Reply to Reviewers' comments (Reviewer#1)

Legend

Reviewers' comments

Authors' responses

Direct quotes from the revised manuscript

Reviewer #1:

The paper proposes a new CHUP-BMA ensemble forecasting method by incorporating the CHUP-derived posterior distribution of the observed flow into the BMA framework. It has the advantage that the initial state constraints can be considered in the BMA while avoiding the normal quantile transformation of the HUP-BMA method. Based on deep learning, an ensemble forecasting scheme considering input, model structure, and parameter uncertainty is constructed in Three Gorges Reservoir, China, and the effectiveness of the CHUP-BMA method in reducing forecast uncertainty is verified. The study is innovative and theoretically rigorous and has promising results with solid application potential. Some questions need further discussion.

Response: We deeply appreciate your constructive comments and the time you spent on reviewing the paper. We have accepted all the revision comments. Point-by-point replies to the comments or suggestions made can be found below.

1. The sources of Figure 3 and Table 2 need to be explained to improve the reasonableness of the paper.

Response: The values of Figure 3 and Table 2 are obtained from the Hydrological Bureau of the Changjiang Water Resources Commission (HBCWRC). In addition, Figure 3 can be referred to the paper by Zhong et al. (2018b). We make the following changes in subsection 3.2.1:

The rainfall-runoff relationship graph method (Fedora and Beschta, 1989) commonly used in the Yangtze River basin can calculate the effective precipitation. The antecedent precipitation index, which is the key variable of the method, can be calculated by the following equation to represent the soil moisture content (Zhong et al., 2018b).

$$P_{a,t+1} = k(P_{a,t} + P_t) \tag{16}$$

$$P_{a,t+1} \leq I_m \tag{17}$$

where P_a denotes the antecedent precipitation index, P_t is the daily precipitation, I_m is the water storage capacity of the basin, k denotes evaporation reduction index.

The values of k and I_m for these three sub-basins are listed in Table 2, which are obtained from the Hydrological Bureau of the Changjiang Water Resources Commission (HBCWRC). Since the rainfall-runoff relationship graph method have been widely used for runoff generation calculation in the Yangtze River basin, the rainfall-runoff relationship between Xiangjiaba and Three Gorges Dam-site uncontrolled interval basin are established and plotted in Fig. 3, which is used to calculate the effective precipitation based on the antecedent precipitation index (P_a) and observed (or forecasted) precipitation for these three sub-basins.

Reference

Zhong, Y., Guo, S., Liu, Z., Wang, Y., and Yin, J. Quantifying differences between reservoir inflows and dam site floods using frequency and risk analysis methods. Stoch. Environ. Res. Risk Assess., 32, 419-433. https://doi.org/10.1007/s00477-017-1401-4, 2018b.

2. Various model inputs (e.g., rainfall, tributary flows, etc.) exist in the interval basins. The article only considers the input uncertainty of rainfall, and it is suggested to add a reason for this in subsection 3.2.1.

Response: Thanks to the reviewers for the constructive comments. There are five flow discharge inputs from five large tributaries (Jinsha, Min, Jialing, Tuo, and Wu Rivers) in our case study. The flow discharges are observed at the Pingshan, Gaochang, Fushun, Beibei, and Wulong hydrological controlled stations, respectively. Since these observed (or forecasted) flows are respectively regulated by their upstream cascade reservoirs, these flow data inputs are more accurate than the rainfall inputs.

