Title
Comment on “Improving Bayesian Model Averaging for Ensemble Flood Modeling Using Multiple Markov Chains Monte Carlo Sampling”

Journal
Water Resources Research, 60(11)

ISSN
0043-1397

Author
Vrugt, Jasper A

Publication Date
2024-11-01

DOI
10.1029/2023wr036862

Water Resources Research

COMMENT
10.1029/2023WR036862

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

Jasper A. Vrugt[1]

[1]Department of Civil and Environmental Engineering, University of California, Irvine, CA, USA

Key Points:

- This comment criticizes the work of Huang & Merwade (2023) or HM23 on postprocessing water stage model predictions using Bayesian Model Averaging (BMA)
- The methodology of HM23 is fundamentally flawed and provides erroneous estimates of BMA parameter and predictive uncertainty
- The DREAM$_{(BMA)}$ algorithm in the MODELAVG toolbox of Vrugt (2018) guarantees a robust Bayesian training of mixture models

Correspondence to:

J. A. Vrugt,
jasper@uci.edu

Citation:

Vrugt, J. A. (2024). Comment on "improving Bayesian Model averaging for ensemble flood modeling using multiple Markov Chains Monte Carlo sampling". *Water Resources Research*, *60*, e2023WR036862. https://doi.org/10.1029/2023WR036862

Received 3 DEC 2023
Accepted 13 SEP 2024

Abstract Huang and Merwade (2023), https://doi.org/10.1029/2023wr034947, hereafter conveniently referred to as HM23, wrongly claim improvement of their method for postprocessing multi-model water stage predictions using Bayesian Model Averaging (BMA). Their results show all signs of a flawed implementation of the Metropolis algorithm. In this comment I will point out the many mistakes and shortcomings of the BMA methodology of HM23. Their method is deficient, inefficient and ineffective and wrongly quantifies BMA model parameter and predictive uncertainty. Furthermore, HM23 misrepresent BMA literature, articulate a poor understanding of Markov chain Monte Carlo methods and misuse the autocorrelation function for monitoring convergence of the sampled Markov chains. A proper implementation of the random walk Metropolis algorithm would have led HM23 to substantially different results and findings about their ensemble of water stage predictions. The MODELAVG toolbox of Vrugt (2018) https://www.researchgate.net/publication/299458373_MODELAVG_A_MATLAB_Toolbox_for_Postprocessing_of_Model_Ensembles satisfies all requirements of HM23 and provides robust estimates of BMA model parameter and prediction uncertainty for symmetric, skewed and truncated conditional forecast distributions of the ensemble members.

Plain Language Summary This comment criticizes the work of Huang and Merwade (2023), https://doi.org/10.1029/2023wr034947 or HM23 on postprocessing forecast ensembles of flood models using Bayesian Model Averaging (BMA). The flawed implementation and use of statistical methods by HM23 has led to nonsensical results, erroneous findings, and invalid statements and conclusions. I will point some of the main flaws and demonstrate how existing methods would have led to fundamentally different findings for their ensemble of water stage predictions.

1. Introduction

Huang and Merwade (2023), hereafter referred to as HM23, claim improvement to Bayesian model averaging (BMA) for postprocessing multi-model ensembles of water stage predictions. I have great concerns about this claim and the methodology presented by HM23. This flawed methodology and inadequate implementation of the BMA method has led HM23 to nonsensical results, erroneous findings and many invalid statements and conclusions. Strictly speaking, the HM23 ensemble is not a multi-model ensemble as the water stage predictions are from only one hydraulic model, HEC-RAS, by varying channel roughness, geometry, and upstream flow.

This comment documents my concerns about HM23. To provide context for this critique, I first summarize the BMA methodology in Section 2. Then, in Section 3, I present my comments and concerns about the work of HM23. Section 4 summarizes my main concerns.

2. Bayesian Model Averaging

Standard statistical practice usually ignores model uncertainty. It is common to select a model from some class of models and then proceed as if this model had generated the data. This promotes over-confident inferences and decision-making. Bayesian Model Averaging, or BMA, is a standard approach to inference in the presence of multiple competing statistical models. This method was introduced by Hoeting et al. (1999) and provides a coherent mechanism to account for model uncertainty given that one of the models considered is *true*, or alternatively that the available model class admits the data-generating process. Bayesian Model Averaging possesses a range of theoretical optimality properties and has shown good performance in a variety of simulated and real data situations (Raftery & Zheng, 2003).

Raftery et al. (2005) extended the BMA method to forecasts of dynamical models. Called ensemble BMA, this statistical postprocessing method aims to generate sharp calibrated probabilistic forecasts from an ensemble of point forecasts that honor the spread-skill relationship (Raftery et al., 2005; Vrugt, 2024). Strictly speaking, the ensemble BMA method does not have a formal Bayesian interpretation, nor do we want to assume that any of the models in hand is *true* in any sense. Several authors have therefore criticized the label of ensemble BMA assigned by Raftery et al. (2005) to their methodology. Kernel dressing (Bröcker & Smith, 2008) and Gaussian mixture model averaging (Höge et al., 2019) replace individual ensemble members by kernel functions and are a more accurate reflection of the probability density function (PDF) averaging method of Raftery et al. (2005). I did not make this distinction in my previous publications on ensemble BMA dating back to Vrugt et al. (2006), Vrugt and Robinson (2007) and Vrugt, ter Braak, et al. (2008), but one of the reviewers of this commentary rightly pointed out the importance of correctly labeling statistical post-processing methods to avoid potential confusion about the use and interpretation of BMA in a hydrological context. I think that mixture model averaging or MMA (not to be confused with Mallow's model averaging of Hansen, 2007) best summarizes the ensemble BMA methodology and leaves open the choice of conditional PDF for the ensemble members.

To provide context for my comments about HM23, I briefly review the ensemble BMA method. In doing so, I hold on to the misnomer ensemble BMA as this terminology was used by HM23.

2.1. Mathematical Notation

I use a lowercase italic font (a) for scalars, a lowercase bold font ($\mathbf{a}$) for vectors and an uppercase bold font ($\mathbf{A}$) for matrices. The symbol $\tilde{y}$ is used for a measurement of the quantity y of interest; therefore, I write $\tilde{\mathbf{y}} = \left(\tilde{y}_1, \dots, \tilde{y}_n\right)^\top$ for a collection of materialized outcomes, where the symbol $\top$ denotes transpose. An ensemble of K models is used to predict quantity y. The point forecasts of each model are stored in a $n \times 1$ vector $\mathbf{y} = \left(y_1, \dots, y_n\right)^\top$ and these column vectors are combined into a $n \times K$ matrix $\mathbf{Y}$, where y_{tk} stores the prediction of the kth model $k = 1, \dots, K$ at $t = 1, \dots, n$. Without loss of generality, I ignore the spatial coordinates and restrict attention to a time series of outcomes, and thus t signifies time. I designate a PDF with lowercase f and a cumulative distribution function (CDF) with uppercase F. Thus, $f_{\mathcal{N}}(x, \mu, \sigma^2)$ and $F_{\mathcal{N}}(x, \mu, \sigma^2)$ signify the PDF and CDF of the normal distribution with mean μ and variance σ^2, respectively. The vertical bar "|" denotes the conditional expectation. Thus, $p(x|\tilde{\mathbf{y}})$ is the conditional PDF of x given data $\tilde{\mathbf{y}} = \left(\tilde{y}_1, \dots, \tilde{y}_n\right)^\top$ with $p(x|\tilde{\mathbf{y}}) \geq 0$ and $\int_{\mathcal{X}} p(x|\tilde{\mathbf{y}}) \mathrm{d}x = 1$.

