remote sensing

MDPI

Article

Finding the Optimal Multimodel Averaging Method for Global Hydrological Simulations

Wenyan Qi [1,2], Jie Chen [1,2,*], Chongyu Xu [3] and Yongjing Wan [1,2]

[1] State Key Laboratory of Water Resources & Hydropower Engineering Science, Wuhan University, Wuhan 430072, China; wenyan.qi@whu.edu.cn (W.Q.); wanyongjing@whu.edu.cn (Y.W.)
[2] Hubei Key Laboratory of Water System Science for Sponge City Construction, Wuhan University, Wuhan 430072, China
[3] Department of Geosciences, University of Oslo, 0316 Oslo, Norway; c.y.xu@geo.uio.no
* Correspondence: jiechen@whu.edu.cn

Abstract: Global gridded precipitations have been extensively considered as the input of hydrological models for runoff simulations around the world. However, the limitations of hydrologic models and the inaccuracies of the precipitation datasets could result in large uncertainty in hydrological forecasts and water resource estimations. Therefore, it is of great importance to investigate the hydrological value of a weighted combination of hydrological models driven by different precipitation datasets. In addition, due to the diversities of combination members and climate conditions, hydrological simulation for watersheds under different climate conditions may show various sensitivities to the weighted combinations. This study undertakes a comprehensive analysis of various multimodel averaging methods and schemes (i.e., the combination of the members in averaging) to identify the most skillful and reliable multimodel averaging application. To achieve this, four hydrological models driven by six precipitation datasets were used as averaging members. The behaviors of 9 averaging methods and 11 averaging schemes in hydrological simulations were tested over 2277 watersheds distributed in different climate regions in the world. The results show the following: (1) The multi-input averaging schemes (i.e., members consist of one model driven by multiple precipitation datasets) generally perform better than the multimodel averaging schemes (i.e., members consist of multiple models driven by the same precipitation dataset) for each averaging method; (2) The use of multiple members can improve the averaging performances. Six averaging members are found to be necessary and advisable, since using more than six members only imrpoves the estimation results slightly, as compared with using all 24 members; (3) The advantage of using averaging methods for hydrological modeling is region dependent. The averaging methods, in general, produced the best results in the warm temperate region, followed by the snow and equatorial regions, while a large difference among various averaging methods is found in arid and arctic regions. This is mainly due to the different averaging methods being affected to a different extent by the poorly performed members in the arid and arctic regions; (4) the multimodel superensemble method (MMSE) is recommended for its robust and outstanding performance among various climatic regions.

Keywords: multimodel averaging methods; precipitation datasets; hydrological models; global; climate regions

check for updates

Citation: Qi, W.; Chen, J.; Xu, C.; Wan, Y. Finding the Optimal Multimodel Averaging Method for Global Hydrological Simulations. *Remote Sens.* **2021**, *13*, 2574. https://doi.org/10.3390/rs13132574

Academic Editor: Guy J.-P. Schumann

Received: 21 May 2021
Accepted: 28 June 2021
Published: 1 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The intelligent management of water resources plays a critical role in promoting social and economic development, which needs to be established on the basis of a full understanding of the spatial and temporal distribution of water resources [1,2]. Hydrological models are useful tools to provide hydrological information for water resource management [3]. In the past few decades, numerous hydrological models, from lumped empirical to fully distributed physically based models, have been developed [4]. However, the best-performed models were not consistent under different basin characteristics and various climatologies [5,6]. Multimodel averaging methods, defined as using the outputs from multiple models to obtain one output, are proved to be more efficient in hydrological modeling than their individual members, by numerous studies [7–9]. Since the early paper of Cavadias and Morin [10] introduced the concept of weighted averaging for streamflow simulation, various multimodel averaging approaches have been proposed to find the optimal weights for each member, to minimize the error between the combined and the observed streamflow time series [7,11].

Numerous studies have been conducted to compare various averaging methods in different regions [5,7,12–14]. For example, Diks and Vrugt [5] compared seven averaging methods by using eight conceptual watershed models, and found that the Granger–Ramanathan averaging (GRA) method is superior to other methods. Arsenault et al. [7] compared 9 averaging methods over 429 catchments in the United States, and concluded that the Granger–Ramanathan averaging (GRA, GRB, GRC) methods perform better than any individual member. These studies contributed much to the research in multimodel averaging. However, most of them used a limited number of catchments for a specific region, which did not consider the merits and shortcomings of different averaging methods from a global perspective. Furthermore, the effect on the performance of different averaging methods, caused by various climate conditions and basin attributes, also cannot be revealed.

In addition, with the development of multimodel averaging, many attempts have been made to find the best average scheme. For example, Clark et al. [15] concluded that the outputs from one model, calibrated with different objective functions, could be considered as different models and be used to improve the performance of averaging methods. Arsenault et al. [8] found that promising model averaging results could be achieved by using the outputs from one model, driven by different climate datasets. In recent years, precipitation has been considered as one of the major sources of uncertainty in water resource estimates and may significantly impact the performance of hydrological models in runoff simulations [16–22]. Gauged observations are usually considered as the most accurate estimation for precipitation. However, there are plenty of places with sparsely distributed rain gauges that lack accurate precipitation data for hydrological modeling [23–25]. Therefore, various global-scale gridded precipitation datasets have been developed in the past few decades, to provide precipitation with high temporal and spatial resolution across the world, especially for the ungauged regions [26–31]. However, compared to the real historic precipitation, the gridded precipitation datasets suffered from errors [32–34]. Therefore, compared to the use of a single precipitation dataset, the use of hydrological model outputs, driven by various precipitation datasets as ensemble members, can add the diversity of ensembles and may create a more precise combination for data-sparse regions [8,35,36].

Hydrological model outputs driven by various precipitation datasets are commonly used for uncertainty analysis or climate change impacts in previous studies [3,35,37]. There is limited research on the application of multi-input averaging in the hydrological continuous streamflow simulation [8]. Najafi and Moradkhani [36] used the Bayesian model averaging (BMA) method to estimate runoff extremes, using a single hydrologic model and multiple regional climate model outputs as forcing data, and concluded that the merged signal generally outperforms the best individual signal. Arsenault et al. [8] compared the performance of multimodel and multi-input over the continental United States by using the Granger–Ramanathan C (GRC) method. They found that multi-input averaging provides higher skill than multimodel averaging. Sun et al. [35] used the BMA method to merge streamflows from three global precipitation datasets. They concluded that the hydrologic ensemble using multiple global precipitation products can provide a promising streamflow prediction. However, only one averaging method has been used in the above studies, and whether the improvement in the hydrological runoff simulation of multi-input averaging is independent of averaging methods was not considered. In addition, the number of members used in the multi-input averaging and multimodel averaging was not consistent, which may affect the performance of the averaging methods [8,35].

Accordingly, the first objective of this study is to evaluate and compare the performance of different averaging schemes, i.e., multimodel, multi-input and multi-input model (i.e., members consist of multiple models driven by multiple precipitation datasets). The second objective is to quantify the performances of various averaging methods under different climate regions, to find the optimal averaging methods for global hydrological streamflow simulation. Specifically, four hydrological models, driven by six gridded precipitation datasets (24 combination members) and nine averaging methods, were used to evaluate the performance of different averaging schemes. In addition, the impact of climate conditions on the performance of the averaging methods is investigated by using 2277 watersheds distributed in different climate regions. The large sample size will allow a better understanding of the usage of averaging methods, and thus improving the performance of hydrological runoff simulations, especially for data-sparse regions.

2. Materials and Methods

2.1. Meteorological Data

Various precipitation datasets have been developed in the past few decades. However, some datasets are limited in spatial and temporal coverages. Given the necessity for precipitation data with high resolution, long time period (more than 30 years) and at the global scale, 6 most commonly used gridded precipitation datasets (Table 1) were selected in this study for runoff simulation. The Climate Precipitation Center dataset (CPC) was used [38], which is constructed from global station data and is available from 1979 to present. The Global Precipitation Climatology Center dataset (GPCC) [27] is constructed from global station data and is available from 1981 to 2016. The multi-source weighted-ensemble precipitation V1 (MSWEP) [33] was used, which is based on weighted averaging of several satellites, gauge, and reanalysis products and includes several corrections to improve data quality. This product is available from 1979 to present. The Japanese 55 year ReAnalysis (JRA55) [39] was generated by the Japan Meteorological Agency (JMA) for the period from 1958 to present. The European Centre for Medium-Range Weather Forecast Reanalysis 5 (ERA5) is a reanalysis product and available from 1979 to present. In addition, WATCH forcing data methodology was applied to ERA-Interim dataset (WFDEI) [28], which is available from 1979 to 2016. Thus, these datasets were classified as gauged observation (i.e., CPC and GPCC), satellite-gauge reanalysis (i.e., MSWEP) and reanalysis (i.e., JRA55, ERA5 and WFDEI).

Table 1. Overview of six precipitation datasets.

Dataset	Data Source	Spatial Resolution	Period
CPC	Gauged-based	0.25° × 0.25°	1979–present
GPCC	Gauged-based	0.5° × 0.5°	1981–2016
MSWEP	Satellite-gauge reanalysis	0.25° × 0.25°	1979–present
JRA55	Reanalysis	1.25° × 1.25°	1958–present
ERA5	Reanalysis	0.5° × 0.5°	1979–present
WFDEI	Reanalysis	0.5° × 0.5°	1979–2016

Except for the precipitation datasets, the 0.25° global land evaporation Amsterdam model (GLEAM v3) potential evaporation dataset (1980–2015) was also used for running hydrological models. The potential evaporation of GLEAM v3 is calculated by using the Priestley and Taylor equation [40]. Those datasets with spatial resolutions $< 0.5°$ were resampled to 0.5° using bilinear averaging. The dataset with spatial resolutions $> 0.5°$ was interpolated to 0.5° using the inverse distance weighting method.

2.2. Observed Streamflow Data

The observed daily streamflow data of 2277 watersheds used in this study originate from the Global Runoff Data Centre (GRDC; http://www.bafg.de/GRDC/ (accessed on 28 June 2021)), the Canadian model parameter experiment (CANOPEX) [41] database and some watersheds of China. The size of those watersheds ranges from 2500 km^2 to 50,000 km^2. The streamflow dataset covers the 1982–2015 period, but some of these years are incomplete. All available data were used for each of the watersheds. Figure 1 shows the distribution of these watersheds.

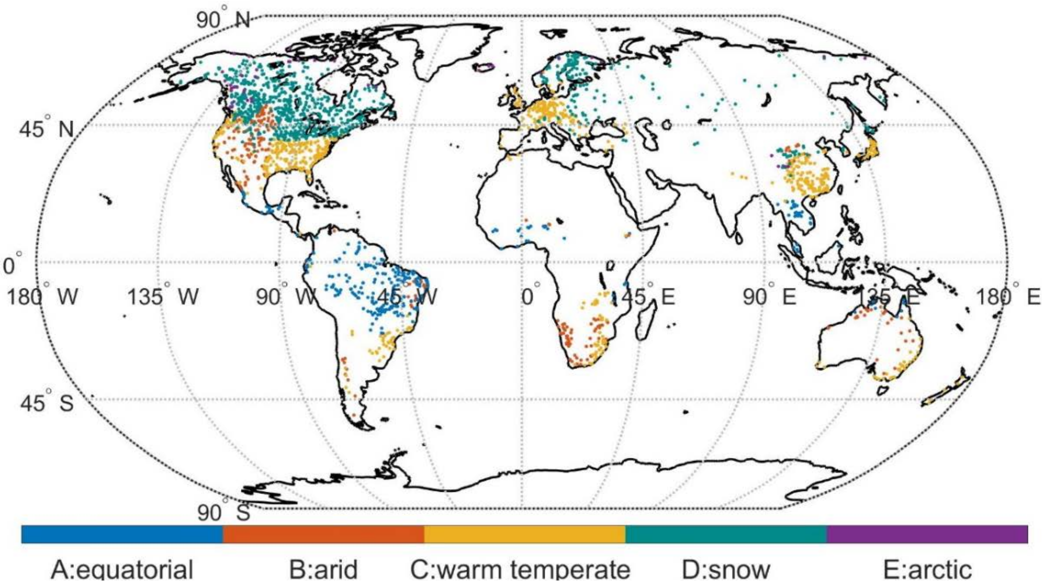

Figure 1. Köppen–Geiger climate classification of the watersheds used in this study. Each point represents the outlet of the watershed.

