Quantify and reduce flood forecast uncertainty by the CHUP-BMA method

Zhen Cui[1], Shenglian Guo[1*], Hua Chen[1], Dedi Liu[1], Yanlai Zhou[1], Chong-Yu Xu[2]

[1] State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan, China

[2] Department of Geoscience, University of Oslo, Oslo, Norway

Correspondence to: Shenglian Guo (slguo@whu.edu.cn)

Abstract. The Bayesian model averaging (BMA), hydrological uncertainty processor (HUP), and HUP-BMA methods have been widely used to quantify flood forecast uncertainty. This study proposed the CHUP-BMA method by introducing a copula-based HUP in the framework of BMA to bypass the need for normal quantile transformation of the HUP-BMA method. The proposed ensemble forecast scheme consists of 8 members (two forecast precipitation inputs, two advanced long short-term memory (LSTM) models, and two objective functions used to calibrate parameters) and is applied to the interval basin between Xiangjiaba and Three Gorges Reservoir (TGR) dam-site. The ensemble forecast performance of the HUP-BMA and CHUP-BMA methods is explored in the 6-168h forecast horizons. The TGR inflow forecasting results show that the two methods can improve the forecast accuracy over the selected member with the best forecast accuracy, and the CHUP-BMA performs much better than the HUP-BMA. Compared with the HUP-BMA method, the forecast interval width with the 90% confidence level and continuous ranked probability score metrics of the CHUP-BMA method are highest reduced by 28.42% and 17.86%, respectively. The probability forecast of the CHUP-BMA method has better reliability and sharpness and is more suitable for flood ensemble forecasts, providing reliable risk information for flood control decision-making.

1 Introduction

Accurate and reliable flood forecasting is one of the necessary measures to reduce flood disasters and improve water resource utilization (Zhou et al., 2019; Vegad and Mishra, 2022). With the development of hydrological theory and flood forecasting techniques, the flood forecasting accuracy and lead time have been significantly improved in recent years (Xu et al., 2022; Cui et al., 2023). However, neither physically-based and conceptual hydrological models nor data-driven models can guarantee to obtain perfect forecasting in real conditions. Because of the influence of the changing environment and the limitations of human perception of complex hydrological processes, the meteorological forcing and other inputs, hydrological model structure, and parameters, etc., contain significant uncertainties (Cloke et al., 2009), which leads to the simulation and forecast results of the model inevitably containing integrated uncertainties from multiple sources (Liu et al., 2022). Traditional flood forecasting schemes are mostly deterministic forecast results without considering forecast

30 uncertainty (Zhong et al., 2018a; Gelfan et al., 2018), which makes decision-makers unable to grasp useful risk information

31 beyond the forecast value. Excessive superstition on a single forecast value will likely lead to poor decision-making

32 (Krzysztofowicz et al., 1999). Therefore, it is essential to quantify and reduce flood forecast uncertainty in practical

33 applications.

34   Probabilistic flood forecasting is one of the effective methods to quantify integrated forecast uncertainty (Matthews et

35 al., 2022). It not only provides a deterministic forecast value, but also provides forecast uncertainty (or risk) information by

36 means of quantile, confidence interval, or density function (Biondi and Todini, 2018; Ferretti et al., 2020; Zhou et al., 2022),

37 which is more scientifically reasonable and practically useful compared with deterministic forecasts and helps decision-

38 makers consider forecast risk quantitatively (Todini, 2008). Various probabilistic forecasting methods based on statistical

39 post-processing of numerical forecast data have been developed in recent years. Among these methods, probabilistic

40 ensemble forecasting is considered to overcome the limitations of a single model or a simple average with fixed model

41 weights (Han and Coulibaly, 2017) and contains richer forecast information because it can consider the ensemble forecast

42 results of multiple models to quantify and reduce integrated uncertainty that contains uncertainties in the inputs, model

43 structure, and parameters (Li et al., 2017; Saleh et al., 2016). Bayesian model averaging (BMA), proposed by Raftery et al.

44 (2005), uses the Bayesian theory and a total probability formulation to transform ensemble forecasts into probabilistic

45 forecasts and is one of the most representative and reliable methods that has been widely used to supplement uncertainty

46 information beyond point estimates (Shu et al., 2022).

47   The BMA method is initially successfully applied to the ensemble forecast of meteorological elements such as

48 temperature, precipitation, and wind speed (Raftery et al 2005; Sloughter et al, 2007; Sloughter et al, 2010). After

49 confirming that the BMA method can effectively quantify forecast uncertainty and obtain highly accurate deterministic

50 forecasts, it is widely used in hydrological forecasting to quantify forecast uncertainty from different sources, such as model

51 inputs, structure, and parameters. The standard BMA method assumes that each member's posterior probability distribution

52 approximately obeys a normal distribution (Huang et al., 2019; Guo et al., 2021). However, some variables, such as wind

53 speed, rainfall, runoff, etc., usually obey skewed distributions and require methods such as Box-Cox to convert non-