2.2. Theory

Let β_k and $f_{tk}(y|\mu_{tk}, \psi_k)$ denote the weight and PDF of the kth model of the ensemble at time t. The mean of the forecast distribution $f_{tk}(y|\mu_{tk}, \psi_k)$ is affixed to the prediction y_{tk} of the kth model. The PDF of the multi-model forecast distribution at time t now equals a mixture distribution of the models' conditional PDFs

$$g_t(y|\mathbf{y}_t, \boldsymbol{\beta}, \boldsymbol{\psi}) = \sum_{k=1}^{K} \beta_k f_{tk}(y|y_{tk}, \psi_k), \tag{1}$$

and weighted-average forecast of the ensemble

$$\dot{g}_t(\boldsymbol{\beta}, \mathbf{y}_t) = \sum_{k=1}^{K} \beta_k y_{tk}, \tag{2}$$

with weights $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_K\right)^\top$ restricted to the probability simplex

$$\Delta^{K-1} = \left\{\boldsymbol{\beta} \in \mathbb{R}^K : \beta_1 + \cdots + \beta_K = 1;\ \beta_k \geq 0 \text{ for } k = 1, \dots, K\right\}, \tag{3}$$

and shape parameters ψ_k of each model's predictive PDF assembled in the array $\boldsymbol{\psi} = \left(\psi_1, \dots, \psi_K\right)^\top$. The weighted-average ensemble prediction $\dot{g}_t$ is a deterministic forecast in its own right, whose performance can be

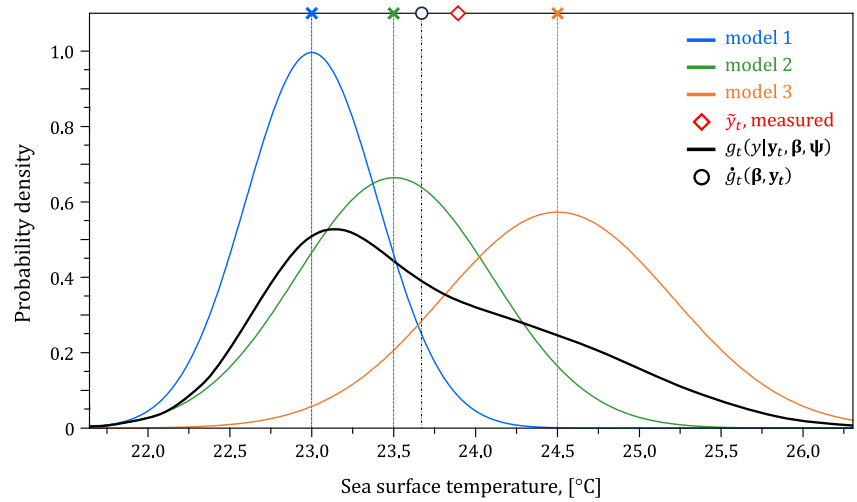

Figure 1. Schematic illustration of a $K = 3$ member ensemble for the sea surface temperature and verifying measurement, "red." The mixture density $g_t(y|\cdot)$ is indicated with the solid black line and equivalent to a weighted average of the conditional PDFs $f_{tk}(y|\cdot)$ of the members of the ensemble (displayed with solid blue, green and orange lines). The colored crosses "×" portray the individual model predictions and the weighted average point forecast $\hat{g}_t(\boldsymbol{\beta}, \boldsymbol{\psi})$ is indicated with the "black" symbol. This deterministic point forecast is closer to the measured sea surface temperature "red"than any of the model predictions. The mixture density can be used to compute $\gamma = 100(1 - \alpha)\%$ prediction intervals of the sea surface temperature at any desired significance level $\alpha = 0.1$, 0.05 or 0.01.

compared to the ensemble mean or median and different models of the ensemble. The black line in Figure 1 illustrates the BMA density of Equation 1 when each model's predictive PDF

$$f_{tk}\left(y|\, y_{tk},\, \sigma_k^2\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2}\sigma_k^{-2}\left(y - y_{tk}\right)^2\right), \tag{4}$$

is a normal distribution $\mathcal{N}(y_{tk},\, \sigma^2)$ centered at its point forecast y_{tk} and with variance σ_k^2.

Then, $g_t(y|\,\mathbf{y}_t,\, \boldsymbol{\beta},\, \boldsymbol{\psi})$ is equal to a Gaussian mixture distribution. The weights $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_K\right)^\top$ and variances $\boldsymbol{\sigma}^2 = \left(\sigma_1^2, \ldots, \sigma_K^2\right)^\top$ of each model must be estimated from a training data record of materialized outcomes, $\tilde{\mathbf{y}} = \left(\tilde{y}_1, \ldots, \tilde{y}_n\right)^\top$. This involves the estimation of $d = K + 1$ parameters if a common constant variance $\sigma_1^2 = \ldots = \sigma_K^2$ is assumed for all models of the ensemble and, thus, $\boldsymbol{\Phi} = \left(\beta_1, \ldots, \beta_K, \sigma^2\right)^\top$ or $d = 2K$ parameters if a constant but model-dependent variance $\sigma_k^2;\ k = 1, \ldots, K$ is used and $\boldsymbol{\Phi} = \left(\beta_1, \ldots, \beta_K, \sigma_1^2, \ldots, \sigma_K^2\right)^\top$. We can also assume a heteroscedastic forecast variance $\sigma_{tk}^2 = c_k\, y_{tk}$ and estimate $c_k;\ k = 1, \ldots, K$ instead (J. M. L. Sloughter et al., 2007; Vrugt, Diks, & Clark, 2008), of which more later. Equations 1 and 2 are also known as the BMA forecast density and BMA model forecast, respectively (Raftery et al., 2005).

The weights are restricted to the unit simplex $\Delta^{K-1} \in \mathbb{R}_+^K$ otherwise the mixture density of Equation 1 does not integrate to one and, possibly, could even produce negative densities. The BMA weight of each ensemble member can then be viewed as each model's relative contribution to predictive skill over a training period. Lower l_t and upper u_t limits of the $\gamma = 100(1 - \alpha)\%$ prediction interval $[l_t, u_t]_\alpha$ can be calculated from the CDF of the BMA mixture density

$$G_t(y|\,\mathbf{y}_t,\, \boldsymbol{\beta},\, \boldsymbol{\psi}) = \sum_{k=1}^{K} \beta_k\, F_{tk}(y|\, y_{tk},\, \psi_k) \tag{5}$$

so that $G_t(l_t|\,\mathbf{y}_t,\, \boldsymbol{\beta},\, \boldsymbol{\psi}) = \alpha/2$ and $G_t(u_t|\,\mathbf{y}_t,\, \boldsymbol{\beta},\, \boldsymbol{\psi}) = 1 - \alpha/2$, where $\alpha \in (0, 1)$ is the significance level. The BMA prediction intervals in Figures 9b–9c and 12a–12d of HM23 suggest a problem with the computation of the prediction intervals. Therefore, in Section 3.4, I will revisit this topic of how to calculate the prediction limits.

