Improving Bayesian Model Averaging for Ensemble Flood Modeling Using Multiple Markov Chains Monte Carlo Sampling

Tao Huang[1] and Venkatesh Merwade[1]

[1]Lyles School of Civil Engineering, Purdue University, West Lafayette, Indiana, USA

Corresponding author: Tao Huang (huan1441@purdue.edu)

Key Points:

- Multiple Markov Chains Monte Carlo method is applicable for estimating Bayesian model averaging weights and variances.
- The performance of Metropolis-Hastings algorithm is better than that of Expectation-Maximization algorithm in ensemble flood modeling.
- Multiple Markov Chains Monte Carlo method can provide more information about the uncertainty of Bayesian model averaging parameters.

Abstract

As all kinds of physics-based and data-driven models are emerging in hydrologic and hydraulic engineering, Bayesian model averaging (BMA) is one of the popular multi-model methods used to account for various uncertainty sources in the flood modeling process and generate robust ensemble predictions. The reliability of BMA parameters (weights and variances) determines the accuracy of BMA predictions. However, the uncertainty in BMA parameters with fixed values, which are usually obtained from Expectation-Maximization (EM) algorithm, has not been adequately investigated in BMA-related applications over the past few decades. Given the limitations of the commonly used EM algorithm, Metropolis-Hastings (M-H) algorithm, which is one of the most widely used algorithms in Markov Chain Monte Carlo (MCMC) method, is proposed to estimate BMA parameters. Both numerical experiments and one-dimensional HEC-RAS models are employed to examine the applicability of M-H algorithm with multiple independent Markov chains. The performances of EM and M-H algorithms are compared based on the daily water stage predictions from 10 model members. Results show that BMA weights estimated from both algorithms are comparable, while BMA variances obtained from M-H algorithm are closer to the given variances in the numerical experiment. Moreover, the normal proposal used in M-H algorithm can yield narrower distributions for BMA weights than those from the uniform proposal. Overall, MCMC approach with multiple chains can provide more information associated with the uncertainty of BMA parameters and its performance is better than the default EM algorithm in terms of multiple evaluation metrics as well as algorithm flexibility.

1 Introduction

Application of different numerical models based on physical processes as well as data driven approaches are playing a crucial role in simulating hydrologic and hydraulic systems to provide flood risk information to the public (Dottori et al., 2021; FEMA, 2018; Teng et al., 2017; Xie et al., 2021). Given the temporal and spatial variability of flood events as well as the complexity of watersheds, the pursuit of a "perfect" model that can incorporate all hydrologic and hydraulic processes and handle different scenarios is encouraging, but this pursuit faces multiple challenges. The challenges in simulating flooding processes, including the limitations of governing mathematic principles, estimation of parameters, measurement of driving forces, and computational efficiency, are important issues that modelers need to take into consideration in order to provide reliable and robust predictions about flooding (Jafarzadegan et al., 2021; Kobarfard et al., 2022; Liu and Merwade, 2018; Merwade et al., 2008; F Pappenberger et al., 2005; Florian Pappenberger et al., 2006; Sharma et al., 2022; Teng et al., 2017). Additionally, "equifinality" in hydrologic and hydraulic modeling may lead to multiple models or different model configurations to yield similar results that match the observations equally well (Beven and Binley, 1992; Refsgaard et al., 2012; Von Bertalanffy, 1972). Thus, it is not recommended to rely on streamflow predictions and flood inundation maps obtained by a single model implementation (Duan et al., 2007; Huang and Merwade, 2023; Liu and Merwade, 2018; Romanowicz and Beven, 2003; Zounemat-Kermani et al., 2021).

Considering the uncertainty involved in simulating the hydrology and hydraulics of flooding, multi-model ensemble methods should be applied to capture various uncertainty sources for making robust predictions (Bates and Granger, 1969; Dickinson, 1973; Newbold and Granger,

1974; Teng et al., 2017; Zounemat-Kermani et al., 2021). Over the past few decades, different kinds of multi-model methods, for example the simple average method (Makridakis et al., 1982), the weighted average method (Makridakis and Winkler, 1983; Singh et al., 2010; Tebaldi and Knutti, 2007; Winkler, 1989), the multi-model super-ensemble method (Ajami et al., 2006; Krishnamurti et al., 1999) and the neural network method (Acharya et al., 2014; Chattopadhyay and Chattopadhyay, 2008; Shamseldin and O'Connor, 1999; Shamseldin et al., 1997; Zounemat-Kermani et al., 2021), have been widely used in climate projections and hydrologic predictions. Among these multi-model ensemble approaches, the Bayesian model averaging (BMA) approach (Kass and Raftery, 1995; Leamer, 1978; Raftery et al., 2005) has been successfully applied to the flood modeling (Duan et al., 2007; Huang and Merwade, 2023; Liu and Merwade, 2018, 2019; Rings et al., 2012; Vrugt et al., 2008). More importantly, the BMA approach is able to produce accurate and reliable predictions due to its advantages over other multi-model techniques. Specifically, the BMA probability density function (PDF) of a predictive variable is a weighted average of PDFs from a model ensemble that covers multiple significant uncertainty sources. From the Bayesian perspective, the weight of a model member represents its relative prediction performance compared to the other model members and the sum of the nonnegative weights is equal to one, which is easy to interpret when performing the model evaluation and comparison (Raftery et al., 2005). Additionally, the BMA procedure can produce the prediction probability distribution which reflects the uncertainty associated with the mean prediction (Raftery et al., 1997). Finally, the relative impacts of individual uncertainty sources and the uncertainty propagation in the modeling process can be demonstrated and quantified through a hierarchical BMA (HBMA) framework (Chitsazan and Tsai, 2015; Huang and Merwade, 2023; Liu and Merwade, 2019).

The performance of BMA predictions is primarily dependent on the accurate estimation of BMA parameters (weights and variances). However, the reliability associated with BMA parameters has not received enough attention. In past literature relevant to the applications of BMA approach, both the weights and variances obtained through the Expectation-Maximization (EM) algorithm are assigned fixed values (Cao et al., 2021; Darbandsari and Coulibaly, 2021; Dormann et al., 2018; Duan et al., 2007; Huang and Merwade, 2023; Liu and Merwade, 2018; Madadgar and Moradkhani, 2014; Moknatian and Mukundan, 2023; Tsai, 2010). These fixed values cannot represent the uncertainty in the BMA parameters, especially when different datasets are used for training or a specific training dataset does encompass the overall prediction capacity of certain models (Madadgar and Moradkhani, 2014; Refsgaard et al., 2012; Rojas et al., 2010; Tebaldi and Knutti, 2007). Although the EM algorithm has been provided to be able to provide good estimates of the BMA weight and variance for each model member with satisfactory computational efficiency (McLachlan and Krishnan, 2007; Vrugt et al., 2008), a few issues need to be addressed. First, the global optimal estimates of BMA parameters cannot be guaranteed especially for solving some high-dimensional problems (Duan et al., 2007; Vrugt et al., 2008). Second, the assumption that the conditional PDF of the variable of interest follows a normal distribution in the application of the default EM algorithm limits its wide application in other fields. Finally, a fixed value for the pre-processing parameter (e.g., transformation parameter in the Box-Cox method) used to transform the original datasets into the Gaussian space can add more uncertainty in the final predictions (Liu and Merwade, 2018).

Given the limitations of the default EM algorithm in the BMA analysis discussed above, the Markov Chain Monte Carlo (MCMC) sampling method (Robert et al., 1999) is proposed in this study to estimate the BMA parameters. It has been shown that the application of MCMC method for hydrology can provide a full view of the unknown parameter's posterior probability distribution (Gaume et al., 2010; Nguyen et al., 2021; Reis Jr and Stedinger, 2005; Sharma et al., 2022; Vrugt et al., 2008; Wang et al., 2017; Zhao et al., 2021), which can provide an explicit representation and quantification for the parameter uncertainty. Specifically, this method can be implemented by different algorithms to generate a sequence of stochastic samples that converge to the target probability density function of the unknown parameter (Luengo et al., 2020). In a previous study, differential evolution adaptive metropolis algorithm (Vrugt, 2016; Vrugt et al., 2008) was developed to estimate the BMA weights and an integrated variance of the hydrologic model ensemble. This study also introduced more parameters, such as the number of chain pairs used to generate the proposed sample and the jump size among different modes, in the algorithm. Furthermore, an overall variance across different model members was assumed and the autocorrelation of the samples in each chain was not evaluated, which would add more uncertainty to the estimates of BMA parameters.

The Metropolis-Hastings (M-H) algorithm (Hastings, 1970; Metropolis et al., 1953) is one of the most widely used algorithms in the MCMC method, but its feasibility for estimating BMA parameters has not been investigated in previous literature. For the solution to some high-dimensional and multimodal problems (e.g., likelihood function in the BMA analysis), one single Markov chain may get stuck in a local optimal mode (Vats and Knudson, 2021), and hence multiple independent Markov chains is a possible option to explore the whole value space of

unknown parameters. Additionally, the MCMC method with the M-H algorithm does not require too many subjective judgments for setting-up of the sampling process and it is flexible in terms of modification based on the assumption of the target posterior distribution of the unknown parameter. Accordingly, the overall goal of this study is to investigate the feasibility of the M-H MCMC algorithm and assess any advantages over the default EM algorithm for estimating BMA parameters (weights and variances) of ensemble flood modeling. This broader goal is accomplished through the following objectives: (i) Apply BMA approach in flood modeling to estimate BMA parameters using different numbers of samples in each Markov chain with the M-H algorithm; (ii) compare the performance of EM and M-H MCMC algorithms for estimating the BMA parameters; (iii) estimate the BMA weights using different proposal distributions in the M-H algorithm; and (iv) investigate the impact of different conditional PDFs of the predictor variable on the BMA parameters. To examine the validity and applicability of the MCMC method with the M-H algorithm in the BMA analysis, both numerical experiments and hydraulic modeling using Hydrologic Engineering Center-River Analysis System (HEC-RAS) (Brunner, 2016b; FEMA, 2018) are carried out in this study.

2 Study area and data

For hydraulic modeling to evaluate the uncertainty in the BMA parameters and verify the applicability of the MCMC method on the parameter estimation, two river reaches (see Figure 1 and Table 1) in the states of Indiana and Texas of the United States are selected. Both reaches have existing HEC-RAS models from the Federal Emergency Management Agency's (FEMA) Flood Insurance Rate Map program and have both upstream and downstream streamflow measurements from the United States Geological Survey (USGS) gauges. These study reaches

148 are located in the midwestern and southern regions of the United States and provide distinct
149 geomorphic characteristics, thus making them good test beds for the BMA analysis. The HEC-
150 RAS models for all counties in Indiana are available from the Indiana Department of Natural
151 Resources' hydrology and hydraulics library (INDNR, 2018), and the HEC-RAS models for the
152 central and western regions in Texas can be accessed from FEMA's Estimated Base Flood
153 Elevation Viewer (FEMA, 2015). Available HEC-RAS models from FEMA studies are 1D
154 steady-state, but to reduce the uncertainty from the steady-flow assumption and also enable
155 comparison of simulation output from the model ensemble to the observation from the USGS
156 gauges, they are modified to perform 1D unsteady-flow simulation. The unsteady state
157 modification did not involve any changes to the model structure and parameter, including the
158 layout and geometry of river cross-sections and Manning's n for the main channel and the
159 floodplain. Unsteady state simulations for 100 days from March to July 2021 covering a couple
160 of flood events are performed. Daily streamflow data (upstream input for HEC-RAS) and water
161 stage data (downstream output for the BMA analysis) used in this study are obtained from the
162 corresponding USGS gauges at each study reach.

