Water Resources Research

RESEARCH ARTICLE
10.1029/2018WR023197

Hydrological Model Diversity Enhances Streamflow Forecast Skill at Short- to Medium-Range Timescales

Sanjib Sharma[1], Ridwan Siddique[2], Seann Reed[3], Peter Ahnert[3], and Alfonso Mejia[1]

[1]Department of Civil and Environmental Engineering, The Pennsylvania State University, State College, PA, USA, [2]Northeast Climate Science Center, University of Massachusetts, Amherst, MA, USA, [3]National Weather Service, Middle Atlantic River Forecast Center, State College, PA, USA

Key Points:
- A new technique (QR-BMA) is implemented to combine multimodel ensemble streamflow predictions at short- to medium-range timescales
- Multimodel ensemble streamflow forecasts are more skillful than single-model ensemble forecasts under various experimental conditions
- Skill enhancements in the multimodel streamflow forecasts are dominated by model diversity rather than by increased ensemble size alone

Supporting Information:
- Supporting Information S1

Correspondence to:
A. Mejia,
amejia@engr.psu.edu

Citation:
Sharma, S., Siddique, R., Reed, S., Ahnert, P., & Mejia, A. (2019). Hydrological model diversity enhances streamflow forecast skill at short- to medium-range timescales. *Water Resources Research*, *55*, 1510–1530. https://doi.org/10.1029/2018WR023197

Received 24 APR 2018
Accepted 25 JAN 2019
Accepted article online 29 JAN 2019
Published online 21 FEB 2019

Abstract We investigate the ability of hydrological multimodel ensemble predictions to enhance the skill of streamflow forecasts at short- to medium-range timescales. To generate the multimodel ensembles, we implement a new statistical postprocessor, namely, quantile regression-Bayesian model averaging (QR-BMA). Quantile regression-Bayesian model averaging uses quantile regression to bias correct the ensemble streamflow forecasts from the individual models and Bayesian model averaging to optimally combine their probability density functions. Additionally, we use an information-theoretic measure, namely, conditional mutual information, to quantify the skill enhancements from the multimodel forecasts. We generate ensemble streamflow forecasts at lead times from 1 to 7 days using three hydrological models: (i) Antecedent Precipitation Index-Continuous, (ii) Hydrology Laboratory-Research Distributed Hydrologic Model, and (iii) Weather Research and Forecasting Hydrological modeling system. As forcing to the hydrological models, we use weather ensemble forecasts from the National Centers for Environmental Prediction 11-member Global Ensemble Forecast System Reforecast version 2. The forecasting experiments are performed for four nested basins of the North Branch Susquehanna River, USA. We find that after bias correcting the streamflow forecasts from each model, their skill performance becomes comparable. We find that the multimodel ensemble forecasts have higher skill than the best single-model forecasts. Furthermore, the skill enhancements obtained by the multimodel ensemble forecasts are found to be dominated by model diversity, rather than by increased ensemble size alone. This result, obtained using conditional mutual information, indicates that each hydrological model contributes additional information to enhance forecast skill. Overall, our results highlight benefits of hydrological multimodel forecasting for improving streamflow predictions.

1. Introduction

Multimodel forecasting is a well-established technique in atmospheric science (Bosart, 1975; Gyakum, 1986; Krishnamurti, 2003; Sanders, 1973; Weisheimer et al., 2009), which consists of using the outputs from several models to make and improve predictions about future events (Fritsch et al., 2000). The motivation for multimodel forecasting is that for a complex system, such as the atmosphere or a river basin, comprised by multiple processes interacting nonlinearly and with limited observability, predictions solely based on the outputs from a single model will be prone to errors and biases (Fritsch et al., 2000). Indeed, early experiments comparing blended forecasts from different weather models against single-model predictions demonstrated the ability of multimodel predictions to improve the skill and reduce the errors of weather forecasts (Bosart, 1975; Gyakum, 1986; Sanders, 1973; Thompson, 1977; Winkler et al., 1977). This was found to be the case for both forecasts issued by humans (Sanders, 1963, 1973) and from numerical models (Bosart, 1975; Fraedrich & Leslie, 1987; Fraedrich & Smith, 1989; Fritsch et al., 2000; Gyakum, 1986; Krishnamurti et al., 1999, 2000; Sanders, 1973).

Initial meteorological multimodel experiments accounted for model-related uncertainties but not for uncertainties in the initial states. To account for the latter, multimodel ensembles were introduced, where multiple ensemble members from individual models are generated for the same lead time and geographic area by perturbing the models' initial states (Hamill & Colucci, 1997; Stensrud et al., 1999; Toth & Kalnay, 1993). An illustrative example of a recent, successful multimodel framework is the North American Multimodel Ensemble experiment for subseasonal to seasonal timescales (Bastola et al., 2013; Becker et al., 2014;

Kirtman et al., 2013). Indeed, most of the established operational systems across the globe for short- to medium-range weather forecasting are multimodel, multiphysics ensemble systems (Buizza et al., 2005; Du et al., 2003; Hamill et al., 2013; Palmer et al., 2004). In contrast, hydrological multimodel ensemble prediction systems (HMEPS) have not been widely implemented and remain an underexplored area of research. To our knowledge, there is currently no operational HMEPS in the world, despite their success in weather (Hagedorn et al., 2012; Hamill et al., 2013) and climate forecasting (Bastola et al., 2013; Becker et al., 2014; Kirtman et al., 2013).

HMEPS can be classified into the following three general categories, depending on whether multiple weather and/or hydrological models are used: (i) a single hydrological model forced by outputs from multiple numerical weather prediction (NWP) models (Thirel et al., 2008, 2010), (ii) multiple hydrological models forced by outputs from a single NWP model (Randrianasolo et al., 2010), and (iii) multiple hydrological models forced by outputs from multiple NWP models (Velázquez et al., 2011). As is the case in meteorology, hydrological multimodel outputs can be deterministic or probabilistic, depending on how many and the manner in which ensembles are generated from each model (Davolio et al., 2008). It is important to note that although hydrological multimodel approaches have been investigated before (Ajami et al., 2007; Duan et al., 2007; Vrugt & Robinson, 2007), the vast majority of those studies have been performed in simulation mode (i.e., by forcing the hydrological models with observed weather variables), as opposed to forecasting mode. Simulation studies may provide useful information about near-real-time hydrological forecasting conditions. However, at medium-range timescales ($\geq$ 3 days), where weather uncertainties tend to be as important or more dominant than hydrological uncertainties, hydrological simulations provide considerably less information about forecast behavior (Sharma et al., 2018; Siddique & Mejia, 2017).

One of the earliest attempt at hydrological multimodel prediction is that of Shamseldin and O'Connor (1999). They combined streamflow simulations from different rainfall-runoff models by assigning different weights to the models based on their performance during historical runs. Since then, several simulation studies have been performed to address the potential of hydrological multimodel approaches to improve understanding and prediction of hydrological variables (Ajami et al., 2007; Bohn et al., 2010; Duan et al., 2007; Georgakakos et al., 2004; Regonda et al., 2006; Vrugt & Robinson, 2007). In hydrological forecasting, recent implementations of the multimodel approach have been focused on seasonal or longer timescales (Nohara et al., 2006; Yuan & Wood, 2013), while very few studies are available at short- to medium-range timescales (Hopson & Webster, 2010; Velázquez et al., 2011). Furthermore, a shortcoming of the latter studies has been the use of similar hydrological models to generate the multimodel forecasts. For example, Hopson and Webster (2010) as well as Velázquez et al. (2011) used similar spatially lumped or semidistributed hydrological models for their respective multimodel experiments.

To maximize the benefits from a multimodel approach, it is critical to use dissimilar models (Thompson, 1977), a property that is referred to as model diversity (DelSole et al., 2014). In hydrological science, different model types are available that could be used to fulfill model diversity, for example, spatially lumped, spatially distributed, process-based, or land-surface models (Reed et al., 2004; Smith et al., 2012). These different types of models tend to differ markedly in their spatial discretization, physical parameterizations, and numerical schemes (Kollet et al., 2017), potentially making them good candidates for multimodel forecasting. Another important concern with the multimodel approach is that of distinguishing whether any gains in skill from the multimodel are due to model diversity itself or are related to increases in the ensemble size. Recently, an information-theoretic measure, namely, conditional mutual information (*CMI*), was proposed to address this issue in climate forecasts (DelSole et al., 2014). *CMI* is implemented here for the first time with hydrological multimodel forecasts.

Any multimodel forecast requires some type of statistical technique (with simple averaging being the simplest approach; DelSole, 2007; DelSole et al., 2013) or postprocessor (Duan et al., 2007; Fraley et al., 2010; Gneiting et al., 2005; Raftery et al., 1997) to optimally combine the ensemble forecasts from the individual models. Multimodel postprocessing is typically employed to accomplish several objectives: (i) reduce systematic biases in the outputs from each model, (ii) assign each model a weight that measures its contribution to the final multimodel forecast, and (iii) quantify the overall forecast uncertainty. Although a number of multimodel postprocessors have been developed and implemented for dealing with hydrological simulations (Duan et al., 2007; Hsu et al., 2009; Madadgar & Moradkhani, 2014; Najafi et al.,

2011; Shamseldin et al., 1997; Steinschneider et al., 2015; Vrugt & Robinson, 2007; Xiong et al., 2001), few have been applied in a forecasting context (Hopson & Webster, 2010). In this study, we implement a new quantile regression-Bayesian model averaging (QR-BMA) postprocessor. The postprocessor uses QR to bias correct the streamflow forecasts from the individual models (Sharma et al., 2018) and BMA to optimally combine their probability density functions (pdfs; Duan et al., 2007; Vrugt & Robinson, 2007). QR-BMA takes advantage of the proven effectiveness and simplicity of QR to remove systematic biases (Gomez et al., 2019; Sharma et al., 2018) and of BMA to produce optimal weights (Duan et al., 2007; Liang et al., 2013).

Our primary goal with this study is to understand the ability of hydrological multimodel ensemble predictions to improve the skill of streamflow forecasts at short- to medium-range timescales. With this goal, we seek to answer the following two main questions: Are multimodel ensemble streamflow forecasts more skillful than single-model forecasts? Are any skill improvements from the multimodel ensemble streamflow forecasts dominated by model diversity or the addition of new ensemble members (i.e., increasing ensemble size)? Answering the latter is relevant to operational forecasting because generating many ensemble members in real time is often not feasible or realistic and may not be as effective if skill enhancements are dominated by model diversity. The paper is structured as follows. Section 2 describes our methodology. Section 3 describes the experimental setup. The main results and their implications are presented in section 4. Lastly, section 5 summarizes our conclusions.

2. Methodology

2.1. Statistical Multimodel Postprocessor

The proposed postprocessor uses QR to bias correct the ensemble forecasts from individual models and BMA to combine the bias-corrected forecasts. We begin by briefly revisiting the BMA technique. BMA generates an overall forecast pdf by taking a weighted average of the conditional pdfs associated with the individual model forecasts. Letting Δ be the forecasted variable, D the training data, and $M = [M_1, M_2, ..., M_k]$ the independent predictions from a total of K hydrological models, the pdf of the BMA probabilistic prediction of Δ can be expressed by the law of total probability as

$$P(\Delta|(M_1, M_2, ..., M_K)) = \sum_{k=1}^{K} P(\Delta|M_k)P(M_k|D), \tag{1}$$

where $P(\Delta|\, M_k)$ is the posterior distribution of Δ given the model prediction M_k and $P(M_k|\, D)$ is the posterior probability of model M_k being the best one given the training data D. $P(M_k|\, D)$ reflects the performance of model M_k in predicting the forecast variable during the training period.