Successful implementation of the BMA method requires estimates of the weights $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\top$ and shape parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$ of the conditional PDFs of the members of the ensemble. If we assume that the forecast errors of the models are independent, then the d-vector $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \boldsymbol{\psi})$ of BMA weights $\boldsymbol{\beta}$ and shape parameters $\boldsymbol{\psi}$ can be determined by maximization of the likelihood of the mixture distribution in Equation 1 (Raftery et al., 2005)

$$L(\boldsymbol{\Phi}|\mathbf{Y}, \tilde{\mathbf{y}}) = \prod_{t=1}^{n} \sum_{k=1}^{K} \beta_k f_{tk}(\tilde{y}_t | y_{tk}, \psi_k). \tag{6}$$

where $\boldsymbol{\beta} \in \Delta^{K-1}$ and $\boldsymbol{\psi} \in \mathbb{R}^K$. The summation is for all models $k = 1, \ldots, K$ of the ensemble and the product operator cycles through all observations $\tilde{y}_1, \ldots, \tilde{y}_n$ of the training data set. Raftery et al. (2005) recommends using the Expectation Maximization (EM) algorithm (Dempster et al., 1977; McLachlan & Krishnan, 2008) for estimating the BMA weights $\boldsymbol{\beta}$ and variances $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_k^2)$ (shape parameters). This method is relatively easy to implement, computationally efficient and the algorithmic steps constrain the BMA weights to the unit simplex, Δ^{K-1}. But convergence of the EM algorithm to the maximum likelihood BMA weights and variances $\hat{\boldsymbol{\Phi}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}})$ cannot be guaranteed and becomes more and more difficult with increasing size K of the ensemble. Furthermore, for conditional PDFs other than the normal distribution in Equation 4, the EM algorithm does not have a convenient analytic expression for the variances $\sigma_1^2, \ldots, \sigma_K^2$ and/or other shape parameters and so they must be estimated numerically by optimizing the BMA likelihood function (6) using the weights $\beta_1, \ldots, \beta_K$ from the expectation step. The `MODELAVG` toolbox of Vrugt (2018) presents a general implementation of the EM algorithm for the normal, lognormal, truncated normal and gamma conditional distributions. The maximum likelihood solution of the EM algorithm serves as benchmark for the posterior BMA distribution of weights and shape parameters.

3. Comments and Concerns

In this Section 1 articulate my concerns about the work of HM23. The mistakes are too numerous to address all of them. I organize my comments in different subsections to systematically address the most critical flaws of their methodology. First, I will provide some general comments about the premise of HM23 and their misleading characterization of published literature and misuse of statistical methods before discussing the deficiencies and shortcomings of their BMA methodology.

3.1. Misleading Description of Literature and Methods

At numerous places in their manuscript, HM23 misrepresent BMA literature and methods. For example, in the abstract they wrongly assert that "…*However, the uncertainty in BMA parameters with fixed values, which are usually obtained from the Expectation-Maximization (EM) algorithm, has not been adequately investigated in BMA-related applications over the past few decades*." Leaving aside the curious wording "…*the uncertainty in BMA parameters with fixed values*…", this statement is false. If uncertainty is meant to refer to the temporal stability of the BMA parameters, then HM23 ignore the contributions of (among others) Raftery et al. (2005), J. M. L. Sloughter et al. (2007), and Vrugt and Robinson (2007) who studied the relationship between the length n of the training data period and the temporal stability and robustness of the parameters of the BMA model for multi-model ensembles of sea level pressure and temperature, wind speed, and river discharge. If, instead, uncertainty is meant to refer to the confidence intervals of the BMA parameters, then HM23 overlook the contribution of Vrugt, Diks, and Clark (2008) who framed BMA model training in a Bayesian context and presented a general purpose methodology to facilitate posterior exploration of the BMA weights and shape parameters. HM23 repeat the premise of Vrugt, Diks, and Clark (2008) but do not adequately inform readers about the results and/or outcomes of this paper. Specifically, the DREAM$_{(\mathrm{BMA})}$ algorithm will yield credible regions of the BMA weights and shape parameters that HM23 desire; more about this later.

On Page 2 of the introduction, HM23 write "…*However, the reliability associated with BMA parameters has not received enough attention*." I respectfully beg to differ in opinion with the authors. Vrugt, Diks, and Clark (2008) compared the performance of the EM algorithm and the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm for estimating BMA weights and shape parameters. The abstract of this paper reads "…*Simulation*

experiments using 48-hr ensemble data of surface temperature and multi-model stream ow forecasts show that both methods produce similar results, and that their performance is unaffected by the length of the training data set. However, Markov chain Monte Carlo (MCMC) simulation with DREAM is capable of efficiently handling a wide variety of BMA predictive distributions, and provides useful information about the uncertainty associated with the estimated BMA weights and variances." The robustness and reliability of the BMA parameters has been studied in sufficient detail (Raftery et al., 2005; J. M. L. Sloughter et al., 2007; Vrugt & Robinson, 2007; Vrugt, Diks, & Clark, 2008; J. M. Sloughter et al., 2010) (among others) and the DREAM$_{(BMA)}$ algorithm offers a general methodology for post-processing forecast ensembles using the BMA method. Then, Vrugt, Diks, and Clark (2008) also clarifies a fundamental misunderstanding of the BMA method by HM23. They confuse confidence intervals and prediction intervals of the BMA model. Other than HM23 suggests, the predictive uncertainty of the BMA mixture model requires knowledge only of the maximum likelihood values $\hat{\mathbf{\Phi}}$ of the BMA weights and shape parameters.

The next paragraph of HM23 acknowledges the capabilities of the EM algorithm "…*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 & Krishnan, 2008; Vrugt, ter Braak, et al., 2008; Vrugt et al., 2008, 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, ter Braak, et al., 2008; Vrugt et al., 2008, 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*." but repeats the arguments of Vrugt, Diks, and Clark (2008) to justify publication of their MCMC-based BMA parameter estimation method. HM23 cite and acknowledge the work of Vrugt, Diks, and Clark (2008) but fail to mention that the DREAM$_{(BMA)}$ methodology overcomes the aforementioned limitations of the EM algorithm and provides a robust and efficient characterization of the posterior distribution of the BMA model parameters. Furthermore, HM23 conveniently ignore the contributions by J. M. L. Sloughter et al. (2007) and J. M. Sloughter et al. (2010) on postprocessing of forecast ensembles of lower-tail bounded variables such as wind speed and rainfall. Their formulation of the EM algorithm for the gamma distribution would have sufficed for the ensemble of water stage predictions analyzed by HM23.

In the next paragraph of their introduction, HM23 misrepresent BMA literature and methods and lead readers into thinking that available training methods are not powerful enough to robustly estimate the weights and shape parameters of the BMA mixture distribution. Specifically, HM23 write "…*In a previous study, differential evolution adaptive metropolis algorithm (Vrugt,* 2016; Vrugt, ter Braak, et al., 2008; Vrugt et al., 2008, 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*." This description of the DREAM algorithm is wrong and nonsensical. In the first place, the DREAM algorithm of Vrugt (2016) does not introduce additional parameters nor inflate the uncertainty of the BMA parameters. The number of chain pairs δ and the jump rate $\gamma = 2.38/\sqrt{2\delta d}$ control the convergence rate of the sampled chains and not the BMA posterior distribution itself. The default values of these and other algorithmic variables have shown to work well for a large range of problems and target dimensionalities (Laloy & Vrugt, 2012; Lochbühler et al., 2015; Vrugt, 2016). This is easy to confirm with numerical experiments and a comparison of the maximum likelihood solutions derived from the DREAM$_{(BMA)}$ and EM algorithms. Then, with respect to "…*an overall variance across different model members was assumed*.," the DREAM$_{(BMA)}$ algorithm in the `MODELAVG` toolbox (Vrugt, Diks, & Clark, 2008) allows users to select among four different options for the variance of the predictive distributions of the ensemble members, including (a) a constant common variance, $\sigma_1^2 = \ldots = \sigma_K^2$ (b) a constant single variance, $\sigma_k^2; k = 1 \ldots, K$ (c) a nonconstant common variance $\sigma_{tk}^2 = c \cdot y_{tk}$ and (d) a nonconstant single variance, $\sigma_{tk}^2 = c_k\, y_{tk}$. Strictly proper scoring rules such as the quadratic, logarithmic, spherical, and continuous ranked probability scores may be used to determine which variance treatment maximizes the performance of the distribution forecasts of the BMA model (Vrugt, 2023; Vrugt et al., 2006). Lastly, "…*the autocorrelation of the samples in each chain was not evaluated, which would add more uncertainty to the estimates of BMA parameters.*" These types of remarks by HM23 signal a poor understanding of statistical concepts and methods. The autocorrelation functions of the chain samples say

nothing at all about parameter uncertainty and should not be used as HM23 do to assess chain convergence. This is discussed later. The autocorrelation function is only of secondary interest, for example, to determine the effective sample size of the collection of posterior realizations. The MODELAVG toolbox visualizes the sample autocorrelation functions of the BMA parameters in each of the sampled Markov chains, but this information is rarely used. In addition, chain thinning is an effective remedy for reducing the autocorrelation between the values of the sampled parameters.