2.3. Hydrological Models

Four conceptual hydrological models were used to simulate runoff at daily time steps in the present study, as follows: the Génie Rural à 4 paramètres Journalier model (GR4J) [42], the simple lumped conceptual daily rainfall–runoff model (SIMHYD) [43,44], the Xinanjiang model (XAJ) [45,46] and the hydrological model of École de technologie supérieure model (HMETS) [3,47]. These hydrological models were chosen because of the proven effectiveness around the world [37,48,49]. The model structures of these hydrological models differ from each other and the number of the model parameters varies from 6 to 21. Table 2 summarizes the most important information of those 4 hydrological models, e.g., hydrological components, and the number of calibration parameters.

Table 2. Overview of the four hydrological models.

Model	Snow Module	Calibration Parameters	Characteristics of the Model
GR4J	CemaNeige	6	A nonlinear production reservoir with two-unit hydrographs A routing reservoir
SIMHYD	CemaNeige	11	Precipitation loss calculation Surface runoff calculation Two linear reservoirs for the calculation of interflow and base flow
XAJ	CemaNeige	17	Three-layer evapotranspiration system Linear reservoirs for surface flow routing Two recession coefficients for interflow and groundwater flow routing
HMETS	HMETS	21	Generation of surface runoff and delayed runoff after evapotranspiration and infiltration Generation of hypodermic flow and groundwater flow with two reservoirs A routing module

For each watershed, the record of observed streamflow data was split into calibration and validation periods. The first 70% of the record was used for model calibration and the remaining 30% of the record was used for validation. The shuffled complex evolution method optimization algorithm (SCE-UA) [50,51] was used to optimize the hydrological model parameters with the use of the Kling–Gupta efficiency (KGE) [52] as an objective function for calibration. In addition, the 4 conceptual models were calibrated against the observed daily streamflow using each daily precipitation time series as an input. Therefore, there were a total of 54,648 model calibrations (2277 watersheds ×4 hydrological models ×6 precipitation datasets).

2.4. Multimodel Averaging Methods

Nine most commonly used averaging methods were used in this study. They include equal weights averaging (EWA), Bates and Granger averaging (BGA), akaike information criterion averaging (AICA), Bayes information criterion averaging (BICA), Granger–Ramanathan average variant A, B and C (GRA, GRB and GRC), Bayesian model averaging (BMA) and multimodel superensemble (MMSE). These methods are chosen because of the great performance in averaging members [7,9,35,53]. The 9 methods are practical and have different mechanisms in terms of calculating the weight of each member [7,35,53–56]. A summary of these methods is shown in Table 3.

Table 3. Summary of the model averaging methods used in this study.

Name	Method Description	Citation	Sums to Unity	Negative Weights Possible	Bias Correction
EWA	Equal weighted	–	Yes	No	No
BGA	Minimizing the root mean square error	Bates and Granger [57]	Yes	No	No
AICA	Mean of the logarithm of the member variances added a penalty equalling to double the number of calibrated parameters	Akaike [58]	Yes	No	No
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 [56]	Yes	No	No
GRA	Based on the ordinary least squares (OLS) algorithm	Granger and Ramanathan [59]	No	Yes	No
GRB	Based on the OLS algorithm and constrained the weights		Yes	Yes	No
GRC	Based on the OLS algorithm and bias-corrected the results		No	Yes	Yes
BMA	Based on the members' probability distribution functions	Neuman [60]	Yes	No	Yes
MMSE	Based on the OLS algorithm and using the logic of bias reduction with respect to individual member models along with variance reduction in simulation	Vapnik [61], Sivapragasam et al. [62]	No	Yes	Yes

2.5. Multimodel Averaging Schemes

Four hydrological models were driven by 6 precipitation datasets for each watershed, thereby generating 24 hydrological simulations for both calibration and validation periods. The descriptions of different schemes are shown in Table 4 (a total of 11 schemes). Then, the 9 averaging methods were used to generate the weights based on the simulated daily hydrographs of each scheme and observed counterpart for the calibration period. Based on the weights calculated in the calibration period, different members were weighted to generate a single hydrograph for the validation period.

Table 4. Summary of the model averaging schemes used in this study.

Name	Schemes	Members	Declaration
Multi-input	GR4J-COMBINE	GR4J-CPC, GR4J-GPCC, GR4J-MSWEP, GR4J-JRA55, GR4J-ERA5, GR4J-WFDEI	Each scheme runs 15 times based on the combination of 4 out of 6 available members
	SIMHYD-COMBINE	SIMHYD-CPC, SIMHYD-GPCC, SIMHYD-MSWEP, SIMHYD-JRA55, SIMHYD-ERA5, SIMHYD-WFDEI	
	XAJ-COMBINE	XAJ-CPC, XAJ-GPCC, XAJ-MSWEP, XAJ-JRA55, XAJ-ERA5, XAJ-WFDEI	
	HMETS-COMBINE	HMETS-CPC, HMETS-GPCC, HMETS-MSWEP, HMETS-JRA55, HMETS-ERA5, HMETS-WFDEI	
Multimodel	CPC-COMBINE	GR4J-CPC, SIMHYD-CPC, XAJ-CPC, HMETS-CPC	
	GPCC-COMBINE	GR4J-GPCC, SIMHYD-GPCC, XAJ-GPCC, HMETS-GPCC	
	MSWEP-COMBINE	GR4J-MSWEP, SIMHYD-MSWEP, XAJ-MSWEP, HMETS-MSWEP	
	JRA55-COMBINE	GR4J-JRA55, SIMHYD-JRA55, XAJ-JRA55, HMETS-JRA55	
	ERA5-COMBINE	GR4J-ERA5, SIMHYD-ERA5, XAJ-ERA5, HMETS-ERA5	
	WFDEI-COMBINE	GR4J-WFDEI, SIMHYD-WFDEI, XAJ-WFDEI, HMETS-WFDEI	
Multi-input model	ALL	All 24 members	

In total, there are 4 members for multimodel schemes and 6 members for multi-input schemes. To reduce the effects of the different number of ensemble members between multi-input and multimodel schemes, each of the multi-input schemes runs 15 times based on the combination of 4 out of 6 available members for each watershed (Table 4). The average of the 15 combinations represents the final result of the multi-input schemes.

2.6. Statistical Analysis Methods

To evaluate the performance of averaging methods in representing watershed runoff, three statistical indices are utilized, i.e., KGE, Nash–Sutcliffe efficiency (NSE) and accuracy of volume estimates (AVE) [63]. These evaluation criteria were selected for their efficiency to obtain reliable parameter estimation with reasonable performance regarding different parts of the hydrograph [4]. The value of these indices ranges from −∞ to 1, with index value = 1 indicating a perfect fit between the observed and simulated series. The KGE, NSE and AVE are defined as follows:

$$\mathrm{KGE} = 1 - \sqrt{(R-1)^2 + \left(\frac{\overline{Q}_{\mathrm{sim}}}{\overline{Q}_{\mathrm{obs}}} - 1\right)^2 + \left(\frac{\mathrm{CV}_{\mathrm{sim}}}{\mathrm{CV}_{\mathrm{obs}}} - 1\right)^2} \tag{1}$$

$$\mathrm{NSE} = 1 - \frac{\sum(Q_{\mathrm{sim}} - Q_{\mathrm{obs}})^2}{\sum(Q_{\mathrm{obs}} - \overline{Q}_{\mathrm{obs}})^2} \tag{2}$$

$$\mathrm{AVE} = 1 - VE = 1 - \frac{|\sum(Q_{\mathrm{obs}} - Q_{\mathrm{sim}})|}{\sum(Q_{\mathrm{obs}})} \tag{3}$$

where Q_{obs} is the observed runoff and Q_{sim} is the simulated runoff. $\overline{Q}_{\mathrm{obs}}$ is the mean observed runoff and $\overline{Q}_{\mathrm{sim}}$ is the mean simulated runoff. R is the Pearson correlation between observed and simulated runoff. CV_{sim} represents the standard deviation of observed and CV_{obs} represents the standard deviation of simulated discharges.

3. Results

3.1. Performance of Ensemble Members

Figure 2 shows the box plot of the KGE, NSE and AVE values of the hydrological models driven by six precipitation datasets. In the calibration period, the models driven by MSWEP outperform others, and the models driven by JRA55 and CPC perform worse than other precipitation datasets. The performances of ERA5, GPCC and WFDEI are similar, at a median level. Similar results are also observed for the validation period. As for the hydrological models, SIMHYD ranks first, in terms of its good and stable performance in terms of the evaluation criteria for all precipitation datasets. The performances of XAJ and GR4J are similar, at a median level, followed by HMETS.

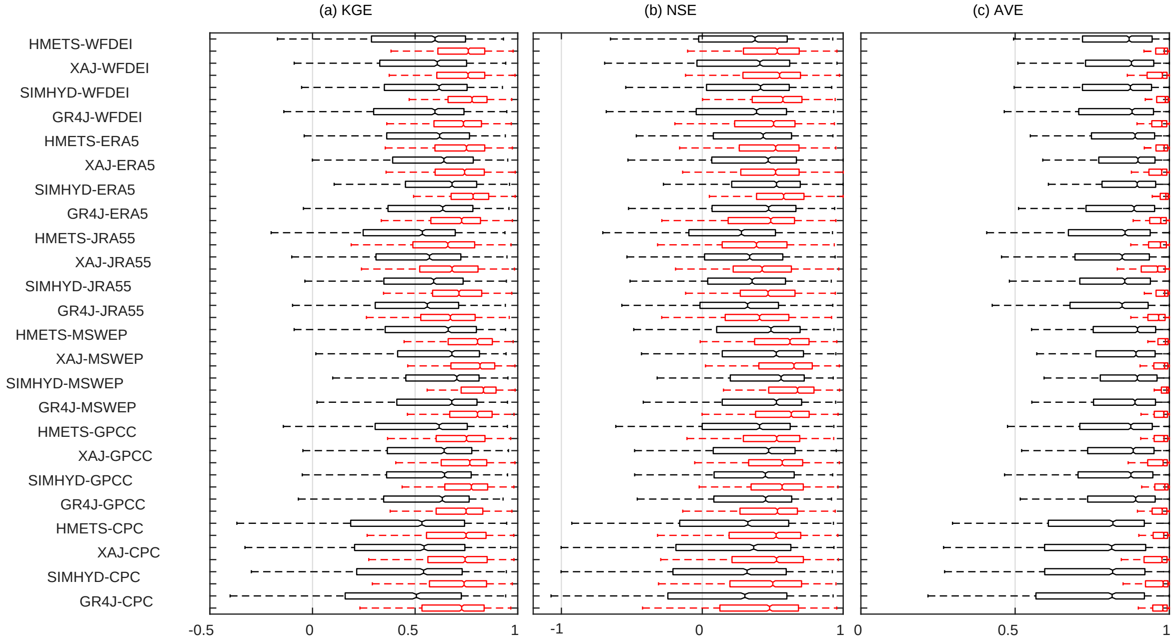

Figure 2. Box plots ofKGE (**a**), NSE (**b**) and AVE (**c**) values for different members. The red and black boxes represent the calibration and validation periods, respectively. The left and right edges of the boxes represent the 25th and 75th percentile values, respectively, while the "whiskers" represent the 5th and 95th percentile values.

