Reply to Reviewers' comments (Reviewer#2)

Legend

Reviewers' comments

Authors' responses

Direct quotes from the revised manuscript

Reviewer #2:

The authors combine the copula-based hydrological uncertainty processor (CHUP) and Bayesian model averaging (BMA) to obtain a novel approach to statistical post-processing of hydrological ensemble forecasts. The proposed approach is promising and the presented results are fair, but the paper needs some improvement and it also raises some questions.

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.

Major comments:

1. L28: The cited paper Sloughter et al. (2010) deals with post-processing wind speed forecasts. The BMA model for precipitation is introduced in Sloughter et al. (2007).

Response: Firstly, thank you very much for your careful and detailed suggestions. We have found the equivalent in line 48 and have made the following supplementary revisions:

The BMA method is initially successfully applied to the ensemble forecast of meteorological elements such as temperature, precipitation, and wind speed (Raftery et al 2005; Sloughter et al, 2007; Sloughter et al, 2010).

Additional references:

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon. Weather Rev. 135, 3209–3220.

2. I am also missing references to BMA models for hydrological forecasts, e.g. Hemri et al. (2013) or Baran et al. (2019).

Response: Thanks for your constructive comments. References have been added to the paper.

Hemri et al. (2013) introduced the principle of Geostatistical output perturbation (GOP) into the BMA method, and extended the membership probability distribution into a multivariate normal distribution function, proposing a multivariate BMA. Relative to the univariate BMA method, this method can not only consider the temporal correlation between forecast flows, but also improve the forecast reliability when the forecast system was changing, i.e., fewer models were available due to dropping out at particular lead times. In order to ensure that the quantiles of forecast distributions after Box-Cox transformation are within the actual physical range, Baran et al. (2013) introduced upper and lower truncated normal distributions into the BMA, they found that the double truncated BMA had reliable forecasting ability compared to ensemble model output statistics, and the advantage was more obvious when rolling window training periods are used.

Additional references:

Baran, S., Hemri, S. and El Ayari, M. (2019) Statistical post-processing of water level forecasts using Bayesian model averaging with doubly-truncated normal components. Water Resour. Res. 55, 3997–4013.

Hemri, S., Fundel, M. and Zappa, M. (2013) Simultaneous calibration of ensemble river flow predictions over an entire range of lead times. Water Resour. Res. 49, 6744–6755.

3. Eq.4: In the original description of the HUP, different CDFs are considered for the forecasts and the observations, moreover, in the former case it is considered as an initial estimate. Does such a relaxation make sense here as well?

Response: Thanks for your insightful suggestions. We will add the following changes to the paper.

The HUP method is a meta-Gaussian model assuming that flow series transformed to

normal space obey the Gaussian distribution. The cumulative distribution function is different for forecasted and observed flows. The common normal quantile transformation is key to the application of the HUP method, and its significance is to make the HUP method applicable to variables with any marginal distributions, heteroskedasticity, and nonlinear dependence structures (Krzysztofowicz and Kelly, 2000; Darbandsari and Coulibaly, 2021).

4. Section 3.1.2. introducing the HUP follows the structure of Sections 2.1.2 – 2.1.4 of Darbandsari and Coulibaly (2021); however, one should mention that the Markov process of Eq.5 is stationary and define exactly how θt in L169 is related to Eq.7 (see Darbandsari and Coulibaly, 2021, Eq.10).

Response: Thanks for your perceptive suggestions. We will add the following changes:

The HUP method assumes that the observed flow obeys the strictly stationary first-order Markov process (Krzysztofowicz and Kelly, 2000)

$\hat{Q}_b$, $\hat{Q}_o$, and $\hat{Q}_{f,i}$ are assumed to obey a linear relationship. The expression of the likelihood function in normal space is as follows.

$$\hat{Q}_{f,i,t} = a_t \times \hat{Q}_{o,t} + d_t \times \hat{Q}_b + b_t + \theta_t$$

$$p\left(\hat{Q}_{f,i,t} \middle| \hat{Q}_{o,t}, \hat{Q}_b\right) = \frac{1}{\sigma_t} n\left\{\frac{\hat{Q}_{f,i,t} - \left(a_t \times \hat{Q}_{o,t} + d_t \times \hat{Q}_b + b_t\right)}{\sigma_t}\right\} \quad (7)$$

where, θ_t is an independent variable obeying N(0,σ_t^2). a_t, d_t, and b_t are regression coefficients.

5. L310: “The IGS metric indicates the sharpness of the probabilistic forecast”. The IGS, similar to the CRPS addresses simultaneously both calibration and sharpness, as indicated in the cited work of Gneiting et al. (2005). Hence, I think referring to IGS as a measure of concentration is slightly misleading.

Response: Thanks for your valuable comments. We will correct the misleading content

and make a corresponding change in line 310:

The IGS and CRPS metrics can reflect the reliability and sharpness of the probabilistic forecast. The former can quantify the forecast probability density at the observation, while the latter can indicate the fit performance between the posterior probabilistic distribution and the actual probabilistic distribution of Qo (Raftery et al., 2005). Both CRPS and IGS are negative scores, i.e., the smaller the value, the better. The IGS imposes severe penalties for particularly poor probabilistic predictions and may be extremely sensitive to outliers and extreme events, yet also lacks robustness (Raftery et al., 2005).

A corresponding change in line 465:

Meanwhile, Fig. 13 (a), (b), and (c) show the evaluation metrics of the ensemble probabilistic forecast.