HM23 continue to misrepresent BMA literature on Page 6 by writing "…*A few previous studies pointed out that the EM algorithm can only converge to the local optimum rather than the global optimal results* (Duan et al., 2007; McLachlan & Krishnan, 2008; Vrugt, ter Braak, et al., 2008; Vrugt et al., 2008, 2008), *but this issue has not been addressed adequately in the literature*." The last part of this statement "……, *but this issue has not been addressed adequately in the literature*." is not true. The authors fail to inform the audience that the DREAM$_{(BMA)}$ algorithm of Vrugt, Diks, and Clark (2008) does not suffer premature convergence and provides users a large sample approximation of the posterior distribution of the BMA weights and shape parameters. HM23 have also overlooked the MODELAVG toolbox of Vrugt (2018) which implements the DREAM$_{(BMA)}$ algorithm in MATLAB and offers users many different options for postprocessing forecast ensembles using BMA and other model averaging methods. Users can choose between different parametric forms for the conditional PDFs of the ensemble members and decide on a constant or nonconstant variance for the predictive PDFs. Performance metrics, prediction quantiles and scoring rules of the BMA distribution forecasts are computed, visualized and returned to the user. This toolbox would have led HM23 to substantially different conclusions about their methodology and multi-model ensemble of water stage predictions.

On the same Page, HM23 write "…*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*." I fail to see how the use of independent chains and a non-adaptive proposal distribution by HM23 is an improvement over much more efficient multi-chain adaptive MCMC methods. Then, the claim of ergodicity (irreducibility) by HM23 appeals to fundamental theory about the equilibrium distribution of a Markov chain but is rather meaningless by itself. A Markov chain must also satisfy detailed balance for its limiting distribution to equal the posterior distribution of the BMA model parameters. Yet, HM23 do not address detailed balance of their sampled chains. This begs the question whether the chains of HM23 even sample the correct target distribution. Particularly, how did HM23 enforce the weights to lie on the probability simplex $\Delta^{K-1} \in \mathbb{R}_{+}^{K}$. A weight normalization step will violate detailed balance of the sampled chains.

Then, in the last sentence of their abstract HM23 conclude that "…*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 both deterministic and probabilistic evaluation metrics as well as algorithm flexibility*." This conclusion does not make much sense. Strictly speaking, the authors have not shown that their multi-chain method has desirable advantages. Their implementation of the Metropolis algorithm is deficient. Multi-chain methods do not provide more information about parameter uncertainty, rather they make it easier to accurately sample the posterior parameter distribution. This was already shown in Vrugt, Diks, and Clark (2008) and has to do with adaptation of the proposal distribution and how trial moves may be created in the Markov chains. Single-chain methods will yield the same results, but are less robust against local optima and multi-modal posterior surfaces. Then, HM23 wrongfully assert that, "…*its performance is better than the default EM algorithm in terms of both deterministic and probabilistic evaluation metrics as well as algorithm flexibility*." This remark highlights an inadequate implementation and/or use of the EM algorithm. If properly implemented for the normal and gamma conditional PDFs, the EM algorithm will find the same solution of the maximum likelihood BMA weights and shape parameters as MCMC-based inference. This is certainly true for multiple different EM trials and the ensemble sizes considered in HM23. As a result, both methods should yield approximately equal performance metrics, prediction intervals and scoring rules of the BMA distribution forecasts. Now, the EM algorithm is much more difficult to implement for conditional PDFs that have more than one shape parameter. This is the advantage of MCMC-based inference of the BMA model parameters. But this point was already made in Vrugt, Diks, and Clark (2008).

3.2. Misuse of Statistical Methods and Terminology

In the abstract HM23 write "…*Moreover, the normal proposal distribution used in the M-H algorithm can yield narrower distributions for BMA weights than those from the uniform proposal distribution*." This conclusion is nonsensical and testifies to a poor understanding of the Metropolis algorithm. The target distribution of the BMA weights and shape parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$, respectively, is solely determined by the BMA likelihood function $L(\boldsymbol{\beta}, \boldsymbol{\psi}|\mathbf{Y}, \widetilde{\mathbf{y}})$ and thus is invariant to the choice of jump distribution. The proposal distribution is just a vehicle that helps the Markov chain transition from an arbitrary initial state to the target distribution. Whether this vehicle is efficient or not is of secondary importance and should not affect the posterior parameter estimates. On a related note, HM23 refer to their MCMC method as a Metropolis-Hastings algorithm. This is not accurate. They use a symmetric proposal distribution and thus implement the Metropolis algorithm.

On Page 7, HM23 write "…*Since the dimensions and units of the 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*." This statement articulates unfamiliarity with the MCMC literature, particularly with adaptive proposal estimation methods. A well-known remedy to this problem is to adaptively tune the jump rate of the transition density so as to achieve a desired acceptance rate of about 23%–24% (Roberts & Rosenthal, 2001). This is the foundation of the adaptive Metropolis algorithm of Haario et al. (2005) and substantially enhances the convergence rate to the posterior target distribution. Alternatively, one can periodically adapt the covariance matrix of the proposal distribution using the chain samples. The entire history of the sampled chains must be used to guarantee an ergodic Markov chain. This approach would dramatically increase the acceptance rate of candidate points and change the results of HM23.

Figure 3 of HM23 signals a very poor sampling efficiency of their MCMC method. Only a handful of candidate points are accepted in the chain of $T = 100,000$ samples. The culprit is a bad proposal distribution. This makes MCMC simulation highly ineffective and casts serious doubts on the supposed posterior realizations of the BMA weights and shape parameters. What is more, the number of chains N and value of T are case study dependent. For example, if one were to change the conditional PDF of the ensemble members to a generalized normal, generalized Pareto or generalized extreme value distribution then the number of BMA parameters would increase to $d = 3K$. In the present implementation, one cannot expect $N = 100$ chains (=samples) to adequately approximate the posterior distribution of the BMA weights, scale and shape parameters.

On repeated occasions, HM23 demonstrate a poor understanding of MCMC methods. Their results show all the signs of a flawed implementation of the Metropolis algorithm, but HM23 defend their findings in the results and discussion section and explain the complete immobility of their sampled chains with local maxima of the BMA likelihood function. This explanation may suit the premise of HM23 but is erroneous. Specifically, on Page 8, they write "…*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.*" A Markov chain should not reject all its candidate points. This alerts to an inadequate scale and/or orientation of the proposal distribution and poor implementation of the Metropolis algorithm. HM23 continue this thread on Page 8 with "…*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 2,000 iterations. Thus, 2,000 samples are generated in each chain and the number of chains is set to be 100.*" No chain should settle on a single point in the BMA parameter space but rather explore the surface of the likelihood function in vicinity of its current state. An adequate implementation of the Metropolis algorithm would have solved the aforementioned problems and lead to substantially different conclusions.