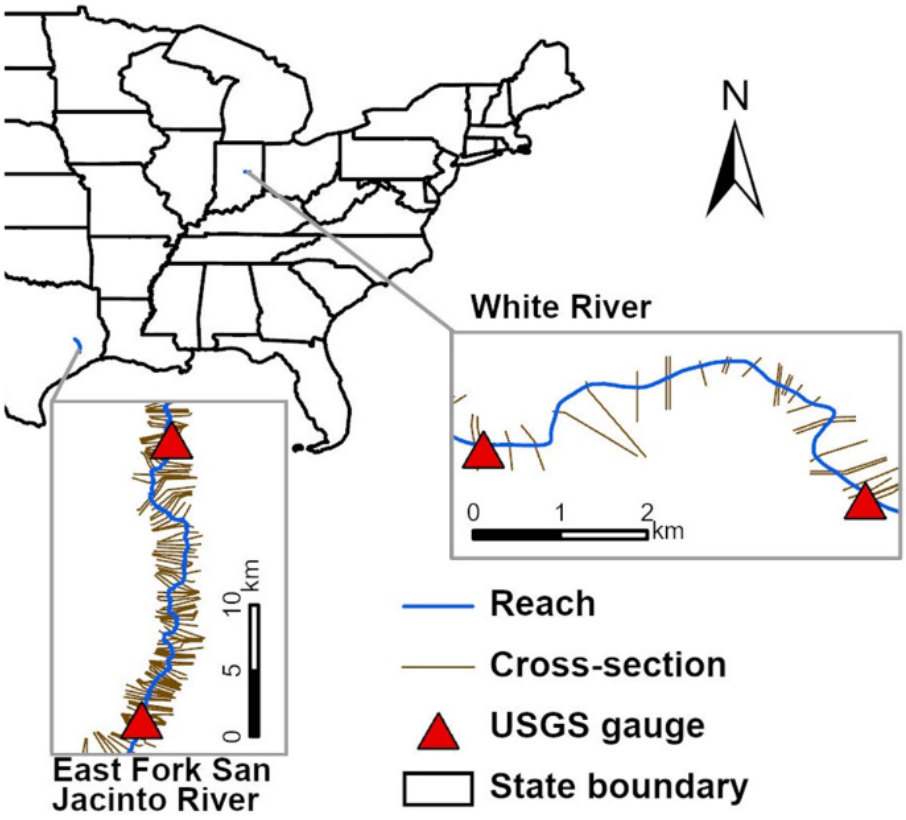


Figure 1. Layout map of study areas in Indiana and Texas, USA.

Table 1. Basic information of study area.

Study stream	Channel length (km)	Average channel width (m)	Channel slope (%)	Upstream USGS gauge	Downstream USGS gauge	Simulation Period (100 days)
White	6.76	64	0.0631	03348000	03348130	2021-3-15 to 2021-6-22
East Fork San Jacinto	50.11	76	0.0438	08070000	08070200	2021-4-15 to 2021-7-23

3 Methodology

3.1 Numerical Experiment and Hydraulic Modeling

(1) Numerical experiment

Considering the complexity of residual patterns in hydrologic and hydraulic models (Beven, 2016), it is necessary to conduct numerical experiments based on pre-defined model errors as

this is the first study to apply the MCMC method for estimating BMA parameters. Given the basic assumption of EM algorithm and the relatively small size (less than 10) of the model ensemble in practice (Cao et al., 2021; Huo et al., 2019; Moknatian and Mukundan, 2023; Raftery et al., 2005; Tian et al., 2021; Tsai, 2010; Vrugt et al., 2008), a total of 10 sets of hydrologic data (100 days of daily observed water stage data, denoted as D, of the White river in Indiana) that contain random errors (denoted as ε) are generated. These synthetic datasets are created assuming a normal distribution with zero mean and standard deviations ranging from 0.06 m (0.2 ft) to 0.30 m (1 ft) with an increment of 0.06m (0.2 ft). These 10 datasets (see Table 2) will serve as the predictions obtained from the flood model ensemble. Next, both EM and M-H MCMC algorithms are applied to these predictions to estimate the BMA parameters (weight and variance) of each candidate model. Due to the randomness in model errors, it is expected that the value of BMA parameters will vary to some extent even though the standard deviations of the normal distribution are the same. For example, both Model 1 (f_1) and Model 2 (f_2) have the same standard deviation of 0.06 m in Table 2.

Table 2. BMA ensemble model members for numerical experiments.

No.	Model Predictions (m)	No.	Model Predictions (m)
1	$f_1 = D + \varepsilon$, where $\varepsilon \sim N(0, 0.06^2)$	2	$f_2 = D + \varepsilon$, where $\varepsilon \sim N(0, 0.06^2)$
3	$f_3 = D + \varepsilon$, where $\varepsilon \sim N(0, 0.12^2)$	4	$f_4 = D + \varepsilon$, where $\varepsilon \sim N(0, 0.12^2)$
5	$f_5 = D + \varepsilon$, where $\varepsilon \sim N(0, 0.18^2)$	6	$f_6 = D + \varepsilon$, where $\varepsilon \sim N(0, 0.18^2)$
7	$f_7 = D + \varepsilon$, where $\varepsilon \sim N(0, 0.24^2)$	8	$f_8 = D + \varepsilon$, where $\varepsilon \sim N(0, 0.24^2)$
9	$f_9 = D + \varepsilon$, where $\varepsilon \sim N(0, 0.30^2)$	10	$f_{10} = D + \varepsilon$, where $\varepsilon \sim N(0, 0.30^2)$

Note: D = 100 days of daily water stage data, and ε = random model errors that follow a normal distribution.

(2) Ensemble Flood Modeling in 1D HEC-RAS

A 1D HEC-RAS model is a simplified representation of the river network, i.e., river centerlines, left and right bank lines, and horizontal cross-sections along the reach. During the simulation process, the water surface elevation at discrete cross-sections is estimated for given boundary conditions. Using the water surface elevations from all cross-sections, a 2D inundation extent map can be generated by interpolating the positive differences between the topographical data (e.g., digital elevation model) and the water surface layer. More technical details about the 1D HEC-RAS can be referred to the software manual (Brunner, 2016a, 2016b). In terms of the model configuration, a HEC-RAS project is a collection of multiple files including the river geometry, relevant parameters, and boundary conditions. Based on previous studies about the uncertainty in the modeling of FEMA's flood inundation maps, an uncertainty range of ± 20% around values of channel roughness and upstream flow input used in the model are considered in simulations (Huang and Merwade, 2023; Liu and Merwade, 2019). Accordingly, the ensemble is made up of 10 sets of model predictions from the HEC-RAS ensemble (see Table 3), 9 of which are obtained based on different combinations of geometry files (including channel roughness) and unsteady flow files (including upstream flow series), and the last one is the average of simulations from the other 9 members.

Table 3. Model configurations in HEC-RAS for BMA analysis.

Model Configuration No.	**Channel Roughness**	**Upstream Flow Input**	**HEC-RAS Plan Files**
1	0.8n	0.8Q	g01 & u01
2	0.8n	Q	g01 & u02
3	0.8n	1.2Q	g01 & u03
4	N	0.8Q	g02 & u01
5	N	Q	g02 & u02
6	N	1.2Q	g02 & u03
7	1.2n	0.8Q	g03 & u01

Model Configuration No.	Channel Roughness	Upstream Flow Input	HEC-RAS Plan Files
8	1.2n	Q	g03 & u02
9	1.2n	1.2Q	g03 & u03
10	Average of simulations from No.1-No.9		

Note: n is the Manning's n value for the main channel in the original HEC-RAS models, Q is the streamflow from USGS gauge stations, g** represents a geometry file of a HEC-RAS project, and u** represents a flow data file of a HEC-RAS project.

3.2 Bayesian modeling averaging (BMA) analysis

BMA (Kass and Raftery, 1995; Raftery et al., 2005) is a statistical method that averages the predictions from an ensemble of multiple competing models instead of from a single "perfect" model through the corresponding BMA weights. According to the law of total probability, the PDF of the BMA probabilistic prediction of the variable of interest (daily water stage in this study) is given by Equations (1) - (2).

$$p(y \mid D) = \sum_{k=1}^{K} p(f_k \mid D) \cdot p_k(y \mid f_k, D) = \sum_{k=1}^{K} w_k \cdot p_k(y \mid f_k, D) \quad (1)$$

$$\sum_{k=1}^{K} w_k = 1 \quad (2)$$

where y is the predictor variable; $D = [y_1^{obs}, y_2^{obs}, ... y_T^{obs}]$ is the observed hydrologic data (daily water stage data from the USGS gauges) with data length T; f_k is the prediction of the k^{th} model. $p(f_k \mid D) = w_k$ is the posterior probability of the predictions of k^{th} model given the observation data. The weight (w_k) reflects how well a specific model prediction matches the observed data, and hence better-performance models have higher weights, which is nonnegative and goes up to

one. $p_k(y|f_k,D)$ is the posterior distribution of y given both the predictions of the k^{th} model and the observation data.

Based on Equation (1), a log-likelihood function is constructed for estimating the BMA weight and variance that can maximize the likelihood. By convention, it is assumed that the conditional probability $p_k(y|f_k,D)$ follows a normal distribution. Thus, the log-likelihood function is shown in Equation (3).

$$L(\theta)=\sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}w_k\cdot p_k(y_t\mid f_{k,t},\sigma_k)\right)=\sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}w_k\cdot\frac{1}{\sigma_k\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y_t-f_{k,t}}{\sigma_k}\right)}\right) \quad (3)$$

where θ is the unknown BMA parameters (weight and variance of each model member), w_k is the BMA weight of k^{th} model, $f_{k,t}$ is the prediction of the k^{th} model at time step t, y_t is the predictor variable at time step t, and the mean and standard deviation of the normal distribution, $f_{k,t}$ and σ_k.

Because some of the hydrologic variables (e.g., water stage, streamflow, rainfall, etc.) are nonnegative, the corresponding distributions tend to be skewed to some extent (Sloughter et al., 2007). Hence the assumption of normal distribution does not rigorously hold in this case. To explore the effect of conditional PDFs on the estimates of BMA parameters, it is assumed that the conditional probability $p_k(y|f_k,D)$ follows a gamma distribution (PDF = 0 for y ≤ 0) with two parameters (α and β) (Qi et al., 2021; Vrugt et al., 2008; Vrugt and Robinson, 2007), and the log-likelihood function is shown in Equation (4). The gamma distribution is not symmetric

around the mean like the normal distribution. Specifically, its mean is identical to $\mu=\alpha\beta=f_k$, and its variance is identical to $\sigma_k^2=\alpha\beta^2=\beta f_k$, which depends on the specific prediction (f_k) and hence is heteroscedastic.

$$L(\theta)=\sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}w_k\cdot p_k(y_t\mid f_{k,t},\sigma_k)\right)=\sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}w_k\cdot\frac{1}{\Gamma(\alpha)\beta^{\alpha}}y_t^{\alpha-1}e^{-\frac{y_t}{\beta}}\right) \quad (4)$$

where α is the shape parameter of gamma distribution and $\alpha=f_{k,t}^2/\sigma_k^2$, and β is the scale parameter of gamma distribution and $\beta=\sigma_k^2/f_{k,t}$.

3.3 Expectation-Maximization (EM) Algorithm

Generally, it is not easy to obtain an analytical solution for the BMA log-likelihood function (see Equation (3)) so the EM algorithm is recommended, and has been commonly used in previous BMA applications (Duan et al., 2007; Huang and Merwade, 2023; Liu and Merwade, 2018; Madadgar and Moradkhani, 2014; McLachlan and Krishnan, 2007; Moknatian and Mukundan, 2023; Parrish et al., 2012; Raftery et al., 2005). To find a numerical solution that can maximize the log-likelihood function, this algorithm alternates iteratively between the E (i.e., Expectation) step (see Equation (5)) and the M (i.e., Maximum) step (see Equations (6) and (7)). The value of log-likelihood function is updated with the weight and variance estimated through each iteration. The iteration will not stop until the difference between the previous value and the current value of the log-likelihood function is within a pre-assigned threshold (10^{-4} is taken in the study). The EM algorithm can guarantee that the likelihood function will be increasing monotonically at each iteration (Saul and Lee, 2002; Wu, 1983), and thus the parameters that maximize the likelihood can be obtained from the last M step. A few previous studies pointed out that the EM algorithm

265 can only converge to the local optimum rather than the global optimal results (Duan et al., 2007;
266 McLachlan and Krishnan, 2007; Vrugt et al., 2008), but this issue has not been addressed
267 adequately in the literature.

268 $$z_{k,t}^{i} = \frac{w_k^{i-1} g_k(y_t \mid f_{k,t}, \sigma_k^{i-1})}{\sum_{k=1}^{K} w_k^{i-1} \cdot g_k(y_t \mid f_{k,t}, \sigma_k^{i-1})} \quad (5)$$

269 $$w_k^i = \frac{1}{T} \sum_{t=1}^{T} z_{k,t}^i \quad (6)$$

270 $$\sigma_k^i = \sqrt{\frac{\sum_{t=1}^{T} z_{k,t}^i \cdot (y_t - f_{k,t})^2}{\sum_{t=1}^{T} z_{k,t}^i}} \quad (7)$$

271 where $z_{k,t}^i$ is the latent variable depending on the performance of the k^{th} model at time t in the i^{th}
272 iteration of the EM algorithm, w_k^i and σ_k^i is the BMA weight and standard deviation of the k^{th}
273 model in the i^{th} iteration, respectively.