The posterior model probabilities are nonnegative and add up to one (Raftery et al., 2005), such that

$$\sum_{k=1}^{K} P(M_k|D) = 1. \tag{2}$$

Thus, $P(M_k|\, D)$ can be viewed as the model weight, w_k, reflecting an individual model's relative contribution to predictive skill over the training period. The BMA pdf is therefore a weighted average of the conditional pdfs associated with each of the individual model forecasts, weighted by their posterior model probabilities. Since model predictions are time variant, letting t be the forecast lead time, equation (1) can be written as

$$P\left(\Delta^t|\left(M_1^t, M_2^t, ..., M_K^t\right)\right) = \sum_{k=1}^{K} w_k^t P\left(\Delta^t|M_k^t\right). \tag{3}$$

The efficient application of BMA requires bias correcting the ensemble forecasts from the individual models and optimizing their weights w_k^t (Raftery et al., 2005). We used QR to bias correct the forecasts. QR has several advantages as compared to the linear regression bias correction used in the original BMA approach (Raftery et al., 2005). It does not make any prior assumptions regarding the shape of the distribution, and, since QR results in conditional quantiles rather than conditional means, QR is less sensitive to the tail behavior of the streamflow data and, consequently, more robust to outliers.

To implement QR, the bias-corrected ensemble forecasts from each model k and forecast lead time t, $f_{k,}^{t}\tau$, are determined using

$$f_{k,\tau}^{t} = \overline{f}_{k}^{t} + \widehat{\xi}_{k,\tau}^{t}, \tag{4}$$

where $\overline{f}_{k}^{t}$ is the ensemble mean forecast of model k at time t and $\widehat{\xi}_{k,\tau}^{t}$ is the error estimate at the quantile interval τ defined as

$$\widehat{\xi}_{k,\tau}^{t} = a_{k,}^{t}\tau + b_{k,}^{t}\tau\overline{f}_{k}^{t}. \tag{5}$$

In equation (5), $a_{k,\tau}^{t}$ and $b_{k,\tau}^{t}$ are the regression parameters for model k and quantile interval τ at time t. The parameters associated with each model are determined separately by minimizing the sum of the residuals from a training data set as follows:

$$\arg\min_{f\in\mathbb{R}} \sum_{j=1}^{J} \Gamma_{\tau}^{t}\left[\xi_{\tau,j}^{t} - \widehat{\xi}_{\tau,j}^{t}\left(j,\overline{f}_{j}\right)\right]. \tag{6}$$

$\xi_{\tau,j}^{t}$ and $\overline{f_{j}}$ are the jth paired samples from a total of J samples, $\widehat{\xi}_{\tau,j}^{t}$ is computed as the observed flow minus the forecasted one at time t, Γ_{τ}^{t} is the QR function for the τth quantile at time t defined as

$$\Gamma_{\tau}^{t}\left(\Psi_{j}^{t}\right) = \begin{Bmatrix} (\tau-1)\,\Psi_{j}^{t} & \text{if } \Psi_{j}^{t}\le 0 \\ \tau\,\Psi_{j}^{t} & \text{if } \Psi_{j}^{t}>0 \end{Bmatrix}, \tag{7}$$

and Ψ_{j}^{t} is the residual term computed as the difference between $\xi_{\tau,j}^{t}$ and $\widehat{\xi}_{\tau,j}^{t}\left(j,\overline{f_{j}}\right)$ for any quantile $\tau \in [0, 1]$. The resulting minimization problem in equation (6) is solved using linear programming via the interior point method (Koenker, 2005). Note that the τ values were chosen to cover the domain [0, 1] sufficiently well, so that the lead time-specific error estimate in equation (5) is a continuous distribution. Specifically, the number of τ values were based on the number of ensemble members required by a particular forecasting experiment and were chosen to vary uniformly between 0.06 and 0.96.

After bias correcting the single-model forecasts using equations (4)–(7), the posterior distribution of each model is assumed Gaussian. Thus, before implementing equation (3), both the observations and bias-corrected forecasts are transformed into standard normal deviates using the normal quantile transformation (NQT; Krzysztofowicz, 1997). The NQT matches the empirical cumulative distribution function (cdf) of the marginal distribution to the standard normal distribution such that

$$f_{k,NQT}^{t} = G^{-1}\left(cdf\left(f_{k}^{t}\right)\right), \tag{8}$$

where $cdf(.)$ is the cdf of the bias-corrected forecasts from model k at time t, f_{k}^{t}; G is the standard normal distribution and G^{-1} its inverse; and $f_{k,NQT}^{t}$ is the transformed, bias-corrected forecasts from model k at time t. When applying the NQT, extrapolation is used to model the tails of the forecast distribution for those cases where a sampled data point in normal space falls outside the range of the training data maxima or minima. For the upper tail, a hyperbolic distribution (Journel & Huijbregts, 1978) is used while linear extrapolation is used for the lower tail.

Lastly, to determine the BMA probabilistic prediction in equation (3), the weight w_{k}^{t} and variance $\sigma_{k}^{2,t}$ of model k at the forecast lead time t are estimated using the log likelihood function. Note that $\sigma_{k}^{2,t}$ is the variance associated with the Gaussian posterior distribution of model k. Setting the parameter vector $\theta = \left\{w_{k}^{t}, \sigma_{k}^{2,t}, k = 1, 2, ..., K\right\}$, the log likelihood function of θ at the forecast lead time t is approximated as

$$l(\theta) = \log\left(\sum_{k=1}^{K} w_{k}^{t} g\left(\Delta_{NQT}^{t} | f_{k,NQT}^{t}\right)\right), \tag{9}$$

where $g(.)$ denotes a Gaussian pdf and Δ_{NQT}^{t} is the forecasted variable in Gaussian space. Because of the high dimensionality of this problem, the log likelihood function typically cannot be maximized analytically. Thus, the maximum likelihood estimates of θ are determined using the expectation maximization (EM)

optimization algorithm (Bilmes, 1998). The steps required to implement the EM algorithm are provided in Appendix A. Finally, discrete ensembles are sampled from the postprocessed predictive distribution using the equidistant quantiles sampling approach (Schefzik et al., 2013).

Our proposed QR-BMA approach consists of implementing equations (3)–(9). To apply QR-BMA, we used a leave-one-out approach where part of the forecast data set was used to train QR-BMA and the rest to verify the multimodel ensemble forecasts. We applied QR-BMA at each forecast lead time t of interest for selected forecast locations. As part of our forecast experiments, we generated both single-model and multimodel ensemble forecasts. The single-model streamflow forecasts were generated from GEFSRv2, while the multimodel forecasts were generated using the QR-BMA technique to optimally combine the single-model forecasts. The single-model forecasts were postprocessed using QR, following the same leave-one-out approach used with QR-BMA. Note that QR-BMA was applied here independently at each lead time; thus, it is suitable for generating forecasts when predictions are needed for a single time.

2.2. Measures of Forecast Skill

2.2.1. Conditional Mutual Information

CMI is used as a measure of skill improvement following the approach by DelSole et al. (2014). The approach allows to distinguish whether multimodel skill improvements are dominated by model diversity (i.e., additional information provided by the different models) or increased ensemble size (i.e., the addition of new ensemble members). To present the *CMI* measure, we first introduce three related information-theoretic measures: entropy, conditional entropy, and mutual information (*MI*).

In the case of a continuous random variable (e.g., the streamflow forecasts F with pdf $P(f)$, where uppercase is used to denote the random variable and lowercase its realizations), the amount of average information required to describe F is given by the entropy $\mathrm{H}(F)$ defined as

$$\mathrm{H}(F) = -\int P(f)\ln P(f)df. \tag{10}$$

Entropy measures the uncertainty of F (Cover & Thomas, 1991). The entropy of a random variable conditional upon the knowledge of another can be defined by the conditional entropy. The conditional entropy between the streamflow observations O and forecasts F can be calculated using the chain rule:

$$\mathrm{H}(O|F) = \mathrm{H}(O,F) - \mathrm{H}(F). \tag{11}$$

With equations (10) and (11), the *MI* between the streamflow observations and the forecasts, $MI(O;F)$, is given by (Cover & Thomas, 1991)

$$\begin{aligned} MI(O;F) &= \mathrm{H}(O) + \mathrm{H}(F) - \mathrm{H}(O,F) \\ &= \iint P(o,f)\log\left[\frac{P(o,f)}{P(o)P(f)}\right] do\, df, \end{aligned} \tag{12}$$

where $P(o,f)$ is the joint pdf of O and F, with marginal pdfs $P(o)$ and $P(f)$, respectively. *MI* is an elegant and powerful measure to quantify the amount of information that one random variable contains about another random variable. It is nonnegative and equal to zero if and only if O and F are independent from each other. *MI* has several important benefits. It is a domain-independent measure such that the information provided is relatively insensitive to the size of data sets and outliers, unaffected by systematic errors, and invariant to any nonlinear transformations of the variables (Cover & Thomas, 1991; Kinney & Atwal, 2014).

In the case of multimodel combinations, where F_1 represents the single-model ensemble mean and F_2 represents the multimodel mean of the remaining models, the *CMI* between O and F_2, conditioning out F_1, is given by

$$CMI(O;F_2|F_1) = MI(O;(F_1,F_2)) - MI(O;F_1), \tag{13}$$

where the mutual information $MI(O;(F_1,F_2))$ measures the degree of dependence between the observation and the joint variability of the forecasts F_1 and F_2. According to equation (13), *CMI* quantifies the additional

decrease in uncertainty due to adding a single-model forecast to the multimodel forecast mean of the other models. When the distributions are Gaussian, the *CMI* reduces to a simple function of partial correlation as follows (Sedghi & Jonckheere, 2014):

$$CMI(O;F_2|F_1) = -\frac{1}{2}\log\left(1-\rho^2_{O2|1}\right), \tag{14}$$

where $\rho_{O2\,|\,1}$ denotes the partial correlation between O and F_2 conditioned on F_1. The partial correlation is related to the pairwise correlations by (Abdi, 2007)

$$\rho_{O2|1} = \frac{\rho_{O2}-\rho_{O1}\rho_{12}}{\sqrt{\left(1-\rho^2_{O1}\right)\left(1-\rho^2_{12}\right)}}, \tag{15}$$

where ρ_{O1} and ρ_{O2} are the correlation skills of F_1 and F_2, respectively, and ρ_{12} is the correlation between F_1 and F_2. Hereafter, the subscript 1 denotes single-model forecasts, and the subscript 2 denotes either single-model forecasts or multimodel forecasts, depending on whether one is assessing the skill of single-model or multimodel forecasts.

To further understand any skill enhancements provided by a multimodel forecast, the streamflow forecasts and observations can be partitioned into a conditional mean, called the signal variable α, and a deviation about the conditional mean, called the noise variable β. As shown by DelSole et al. (2014), in the case that all the ensemble members are drawn from the same model and the forecasts are computed with means of ensemble size E_1 and E_2, the partial correlation in equation (15) becomes

$$\rho^{noise}_{O2|1} = \frac{\rho_{\alpha O}}{E_1}\frac{\sqrt{SNR}}{\sqrt{SNR\left(\frac{1}{E_1}+\frac{1}{E_2}\right)+\frac{1}{E_1E_2}}\sqrt{SNR\left(1-\rho^2_{\alpha O}\right)+\frac{1}{E_1}}}, \tag{16}$$

where the signal-to-noise ratio *SNR* is defined as the ratio of signal variance to noise variance and $\rho_{\alpha O}$ is the correlation between the signal variable and streamflow observation. The partial correlation in equation (16) is nonzero when a predictable signal exists (i.e., $SNR \neq 0$), forecast skill exists ($\rho_{\alpha O} \neq 0$), and the ensemble sizes are finite. To the extent that forecast skill exceeds predictability skill,

$$|\rho_{\alpha O}| \leq \sqrt{\frac{SNR}{SNR+1}}. \tag{17}$$

Equation (17) implies that an upper bound on $\rho_{\alpha O}$ results in an upper bound on the partial correlation in equation (16). Thus, an upper bound on the skill improvement due to adding new ensemble members from the same model can be estimated by combining equations (16) and (17) and taking the limit $SNR \to \infty$,

$$\rho^{noise}_{O2|1} \leq \sqrt{\frac{E_2}{(E_1+E_2)(E_1+1)}}. \tag{18}$$

Thus, any skill enhancement measured by equation (15) that exceeds the upper bound of equation (18) is dominated by the addition of new predictable signals (DelSole et al., 2014).