In Figure 4 and 5, HM23 present sample autocorrelation functions for an unspecified BMA parameter (weight or variance) using the N last chain states. These two graphs are meaningless as the different Markov chains evolve independently and do not share information. Thus, the final chain states have nothing to do with each other. The autocorrelation function should only be computed for the samples within a Markov chain as they relate to each other. This misuse of statistical methods and concepts leads to erroneous results and wrong claims. For example, on Page 8 HM23 conclude that "…*The trace plots and autocorrelation functions (ACF) of the BMA weights and*

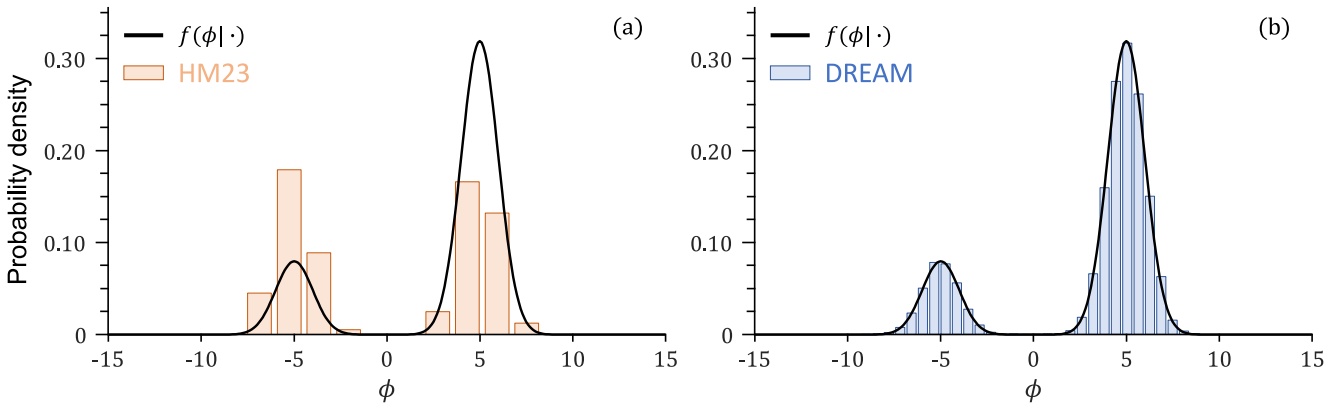

Figure 2. (a) Application of the Markov chain Monte Carlo sampling methodology of HM23 to a bimodal target distribution (solid black line). The orange bins provide an empirical estimate of the density of parameter ϕ; (b) Marginal posterior distribution of parameter ϕ derived from the DREAM algorithm.

variances …(snip) …indicate that the samples drawn from multiple independent MCMC chains are mixing very well and the autocorrelation is quickly dropping within the 5% significant interval (shown in the light blue region in the ACF in Figures 4 and 5*), which means the entire space of the BMA parameters has been fully explored, and individual samples in the MCMC chain follow stationary and independent identical distributions.*" The final samples of the chains are unrelated and do not tell us anything about the mixing of the chains. The authors state that the chains mix well, whereas Figure 3 suggests the opposite. The single chain shown does not explore the target distribution but rather converges to a point. Thus, if we were to follow the algorithmic recipe of HM23 and repeat this exercise for N chains, there will be no mixing between the chains at all. Mixing of the chains can only be assessed visually as I will demonstrate later. Furthermore, a low autocorrelation between the sampled chains does not tell us anything about whether the chains have been able to thoroughly explore the parameter space or whether the samples have independent identical distributions. The authors misuse the sample autocorrelation function as a convergence diagnostic.

Then, the use of MCMC simulation cannot go without a proper quantitative assessment of the convergence of the sampled chains. At a minimum HM23 should have plotted the sampled traces of the individual chains. This would

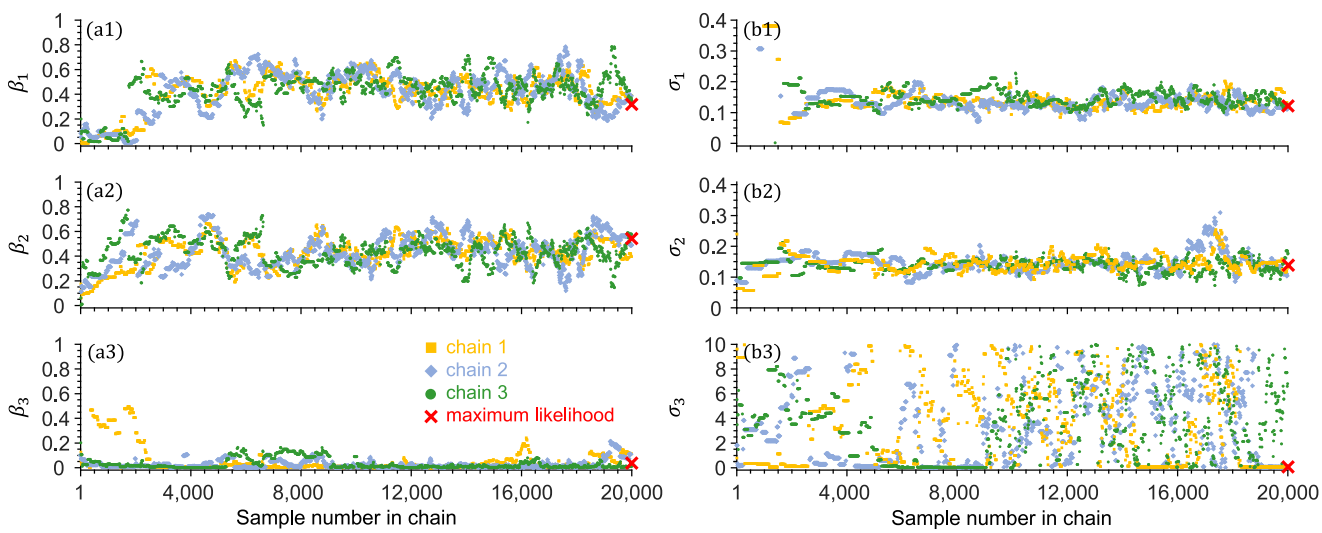

Figure 3. 10-model ensemble of water stage predictions of HM23: Traces of the Bayesian Model Averaging (BMA) weights (β_1, β_2 and β_3) and BMA standard deviations (σ_1, σ_2 and σ_3) of the first three ensemble members derived from the DREAM$_{(BMA)}$ algorithm of Vrugt, Diks, and Clark (2008) using the `MODELAVG` toolbox (Vrugt, 2018). The Markov chains are color coded. The red crosses portray the values of the BMA weights and standard deviations that maximize the likelihood of Equation 6. Only the first two ensemble members effectively contribute to the BMA model and receive an appreciable weight. All other ensemble members attain near zero weights $\beta_3, \ldots, \beta_{10}$ and, thus, their standard deviations $\sigma_3, \ldots, \sigma_{10}$ cannot be determined. This explains the sampled chain trajectories of σ_3 and the almost uniform marginal posterior distribution (not shown) of $\sigma_3, \ldots, \sigma_{10}$.

