water

MDPI

Article

Comparing Bayesian Model Averaging and Reliability Ensemble Averaging in Post-Processing Runoff Projections under Climate Change

Kai Duan [1,2,*], Xiaola Wang [1], Bingjun Liu [1], Tongtiegang Zhao [1] and Xiaohong Chen [1]

[1] School of Civil Engineering, Sun Yat-Sen University, Guangzhou 510275, China; wangxla@mail2.sysu.edu.cn (X.W.); liubj@mail.sysu.edu.cn (B.L.); zhaottg@mail.sysu.edu.cn (T.Z.); eescxh@mail.sysu.edu.cn (X.C.)

[2] Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519082, China

* Correspondence: duank6@mail.sysu.edu.cn

Abstract: This study investigated the strength and limitations of two widely used multi-model averaging frameworks—Bayesian model averaging (BMA) and reliability ensemble averaging (REA), in post-processing runoff projections derived from coupled hydrological models and climate downscaling models. The performance and weight distributions of five model ensembles were thoroughly compared, including simple equal-weight averaging, BMA, and REAs optimizing mean (REA-mean), maximum (REA-max), and minimum (REA-min) monthly runoff. The results suggest that REA and BMA both can synthesize individual models' diverse skills with comparable reliability, despite of their different averaging strategies and assumptions. While BMA weighs candidate models by their predictive skills in the baseline period, REA also forces the model ensembles to approximate a convergent projection towards the long-term future. The type of incorporation of the uncertain future climate in REA weighting criteria, as well as the differences in parameter estimation (i.e., the expectation maximization (EM) algorithm in BMA and the Markov Chain Monte Carlo sampling method in REA), tend to cause larger uncertainty ranges in the weight distributions of REA ensembles. Moreover, our results show that different averaging objectives could cause much larger discrepancy than that induced by different weighting criteria or parameter estimation algorithms. Among the three REA ensembles, REA-max most resembled BMA because the EM algorithm of BMA converges to the minimum aggregated error, and thus emphasize the simulation of high flows. REA-min achieved better performance in terms of inter-annual temporal pattern, yet at the cost of compromising accuracy in capturing mean behaviors. Caution should be taken to strike a balance among runoff features of interest.

Keywords: runoff projection; probabilistic multi-model ensemble; Bayesian model averaging; reliability ensemble averaging; climate change

Citation: Duan, K.; Wang, X.; Liu, B.; Zhao, T.; Chen, X. Comparing Bayesian Model Averaging and Reliability Ensemble Averaging in Post-Processing Runoff Projections under Climate Change. *Water* **2021**, *13*, 2124. https://doi.org/10.3390/w13152124

Academic Editor: George Arhonditsis

Received: 12 July 2021
Accepted: 30 July 2021
Published: 1 August 2021

1. Introduction

Climate change is significantly altering runoff characteristics and thus affects water availability for both ecosystem and humans [1–5]. Due to the gap between spatial scales in global climate models (GCMs) and regional hydrological simulations, climate under future scenarios are usually reanalyzed with downscaling tools before fed into hydrological models for runoff projections. Various downscaling techniques and hydrological models have been developed and applied for understanding and quantifying climate change impacts on runoff [6,7]. Downscaling methods fall into two categories that focus on atmospheric physics and empirical statistics respectively, known as dynamical downscaling and statistical downscaling, between which statistical downscaling methods have been widely adopted by hydrologists for their accessibility, efficiency, and flexibility [8–10]. Hydrological models range from simple water balance models (Mpelasoka et al., 2008), large

scale energy-water balance equations [11], and conceptual rainfall-runoff models [12,13] to more complex landscape distributed models [14]. Previous studies suggest that no single model is perfect, and it is necessary to compare the strength and weakness of diverse downscaling methods and hydrological models before applying them on specific circumstances [9,15–17].

Multi-model-ensemble approaches have been employed to contain model biases and evaluate uncertainties. Multi-model-ensemble strategies can exploit the diversity of skillful predictions and enhance the predictive capabilities from a perspective of either point forecast [18,19] or density forecast [20,21]. Particularly, probabilistic multi-model-ensemble (PMME) approaches that synthesize outputs from different GCMs [22–24], regional climate models (RCMs) [25,26], statistical downscaling models [27,28], or hydrological models [20] are receiving a surge of attention in the recent decades. The key of PMME is to measure the weights of each model and to produce proper probability density functions (pdfs) of the variables of interest, which is usually achieved through Bayesian approaches [29]. The basic idea is to obtain the posterior distributions of parameters of interest from pre-specified prior probability distributions and likelihood functions based on the Bayes' theorem.

It has been well established that PMME methods can improve hydrological modeling across different climates or terrestrial environments. However, compared to simulations aiming at reproducing historical hydrological variations, some particular issues are involved in future-oriented runoff projections. For example, the performance of hydrological models can be largely affected by the adaptability of downscaling techniques in regional simulations [8]. PMME approaches such as Bayesian Model Averaging (BMA) have been used to generate predictive distributions from either different climate models [30] or hydrological models [20,31]. However, rarely have there been experiments of probabilistically combining downscaling and hydrological modeling approaches based on different statistical philosophy. The evaluation of hydrological models can be largely distorted by uncertainties in climate downscaling, which can be much larger than uncertainties in retrospective hydrological modeling [32]. Besides, multi-model-ensemble hydrological modeling usually uses weighting criteria that measures accuracy in capturing magnitudes and timings of historical runoff [18,21]. However, the capability in reproducing historical events is not the only expected merit in hydro-climatic simulations. Convergence in the multi-model projections of future changes are also considered critical for model selection in PMME method such as the Reliability Ensemble Averaging (REA) [6,33]. It is not clear how such different weighting strategies would affect the overall performance of PMME modeling and the relative contributions of individual models.

This study aims to investigate the merits of BMA and REA in post-processing runoff projections derived from multiple combinations of downscaling models and hydrological models. An integrated evaluation framework was established for exploring the usefulness of individual climate downscaling models, hydrological models, and their probabilistic ensembles in regional runoff projection. We used two statistical downscaling models, including a regression-based approach (Statistical Down Scaling Model, SDSM) [34] and a weather generator (Long Ashton Research Station Weather Generator, LARS-WG) [35], and three hydrological models developed in Sweden (i.e., the Hydrologiska Byråns Vattenbalansavdelning model, HBV-light) [36,37], Australia (i.e., the SIMHYD model) [9,38], and China (i.e., the Xinanjiang, XAJ model) [39] respectively, for modeling experiments. Combinations of these downscaling models and hydrological models were synthesized through the BMA and REA schemes to explore the effectiveness of probabilistic ensembles.

2. Methods

2.1. Data

The upper Huai River basin above Bengbu station in eastern China (121,330 km^2) was selected as a case study for model testing (Figure 1). The Huai River basin (30°55′–36°36′ N, 111°55′–121°25′ E) is located between the Yangtze River basin and the Yellow River basin, with subtropical and temperate monsoon climates in the south and north of

the Huai River, respectively [40]. The mean annual precipitation reaches 883 mm and most rainfall (50–80%) occurs between June and September. The mean annual temperature and evaporation are 11–16 °C and 900–1500 mm.

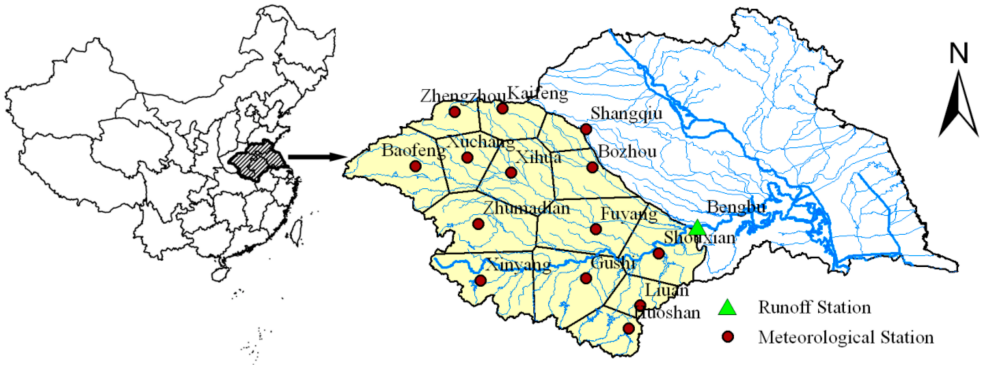

Figure 1. Distribution of meteorological stations, runoff station and Thiessen polygons in the upper Huai River basin. Figure adapted from Duan et al. [32].

Datasets used in this study include: (1) daily precipitation, temperature, and evaporation at 14 meteorological stations from January 1961 to December 2000 (China Meteorological Administration, http://data.cma.cn/data, accessed on 19 July 2020); (2) monthly runoff at the Bengbu station (Anhui Water Bureau, http://slt.ah.gov.cn/, accessed on 19 July 2020) from January 1961 to December 2000; (3) large-scale atmospheric predictors (i.e., air pressure, velocity, vorticity, wind, humidity) obtained from the NCEP/NCAR reanalysis data [41] for the time periods of 1961–2000 and 2060–2099; and (4) daily precipitation, temperature, and large-scale atmospheric predictors projected by the Met Office climate prediction model (HadCM3) model in the historical period 1961–2000 and future period 2060–2099 under the A2 scenario of Special Report on Emissions Scenarios (https://www.cics.uvic.ca/scenarios/index.cgi, accessed on 19 July 2020).

2.2. Statistical Downscaling of Regional Climate

SDSM is a hybrid downscaling technique based on multiple linear and exponential regression equations [34]. A major improvement distinguishing SDSM from traditional regression methods is that a precipitation occurrence module similar to stochastic weather generators was nested in the precipitation downscaling process. A linear regression equation with a uniformly distributed random number was used to represent the probability of precipitation occurrence, and an exponential regression model was used to simulate the precipitation magnitude. The downscaling using SDSM was performed with the following steps: (1) A total of 26 large-scale atmospheric variables were obtained as predictor candidates, including air pressure, airflow strength, velocity, vorticity, wind direction, divergence, and humidity at the heights of near surface, 500 hPa, and 850 hPa. (2) Data of 390 (26×15) candidates at the 15 grids (3×5, $3.75° \times 2.5°$) overlaying each target meteorological station were collected as initial predictors, and then the 30 most correlated candidates were extracted based on Spearman's rank correlation coefficient. (3) The 30 candidates were further screened by a principal component analysis, and the numbers of predictors were reduced to less than 10 with the cumulative explained variance ratio reaching 90%. (4) The selected predictors were used to establish their regression relationships with observed precipitation and temperature at each station.