We computed *CMI* using equations (14) and (15), together with the streamflow ensemble forecasts and observations. We used equation (18) to obtain an upper bound for the skill improvement due to increased ensemble size. Any improvements beyond this upper bound, we attributed to the addition of new signals or model diversity. When using equations (14), (15), and (18), the subscript 1 refers to the single-model forecasts F_1 that one is trying to improve, and the subscript 2 the multimodel forecasts F_2 or, in the case of a single-model experiment, the addition of new members from the same model. *CMI* was computed for each individual model and multimodel combination at every lead time of interest for selected forecast locations. Before computing *CMI*, both the streamflow observations and forecasts were transformed into Gaussian space using NQT.

To implement *CMI*, three different experiments were performed: (i) 9-member single model, (ii) 9-member multimodel, and (iii) 33-member multimodel. The 9-member single-model experiment consists of a 3-

member single-model forecast (F_1) combined with a 6-member ensemble from the same model (F_2). Note that this 6-member ensemble may be treated as proxy for adding members from hydrological models with very similar structures. This experiment was repeated for each of the models used. In the 9-member multimodel experiment, a 3-member single-model ensemble from one of the models (F_1) was combined with a 6-member multimodel ensemble obtained using the remaining two other models (F_2). This 6-member multimodel ensemble was generated as follows: (i) Three raw members from each of the remaining two models were randomly selected, and (ii) the selected members were combined using the QR-BMA postprocessor to generate a 6-member multimodel ensemble. Note that the number of ensemble members from each model are equal only in relation to the number of raw forecast members sampled from each model. Additionally, in both the 9-member single-model and 9-member multimodel experiments, the values of E_1 and E_2 in equation (18) are 3 and 6, respectively. The last experiment, 33-member multimodel, was the same as the 9-member multimodel experiment but using instead 33 members. That is, an 11-member single-model ensemble from one of the models (F_1) was combined with a 22-member multimodel ensemble obtained by postprocessing the remaining two other models (F_2). For the *CMI* experiments, raw single-model forecasts were used for F_1 to emulate basic operational conditions. The *CMI* values for the different experiments were computed by first randomly selecting raw ensemble members from each hydrological model. This process of randomly selecting raw forecasts from each model was repeated several times for each *CMI* value, so that the reported *CMI* value is the average from multiple realizations.

Additionally, we estimated *CMI* in streamflow space using the approach discussed by Meyer (2008). The approach relies on the Miller-Madow asymptotic bias-corrected empirical estimator for entropy estimation (Meyer, 2008; Miller, 1955) and an equal frequency binning algorithm for data discretization (Meyer, 2008). This approach does not require transforming streamflow into Gaussian space but has the drawback that an exact upper bound, akin to equation (18), is not available. The *CMI* in streamflow space was computed using the same experimental conditions described before for *CMI* in Gaussian space.

2.2.2. Continuous Ranked Probability Skill Score

Besides using *CMI* to measure skill improvements, we used the mean Continuous Ranked Probability Skill Score (*CRPSS*; Hersbach, 2000) since this is a commonly used verification metric to assess the quality of ensemble forecasts (Brown et al., 2014). The *CRPSS* is derived from the Continuous Ranked Probability Skill Score (*CRPS*). The *CRPS* evaluates the overall accuracy of a probabilistic forecast by estimating the quadratic distance between the forecasts' cdf and the corresponding observations. The *CRPS* is defined as

$$CRPS = \int_{-\infty}^{\infty} [cdf(f) - \Pi(f-o)]^2 df, \tag{19}$$

where

$$\Pi(f) = \begin{cases} 0 & \text{for} f<0 \\ 1 & \text{otherwise} \end{cases}. \tag{20}$$

$\Pi(.)$ is the Heaviside step function.

To evaluate the skill of the forecasting system relative to a reference system, the associated skill score or *CRPSS* is computed as

$$CRPSS = 1 - \frac{\overline{CRPS}_m}{\overline{CRPS}_r}, \tag{21}$$

where the *CRPS* is averaged across n pairs of forecasts and observations to calculate the mean *CRPS* of the main forecast system, $\overline{CRPS}_m$, and reference forecast system, $\overline{CRPS}_r$. The *CRPSS* ranges from $[-\infty, 1]$. Positive *CRPSS* values indicate the main forecasting system has higher skill than the reference forecasting system, with 1 indicating perfect skill. In this study, we used sampled climatology as the reference forecasting system. Similar to our implementation of *CMI*, the *CRPSS* was computed for both single-model and multimodel ensemble streamflow forecasts at each lead time of interest for selected forecast locations.

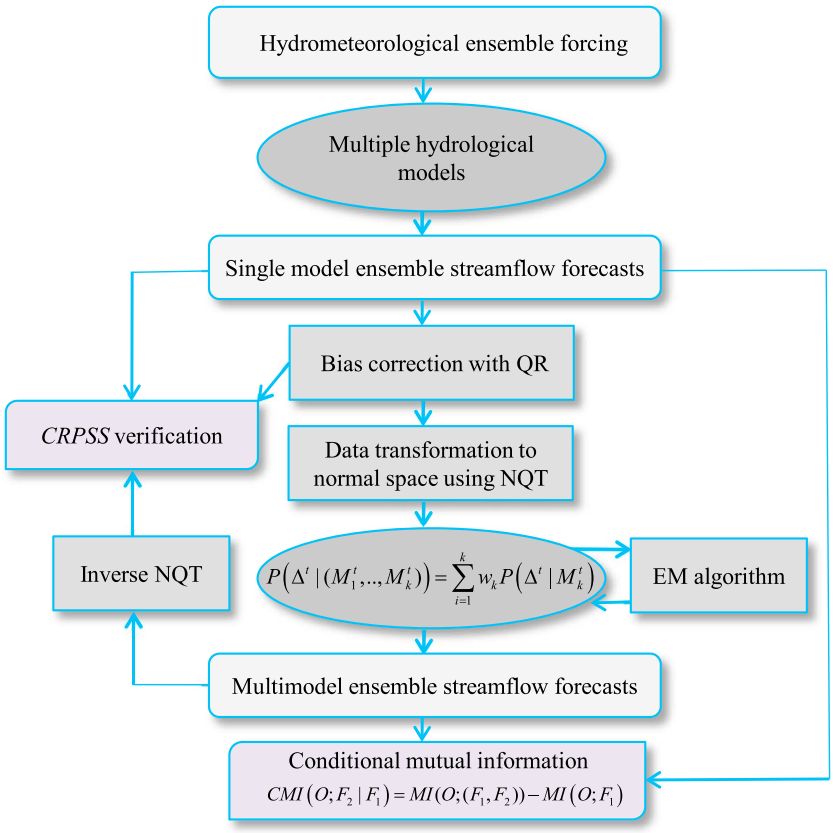

Figure 1. Diagrammatic representation of the proposed multimodel forecasting approach. The approach starts with the hydrometeorological ensemble forcing. The forcing is used to drive different hydrological models to generate single-model ensemble streamflow forecasts. The single-model forecasts are subsequently bias corrected, transformed to Gaussian space, and combined using BMA to generate multimodel ensemble streamflow forecasts. Lastly, both the single-model and multimodel forecasts are verified using the *CRPSS* and *CMI*. QR = quantile regression; *CRPSS* = Continuous Ranked Probability Skill Score; NQT = normal quantile transformation; EM = expectation maximization; *CMI* = conditional mutual information; BMA = Bayesian model averaging.

Confidence intervals for the *CRPSS* were determined using the stationary block bootstrap technique (Politis & Romano, 1994). Note that the *CRPSS* represents a quantitative measure of the overall forecast skill relative to the reference system (i.e., sampled climatology), whereas the *CMI* represents the skill improvement or enhancement provided by the multimodel forecasts. Thus, the *CMI* and *CRPSS* are not directly comparable against each other. Our proposed multimodel forecasting approach is summarized in Figure 1.

3. Experimental Setup

3.1. Study Area

The North Branch Susquehanna River (NBSR) basin in the U.S. Middle Atlantic Region was selected as the study area (Figure 2; Nelson, 1966). Severe weather and flooding hazards are an important concern in the NBSR, for example, the City of Binghamton, New York, has been affected by multiple damaging flood events over recent years (Gitro et al., 2014; Jessup & DeGaetano, 2008). In the NBSR, four different U.S. Geological Survey (USGS) daily gauge stations were selected as the forecast locations (Figure 2). The selected locations are the Ostelic River at Cincinnatus (USGS gauge 01510000), Chenango River at Chenango Forks (USGS gauge 01512500), Susquehanna River at Conklin (USGS gauge 01503000), and Susquehanna River at Waverly (USGS gauge 01515000). These forecast locations represent a system of nested subbasins with drainage areas ranging from ~381 to 12,362 km^2. A summary of the main characteristics of the selected gauge locations is provided in Table 1.

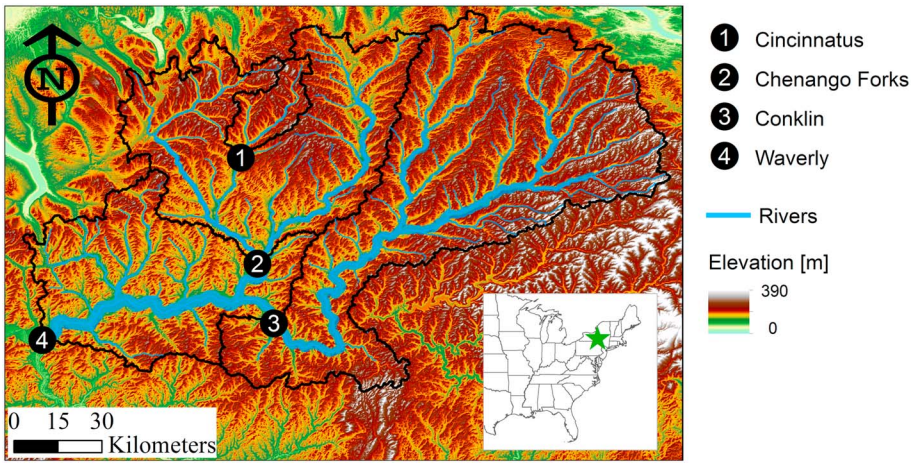

Figure 2. Map of the study area showing the terrain elevations, stream network, and the location of the selected gauged stations. The inset map shows the approximate location of the study area in the United States.

3.2. Data Sets

3.2.1. Meteorological Forecasts

NOAA's latest global, medium-range ensemble reforecast data set, the Global Ensemble Forecast System Reforecast version 2 (GEFSRv2; https://www.esrl.noaa.gov/psd/forecasts/reforecast2/), was used as the forecast forcing. The following GEFSRv2 variables were used: precipitation, specific humidity, surface pressure, downward shortwave and longwave radiation, u-v components of wind speed, and near-surface air temperature. The GEFSRv2 is an 11-member ensemble forecast generated by stochastically perturbing the initial NWP model conditions using the ensemble transform technique with rescaling (Wei et al., 2008). The GEFSRv2 data are based on the same atmospheric model and initial conditions as the version 9.0.1 of the NOAA's Global Ensemble Forecast System and run at T254 L42 (0.50° Gaussian grid spacing or ~55 km) resolution up to day 8. The 11-member reforecasts are generated every day at 00 Coordinated Universal Time. The GEFSRv2 forecast cycle consists of 3-hourly accumulations for the first 3 days and 6-hourly accumulations after that. To generate the ensemble streamflow forecasts, we used the first 7 days of GEFSRv2 data for the period 2004–2009. The GEFSRv2 data were bilinearly interpolated onto the regularly spaced grid required by the hydrological models. Table 2 summarizes key information about the GEFSRv2 data set. Additional details about the GEFSRv2 can be found elsewhere (Hamill et al., 2013).

