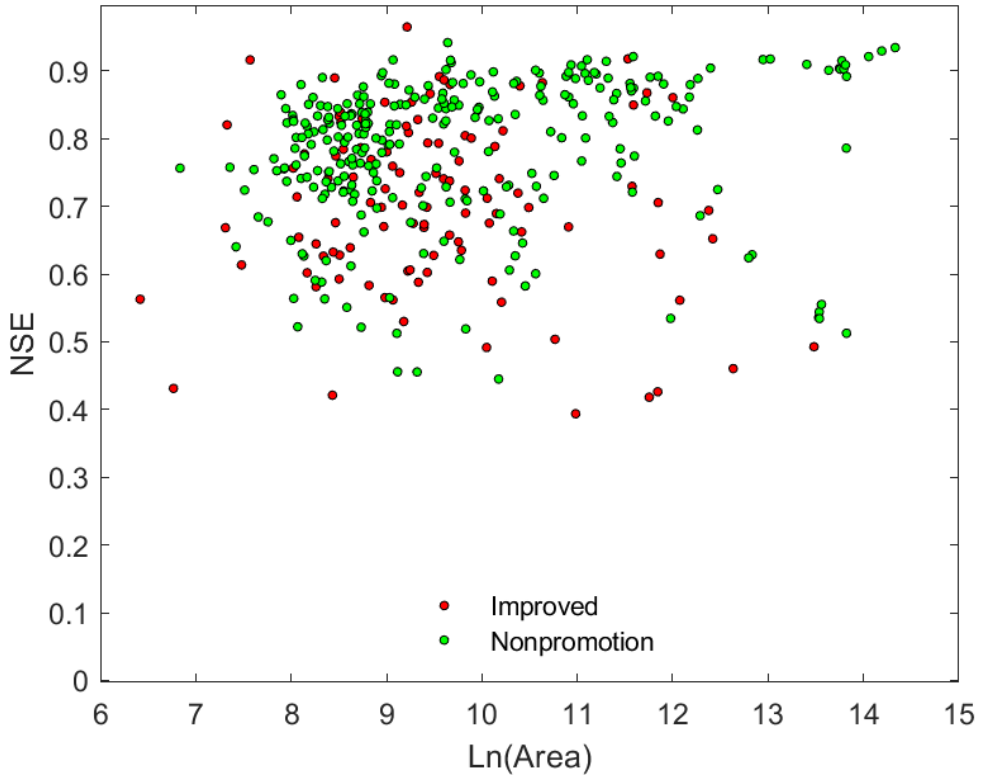

Figure R1. The relationship between the NSE value and catchment area. The green/red markers represent the combination performed better/worse than the best individual member.

Figure

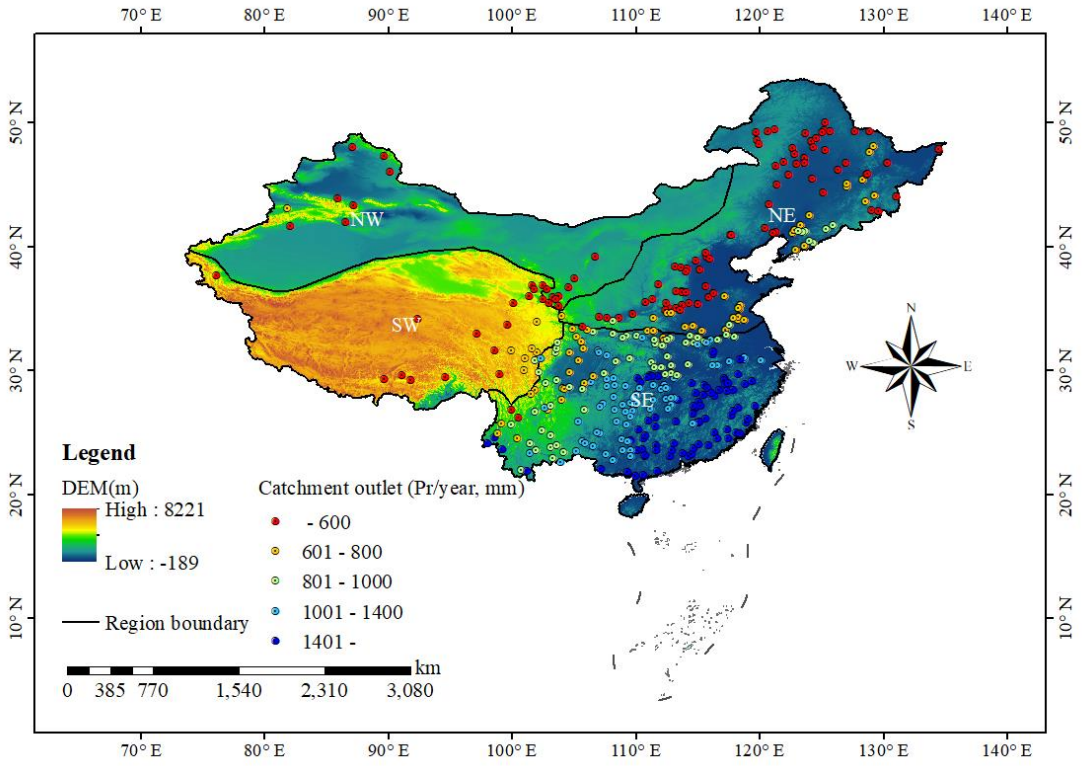

Figure 1. Spatial distribution of the outlets and total annual precipitation (Pr, mm) for 383 catchments in China.