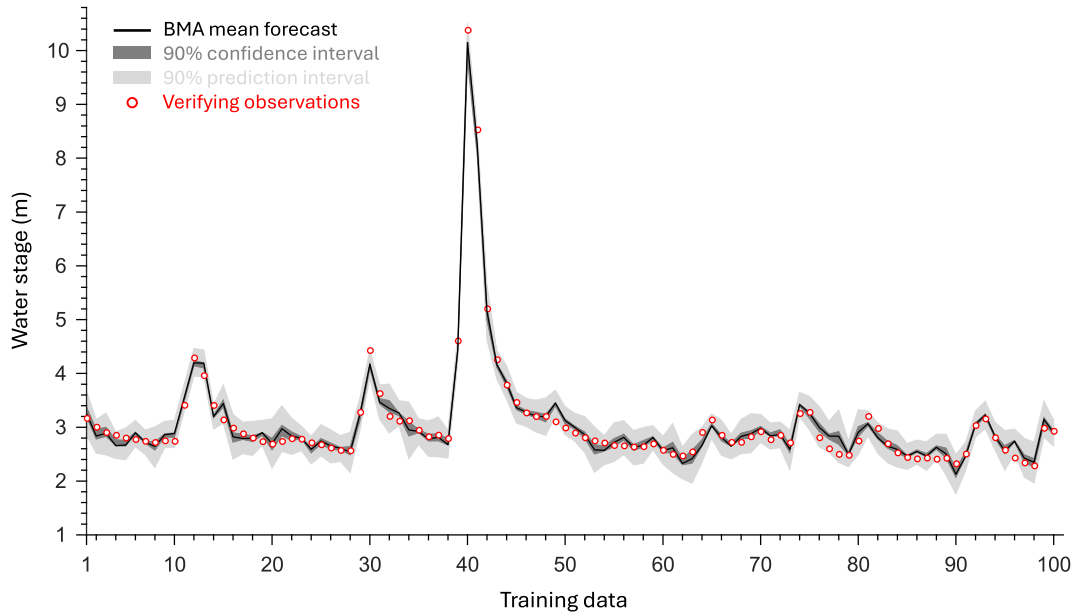

Figure 4. 10-model ensemble of water stage predictions of HM23: 90% Bayesian Model Averaging confidence (dark gray) and prediction (light gray) intervals derived from the `MODELAVG` toolbox of (Vrugt, 2018) using a normal forecast distribution with constant forecast variance.

have given insights into the mixing of the chains. But multi-chain methods admit convergence monitoring through application of the univariate $\hat{R}$ and multivariate $\hat{R}^d$ scale reduction factors (Brooks & Gelman, 1998; Gelman & Rubin, 1992). These diagnostics compare the within-chain and between-chain (co)variances of the sampled parameters. Convergence is achieved when $\hat{R}_j \leq 1.2$ for all $j = 1, \ldots, d$ or $\hat{R}^d \leq 1.2$. After, the chains have converged to a limiting distribution, the sample autocorrelation function of the parameters can be investigated. Chain thinning is a commonly used remedy to reduce autocorrelation between the sampled parameter values (Vrugt, 2016).

The MCMC-based method of HM23 and summarized in the flowchart of Figure 2 is predicated on the assumption that the collection of final chain states makes up the posterior distribution of the BMA weights and shape parameters. This is true in theory, but the MCMC implementation of MH23 is woefully inadequate and wrong. I will demonstrate this in more detail next.

3.3. BMA Model Parameter Estimation

HM23 resort to Markov chain Monte Carlo (MCMC) simulation to determine the BMA weights and shape parameters, $\mathbf{\Phi} = (\beta_1, \ldots, \beta_K, \psi_1, \ldots, \psi_K)^\top$. They simulate N independent Markov chains $\{\mathbf{\Phi}^j_{(1)}, \mathbf{\Phi}^j_{(2)}, \ldots, \mathbf{\Phi}^j_{(T)}\}$ of T samples with the Metropolis algorithm (and not the Metropolis-Hastings method, as the authors write) and use the last sample of each chain $j = 1, \ldots, N$ to approximate the posterior distribution of the BMA parameters $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$. This implementation is highly inefficient, nevertheless, would work in theory if each chain (a) is able to explore the complete target distribution, (b) is given sufficient opportunity to do so, and (c) the number of chains N is set large enough. If these conditions are satisfied, then the final chain states $\{\mathbf{\Phi}^1_{(T)}, \mathbf{\Phi}^2_{(T)}, \ldots, \mathbf{\Phi}^N_{(T)}\}$ are N independent samples of the target distribution. But Figures 3, 4 and 5 of HM23 demonstrate that conditions (a) and (b) are violated. The chains do not explore the target distribution and instead settle on a fixed point. Furthermore, a collection of only $N = 100$ samples (chains) is insufficient to robustly characterize the posterior mean and standard deviation of the BMA parameters, let alone accurately depict the marginal distribution of the weight and shape parameters. I will demonstrate this with a simple univariate bimodal target density

$$f(\phi|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, w_1) = w_1 f_{\mathcal{N}}(\phi|\mu_1, \sigma_1^2) + (1 - w_1) f_{\mathcal{N}}(\phi|\mu_2, \sigma_2^2) \tag{7}$$

of two normal distributions with means $\mu_1 = -5$ and $\mu_2 = 5$, unit variances, $\sigma_1^2 = \sigma_2^2 = 1$, and weights $w_1 = 0.2$ and $w_2 = 1 - w_1 = 0.8$. The bimodality of this target distribution resemblances local minima of the BMA likelihood function $L(\mathbf{\Phi}|\mathbf{Y}, \tilde{\mathbf{y}})$ typically found from repeated trials of the EM algorithm. The initial state $\phi_{(1)}^j$ of each Markov chain $j = 1, \ldots, N$ is drawn at random from a uniform prior distribution, $\phi_{(1)}^j \sim \mathcal{U}[-20, 20]$. Candidate points ϕ_p^j in each chain are sampled using a standard normal jump distribution $q(\phi_{(i-1)}^j \rightarrow \phi_p^j) = q(\phi_p^j \rightarrow \phi_{(i-1)}^j)$ and accepted with Metropolis probability, $\alpha_j = \min\left(1, f(\phi_p^j|\cdot)/f(\phi_{(i-1)}^j|\cdot)\right)$. Figure 2a plots a histogram of the $T = 200$th (=last) samples of the $N = 1{,}000$ Markov chains. The PDF of the target distribution $f(\phi| -5, 5, 1, 1, 0.2)$ is separately indicated with a solid black line.

The histogram of the N last samples of the Markov chains corresponds poorly to the target distribution $f(\phi| -5, 5, 1, 1, 0.2)$ (solid black line). We can repeat this experiment with a much larger number of chains N and samples T but the results remain the same. The method of HM23 is fundamentally flawed and does not lead us to the correct target distribution. The proposal distribution of HM23 is incapable of exploring the bimodal target distribution and this problem only increases with the number d of BMA parameters. Thus, the histograms of the BMA weight and standard deviation presented in Figures 4 and 5 of HM23 are NOT valid posterior distributions. Figure 2b presents a histogram of ϕ derived from the DREAM$_{(\text{BMA})}$ algorithm. The $N = 3$ Markov chains visit ϕ with frequency proportional to the target density.

To demonstrate what proper chain convergence looks like for the BMA method, I apply the `MODELAVG` toolbox of Vrugt (2018) to the 10-member ensemble of water stage predictions analyzed by HM23. The first author of HM23, Tao Huang, was kind enough to share a one hundred-day sample of their data set. Except for its length of $n = 100$ days, this data set does not match any of the water stage records shown in Figures 9 and 12 of HM23. In analogy to HM23, I execute the BMA method using a normal conditional PDF for the $K = 10$ models of the ensemble with constant forecast variances $\sigma_{tk}^2 = \sigma_k^2$. Figure 3 displays traces of the DREAM$_{(\text{BMA})}$ sampled values of the BMA weights and standard deviations $\mathbf{\Phi} = (\beta_1, \ldots, \beta_{10}, \sigma_1, \ldots, \sigma_{10})^{\top}$ using the likelihood function of Equation 6.

The chains of the DREAM$_{(\text{BMA})}$ algorithm mix well and require about 5,000 function evaluations (=samples) to converge to the limiting distribution of the $d = 20$ BMA parameters. This is far superior to any non-adaptive implementation of the Metropolis algorithm as attempted by HM23. Similar trace plots are obtained for other conditional PDFs and/or nonconstant forecast variances. Convergence is monitored using the univariate $\hat{R}$ and multivariate $\hat{R}^d$ scale reduction factors of Gelman and Rubin (1992) and Brooks and Gelman (1998), respectively. I do not display the posterior marginal distributions for the BMA parameters of HM23. This is standard screen output of the `MODELAVG` toolbox and interested readers can download this toolbox and execute the built-in examples.