3.2.2. Hydrometeorological Observations

Four main observational data sets were used: multisensor precipitation estimates (MPEs), gridded near-surface air temperature, phase 2 of the North American Land Data Assimilation System (NLDAS-2; https://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php), and daily streamflow. These observational data sets were used to calibrate and verify the hydrological models, perform the hydrological model simulations,

Table 1
Characteristics of the Selected Gauged Locations

Location of outlet	Cincinnatus, New York	Chenango Forks, New York	Conklin, New York	Waverly, New York
NWS id	CINN6	CNON6	CKLN6	WVYN6
USGS id	01510000	01512500	01503000	01515000
Area (km^2)	381	3,841	5,781	1,2362
Outlet latitude (North)	42°32′28″	42°13′05″	42°02′07″	41°59′05″
Outlet longitude (West)	75°53′59″	75°50′54″	75°48′11″	76°30′04″
Minimum daily flow[a] (m^3/s)	0.31 (0.11)	4.05 (2.49)	6.80 (5.32)	13.08 (6.71)
Maximum daily flow[a] (m^3/s)	172.73 (273.54)	1,248.77 (1,401.68)	2,041.64 (2,174.734)	4,417.42 (4,417.42)
Mean daily flow[a] (m^3/s)	8.89 (9.17)	82.36 (81.66)	122.93 (121.99)	277.35 (215.01)

[a]The number in parenthesis is the historical (based on the entire available record, as opposed to the period 2004–2009 used in this study) daily minimum, maximum, or mean recorded flow.

Table 2
Summary and Main Characteristics of the Data Sets Used in This Study

Data set	Source	Horizontal resolution (km^2)	Temporal resolution (hour)	Variables
Meteorological forecasts				
GEFSRv2	NCEP	~55 × 55 (0.5° × 0.5°)	3 (days 1–3) and 6 (days 4–7) hourly accumulations	Precipitation, near-surface temperature, specific humidity, surface pressure, downward shortwave and longwave radiation, and u-v components of wind speed
Hydrometeorological observations				
NLDAS-2	NASA	~13 × 13 (0.125° × 0.125°)	Hourly	Near-surface temperature, specific humidity, surface pressure, downward longwave and shortwave radiation, and u-v components of wind speed
MPEs	MARFC	~4 × 4	Hourly	Gridded precipitation
Temperature	MARFC	~4 × 4	Hourly	Gridded temperature
Gauge discharge	USGS	—	Hourly	Streamflow

and obtain initial conditions for the forecasting runs for the period 2004–2009. Both the MPEs and gridded near-surface air temperature data at 4 × 4 km^2 were obtained from the MARFC. Similar to the NCEP stage IV MPEs (Moore et al., 2014; Prat & Nelson, 2015), the MARFC MPE product combines radar estimated precipitation with in situ gauge measurements to create a continuous time series of hourly, gridded precipitation observations. The gridded near-surface air temperature data were produced by the MARFC using multiple observation networks, including the meteorological terminal aviation routine weather report (METAR), USGS stations, and National Weather Service Cooperative Observer Program (Siddique & Mejia, 2017). Additionally, we used NLDAS-2 data for near-surface air temperature, specific humidity, surface pressure, downward longwave and shortwave radiation, and u-v components of wind speed. The spatial resolution of the NLDAS-2 data is 1/8th-degree grid spacing while the temporal resolution is hourly. Further details about the NLDAS-2 data can be found elsewhere (Mitchell et al., 2004). To calibrate the hydrological models and verify the streamflow simulations and forecasts, daily streamflow observations for the selected gauged locations were obtained from the USGS. In total, 6 years (2004–2009) of hydrometeorological observations were used. Table 2 summarizes the observational data sets.

3.3. Hydrological Models

To generate the multimodel forecasts, we used the following three hydrological models: Antecedent Precipitation Index (API)-Continuous (Moreda et al., 2006), NOAA's Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM; Koren et al., 2004), and the Weather Research and Forecasting Hydrological (WRF-Hydro) modeling system (Gochis et al., 2015). We selected these three hydrological models because they are relevant to operational forecasting in the United States and represent varying levels of model structural complexity as well as different spatial resolutions and parameterizations. The selected models collectively represent a sufficiently diverse set of models favorable for multimodel forecasting. The description of each model and the details about the configuration, calibration, and performance of the models in simulation mode are provided in Text S1 in the supporting information. The parameters selected for calibration, and the parameters' feasible ranges and calibrated values for the HL-RDHM and WRF-Hydro models are summarized in Table S1.

The models were used to simulate and forecast flows over the entire period of analysis (years 2004–2009) at the selected gauge locations (Figure 1) but were verified for the warm season only (May–October). We focused on the warm season because flood events are more prevalent in our study area during these months. The simulated flows were obtained by forcing the hydrological models with meteorological observations. The streamflow simulations were verified against daily observed flows for the entire period of analysis, warm season only (years 2004–2009, May–October). The HL-RDHM simulations were performed for the period 2004–2009, with the year 2003 used as warm-up. To calibrate HL-RDHM, we first manually adjusted the a priori parameter fields through a multiplying factor; once the manual changes did not yield noticeable

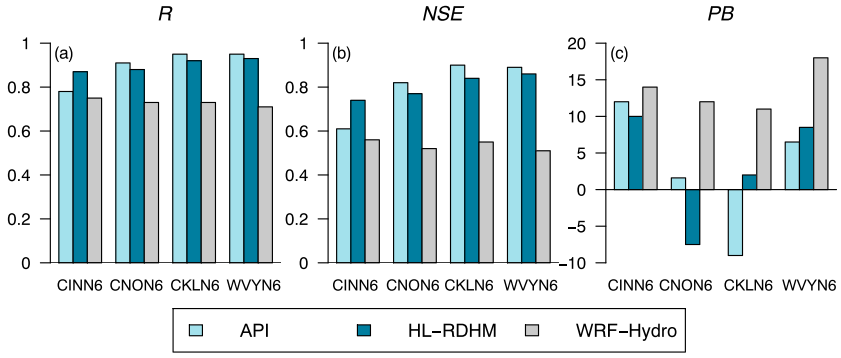

Figure 3. Performance of the hydrological models in simulation mode over the entire period of analysis (2004–2009, May–October): (a) Pearson's correlation coefficient, *R*; (b) Nash-Sutcliffe efficiency, *NSE*; and (c) percent bias, *PB*, between the daily simulated and observed flows. API = Antecedent Precipitation Index; HL-RDHM = Hydrology Laboratory-Research Distributed Hydrologic Model; WRF-Hydro = Weather Research and Forecasting Hydrological modeling system.

improvements in model performance, the multiplying factors were tuned up using the stepwise line search algorithm (Kuzmin, 2009; Kuzmin et al., 2008). Out of all the HL-RDHM adjusted parameters, the most sensitive parameters were found to be the upper and lower soil zones transport and storage parameters, as well as the stream routing parameters. The WRF-Hydro simulations were performed for the period 2004–2009, with the first year used as warm-up. To calibrate WRF-Hydro, we implemented a stepwise manual adjustment approach (Yucel et al., 2015); that is, once a parameter value was calibrated, its value was kept fixed during the calibration of subsequent parameters. Out of all the adjusted parameters, the most sensitive parameters were the soil, groundwater, and runoff parameters. After manually calibrating the WRF-Hydro parameters, the most sensitive parameter values were fine tuned using an optimization algorithm, namely, dynamically dimension search (Tolson & Shoemaker, 2007). The API-Continuous model was previously calibrated by the MARFC for operational forecasting purposes using a manual approach.

Figure 3 summarizes the models' performance in simulation mode using the Pearson's correlation coefficient, *R*; Nash-Sutcliffe efficiency, *NSE*; and percent bias, *PB*, between the simulated and observed streamflows at daily resolution for the entire analysis period. The overall performance of the models was satisfactory (Figures 3a and 3b). API and HL-RDHM exhibited comparable performance while WRF-Hydro tended to underperform relative to API and HL-RDHM. The performance of the models is discussed further in section 4.

3.4. Ensemble Streamflow Forecasts

To perform our forecast experiments, we generated and verified the following three different data sets of ensemble streamflow forecasts: (i) raw single model, (ii) postprocessed single model, and (iii) multimodel. The raw single-model data set consisted of ensemble streamflow forecasts from each hydrological model without postprocessing. The postprocessed single-model data set was generated by using QR to postprocess the raw ensemble streamflow forecasts from each hydrological model. Lastly, the multimodel data set was generated by optimally combining the ensemble forecasts from the different hydrological models using QR-BMA. As part of the multimodel data set, we also generated an equal-weight multimodel forecast by using the same weight, $1/K$, to combine the models rather than the optimal weights from QR-BMA. Additionally, for both the single-model and multimodel forecast data sets, we varied the number of ensemble members used (9 to 33 members) to perform different experiments.

All the forecast data sets were verified across lead times of 1 to 7 days using 6 years of data (2004–2009) for the warm season only (May–October). To postprocess and verify both the single-model and multimodel ensemble streamflow forecasts, a leave-one-out approach was implemented by using 4 years of forecast data (training period) to train the postprocessor and the remaining 2 years to verify the forecasts. This was repeated until all the 6 years of forecast data were postprocessed and verified independently of the training period. The subdaily streamflow forecasts generated by the hydrological models were averaged over 24 hr to get the mean daily flow. Six-hourly streamflow forecasts were generated from API and HL-RDHM, and

3-hourly forecasts from WRF-Hydro. The mean daily ensemble streamflow forecasts were verified against mean daily streamflow observations for the selected gauged locations.

4. Results and Discussion

4.1. CRPSS Verification of the Single-Model Forecasts

4.1.1. Raw Ensemble Streamflow Forecasts

In terms of the *CRPSS* (relative to sampled climatology), the raw single-model ensemble streamflow forecasts remain skillful across lead times (1–7 days) and basins (Figures 4a–4d), with the exception of WRF-Hydro that has slightly negative *CRPSS* values at the longer lead times (6–7 days). In Figures 4a–4d, the *CRPSS* values tend overall to decline with increasing lead time, as might be expected since the weather uncertainties tend to grow and become more dominant of forecast skill as the lead time progresses (Siddique & Mejia, 2017). There is also a slight tendency for the *CRPSS* values to exhibit spatial scale dependency. The *CRPSS* values for each model tend to increase from the smallest (Figure 4a) to the largest (Figure 4d) basin across lead times. This tendency is, however, rather weak throughout all of our forecasts, and it is somewhat more apparent for the API and HL-RHDM forecasts than for the WRF-Hydro (Figures 4a–4d).