We collected the forecasted precipitation data from the European Centre for

Medium-Range Weather Forecasts (ECMWF) and HBCWRC during the 2017-2021 flood season in the three sub-basins. Since the rainfall data is more diverse and has relatively large uncertainty, so the forecast rainfall input variable is used to explore the impact of forecast rainfall uncertainty on the Three Gorges reservoir inflow forecasts. We also make the following changes in subsection 3.2.1:

There are five flow discharge inputs from five large tributaries (Jinsha, Min, Jialing, Tuo, and Wu Rivers) in this case study. The flow discharges are observed at the Pingshan, Gaochang, Fushun, Beibei, and Wulong hydrological controlled stations, respectively. Since these observed (or forecasted) flows are respectively regulated by their upstream cascade reservoirs, these flow data inputs are more accurate than the rainfall inputs. This study collected the forecasted precipitation data from the European Centre for Medium-Range Weather Forecasts (ECMWF) and HBCWRC during the 2017-2021 flood season in these three sub-basins. Since the rainfall data is more diverse and has relatively large uncertainty, only the forecast rainfall input variable is used to explore the impact of forecast rainfall uncertainty on the Three Gorges reservoir inflow forecasts.

3. Line 255. You should briefly introduce the LSTM in subsection 3.2.2 to improve the paper's readability. In addition, it is recommended to cite references more relevant to the LSTM.

Response: Thanks for your valuable comments. We will add a brief introduction to LSTM neural networks and cite more relevant literatures as follows:

The structure of LSTM neural network includes forgetting gate, input gate, updating the state of the memory unit, and output gate (Hochreiter and Schmidhuber, 1997). The forgetting gate can select the relatively important information in the previous memory unit. The input gate can select useful information from the input variables at the current moment. The memory unit state can store relatively important information extracted from historical moments, which is updated under the control of the forgetting gate and the input gate. The output gate selects and outputs useful information from the memory cell state. More detailed procedures of the LSTM neural network formulation have been described by Kratzert et al. (2018).

Additional references:

Hochreiter, S., Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M. Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005-6022. https://doi.org/10.5194/hess-22-6005-2018, 2018.

4. Deep learning parameters significantly impact forecast accuracy, so it is recommended to show the values of deep learning parameters. The study should concentrate on ensemble forecasting methods rather than deep learning models. Therefore, the model parameter values can be shown in the appendix.

Response: Thanks for your valuable comments. We supplement the appendix with model parameter values for the ensemble members. An appendix will be added in the revised manuscript as follow:

Appendix:

We set the number of neural network layers and neurons to be the same for the encoding and decoding processes, with trial-and-error preferences for the number of hidden layers, neurons, and dropout. Meanwhile, the batch size, epoch, and learning rate are set to 100, 500, and 0.001, respectively. The different model parameters are shown in Table A.

Table A The different model parameters for ensemble membership

Ensemble member type	Neuron	Hidden layers	Dropout
ECMWF&DA-LSTM-RED&MSE	64	1	0.001
ECMWF&LSTM-RED&MSE	64	1	0.001

ECMWF&DA-LSTM-RED&MAE	32	1	0.01
ECMWF&LSTM-RED&MAE	64	1	0.1
HBYRWRC &DA-LSTM-RED&MSE	32	1	0.1
HBYRWRC &LSTM-RED&MSE	32	1	0.001
HBYRWRC &DA-LSTM-RED&MAE	64	1	0.001
HBYRWRC &LSTM-RED&MAE	48	1	0.01

5. Line 369, add a description of the member type with better forecast accuracy, i.e., the input composition, the model structure, and the objective function of the selected parameters.

Response: Thanks for your insightful comments. We have added relevant content to the article:

The member with relatively optimal forecast accuracy is composed of the forecast rainfall from ECWMF, the DA-LSTM-RED model, and the objective function with mean square error to optimize the parameters.

6. There are numerous evaluation metrics in deterministic and probabilistic forecasting. Briefly explain the reasons for the metrics chosen in the paper.