LARS-WG is "serial" stochastic weather generator that uses semi-empirical distributions with pre-defined intervals to simulate the temporal distributions of predictands [35]. LARS-WG models for daily precipitation and temperature were specifically established for each site and each month in this study. We first separated the historical daily series into wet and dry days and calculated the probability of precipitation occurrence in each month. In wet days, 23 intervals were used to approximate the distributions of daily precipitation

magnitudes with particular emphasis on extreme events, including 19 intervals dividing 2–98% evenly, two intervals close to 0 (0–1% and 1–2%), and two intervals close to 1 (98–99% and 99–100%). The time series of wet and dry days were also used to establish temperature downscaling models separately. For generations of future climate, the parameters were first corrected using the delta change method based on changes in mean values projected by the GCM. More details of data processing and model validation of SDSM and LARS-WG can be found in our earlier studies [19,42].

2.3. Hydrological Modeling

We divided the study area into 14 Thiessen polygons with one meteorological station representing each polygon, and the hydrological processes from precipitation, evapotranspiration, and infiltration to runoff were simulated for each polygon separately at daily scale. The runoffs generated from each polygon were summed up as the total runoff to proximate the streamflow observed at the Bengbu station. All of the three models have showed good performance in the study area with the Nash-Sutcliff coefficient exceeding 0.8, and thus are suitable for exploring the uncertainties derived from different hydrological model structures [32]. Statistically downscaled climate data were used to drive the calibrated hydrological models to simulate runoff under the future scenario. Six combinations of the statistical downscaling models and hydrological models were established for model inter-comparison and probabilistic runoff projection, including SDSM and HBV (SD-HBV), SDSM and SIMHYD (SD-SIMHYD), SDSM and XAJ (SD-XAJ), LARS-WG and HBV (LW-HBV), LARS-WG and SIMHYD (LW-SIMHYD), and LARS-WG and XAJ (LW-XAJ).

2.4. Probabilistic Multi-Model-Ensemble Runoff Projection

We used two PMME methods with different underlying assumptions, i.e., BMA and REA, to synthesize the runoff projections from the baseline period 1961–2000 to the future period 2060–2099 derived from the six model combinations. Main features of the climate downscaling models, hydrological models, and model ensemble methods are summarized in Table 1.

Table 1. Comparison of the methods used in this study.

Procedure	Model	Features
Climate downscaling	SDSM [34]	Data required—Historical precipitation, temperature, and large-scale predictors Model structure—Precipitation occurrence and magnitude fitted by multiple regressions
	LARS-WG [35]	Data required—Historical precipitation and temperature Model structure—Precipitation occurrence and magnitude fitted by semi-empirical distributions; extreme events represented by specifying precipitation intervals
Hydrological modeling	HBV [36]	Water storage components—Upper and lower groundwater storage Runoff generation regime—Saturation excess Runoff components—Peak, intermediate, and baseflow
	SIMHYD [38]	Water storage components—Soil moisture storage and groundwater storage Runoff generation regime—Saturation excess and infiltration excess Runoff components—Infiltration excess runoff, interflow, and baseflow
	XAJ [39]	Water storage components—Upper, lower, and deep tension water storage, free water storage Runoff generation regime—Saturation excess Runoff components—surface runoff, interflow, groundwater flow
Model ensemble	BMA [20,30]	Uncertainty interpretation—Bayesian probability Weighting criteria—Models' relative contributions to predictive skill Parameter estimation—Expectation maximization algorithm
	REA [33,43]	Uncertainty interpretation—Bayesian probability Weighting criteria—Model bias and model convergence Parameter estimation—Gibbs-Metropolis algorithm

2.4.1. Bayesian Model Averaging

BMA was proposed as a post-processing ensemble approach to correct the under-dispersion of spread-error correlation in probabilistic weather forecasting [30]. It was assumed that more reliable prediction on the point estimate can be acquired by weighing and combining several ensemble members of interest according to their bias against the observations. The input training datasets of BMA include prediction from each individual models and the observed series. The output is the calibrated set of weights reflecting the relative contributions to predictive skill of each ensemble member and the corresponding variance, which are used to construct a probability density function (pdf) accompanied by the weights and produce ensemble predictions. In a situation $\{M_1, M_2, \ldots, M_k\}$ are the candidate predictions, the pdf of the BMA prediction can be obtained based on the law of total probability:

$$p(y|M_1, M_2, \ldots, M_k) = \sum_{k=1}^{K} p(M_k|D) \cdot p_k(y|M_k, D) \quad (1)$$

where D is the training dataset; $p(M_k|D)$ is the posterior probability of M_k, which equals the weight w_k, and $\sum_{k=1}^{K} w_k = 1$. The posterior mean and variance of the ensemble prediction are as follows:

$$E[y|M_1, M_2, \ldots, M_k] = \sum_{k=1}^{K} w_k \cdot M_k \quad (2)$$

$$Var[y|M_1, M_2, \ldots, M_k] = \sum_{k=1}^{K} w_k \cdot \left(M_k - \sum_{i=1}^{K} w_i \cdot M_i \right) + \sum_{k=1}^{K} w_k \cdot \sigma_k^2 \quad (3)$$

where σ_k^2 is the variance of a single model. The BMA variance is interpreted as the sum of variances contributed by between-model uncertainty and within-model uncertainty.

Both simulated and observed runoff data are first transformed to the Gaussian distribution using the Box-Cox method prior to the BMA procedure, so that the conditional probability distribution of each member $p_k(y|M_k, D)$ can be treated as Gaussian. The BMA weights and variance are then estimated using the expectation maximization (EM) algorithm, which is iterative and converges to a local maximum likelihood. The detailed description of the BMA method and EM algorithm can be found in Raftery et al. [30] and Duan et al. [20].

2.4.2. Reliability Ensemble Averaging

We used a Bayesian-based REA method proposed by Tebaldi et al. [43,44]. A fundamental difference between the Tebaldi method and the original REA method [33] is that the weights are treated as random quantities to account for uncertainty in the estimation. The method has been successfully used to evaluate the probability distributions of future changes in temperature [45] and precipitation [46] projected by different climate models. The underlying assumptions of this method are that the projections have a symmetric distribution centering the "true value", but with an individual variability to be regarded as a measure of how well each model approximates the climate response to the given set of natural and anthropogenic forcings [44]. We assume that such hypothesis also applies to runoff projections, because previous model inter-comparison studies on both climate models [47] and hydrological models [48] have demonstrated that the mean of multi-model ensemble can achieve better overall validation properties than individual members.

(1) Likelihoods

The inputs include three datasets: X_0, observed runoff in the baseline period; X_i, runoff simulated by the ith ensemble member in the baseline period, $i = 1, 2, 3, \ldots, 6$; Y_i,

runoff simulated by the ith ensemble member in the future period. Gaussian distributions are assumed for X_0, X_i and Y_i, and then the likelihoods are specified as:

$$X_0 \sim N\left(\mu, {\lambda_0}^{-1}\right) \tag{4}$$

$$X_i \sim N\left(\mu, {\lambda_i}^{-1}\right) \tag{5}$$

$$Y_i \sim N\left(\nu + \beta(X_i - \mu), (\theta\lambda_i)^{-1}\right) \tag{6}$$

where $N(\mu, \lambda^{-1})$ indicates a Gaussian distribution with mean μ and variance $1/\lambda$. λ_0 is the regional natural variability of observed series, using estimates from Giorgi and Mearns [33,49]; μ and ν represent the true values of runoff characteristics in the baseline and future periods, and $(\nu - \mu)$ can be used to represent the expectation of runoff change in the future. The reciprocal of the variance λ_i is referred to as the precision of the distribution of X_i. β is used to introduce the correlation between baseline and future runoff responses. The random variable θ acts as the inflation-deflation factor in the precision of ensemble member when comparing simulations of the baseline to the future.

(2) Prior distributions

The parameters μ, ν, β, λ_i and θ are assigned to uninformative priors. The true values of μ, ν and β are assumed to have uniform prior densities on the real line. λ_i and θ follow two-parameter Gamma distributions $\lambda_i \sim Ga(a, b)$ and $\theta \sim Ga(c, d)$, where $a = b = c = d = 0.001$, that translate into distributions with mean 1 and variance 1000. Such unity mean and large variance over the positive real line is used to create extremely diffuse distributions that have the required uninformative quality.

(3) Posterior distributions

Through Bayes' theorem, the joint posterior distribution for the parameters μ, ν, β, λ_i, θ resulting from the likelihoods and prior distributions is given by:

$$\prod_{i=1}^{6} \lambda_i^{a-1} e^{-b\lambda_i} \times \lambda_i \theta^{1/2} \exp\left\{-\frac{\lambda_i}{2}\left[(X_i - \mu)^2 + \theta(Y_i - \nu)^2\right]\right\} \times \theta^{c-1} e^{-d\theta} \times \exp\left\{-\frac{\lambda_0}{2}(X_0 - \mu)^2\right\} \tag{7}$$

Then, the forms of marginal posterior distributions for each individual parameter can be specified as functions of other parameters. The conditional posterior distribution of λ_i is a Gamma function with mean:

$$\lambda_i = \frac{a + 1}{b + \left\{(X_i - \mu)^2 + \theta[Y_i - \nu - \beta(X_i - \mu)]^2\right\}/2} \tag{8}$$

Similarly derived, the conditional distribution of μ and ν are Gaussian distributions with mean as:

$$\widetilde{\mu} = \frac{\Sigma\lambda_i X_i - \theta\beta\Sigma\lambda_i(Y_i - v - \beta X_i) + \lambda_0 X_0}{\Sigma\lambda_i + \theta\beta^2\Sigma\lambda_i + \lambda_0} \tag{9}$$

$$\widetilde{\nu} = \frac{\Sigma\lambda_i[Y_i - \beta(X_i - \mu)]}{\Sigma\lambda_i} \tag{10}$$

and variance as:

$$\sigma_\mu = \frac{1}{\Sigma\lambda_i + \theta\beta^2\Sigma\lambda_i + \lambda_0} \tag{11}$$

$$\sigma_v = \frac{1}{\theta\Sigma\lambda_i} \tag{12}$$

(4) Parameter estimation

The joint posterior distribution is too complex to be computed by an analytical solution. Therefore, the Markov Chain Monte Carlo (MCMC) method is used to estimate the posterior distributions statistically by generating a large number of random samples. In this study, 50,000 groups of parameters were stochastically generated through the Gibbs-Metropolis algorithm (http://www.image.ucar.edu/~nychka/REA, accessed on 19 July 2020) [50].