Across all lead times and basins (Figures 4a–4d), the *CRPSS* values vary approximately from −0.15 (WRF-Hydro at the day 7 lead time; Figure 4d) to 0.6 (API at the day 1 lead time; Figure 4d). Contrasting the hydrological models, the performance of API and HL-RDHM is comparable, with the exception of CNON6 (Figure 4b) where API outperforms HL-RDHM. This is due to HL-RDHM having an unusually high percent simulation bias of −14.3 for CNON6 relative to API whose simulation bias is −5.8. The performance of the models in forecasting mode tends to mimic their performance in simulation mode (Figure 3). That is, API tends to perform better than HL-RDHM, and, in turn, both of these models tend to outperform WRF-Hydro. Deviations from this tendency, however, do emerge. For example, WRF-Hydro has similar forecasting skill as HL-RDHM at the day 1 lead time in CINN6 (Figure 4a), even though in this basin HL-RDHM performs better than WRF-Hydro in simulation mode. Similarly, API performs slightly better than HL-RDHM in forecasting mode at the later lead times (>4 days) in CINN6 (Figure 4a), but HL-RDHM shows better performance in simulation mode. Thus, the results obtained here in simulation mode do not always translate to similar performance in forecasting mode. This is not surprising given the nonlinear relationship between hydrological processes and weather forcings. It reinforces the need to verify hydrological models in both simulation and forecasting mode to gain a more complete understanding of model behavior.

The underperformance of WRF-Hydro, in both simulation and forecasting mode, in comparison to API and HL-RDHM may be due to several factors. One factor is likely to be the additional model complexity of WRF-Hydro. That is, WRF-Hydro requires more forcing inputs and parameters to be specified than the other two models. For example, in terms of forcings, HL-RDHM requires only precipitation and near-surface air temperature to be specified, whereas WRF-Hydro requires seven different forcings. It is possible that any biases in the NLDAS-2 or GEFSRv2 forcings used here to configure the WRF-Hydro simulations and forecasts, respectively, could be affecting its performance. However, we evaluated (results not shown) for the WRF-Hydro streamflow forecasts the effect of each individual forcing on the *CRPSS* values and found that precipitation was the most dominant forcing. At least in forecasting mode, the additional forcings used by WRF-Hydro do not seem to have a strong influence on its forecast skill. The relatively low performance of the WRF-Hydro could also be due to restrictions in its ability to represent physical processes because of a priori constraints in model parameter values, which neglect the large uncertainty in parameter estimates and large impact that parameters have on model predictions.

The determination of model parameter values for the WRF-Hydro is another factor that is likely affecting its performance. Although we calibrated selected WRF-Hydro parameter values (see Table S1), both manually and numerically, there is generally less community knowledge about and experience with WRF-Hydro than API and HL-RDHM. The latter two have been around for much longer (e.g., Anderson et al., 2006; Koren et al., 2004; Moreda et al., 2006; Reed et al., 2004) than WRF-Hydro. In the future, a more in-depth sensitivity analysis of the WRF-Hydro model parameters could be beneficial. Nonetheless, the performance of WRF-Hydro in this study is comparable to those previously reported in the literature (Givati et al., 2016; Kerandi et al., 2017; Naabil et al., 2017; Salas et al., 2018; Silver et al., 2017; Yucel et al., 2015).

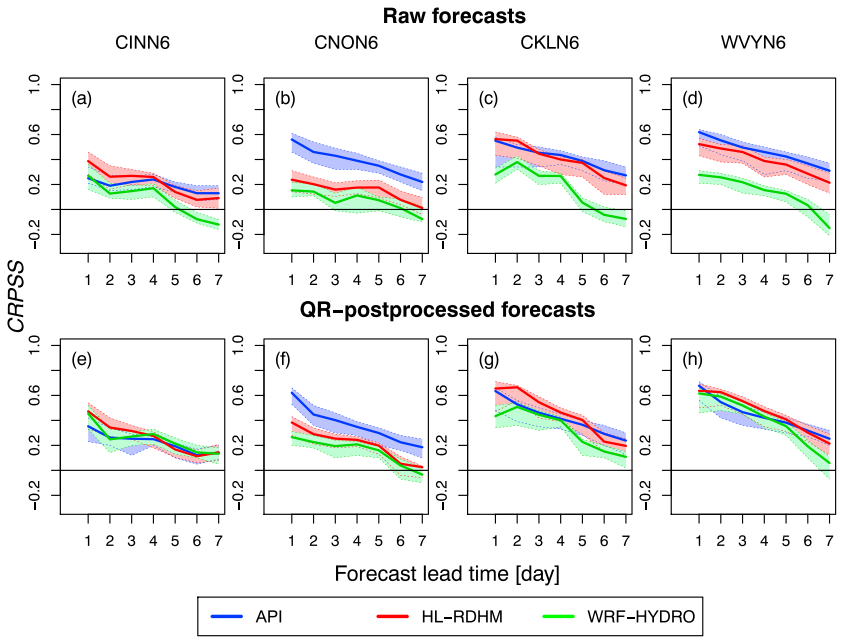

Figure 4. *CRPSS* (relative to sampled climatology) of the (a–d) raw and (e–h) QR-postprocessed singlemodel ensemble streamflow forecasts versus the forecast lead. The *CRPSS* are shown for the four selected basins. *CRPSS* = Continuous Ranked Probability Skill Score; QR = quantile regression; API = Antecedent Precipitation Index; HL-RDHM = Hydrology Laboratory-Research Distributed Hydrologic Model; WRF-Hydro = Weather Research and Forecasting Hydrological modeling system.

4.1.2. Postprocessed (Single-Model) Ensemble Streamflow Forecasts

We used QR to postprocess the raw single-model ensemble streamflow forecasts. Using the *CRPSS* (relative to sampled climatology) to assess the forecast skill (Figures 4e–4h), we found that the postprocessed single-model ensemble streamflow forecasts show, overall, skill improvements relative to the raw forecasts. The relative improvements are more noticeable for the WRF-Hydro. For example, at WVYN6 (Figure 4d), the raw WRF-Hydro forecasts have a *CRPSS* value of ~0.27 at the day 1 lead time, and that value increases to ~0.6 after postprocessing (Figure 4h). However, since the hydrological models are calibrated with data sets used for cross-validating the postprocessor, the absolute *CRPSS* for the postprocessed forecasts are not representative of real-time conditions.

Interestingly, the *CRPSS* values for the postprocessed single-model forecasts reveal that after postprocessing, the models have comparable skill across lead times and basins (Figures 4e–4h), perhaps with the exception of CNON6 (Figure 4f) where API tends to outperform the other models. This indicates that the streamflow forecasts are influenced by systematic biases and, in this case, those biases are stronger in WRF-Hydro than in the other models. Such streamflow forecast biases result from the combined effect of biases in the weather forcings and hydrological models. In regards to the former, precipitation forecasts from the GEFSRv2 are characterized by an underforecasting bias in our study region (Sharma et al., 2017; Siddique et al., 2015), particularly at the longer lead times. This underforecasting bias affects all of our hydrological model forecasts, so it is unlikely to be the cause of the strong biases seen in the WRF-Hydro forecasts.

Hydrological model biases appear to have a strong effect on the performance of WRF-Hydro, given the relatively mild skill gains from postprocessing for the API and HL-RHDM models and the larger gains for WRF-Hydro (Figures 4e–4h). Nonetheless, the QR postprocessor is able in this case to handle those biases. This suggests that models with simple structure (e.g., API, which is spatially lumped and has fewer parameters) may benefit less from postprocessing while models with complex structure (e.g., WRF-Hydro, which is spatially distributed and has more parameters) may be good candidates for postprocessing. It is also possible that systematic biases in the WRF-Hydro could be reduced through improved parameter sensitivity analysis and calibration, as opposed to statistical postprocessing.

Another interesting outcome from the postprocessed single-model results is that the ranking of the models, in terms of the *CRPSS*, varies depending on the lead time and basin. For example, both HL-RDHM and WRF-Hydro tend to slightly outperform API at the day 1 lead time in Figure 4e, but API outperforms both models at the later lead times (>6 days) in Figures 4f–4h. This is important because it indicates that there is no single model that consistently outperforms the other models. In other words, it is not possible, at least in terms of the *CRPSS*, to choose one model as the best in all cases. This suggests that it may be possible to maximize forecast skill across lead times and basins by optimally combining the outputs from the different models, as opposed to relying on a single model. It shows that multimodel forecasting may be a viable option to enhance streamflow predictions.

4.2. CRPSS Verification of the Multimodel Forecasts

We now examine with the *CRPSS* the ability of multimodel forecasts to improve streamflow predictions. For this, the *CRPSS* is again plotted against the forecast lead time for the selected basins (Figure 5). In Figure 5, the following three different multimodel forecasting experiments are shown: (i) equal weight, (ii) 9 members, and (iii) 33 members. For the equal-weight experiment, the same weight, $1/K$, was used to combine the predictive distribution of the streamflow forecasts from each hydrological model. That is, instead of using the optimal weights from QR-BMA, the same weight was used to form a 9-member multimodel forecast. For the 9-member and 33-member experiments, we used 3 and 11 raw members per model, respectively, to obtain a multimodel forecast with QR-BMA; QR-BMA was used to optimize the weights. Additionally, the reference system used to compute the *CRPSS* values in Figure 5 consists of the postprocessed ensemble streamflow forecasts from API, as opposed to sampled climatology. We selected API as the reference system since this is currently the regional operational model being used to generate streamflow forecasts in our study area.

We found that the 33-member multimodel forecasts result in higher *CRPSS* values than API across lead times and basins (Figure 5). The 9-member multimodel forecasts perform similarly to the 33-member forecasts, but in a few cases (e.g., Figure 5c at the day 5 lead time) the 9-member forecasts result in lower (negative) *CRPSS* values than API. The equal-weight experiment is only able to improve the *CRPSS* values at the initial lead times (<3 or 4 days; Figure 5), while at the later lead times its *CRPSS* values are lower than API. CNON6 offers an interesting case to further compare the single-model and multimodel forecasts. In the single-model forecasts for CNON6 (Figure 4f), API tends to clearly outperform the other models. Despite the better performance of API alone, the multimodel forecasts are still able to improve the skill for CNON6 relative to the performance of API, with the largest improvement being ~0.16 at the day 7 lead time for the 33-member experiment.

The BMA weights associated with the multimodel forecasts (see Table S2) tend to reflect the performance of the postprocessed forecasts for the individual models in Figure 4. For example, the API at CNON6 consistently gets a higher weight than the other models, particularly at the longer lead times, while WRF-Hydro at CNON6, CKLN6, and WVYN6 has relatively low BMA weights at the later lead times. Additionally, the weights show that even when the performance of one of the models is dominant, the remaining models may still contribute to improving the multimodel forecasts. This is the case for CNON6 at the later lead times (e.g., days 6 and 7 in Table S2), where despite the higher weights for API, the HL-RDHM and WRF-Hydro are still assigned some weight.

In sum, the multimodel forecasts reveal skill improvements relative to API, which may be considered here the best performing model in terms of the overall simulation and raw forecasts results; the optimal weights from QR-BMA result in more skillful multimodel forecasts than using equal weights, particularly at the later lead times (>3 days); and increasing the ensemble size of the multimodel forecasts results in relatively mild skill gains. We also computed reliability diagrams, as determined by Brown et al. (2014), for the single-model and 9-member multimodel forecasts (see Figures S2 and S3). The reliability diagrams show that the multimodel forecasts tend, for the most part, to display better reliability than the single-model forecasts.