274 **3.4 Metropolis-Hastings (M-H) Algorithm**

275 In the conventional BMA analysis, optimal estimates of BMA weights and variances yielded
276 from the EM algorithm are fixed values. These fixed values cannot be considered with certainty
277 to be the global optimal solutions and provide enough information about the uncertainty
278 associated with the estimates of the BMA parameters. Therefore, the MCMC method with the
279 M-H algorithm is proposed to estimate the most likely values of the BMA parameters and
280 produce their underlying posterior distributions. The MCMC sampling is usually conducted
281 based on one single chain, but one single Markov chain may get stuck in a local mode for a high

dimensional problem (in the order of 20 unknown variables in this study). Thus, multiple independent chains are also tried in this study to explore the parameter space simultaneously and adequately, and then posterior distribution of each parameter is generated by using the local mode yielded by each chain. The flowchart in Figure 2 shows the procedure of the M-H algorithm with multiple MCMC chains. Without enough prior knowledge about the target distribution of the BMA parameters, a uniform distribution ranging from 0 to 1 or the normal distribution with a standard deviation of 0.1 is proposed to generate new samples of weights. A uniform distribution ranging from 0 to 1.5 x *RMSE* (root mean square error) is proposed to generate new samples of standard deviations during the MCMC sampling process. Since the dimensions and units of BMA standard deviations are different for different hydrologic variables of interest, it is difficult to propose another generalized proposal distribution and hence only a uniform proposal distribution is used in this study. The benefits of these assumptions mentioned above are that the proposal (prior) distribution is symmetric, and hence $q(\theta^* | \theta_n) = q(\theta_n | \theta^*)$. Thus, the posterior density distribution of the parameters, $\pi(\theta \mid D)$, is directly equal to the likelihood function (see Equations (3) and (4)). The improved method can maintain the ergodicity based on multiple Markov chains and is expected to provide a full view of the posterior distributions of the BMA parameters. Moreover, multiple sizes of each MCMC chain are attempted to investigate its effect on the estimates of BMA parameters. In addition, compared with the traditional EM algorithm, which requires the assumption of a conditional normal PDF, the MCMC simulation is easy to set up without any major algorithmic modifications when using different conditional PDFs for the variable of interest in the BMA analysis.

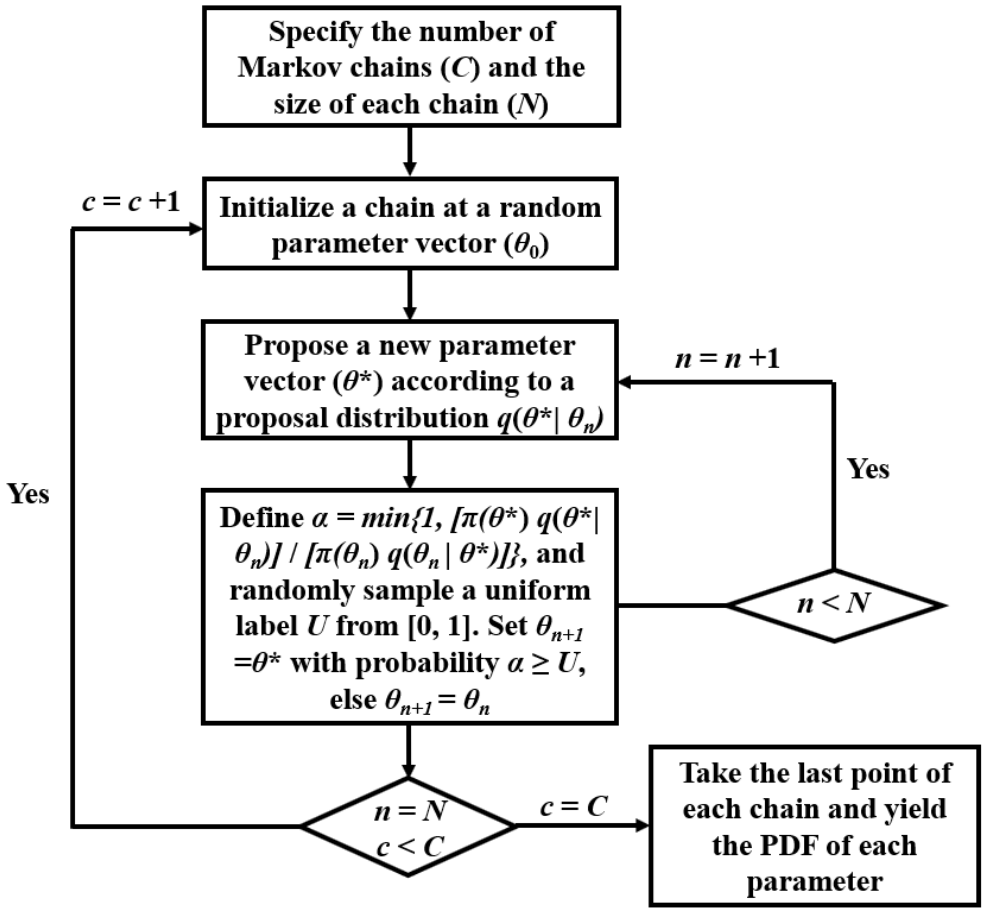


Figure 2. Procedure of Metropolis-Hastings (M-H) algorithm with multiple Markov chains.

3.5 Evaluation Metrics for Model Performance

Both *RMSE* (Equation (8)) and four types of uncertainty coefficients (*UC1* - *UC4*, see Equations (9) - (12)) (Huang and Merwade, 2023) defined based on the reliability of BMA prediction distribution and the accuracy of BMA mean predictions are employed to evaluate the overall model performance under EM and MCMC algorithms. In general, lower values of these metrics indicate less uncertain model predictions, a "perfect" model will have can uncertainty coefficient of zero. It is also important to note that these four types of uncertainty coefficients are independent from each other, thus the model performance and the associated uncertainty can be evaluated and quantified comprehensively from different perspectives.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{obs,i} - y_{BMA,i})^2}{n}} \quad (8)$$

$$UC1 = \frac{N_{obs-90\%}}{n} \cdot 100\% \quad (9)$$

$$UC2 = 1 - NSE = \frac{RMSE^2}{\sigma_{obs}^{\ 2}} \cdot 100\% \quad (10)$$

$$UC3 = 1 - KGE = \sqrt{(r-1)^2 + (\frac{\sigma_{BMA}}{\sigma_{obs}} - 1)^2 + (\frac{\bar{y}_{BMA}}{\bar{y}_{obs}} - 1)^2} \cdot 100\% \quad (11)$$

$$UC4 = (1 - R^2 + |1 - Slope|) \cdot 100\% \quad (12)$$

where n is the total number of data points, $\bar{y}_{obs}$ is the mean value of the observations $y_{obs,i}$ at each step i, $y_{BMA,i}$ is the i^{th} BMA mean prediction value, $N_{obs-90\%}$ is the number of observed data points located outside the 90% prediction interval, σ_{obs} is the standard deviation of the observed data, r is the correlation coefficient between BMA mean predictions and observations, σ_{BMA} is the standard deviation of BMA mean predictions, $\bar{y}_{BMA}$ is the mean of the BMA mean predictions, R^2 is the coefficient of determination of a linear regression equation in the form of "Observation = *Slope**BMA mean prediction", and *Slope* is the slope of the linear regression equation.

4 Results and Discussion

4.1 Effect of Sample Sizes in M-H MCMC Algorithm

The MCMC method aims to find a finite number of local moves that produce samples asymptotically from the "correct" distribution of the variables of interest, which are the BMA weights and variances in this study. However, an appropriate sample size for generating a MCMC chain that mixes rapidly is still based on modelers' experience and this size varies significantly for different cases (Gelman et al., 1995; Robert et al., 1999). Moreover, a strong autocorrelation among the samples within a chain can reduce the effective sample size, thus reducing the efficiency of sampling (Luengo et al., 2020). For the high-dimensional problem in this study (i.e., BMA likelihood function), Figure 3 shows that the conventional MCMC method with one single chain gets trapped in a set of local optimal solutions, and all the trace plots of BMA weights and standard deviations tend to be stagnant. Even after the sample size reaches 100,000, which is much larger than the values used in previous literature, the mixing of any MCMC chain does not improve. In other words, the proposed samples are always rejected due to the low acceptance rate α (see Figure 2) and it is hard for a chain to jump outside the local mode of the posterior distribution of BMA parameters.

Given the poor mixing of the MCMC sampling with one single chain, multiple independent MCMC chains are considered in this study by using the last point of each chain to form a new chain (see Figure 2), which presents the target posterior distribution. The trace plots (see Figure 3) of multiple MCMC chains with different sample sizes show that a chain is very unlikely to accept a new sample after about 2000 iterations. Thus, 2000 samples are generated in each chain and the number of chains is set to be 100. The trace plots and autocorrelation functions (ACF) of

351 the BMA weights and variances of Model 1 (a model member with the lowest variance of model
352 errors) and Model 9 (a model with the largest variance of model errors) in the numerical
353 experiments (see Figures 4 and 5) indicate that the samples drawn from multiple independent
354 MCMC chains are mixing very well and the autocorrelation is quickly dropping within the 5%
355 significant interval (shown in the light blue region in the ACF in Figures 4 and 5), which means
356 the entire space of the BMA parameters has been fully explored, and individual samples in the
357 MCMC chain follow stationary and independent identical distributions. The histograms in
358 Figures 4 and 5 represent the posterior distributions of the BMA weights and standard deviations,
359 and more importantly, these statistical distributions demonstrate the uncertainty of the mean
360 estimates. The histograms of the BMA parameters of Model 1 are close to a normal distribution,
361 while the histograms of the BMA weights and standard deviations of Model 9 seem to follow a
362 positively skewed distribution and a uniform distribution, respectively. It should be noted that
363 similar analyses can also be performed on the posterior distributions of BMA parameters of other
364 model members.

365

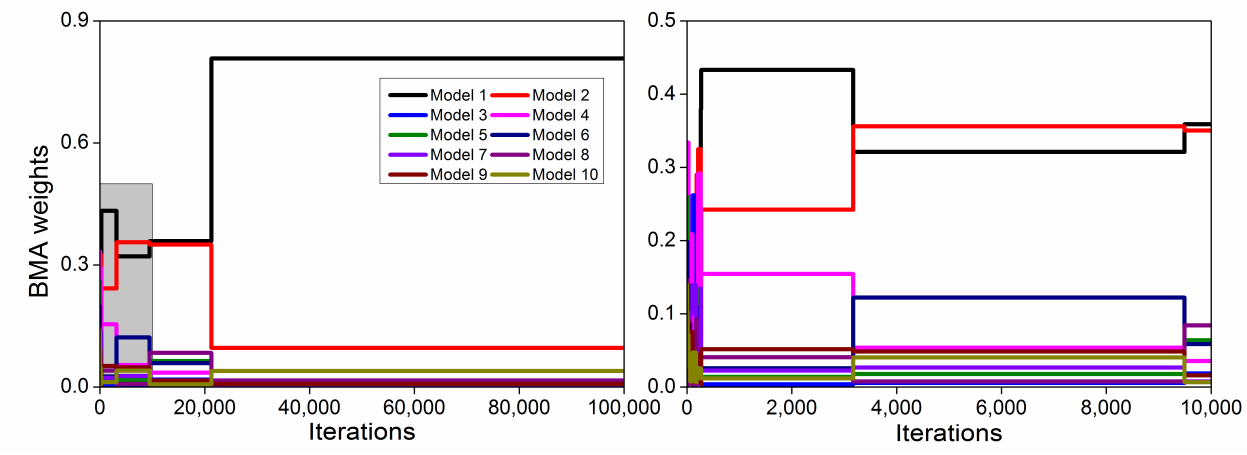
366 As shown in the mixing trace plots and the ACFs in Figures 4 and 5, the posterior distribution of
367 BMA parameters is stationary, and then the results from the 2000^{th} sample in 100 independent
368 chains are taken as a benchmark compared to the results obtained from multiple sizes of each
369 chain. The comparison of mean BMA weights and standard deviations obtained from 100
370 MCMC chains with different sample sizes is shown in Figures 6 and 7. The estimates from
371 multiple sizes (3000, 4000, 5000, and 10000) are linearly regressed with those from 2000
372 samples of each chain. The intercept of the regression equation is set to be zero, so only *Slope* is
373 involved in the equation. The results show that the estimates of BMA parameters do not change

374 significantly (all the data points in Figure 6 are lying near the 45-degree line, and R^2 is within
375 0.95-1 and *Slope* is within 0.95-1.05 as shown in Figure 7) as the sample size increases, while
376 the elapsed time of the sampling process carried out in a local personal computer becomes longer
377 and longer (see Figure 7). Therefore, from the parsimony point of view, 2000 samples for each
378 MCMC chain are adequate to guarantee the convergence of the target distribution of BMA
379 parameters of the flood model ensemble with 10 or fewer members.