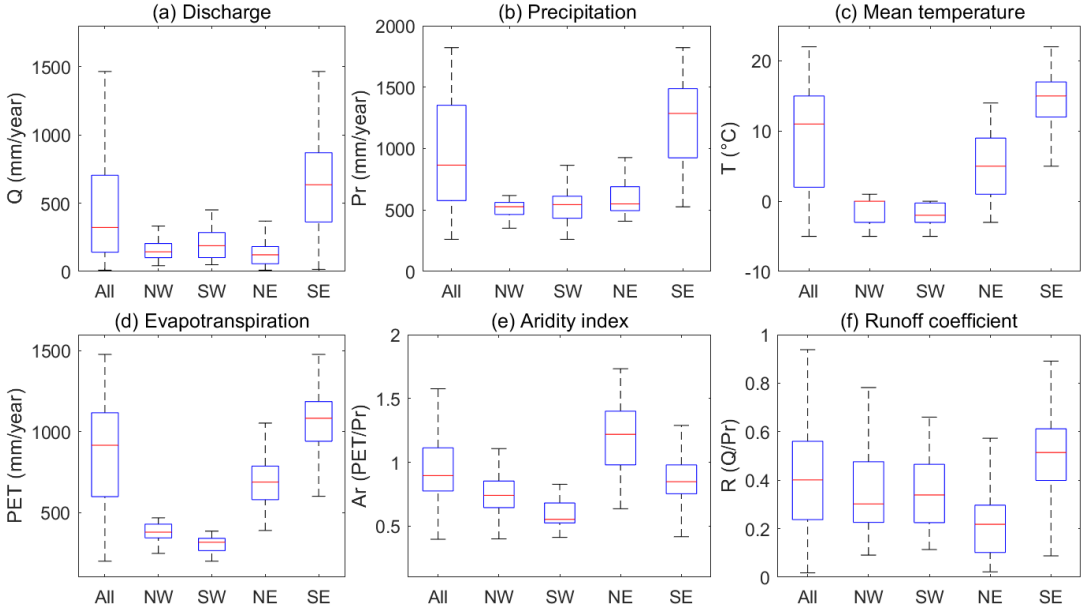

Figure 2. Characteristics of the catchments over different regions. The red line in the boxplots represents the median value, the ends of the boxes represent the 25th and 75th percentiles, the whiskers represent the values at the 5th and 95th percentiles, and outliers are not shown.