Several studies have investigated the source of improvements (skill gains) from multimodel forecasts (Hagedorn et al., 2012; Weigel et al., 2008, 2009). Those studies have found that multimodel forecasts can improve predictions by error cancelation and correcting deficiencies (underdispersion) in the ensemble spread of the single models. These sources of skill gain appear to be mainly statistical. This way of

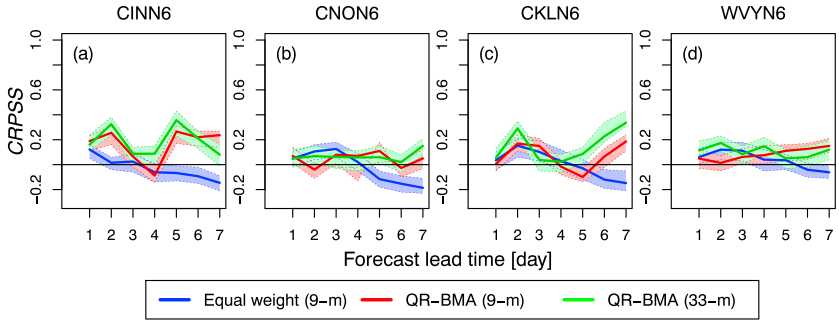

Figure 5. *CRPSS* of the multimodel ensemble streamflow forecasts versus the forecast lead time for (a) CINN6, (b) CNON6, (c) CKLN6, and (d) WVYN6. The *CRPSS* is plotted with reference to the QR-postprocessed API forecasts. Three different experiments are shown: equal weight (9 members), QR-BMA (9 members), and QR-BMA (33 members). The equal-weight experiment uses the same weight to combine the predictive distribution of the streamflow forecasts from each hydrological model. The 9-member and 33-member experiments use 3 and 11 members per model, respectively, to obtain a multimodel forecast with optimal weights using QR-BMA. *CRPSS* = Continuous Ranked Probability Skill Score; QR-BMA = quantile regression-Bayesian model averaging.

understanding the benefits of multimodel forecasts does not consider whether a particular model contributes additional information to the forecasts. Considering the latter is important to be able to justify adding any new models to an existing forecasting system. Another way to assess the source of improvements from multimodel forecasts that accounts for the contribution of model information, signal as opposed to noise, is through *CMI*, which we do next.

4.3. Skill Assessment Using CMI

We used *CMI* to determine whether the skill improvements from the multimodel forecasts are dominated by model diversity or increased ensemble size alone. To this end, *CMI* was computed using equations (14) and (15), together with the ensemble mean forecast, at lead times of 1–7 days for the selected basins (Figure 6). In Figure 6, the following three different experiments are shown: (i) 9-member single model (Figures 6a–6c), (ii) 9-member multimodel (Figures 6d–6f), and (iii) 33-member multimodel (Figures 6g–6i). The experiments are described in subsection 2.2.1.

For the first experiment, we used equations (14) and (18) to obtain a theoretical upper bound for *CMI*. This theoretical bound represents the potential skill gain from the ensemble size alone. We found that the theoretical bound is in this case equal to 0.090. Figures 6a–6c show that indeed the empirical *CMI* values for the 9-mmeber single-model forecasts tend to be less than or around 0.090 for all three models across lead times and basins. The 9-member single-model *CMI* values tend to be greater for API than HL-RDHM and WRF-Hydro. This indicates that the less complex model, API, is able to maximize the skill gains from the ensemble size alone. For example, in terms of the *CRPSS*, the raw single-model forecasts from API and HL-RDHM have comparable skill in the case of CKLN6 (Figure 4c) and WVYN6 (Figure 4d). In contrast, the 9-member single-model *CMI* values tend to be greater for API than HL-RDHM in both cases, CKLN6 and WVYN6 (Figures 6a and 6b), particularly at the longer lead times. This ability of API to maximize the benefits from ensemble size alone may be due to API being more sensitive than the other models to the weather forcing. Also, in Figures 6a–6c, the tendency is for the *CMI* values to increase some with the lead time for all the basins. This is more apparent for API and HL-RDHM than WRF-Hydro.

Contrasting the *CMI* values between the 9-member single-model (Figures 6a–6c) and 9-member multimodel (Figures 6d–6f) experiment, it is apparent that the multimodel forecasts have substantially greater *CMI* values than the single-model forecasts across lead times and basins. This indicates that any of the single-model forecasts (API, HL-RDHM, or WRF-Hydro) can be improved by combining them with forecasts from the other models. Indeed, this improvement is dominated by model diversity rather than increased ensemble size alone. Although the multimodel forecasts show skill gains at all the lead times, the tendency is for the *CMI* values to increase with the lead time, suggesting that the multimodel forecasts may be particularly useful for improving medium-range streamflow forecasts.

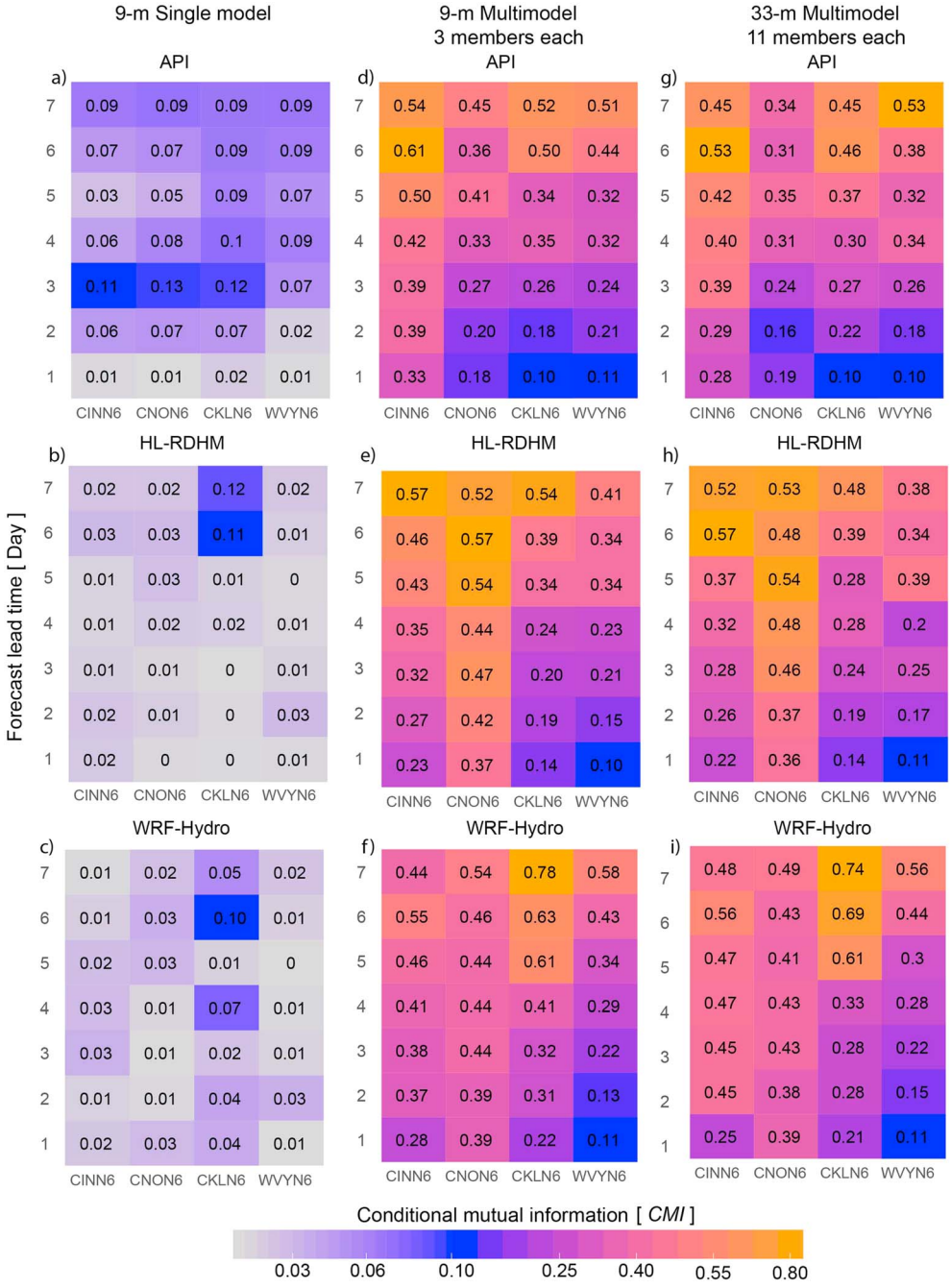

Figure 6. *CMI* of the ensemble streamflow forecasts versus both the basin and forecast lead time for three different experiments: (a–c) 9-member single-model, (d–f) 9-member multimodel, and (g–i) 33-member multimodel forecasts. The 9-member single-model experiment consists of a 3-member single-model forecast from one of the hydrological models combined with a 6-member ensemble from the same model. In the 9-member multimodel experiment, a 3-member single-model ensemble forecast from one of the models is combined with a 6-member ensemble from the remaining other two models (3 raw members from each model). The last experiment, 33-member multimodel, is the same as the 9-member multimodel experiment but using instead 33 members (11 raw members from each model). The standard deviation of the *CMI* values varies from 0.02 to 0.06. *CMI* = conditional mutual information; API = Antecedent Precipitation Index; HL-RDHM = Hydrology Laboratory-Research Distributed Hydrologic Model; WRF-Hydro = Weather Research and Forecasting Hydrological modeling system.

To further examine the hypothesis that improvements in *CMI* are dominated by model diversity rather than the ensemble size alone, the *CMI* values from the 9-member multimodel experiment (Figures 6d–6f) can be compared against the values from the 33-member multimodel experiment (Figures 6g–6i). From this comparison, it is seen that the *CMI* values for these two experiments are, overall, very similar across lead times and basins. This further supports that incorporating additional information by adding new models plays an important role in enhancing the skill of the multimodel forecasts. The results in Figure 6 indicate that hydrological multimodel forecasting can be a viable approach to improve streamflow forecasts at short- and medium-range timescales. They suggest that model diversity is a relevant consideration when trying to enhance the skill of streamflow forecasts. Although this is the case here for forecast skill, one would like in the future to examine whether these results apply to other attributes of forecast quality. In particular, metrics that are more responsive to the ensemble size than the adopted *CMI* formalism, which was based on the ensemble mean, could be tried.

We also tested the effect on the *CMI* values of using postprocessed single-model forecasts, as opposed to raw forecasts. Thus, we calculated *CMI* (results not shown) for each basin and lead time using the QR postprocessed single-model forecasts, that is, the experiments in Figure 6 were repeated using the postprocessed single-model forecasts. We found that as was the case with the raw forecasts, the *CMI* values for the multimodel combinations exceeded the theoretical upper bound of 0.090 and the *CMI* values remained very similar after increasing the ensemble size, that is, between the 9-member and 33-member multimodel experiments. Thus, the ability of model diversity to enhance the skill of the streamflow forecasts is independent of whether raw or postprocessed single-model forecasts are used.

Additionally, the *CMI* values for all the different experiments in Figure 6 were recomputed (results not shown) in streamflow space using the approach by Meyer (2008). Although a theoretical upper bound is not available for this approach, the *CMI* values in streamflow space for the multimodel forecasts tended to be noticeably greater than the values for the single-model forecasts for most lead times. Moreover, differences in the *CMI* values between the 9-member and 33-member multimodel forecasts were only marginal. Thus, the results for the experiments in Figure 6 using *CMI* values computed in both real (streamflow) and Gaussian space, overall, exhibited similar trends. This is again indicative of the ability of model diversity to enhance forecast skill beyond the improvements achievable by ensemble size alone.

5. Summary and Conclusions

In this study, we generated single-model ensemble streamflow forecasts at short- to medium-range lead times (1–7 days) from three different hydrological models: API, HL-RDHM, and WRF-Hydro. These models were selected because they represent different types of hydrological models with varying structures and parameterizations. API is a spatially lumped model; HL-RDHM is a conceptual, spatially distributed hydrological model; and WRF-Hydro is a land surface model. By forcing each hydrological model with GEFSRv2 data, single-model ensemble streamflow forecasts were generated for four nested basins of the US NBSR basin over the period 2004–2009, and the warm season (May–October). The single-model forecasts were used to generate multimodel forecasts using a new statistical postprocessor, namely, QR-BMA. QR-BMA uses first QR to correct systematic biases in the single-model forecasts and, in a subsequent step, BMA to optimally combine the predictive distribution from each model. To further understand the performance and behavior of the multimodel forecasts, we performed different ensemble streamflow forecast experiments by varying the number of ensemble members, models, and weights used to create the multimodel forecasts.