The first 25,000 samples were discarded as the burn-in period, and the other 25,000 were refined into 5000 samples by saving every 50th to avoid correlation between the successive values [46].

2.4.3. Weighting Strategies

The BMA ensemble simulation is the average of individual simulations weighted by the likelihood that each individual model is correct given the observations. The weight of each BMA member (w_k) is measured by the agreement between the model simulations and the observations, which is calculated as the summation of bias over time, as:

$$\sigma_k^2 = \frac{1}{K}\sum_{t=1}^{T}\frac{\left(\sum_{k=1}^{K}(y_t - f_{k,t})^2\right)}{T} \quad (13)$$

where T is the total number of data points in the training period, K is the number of ensemble members, y_t and $f_{k,t}$ are the observed and simulated runoff by the kth model. The weights are optimized through the EM algorithm towards the minimization of total bias.

Model weighting of the REA method is conducted based on two criteria: (1) Model performance in the baseline period that measures the models' relative skills in reproducing historical runoff, and (2) model convergence that measures agreement among ensemble members. The first criterion is similar to BMA that a lower bias indicates a higher model reliability, and the bias is defined as the difference between the simulated and observed values. The second criterion assumes that the convergence of simulations by different models for a given forcing scenario indicates reliability of robust signals. Higher convergence implies that the predictions are less sensitive to the differences among models. In other words, the REA weighting approach penalizes models that do not predict the same runoff responses to climate change. We used the extracted 5000 samples of the precision parameter λ_i (Equation (8)) to evaluate the usefulness of each ensemble member and the ranges of inherent uncertainty. The precision parameters were converted to relative weights (%) as:

$$W_i = 100 \times \lambda_i / \sum_{i=1}^{6}\lambda_i \quad (14)$$

2.5. Evaluation Metrics of Model Performance

The performance of individual models is evaluated from two perspectives, i.e., accuracy in reproducing historical runoff characteristics, and probability distributions of weights in multi-model ensembles. Traditionally, performance of rainfall-runoff models is evaluated by indices measuring errors in serial runoff forecast, such as Nash-Sutcliffe efficiency (NSE) and root mean square error (RMSE). These indices are sensitive to the time scale of the simulations. However, accurate simulation of precipitation and runoff at daily scale is still challenging in climate change impact studies. Weather generators such as LARS-WG put more emphasis on capturing monthly (or seasonal) means, temporal variations, and extreme events, instead of reproducing daily values with sequence information. Therefore, we here suggest four metrics to evaluate model performance at the monthly scale.

(1) Relative error of mean (REM):

$$\mathrm{REM} = \left(\overline{R_{sim}} - \overline{R_{obs}}\right)/\overline{R_{obs}} \quad (15)$$

where $\overline{R_{sim}}$ and $\overline{R_{obs}}$ are the simulated and observed mean monthly values.

(2) Relative error of standard deviation (RES):

$$\mathrm{RES} = [std(R_{sim}) - std(R_{obs})]/std(R_{obs}) \quad (16)$$

where $std(R_{sim})$ and $std(R_{obs})$ are the standard deviations of monthly runoff derived from simulated and observed series.

(3) Quantile rank score (QRS):

$$\mathrm{QRS} = \frac{\sum_{q=1}^{Q} |N_{sim}(q) - N_{obs}(q)|}{\sum_{q=1}^{Q} N_{obs}(q)} \quad (17)$$

where $N_{sim}(q)$ and $N_{obs}(q)$ are the number of months with a runoff amount lying in the range of the qth category [19]. Six categories divided by 10%, 25%, 50%, 75% and 90% quantiles were used, and the threshold values were obtained from observed runoff. A larger QRS indicates larger discrepancy in quantile distributions of the time series.

(4) Nash-Sutcliffe efficiency (NSE):

$$\mathrm{NSE} = \frac{\sum_{i=1}^{N} \left(R_{obs,i} - \overline{R_{obs}}\right)^2 - \sum_{i=1}^{N} \left(R_{obs,i} - R_{sim,i}\right)^2}{\sum_{i=1}^{N} \left(R_{obs,i} - \overline{R_{obs}}\right)^2} \quad (18)$$

where $R_{obs,i}$ and $R_{sim,i}$ are observed and simulated monthly runoff in the ith month.

3. Results

3.1. Model Performance in the Baseline Period

3.1.1. Individual Downscaling Models and Hydrological Models

SDSM and LARS-WG both reached satisfactory accuracy in capturing precipitation characteristics in the baseline period (Figure 2). SDSM and LARS-WG tended to underestimate and overestimate the total magnitudes of precipitation, respectively. However, relative errors in the simulations of mean and maximum daily precipitation were smaller than 10% in most months. The two downscaling models have shown similar intra-annual variation behaviors, except that LARS-WG predicted larger storm rainfall in July than SDSM.

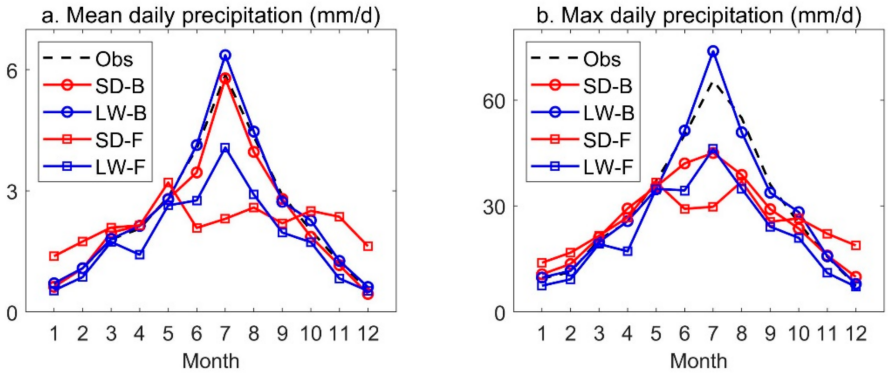

Figure 2. Mean (**a**) and maximum (**b**) daily precipitation in the baseline (1961–2000) and future periods (2060–2099) projected by the SDSM and LARS-WG downscaling models. The on-site observations, baseline simulations by SDSM, baseline simulations by LARS-WG, future simulations by SDSM, and future simulations by LARS-WG are denoted by "Obs", "SD-B", "LW-B", "SD-F", and "LW-F", respectively.

The three hydrological models were calibrated and validated for the Huai River basin in our previous studies [32]. All three hydrological models achieved good overall performance in the baseline periods (Figure 3), with REM and RES less than 5%, QRS less than 0.2, and NSE exceeding 0.8. We grouped the monthly runoff values into 12 months to evaluate the models' performance in each specific month. The results suggest different

model superiority across the months and evaluation metrics, although HBV generally achieved higher NSE and lower REM and RES in a majority of the months. For example, XAJ's performance was comparable or better than the other models in the flooding season (April–September), yet worse in drier months from November to February.

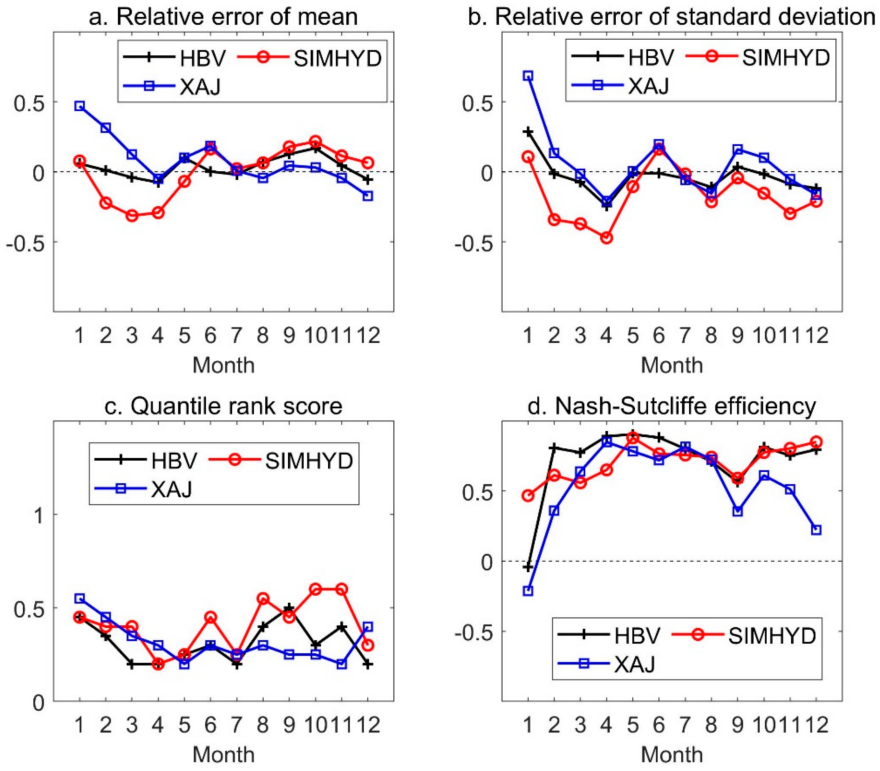

Figure 3. Performance of hydrological models by month in the baseline period driven by observed climate data.

3.1.2. Combinations of Downscaling and Hydrological Models

Compared to the hydrological models driven by observed climate, the results of the six model combinations show much larger discrepancy and intra-annual variability in capturing historical runoff characteristics (Figure 4). The bias in climate downscaling, particularly in reproducing the magnitudes and distributions of regional precipitation, were aggregated to the uncertainties in hydrological modeling (e.g., model structure, parameter estimation) nonlinearly and led to an amplified uncertainty spread. Overall, LW-HBV gained superiority in REM, RES and NSE, while LW-SIMHYD outperformed the other model combinations in QRS. However, this pattern does not hold true when evaluating the models' performance for each individual month instead of over the entire training period. For example, SD-HBV is least biased regarding RES and QRS in half of the twelve months, showing better skills in modeling the characteristics of temporal distribution.