Table 5 shows the median KGE values of the hydrological models driven by different precipitation datasets over 2277 watersheds. The median KGE values range from 0.672 to 0.833 among 24 members in the calibration period, and 0.506 to 0.703 in the validation period. This means that the performance differences of these 24 members are large in both the calibration (median KGE of 0.16) and validation periods (median KGE of 0.20). More specifically, a given model driven by different precipitation datasets performs quite differently, with SIMHYD-MSWEP having a median KGE that is approximately 0.12 greater than that of SIMHYD-JRA55. However, the variability due to the hydrological models is much lower for a given precipitation, with SIMHYD-ERA5 having a median KGE that is approximately 0.06 greater than that of GR4J-ERA5. This indicates that the impact of the predictive skill of streamflow from the usage of different precipitations is larger than that from the usage of different hydrological models in this study.

Table 5. Median KGE values of hydrological models driven by different precipitation datasets over 2277 watersheds.

Calibration	GR4J	SIMHYD	XAJ	HMETS	Difference 3
CPC	0.727	0.738	0.743	0.748	0.021
GPCC	0.748	0.774	0.766	0.750	0.026
MSWEP	0.804	0.833	0.816	0.804	0.029
JRA55	0.672	0.714	0.680	0.659	0.054
ERA5	0.727	0.782	0.741	0.750	0.055
WFDEI	0.735	0.777	0.758	0.759	0.042
Difference 1	0.132	0.119	0.136	0.145	
Validation	**GR4J**	**SIMHYD**	**XAJ**	**HMETS**	**Difference 4**
CPC	0.506	0.541	0.544	0.533	0.037
GPCC	0.632	0.644	0.641	0.617	0.027
MSWEP	0.679	0.703	0.679	0.661	0.042
JRA55	0.559	0.590	0.570	0.535	0.055
ERA5	0.635	0.680	0.640	0.619	0.060
WFDEI	0.597	0.617	0.608	0.597	0.020
Difference 2	0.173	0.162	0.135	0.128	

To further evaluate the effectiveness of the 24 members, in terms of climate regions, Figure 3 summarizes the median KGE, NSE and AVE values for 2277 watersheds, in terms of the five Köppen–Geiger climate types. Generally, hydrological models perform much worse in arid climate regions than in the other four climate types, for all precipitation datasets. The complicated hydrological behavior (extreme floods and droughts) and the high hydroclimatic variability cause some challenges in hydrological modeling for this region [64–66]. In addition, compared to other climate regions, the differences in performance among the 24 members are larger in the arctic region. In general, the MSWEP outperforms other precipitation datasets in all climate types. The performance of the MSWEP in different climate types is relatively stable (except for arid regions). JRA55 performs worse than other precipitation datasets in equatorial, arid and warm-temperate regions. However, in the other two climate types, the performance of JRA55 is comparable to that of the other five precipitation datasets. In general, SIMHYD driven by MSWEP (SIMHYD-MSWEP) shows the best performance in most regions (except for the arctic region), and XAJ driven by MSWEP (XAJ-MSWEP) shows the best performance in the arctic region.

3.2. *Performance of Multimodel Averaging Schemes*

The observed streamflow values in the calibration period and the streamflow series from different averaging schemes (Table 4) for the same period were used to calculate the optimal weights for each method. The optimized weights were then used in the validation period, to calculate the averaged flows. Figure 4 shows the KGE values of each scheme and the best individual member (SIMHYD-MSWEP) in the validation period (if not specified, the period in the following figures represents the validation period). The NSE and AVE values are shown in Figure A1.

The following results can be observed: (1) The best- and worst-performing averaging schemes are consistent among the averaging methods and evaluation criteria. For example, in terms of KGE, NSE, and AVE, the best- and worst-performing schemes for all of the methods are ALL and CPC-COMBINE, respectively. (2) The differences in performance among the schemes are various for different methods. For example, the performance difference of AICA is larger than the other methods. (3) Compared to the KGE value, the numbers of schemes that improved the performance of the averaging methods are larger for AVE and NSE. For example, the performances of HMETS-COMBIN are worse than the best single member (SIMHYD-MSWEP) in terms of KGE. However, in terms of NSE and AVE, HMETS-COMBIN performs better than SIMHYD-MSWEP for most averaging methods (except for AICA and BICA). (4) The MMSE method obviously outperforms the others in terms of AVE for each averaging scheme. As depicted in Table 3, MMSE uses bias correction and variance reduction in the simulation, to further improve the averaging quality. The AVE value is one minus the volume error, which is partially corrected by this method [67,68]. Therefore, the MMSE method shows great performance in terms of AVE.

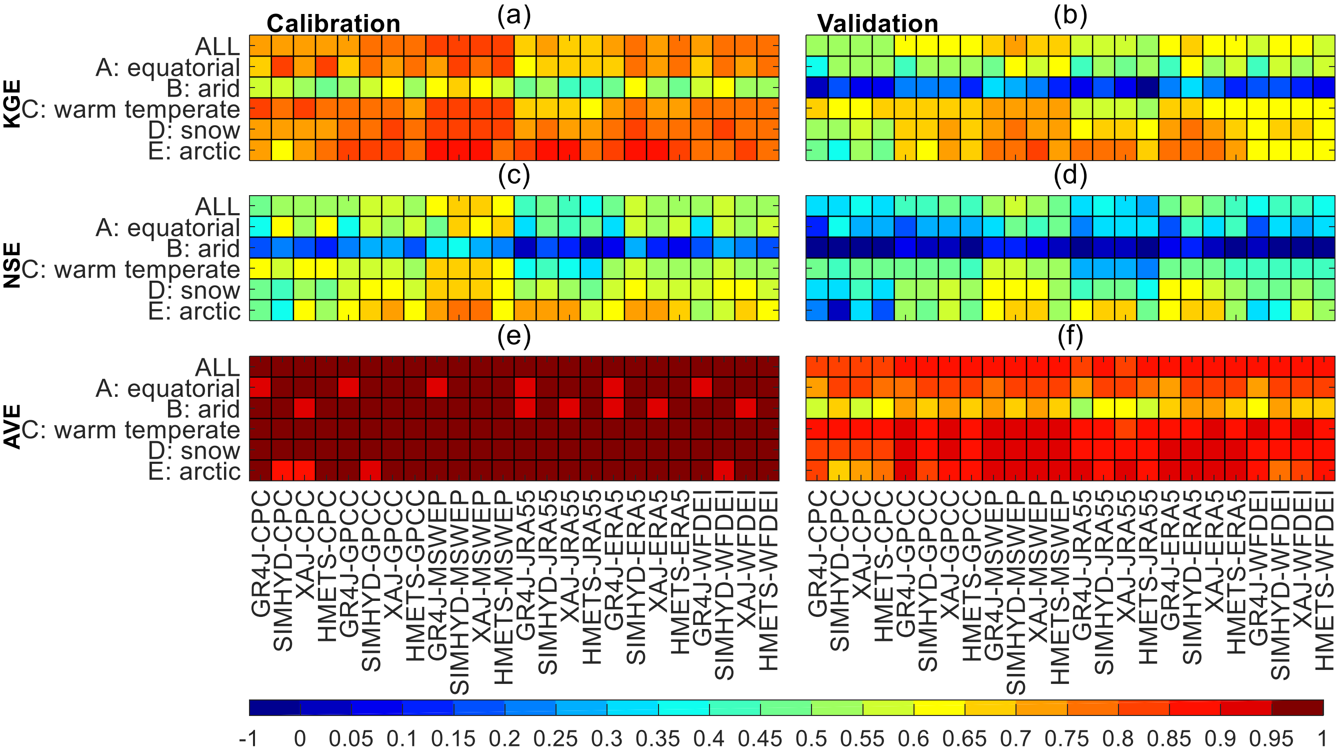

Figure 3. Median KGE, NSE and AVE values of each member in different climate regions for both calibration period (**a**,**c**,**e**) and validation period (**b**,**d**,**f**).

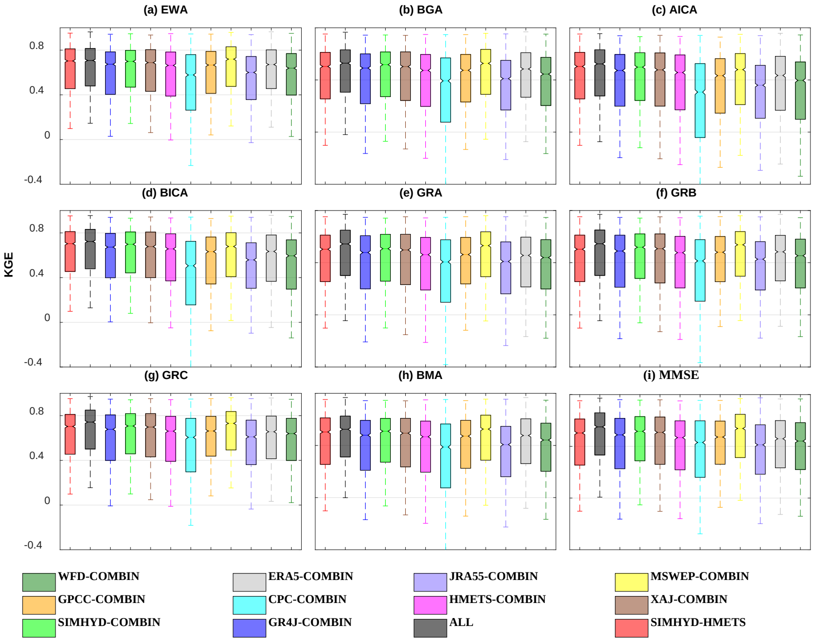

Figure 4. Box plots of KGE valuesof different averaging schemes for each averaging method (**a**–**i**).

The graphic demonstration of the comparison results of various averaging schemes is shown in Figure 5. In general, the multi-input averaging schemes perform better than the multimodel averaging schemes, especially for AICA and BICA. Table 6 further shows the median values of KGE, NSE, and AVE over nine averaging methods for different averaging schemes. The median KGE values across 2277 watersheds and nine averaging methods (2277 × 9) are 0.68 for GR4J-COMBIN, 0.71 for SIMHYD-COMBIN, 0.70 for XAJ-COMBIN, and 0.66 for HMETS-COMBIN. Only one multimodel averaging scheme (i.e., MSWEP-COMBIN) performs better than HMETS-COMBIN (the worst-performed multi-input averaging scheme). Overall, the multi-input model averaging scheme (ALL) shows the best performance for different averaging methods and evaluation criteria. As shown in other studies, using more members can increase the averaging performance [7–9,69].