3.4. BMA Quantile Calculation

The 90% BMA prediction intervals of water stage displayed in Figures 9b, 9c and 12a–12d of HM23 appear overdispersed. HM23 do not report the spread and coverage of their prediction intervals, but the gray intervals appear to be much too wide. In fact, the 90% prediction intervals contain all the training observations (which HM23 should have plotted as points and not a dashed line). This raises serious questions on how these intervals were computed. Also, unexpectedly often, the weighted average BMA forecast (solid black line) for the normal conditional PDF is found near the upper prediction limit (Figures 12a, and 12c). This behavior is expected for an asymmetric conditional distribution such as the gamma PDF in Figures 12b, and 12d of HM23, but is much more uncommon for a normal PDF with constant forecast variance, as I will demonstrate later.

HM23 hardly explain how they compute the 90% BMA prediction intervals, but I suspect that what they have instead shown are confidence intervals of the BMA weighted average forecast in Equation 2. This would explain why HM23's supposed prediction intervals do not center sufficiently on their weighted average BMA forecasts. Furthermore, this may also explain HM23's erroneous claim and/or finding that MCMC-based prediction intervals are substantially larger than their counterparts of the EM method. This is a result of their deficient

Metropolis algorithm, the result of which is a collection of N final chain states which will have substantially overestimated the posterior uncertainty of the BMA weights and variances.

HM23 should not confuse confidence intervals and prediction intervals. The predictive uncertainty of the BMA model is exactly described by the CDF in Equation 5. This only requires knowledge of the maximum likelihood values of the BMA weights $\hat{\boldsymbol{\beta}}$ and shape parameters $\hat{\boldsymbol{\psi}}$. Hence, MCMC simulation (Bayesian method) and the EM algorithm (frequentist method) are expected to yield an equal BMA forecast distribution and predictive uncertainty. Unless, of course, the EM algorithm does not return the proper maximum likelihood values of the BMA parameters. My own experience with the EM algorithm suggests that it has a very high success in locating the maximum likelihood BMA model parameters in a handful of trials when the ensemble is not too large (say $K \leq 10$). This is true for the normal, lognormal, truncated normal, and gamma conditional PDFs with a constant or non-constant forecast variance. For larger ensemble sizes, say $K > 15$, and/or conditional PDFs that have more than 1 shape parameter, such as the generalized normal, generalized Pareto and generalized extreme value distributions, I can only recommend using MCMC simulation.

The credible regions of the BMA parameters obtained from MCMC simulation will provide valuable information on the importance of the ensemble members and their interactions in the BMA model, as demonstrated in Vrugt, Diks, and Clark (2008), but this knowledge is not required for ensemble forecasting. For small data sets, the credible regions of the BMA model parameters sampled by MCMC simulation may deviate from frequentist confidence regions obtained from a first-order approximation around the maximum likelihood solution of the EM algorithm. The reason is the prior distribution. The first-order approximation assumes a multi-normal BMA parameter distribution around the maximum likelihood solution, whereas the posterior BMA parameter distribution is truncated at zero by the prior distribution. This mismatch of credible and confidence regions is expected to decrease with increasing length n of the training data record as the credible (confidence) intervals concentrate more and more on the maximum likelihood parameter values. Thus, the longer the training data record, the smaller the fraction of the prediction uncertainty of the BMA mixture distribution that is explained by parameter uncertainty.

Figure 4 displays 90% confidence intervals (dark gray) and 90% prediction intervals (light gray) of the BMA mixture model for the HM23 training data set. The prediction intervals appear noticeably tighter than their counterparts presented in Figures 9b, 9c and 12a–12d of HM23. The 90% prediction intervals are a bit too dispersed and contain 97% of the water stage measurements. This is the result of an insufficiently long training data record. Note that the BMA prediction intervals much better center on the weighted-average BMA forecast (solid black line). The midpoint of the BMA prediction interval is in close vicinity of the BMA mean forecast of Equation 2. This is certainly not imposed by the mixture CDF of Equation 5 but is a common finding for symmetric conditional PDFs with a constant variance.

The BMA prediction limits in Figure 4 are much more bumpy than their counterparts in Figures 9b–9c and 12a–12c of HM23. The overly smooth prediction limits of HM23 are suspicious in the face of the error-corrupted water stage training data of HM23 and, possibly, a BMA distribution forecast that depends only on two ensemble members. As I contemplated earlier, HM23 may have treated the prediction intervals of the BMA model as confidence intervals of the weighted average BMA forecast. As I pointed out above, evidence for this conjecture is found in the text and figures of HM23. Instead, the lower l_t and upper u_t prediction limits of the BMA forecast distribution at time t should be calculated as follows. After we have determined the maximum likelihood values $\hat{\boldsymbol{\Phi}} = \left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right)$ of the BMA parameters, a $\gamma = 100(1-\alpha)\%$ prediction interval $[l_t, u_t]_\alpha$ of the BMA distribution forecast can be calculated from its CDF $G_t\left(y|\,\mathbf{y}_t, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right)$ in Equation 5 so that $G_t\left(l_t|\,\mathbf{y}_t, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right) = \tau_1$ and $G_t\left(u_t|\,\mathbf{y}_t, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right) = \tau_2$, where $\tau_1 = \alpha/2$ and $\tau_2 = 1-\alpha/2$ are quantiles at the significance level $\alpha \in (0,1)$. As we do not have a convenient analytic expression for the quantile function (inverse CDF) $G_t^{-1}\left(\tau|\,\mathbf{y}_t, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right)$ of the BMA mixture distribution, we must use an iterative recipe to find the values of l_t and u_t. As $G_t\left(y|\,\mathbf{y}_t, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right)$ will be strictly increasing on the support $y \in [a,b]$, root finding with the secant method will find the exact prediction limits in only a handful of iterations. This methodology is implemented in the `MODELAVG` toolbox, the results of which are shown in Figure 4. A less elegant solution of l_t and u_t is linear interpolation of a large collection of (y,τ) pairs computed from the BMA CDF. This will return approximate values of l_t and u_t at which $G_t\left(y|\,\mathbf{y}_t, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right)$ is supposed to be $\alpha/2$ and $1-\alpha/2$, respectively.

3.5. Bias Correction

HM23 do not make any mention of the spread-skill relationship of the multi-model ensemble of water stages. The forecast ensemble must be properly calibrated and generally envelope the training data $\tilde{y}_1, \ldots, \tilde{y}_n$ for a successful application of the BMA method. Bias correction is critically important before applying the BMA method and can be considered a very simple form of model output statistics (Carter et al., 1989; Glahn & Lowry, 1972). Raftery et al. (2005) recommends a simple linear transformation of the raw point forecasts of each model

$$y_{tk}^{b} = a_k + b_k\, y_{tk}. \tag{8}$$

with intercept a_k and slope b_k that can be derived from ordinary least squares using the $n \times 1$ vector of training observations $\tilde{\mathbf{y}} = \left(\tilde{y}_1, \ldots, \tilde{y}_n\right)^{\top}$ to yield

$$\begin{bmatrix} \hat{a}_k \\ \hat{b}_k \end{bmatrix} = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\tilde{\mathbf{y}}. \tag{9}$$

The $n \times 2$ design matrix $\mathbf{X}$ has unit entries in the first column and the raw forecasts of the kth model $\left(y_{1k}, \ldots, y_{nk}\right)^{\top}$ in the second column. This bias correction step is rather simplistic, but deemed necessary for removing systematic bias between the models' point forecasts and the verifying observations.