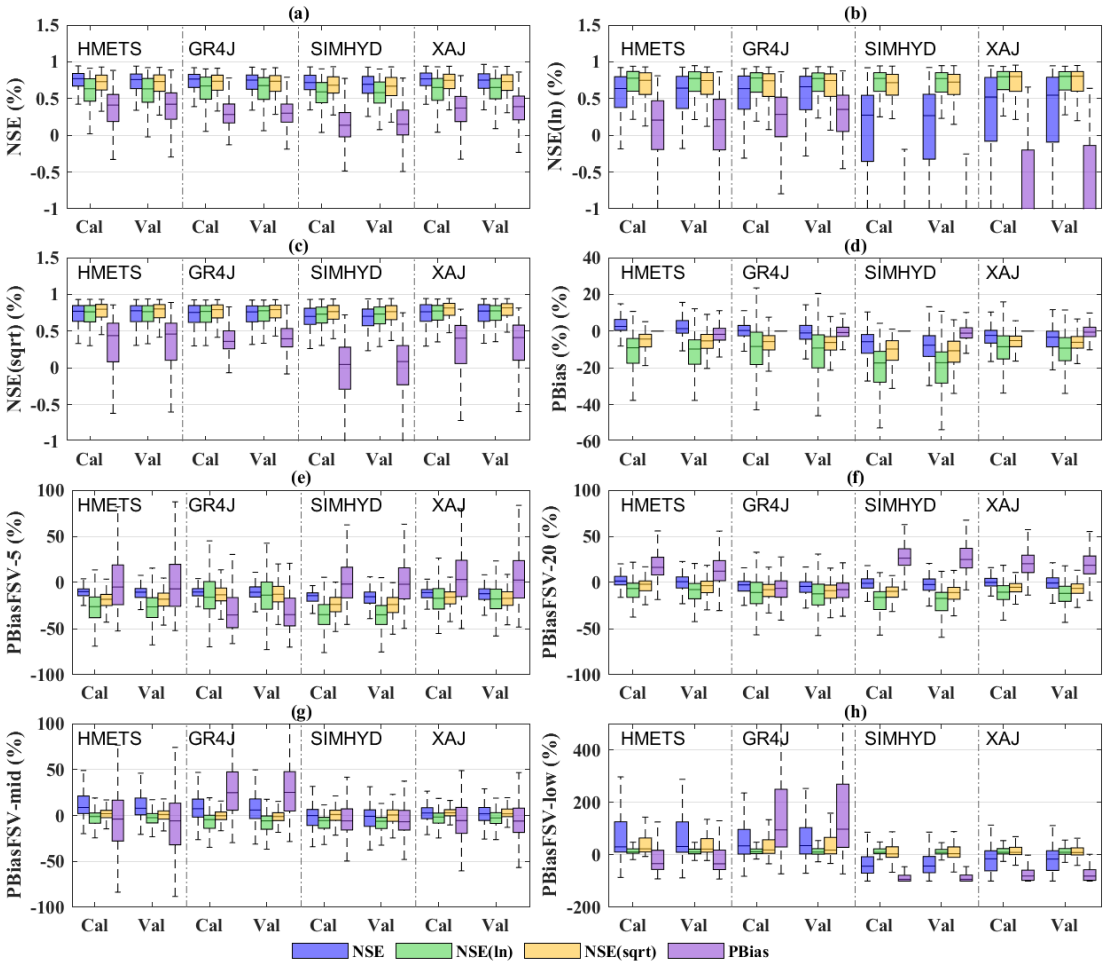

Figure 3. NSE, NSE(ln), NSE(sqrt), PBias, and four PBiasFSV values for 16 ensemble members (model/objective function pairs). The boxplots indicate the spread of performance values of the evaluation metrics over 383 catchments in the calibration and validation periods.

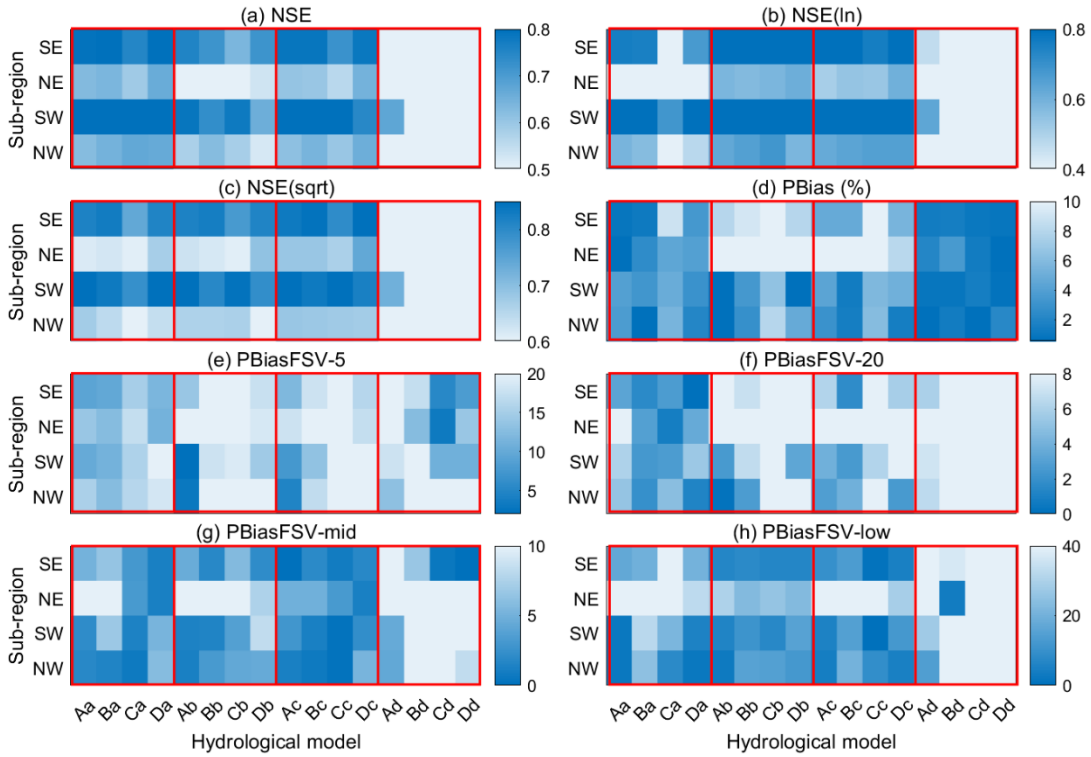

Figure 4. The median value of NSE, NSE(ln), NSE(sqrt), absolute of PBias and four PBiasFSV values in the validation period for the catchments over the four sub-regions. The four hydrological models GR4J, HMETS, SIMHYD and XAJ are denoted by A, B, C and D, respectively. The four objective functions NSE, NSE(ln), NSE(sqrt), and PBias are denoted by a, b, c and d, respectively.