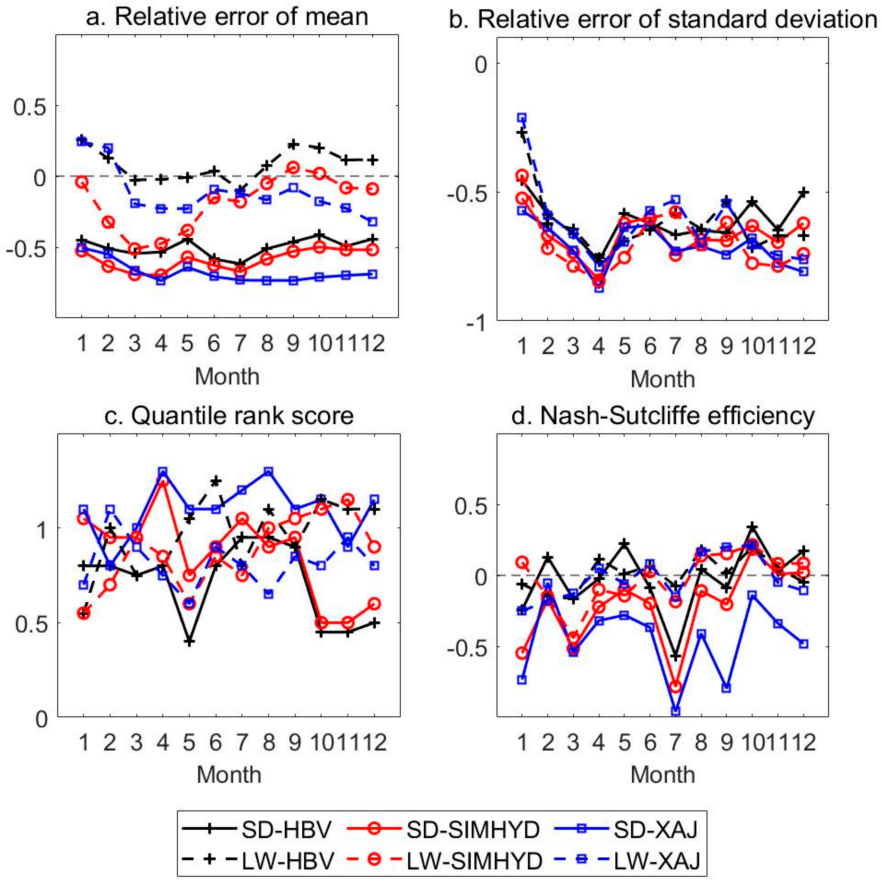

Figure 4. Performance of hydrological models in the baseline period driven by climate data derived from the SDSM and LARS-WG downscaling models.

3.2. Projected Runoff Changes in the Future

3.2.1. Projected Changes by Individual Models

We then examined the runoff changes from 1961–2000 to 2060–2099 derived from the model combinations and their multi-model ensembles. The two climate downscaling models project significantly discrepant precipitation due to the differences in the model structures and underlying assumptions (Figure 2). LARS-WG follows the changing patterns of precipitation provided by GCMs and predicts widespread decreases in precipitation through the year. On the other hand, SDSM responds to the changes in the relevant atmospheric predictors and predicts larger intra-annual variations of precipitation, suggesting increases in precipitation from October to May and decreases from June to September.

Driven by the decreasing precipitation and rising temperature, a decline in mean monthly runoff (Figure 5) is projected by all the six model combinations. The only increases are found in May and June in the SDSM derived results. LW-XAJ suggests the most notable decrease (-17.45×10^8 m^3) in annual average runoff, followed by LW-SIMHYD (-16.24×10^8 m^3) and LW-HBV (-12.31×10^8 m^3). More dramatic changes are expected in maximum monthly runoff, ranging from -198.67×10^8 m^3 (−87%, LW-XAJ in August) to 114.00×10^8 m^3 (+123%, SD-SIMHYD in May). While SDSM-derived projections suggest an increase in April–June, particularly in May, LARS-WG suggests decreases in the maximum monthly runoff in nearly all months. Future change in minimum monthly runoff is expected to vary from -10.81×10^8 m^3 (−83%, SD-SIMHYD in July) to 11.11×10^8 m^3 (+149%, LW-HBV in August). LW-HBV and LW-SIMHYD both suggest significant increases in minimum runoff in May and August–October.

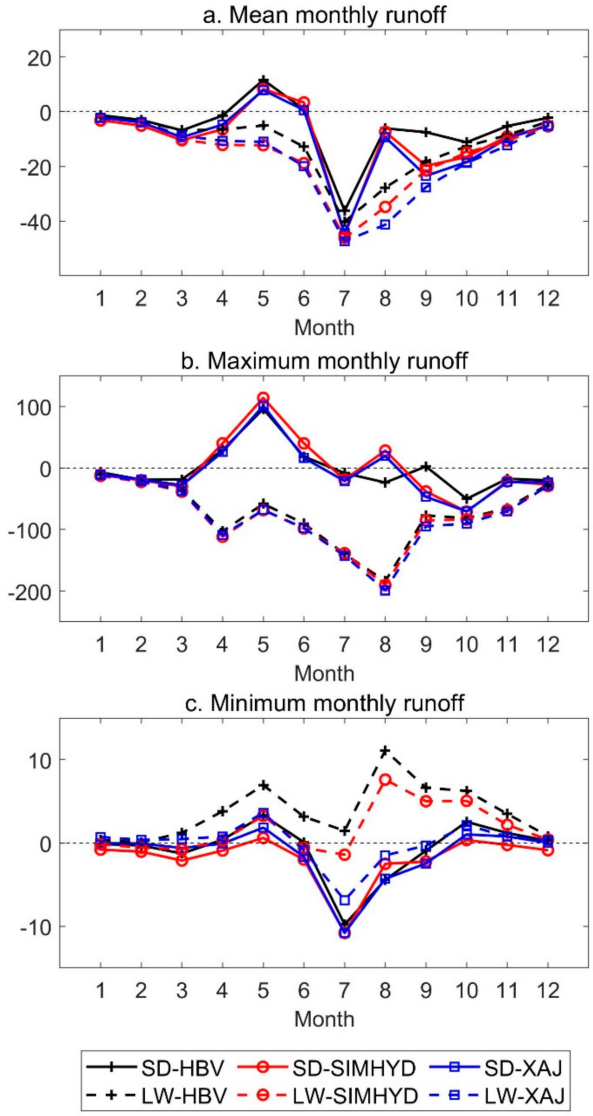

Figure 5. Changes in mean monthly runoff (**a**), maximum monthly runoff (**b**) and minimum monthly runoff (**c**) (unit: 10^8 m^3/month) from the baseline period to the future period projected with different model combinations.

3.2.2. Posterior Distributions of Runoff Change

Figure 6 displays the expectation and spreads of the 90% confidence intervals of the BMA ensemble predictions. The intra-annual variations of expected changes in mean, maximum, and minimum runoff indicate consistent increases in April–June and decreases in the remaining months. The 90% confidence interval encompasses the historical observation very well, except for a dozen outliers. In the future period, the 90% confidence interval covers much wider spreads than that in the historical period due to the inconsistent results from different model combinations. The uncertainty spread seems to expand upon time in the farther future along with the increasing extreme values projected in the 2090s. The expectation of changes in monthly runoff essentially reflects the optimized averaging of each model's results, and thus narrows down the uncertainty ranges of runoff change from the individual ensemble members.

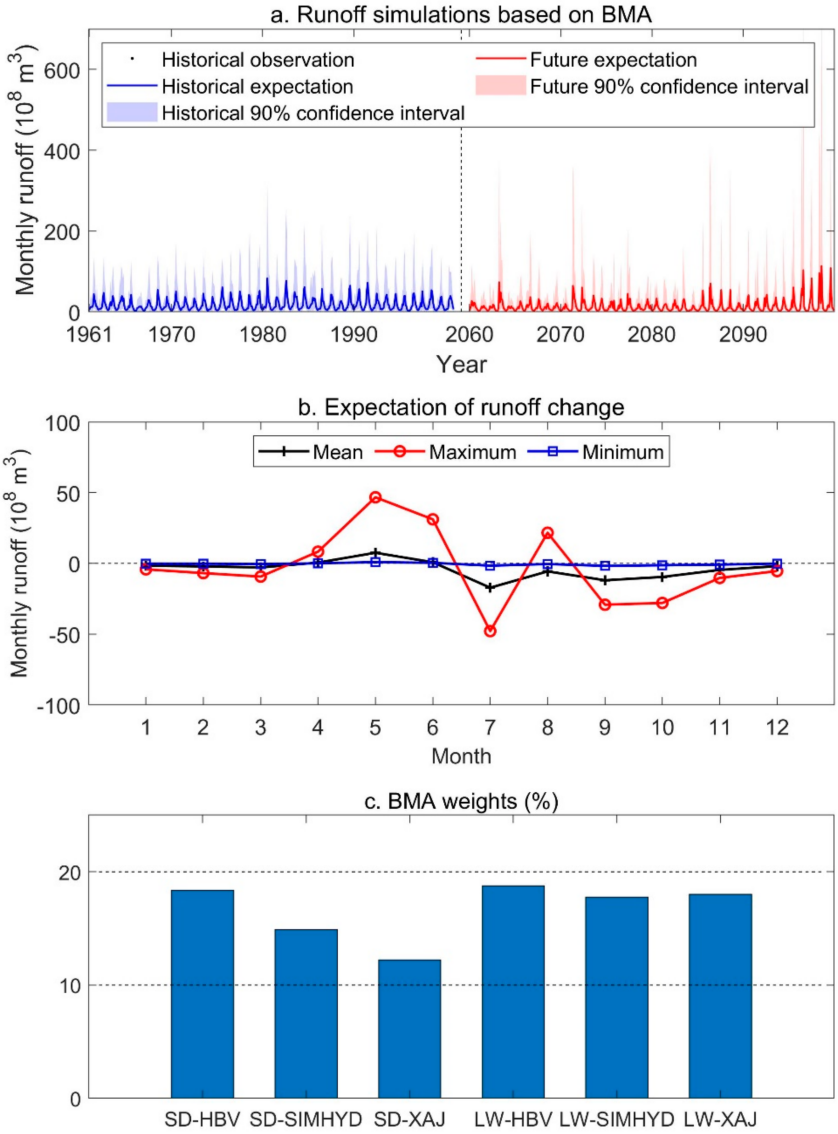

Figure 6. Runoff simulations and BMA weights computed over the entire baseline period. (**a**), Expectation and ensemble spread within the 90% confidence interval in the baseline and future periods; (**b**), Projected changes in mean, maximum, and minimum monthly runoff from the baseline to the future; (**c**), BMA weights of the six ensemble members.