54 Gaussian variables to standard normal variables that affect the accuracy of probability distribution estimation (Duan et al.,

55 2007; Liu et al., 2018). Many authors have investigated the applicability of BMA in flood ensemble forecasting and tried to

56 overcome its limitations (Madadgar and Moradkhani, 2014; Darbandsari and Coulibaly,2020). Sloughter et al. (2010)

57 proposed an improved BMA method by assuming that the posterior probability distribution of each member could obey a

58 specific non-normal distribution (e.g., Gamma distribution) and using the member forecast values to estimate the mean and

59 variance of the distribution. Madadgar and Moradkhani (2014) introduced the Copula function to solve the posterior

60 probability distribution of members in the BMA method and proposed the Copula-based BMA method, which avoids the

61 assumption of the posterior probability distribution and further reduces the application limitation of the BMA method. In

62 order to ensure that the quantiles of forecast distributions after Box-Cox transformation are within the actual physical range,

63 Baran et al. (2019) introduced upper and lower truncated normal distributions into the BMA, and found that the double

truncated BMA had reliable forecasting ability compared to ensemble model output statistics. The advantage was more obvious when rolling window training periods are used. Hemri et al. (2013) introduced the principle of geostatistical output perturbation into the BMA method and proposed a multivariate BMA, which extended the membership probability distribution into a multivariate normal distribution function. Relative to the univariate BMA method, the multivariate BMA 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. Meanwhile, the BMA method usually ensembles the forecast results of multiple models to be as close to the actual values as possible. However, too many ensemble members may generate redundant information. Darbandsari and Coulibaly (2020) introduced the Shannon entropy theory to select the forecast members that satisfy the above conditions before applying BMA. Their results showed that the BMA method incorporating entropy could improve the probabilistic forecasting performance for high flows over the standard BMA method. In addition, some studies have developed various methods based on the BMA principle, such as the multi-model ensemble forecasting method based on Vine Copula (Zhang et al., 2022) and the combination of BMA and data assimilation techniques (Parrish et al., 2012).

However, most studies ignore an essential issue: the BMA does not consider the constraint of initial conditions (i.e., observed flow at the start of the forecast). It can be shown from Raftery et al. (2005) that the conditional distribution of the member ($Q_{f,i}$) in the BMA is assumed to follow the normal distribution with expectation $\mu_i = a_i + b_i \cdot Q_{f,i}$ (a_i and b_i are the bias correction coefficients) and variance σ_i, which implies that the conditional distribution is only related to the member's forecasted flow and not affected by the observed flow at the forecast start time. It is unreasonable to produce the same posterior distribution when the forecast results are the same at different moments.

The hydrological uncertainty processor (HUP) can obtain the posterior distribution function of the actual value under the condition of the forecast value and the observed flow at the start time based on Bayesian principles and the assumption of perfect rainfall forecasting (Krzysztofowicz and Kelly, 2000). Darbandsari and Coulibaly (2021) firstly utilized the HUP method to derive the posterior distribution of each member considering the initial constraints, and then used the BMA method to weight the conditional distribution of all members to obtain the final posterior distribution, which is called the HUP-BMA method. Their results showed that the HUP-BMA method outperforms the HUP method and improves the BMA method in short-term probabilistic forecasting. In addition, the derivability of the posterior distribution for the ensemble members is theoretically enhanced, the heteroskedasticity of the ensemble members is considered, and the interpretability and logical rationality of the BMA method are improved.

Although it has been demonstrated that considering initial conditions in the BMA method can improve ensemble forecast performance, there are still issues to be explored. The HUP-BMA method requires a normal quantile conversion method to convert the flow data series to Gaussian space to solve the posterior distribution. The process is not only tedious and complicated, but also prone to bias in the inverse conversion. To this end, Liu et al. (2018) adopted the copula to derive the conditional distribution of the observed flow under the conditions of the forecasted flow, which avoids the assumption that the flow series obeys a normal distribution in the HUP and relaxes the application limitation. The study shows that the

CHUP can improve the probabilistic forecasting performance of the HUP method. It is anticipated that coupling CHUP to the BMA may improve the HUP-BMA accuracy and applicability, which motivates the current study.

The main innovations and research steps are shown as follows: (1) A novel CHUP-BMA method is proposed for the first time by coupling CHUP into BMA, which not only avoids the normal distribution assumption in HUP-BMA, but also considers the constraints of the initial condition of the forecast. (2) An ensemble forecast containing eight members is constructed by combining two types of forecast precipitation, two long short-term memory (LSTM) models, i.e., the recursive encoder-decoder structure-based LSTM-RED model and the feature-temporal dual attention-based DA-LSTM-RED model, and two objective functions of model calibration. (3) The ensemble forecast performance of the proposed method is analysed and discussed in comparison to the HUP-BMA benchmark method in terms of the deterministic and probabilistic forecasts. The interval basin between Xiangjiaba Dam and the Three Gorges Dam in the Yangtze River, China, is selected as case study.

The rest of the paper is organized as follows. Section 2 introduces the case study and materials. The methods are presented in Section 3. Section 4 evaluates the deterministic and ensemble forecast results. Conclusions and prospects are given in Section 5.