From the forecast experiments performed, we found that the raw single-model ensemble streamflow forecasts from both API and HL-RHDM tended to outperform, in terms of the CRPSS, the forecasts from WRF-Hydro across lead times and basins. However, after postprocessing the raw single-model forecasts using QR, we found that the CRPSS performance of the individual models was mostly comparable across lead times and basins. In terms of the multimodel ensemble streamflow forecasts, we found that the implementation of QR-BMA tended to improve the skill of the forecasts relative to the performance of API, which can be considered here the best performing model in terms of the raw single-model forecasts. Additionally, we compared the forecasts from QR-BMA against an equal-weight experiment, where each model was assigned the same weight. We found from this experiment that the optimal-weight forecasts from QR-BMA outperform the equal-weight forecasts. The latter was particularly evident at the later lead times (> 3 days).

Lastly, we used *CMI* to distinguish the source of the improvements for the multimodel forecasts. Although the adopted *CMI* formalism does not capture all aspects of ensemble forecasts, it allows a robust analysis to decide whether the skill enhancement from multimodel forecasts is dominated by model diversity or is only due to the reduction of noise associated with the ensemble size. We found that skill enhancements across lead times and basins are largely dominated by model diversity and that increasing the ensemble size has only a small influence on the *CMI* values. This is important because it indicates that in an operational setting the combination of different hydrological models, as opposed to only increasing the ensemble size of a single model, may be an effective approach to improve forecast skill. It also highlights that there is no single model that can be considered best in all forecasting cases, instead the benefits or strengths of different models can be combined to produce the best forecast. Importantly, the benefits from using different models are, in this case, not only due to the noise reduction associated with the ensemble size but with the ability of each model to contribute additional information to the forecasts.

Appendix A: Implementation of the EM Algorithm

We describe here the steps followed to implement the EM algorithm. The description uses the variables and notation previously defined in subsection 2.1. To implement the EM algorithm, the latent variable $z_k^{t,i}$ is introduced, which has a value of 1 if the kth model ensemble is the best prediction at time step i and a value of 0 otherwise. The EM algorithm starts with an initial weight and variance for each model set to

$$w_{k,Iter-1}^{t} = \frac{1}{K}, \tag{A1}$$

and

$$\sigma_{k,Iter-1}^{2,t} = \frac{1}{K}\sum_{i=1}^{T}\frac{\sum_{k=1}^{K}\left(\Delta_{NQT}^{t,i} - f_{k,NQT}^{t,i}\right)^2}{T}, \tag{A2}$$

allowing the calculation of an initial log likelihood

$$l(\theta_{Iter-1}) = \sum_{i=1}^{T}\log\left(\sum_{k=1}^{K} w_{k,Iter-1}^{t}\, g\left(\Delta_{NQT}^{t,i} | f_{k,NQT}^{t,i}, \sigma_{k,Iter-1}^{2,t}\right)\right), \tag{A3}$$

where T is the length of the training period extending over the time steps $i \in [1, T]$. After initializing the weight and variance for each model, the EM algorithm alternates iteratively between an expectation and maximization step until a convergence criteria is satisfied. In the expectation step, the $\widehat{z}_k^{t,i}$ for each time step is estimated given the initial values of the weight and variance as

$$\widehat{z}_{k,Iter}^{t,i} = \frac{w_{k,Iter-1}^{t}\, g\left(\Delta_{NQT}^{t,i} | f_{k,NQT}^{t,i}, \sigma_{k,Iter-1}^{2,t}\right)}{\sum_{k=1}^{K} w_{k,Iter-1}^{t}\, g\left(\Delta_{NQT}^{t,i} | f_{k,NQT}^{t,i}, \sigma_{k,Iter-1}^{2,t}\right)}. \tag{A4}$$

In the subsequent maximization step, the values of the weight and variance are updated using the current estimate of $z_{k,Iter}^{t,i}$ as follows:

$$\begin{aligned} w_{k,Iter}^{t} &= \frac{1}{T}\sum_{i=1}^{T}\widehat{z}_{k,Iter}^{t,i}, \text{and} \\ \sigma_{k,Iter}^{2,t} &= \frac{\sum_{i=1}^{T}\widehat{z}_{k,Iter}^{t,i}\left(\Delta_{NQT}^{t,i} - f_{k,NQT}^{t,i}\right)^2}{\sum_{i=1}^{T}\widehat{z}_{k,Iter}^{t,i}}. \end{aligned} \tag{A5}$$

The log likelihood function in equation (A3) is then recomputed using the updated weight and variance as

$$l(\theta_{Iter}) = \sum_{i=1}^{T} \log\left(\sum_{k=1}^{K} w^{t}_{k,\,Iter}\, g\left(\Delta^{t,i}_{NQT} | f^{t,i}_{k,NQT}, \sigma^{2,t}_{k,Iter}\right)\right). \quad \text{(A6)}$$

The expectation and maximization steps are iterated until the improvement in the log likelihood is no less than some predefined tolerance, that is, $(|\,(l(\theta_{Iter}) - l(\theta_{Iter\,-\,1}))|\,) < tol$, in this case $tol = 10^{-6}$.

Acknowledgments
We are thankful to the Editor, Martyn Clark, Associate Editor, Jonathan J. Gourley, and three anonymous reviewers for their comments and suggestions, which helped to improve the overall quality of the manuscript. We acknowledge the funding support provided by the NOAA/NWS through Award NA14NWS4680012 and the computational support provided by the Institute for CyberScience at The Pennsylvania State University. Daily streamflow observation data for the selected forecast stations can be obtained from the USGS (https://waterdata.usgs.gov/nwis/). Both the multisensor precipitation estimates and gridded near-surface air temperature data were provided by the NOAA's Middle Atlantic River Forecast Center (MARFC). The Phase 2 of the North American Land Data Assimilation System (NLDAS-2) data can be obtained from the LDAS website (https://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php). Forecast forcings from the Global Ensemble Forecast System Reforecast version 2 (GEFSRv2) can be obtained from the NOAA Earth System Research Laboratory website (https://www.esrl.noaa.gov/psd/forecasts/reforecast2/).

References

Abdi, H. (2007). Part (semi partial) and partial regression coefficients. In *Encyclopedia of measurement and statistics* (pp. 736–740). New York: SAGE Publications.

Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, *43*, W01403. https://doi.org/10.1029/2005WR004745

Anderson, R. M., Koren, V. I., & Reed, S. M. (2006). Using SSURGO data to improve Sacramento Model a priori parameter estimates. *Journal of Hydrology*, *320*, 103–116.

Bastola, S., Misra, V., & Li, H. (2013). Seasonal hydrological forecasts for watersheds over the southeastern United States for the boreal summer and fall seasons. *Earth Interactions*, *17*(25), 1–22.

Becker, E., van den Dool, H., & Zhang, Q. (2014). Predictability and forecast skill in NMME. *Journal of Climate*, *27*(15), 5891–5906.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, *4*(510), 126.

Bohn, T. J., Sonessa, M. Y., & Lettenmaier, D. P. (2010). Seasonal hydrologic forecasting: Do multimodel ensemble averages always yield improvements in forecast skill? *Journal of Hydrometeorology*, *11*(6), 1358–1372. https://doi.org/10.1175/2010JHM1267.1

Bosart, L. F. (1975). SUNYA experimental results in forecasting daily temperature and precipitation. *Monthly Weather Review*, *103*(11), 1013–1020.

Brown, J. D., He, M., Regonda, S., Wu, L., Lee, H., & Seo, D.-J. (2014). Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *Journal of Hydrology*, *519*, 2847–2868.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., & Wei, M. (2005). A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Monthly Weather Review*, *133*(5), 1076–1097. https://doi.org/10.1175/MWR2905.1

Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. In *Elements of information theory* (Vol. 2, pp. 1–55). Hoboken, NJ: John Wiley & Sons, Inc.

Davolio, S., Miglietta, M. M., Diomede, T., Marsigli, C., Morgillo, A., & Moscatello, A. (2008). A meteo-hydrological prediction system based on a multi-model approach for precipitation forecasting. *Natural Hazards and Earth System Sciences*, *8*(1), 143–159.

DelSole, T. (2007). A Bayesian framework for multimodel regression. *Journal of Climate*, *20*(12), 2810–2826. https://doi.org/10.1175/JCLI4179.1

DelSole, T., Nattala, J., & Tippett, M. K. (2014). Skill improvement from increased ensemble size and model diversity. *Geophysical Research Letters*, *41*, 7331–7342. https://doi.org/10.1002/2014GL060133

DelSole, T., Yang, X., & Tippett, M. K. (2013). Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quarterly Journal of the Royal Meteorological Society*, *139*(670), 176–183. https://doi.org/10.1002/qj.1961

Du, J., DiMego, G., Tracton, S., & Zhou, B. (2003). NCEP Short-Range Ensemble Forecasting (SREF) system: Multi-IC, multi-model and multi-physics approach. In J. Cote (Ed.), Report 33, CAS/JSC Working Group Numerical Experimentation (WGNE), WMO/TD-No. 1161*Research activities in atmospheric and oceanic modelling* (Vol. 5, pp. 09–5.10). Geneva, Switzerland: WMO.

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Fraedrich, K., & Leslie, L. M. (1987). Combining predictive schemes in short-term forecasting. *Monthly Weather Review*, *115*(8), 1640–1644. https://doi.org/10.1175/1520-0493(1987)115<1640:CPSIST>2.0.CO;2

Fraedrich, K., & Smith, N. R. (1989). Combining predictive schemes in long-range forecasting. *Journal of Climate*, *2*(3), 291–294. https://doi.org/10.1175/1520-0442(1989)002<0291:CPSILR>2.0.CO;2

Fraley, C., Raftery, A. E., & Gneiting, T. (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, *138*(1), 190–202.

Fritsch, J. M., Hilliker, J., Ross, J., & Vislocky, R. L. (2000). Model consensus. *Weather and Forecasting*, *15*(5), 571–582. https://doi.org/10.1175/1520-0434(2000)015<0571:MC>2.0.CO;2

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J., & Butts, M. B. (2004). Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology*, *298*(1–4), 222–241. https://doi.org/10.1016/j.jhydrol.2004.03.037

Gitro, C. M., Evans, M. S., & Grumm, R. H. (2014). Two major heavy rain/flood events in the Mid-Atlantic: June 2006 and September 2011. *Journal of Operational Meteorology*, *2*(13), 152–168. https://doi.org/10.15191/nwajom.2014.0213

Givati, A., Gochis, D., Rummler, T., & Kunstmann, H. (2016). Comparing one-way and two-way coupled hydrometeorological forecasting systems for flood forecasting in the Mediterranean region. *Hydrology*, *3*(2), 19. https://doi.org/10.3390/hydrology3020019

Gneiting, T. A., Raftery, E., Westveld, A. H. III, & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*, 1098–1118.

Gochis, D., Yu, W., & Yates, D. (2015). The WRF-Hydro model technical description and user's guide, version 3.0, NCAR technical document (120 pp.). Boulder, CO: NCAR.

Gomez, M., Sharma, S., Reed, S., & Mejia, A. (2019). Skill of ensemble flood inundation forecasts at short-to medium-range timescales. *Journal of Hydrology*, *568*, 207–220.

Gyakum, J. R. (1986). Experiments in temperature and precipitation forecasting for Illinois. *Weather and Forecasting*, *1*(1), 77–88.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., & Palmer, T. (2012). Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *138*(668), 1814–1827.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., et al. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, *94*(10), 1553–1565. https://doi.org/10.1175/BAMS-D-12-00014.1

Hamill, T. M., & Colucci, S. J. (1997). Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, *125*(6), 1312–1327. https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570.