The REA ensemble optimized through an MCMC approach projects future runoff changes by randomly generating numerous samples. We here focus on the posterior distributions of changes in mean, maximum, and minimum runoff (Figure 7). Changes in mean monthly runoff demonstrate consistent decrease across the entire year. Stronger decreasing signals are observed in June-November, where both the medians and the interquartile ranges (IQRs) are below −60%. The relatively smaller uncertainty ranges in March, July, and October indicate higher degree of consensus among the model combinations. Contrarily, wider uncertainty spread (e.g., in May) can be explained by the contradictory changes projected by individual models. In terms of changes in maximum runoff, positive median values are found in April, May, and August. The divergent results in April and August lead to larger spreads straddling the zero line. May is an exceptional case where the box of IQR stays on the positive side. Meanwhile, the narrowest IQRs are observed in January–March and October due to the consistent decreasing signals projected by all the six ensemble members. On the other hand, the minimum monthly runoff is more likely to

increase in May, October, and November, yet decrease in the remaining months with the IQRs varying between −50% and zero.

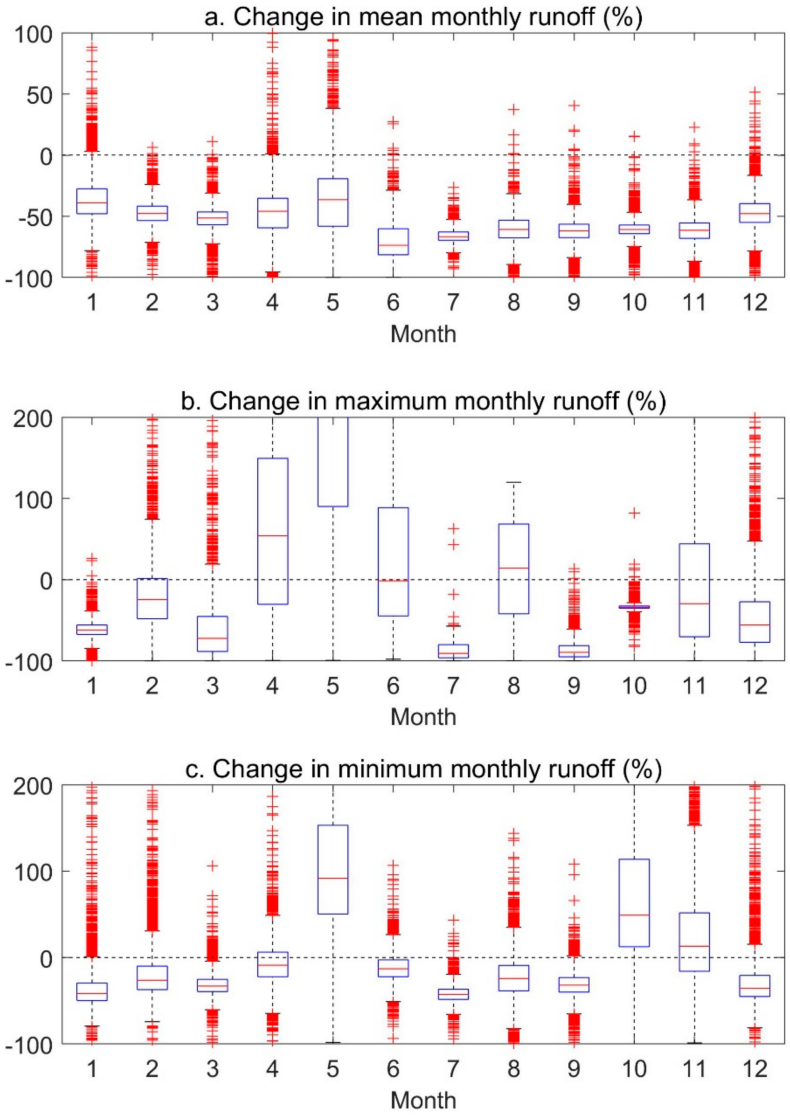

Figure 7. Posterior distributions of percent change in mean (**a**), maximum (**b**), and minimum (**c**) monthly runoff (in *y*-axis, %) from the baseline to the future period. The vertical spread of the box–whisker plots shows the variations in runoff change randomly generated by the Markov Chain Monte Carlo method. The boxes cover the ranges from the 25% quartile to the 75% quartile of the distributions, with the median values marked by red lines within each box and outliers marked by plus signs.

3.3. Uncertainties in Model Weighting

The rankings of the BMA weights of the six model combinations computed over the baseline period (Figure 6c) are consistent with their overall performance. LW-HBV and SD-HBV gained the largest weights of 18.8% and 18.4% respectively, followed by LW-XAJ (18.0%), LW-SIMHYD (17.8%), and SD-SIMHYD (14.9%). It is worth noticing that SD-XAJ also contributed 12.2%, although it seemed to be constantly outperformed by other models with respect to the evaluation metrics. We also calculated the BMA weights for records in each individual month to better represent different runoff regimes (Figure 8). The BMA weights varied between 10% and 21% among the months and ensemble members, indicating that the less effective models can also contribute significantly to the PMME runoff projection. LW-HBV was identified as the largest contributor in March–April and June–

August with a weight of 19%–21%. SD-HBV prevailed in May and September–November, while LW-SIMHYD was most heavily weighted in January–February and December.

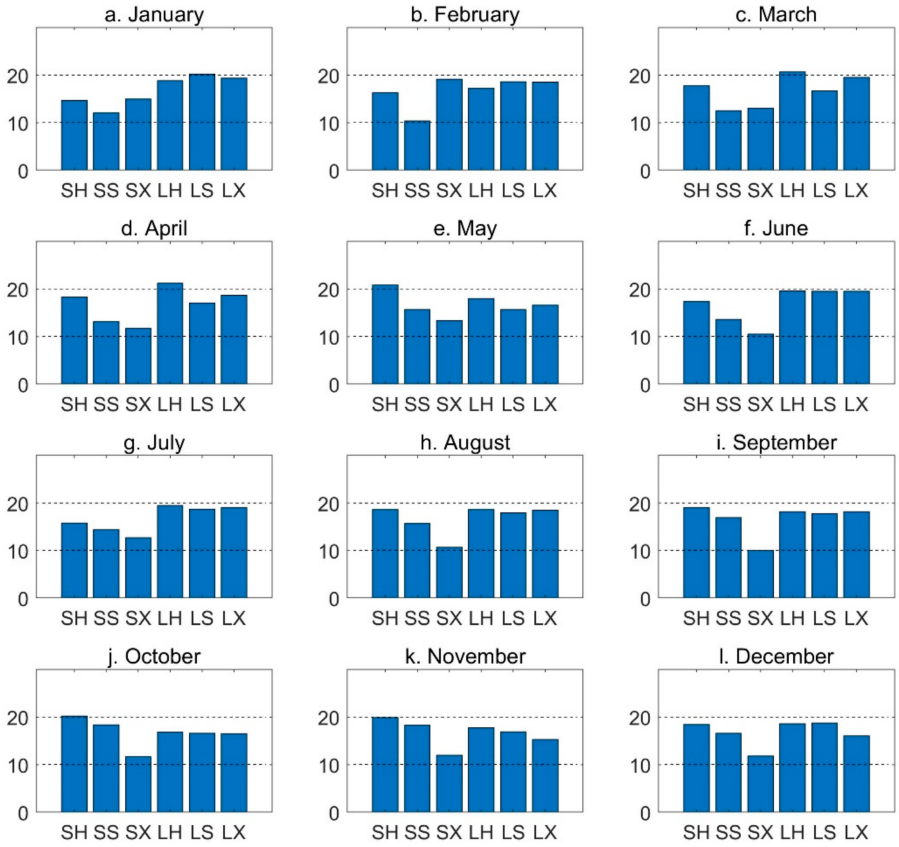

Figure 8. BMA weights (%) computed for runoff records in different months.

We calculated three sets of REA weights by optimizing the simulations of mean (REA-mean), maximum (REA-max), and minimum (REA-min) monthly runoff, respectively (Figures 9–11). These different optimization objectives forced the REA ensemble to put more weight on models' various skills. The three model combinations incorporating LARS-WG gained significantly larger weights when emphasizing on mean runoff. The median REA-mean weights of the largest contributor (LW-HBV in February–May, LW-XAJ in June–July, and LW-SHIMHYD in August–January) range between 18% and 40%. Meanwhile, models driven by SDSM gained comparable or larger weights in REA-max and REA-min than those by LARS-WG. Among the three hydrological models, HBV and XAJ outperformed the other models in maximum and minimum runoff projections, respectively. SD-HBV and LW-HBV were identified as the largest contributor in five and six out of the 12 months respectively, with the highest weight reaching 40% (LW-HBV in September). SD-XAJ contributed substantially in to projecting minimum runoff with a weight over 20% in six months, although it was least weighted in the REA-mean and REA-max ensembles in most of the months.

Between the schemes of BMA and REA, the different weighting criteria clearly lead to different evaluations of the overall superiority or inferiority of each ensemble member. Also, the Gibbs-Metropolis sampling technique used in the REA ensembles caused wider uncertainty ranges in weight estimation. The distributions of BMA weights seem to be more similar to the REA weights for maximum runoff. This could be explained by the fact that the BMA weights are optimized towards minimizing the aggregated error against the observed records in the training period. Therefore, BMA tends to force the ensemble average to fit the high flows well, rather than mean or low flows.