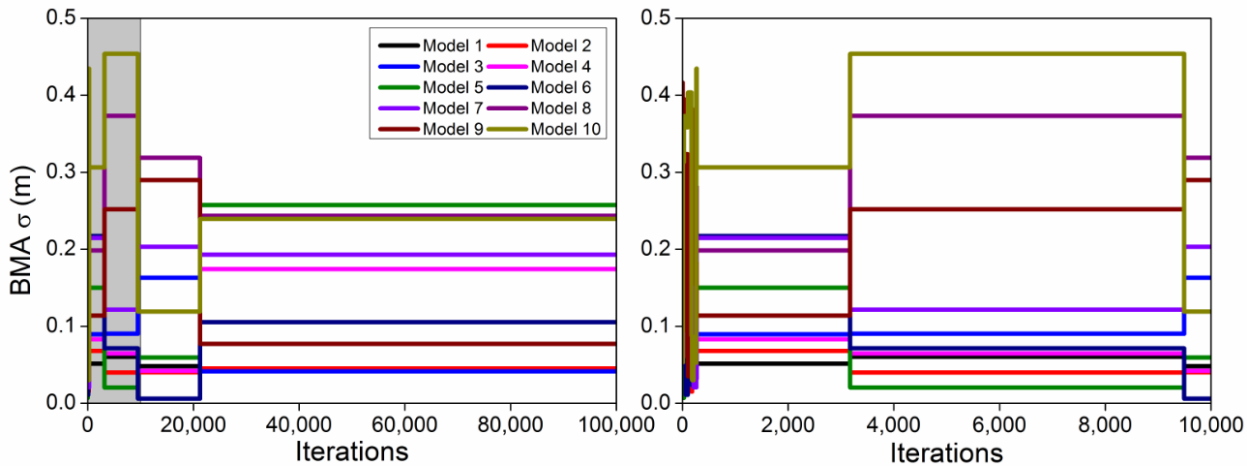
380

381 Table 4 and Table 5 present the BMA parameters estimated through EM and MCMC algorithms,
382 respectively. Because the model errors of Model 1 and Model 2 have the smallest variance, the
383 results from different algorithms show good agreement that these two models are assigned the
384 highest BMA weight among the 10 model members. The EM algorithm can quickly identify the
385 model member with lower-variance errors (i.e., Model 1 and Model 2) and a small weight that is
386 close to zero is assigned to the other model members with poorer prediction performances (i.e.,
387 Model 3 - Model 10). Since the prior distribution of BMA weights spans from 0 to 1, it is
388 unlikely for the weight to be zero in the MCMC sampling process. However, it is important to
389 highlight that the magnitudes of BMA weights obtained via these two algorithms are similar and
390 the 90% confidence intervals from the MCMC algorithm do contain the deterministic value from
391 the EM algorithm. The histograms in Figures 4 and 5 also show that the EM estimates are
392 located around the mode of the statistical distribution. On the other hand, the BMA standard
393 deviations of Model 3 - Model 10 estimated from the EM algorithm is quite close to zero, which
394 is far away from the “true” standard deviation of the given model errors. This issue may be due
395 to the original set up of the EM algorithm for the standard deviation (see Equations (5) and (7))
396 and hence it makes the estimates difficult to interpret. On the contrary, corresponding estimates

from the MCMC algorithm are quite close to the known standard deviation and it makes much more sense that a model member with a larger BMA variance usually gets a lower BMA weight. Additionally, it is important to note that even though the given errors have the same variance for two model members (e.g., Model 1 and Model 2, Model 3 and Model 4, and so on) in the numerical experiment, the BMA parameters of each pair are slightly different due to the randomness of the model errors. This further implies that fixed values of BMA parameters would introduce more uncertainty to these estimates. Thus, the posterior statistical distribution obtained through the MCMC algorithm is a more reliable way to demonstrate and quantify the uncertainty associated with BMA parameters.

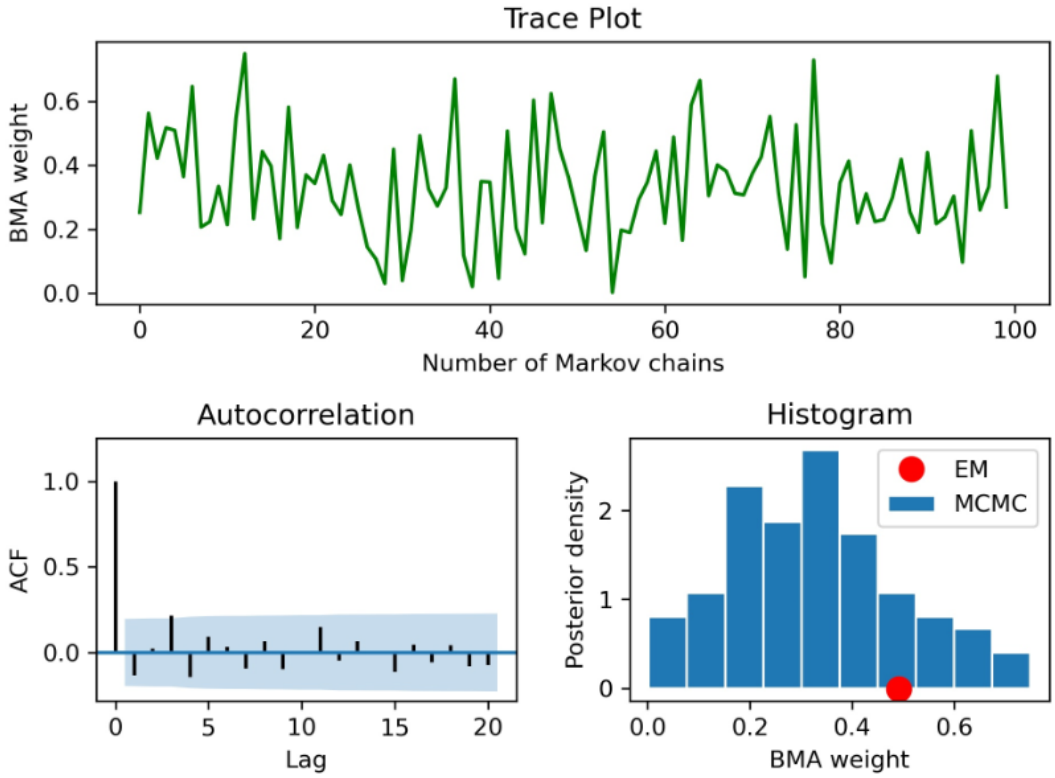


(a) BMA weights obtained from a single MCMC chain (100k samples & top 10k samples in the grey region).

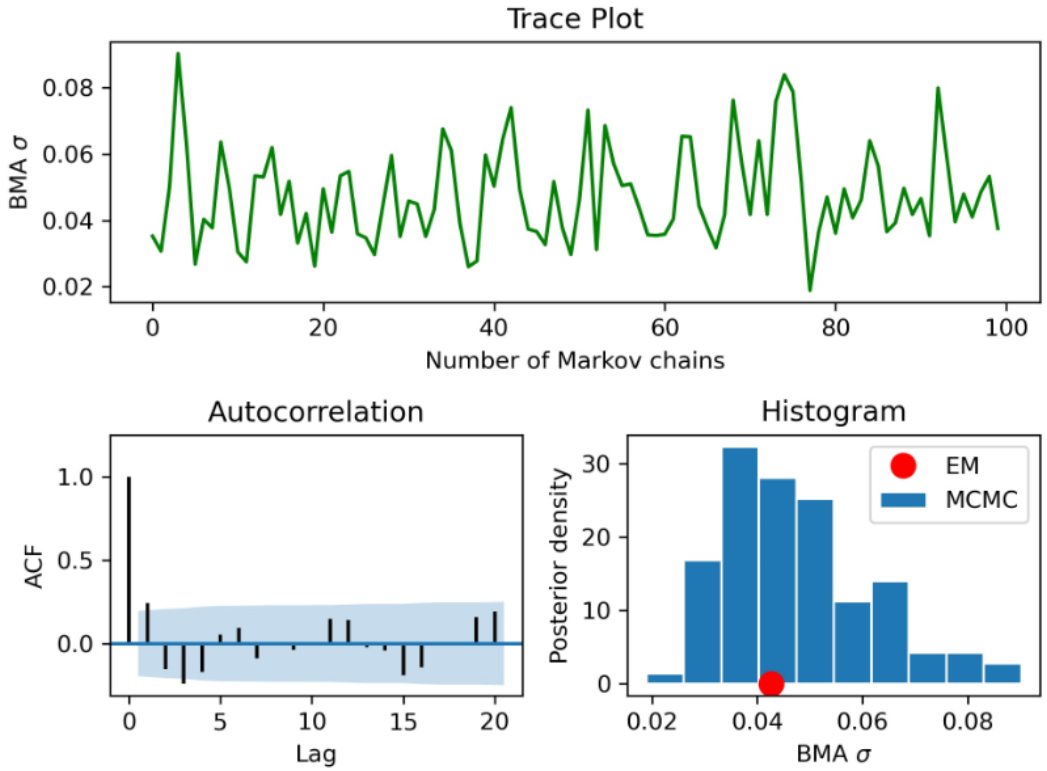


(b) BMA standard deviations obtained from a single MCMC chain (100k samples and top 10k samples in the grey region).

Figure 3. Trace plots of BMA parameters from M-H algorithm with a single MCMC chain with 100k samples.

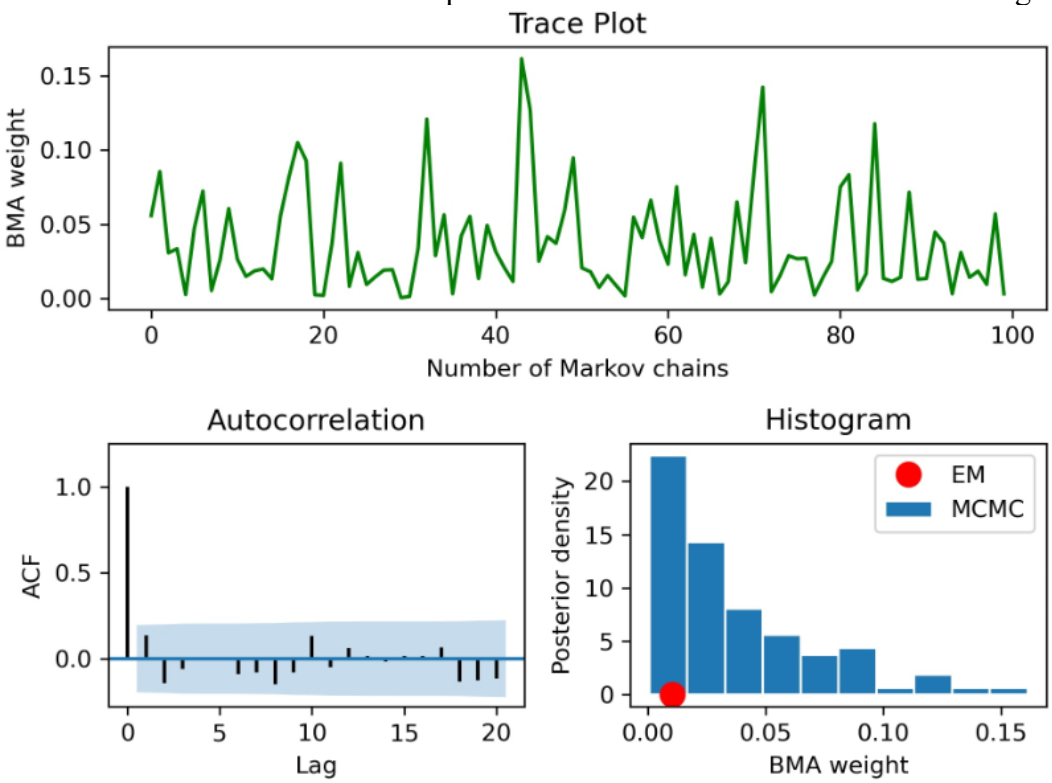


(a) BMA weight.

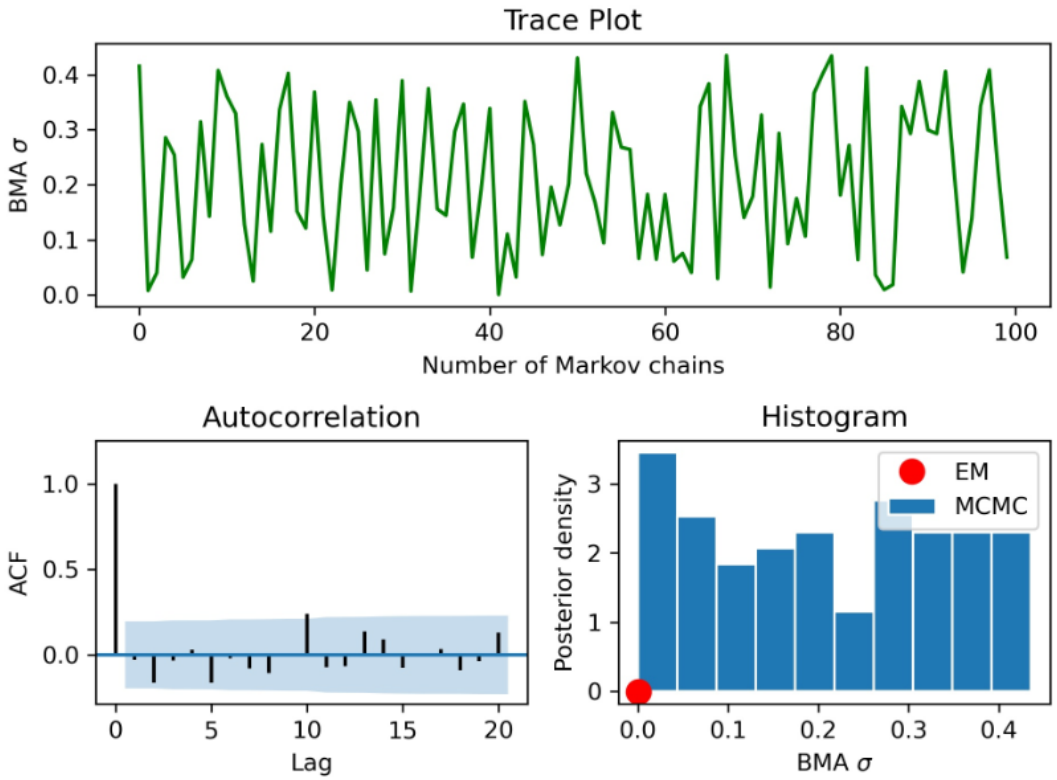


(b) BMA standard deviation (m).