A corresponding change in line 480:

It can be seen from Fig. 13(b) that the IGS values of the two methods gradually increase with the increase of the forecast horizon, indicating that the forecast uncertainty gradually increases. The maximum, minimum, and mean of the IGS metric for the CHUP-BMA method are 9.10, 8.33, and 8.87, respectively, and 9.16, 8.59, and 8.98 for the HUP-BMA method, respectively. It can be seen that the IGS metrics of the CHUP-BMA method are consistently lower than those of the HUP-BMA method, which indicates that the CHUP-BMA method has better ensemble forecast performance relative to the HUP-BMA method by assigning a higher probability density around the actual values.

6. In Section 4, I would definitely consider the corresponding scores (or at least some of them) for the ensemble forecasts as well.

Response: Thanks very much for your insightful suggestions. Some of the evaluation metrics with a high degree of acceptance corresponding to ensemble forecasts, such as the IGS and CRPS metrics, have been used in the paper, supplemented by the

probability integral transform (PIT) histogram, which is more intuitive relative to the Q-Q diagram.

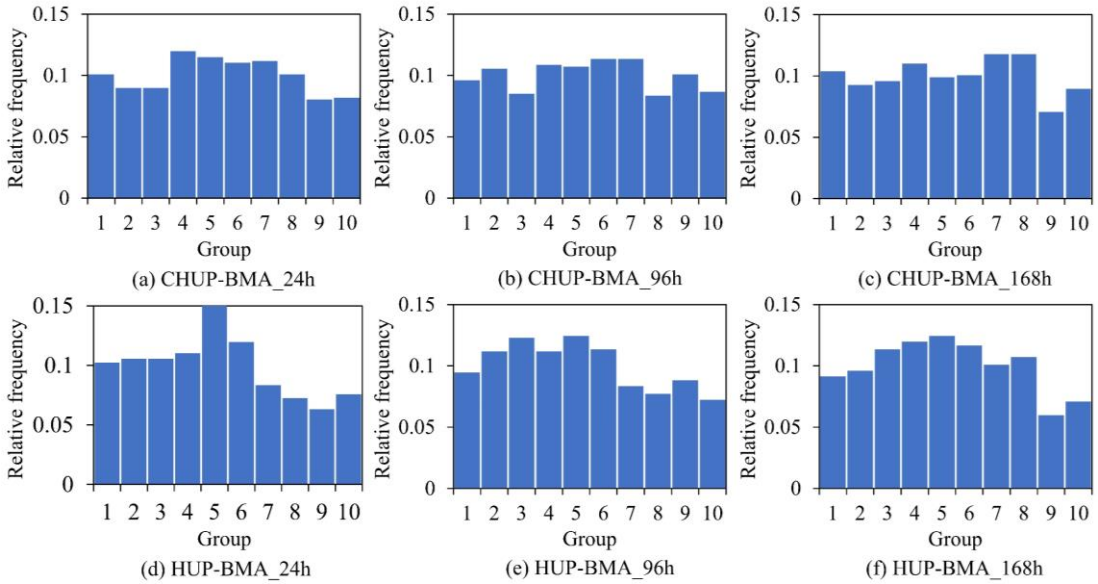

Fig. 12 The probability integral transform (PIT) histograms of the HUP-BMA and CHUP-BMA methods for the ensemble forecasts of the 24, 96, and 168h forecast horizons.

Fig. 12 shows the PIT histograms of the HUP-BMA and CHUP-BMA methods for 24, 96, and 168h forecast horizons. It can be significantly observed that the PIT plots of the HUP-BMA method show a ∩-shaped distribution, which indicates that the forecast distribution is over-dispersed and overestimates the forecast uncertainty, explaining the phenomenon of wide intervals. Meanwhile, the PIT plot of CHUP-BMA is more uniformly distributed than that of the HUP-BMA method, which can obtain a better calibration performance.

7. What can be said about the statistical significance of the score differences between HUP-BMA and CHUP-BMA?

Response: Thanks for your valuable suggestions. We supplemented the statistical significance of the score differences between HUP-BMA and CHUP-BMA.

Table 5 T-test results of ensemble forecast metrics at 0.05 significance level

Metric	α_index		IGS		CRPS	
	HUP-BMA	CHUP-BMA	HUP-BMA	CHUP-BMA	HUP-BMA	CHUP-BMA
Mean	0.93	0.97	8.98	8.87	1188	1074
Variance	0.0003	0.0001	0.02	0.03	32247	33716
Degree of freedom	46.00		52.00		54.00	
T-statistic	-10.76		2.36		2.34	
T-threshold	1.68		1.67		1.67	
Difference significance analysis	Significant		Significant		Significant	

From the Table 5, it can be seen that the T-statistics at the 0.05 significance level for all three metrics are higher than the threshold value, indicating that there is a significant difference between the scores of the CHUP-BMA and HUP-BMA methods, i.e., the CHUP-BMA method is significantly better than the HUP-BMA method for ensemble forecasting metrics and performance.

Minor remarks, typos:

1. L205-206: “It has been studied that the BMA method with sliding windows can obtain better probabilistic forecast performance”. Better compared to what?

Response: Thanks for your thoughtful suggestions for changes. The following changes have been made:

Parrish et al. (2012) and Darbandsari and Coulibaly (2019) have shown that the BMA method with the sliding window can obtain better probabilistic forecast performance compared to the method without the sliding window.

2. L307: “indicative function” → “indicator function”

Response: Thanks for your detailed suggestions for changes. The following changes have been made:

$I(\cdot)$ denotes the indicator function.