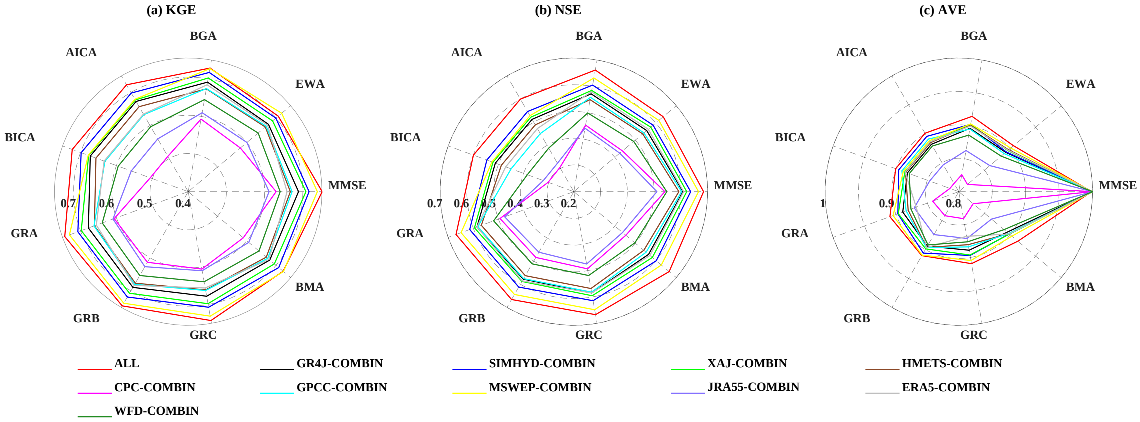

Figure 5. Median values of KGE (**a**), NSE (**b**), and AVE (**c**) over all watersheds for all multimodel averaging methods under different schemes.

Table 6. The median value of three criteria over 2277 watersheds and 9 averaging methods for different averaging schemes.

Averaging Schems	SIMHYD-MSWEP	ALL	GR4J-COMBINE	SIMHYD-COMBINE	XAJ-COMBINE	HMETS-COMBINE
KGE	0.70	0.73	0.68	0.71	0.70	0.66
NSE	0.56	0.66	0.57	0.60	0.58	0.55
AVE	0.90	0.91	0.89	0.90	0.90	0.90
Averaging Schems	**CPC-COMBINE**	**GPCC-COMBINE**	**MSWEP-COMBINE**	**JRA55-COMBINE**	**ERA5-COMBINE**	**WFDEI-COMBINE**
KGE	0.59	0.67	0.73	0.61	0.66	0.64
NSE	0.45	0.56	0.63	0.44	0.58	0.50
AVE	0.83	0.89	0.90	0.86	0.90	0.88

3.3. Impacts of Averaging Size on Performances of Multimodel Averaging Methods

Based on the comparison of different averaging schemes, the multi-input model averaging scheme (ALL), which includes the largest number of members (24 members), shows the best performance (largest median KGE, NSE and AVE). This is in line with the previous studies, which concluded that a large number of members leads to an improvement in hydrological simulating abilities [8,37,70]. However, how the number of members used in the averaging influences the performance of the averaging methods deserves further investigation. Figure 6 shows the KGE values of averaging members, ranging from 2 to 24 for different averaging methods. The members of different averaging numbers are generated by resampling all of the available members (24 in total) 100 times for 2277 watersheds. The 10th and 25th percentiles represent poor simulation performances, while the 75th and 90th quantiles represent good simulation performances. The average performance is represented by the 50th quantile [37].

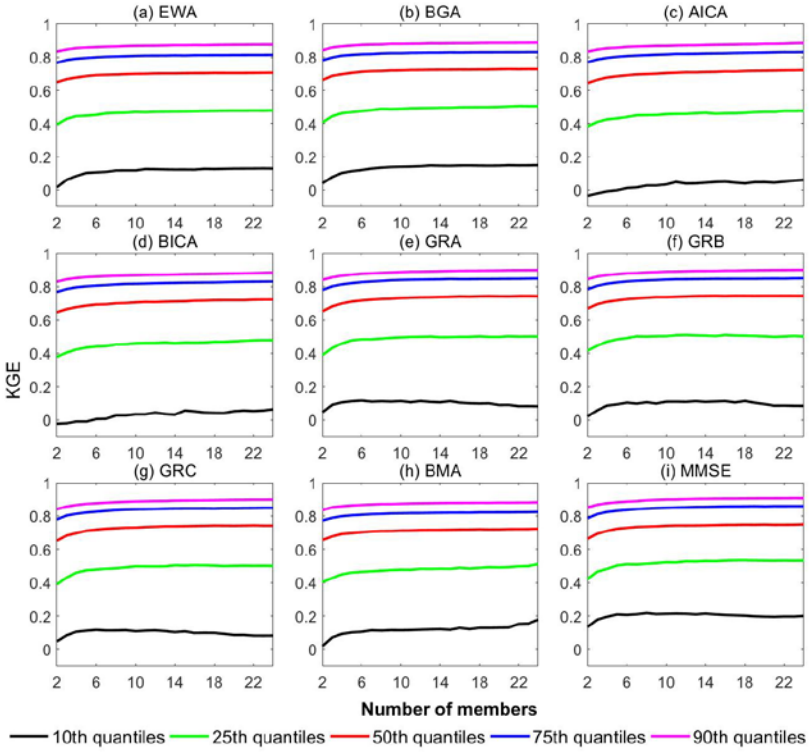

Figure 6. The relationship between the averaging size and the median KGE values over 2277 water sheds for different averaging methods (**a**–**i**).

Generally, the median value of KGE is improved from two to multiple averaging members for all of the averaging methods. In particular, the improvement is very significant when the number of members increases from two to six. However, when the number of averaging members is larger than six, the improvement tends to be minimal. For example, the difference in the median KGE value is no more than 0.05 between the use of 6 members and all 24 members. Taking into account how time-consuming using all of the available observations is, we recommend using 6 to 10 members for each averaging method. The same conclusions can be held with the other percentiles, except the 10th percentiles. For the 10th percentiles, the optimal number of averaging members varies greatly for the different averaging methods. Thus, a consistent conclusion cannot be drawn on the optimal number of members for this percentile. The same conclusions can be drawn for NSE and AVE (see Figure A2).

3.4. Comparison of Multimodel Averaging Methods

3.4.1. Performance of Multimodel Averaging Methods over Global

To further evaluate the performance of different averaging methods, Figure 7 shows the comparison of the best individual member (SIMHYD-MSWEP) and different averaging methods under the multi-input model averaging scheme. This averaging scheme was chosen because of its appreciable performance over different methods (Figures 4 and 5). More than half of the watersheds show that the averaging methods outperform the individual members, in terms of KGE. The same conclusions can be drawn for NSE and AVE (Figure A3). Overall, the MMSE and the Granger–Ramanathan average group (GRA, GRB and GRC) are more efficient than the other methods. Figure 8 shows the geographical information of each averaging method. The points in blue represent the watersheds that obtained improved hydrological performance by using the averaging methods. The results show that the KGE values increase for most of the watersheds in Northeast America, Southern China, and along the Atlantic coast of Europe, when using averaging methods. However, these methods perform inadequately for watersheds in the American tropics, Northwestern America and the Middle East. Considering the high variety in the performances of the averaging methods around the world, it is worth investigating whether the well-performed averaging methods are climate dependent and how the performances of the averaging methods vary with watersheds under different climate conditions.

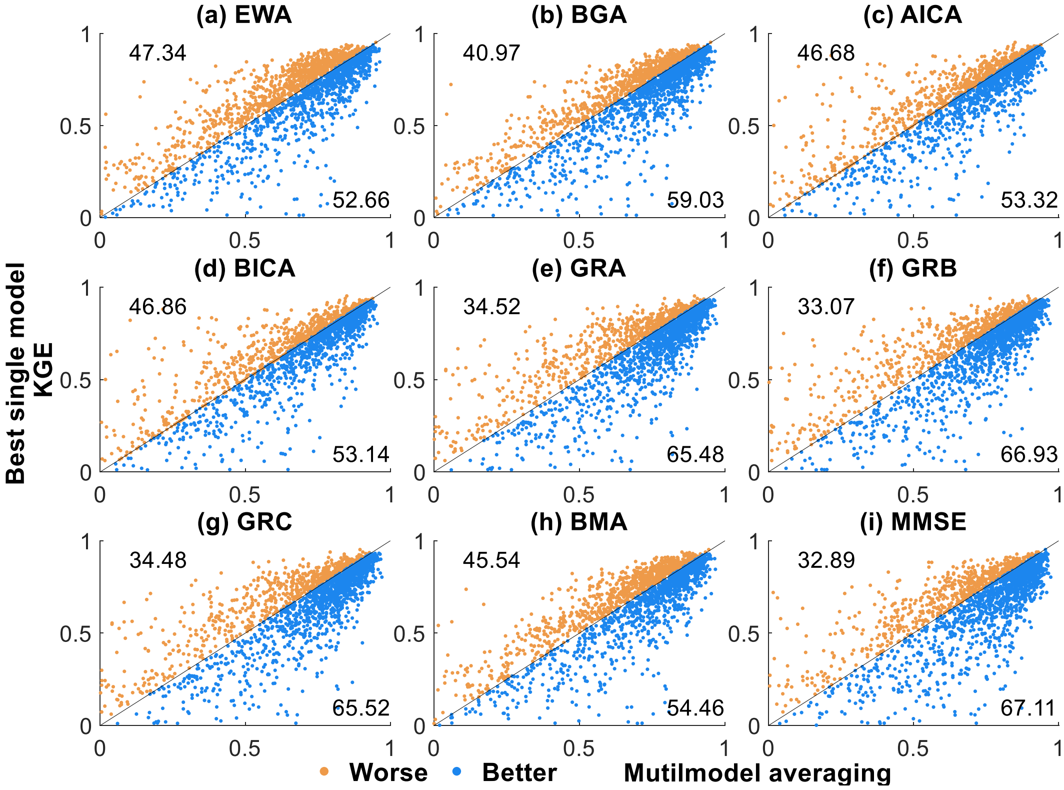

Figure 7. Comparison of nine multimodel averaging methods (**a**–**i**) and the best-performing model (SIMHYD-MSWEP) for each of the watersheds in this study. Model averaging that produced results better than the best member will generate markers under (or to the right of) the 45-degree line. The number in the upper left corner represents the percent of watersheds that, using the multimodel averaging method, perform worse than the best-performing model. The number in the lower right corner represents the percent of watersheds that, using the multimodel averaging method, perform better than the best-performing model.

3.4.2. Performance of Multimodel Averaging Methods in Multiple Climatic Regions

To better understand the impact of climate conditions on the performance of averaging methods, Figure 9 shows the KGE, NSE and AVE values for each averaging method, and the best individual member (SIMHYD-MSWEP) under different climate regions. The averaging methods perform much worse in the arid region than in the other four climate regions, which is consistent with the performance of single hydrological models. The nine averaging methods all perform better than, or are comparable to, the best individual member for most of the climate regions, except for the arid and arctic regions. For the arid region, the GRA, GRB and GRC consistently perform worse than the best individual member for different criteria. Generally, the Granger–Ramanathan average group are approximately equal under different climate regions and criteria. For the arctic region, EWA, BGA and BMA perform worse than the best individual member, in terms of KGE and AVE. In terms of NSE, the performance of the averaging methods is better than the other criteria in this region.