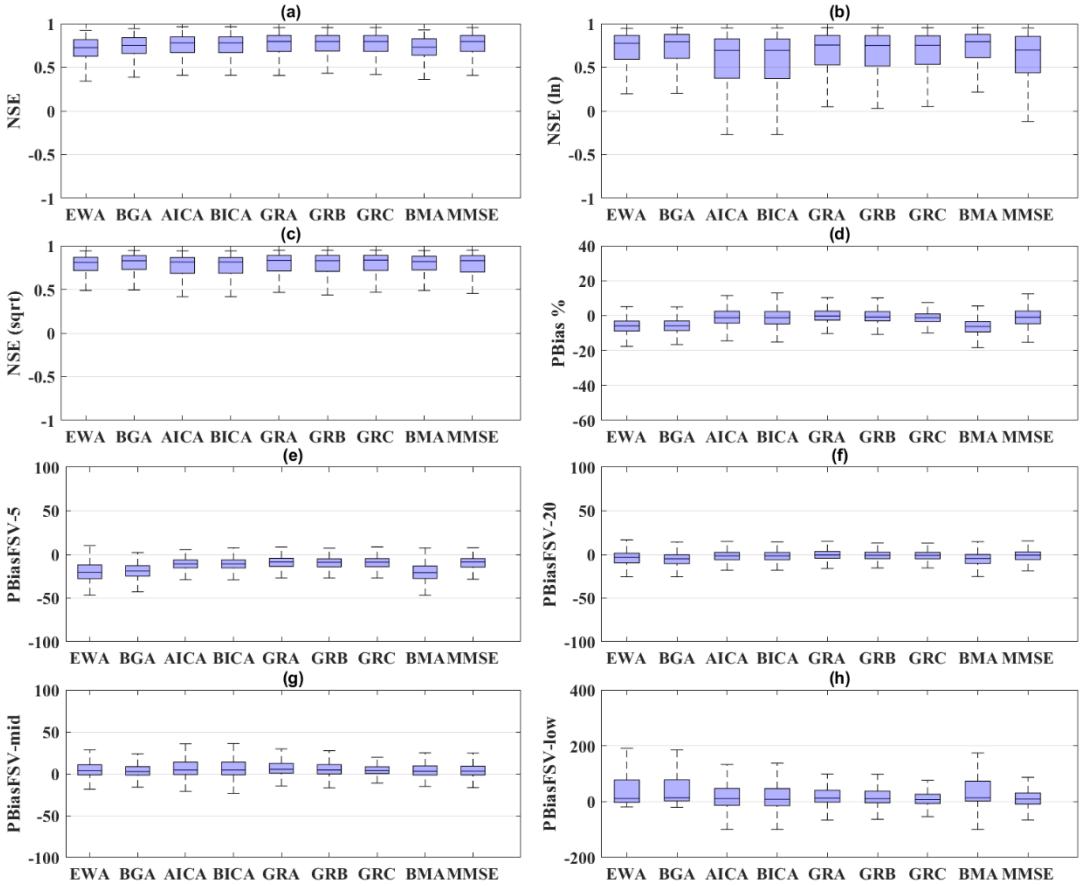

Figure 5. NSE, NSE(ln), NSE(sqrt), PBias, and four PBiasFSV values over 383 catchments for nine multi-model ensemble techniques with combinations of the 16 ensemble members in the validation period.