Hopson, T. M., & Webster, P. J. (2010). A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *Journal of Hydrometeorology*, *11*(3), 618–641.

Hsu, K.-l., Moradkhani, H., & Sorooshian, S. (2009). A sequential Bayesian approach for hydrologic model selection and prediction. *Water Resources Research*, *45*, W00B12. https://doi.org/10.1029/2008WR006824

Jessup, S. M., & DeGaetano, A. T. (2008). A statistical comparison of the properties of flash flooding and nonflooding precipitation events in portions of New York and Pennsylvania. *Weather and Forecasting*, *23*(1), 114–130. https://doi.org/10.1175/2007waf2006066.1

Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. London: Academic Press.

Kerandi, N., Arnault, J., Laux, P., Wagner, S., Kitheka, J., & Kunstmann, H. (2017). Joint atmospheric-terrestrial water balances for East Africa: a WRF-Hydro case study for the upper Tana River basin. *Theoretical and Applied Climatology*, 1–19. https://doi.org/10.1007/s00704-017-2050-8

Kinney, J. B., & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, *111*(9), 3354–3359.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., et al. (2013). The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1

Koenker, R. (2005). *Quantile regression* (Vol. 38). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511754098

Kollet, S., Sulis, M., Maxwell, R. M., Paniconi, C., Putti, M., Bertoldi, G., et al. (2017). The integrated hydrologic model intercomparison project, IH-MIP2: A second set of benchmark results to diagnose integrated hydrology and feedbacks. *Water Resources Research*, *53*, 867–890. https://doi.org/10.1002/2016WR019191

Koren, V., Reed, S., Smith, M., Zhang, Z., & Seo, D.-J. (2004). Hydrology laboratory research modeling system (HL-RMS) of the US national weather service. *Journal of Hydrology*, *291*(3–4), 297–318. https://doi.org/10.1016/j.jhydrol.2003.12.039

Krishnamurti, T. N. (2003). Methods, systems and computer program products for generating weather forecasts from a multi-model superensemble. U.S. Patent 6535817 B1, filed 13 November 2000, issued 18 Mar 2003.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., et al. (1999). Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, *285*(5433), 1548–1550. https://doi.org/10.1126/science.285.5433.1548

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., et al. (2000). Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, *13*(23), 4196–4216. https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2

Krzysztofowicz, R. (1997). Transformation and normalization of variates with specified distributions. *Journal of Hydrology*, *197*(1), 286–292.

Kuzmin, V. (2009). Algorithms of automatic calibration of multi-parameter models used in operational systems of flash flood forecasting. *Russian Meteorology and Hydrology*, *34*, 473–481.

Kuzmin, V., Seo, D.-J., & Koren, V. (2008). Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search. *Journal of Hydrology*, *353*, 109–128.

Liang, Z., Wang, D., Guo, Y., Zhang, Y., & Dai, R. (2013). Application of Bayesian model averaging approach to multimodel ensemble hydrologic forecasting. *Journal of Hydrologic Engineering*, *18*(11), 1426–1436. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000493

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

Meyer, P. E. (2008). Information-theoretic variable selection and network inference from microarray data. (PhD thesis). Université Libre de Bruxelles.

Miller, G. A. (1955). Note on the bias of information estimates. In *Information theory in psychology: Problems and methods* (Vol. 2, pp. 95–100). Glencoe, IL: Free Press.

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Cosgrove, B. A., et al. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research*, *109*, D07S90. https://doi.org/10.1029/2003JD003823

Moore, B. J., Mahoney, K. M., Sukovich, E. M., Cifelli, R., & Hamill, T. M. (2014). Climatology and environmental characteristics of extreme precipitation events in the Southeastern United States. *Monthly Weather Review*, *143*(3), 718–741. https://doi.org/10.1175/MWR-D-14-00065.1

Moreda, F., Koren, V., Zhang, Z., Reed, S., & Smith, M. (2006). Parameterization of distributed hydrological models: Learning from the experiences of lumped modeling. *Journal of Hydrology*, *320*(1), 218–237.

Naabil, E., Lamptey, B., Arnault, J., Olufayo, A., & Kunstmann, H. (2017). Water resources management using the WRF-Hydro modelling system: Case-study of the Tono dam in West Africa. *Journal of Hydrology: Regional Studies*, *12*, 196–209.

Najafi, M. R., Moradkhani, H., & Jung, I. W. (2011). Assessing the uncertainties of hydrologic model selection in climate change impact studies. *Hydrological Processes*, *25*(18), 2814–2826. https://doi.org/10.1002/hyp.8043

Nelson, J. G. (1966). Man and geomorphic process in the Chemung River Valley, New York and Pennsylvania. *Annals of the Association of American Geographers*, *56*(1), 24–32. https://doi.org/10.1111/j.1467-8306.1966.tb00541.x

Nohara, D., Kitoh, A., Hosaka, M., & Oki, T. (2006). Impact of climate change on river discharge projected by multimodel ensemble. *Journal of Hydrometeorology*, *7*(5), 1076–1089. https://doi.org/10.1175/JHM531.1

Palmer, T. N., Alessandri, A., Andersen, U., & Cantelaube, P. (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society*, *85*(6), 853.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*, 1303–1313.

Prat, O. P., & Nelson, B. R. (2015). Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). *Hydrology and Earth System Sciences*, *19*(4), 2037–2056.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/mwr2906.1

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191.

Randrianasolo, A., Ramos, M. H., Thirel, G., Andréassian, V., & Martin, E. (2010). Comparing the scores of hydrological ensemble forecasts issued by two different hydrological models. *Atmospheric Science Letters*, *11*(2), 100–107.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., & Participants, D. (2004). Overall distributed model intercomparison project results. *Journal of Hydrology*, *298*(1), 27–60.
Regonda, S. K., Rajagopalan, B., Clark, M., & Zagona, E. (2006). A multimodel ensemble forecast framework: Application to spring seasonal flows in the Gunnison River Basin. *Water Resources Research*, *42*, W09404. https://doi.org/10.1029/2005WR004653
Salas, F. R., Somos-Valenzuela, M. A., Dugger, A., Maidment, D. R., Gochis, D. J., David, C. H., et al. (2018). Towards real-time continental scale streamflow simulation in continuous and discrete space. *JAWRA Journal of the American Water Resources Association*, *54*(1), 7–27. https://doi.org/10.1111/1752-1688.12586
Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, *2*(2), 191–201.
Sanders, F. (1973). Skill in forecasting daily temperature and precipitation: Some experimental results. *Bulletin of the American Meteorological Society*, *54*(11), 1171–1178.
Schefzik, R., Thorarinsdottir, T. L., & Gneiting, T. (2013). Uncertainty quantification in complex simulation models using Ensemble Copula Coupling. *Statistical Science*, *28*(4), 616–640.
Sedghi, H., & Jonckheere, E. (2014). On the conditional mutual information in the Gaussian–Markov structured grids. In *Information and control in networks* (pp. 277–297). New York: Springer.
Shamseldin, A. Y., & O'Connor, K. M. (1999). A real-time combination method for the outputs of different rainfall-runoff models. *Hydrological Sciences Journal*, *44*(6), 895–912.
Shamseldin, A. Y., O'Connor, K. M., & Liang, G. C. (1997). Methods for combining the outputs of different rainfall-runoff models. *Journal of Hydrology*, *197*(1), 203–229. https://doi.org/10.1016/S0022-1694(96)03259-3
Sharma, S., Siddique, R., Balderas, N., Fuentes, J. D., Reed, S., Ahnert, P., et al. (2017). Eastern U.S. verification of ensemble precipitation forecasts. *Weather and Forecasting*, *32*(1), 117–139. https://doi.org/10.1175/waf-d-16-0094.1
Sharma, S., Siddique, R., Reed, S., Ahnert, P., Mendoza, P., & Mejia, A. (2018). Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system. *Hydrology and Earth System Sciences*, *22*(3), 1831.
Siddique, R., & Mejia, A. (2017). Ensemble streamflow forecasting across the U.S. Mid-Atlantic region with a distributed hydrological model forced by GEFS reforecasts. *Journal of Hydrometeorology*, *18*(7), 1905–1928. https://doi.org/10.1175/jhm-d-16-0243.1
Siddique, R., Mejia, A., Brown, J., Reed, S., & Ahnert, P. (2015). Verification of precipitation forecasts from two numerical weather prediction models in the Middle Atlantic Region of the USA: A precursory analysis to hydrologic forecasting. *Journal of Hydrology*, *529*, 1390–1406.
Silver, M., Karnieli, A., Ginat, H., Meiri, E., & Fredj, E. (2017). An innovative method for determining hydrological calibration parameters for the WRF-Hydro model in arid regions. *Environmental Modelling & Software*, *91*, 47–69.
Smith, M. B., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., et al. (2012). The distributed model intercomparison project—Phase 2: Motivation and design of the Oklahoma experiments. *Journal of Hydrology*, *418*, 3–16.
Steinschneider, S., Wi, S., & Brown, C. (2015). The integrated effects of climate and hydrologic uncertainty on future flood risk assessments. *Hydrological Processes*, *29*(12), 2823–2839. https://doi.org/10.1002/hyp.10409
Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S., & Rogers, E. (1999). Using ensembles for short-range forecasting. *Monthly Weather Review*, *127*(4), 433–446. https://doi.org/10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2
Thirel, G., Regimbeau, F., Martin, E., Noilhan, J., & Habets, F. (2010). Short- and medium-range hydrological ensemble forecasts over France. *Atmospheric Science Letters*, *11*(2), 72–77. https://doi.org/10.1002/asl.254
Thirel, G., Rousset-Regimbeau, F., Martin, E., & Habets, F. (2008). On the impact of short-range meteorological forecasts for ensemble streamflow predictions. *Journal of Hydrometeorology*, *9*(6), 1301–1317.
Thompson, P. D. (1977). How to improve accuracy by combining independent forecasts. *Monthly Weather Review*, *105*(2), 228–229. https://doi.org/10.1175/1520-0493(1977)105<0228:HTIABC>2.0.CO;2
Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*, W01413. https://doi.org/10.1029/2005WR004723
Toth, Z., & Kalnay, E. (1993). Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of the American Meteorological Society*, *74*(12), 2317–2330. https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2
Velázquez, J. A., Anctil, F., Ramos, M. H., & Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. *Advances in Geosciences*, *29*, 33–42. https://doi.org/10.5194/adgeo-29-33-2011
Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, *43*, W01411. https://doi.org/10.1029/2005WR004838
Wei, M., Toth, Z., Wobus, R., & Zhu, Y. (2008). Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A*, *60*(1), 62–79.
Weigel, A. P., Liniger, M., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241–260.
Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2009). Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Monthly Weather Review*, *137*(4), 1460–1479.
Weisheimer, A., Doblas-Reyes, F. J., Palmer, T. N., Alessandri, A., Arribas, A., Déqué, M., Keenlyside, N., et al. (2009). ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical Research Letters*, *36*, L21711. https://doi.org/10.1029/2009GL040896
Winkler, R. L., Murphy, A. H., & Katz, R. W. (1977). The consensus of subjective probability forecasts: Are two, three, ..., heads better than one? In *Preprints of the fifth conference on probability and statistics in atmospheric sciences* (pp. 57–62). Boston: American Meteorological Society.
Xiong, L., Shamseldin, A. Y., & O'Connor, K. M. (2001). A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi-Sugeno fuzzy system. *Journal of Hydrology*, *245*(1), 196–217. https://doi.org/10.1016/S0022-1694(01)00349-3
Yuan, X., & Wood, E. F. (2013). Multimodel seasonal forecasting of global drought onset. *Geophysical Research Letters*, *40*, 4900–4905. https://doi.org/10.1002/grl.50949
Yucel, I., Onen, A., Yilmaz, K., & Gochis, D. (2015). Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall. *Journal of Hydrology*, *523*, 49–66.