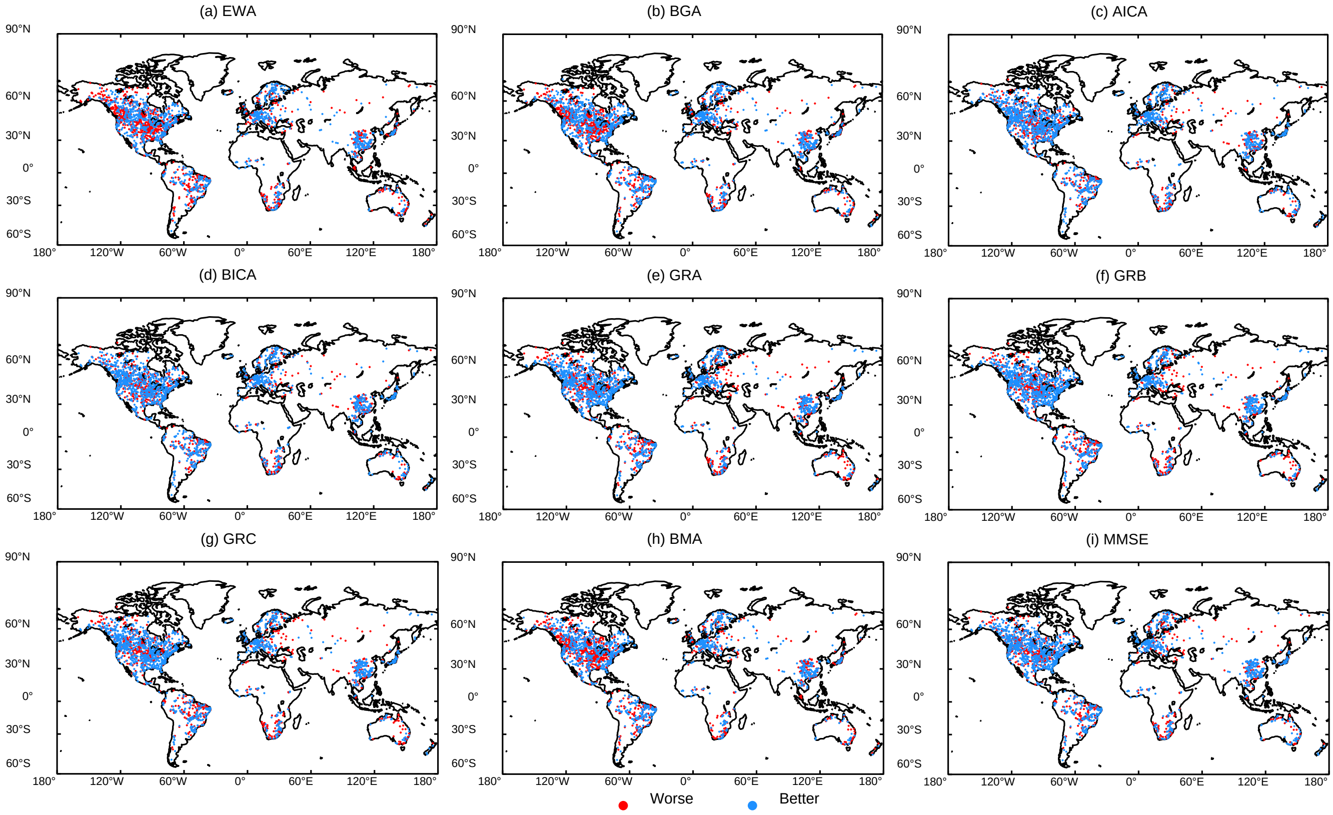

Figure 8. The spatial distribution of watersheds for which the averaging methods (**a**–**i**) performed better than the best single model (blue) and the watersheds for which the averaging methods were not as good as the best single model (red).

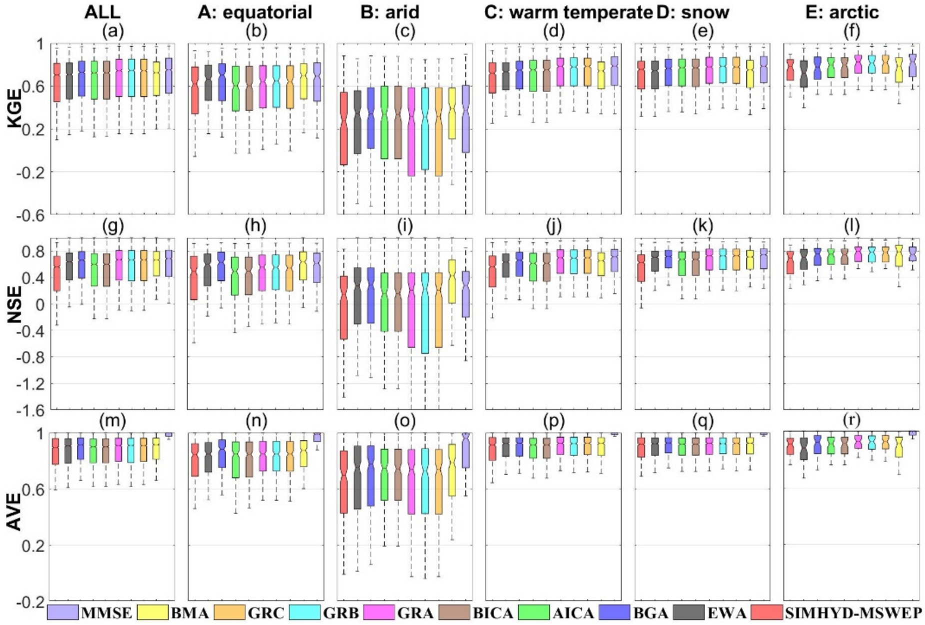

Figure 9. Box plots of KGE (**a**–**f**), NSE (**g**–**l**) and AVE (**m**–**r**) values of different averaging methods in different climate regions.

The performance of the averaging methods is different among the five climate regions. According to both the KGE and NSE results, it is apparent that the MMSE and BMA methods outperform others in the equatorial and arid regions. The performances of the averaging methods are very similar in warm-temperate and snow regions. The MMSE method marginally performs better than the others in these regions. In the arctic region, the Granger–Ramanathan group performs better than the other methods. The AICA and BICA methods are approximately equal and their performances are lower than others in equatorial, warm-temperate and snow regions. It is the same in terms of AVE, except for the MMSE method, which is the best averaging method in all of the climate regions. The reason for the MMSE method being outstanding in terms of AVE is the bias correction and variance reduction used in this method, as mentioned in Section 3.2.

Table 7 shows the frequency of watersheds where the KGE value of the different averaging methods is larger than the best-performing individual model (SIMHYD-HMETS) in different climate regions. It can be seen that the outperformance of the different averaging methods exceeds 50% of the watersheds for most of the climate regions, except AICA and BICA in the equatorial region, EWA and BMA in the arctic region, and seven out of nine methods in the arid region. In general, the improvements in hydrological simulation, caused by the different averaging methods, are the largest in the warm-temperate region, followed by the snow and equatorial regions. The differences in the averaging method performances in arid and arctic regions are more significant than those in equatorial, warm-temperate and snow regions. This indicates that more consideration should be given to the selection of the averaging method used in these two regions.

Table 7. The percentage of watersheds where the KGE value of different averaging methods is bigger than the best-performing individual model (SIMHYD-HMETS) in different climate regions.

Climate Reions (Number of Cathments)	EWA	BGA	AICA	BICA	GRA	GRB	GRC	BMA	MMSE
All (*n* = 2277)	52.7	59.0	53.3	53.1	65.5	66.9	65.5	54.5	67.1
A:equatorial (*n* = 293)	56.0	65.5	47.4	46.4	55.6	57.7	55.6	61.1	58.7
B:arid (*n* = 247)	48.6	54.7	48.2	47.8	45.3	47.0	45.3	44.5	56.7
C:warm temperate (n = 717)	56.5	62.3	56.9	56.9	73.5	74.5	73.5	58.3	72.8
D:snow (*n* = 970)	50.3	55.8	53.6	53.5	67.3	69.1	67.3	52.6	68.2
E:arctic (*n* = 50)	44.0	58.0	56.0	58.0	72.0	70.0	74.0	46.0	64.0

4. Discussion

This study used 6 global gridded precipitation datasets to drive 4 hydrological models for streamflow simulations over 2277 watersheds around the world, and took each of the outputs as a member for model averaging. To find the best combination of different members and improve the predictive skill in hydrological runoff modeling, eleven averaging schemes classified as multi-input, multimodel and multi-input model, and nine averaging methods were considered for streamflow averaging. The results show that the combination of different members may largely impact the performance of the averaging methods. The performance of multimodel averaging schemes largely depends on the input data. In general, the multi-input averaging schemes perform better than multimodel averaging schemes. Global gridded precipitation datasets are laden with intrinsic and structural errors, due to the different interpolation schemes, and they are likely all different from the real climate data [32,33]. Therefore, a given model driven by different precipitation datasets performs quite differently (Table 5). For example, the median KGE value of SIMHYD-MSWEP is approximately 0.12 greater than that of SIMHYD-JRA55. The improvement in the multi-input schemes may be partly because of the reduction in the uncertainties caused by the inputs between the simulated and observed hydrograph [7,35]. Theoretically, using real climate data may reduce the advantage of the multi-input schemes. However, real precipitation varies greatly in time and space, and therefore is extremely challenging to observe and estimate [32]. Therefore, multi-input averaging schemes can be a powerful tool for hydrograph simulations, and can provide an advantageous way to support reasonable runoff prediction and water management, especially in ungauged basins [35].

Equifinality is defined as a hydrological model having multiple sets of parameters that lead to equally acceptable model performance, which is considered to be one of the uncertainties in hydrological modeling [71,72]. Theoretically, using the outputs from equifinal parameter sets as averaging members may improve the performance of averaging methods, by reducing the errors caused by the parameter set uncertainty. The performance of averaging methods, by combining the outputs of 10 equifinal parameter sets, was tested. Four models driven by MSWEP were calibrated ten times by the shuffled complex evolution method (SCE-UA), with different initial random seeds. The results show that using the outputs of 10 equifinal parameter sets, calibrated from a hydrological model driven by specific precipitation as averaging members, cannot improve the performance of the averaging methods (Figure 10). This conclusion is consistent with that in Arsenault et al. [8].

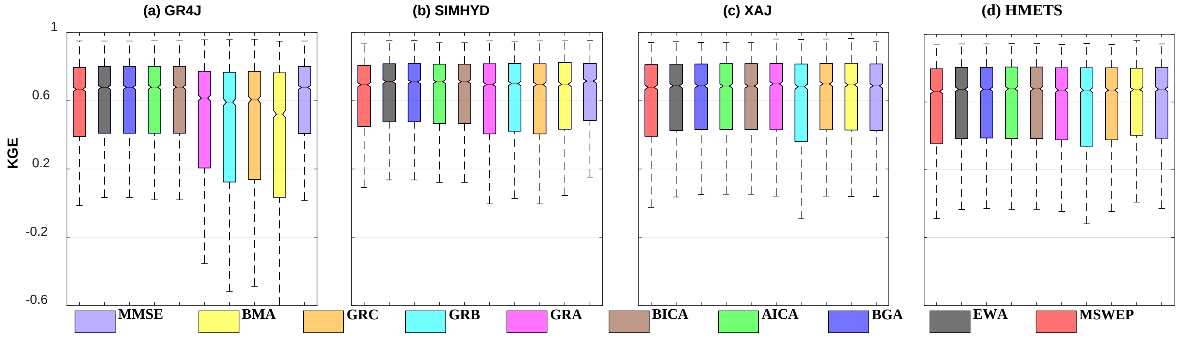

Figure 10. Box plots of KGE values of different averaging methods calculated by the combination of 10 equifinal parameter sets of four hydrological models (**a**–**d**).