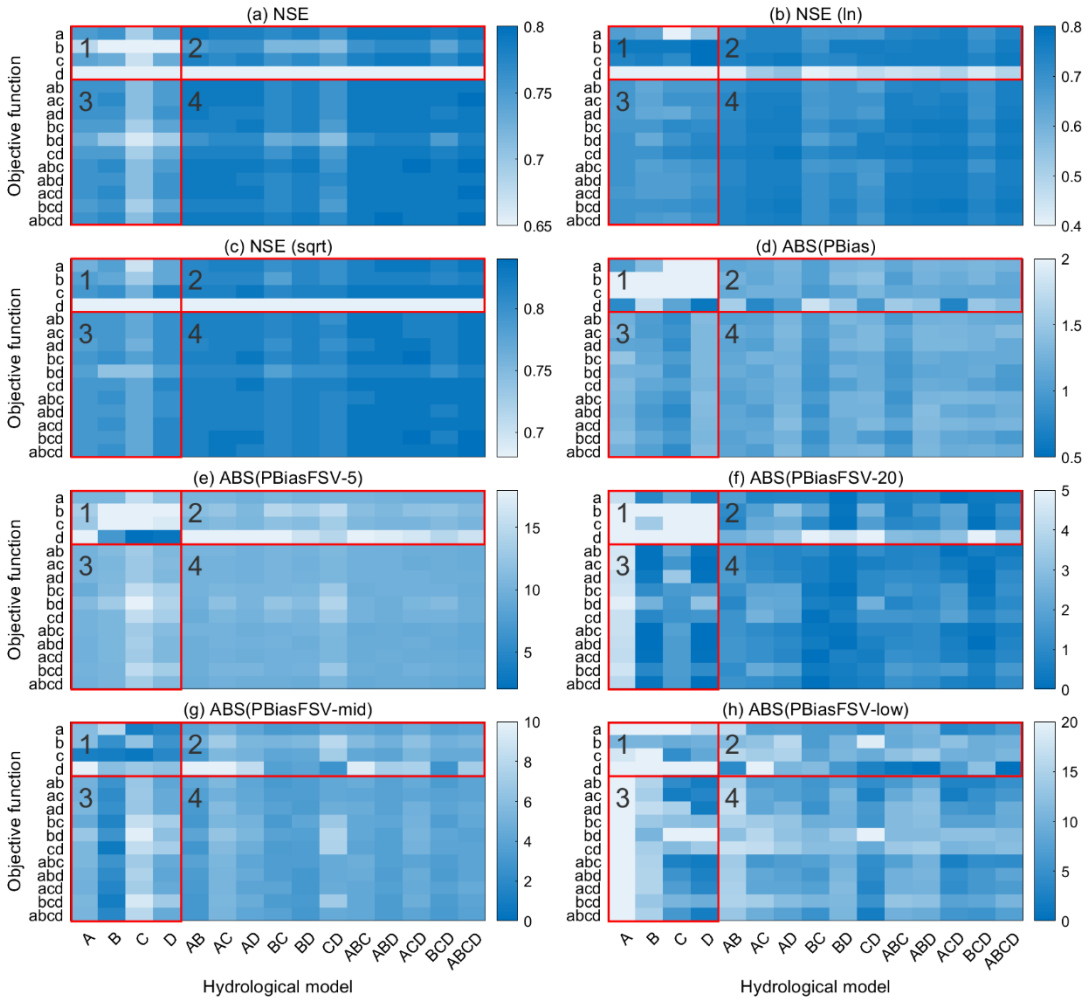

Figure 6. The median value of NSE, NSE(ln), NSE(sqrt), and absolute values of PBias four PBiasFSV over 383 catchments in the validation period based on different combinations in terms of ensemble members. The four hydrological models GR4J, HMETS, SIMHYD and XAJ are denoted by A, B, C and D, respectively. The four objective functions NSE, NSE(ln), NSE(sqrt), and PBias are denoted by a, b, c and d, respectively.

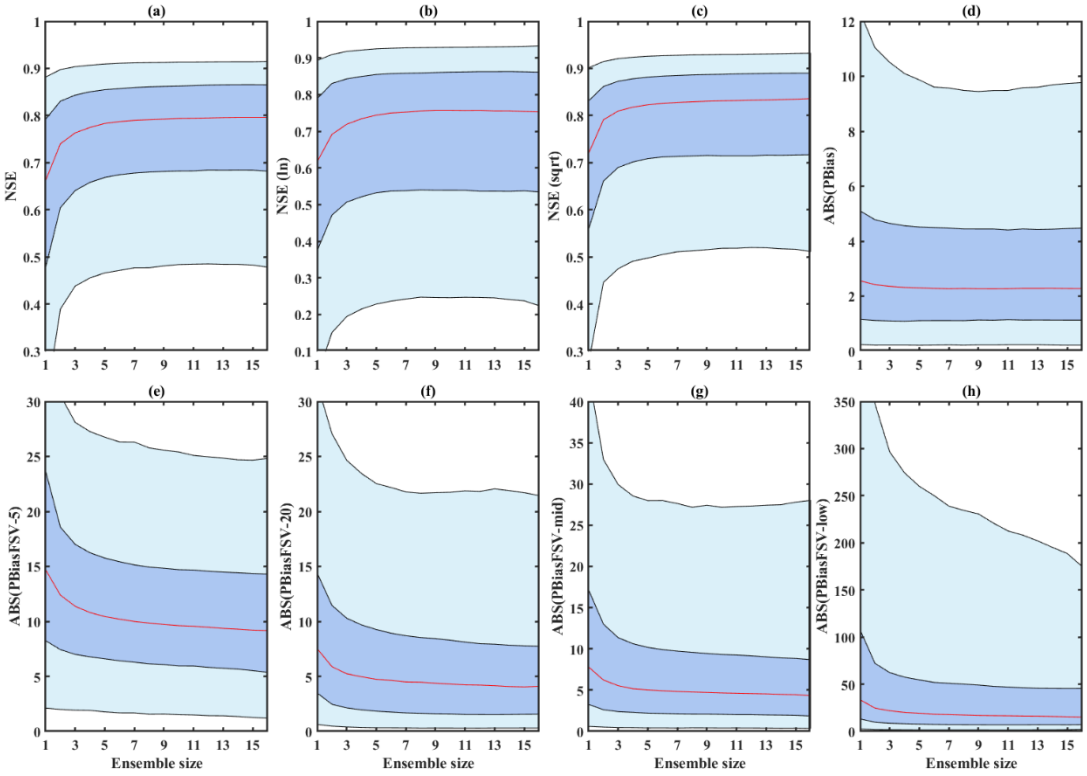

Figure 7. The relationship between the ensemble size and the performance metrics (NSE, NSE(ln), NSE(sqrt), absolute values of PBias and four PBiasFSV) over 383 catchments in the validation period. The light-colored envelopes: the 0.05 and 0.95 interquartile range values. The dark-colored envelopes: the 0.25 and 0.75 interquartile range values. The red lines are the median values.