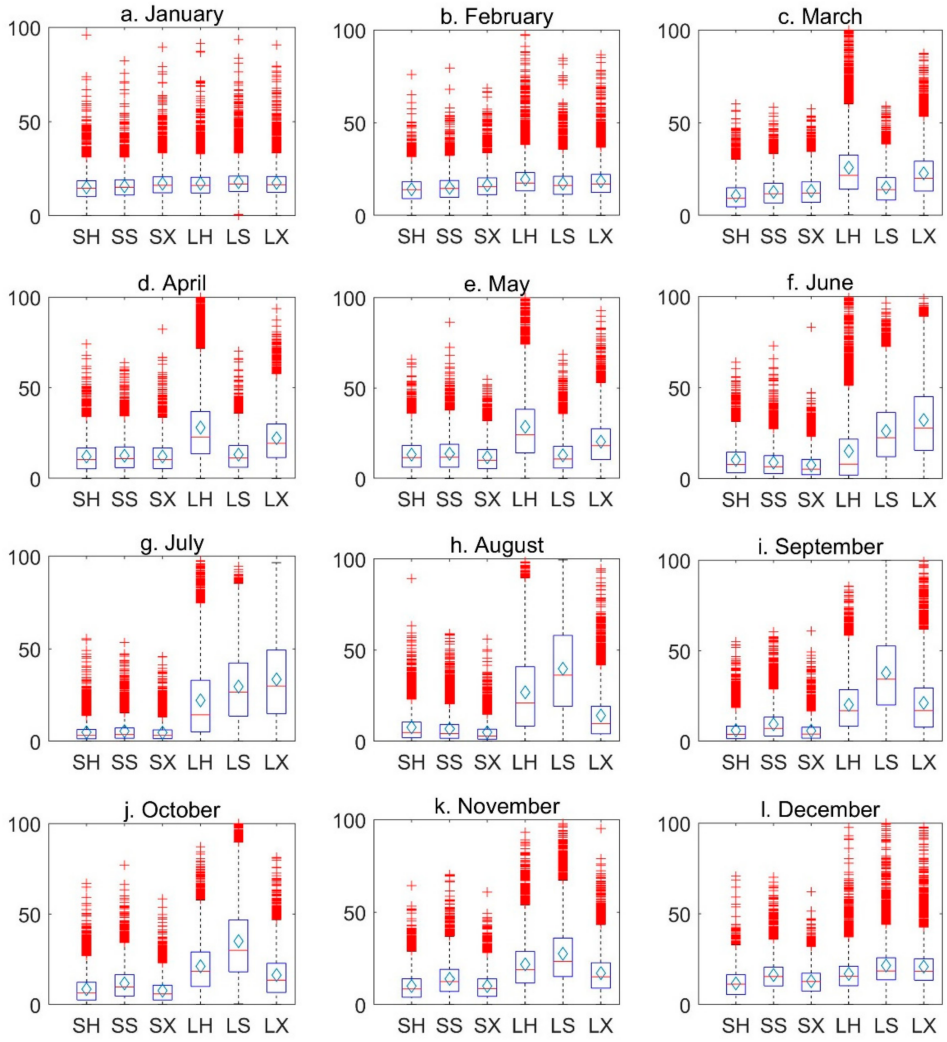

Figure 9. REA weights (%) of the six model-ensemble members in mean monthly runoff projection derived from the posterior distributions of precision parameter. The model-ensemble members are denoted by "SH" (SD-HBV), "SS" (SD-SIMHYD), "SX" (SD-XAJ), "LH" (LW-HBV), "LS" (LW-SIMHYD), and "LX" (LW-XAJ) on the *x*-axis, respectively. The mean weights of each member are marked by diamonds.

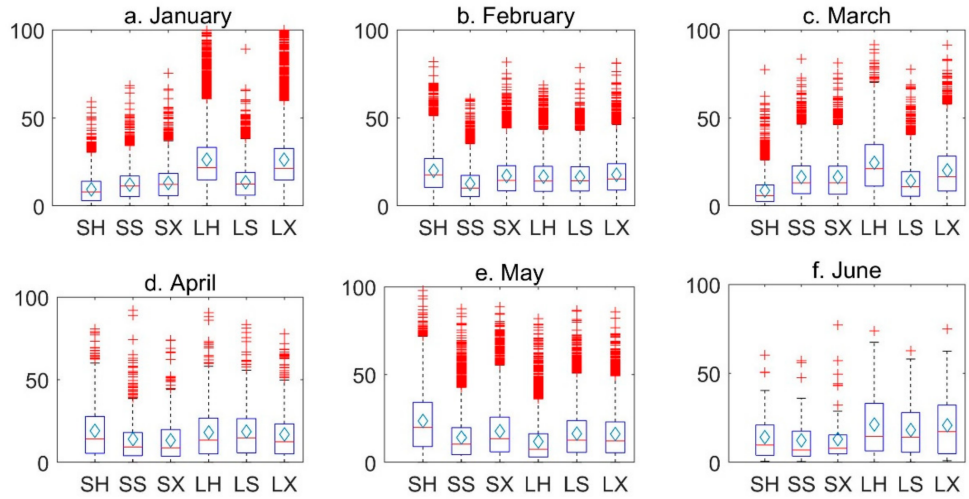

Figure 10. *Cont.*

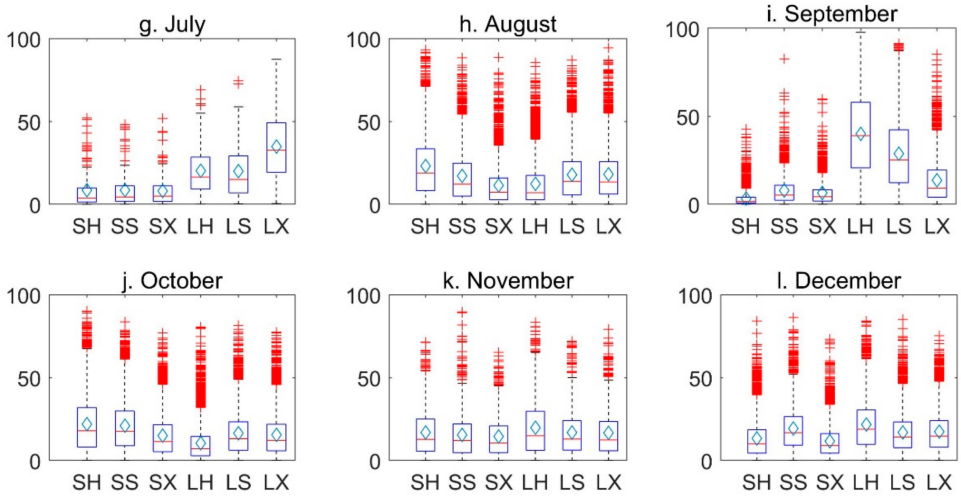

Figure 10. Same as Figure 7 but for maximum monthly runoff.

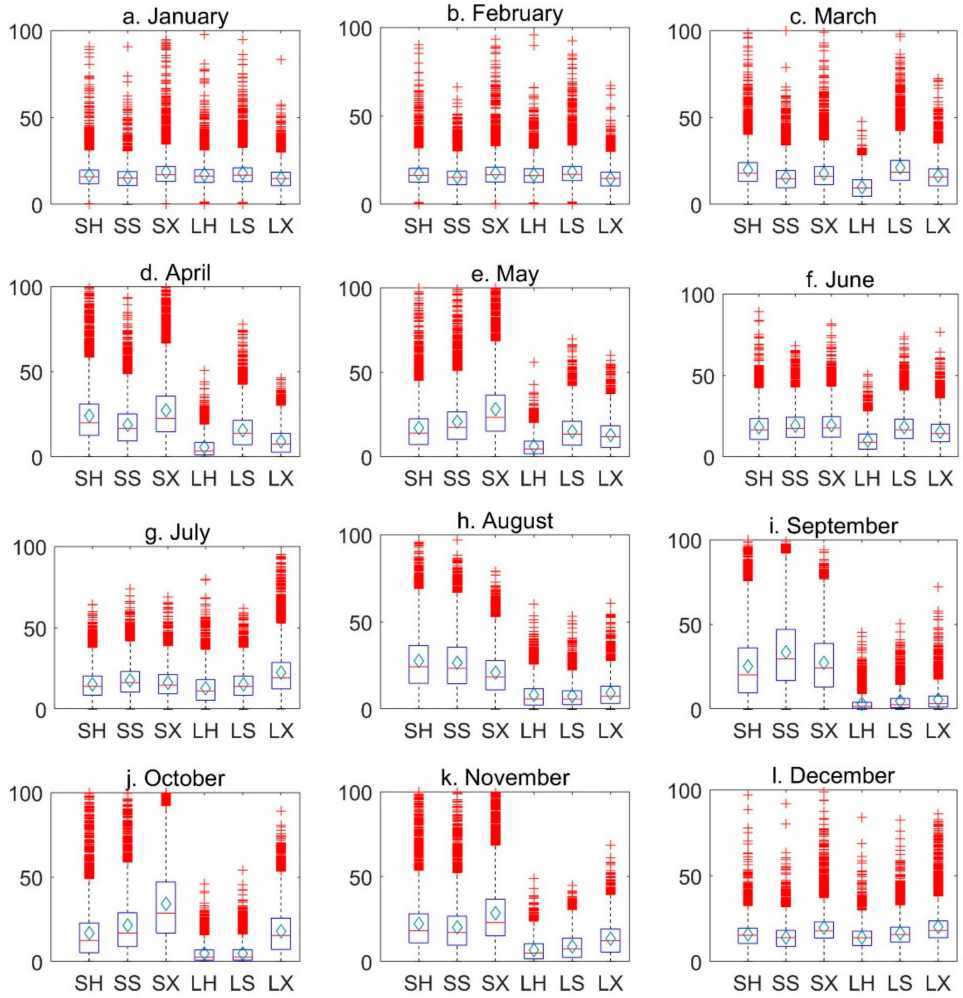

Figure 11. Same as Figure 7 but for minimum monthly runoff.

3.4. Performance of Probabilistic Multi-Model Ensembles

We cross-compared the performance of five multi-model ensembles in the baseline period, including simple equal-weight averaging (EW), BMA, REA-mean, REA-max, and REA-min (Figure 12). The results confirmed that both BMA and REA schemes have the potential of generating more skillful and reliable simulations than individual models. In

terms of NSE, BMA, REA-mean, and REA-max all provided more reliable simulation than individual models. The only exceptions were March and July, in which all the six members failed to achieve a positive NSE. Regarding the errors in simulating mean values (REM), LW-HBV was the only individual model overestimating mean runoff. The ensembles agreed with the majority of the individual models that underestimated mean runoff and showed negative REM but showed consistently larger bias than LW-HBV.