The KGE was used to calibrate the models. The KGE metric is one of the most common metrics used in hydrological modeling. It puts more emphasis on the simulation of flow variability and correlation [73,74]. Compared to the best single model (SIMHYD-MSWEP), the KGE values improved for each averaging method for most schemes. When it comes to NSE and AVE, the improvement in the averaging methods is more obvious (Figure 4, Figure A1, and Figure 9). The NSE metric focuses more on the peak flows and less on the low flows [73]. Therefore, most aspects of the hydrograph simulated by averaging methods are improved compared to the specific hydrograph simulated by one objective function. In addition, previous studies indicated that using the outputs from one model, calibrated with different objective functions as averaging members, can improve the performance of the averaging methods [15]. Therefore, a more comprehensive study is needed to investigate how a large ensemble containing multiple model structures, each with multiple objective functions and driving datasets, impacts the performance of averaging methods.

When compared to the best individual member, even though the simplest equal weights averaging methods (EWA) can improve the simulation performance for more than 40% of the watersheds. However, the performance of different methods is not consistent among climate regions. The AICA, BICA and Granger–Ramanathan group are in the lead group in the arctic region; however, they show poor performance in other climate regions, especially in the equatorial and arid regions. In fact, AICA and BICA tend to put more weight on the best individual member and neglect others [5]. Therefore, the high performance of AICA and BICA in the arctic region could be due to the large differences in performance among 24 members in this region. It is the same for the Granger–Ramanathan group. The Granger–Ramanathan group allows negative weights; therefore, these methods are able to hedge against the use of a bad model [5,54]. The BGA and EWA methods are in the middle level compared with other averaging methods for most regions. In addition, they are more robust than the AICA and BICA methods. The stable performance of these two methods may be due to the fact that these methods distribute the weights fairly. The BMA method is amongst the best methods for most climate regions (except the arctic region). The fact that the performance of BMA would be affected by the poorly performed members may be the reason for the relatively poor performance in the arctic region [36]. In addition, the BMA method is the longest to execute among these averaging methods, because of its iterative nature [5,7]. Therefore, the MMSE averaging method is recommended for its speed of execution, simplicity and stable performance among climate regions.

5. Conclusions

Nine multimodel averaging methods and 11 averaging schemes have been compared, using the simulations of 4 hydrological models driven by 6 precipitation datasets, to find the most suitable multimodel averaging application under different climate regions. The study was conducted over 2277 watersheds around the globe, covering 5 main climatic groups, according to the Köppen–Geiger classification. The following paragraphs outline the results.

The performances of multimodel averaging schemes are closely related to the precipitation used in the hydrological simulation, with a 0.14 difference of the median KGE values between the worst (CPC-COMBINE) and the best (MSWEP-COMBINE) multimodel averaging schemes. Using models driven by different gridded precipitation datasets as ensemble members allows for improving the performance of different averaging methods compared to the multimodel averaging schemes.

Merging multiple members can lead to a significant improvement in hydrological simulations for up to six members. The use of more than 6 members only improves the estimation results slightly, as compared with using all 24 members.

Clear differences in the performance of averaging methods were displayed for different climatic regions. The warm-temperate climatic regions provided the best performance for the averaging methods, with at least 61% of the watersheds having experienced improvements in runoff prediction skills compared to the best single member. Equatorial and snow regions follow closely behind. Moreover, the differences in hydrological model performance among the various averaging methods in arid and arctic regions are more significant than the others.

The best-performing averaging method was different among different climate regions. The MMSE method shows the best performance in most climate regions, except for the arctic region. It is the Granger–Ramanathan average group that outperforms others in the arctic region. In general, the MMSE averaging method shows more advantages over other averaging methods because it is simple to implement, and is always amongst the leading groups.

Author Contributions: Conceptualization, W.Q. and J.C.; methodology, W.Q., J.C. and Y.W.; software, W.Q. and Y.W.; validation, W.Q.; formal analysis, W.Q. and J.C.; investigation, W.Q., J.C. and Y.W.; resources, W.Q. and Y.W.; data curation, W.Q.; writing—original draft preparation, W.Q.; writing—review and editing, J.C. and C.X.; visualization, W.Q.; supervision, J.C.; project administration, J.C.; funding acquisition, J.C. and C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2017YFA0603704; the National Natural Science Foundation of China, grant number 52079093; the Hubei Provincial Natural Science Foundation of China, grant number 2020CFA100; the Overseas Expertise Introduction Project for Discipline Innovation (111 Project), grant number B18037; and the Research Council of Norway (FRINATEK Project), grant number 274310.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

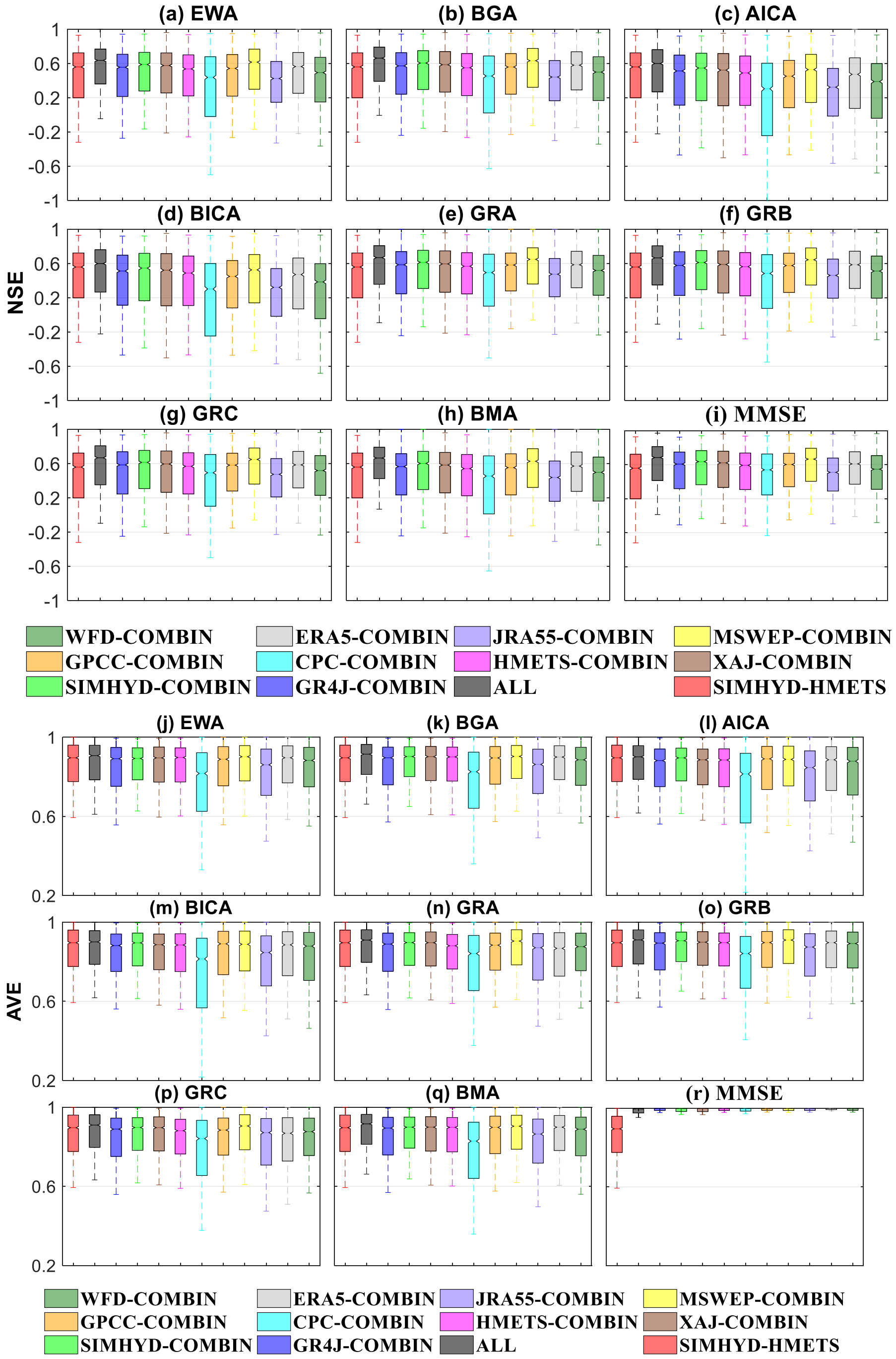

Figure A1. Box plots of NSE (**a**–**i**) and AVE (**j**–**r**) values of different averaging schemes for each averaging method (**a**–**i**).

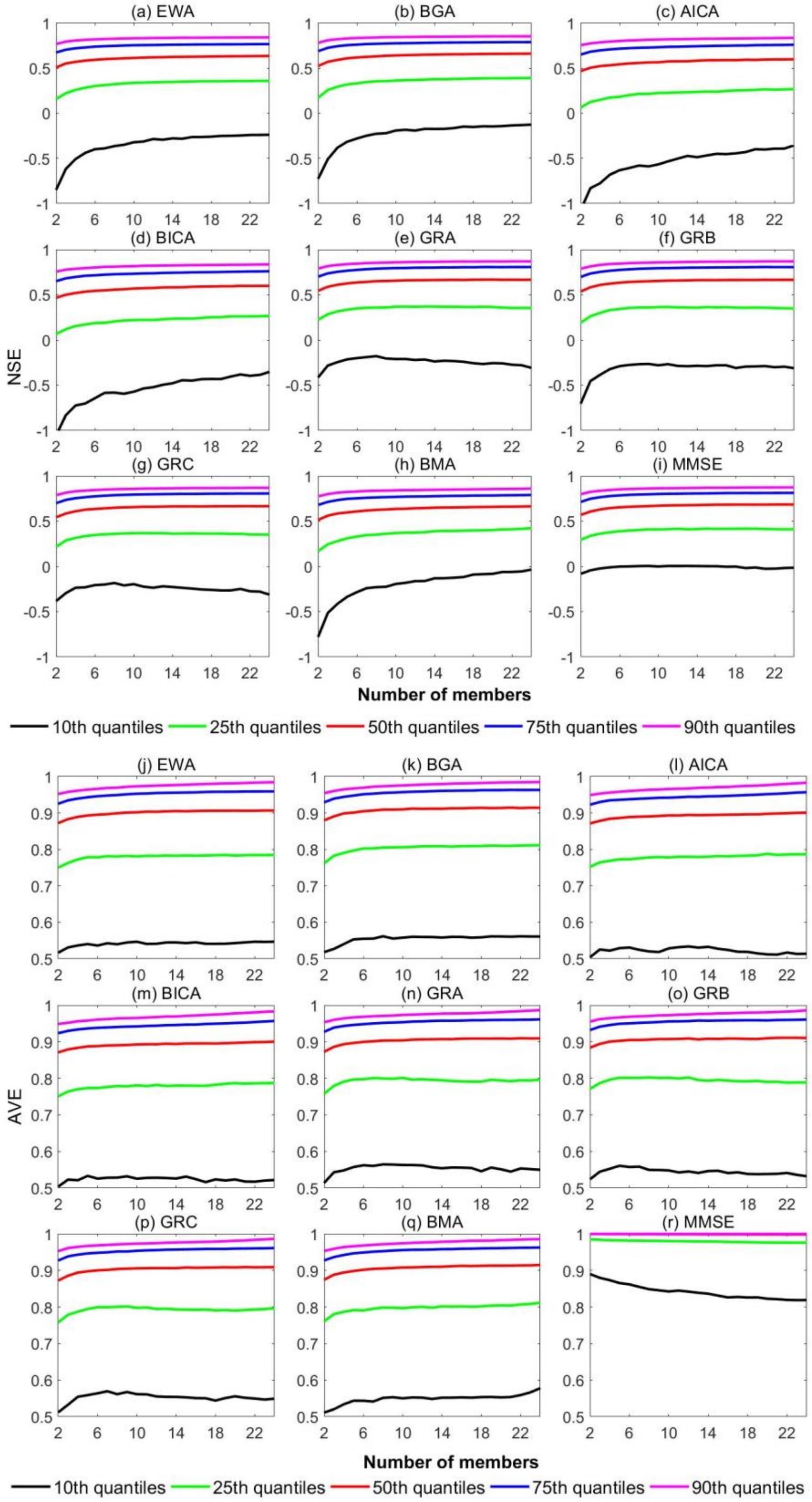

Figure A2. The relationship between the averaging size and the median NSE (**a**–**i**) and AVE (**j**–**r**) values over 2277 watersheds for different averaging methods.