Response: Thanks for your suggestion. The article is added below:

The Nash-Sutcliffe efficiency (NSE) is one of the most important metrics in flood forecasting, reflecting the degree of fit between forecasted and observed flows (Nash & Sutcliffe, 1970). Since the accurate runoff volume predictions is more important than peak discharge for the operation of a large reservoir (Cui et al., 2023), the relative error for total runoff volume (RE) is also chosen. The mean absolute error (MAE) can reflect the forecast error for each moment, and compared with the continuous ranked probability score (CRPS) of the ensemble forecast (Raftery et al., 2005), which can reflect the effectiveness of the ensemble forecast correction. Therefore, three metrics, NSE, RE, and MAE, are selected to evaluate the deterministic forecast results.

The average coverage rate (CR) is one of the most necessary metrics for evaluating the reliability of forecast intervals (Li et al., 2021). The average interval width (IW) is the metric that directly reflects the level of forecast uncertainty, which is an important metric for evaluating the effectiveness of the proposed methods. The percentage of observations bracketed by the unit confidence Interval (PUCI) is a comprehensive metric for evaluating the performance of forecast intervals in quantifying uncertainty (Xiong et al., 2009). Therefore, the CR, RB, and PUCI metrics are selected to evaluate the forecast intervals performance.

The α_index metric can quantitatively assess the reliability of ensemble probabilistic forecasts from the perspective of distribution function values (Renard et al., 2010). The ignorance score (IGS) metric can quantitatively assess the sharpness of the posterior density function and quantify forecast uncertainty from the perspective of the probability density (Gneiting et al., 2005). The continuous ranked probability score (CRPS) is one of the most important composite metrices for assessing the overall performance of probabilistic forecasts (Raftery et al., 2005) and can represent both the reliability and sharpness of forecasted posterior distribution function. Therefore, the α_index, IGS, and CRPS metrics are selected to evaluate the probabilistic forecast performance.

Additional references:

Li, D., Marshall, L., Liang, Z., Sharma, A., Zhou, Y. Bayesian LSTM with stochastic variational inference for estimating model uncertainty in process-based hydrological models. *Water Resources Research*, 57(9), e2021WR029772. https://doi.org/10.1029/2021WR029772, 2021.

Xiong, L., Wan, M. I. N., Wei, X., O'connor, K. M. Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation. *Hydrological sciences journal*, 54(5), 852-871. https://doi.org/10.1623/hysj.54.5.852, 2009.

7. Line 465, replacing 'concentration' with 'sharpness' as 'reliability (α_index), concentration (IGS),' should correspond to the name of Figure 13.

Response: Thank you for reminder. We have changed "concentration" to "sharpness" in the revised manuscript.

8. To improve modeling rationality, explain why observations are used as model tributary inputs in training and validation periods.

Response: Thanks for your well-considered suggestions. Three Gorges Reservoir (TGR) is the largest hydraulic project in the world and controls a watershed area of 1 million km^2. There are more than 300 larger-scale reservoirs have been built in the upstream Yangtze River basin with a total storage of 163.3 billion m^3. The operational flow forecasting procedure is from the sources of tributaries, each larger-scale reservoir inflows, interval basin flow forecasting and river flow routing to downstream sections, and so on. Since the forecast data series at the outlets of five large tributaries (Jinsha, Min, Jialing, Tuo, and Wu Rivers) are inconsistent, we used the observed flows to train and validate the proposed models or forecasting schemes. We have added the following in subsection 3.2.2:

In the actual TGR inflow forecasting model, the observed flow discharge data at the outlets of five large tributaries (Jinsha, Min, Jialing, Tuo, and Wu Rivers) in the interval-basin between Xiangjiaba and TGR dam-site are used to train and validate the proposed models or forecasting schemes since the forecast data series at the outlets of tributaries are inconsistent.

9. In the outlook, adding the construction of the CHUP-BMA method using a more flexible vine copula will make the CHUP-BMA method more competitive.

Response: Thank you for your foresighted suggestions. Our additions to the article are as follows:

In the future, the vine copula, which facilitates multivariate joint distribution modelling, can be considered for constructing the CHUP-BMA method and exploring its advantages and effectiveness in ensemble flood forecasting.