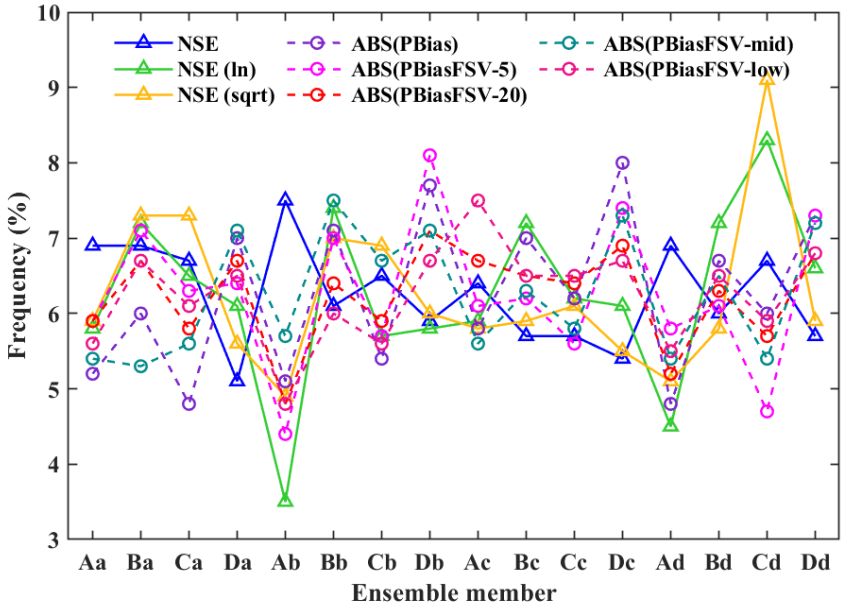

Figure 8. Frequency of individual members is selected in the best combination scheme in terms of different

evaluation metrics (NSE, NSE(ln), NSE(sqrt), absolute values of PBias and four PBiasFSV). The four hydrological models GR4J, HMETS, SIMHYD and XAJ are denoted by A, B, C and D, respectively. The four objective functions NSE, NSE(ln), NSE(sqrt), and PBias are denoted by a, b, c and d, respectively.

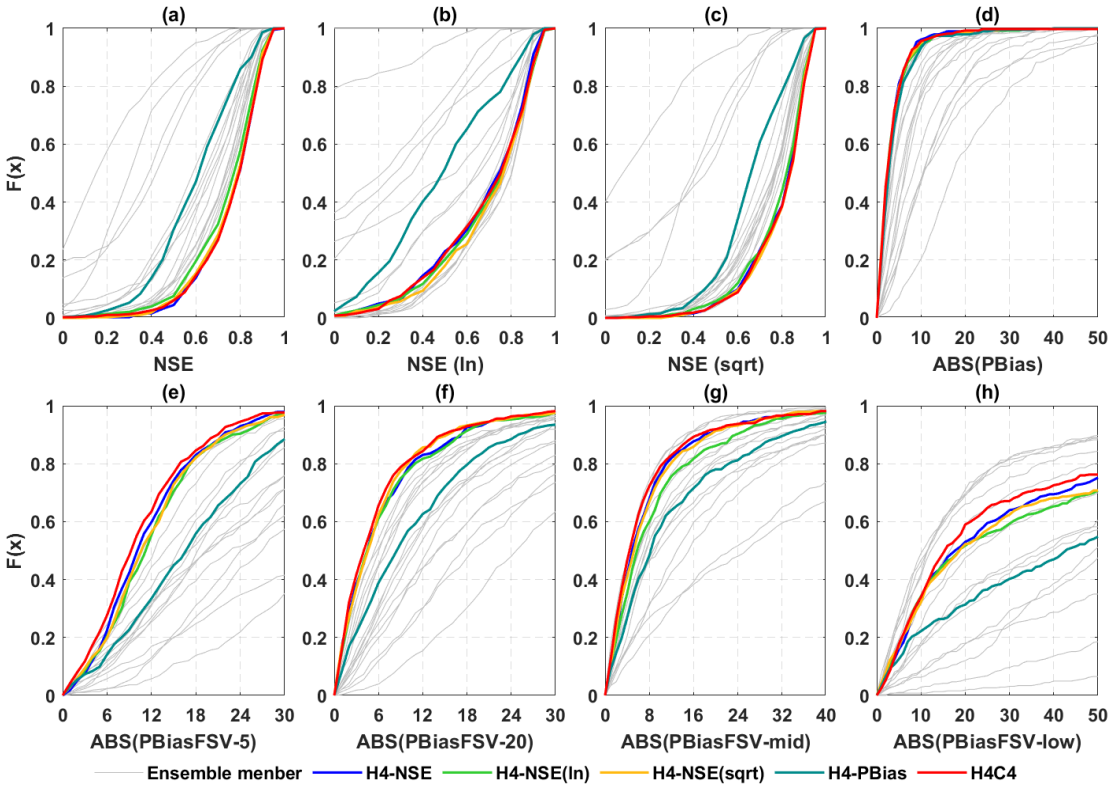

Figure 9. Cumulative distributions of NSE, NSE(ln), NSE(sqrt), absolute values of PBias and four PBiasFSV over the 383 catchments in the validation period.