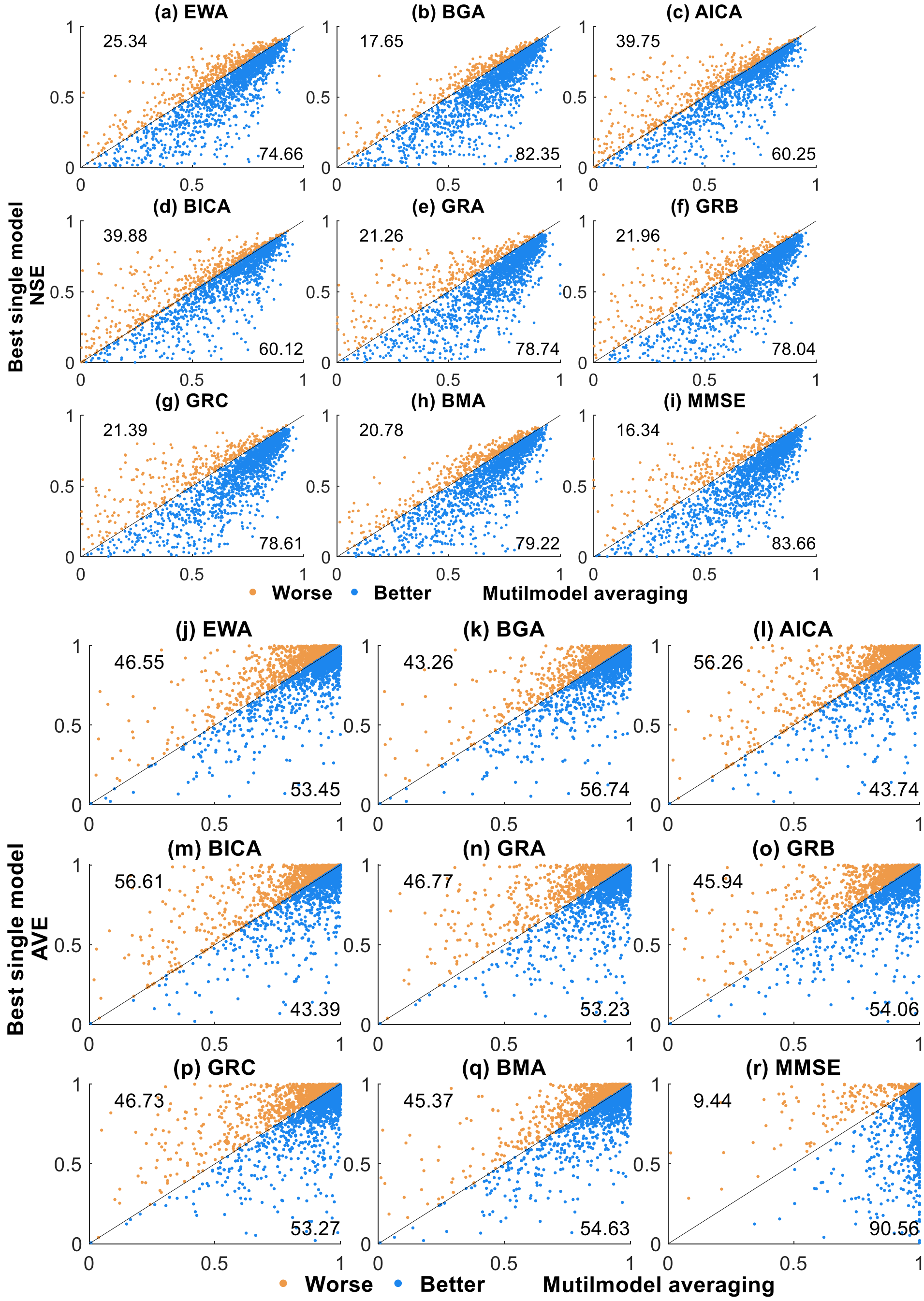

Figure A3. Comparison of 9 multimodel averaging methods and the best overall performing model (SIMHYD-MSWEP) in terms of NSE (**a**–**i**) and AVE (**j**–**r**) values over 2277 watersheds. Model averaging that produces results better than the best member will generate markers under (or to the right of) the 45-degree line. The number in the upper left corner represents the percent of watersheds that, using the multimodel averaging method, perform worse than the best-performing model. The number in the lower right corner represents the percent of watersheds that, using the multimodel averaging method, perform better than the best-performing model.

References

1. Xu, C.-Y.; Tunemar, L.; Chen, Y.D.; Singh, V.P. Evaluation of seasonal and spatial variations of lumped water balance model sensitivity to precipitation data errors. *J. Hydrol.* **2006**, *324*, 80–93. [CrossRef]
2. Xu, C.-Y. Statistical analysis of parameters and residuals of a conceptual water balance model–methodology and case study. *Water Resour. Manag.* **2001**, *15*, 75–92. [CrossRef]

3. Chen, J.; Brissette, F.P.; Poulin, A.; Leconte, R. Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed. *Water Resour. Res.* **2011**, *47*, W12509. [CrossRef]
4. Darbandsari, P.; Coulibaly, P. Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario, Canada. *J. Hydrol. Reg. Stud.* **2020**, *31*, 100730. [CrossRef]
5. Diks, C.G.H.; Vrugt, J.A. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stoch. Environ. Res. Risk. A.* **2010**, *24*, 809–820. [CrossRef]
6. Seifert, D.; Sonnenborg, T.O.; Refsgaard, J.C.; Højberg, A.L.; Troldborg, L. Assessment of hydrological model predictive ability given multiple conceptual geological models. *Water Resour. Res.* **2012**, *48*, W06503. [CrossRef]
7. Arsenault, R.; Gatien, P.; Renaud, B.; Brissette, F.; Martel, J.-L. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. *J. Hydrol.* **2015**, *529*, 754–767. [CrossRef]
8. Arsenault, R.; Essou, G.R.C.; Brissette, F.P. Improving Hydrological Model Simulations with Combined Multi-Input and Multimodel Averaging Frameworks. *J. Hydrol. Eng.* **2017**, *22*, 04016066. [CrossRef]
9. Kumar, A.; Singh, R.; Jena, P.P.; Chatterjee, C.; Mishra, A. Identification of the best multi-model combination for simulating river discharge. *J. Hydrol.* **2015**, *525*, 313–325. [CrossRef]
10. Cavadias, G.; Morin, G. The Combination of Simulated Discharges of Hydrological Models: Application to the WMO Intercomparison of Conceptual Models of Snowmelt Runoff. *Hydrol. Res.* **1986**, *17*, 21–32. [CrossRef]
11. Anctil, F.; Lauzon, N. Generalisation for Neural Networks Through Data Sampling and Training Procedures, With Applications to Streamflow Predictions. *Hydrol. Earth. Syst. Sc* **2004**, *8*, 940–958. [CrossRef]
12. Bowler, N.E.; Arribas, A.; Mylne, K.R. The benefits of multianalysis and poor man's ensembles. *Mon. Weather. Rev.* **2008**, *136*, 4113–4129. [CrossRef]
13. Mylne, K.R.; Evans, R.E.; Clark, R.T. Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Q. J. Roy. Meteor. Soc.* **2002**, *128*, 361–384. [CrossRef]
14. Raftery, A.E.; Gneiting, T.; Balabdaoui, F.; Polakowski, M. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather. Rev.* **2005**, *133*, 1155–1174. [CrossRef]
15. Clark, M.P.; Kavetski, D.; Fenicia, F. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.* **2011**, *47*, W09301. [CrossRef]
16. Nijssen, B.; O'Donnell, G.M.; Lettenmaier, D.P.; Lohmann, D.; Wood, E.F. Predicting the Discharge of Global Rivers. *J. Clim.* **2001**, *14*, 3307–3323. [CrossRef]
17. Sawunyama, T.; Hughes, D.A. Application of satellite-derived rainfall estimates to extend water resource simulation modelling in South Africa. *Water SA* **2018**, *34*, 1–10. [CrossRef]
18. Tuo, Y.; Duan, Z.; Disse, M.; Chiogna, G. Evaluation of precipitation input for SWAT modeling in Alpine catchment: A case study in the Adige river basin (Italy). *Sci. Total Environ.* **2016**, *573*, 66–82. [CrossRef] [PubMed]
19. Douglas-Mankin, K.R.; Srinivasan, R.; Arnold, J.G. Soil and Water Assessment Tool (SWAT) Model: Current Developments and Applications. *Trans. Asabe* **2010**, *53*, 1423–1431. [CrossRef]
20. Senent-Aparicio, J.; López-Ballesteros, A.; Pérez-Sánchez, J.; Segura-Méndez, F.; Pulido-Velazquez, D. Using Multiple Monthly Water Balance Models to Evaluate Gridded Precipitation Products over Peninsular Spain. *Remote Sens.* **2018**, *10*, 922. [CrossRef]
21. Zhang, D.; Liu, X.; Bai, P.; Li, X.-H. Suitability of Satellite-Based Precipitation Products for Water Balance Simulations Using Multiple Observations in a Humid Catchment. *Remote Sens.* **2019**, *11*, 151. [CrossRef]
22. Tang, X.; Zhang, J.; Gao, C.; Ruben, G.; Wang, G. Assessing the Uncertainties of Four Precipitation Products for Swat Modeling in Mekong River Basin. *Remote Sens.* **2019**, *11*, 304. [CrossRef]
23. Ahmed, K.; Shahid, S.; Ali, R.O.; Bin Harun, S.; Wang, X.-j. Evaluation of the performance of gridded precipitation products over Balochistan Province, Pakistan. *Desalination Water Treat.* **2017**, *79*, 73–86. [CrossRef]
24. Gampe, D.; Ludwig, R. Evaluation of Gridded Precipitation Data Products for Hydrological Applications in Complex Topography. *Hydrology* **2017**, *4*, 53. [CrossRef]
25. Bai, L.; Wen, Y.; Shi, C.; Yang, Y.; Zhang, F.; Wu, J.; Gu, J.; Pan, Y.; Sun, S.; Meng, J. Which Precipitation Product Works Best in the Qinghai-Tibet Plateau, Multi-Source Blended Data, Global/Regional Reanalysis Data, or Satellite Retrieved Precipitation Data? *Remote Sens.* **2020**, *12*, 683. [CrossRef]
26. Chen, H.; Yong, B.; Shen, Y.; Liu, J.; Hong, Y.; Zhang, J. Comparison analysis of six purely satellite-derived global precipitation estimates. *J. Hydrol.* **2020**, *581*, 124376. [CrossRef]
27. Schneider, U.; Becker, A.; Finger, P.; Meyer-Christoffer, A.; Ziese, M.; Rudolf, B. GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Appl. Clim.* **2014**, *115*, 15–40. [CrossRef]
28. Weedon, G.P.; Balsamo, G.; Bellouin, N.; Gomes, S.; Best, M.J.; Viterbo, P. The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resour. Res.* **2014**, *50*, 7505–7514. [CrossRef]
29. Abramowitz, G.; Hobeichi, S.; Contractor, S.; Evans, J. Evaluating Precipitation Datasets Using Surface Water and Energy Budget Closure. *J. Hydrometeorol.* **2020**, *21*, 989–1009.
30. Sharifi, E.; Eitzinger, J.; Dorigo, W. Performance of the State-of-the-Art Gridded Precipitation Products over Mountainous Terrain: A Regional Study over Austria. *Remote Sens.* **2019**, *11*, 2018. [CrossRef]
31. Wang, S.; Liu, J.; Wang, J.; Qiao, X.; Zhang, J. Evaluation of GPM IMERG V05B and TRMM 3B42V7 Precipitation Products over High Mountainous Tributaries in Lhasa with Dense Rain Gauges. *Remote Sens.* **2019**, *11*, 2080. [CrossRef]