Figure 4. Posterior distribution of BMA parameters of Model 1 from M-H MCMC algorithm.

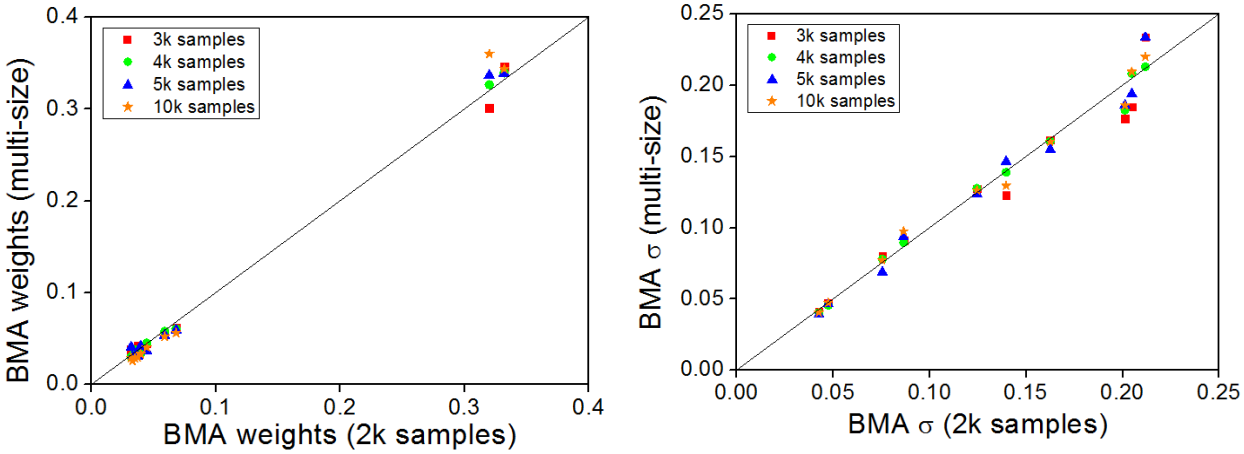


(a) BMA weight.



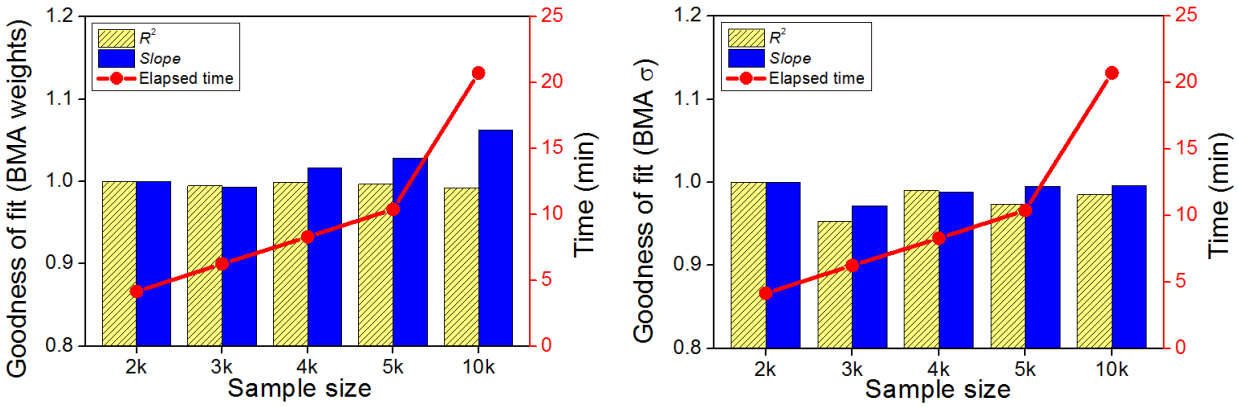
(b) BMA standard deviation (m).

Figure 5. Posterior distribution of BMA parameters of Model 9 from M-H MCMC algorithm.



(a) BMA weights. (b) BMA standard deviations (m).

Figure 6. Comparison of BMA weights from M-H algorithm with 100 MCMC chains with different sample sizes.



(a) BMA weights. (b) BMA standard deviations.

Figure 7. Comparison of elapsed time (Desktop processor specifications: Intel(R) Core(TM) i7-9700 CPU @ 3.00GHz) and R^2 & *Slope* of linear regression from M-H algorithm with 100 MCMC chains with different sample sizes.

Table 4. BMA parameters obtained from EM algorithm.

No.	Model	Weight	σ (m)
1	$\varepsilon \sim N(0, 0.06^2)$	0.492	0.04
2	$\varepsilon \sim N(0, 0.06^2)$	0.497	0.04
3	$\varepsilon \sim N(0, 0.12^2)$	0.0002	0.02
4	$\varepsilon \sim N(0, 0.12^2)$	0	0.05
5	$\varepsilon \sim N(0, 0.18^2)$	0	0.08
6	$\varepsilon \sim N(0, 0.18^2)$	0	0.02
7	$\varepsilon \sim N(0, 0.24^2)$	0	0.11
8	$\varepsilon \sim N(0, 0.24^2)$	0	0.05
9	$\varepsilon \sim N(0, 0.3^2)$	0.01	0.001
10	$\varepsilon \sim N(0, 0.3^2)$	0	0.09

Table 5. BMA parameters obtained from MCMC algorithm.

No.	Model	Mean weight	90% interval of weight	Mean σ (m)	90% interval of σ (m)
1	$\varepsilon \sim N(0, 0.06^2)$	0.332	[0.05, 0.65]	0.05	[0.02, 0.08]
2	$\varepsilon \sim N(0, 0.06^2)$	0.32	[0.07, 0.60]	0.04	[0.01, 0.07]
3	$\varepsilon \sim N(0, 0.12^2)$	0.068	[0, 0.17]	0.08	[0.02, 0.14]
4	$\varepsilon \sim N(0, 0.12^2)$	0.059	[0, 0.16]	0.09	[0.01, 0.17]
5	$\varepsilon \sim N(0, 0.18^2)$	0.04	[0, 0.11]	0.14	[0.02, 0.27]
6	$\varepsilon \sim N(0, 0.18^2)$	0.044	[0, 0.12]	0.12	[0.02, 0.23]
7	$\varepsilon \sim N(0, 0.24^2)$	0.035	[0, 0.10]	0.16	[0.02, 0.30]
8	$\varepsilon \sim N(0, 0.24^2)$	0.032	[0, 0.09]	0.2	[0.04, 0.35]
9	$\varepsilon \sim N(0, 0.3^2)$	0.037	[0, 0.11]	0.2	[0.01, 0.41]
10	$\varepsilon \sim N(0, 0.3^2)$	0.033	[0, 0.09]	0.21	[0.02, 0.43]

438

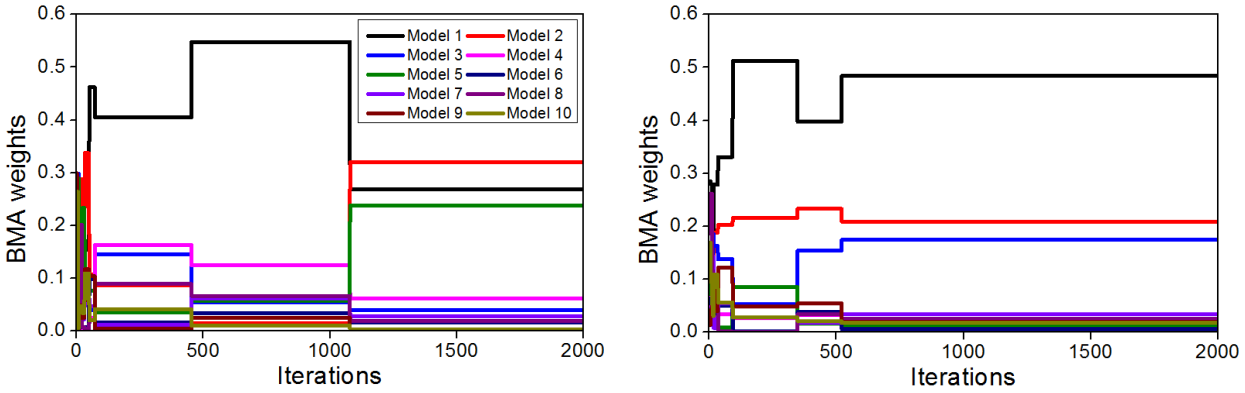
439 **4.2 Impact of Proposal Distributions in M-H MCMC Algorithm**

440 The proposal distribution in the MCMC method is a conditional distribution (see $q(\theta^*|\ \theta_n)$ and
441 $q(\theta_n\ |\theta^*)$ in Figure 2) that is used to generate a new sample θ^* given the current sample θ_n, and
442 vice versa. However, the selection of the proposal distribution tends to be subjective in practice.
443 To investigate the effect of proposal distributions on the estimates of BMA weights, a normal
444 proposal distribution is proposed to compare with the uniform proposal distribution used in the
445 previous section. The standard deviation of the normal prior distribution is taken as 0.1 through a
446 trial and error procedure. If the standard deviation is too small, it is very likely that a new sample
447 BMA weight to be rejected; if the standard deviation is assigned a large value, a new sample
448 BMA weight might be negative, which would contradict the assumption of nonnegative BMA
449 weights. Similarly, the number of samples in each chain is assigned to be 2000 since the larger
450 sample size does not change the estimates significantly. Figure 8 shows that the trace plots of
451 BMA weights generated based on both proposal distribution are already stagnant when the
452 sample size reaches 2000. To be consistent with the sampling based on a uniform proposal
453 distribution, the number of independent chains is also set to be 100. The estimates of BMA
454 weights based on the normal proposal distributions in the M-H MCMC sampling are presented in
455 Table 6. Comparing Tables 5 and 6, the mean estimates of BMA weights obtained from both
456 uniform and normal proposal distributions are quite close to each other. However, the 90%
457 confidence interval of the BMA weights based on the normal proposal distribution is narrower
458 compared to that from the uniform proposal distribution, which means the former one gives a
459 more precise and confident assessment on the performance of individual model members.

460

461 Based on the BMA parameters estimated from the EM algorithm and the MCMC method with
462 different proposal distributions, the prediction distributions of water stage produced through a
463 Monte Carlo sampling procedure are shown in Figure 9 and the model uncertainty quantified
464 based on multiple independent evaluation metrics are presented in Table 7. Figure 9 shows that
465 these three sets of BMA mean predictions match the observations very well and the difference in
466 model performance is similar visually. However, it is obvious that the 90% confidence interval
467 of the BMA prediction distribution obtained from the EM algorithm is narrower due to its
468 extreme low values of the BMA variances. The values of *RMSE* and *UC1* - *UC4* indicate that the
469 overall model performance based on MCMC method is better than the EM algorithm, and the
470 normal proposal distribution in the M-H MCMC algorithm is slightly better than the uniform
471 proposal distribution. Therefore, the results of the numerical experiment demonstrate that the
472 MCMC method with multiple independent chains and a normal proposal distribution in the M-H
473 algorithm can be a better alternative to the default EM algorithm for estimating the reliable BMA
474 parameters (weights and variances) in the BMA analysis.

475



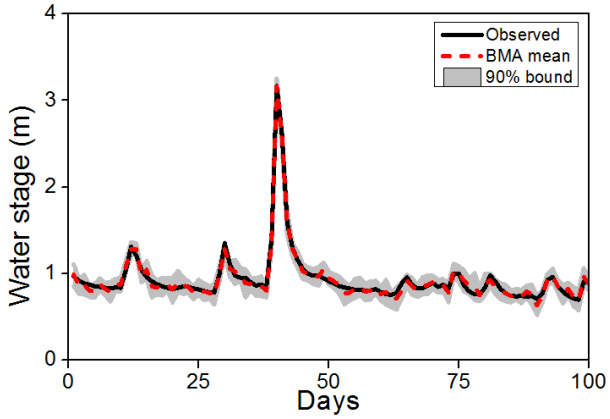
(a) Uniform proposal distribution. (b) Normal proposal distribution.