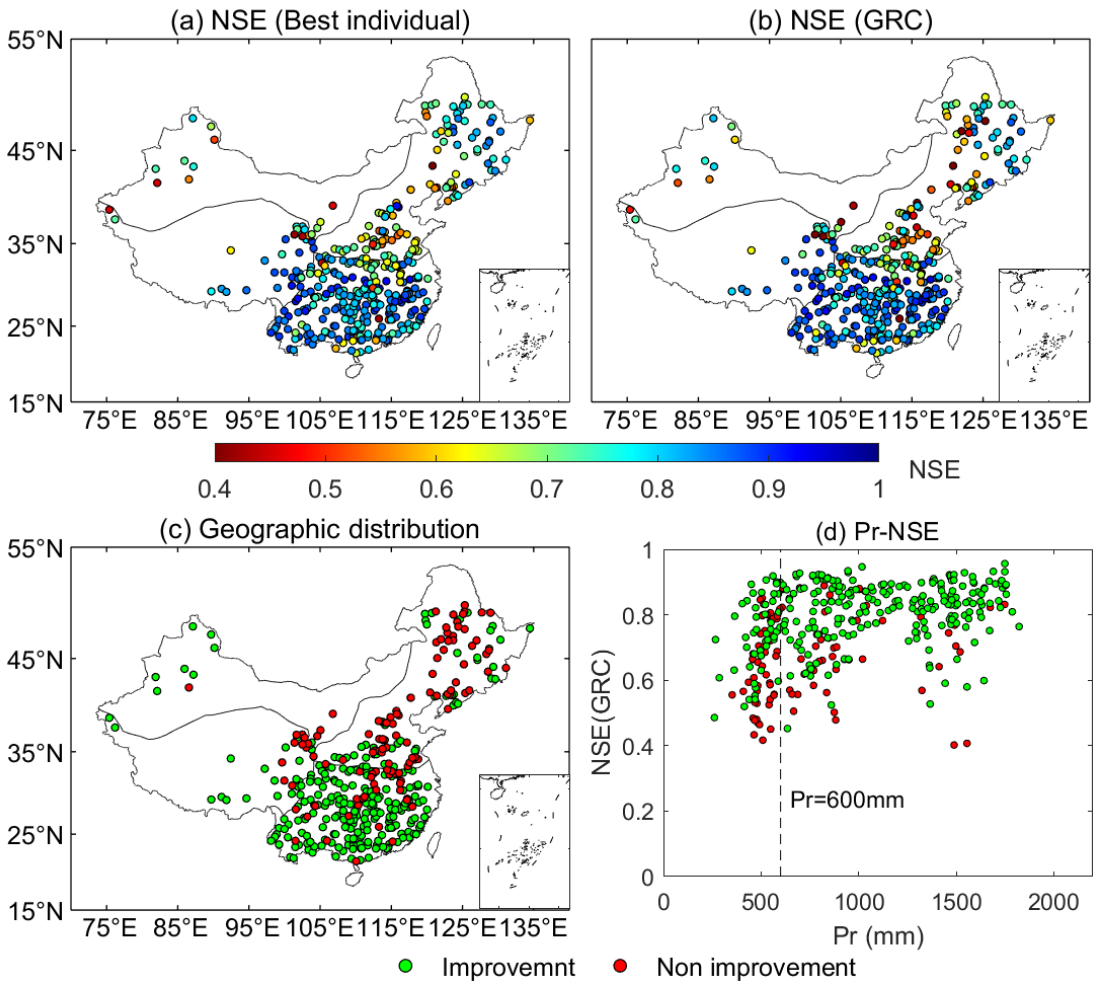

Figure 10. Geographic distribution of the NSE value of (a) the best individual member and (b) the combination over the 383 catchments in the validation period. Comparison of the best individual and GRC averaging method (c) on geographic distribution and (d) related to annual precipitation. The green/red markers present GRC produced results better/worse than the best individual member.

Table 1. Structures of the four lumped conceptual rainfall-runoff models

ID	Model	Number of parameters	Characteristics of the model	References
A	GR4J	6	The effective rainfall is partitioned as a 10:90 split representing direct runoff and delayed runoff, a nonlinear production reservoir with two-unit hydrographs, a routing reservoir	Edijatno et al. (1999); Perrin et al. (2003)
B	HMETS	21	Generation of hypodermic flow and groundwater flow with two linear reservoirs, a routing module with two unit hydrographs, a snowmelt module, an evapotranspiration calculation module	Martel et al. (2017)
C	SIMHYD	11	Two linear reservoirs for the calculation of interflow and base-flow, a nonlinear routing reservoir, an evapotranspiration calculation module	Chiew et al. (2002)
D	XAJ	17	Linear reservoirs for surface flow routing, two recession coefficients for interflow and groundwater flow routing, three-layer evapotranspiration system	Zhao (1992); Zhao et al. (1980)

Table 2. Performance metrics for the evaluation of different phases of the hydrograph.