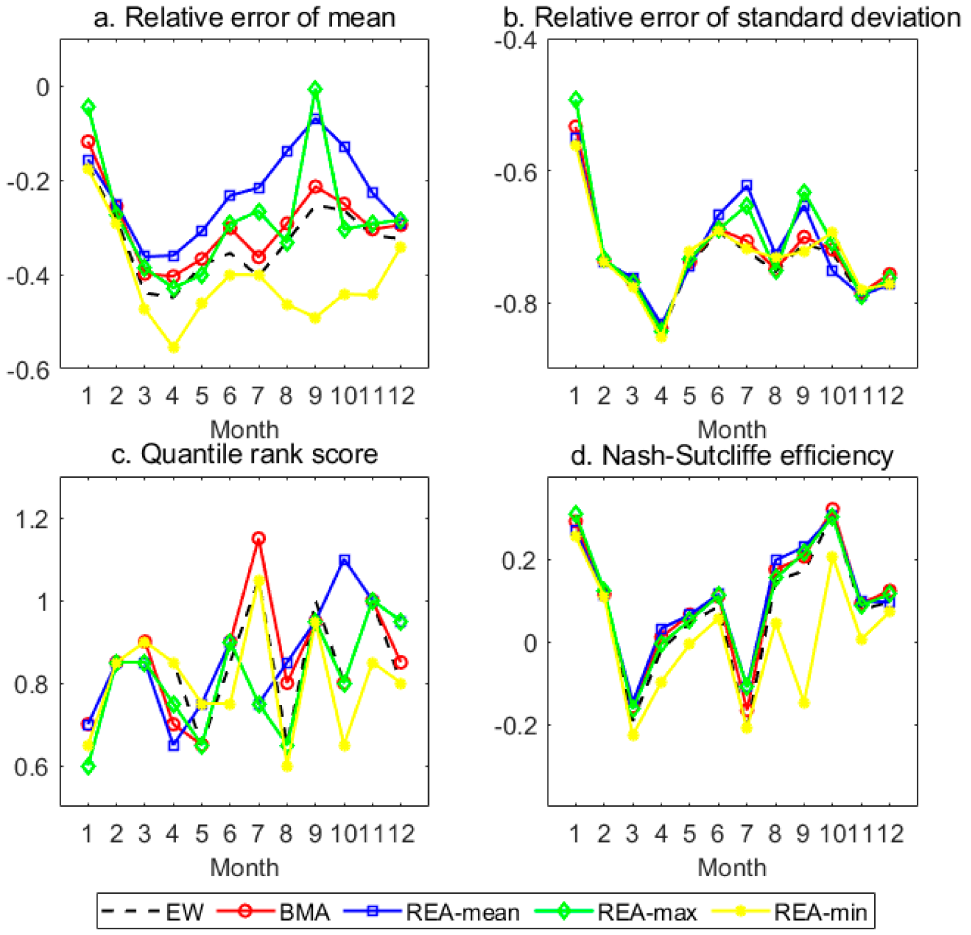

Figure 12. Performance of multi-model averaging of the six model combinations in the baseline period. Simple equal-weight averaging, Bayesian Model Averaging, Reliability Ensemble Averaging that optimizes the simulations of mean, maximum, and minimum monthly runoff are denoted by "EW", "BMA", "REA-mean", "REA-max", and "REA-min", respectively.

Among the five experiments of multi-model ensembles, different weighting strategies and optimization objectives have largely affected their performance. REA-mean put more emphasis on the simulation of mean behaviors and obtained the highest accuracy in REM in most of the months, followed by the ensembles of REA-max, BMA, EW, and REA-min. BMA and REA-max can achieve a similar NSE to that of REA-mean. In terms of both REM and NSE, REA-min was not only less efficient than REA-mean and REA-max, but also the EW averaging. However, REA-min was superior to other ensembles in terms of QRS, particularly in October–December. This inconsistent performance of the REA ensembles suggests that using different optimization objectives could cause comparable or larger discrepancy than using different averaging methods.

Multiple sets of weights have been employed in PMME schemes to enhance the performance, such as using different weights for certain aspects of the hydrograph (i.e., peak flow, mid-flow, and low flow), or for certain seasons (i.e., flooding season and dry season). In this case, we used different weights for each specific month to exploit the

diversity of flow regimes and seasonality. However, these multiple-weight strategies still cannot guarantee a 'perfect' ensemble averaging due to the inevitable tradeoffs among the optimization objectives, which are particularly notable when hydrological models are driven by downscaled climate inputs instead of historical observations.

4. Summary and Conclusions

Global and regional climate models often provide different or even conflicting climate inputs for hydrological models in future-oriented studies [51–53]. Consequently, runoff projection under climate change is much more challenging than retrospective rainfall-runoff simulations, as uncertainties from climate models and emission scenarios compound with those inherited in the structure of downscaling models and hydrological models. In this study, we addressed the usefulness and limitation of BMA and REA multi-model ensembles in post-processing runoff projections derived from diverse climate downscaling models and hydrological models. The conclusions are as follows:

(1) The compatibility of hydrological models and downscaling methods should be incorporated as an important indicator in the selection of hydrological models for runoff projection under climate change. Although the three used hydrological models can achieve similar accuracy in reproducing historical runoff, their merits in runoff projection varied with the coupled climate downscaling techniques. The uncertainties inherited from statistical downscaling tend to be accumulated nonlinearly in the hydro-climatic response. The usefulness of hydrological models in runoff projection needs to be interpreted cautiously from a broader interdisciplinary perspective.

(2) REA and BMA both can improve performance of runoff projection by synthesizing individual models' diverse skills. Our results confirmed the competing model combinations' various strength and weakness in capturing different runoff characteristics, such as bias in mean, standard deviation, quantile distribution, and Nash-Sutcliffe efficiency. Particularly, climate downscaling models based on different assumptions and inputs (as LARS-WG and SDSM in this study) lead to much larger uncertainty spreads associated with runoff projections than that with retrospective hydrological simulations. These results indicate the importance of employing valid PMME approaches for combining models of multiple processes and quantifying uncertainties from a complex set of sources. The useful information provided by less effective models can be incorporated in a probabilistic way to obtain a more reliable projection of future runoff change. Performance of the BMA, REA-mean, and REA-max ensembles in the baseline period suggest comparable accuracy in capturing monthly runoff characteristics. Among the three REA ensembles, BMA results resembled REA-max because BMA's weighting procedure forces the ensemble average to fit the high flows well.

(3) Different weighting criteria and parameter estimation methods lead to larger uncertainty ranges in the weight distributions of REA ensembles, although the reliability of BMA and REA in runoff projection is comparable. Weight distributions derived from the REA and BMA schemes both confirm that less effective models (e.g., SD-XAJ in this case) can also contribute significantly. Individual models' BMA weights were limited between 10%–21%, but mean REA weights varied from 2% to 40%. Compared to BMA's weighting criteria (i.e., models' relative contributions to predictive skill in the training period), REA puts more emphasis on the convergent projection of the future, which brings in uncertainties in future climate into the weight estimation. Besides, the REA ensembles used a MCMC sampling approach (the Gibbs-Metropolis algorithm) to inflate the posterior distributions of parameters, while BMA uses the Expectation maximization algorithm to reach a minimum aggregated error. These differences resulted in wider uncertainty ranges in REA weight estimation.

(4) The performance of the model ensembles in uncertain runoff projection is largely restrained by the tradeoffs among different averaging objectives, even though multi-weight strategies (distinct sets of weight for each month in this case) are applied to exploit the diversity of flow regimes and seasonality. Runoff projection is essentially a multi-objective

task of capturing various runoff features (e.g., means, temporal distributions, and extreme events) with large uncertainties. Our results highlight the limitation of multi-model averaging approaches. Although multiple-weight strategies can better represent the different flow regimes, it is still difficult to capture various runoff features simultaneously with highly uncertain climate inputs. Therefore, the averaging ensemble could not guarantee an improvement from the best individual models on every aspect. The predictive skills in capturing different runoff characteristics vary greatly not only among the individual models, but also among the averaging ensembles. A cross-comparison of the three REA ensembles using different optimization objectives shows large discrepancy in terms of both weight distribution and predictive skills, suggesting that different optimization objectives could cause more significant discrepancy in runoff projection than that induced by different weighting criteria or parameter estimation algorithms.

Several caveats apply to our study. We did not consider climate simulations from different GCMs or under different emission scenarios. The focus of this study is to investigate the effectiveness of multi-model averaging approaches in synthesizing multiple downscaling models and hydrological models, rather than predicting potential changes in regional runoff. Therefore, the modeling experiments were limited to the procedures of downscaling and hydrological simulation. The incorporation of data derived from additional GCMs and scenarios would introduce larger uncertainty spreads to the results but is not likely to invalidate our major conclusions. Besides, we examined the consistency and inconsistency in the averaging results derived from different weighting criteria and optimization goals in the context of BMA and REA. The roles of more aspects of the PMME methods, such as using different procedures of sampling and uncertainty reduction, need to be addressed in further research.

Author Contributions: K.D.: Methodology, analyzing, writing; X.W.: Data curation, analyzing; B.L.: Review & editing; T.Z.: Review & editing; X.C.: Review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (51909285, U1911204), and the Guangdong Provincial Department of Science and Technology (2019ZT08G090).

Data Availability Statement: All data used in this paper is available upon reasonable request from the authors.

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

References

1. Xu, C.Y. From gcms to river flow: A review of downscaling methods and hydrologic modelling approaches. *Prog. Phys. Geogr.* **1999**, *23*, 229–249. [CrossRef]
2. Milly, P.C.; Dunne, K.A.; Vecchia, A.V. Global pattern of trends in streamflow and water availability in a changing climate. *Nature* **2005**, *438*, 347–350. [CrossRef] [PubMed]
3. Arnell, N.W.; Gosling, S.N. The impacts of climate change on river flow regimes at the global scale. *J. Hydrol.* **2013**, *486*, 351–364. [CrossRef]
4. Duan, K.; Sun, G.; McNulty, S.G.; Caldwell, P.V.; Cohen, E.C.; Sun, S.; Aldridge, H.D.; Zhou, D.; Zhang, L.; Zhang, Y. Future shift of the relative roles of precipitation and temperature in controlling annual runoff in the conterminous united states. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 5517–5529. [CrossRef]
5. Duan, K.; Caldwell, P.V.; Sun, G.; McNulty, S.G.; Zhang, Y.; Shuster, E.; Liu, B.; Bolstad, P.V. Understanding the role of regional water connectivity in mitigating climate change impacts on surface water supply stress in the united states. *J. Hydrol.* **2019**, *570*, 80–95. [CrossRef]
6. Fowler, H.J.; Blenkinsop, S.; Tebaldi, C. Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol. A. J. R. Meteorol. Soc.* **2007**, *27*, 1547–1578. [CrossRef]
7. Wilby, R.L.; Charles, S.; Zorita, E.; Timbal, B.; Whetton, P.; Mearns, L. Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods. Supporting Material of the Intergovernmental Panel on Climate Change, Available from the DDC of IPCC TGCIA 27. 2004. Available online: https://www.ipcc-data.org/guidelines/dgm_no2_v1_09_2004.pdf (accessed on 19 July 2021).
8. Chen, J.; Brissette, F.P.; Leconte, R. Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. *J. Hydrol.* **2011**, *401*, 190–202. [CrossRef]