476 Figure 8. Trace plots of BMA weights from M-H algorithm with different proposal distributions.
477

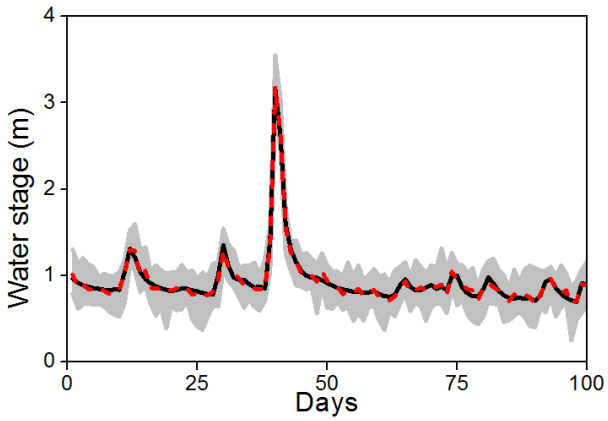
478 Table 6. BMA weights obtained from normal proposal distributions in M-H MCMC algorithm.

No.	Model	Mean weight	90% interval of weight
1	$\varepsilon \sim N(0, 0.06^2)$	0.354	[0.23, 0.50]
2	$\varepsilon \sim N(0, 0.06^2)$	0.35	[0.21, 0.49]
3	$\varepsilon \sim N(0, 0.12^2)$	0.062	[0, 0.14]
4	$\varepsilon \sim N(0, 0.12^2)$	0.053	[0, 0.14]
5	$\varepsilon \sim N(0, 0.18^2)$	0.034	[0, 0.09]
6	$\varepsilon \sim N(0, 0.18^2)$	0.033	[0, 0.09]
7	$\varepsilon \sim N(0, 0.24^2)$	0.029	[0, 0.09]
8	$\varepsilon \sim N(0, 0.24^2)$	0.028	[0, 0.08]
9	$\varepsilon \sim N(0, 0.3^2)$	0.028	[0, 0.07]
10	$\varepsilon \sim N(0, 0.3^2)$	0.028	[0, 0.07]

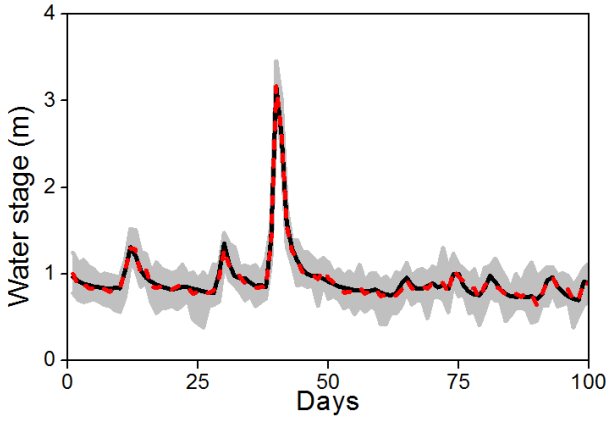
479



(a) EM algorithm.



(b) Uniform proposal distribution in M-H MCMC algorithm.



(c) Normal proposal distribution in M-H MCMC algorithm.

480 Figure 9. Water stage predictions obtained from EM and M-H MCMC algorithms with different
481 proposal distributions.

Table 7. Comparison of model performance under different algorithms.

Algorithm	***RMSE* of mean predictions (m)**	**Average 90% prediction interval (m)**	***UC1* (%)**	***UC2* (%)**	***UC3* (%)**	***UC4* (%)**
EM	0.041	0.19	5.00	1.61	1.06	2.00
M-H MCMC (uniform)	0.037	0.47	0.00	1.33	1.04	1.66
M-H MCMC (normal)	0.037	0.41	0.00	1.33	0.96	1.64

4.3 Influence of Conditional PDFs in BMA analysis

One of the advantages of using the MCMC method, compared to the EM algorithm, in the BMA analysis is its flexibility in assuming a conditional PDF. The EM algorithm is strictly limited to the normal conditional PDF (see Equation (3)). Even though the normal conditional PDF is assumed initially in the BMA analysis (Raftery et al., 2005), it is probably true for some climate variable (e.g., temperature), but some nonnegative hydrologic variables (e.g., rainfall, streamflow, and water stage) may not be normally distributed. Because the probability of the negative hydrologic variable must be equal to zero, the corresponding PDF should be skewed to some extent. To explore the impact the conditional PDFs on the BMA parameters, the results from ten HEC-RAS model configurations of the two study reaches in Indiana and Texas under both normal and gamma conditional distributions are compared. Figures 10 and 11 exhibit the posterior distribution of the BMA parameters of individual model members under different conditional PDFs, and Tables 8 and 9 present their mean estimates. For the HEC-RAS models used in this study, the statistical distributions obtained from both normal and gamma conditional PDFs are slightly different, but the overall patterns are quite similar. For the White River in

499 Indiana, the BMA weight of Model 9 ranks first and its distribution tends to be normal. The
500 BMA weights of Model 6 and Model 8 come second and third, respectively. The statistical
501 distributions of the BMA weights of other model members are highly positively skewed and the
502 mean values are less than 0.1. Figures 10(c) and 10(d) show that the statistical distributions of
503 the BMA standard deviations of the model members with a good performance (i.e., a higher
504 BMA weight) are usually narrowly dispersed and the corresponding mean values in Table 9 are
505 also lower than those of the model members with a poor performance (i.e., a lower BMA weight).

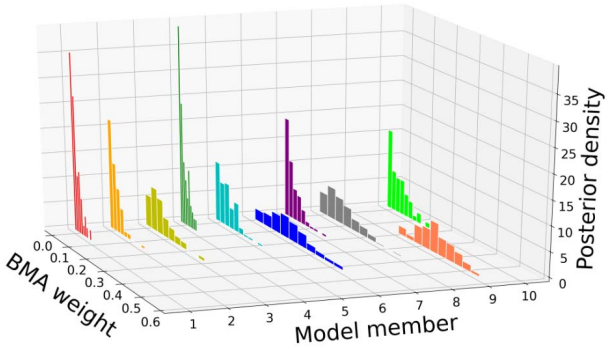
506

507 For the East Fork San Jacinto River in Texas, the performance of most of the model members is
508 very close, and hence produce similar posterior distributions and mean values for the BMA
509 parameters are yielded (see Figure 11). Among the model ensemble, Model 9, Model 3, and
510 model 10 are assigned a relatively higher BMA weight and a relatively lower BMA standard
511 deviation. Overall, this conclusion based on the HEC-RAS models from the two study reaches is
512 consistent with the findings obtained from the numerical experiment. More importantly, the
513 MCMC method can provide a comprehensive view of the uncertainty associated with the BMA
514 parameters, and hence more informed decisions on the flood risk control can be made based on
515 the flood model members with robust and consistent prediction capacities.

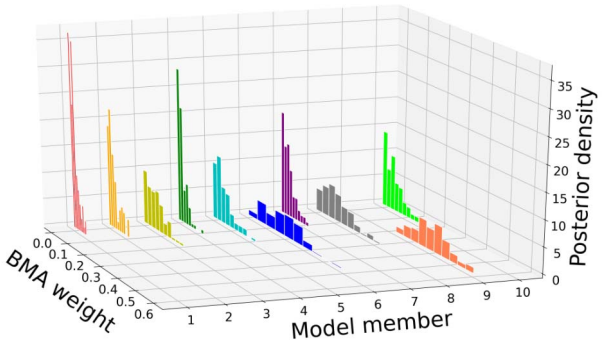
516

517 Based on the BMA parameters estimated from the MCMC method under different conditional
518 PDFs, the prediction distributions of water stage are created through a Monte Carlo sampling
519 procedure and are shown in Figure 12. The rank of water stage *RMSE*s from each model member
520 is presented in Figure 13 and the values of different uncertainty coefficients (Equations 8-12) to

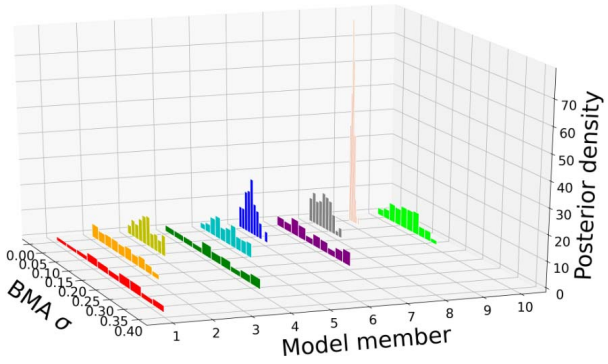
quantify the model performance are presented in Table 10. Both visual and the quantitative comparison of the water stage predictions demonstrate that the model performance under the normal and gamma conditional PDFs are comparable, because all the model members of these two study reaches presents relatively good prediction skills (e.g., *RMSE*s < 0.5m and *UC*s < 15%) and the model residuals do not demonstrate a dominated pattern (e.g., either normal or gamma PDF in this study). It is, however, important to note that that the gamma conditional PDF yields a slightly better performance, which means the pattern of model residuals fit a gamma distribution better than a normal one. Additionally, it is interesting to note in Figure 13 that the *RMSE*s of Model 6 for the White River and Model 3 for the East Fork San Jacinto River rank first compared to other model members and two sets of BMA mean predictions. It implies that some other types of conditional PDF that fits the water stage residuals of the model members better than the normal or gamma PDF used in this study might exist, which, however, has to be found through a trial-and-error procedure.



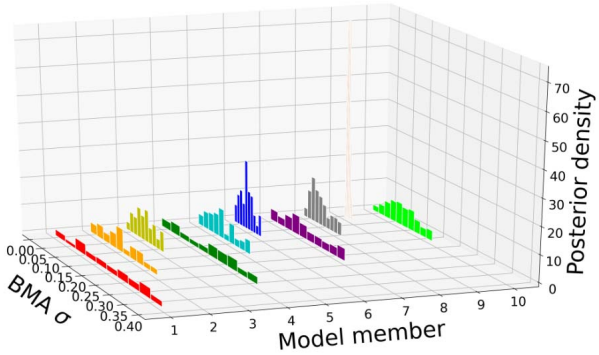
(a) BMA weights (normal conditional PDF).



(b) BMA weights (gamma conditional PDF).



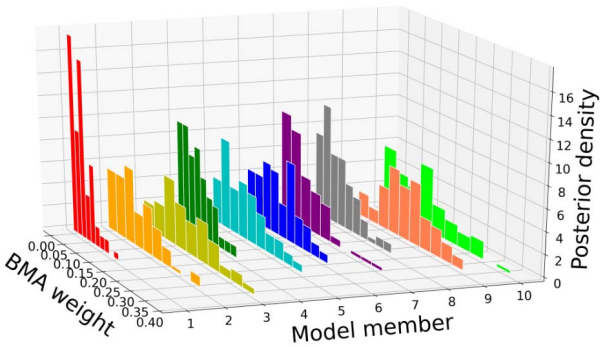
(c) BMA standard deviations in meters (normal conditional PDF).



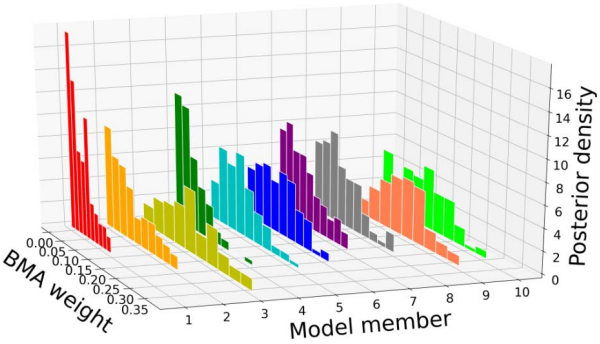
(d) BMA standard deviations in meters (gamma conditional PDF).

534 Figure 10. Comparison of the BMA parameters obtained from different BMA conditional PDFs
535 for White River.
536

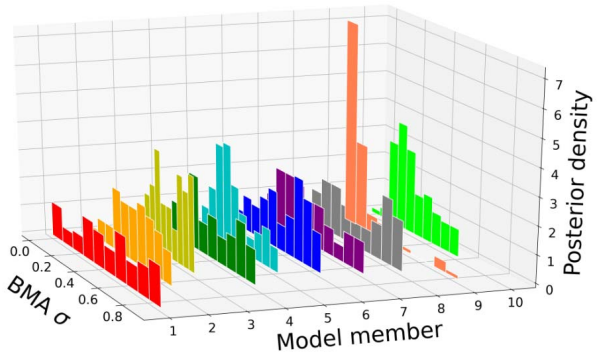
537



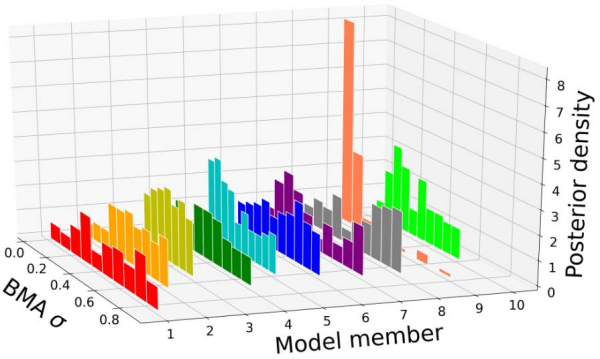
(a) BMA weights (normal conditional PDF).