32. Beck, H.E.; Pan, M.; Roy, T.; Weedon, G.P.; Pappenberger, F.; van Dijk, A.I.J.M.; Huffman, G.J.; Adler, R.F.; Wood, E.F. Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 207–224. [CrossRef]
33. Beck, H.E.; Vergopolan, N.; Pan, M.; Levizzani, V.; van Dijk, A.I.J.M.; Weedon, G.P.; Brocca, L.; Pappenberger, F.; Huffman, G.J.; Wood, E.F. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 6201–6217. [CrossRef]
34. Chen, J.; Li, Z.; Li, L.; Wang, J.; Qi, W.; Xu, C.-Y.; Kim, J.-S. Evaluation of Multi-Satellite Precipitation Datasets and Their Error Propagation in Hydrological Modeling in a Monsoon-Prone Region. *Remote Sens.* **2020**, *12*, 3550. [CrossRef]
35. Sun, R.; Yuan, H.; Yang, Y. Using multiple satellite-gauge merged precipitation products ensemble for hydrologic uncertainty analysis over the Huaihe River basin. *J. Hydrol.* **2018**, *566*, 406–420. [CrossRef]
36. Najafi, M.R.; Moradkhani, H. Multi-model ensemble analysis of runoff extremes for climate change impact assessments. *J. Hydrol.* **2015**, *525*, 352–361. [CrossRef]
37. Wang, H.M.; Chen, J.; Xu, C.Y.; Zhang, J.; Chen, H. A Framework to Quantify the Uncertainty Contribution of GCMs Over Multiple Sources in Hydrological Impacts of Climate Change. *Earth's Future* **2020**, *8*, e2020EF001602. [CrossRef]
38. Chen, M.; Shi, W.; Xie, P.; Silva, V.B.S.; Kousky, V.E.; Wayne Higgins, R.; Janowiak, J.E. Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.* **2008**, *113*, D04110. [CrossRef]
39. Kobayashi, S.; Ota, Y.; Harada, Y.; Ebita, A.; Moriya, M.; Onoda, H.; Onogi, K.; Kamahori, H.; Kobayashi, C.; Endo, H. The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteorol. Soc. Jpn.* **2015**, *93*, 5–48. [CrossRef]
40. Martens, B.; Miralles, D.G.; Lievens, H.; van der Schalie, R.; de Jeu, R.A.M.; Fernández-Prieto, D.; Beck, H.E.; Dorigo, W.A.; Verhoest, N.E.C. GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geosci. Model Dev.* **2017**, *10*, 1903–1925. [CrossRef]
41. Arsenault, R.; Bazile, R.; Ouellet Dallaire, C.; Brissette, F. CANOPEX: A Canadian hydrometeorological watershed database. *Hydrol. Process.* **2016**, *30*, 2734–2736. [CrossRef]
42. Perrin, C.; Michel, C.; Andréassian, V. Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* **2003**, *279*, 275–289. [CrossRef]
43. Chiew, F.H. Lumped Conceptual Rainfall-Runoff Models and Simple Water Balance Methods: Overview and Applications in Ungauged and Data Limited Regions. *Geogr. Compass* **2010**, *4*, 206–225. [CrossRef]
44. Chiew, F.H.; Peel, M.C.; Western, A.W. Application and testing of the simple rainfall-runoff model SIMHYD. In *Mathematical Models of Small Watershed Hydrology and Applications*; Singh, V.P., Frevert, D.K., Eds.; Water Resources Publications: California City, CA, USA, 2002; pp. 335–367.
45. Zhao, R.; Liu, X. *The Xinanjiang Model, Computer Models of Watershed Hydrology*; Singh, V.P., Ed.; Water Resources Publications: California City, CA, USA, 1995; pp. 215–232.
46. Zhao, R.-J. The Xinanjiang model applied in China. *J. Hydrol.* **1992**, *135*, 371–381.
47. Martel, J.-L.; Demeester, K.; Brissette, F.P.; Arsenault, R.; Poulin, A. HMET: A simple and efficient hydrology model for teaching hydrological modelling, flow forecasting and climate change impacts. *Int. J. Eng. Educ.* **2017**, *33*, 1307–1316.
48. Yang, X.; Magnusson, J.; Huang, S.; Beldring, S.; Xu, C.-Y. Dependence of regionalization methods on the complexity of hydrological models in multiple climatic regions. *J. Hydrol.* **2020**, *582*, 124357. [CrossRef]
49. Yin, J.; Guo, S.; Gu, L.; He, S.; Ba, H.; Tian, J.; Li, Q.; Chen, J. Projected changes of bivariate flood quantiles and estimation uncertainty based on multi-model ensembles over China. *J. Hydrol.* **2020**, *585*, 124760. [CrossRef]
50. Duan, Q.; Gupta, V.K.; Sorooshian, S. Shuffled complex evolution approach for effective and efficient global minimization. *J. Optim. Theory. App* **1993**, *76*, 501–521. [CrossRef]
51. Duan, Q.; Sorooshian, S.; Gupta, V. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.* **1992**, *28*, 1015–1031. [CrossRef]
52. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [CrossRef]
53. Muhammad, A.; Stadnyk, T.; Unduche, F.; Coulibaly, P. Multi-Model Approaches for Improving Seasonal Ensemble Streamflow Prediction Scheme with Various Statistical Post-Processing Techniques in the Canadian Prairie Region. *Water* **2018**, *10*, 1604. [CrossRef]
54. Arsenault, R.; Brissette, F. Multi-model averaging for continuous streamflow prediction in ungauged basins. *Hydrol. Sci. J.* **2016**, *61*, 2443–2454. [CrossRef]
55. Zhang, J.; Chen, J.; Li, X.; Chen, H.; Xie, P.; Li, W. Combining Postprocessed Ensemble Weather Forecasts and Multiple Hydrological Models for Ensemble Streamflow Predictions. *J. Hydrol. Eng.* **2020**, *25*, 04019060. [CrossRef]
56. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
57. Bates, J.M.; Granger, C.W. The combination of forecasts. *J. Oper. Res. Soc.* **1969**, *20*, 451–468. [CrossRef]
58. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
59. Granger, C.W.; Ramanathan, R. Improved methods of combining forecasts. *J. Forecast.* **1984**, *3*, 197–204. [CrossRef]
60. Neuman, S.P. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stoch. Environ. Res. Risk Assess.* **2003**, *17*, 291–305. [CrossRef]
61. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Des Moines, IA, USA, 1999.

62. Sivapragasam, C.; Liong, S.-Y.; Pasha, M.J. Rainfall and runoff forecasting with SSA–SVM approach. *J. Hydroinform.* **2001**, *3*, 141–152. [CrossRef]
63. Vis, M.; Knight, R.; Pool, S.; Wolfe, W.; Seibert, J. Model Calibration Criteria for Estimating Ecological Flow Characteristics. *Water* **2015**, *7*, 2358–2381. [CrossRef]
64. Widén-Nilsson, E.; Halldin, S.; Xu, C.-Y. Global water-balance modelling with WASMOD-M: Parameter estimation and regionalisation. *J. Hydrol.* **2007**, *340*, 105–118. [CrossRef]
65. Beck, H.E.; van Dijk, A.I.J.M.; de Roo, A.; Miralles, D.G.; McVicar, T.R.; Schellekens, J.; Bruijnzeel, L.A. Global-scale regionalization of hydrologic model parameters. *Water Resour. Res.* **2016**, *52*, 3599–3622. [CrossRef]
66. Ghebrehiwot, A.A.; Kozlov, D.V. Hydrological modelling for ungauged basins of arid and semi-arid regions: Review. *Vestn. Mgsu.* **2019**, *8*, 1023–1036. [CrossRef]
67. Krishnamurti, T.N.; Kishtawal, C.M.; LaRow, T.E.; Bachiochi, D.R.; Zhang, Z.; Williford, C.E.; Gadgil, S.; Surendran, S. Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble. *Science* **1999**, *285*, 1548–1550. [CrossRef]
68. Krishnamurti, T.; Kishtawal, C.; Zhang, Z.; Larow, T.; Bachiochi, D.; Williford, E.; Gadgil, S.; Surendran, S. Multimodel Ensemble Forecasts for Weather and Seasonal Climate. *J. Clim.* **2000**, *13*, 4196–4216. [CrossRef]
69. Ajami, N.K.; Duan, Q.; Gao, X.; Sorooshian, S. Multimodel Combination Techniques for Analysis of Hydrological Simulations: Application to Distributed Model Intercomparison Project Results. *J. Hydrometeorol.* **2006**, *7*, 755–768. [CrossRef]
70. Awol, F.S.; Coulibaly, P.; Tsanis, I. Identification of Combined Hydrological Models and Numerical Weather Predictions for Enhanced Flood Forecasting in a Semiurban Watershed. *J. Hydrol. Eng.* **2021**, *26*, 04020057. [CrossRef]
71. Beven, K. A manifesto for the equifinality thesis. *J. Hydrol.* **2006**, *320*, 18–36. [CrossRef]
72. Samaniego, L.; Kumar, R.; Attinger, S. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resour. Res.* **2010**, *46*, W05523. [CrossRef]
73. Knoben, W.J.M.; Freer, J.E.; Woods, R.A. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrol. Earth. Syst. Sci.* **2019**, *23*, 4323–4331. [CrossRef]
74. Santos, L.; Thirel, G.; Perrin, C.J.H.; Sciences, E.S. Technical note: Pitfalls in using log-transformed flows within the KGE criterion. *Hydrol. Earth. Syst. Sci.* **2018**, *22*, 4583–4591. [CrossRef]