For an unbiased model, the a_ks should be near zero and the b_ks should attain values close to one. If the training data set is small, the least squares estimates of the intercepts and slopes may become unstable, and bias correction may actually degrade the predictive capabilities of the ensemble (Vrugt & Robinson, 2007).

3.6. Other Methodological Limitations

HM23 assume the use of a fixed variance of the conditional PDFs of the ensemble members. This may suffice for variables such as water stage, but does not work for variables such as wind speed and river discharge that are known to have non-constant measurement error variances. Furthermore, HM23 evaluate only the normal and gamma conditional PDFs as part of their BMA model. The `MODELAVG` toolbox has a large family of built-in predictive distributions for the members of the ensemble. This includes the normal, lognormal, generalized normal, truncated normal, gamma, Weibull, generalized Pareto, and generalized extreme value distributions with a constant or non-constant forecast variance. For each choice of conditional PDF, the code returns time-averaged performance metrics, scoring rules, and prediction limits of the BMA distribution forecasts.

Lastly, the authors methodology is highly inefficient requiring between 3 and 20 min to complete its computation for sample sizes ranging between $N = 2,000$ and $N = 10,000$ chains. The DREAM$_{(\mathrm{BMA})}$ algorithm needs only a handful of seconds to derive a large sample approximation of the posterior distribution of the BMA weights and shape parameters.

4. Summary and Conclusions

The BMA methodology of HM23 amounts to a defective and inefficient implementation of the Metropolis algorithm. The proposal distribution is poorly chosen, the sampled chains do not mix properly, and the last chain sample does not yield an adequate approximation of the posterior distribution of the BMA weights and shape parameters. Unlike HM23 suggest, the choice of proposal distribution should not impact the posterior distribution of the BMA weights and parameters unless the target is multimodal with disconnected modes. HM23 also misapply the sample autocorrelation function of the BMA parameters to judge chain performance, convergence, and mixing and confuse BMA confidence and prediction intervals. In doing so, HM23 wrongfully assert that BMA parameter uncertainty has not been properly investigated in the literature. This claim not only ignores contributions on the temporal stability of BMA parameters but also overlooks software and publications on a Bayesian treatment of the BMA mixture model. But more fundamentally, HM23 suggest that the credible (confidence) regions of the BMA parameters affect the prediction uncertainty of the BMA model. This is not true. The BMA prediction limits are determined by the maximum likelihood values of the BMA weights and shape parameters.

The MODELAVG toolbox of Vrugt (2018) satisfies all requirements of HM23 and provides robust estimates of BMA model parameter and prediction uncertainty in the presence of symmetric, skewed and truncated forecast variables. This toolbox would have led HM23 to fundamentally different conclusions about their ensemble of water stage predictions. These results would have shown that (a) the posterior distribution of BMA weights and shape parameters is invariant to the choice of proposal distribution, (b) the sampled chains mix well and converge to a limiting distribution in only a handful of seconds, (c) the sampled Markov chains yield an accurate description of the marginal posterior distributions, credible intervals and sample autocorrelation functions of the BMA weights and variances of the flood models, (d) BMA prediction intervals of water stage are sharper, more erratic and with an appropriate coverage, (e) BMA prediction intervals do not depend on BMA parameter uncertainty, and (f) MCMC-determined BMA distribution forecasts match their counterparts derived from the EM algorithm. The proper implementation and use of statistical methods would have prevented the erroneous findings and results of HM23.

Data Availability Statement

The one-hundred day data set of water stage measurements and ensemble point forecasts was kindly shared by Tao Huang, the first author of HM23. The MODELAVG toolbox with its built-in examples and manual is available for download from the author's GitHub account http://github.com/jaspervrugt.

Acknowledgments
The first author (JAV) greatly appreciates the review comments of the two anonymous referees. This has led to a substantially improved manuscript.

References

Bröcker, J., & Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus*, *60*(4), 663–678. https://doi.org/10.1111/j.1600-0870.2008.00333.x

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational & Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.2307/1390675

Carter, G. M., Dallavalle, J. P., & Glahn, H. R. (1989). Statistical forecasts based on the national meteorological centers numerical weather prediction system. *Weather and Forecasting*, *4*(3), 401–412. https://doi.org/10.1175/1520-0434(1989)004⟨0401:SFBOTN⟩2.0.CO;2

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, *39*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

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.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511. https://doi.org/10.1214/ss/1177011136

Glahn, H. R., & Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology and Climatology*, *11*(8), 1203–1211. https://doi.org/10.1175/1520-0450(1972)011⟨1203:TUOMOS⟩2.0.CO;2

Haario, H., Saksman, E., & Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC. *Computational Statistics*, *20*(2), 265–273. https://doi.org/10.1007/BF02789703

Hansen, B. E. (2007). Leats squares model averaging. *Econometrica*, *75*(4), 1175–1189. https://doi.org/10.1111/j.1468-0262.2007.00785.x

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Statistical Science*, *14*(4), 382–417. https://doi.org/10.1214/ss/1009212519

Höge, M., Guthke, A., & Nowak, W. (2019). The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, *572*, 96–107. https://doi.org/10.1016/j.jhydrol.2019.01.072

Huang, T., & Merwade, V. (2023). Improving Bayesian model averaging for ensemble flood modeling using multiple Markov chains Monte Carlo sampling. *Water Resources Research*, *59*(10), e2023WR034947. https://doi.org/10.1029/2023WR034947

Laloy, E., & Vrugt, J. A. (2012). High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(ZS)}$ and high-performance computing. *Water Resources Research*, *48*(1). https://doi.org/10.1029/2011WR010608

Lochbühler, T., Vrugt, J. A., Sadegh, M., & Linde, N. (2015). Summary statistics from training images as prior information in probabilistic inversion. *Geophysical Journal International*, *201*(1), 157–171. https://doi.org/10.1093/gji/ggv008

McLachlan, G. J., & Krishnan, T. (2008). *The em algorithm and extensions* (2nd ed.). Wiley. https://doi.org/10.1002/9780470191613

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., & Zheng, Y. (2003). *Long-run performance of Bayesian model averaging (Tech. Rep. No. 433)*. University of Washington.

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, *16*(4), 351–367. https://doi.org/10.1214/ss/1015346320

Sloughter, J. M., Gneiting, T., & Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and bayesian model averaging. *Journal of the American Statistical Association*, *105*(489), 25–35. https://doi.org/10.1198/jasa.2009.ap08615

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. https://doi.org/10.1175/MWR3441.1

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and Matlab implementation. *Environmental Modelling and Software*, *75*, 273–316. https://doi.org/10.1016/j.envsoft.2015.08.013

Vrugt, J. A. (2018). *MODELAVG: A Matlab toolbox for postprocessing of model ensembles (tech. Rep.)*. University of California. https://www.researchgate.net/publication/299458373_MODELAVG_A_MATLAB_Toolbox_for_Postprocessing_of_Model_Ensembles

Vrugt, J. A. (2024). Distribution-based model evaluation and diagnostics: Elicitability, propriety and scoring rules for hydrograph functionals. *Water Resources Research*, *60*, e2023WR036710. https://doi.org/10.1029/2023WR036710

Vrugt, J. A., Clark, M. P., Diks, C. G. H., Duan, Q., & Robinson, B. A. (2006). Multi-objective calibration of forecast ensembles using Bayesian model averaging. *Geophysical Research Letters*, *33*(19). https://doi.org/10.1029/2006GL027126

Vrugt, J. A., Diks, C. G. H., & Clark, M. P. (2008). Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. *Environmental Fluid Mechanics*, *8*(5), 579–595. https://doi.org/10.1007/s10652-008-9106-3

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). https://doi.org/10.1029/2005WR004838

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2007WR006720