(b) BMA weights (gamma conditional PDF).



(c) BMA standard deviations in meters (normal conditional PDF).



(d) BMA standard deviations in meters (gamma conditional PDF).

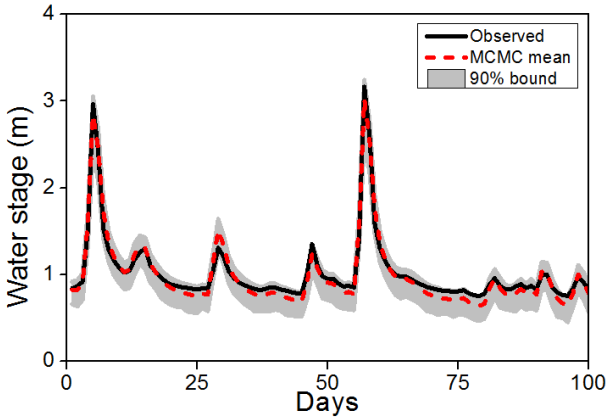
538 Figure 11. Comparison of the BMA parameters obtained from different BMA conditional PDFs
539 for East Fork San Jacinto River.

540 Table 8. Comparison of mean BMA weights under different conditional PDFs.

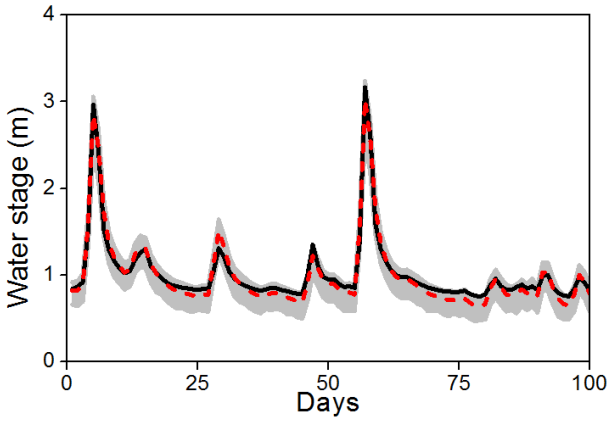
River / Model No.	White		East Fork San Jacinto	
	Normal	**Gamma**	**Normal**	**Gamma**
1	0.022	0.02	0.041	0.045
2	0.032	0.035	0.1	0.083
3	0.083	0.066	0.15	0.17
4	0.023	0.025	0.064	0.054
5	0.057	0.055	0.104	0.097
6	0.193	0.184	0.105	0.108
7	0.039	0.037	0.069	0.076
8	0.11	0.112	0.071	0.085
9	0.387	0.411	0.178	0.162
10	0.054	0.054	0.119	0.121

Table 9. Comparison of mean BMA standard deviations (m) under different conditional PDFs.

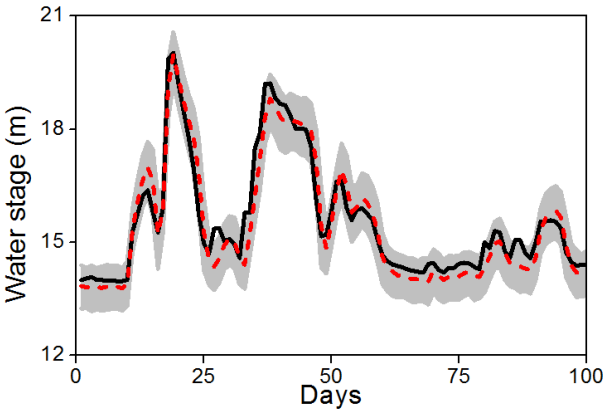
River / Model No.	White		East Fork San Jacinto	
	Normal	**Gamma**	**Normal**	**Gamma**
1	0.22	0.21	0.49	0.5
2	0.12	0.11	0.4	0.4
3	0.07	0.07	0.37	0.38
4	0.18	0.17	0.4	0.38
5	0.1	0.09	0.3	0.33
6	0.05	0.05	0.41	0.41
7	0.14	0.12	0.32	0.35
8	0.05	0.06	0.4	0.45
9	0.03	0.03	0.19	0.16
10	0.09	0.09	0.31	0.32



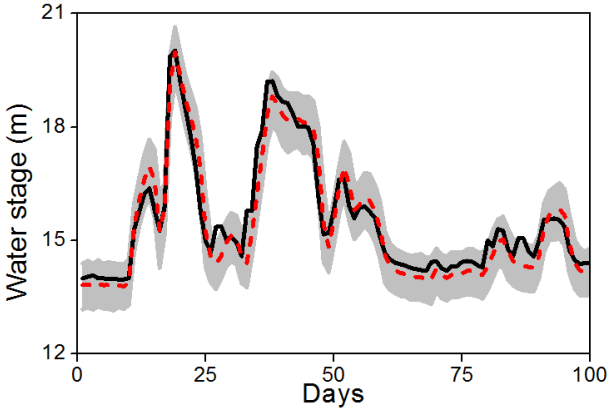
(a) Normal conditional PDF for White River.



(b) Gamma conditional PDF for White River.

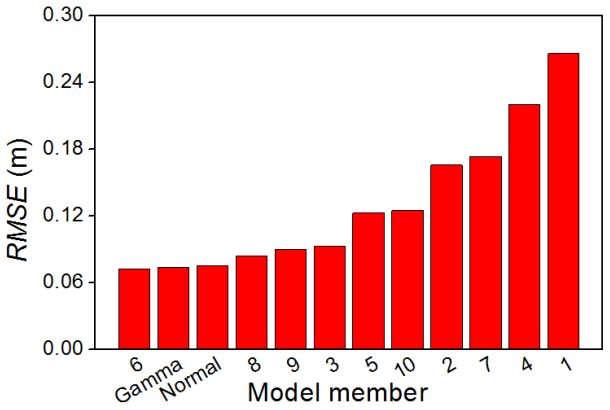


(c) Normal conditional PDF for East Fork San Jacinto River.

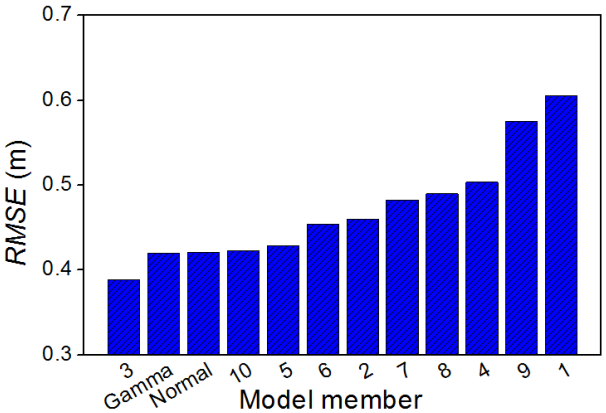


(d) Gamma conditional PDF for East Fork San Jacinto River.

Figure 12. Water stage predictions from different BMA conditional PDFs for study area.



(a) White River.



(b) East Fork San Jacinto River.

Figure 13. Rank of RMSEs of daily water stage from model ensembles for study area.

Table 10. Comparison of prediction performances under different BMA conditional PDFs.

Study	Conditional	*RMSE* of	Average 90%	*UC1*	*UC2*	*UC3*	*UC4*

stream	PDF	mean predictions (m)	prediction interval (m)	(%)	(%)	(%)	(%)
White	Normal	0.075	0.32	5.00	3.43	5.94	4.87
	Gamma	0.074	0.32	5.00	3.34	5.97	4.62
East Fork San Jacinto	Normal	0.421	1.32	13.00	7.54	5.69	7.76
	Gamma	0.419	1.34	11.00	7.48	5.67	7.71

5 Conclusions

Reliable, robust, and accurate flood predictions are critical for understanding flood risk and taking actions. A computational model based on certain assumptions and simplifications of the complicated hydrologic system is subject to uncertainty that must not be ignored. Thus, it is wise to make decisions based on predictions from multiple competing candidate models rather than relying on one single model even if it has been well calibrated. Among all kinds of multi-model methods, quite a few studies have shown successful applications of the BMA method in the fields of hydrologic and hydraulic engineering. Accurate estimates of BMA parameters (weights and variances) determine the performance of BMA predictions. However, the uncertainty associated with BMA parameters estimated through the default EM algorithm has not been investigated systematically. Given the research gap in the previous literature, the M-H MCMC method with multiple independent chains is proposed in this study to address the limitations of the EM algorithm that provide deterministic estimates for the BMA parameters. The applicability of the MCMC method is examined based on both numerical experiment with known patterns of model errors and the case studies of two 1D HEC-RAS models in the states of Indiana and Texas of the United States. Following major conclusions are drawn from this study:

(1) Results of the numerical experiment show that the estimates of BMA parameters obtained from the MCMC method do not change significantly beyond a sample size of 2000, but the computational cost of the sampling process increases as the sample size increases. Considering both the accuracy of estimates and the sampling efficiency, 2000 samples per individual MCMC chains are adequate to generate a stationary distribution of BMA parameters for the model ensemble of 10 or fewer members. As the posterior distribution estimated from the M-H algorithm is found to be stationary, the chain number that is greater than 30 should be sufficient to draw inferences for the population properties of BMA parameters, but 100 independent chains are employed in this study to be more conservative.

(2) The numerical experiment demonstrates that the prediction performance of the M-H MCMC algorithm in the BMA analysis is less uncertain than the default EM algorithm in terms of multiple independent evaluation metrics (*RMSE* and *UC1* - *UC4*). The magnitudes of BMA weights estimated from both algorithms are similar, but the standard deviations estimated from the MCMC method are closer to the "true" values of model errors in Table 2. Furthermore, the normal proposal distribution with a standard deviation of 0.1 for the BMA weight can slightly improve the performance of the MCMC method. Overall, the BMA parameters obtained from the MCMC method are more interpretable in terms of the model performance comparison than the EM algorithm, since the model members with a better prediction performance are assigned a higher BMA weight and a lower BMA variance, and vice versa.

(3) Results of the case studies based on two HEC-RAS models show that the estimates of BMA parameters based on both the normal and gamma conditional PDFs are close to

each other. As a result, BMA predictions under these two assumptions of the posterior distribution of the predictor variable yield a similar performance. However, it should be noted that the gamma conditional PDF is slightly better in terms of multiple evaluation metrics, which implies that the posterior distribution of the water stage data fits the gamma PDF better than the normal PDF for these two study areas.

(4) This study indicates that the M-H MCMC method with multiple independent chains is valid in estimating the BMA parameters and it is superior to the default EM algorithm in the BMA analysis. The application of MCMC method makes it easy and flexible to release and modify the strict assumption of the model residuals which are assumed to follow a normal distribution in the EM algorithm. A better fit of the conditional PDF of the variable of interest will produce more accurate and reliable BMA predictions. Most importantly, the MCMC sampling approach can provide a complete perspective and a full picture of the uncertainty in the BMA parameters through the corresponding posterior distributions.

As per authors' knowledge, this is the first study to apply the M-H MCMC method with multiple independent chains to estimating the BMA parameters of ensemble flood modeling. The basic M-H algorithm is used in this study since it is relatively easy to set up, the computational cost is feasible, and it requires few subjective selections of algorithm parameters. Some more appropriate prior and proposal distributions besides uniform and normal distributions can be attempted in the M-H algorithm to investigate any further improvement in the results. The slight improvement in the performance of the gamma conditional PDF in the case studies also imply

612 the potential room for better BMA predictions due to a better fit of specific hydrologic variables.
613 In the case that the sufficient knowledge of the shape of conditional PDFs is not available, some
614 transformation procedures (Box and Cox, 1964) might be required for pre-processing the raw
615 data, but whether the desired assumption of the conditional PDF is satisfied after the
616 transformation needs to be carefully validated. Moreover, some advanced and more complicated
617 MCMC algorithms, such as the Hamiltonian Monte Carlo (Gebraad et al., 2020; Ulzega and
618 Albert, 2022), particle filtering (Fan et al., 2022; Shen et al., 2022), reversible-jump MCMC
619 (Jiménez et al., 2016; Ouarda and El-Adlouni, 2011), etc., can also be attempted in this specific
620 issue of BMA parameter estimation, but the benefits of more efficient sampling and better
621 hydrologic predictions may not be guaranteed. Furthermore, more hydrologic and hydraulic
622 models of regions with different geomorphic features should be employed to extend and
623 reinforce the findings obtained in this study.