9. Chiew, F.; Kirono, D.; Kent, D.; Frost, A.; Charles, S.; Timbal, B.; Nguyen, K.; Fu, G. Comparison of runoff modelled using rainfall from different downscaling methods for historical and future climates. *J. Hydrol.* **2010**, *387*, 10–23. [CrossRef]
10. Duan, K.; Sun, G.; Zhang, Y.; Yahya, K.; Wang, K.; Madden, J.M.; Caldwell, P.V.; Cohen, E.C.; McNulty, S.G. Impact of air pollution induced climate change on water availability and ecosystem productivity in the conterminous united states. *Clim. Chang.* **2017**, *140*, 259–272. [CrossRef]
11. Zhang, L.; Potter, N.; Hickel, K.; Zhang, Y.; Shao, Q. Water balance modeling over variable time scales based on the budyko framework—model development and testing. *J. Hydrol.* **2008**, *360*, 117–131. [CrossRef]
12. Crosbie, R.S.; Scanlon, B.R.; Mpelasoka, F.S.; Reedy, R.C.; Gates, J.B.; Zhang, L. Potential climate change effects on groundwater recharge in the high plains aquifer, USA. *Water Resour. Res.* **2013**, *49*, 3936–3951. [CrossRef]
13. Driessen, T.; Hurkmans, R.; Terink, W.; Hazenberg, P.; Torfs, P.; Uijlenhoet, R. The hydrological response of the ourthe catchment to climate change as modelled by the HBV model. *Hydrol. Earth Syst. Sci.* **2010**, *14*, 651–665. [CrossRef]
14. Christensen, N.S.; Lettenmaier, D.P. A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the colorado river basin. *Hydrol. Earth Syst. Sci.* **2007**, *11*, 1417–1434. [CrossRef]
15. Chen, H.; Xu, C.-Y.; Guo, S. Comparison and evaluation of multiple gcms, statistical downscaling and hydrological models in the study of climate change impacts on runoff. *J. Hydrol.* **2012**, *434*, 36–45. [CrossRef]
16. Dibike, Y.B.; Coulibaly, P. Hydrologic impact of climate change in the saguenay watershed: Comparison of downscaling methods and hydrologic models. *J. Hydrol.* **2005**, *307*, 145–163. [CrossRef]
17. Tian, Y.; Xu, Y.-P.; Zhang, X.-J. Assessment of climate change impacts on river high flows through comparative use of gr4j, HBV and xinanjiang models. *Water Resour. Manag.* **2013**, *27*, 2871–2888. [CrossRef]
18. Diks, C.G.; Vrugt, J.A. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stoch. Environ. Res. Risk Assess.* **2010**, *24*, 809–820. [CrossRef]
19. Duan, K.; Mei, Y. A comparison study of three statistical downscaling methods and their model-averaging ensemble for precipitation downscaling in china. *Theor. Appl. Climatol.* **2014**, *116*, 707–719. [CrossRef]
20. Duan, Q.; Ajami, N.K.; Gao, X.; Sorooshian, S. Multi-model ensemble hydrologic prediction using bayesian model averaging. *Adv. Water Resour.* **2007**, *30*, 1371–1386. [CrossRef]
21. Wöhling, T.; Vrugt, J.A. Combining multiobjective optimization and bayesian model averaging to calibrate forecast ensembles of soil hydraulic models. *Water Resour. Res.* **2008**, *44*, 12. [CrossRef]
22. Khan, M.S.; Coulibaly, P. Assessing hydrologic impact of climate change with uncertainty estimates: Bayesian neural network approach. *J. Hydrometeorol.* **2010**, *11*, 482–495. [CrossRef]
23. Yang, T.; Hao, X.; Shao, Q.; Xu, C.-Y.; Zhao, C.; Chen, X.; Wang, W. Multi-model ensemble projections in temperature and precipitation extremes of the tibetan plateau in the 21st century. *Glob. Planet. Chang.* **2012**, *80*, 1–13. [CrossRef]
24. Demirel, M.C.; Moradkhani, H.J.C.C. Assessing the impact of cmip5 climate multi-modeling on estimating the precipitation seasonality and timing. *Clim. Chang.* **2016**, *135*, 357–372. [CrossRef]
25. van Vliet, M.T.; Blenkinsop, S.; Burton, A.; Harpham, C.; Broers, H.P.; Fowler, H.J. A multi-model ensemble of downscaled spatial climate change scenarios for the dommel catchment, western europe. *Clim. Chang.* **2012**, *111*, 249–277. [CrossRef]
26. Yang, H.; Wang, B. Reducing biases in regional climate downscaling by applying bayesian model averaging on large-scale forcing. *Clim. Dyn.* **2012**, *39*, 2523–2532. [CrossRef]
27. Nury, A.H.; Sharma, A.; Marshall, L.; Mehrotra, R. Characterising uncertainty in precipitation downscaling using a bayesian approach. *Adv. Water Resour.* **2019**, *129*, 189–197. [CrossRef]
28. Hashmi, M.; Shamseldin, A.; Melville, B. Statistical downscaling of precipitation: State-of-the-art and application of bayesian multi-model approach for uncertainty assessment. *Hydrol. Earth Syst. Sci. Discuss.* **2009**, *6*, 6535–6579.
29. Maraun, D.; Wetterhall, F.; Ireson, A.; Chandler, R.; Kendon, E.; Widmann, M.; Brienen, S.; Rust, H.; Sauter, T.; Themeßl, M. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* **2010**, *48*, RG3003. [CrossRef]
30. 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]
31. Liu, Z.; Merwade, V. Accounting for model structure, parameter and input forcing uncertainty in flood inundation modeling using bayesian model averaging. *J. Hydrol.* **2018**, *565*, 138–149. [CrossRef]
32. Duan, K.; Mei, Y. Comparison of meteorological, hydrological and agricultural drought responses to climate change and uncertainty assessment. *Water Resour. Manag.* **2014**, *28*, 5039–5054. [CrossRef]
33. Giorgi, F.; Mearns, L.O. Calculation of average, uncertainty range, and reliability of regional climate changes from aogcm simulations via the "reliability ensemble averaging"(rea) method. *J. Clim.* **2002**, *15*, 1141–1158. [CrossRef]
34. Wilby, R.L.; Hay, L.E.; Leavesley, G.H. A comparison of downscaled and raw gcm output: Implications for climate change scenarios in the san juan river basin, colorado. *J. Hydrol.* **1999**, *225*, 67–91. [CrossRef]
35. Semenov, M.A.; Stratonovitch, P. Use of multi-model ensembles from global climate models for assessment of climate change impacts. *Clim. Res.* **2010**, *41*, 1–14. [CrossRef]
36. Seibert, J. Estimation of parameter uncertainty in the HBV model: Paper presented at the nordic hydrological conference (akureyri, iceland-august 1996). *Hydrol. Res.* **1997**, *28*, 247–262. [CrossRef]

37. Seibert, J. *HBV-Light, Users Manual Version 2*; Department of Physical Geography, Stockholm University: Stockholm, Sweden, 2005.
38. Tan, K.; Chiew, F.; Grayson, R.; Scanlon, P.; Siriwardena, L. Calibration of a Daily Rainfall-Runoff Model to Estimate High Daily Flows. In Proceedings of the MODSIM 2005 International Congress on Modelling and Simulation, Melbourne, Australia, 12–15 December 2005; pp. 2960–2966. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.375.5625&rep=rep1&type=pdf (accessed on 19 July 2021).
39. Zhao, R.-J. The xinanjiang model applied in china. *J. Hydrol.* **1992**, *135*, 371–381.
40. Duan, K.; Xiao, W.; Mei, Y.; Liu, D. Multi-scale analysis of meteorological drought risks based on a bayesian interpolation approach in huai river basin, china. *Stoch. Environ. Res. Risk Assess.* **2014**, *28*, 1985–1998. [CrossRef]
41. Kalnay, E.; Kanamitsu, M.; Kistler, R.; Collins, W.; Deaven, D.; Gandin, L.; Iredell, M.; Saha, S.; White, G.; Woollen, J. The ncep/ncar 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **1996**, *77*, 437–472. [CrossRef]
42. Duan, K.; Mei, Y. Uncertainty analysis of precipitation change based on tebaldi multimodel ensemble method. *Eng. J. Wuhan Univ.* **2015**, *048*, 433–440.
43. Tebaldi, C.; Mearns, L.O.; Nychka, D.; Smith, R.L. Regional probabilities of precipitation change: A bayesian analysis of multimodel simulations. *Geophys. Res. Lett.* **2004**, *31*, L24213. [CrossRef]
44. Tebaldi, C.; Smith, R.L.; Nychka, D.; Mearns, L.O. Quantifying uncertainty in projections of regional climate change: A bayesian approach to the analysis of multimodel ensembles. *J. Clim.* **2005**, *18*, 1524–1540. [CrossRef]
45. Lopez, A.; Tebaldi, C.; New, M.; Stainforth, D.; Allen, M.; Kettleborough, J. Two approaches to quantifying uncertainty in global temperature changes. *J. Clim.* **2006**, *19*, 4785–4796. [CrossRef]
46. Hashmi, M.Z.; Shamseldin, A.Y.; Melville, B.W. Statistically downscaled probabilistic multi-model ensemble projections of precipitation change in a watershed. *Hydrol. Process.* **2013**, *27*, 1021–1032. [CrossRef]
47. Meehl, G.A.; Boer, G.J.; Covey, C.; Latif, M.; Stouffer, R.J. The coupled model intercomparison project (cmip). *Bull. Am. Meteorol. Soc.* **2000**, *81*, 313–318. [CrossRef]
48. Georgakakos, K.P.; Seo, D.-J.; Gupta, H.; Schaake, J.; Butts, M.B. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *J. Hydrol.* **2004**, *298*, 222–241. [CrossRef]
49. Yu, G.; Fu, Y.; Sun, X.; Wen, X.; Zhang, L. Recent progress and future directions of chinaflux. *Sci. China Ser. D Earth Sci.* **2006**, *49*, 1–23. [CrossRef]
50. Smith, R.L.; Tebaldi, C.; Nychka, D.; Mearns, L.O. Bayesian modeling of uncertainty in ensembles of climate models. *J. Am. Stat. Assoc.* **2009**, *104*, 97–116. [CrossRef]
51. Pachauri, R.K.; Allen, M.R.; Barros, V.R.; Broome, J.; Cramer, W.; Christ, R.; Church, J.A.; Clarke, L.; Dahe, Q.; Dasgupta, P.; et al. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; IPCC: Geneva, Switzerland, 2014; ISBN 9291691437.
52. Giuntoli, I.; Vidal, J.-P.; Prudhomme, C.; Hannah, D.M. Future hydrological extremes: The uncertainty from multiple global climate and global hydrological models. *Earth Syst. Dyn.* **2015**, *6*, 267–285. [CrossRef]
53. Knutti, R.; Sedláček, J. Robustness and uncertainties in the new cmip5 climate model projections. *Nat. Clim. Chang.* **2013**, *3*, 369–373. [CrossRef]