Performance metric	Sensitive hydrograph phase	Range
NSE	Peak and discharge dynamic	$-\infty \sim 1$
NSE(ln)	Low-flow and discharge dynamic	$-\infty \sim 1$
NSE(sqrt)	Discharge dynamic	$0 \sim 1$
PBias	Overall water balance	$-\infty \sim \infty$
PBiasFSV-5	Tendencies of overestimation and underestimation for FDC very high-segment volume	$-\infty \sim \infty$
PBiasFSV-20	Tendencies of overestimation and underestimation for FDC high-	$-\infty \sim \infty$

		segment volume	
PBiasFSV-mid		Tendencies of overestimation and underestimation for FDC mid-segment volume	-∞ ~ ∞
PBiasFSV-low		Tendencies of overestimation and underestimation for FDC very low-segment volume	-∞ ~ ∞

Table 3. Basic characteristics of the nine multi-model averaging techniques used in this study

Combination method	Acronym	Method description	References
Equal Weights Averaging	EWA	Unweighted average	Shamseldin et al. (1997)
Akaike's Information Criterion Averaging	AICA	Mean of the logarithm of the member variances added a penalty equalling to double the number of calibrated parameters	Akaike (1974)
Bayes Information Criterion Averaging	BICA	Mean of the logarithm of the member variances added a penalty equalling to the number of calibrated parameters times the logarithm of the number of time steps	Schwarz (1978)
Bates and Granger Averaging	BGA	Minimizing the Root Mean Square Error	Bates and Granger (1969)
Granger	GRA	Based on ordinary least squares (OLS) algorithm	Granger and

Ramanathan A			Ramanathan (1984)
Granger Ramanathan-B	GRB	Weights based on ordinary least squares (OLS) algorithm and constrained its sum to unity	Granger and Ramanathan (1984)
Granger Ramanathan-C	GRC	Based on ordinary least squares (OLS) algorithm and bias-corrected the results	Granger and Ramanathan (1984)
Bayesian Model Averaging	BMA	Determining weights by probability distribution functions (PDFs)	Neuman (2003)
Multi-model Super Ensemble	MMSE	Based on ordinary least squares (OLS) algorithm and using the logic of bias reduction with respect to individual member models along with variance reduction in simulation	Krishnamurti et al. (2000)

Table 4. The frequency (%) of individual members obtaining the best score for the performance metrics in the validation period.

Model member	NSE	NSE(ln)	NSE(sqrt)	PBias	PBiasFSV			
					5	20	mid	low
GR4J-NSE	17.2	0	0.8	11.2	13.8	7.8	2.9	3.9
GR4J-NSE(ln)	0.8	9.7	0.8	2.1	6.8	3.7	7.6	12.8
GR4J-NSE(sqrt)	5.5	1.3	16.4	3.7	5.7	3.7	10.4	4.4
GR4J-PBias	0.3	0	0	15.4	1.6	6	3.1	2.1
HMETS-NSE	20.1	0	1.6	11.5	12.5	11	4.4	4.2
HMETS-NSE(ln)	0.5	18.8	1.8	2.9	2.1	4.2	9.1	7.6
HMETS-NSE(sqrt)	4.4	3.1	17.8	4.7	4.4	11.7	8.4	3.1

HMETS-PBias	0	0	0	7	6.5	8.4	3.1	5
SIMHYD-NSE	7	0	1.8	4.4	1.6	9.7	4.7	5.2
SIMHYD-NSE(ln)	0.5	17.5	2.1	0.3	0.5	1.6	5.7	13.6
SIMHYD-NSE(sqrt)	2.3	2.3	7.8	1.8	0.8	1.6	7.3	8.9
SIMHYD-PBias	0	0	0	7	12.8	1.8	4.2	1
XAJ-NSE	26.6	2.1	5.2	7.3	9.7	13.3	5.7	3.4
XAJ-NSE(ln)	0.3	23.5	1.3	2.1	7.3	2.6	7.6	13.8
XAJ-NSE(sqrt)	13.6	21.7	42.6	2.6	2.3	8.1	12	7.8
XAJ-PBias	0.8	0	0	15.9	11.5	5	3.7	3.1

Table 5. Comparison of the multi-model combination and the best individual member for each of the 383 catchments in the validation period. Here, H4-NSE, H4-NSE(ln), H4-NSE(sqrt) and H4-PBias donate four hydrological models calibrated with a specified objective function, and H4C4 donates four hydrological models calibrated with four calibration strategies.

Acronym	NSE	NSE(ln)	NSE(sqrt)	PBias	PBiasFSV			
					5	20	mid	low
H4-NSE	70	11.2	45.2	10.7	14.1	13.8	4.4	8.4
H4-NSE(ln)	32.4	18.3	35.2	11.2	10.4	13.1	4.2	5.7
H4-NSE(sqrt)	60.1	17.8	56.4	11.7	13.3	12.5	7	6
H4-PBias	1.8	1	1.6	7.3	3.1	7.8	5	7.6
H4C4	71	17	55.4	12.3	20.4	13.6	8.4	8.4

Highlights (3 to 5 bullet points (maximum 85 characters including spaces per bullet point)

Highlights：

- Four hydrological models calibrated with four objective functions are compared.
- The Granger Ramanathan average variant C (GRC) method performs the best.
- Using more than nine ensemble members does not further improve performance.
- Combinations of models and objective functions are better than the single model and objective.
- Averaging outperforms the ensemble members except in low-flow simulations.

Credit Author Statement

Yongjing Wan: Conceptualization, Software, Formal analysis, Writing-Original Draft.

Jie Chen: Conceptualization, Writing-Review & Editing, Project administration, Funding acquisition.

Chong-Yu Xu: Conceptualization, Writing-Review & Editing.

Ping Xie: Writing-Review & Editing.

Wenyan Qi: Writing-Review & Editing.

Daiyuan Li: Writing-Review & Editing.

Shaobo Zhang: Data curation.