624 **Data Availability**

625 All the data used in this study, including the daily streamflow and water stage data, are publicly
626 available from the United States Geological Survey (USGS) website. FEMA's HEC-RAS
627 models used in this study are available from the Indiana Department of Natural Resources'
628 Hydrology and Hydraulics Model Library at
629 https://dnrmaps.dnr.in.gov/appsphp/model/index.php and FEMA's Estimated Base Flood
630 Elevation Viewer at https://webapps.usgs.gov/infrm/estbfe/.

631 **Acknowledgements**

632 The first author was partially supported by a grant from the Joint Transportation Research
633 Program (JTRP) at Purdue University.

References

Acharya, N., Shrivastava, N. A., Panigrahi, B., Mohanty, U. (2014). Development of an artificial neural network based multi-model ensemble to estimate the northeast monsoon rainfall over south peninsular India: an application of extreme learning machine. *Climate dynamics, 43*(5-6), 1303-1310.

Ajami, N. K., Duan, Q., Gao, X., Sorooshian, S. (2006). Multimodel combination techniques for analysis of hydrological simulations: Application to distributed model intercomparison project results. *Journal of Hydrometeorology, 7*(4), 755-768.

Bates, J. M., Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society, 20*(4), 451-468.

Beven, K. (2016). Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal, 61*(9), 1652-1665.

Beven, K., Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes, 6*(3), 279-298.

Box, G. E., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B, 26*(2), 211-243.

Brunner, G. W. (2016a). HEC-RAS hydraulic reference manual, Version 5.0. In: Hydrologic Engineering Center Davis CA.

Brunner, G. W. (2016b). HEC-RAS River Analysis System. User's Manual. Version 5.0. In: Hydrologic Engineering Center Davis CA.

Cao, C., Yan, B., Guo, J., Jiang, H., Li, Z., Liu, Y. (2021). A framework for projecting future streamflow of the Yalong River basin to climate change. *Stochastic Environmental Research Risk Assessment, 35*(8), 1549-1562.

Chattopadhyay, S., Chattopadhyay, G. (2008). Comparative study among different neural net learning algorithms applied to rainfall time series. *Meteorological Applications, 15*(2), 273-280.

Chitsazan, N., Tsai, F. T. C. (2015). A hierarchical Bayesian model averaging framework for groundwater prediction under uncertainty. *Groundwater, 53*(2), 305-316.

Darbandsari, P., Coulibaly, P. (2021). HUP-BMA: An Integration of Hydrologic Uncertainty Processor and Bayesian Model Averaging for Streamflow Forecasting. *Water Resources Research, 57*(10), e2020WR029433.

Dickinson, J. (1973). Some statistical results in the combination of forecasts. *Journal of the Operational Research Society, 24*(2), 253-260.

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs, 88*(4), 485-504.

Dottori, F., Alfieri, L., Bianchi, A., Salamon, P. (2021). River flood hazard maps for Europe and the Mediterranean Basin region, European Commission, Joint Research Centre (JRC)[Dataset]. In.

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.

Fan, Y., Shi, X., Duan, Q., Yu, L. (2022). Towards reliable uncertainty quantification for hydrologic predictions, Part I: Development of a particle copula Metropolis Hastings method. *Journal of Hydrology, 612*, 128163.

FEMA. (2015). FEMA's Estimated Base Flood Elevation Viewer. Retrieved from https://webapps.usgs.gov/infrm/estbfe/

FEMA. (2018). *Guidance for Flood Risk Analysis and Mapping*. Retrieved from https://www.fema.gov/sites/default/files/2020-02/Base_Level_Engineering_Guidance_Feb_2018.pdf

Gaume, E., Gaál, L., Viglione, A., Szolgay, J., Kohnová, S., Blöschl, G. (2010). Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites. *Journal of Hydrology, 394*(1-2), 101-117.

Gebraad, L., Boehm, C., Fichtner, A. (2020). Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo. *Journal of Geophysical Research: Solid Earth, 125*(3), e2019JB018428.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (1995). *Bayesian data analysis*: Chapman and Hall/CRC.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.

Huang, T., Merwade, V. (2023). Uncertainty Analysis and Quantification in Flood Insurance Rate Maps Using Bayesian Model Averaging and Hierarchical BMA. *Journal of Hydrologic Engineering, 28*(2), 04022038.

Huo, W., Li, Z., Wang, J., Yao, C., Zhang, K., Huang, Y. (2019). Multiple hydrological models comparison and an improved Bayesian model averaging approach for ensemble prediction over semi-humid regions.

Stochastic Environmental Research and Risk Assessment, 33(1), 217-238.
INDNR. (2018). Indiana Hydrology and Hydraulics Model Library, Indiana Department of Natural Resources. Retrieved from https://dnrmaps.dnr.in.gov/appsphp/model/index.php
Jafarzadegan, K., Abbaszadeh, P., Moradkhani, H. (2021). Sequential data assimilation for real-time probabilistic flood inundation mapping. *Hydrology and Earth System Sciences, 25*(9), 4995-5011.
Jiménez, S., Mariethoz, G., Brauchler, R., Bayer, P. (2016). Smart pilot points using reversible-jump Markov-chain Monte Carlo. *Water Resources Research, 52*(5), 3966-3983.
Kass, R. E., Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association, 90*(430), 773-795.
Kobarfard, M., Fazloula, R., Zarghami, M., Akbarpour, A. (2022). Evaluating the uncertainty of urban flood model using glue approach. *Urban Water Journal*, 1-16.
Krishnamurti, T., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., Surendran, S. (1999). Improved weather and seasonal climate forecasts from multimodel superensemble. *Science, 285*(5433), 1548-1550.
Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data* (Vol. 53): John Wiley & Sons Incorporated.
Liu, Z., Merwade, V. (2018). Accounting for model structure, parameter and input forcing uncertainty in flood inundation modeling using Bayesian model averaging. *Journal of Hydrology, 565*, 138-149.
Liu, Z., Merwade, V. (2019). Separation and prioritization of uncertainty sources in a raster based flood inundation model using hierarchical Bayesian model averaging. *Journal of Hydrology, 578*, 124100.
Luengo, D., Martino, L., Bugallo, M., Elvira, V., Särkkä, S. (2020). A survey of Monte Carlo methods for parameter estimation. *EURASIP Journal on Advances in Signal Processing, 2020*(1), 1-62.
Madadgar, S., Moradkhani, H. (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resources Research, 50*(12), 9586-9603.
Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of forecasting, 1*(2), 111-153.
Makridakis, S., Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management science, 29*(9), 987-996.
McLachlan, G. J., Krishnan, T. (2007). *The EM algorithm and extensions* (Vol. 382): John Wiley & Sons.
Merwade, V., Olivera, F., Arabi, M., Edleman, S. (2008). Uncertainty in flood inundation mapping: current issues and future directions. *Journal of Hydrologic Engineering, 13*(7), 608-620.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics, 21*(6), 1087-1092.
Moknatian, M., Mukundan, R. (2023). Uncertainty analysis of streamflow simulations using multiple objective functions and Bayesian Model Averaging. *Journal of Hydrology, 617*, 128961.
Newbold, P., Granger, C. W. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society: Series A, 137*(2), 131-146.
Nguyen, D. H., Kim, S.-H., Kwon, H.-H., Bae, D.-H. J. W. R. M. (2021). Uncertainty Quantification of Water Level Predictions from Radar-based Areal Rainfall Using an Adaptive MCMC Algorithm. *35*(7), 2197-2213.
Ouarda, T. B., El-Adlouni, S. (2011). Bayesian nonstationary frequency analysis of hydrological variables 1. *JAWRA Journal of the American Water Resources Association, 47*(3), 496-505.
Pappenberger, F., Beven, K., Horritt, M., Blazkova, S. (2005). Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations. *Journal of Hydrology, 302*(1-4), 46-69.
Pappenberger, F., Matgen, P., Beven, K. J., Henry, J.-B., Pfister, L. (2006). Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Advances in Water Resources, 29*(10), 1430-1449.
Parrish, M. A., Moradkhani, H., DeChant, C. M. (2012). Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation. *Water Resources Research, 48*(3).
Qi, H., Zhi, X., Peng, T., Bai, Y., Lin, C., Chen, W. (2021). Using stratified Bayesian model averaging in probabilistic forecasts of precipitation over the middle and lower Yangtze River region. *Meteorology and Atmospheric Physics, 133*(4), 961-972.
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.
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.
Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A. L., Troldborg, L. (2012). Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Advances in Water Resources, 36*, 36-50.
Reis Jr, D. S., Stedinger, J. R. (2005). Bayesian MCMC flood frequency analysis with historical information. *Journal of Hydrology, 313*(1-2), 97-116.
Rings, J., Vrugt, J. A., Schoups, G., Huisman, J. A., Vereecken, H. (2012). Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments. *Water Resources Research, 48*(5).
Robert, C. P., Casella, G., Casella, G. (1999). *Monte Carlo statistical methods* (Vol. 2): Springer.
Rojas, R., Batelaan, O., Feyen, L., Dassargues, A. (2010). Assessment of conceptual model uncertainty for the regional aquifer Pampa del Tamarugal–North Chile. *Hydrology and Earth System Sciences, 14*(2), 171-192.
Romanowicz, R., Beven, K. (2003). Estimation of flood inundation probabilities as conditioned on event inundation maps. *Water Resources Research, 39*(3).
Saul, L. K., Lee, D. D. (2002). Multiplicative updates for classification by mixture models. *Adv. Neural Inf. Process Systems, 14*, 897-904.
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. (1997). Methods for combining the outputs of different rainfall–runoff models. *Journal of Hydrology, 197*(1-4), 203-229.
Sharma, S., Ghimire, G. R., Talchabhadel, R., Panthi, J., Lee, B. S., Sun, F., Baniya, R., Adhikari, T. R. (2022). Bayesian characterization of uncertainties surrounding fluvial flood hazard estimates. *Hydrological Sciences Journal*, 1-10.
Shen, Y., Wang, S., Zhang, B., Zhu, J. (2022). Development of a stochastic hydrological modeling system for improving ensemble streamflow prediction. *Journal of Hydrology, 608*, 127683.
Singh, A., Mishra, S., Ruskauff, G. (2010). Model averaging techniques for quantifying conceptual model uncertainty. *Groundwater, 48*(5), 701-715.
Sloughter, J. M. L., Raftery, A. E., Gneiting, T., Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly weather review, 135*(9), 3209-3220.
Tebaldi, C., Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical transactions of the royal society A: mathematical, physical engineering sciences, 365*(1857), 2053-2075.
Teng, J., Jakeman, A. J., Vaze, J., Croke, B. F., Dutta, D., Kim, S. (2017). Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. *Environmental modelling software, 90*, 201-216.
Tian, D., He, X., Srivastava, P., Kalin, L. (2021). A hybrid framework for forecasting monthly reservoir inflow based on machine learning techniques with dynamic climate forecasts, satellite-based data, and climate phenomenon information. *Stochastic Environmental Research Risk Assessment*, 1-23.
Tsai, F. T.-C. (2010). Bayesian model averaging assessment on groundwater management under model structure uncertainty. *Stochastic Environmental Research and Risk Assessment, 24*(6), 845-861.
Ulzega, S., Albert, C. (2022). Bayesian parameter inference in hydrological modelling using a Hamiltonian Monte Carlo approach with a stochastic rain model. *EGUsphere*, 1-23.
Vats, D., Knudson, C. (2021). Revisiting the Gelman–Rubin Diagnostic. *Statistical Science, 36*(4), 518-529.
Von Bertalanffy, L. (1972). The history and status of general systems theory. *Academy of management journal, 15*(4), 407-426.
Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environmental Modelling & Software, 75, 273-316.
Vrugt, J. A., Diks, C. G., Clark, M. P. (2008). Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. Environmental fluid mechanics, 8(5-6), 579-595.
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*(1).
Wang, H., Wang, C., Wang, Y., Gao, X., Yu, C. (2017). Bayesian forecasting and uncertainty quantifying of stream flows using Metropolis–Hastings Markov Chain Monte Carlo algorithm. *Journal of Hydrology, 549*, 476-483.
Winkler, R. L. (1989). Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting, 5*(4), 605-609.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of statistics*, 95-103.

Xie, S., Wu, W., Mooser, S., Wang, Q., Nathan, R., Huang, Y. (2021). Artificial neural network based hybrid modeling approach for flood inundation modeling. *Journal of Hydrology, 592*, 125605.

Zhao, G., Bates, P., Neal, J., Pang, B. (2021). Design flood estimation for global river networks based on machine learning models. *Hydrology and Earth System Sciences, 25*(11), 5981-5999.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology, 598*, 126266.