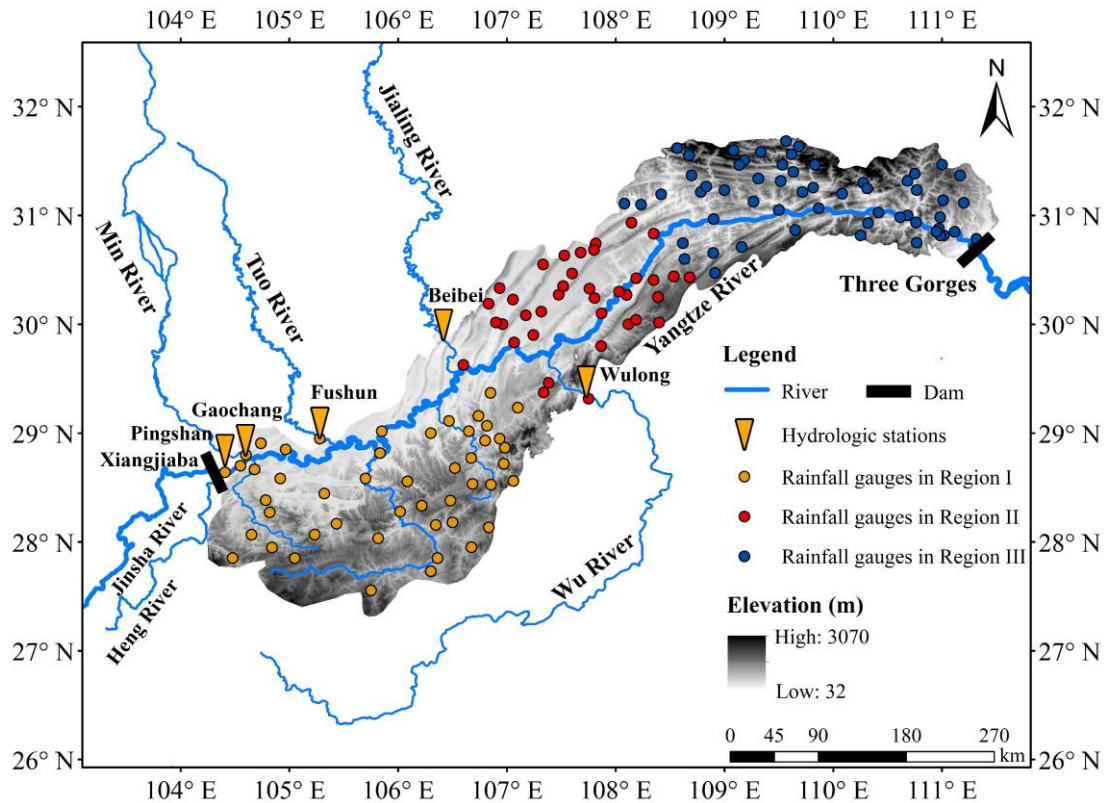
2 Case study and materials

2.1 Study basin

Three Gorges Reservoir (TGR) is the largest hydraulic project in the world and plays a vital role in flood control, power generation, and other water resource management issues (Zhong et al., 2020). The TGR controls a watershed area of about 1 million km^2. The total reservoir capacity is about 39.3 billion m^3, with a flood control capacity of about 22.15 billion m^3.

The TGR inflow is directly influenced by the runoff yield of the cascade reservoir interval-basin between Xiangjiaba and TGR (Fig.1), with a basin area of about 127,400 km^2 (Zhou et al., 2019). The inflow of the TGR consists of the outflow discharge from the Xiangjiaba Reservoir, the inflow of several tributaries such as Min, Tuo, Jialing, and Wu Rivers, and the rainfall of the interval-basin. The flow sources are complex and have different effects on the TGR inflow. Moreover, TGR is a river-type reservoir with a length of about 600 km at the normal storage level (175m) and an average width of only 1.1 km, resulting in uncertainty in rainfall intensity and storm-centre positioning (Zhong et al., 2020). Therefore, there is significant uncertainty in the flood forecast of TGR. It has been a major challenge to quantify and reduce forecast uncertainty.

Table 1 shows the flow propagation time from the hydrological control stations of the mainstream and tributaries to the TGR dam. The outflow discharge of Xiangjiaba Reservoir, located on the Jinsha River, is observed at the Pingshan hydrological station and represents the mainstream flow. The discharge values from large tributaries (Min, Jialing, Tuo, and Wu Rivers) are observed at the Gaochang, Fushun, Beibei, and Wulong hydrological stations, respectively.

129

130 **Fig. 1** Schematic diagram of the interval-basin between Xiangjiaba and TGR dam-site which is divided into three sub-regions.

131

132   Considering the uneven distribution of rainfall intensity because of the narrow and long basin, the interval basin

133 between Xiangjiaba and TGR dam-site is divided into three sub-basins: Pingshan-Cuntan, Cuntan-Wanxian, and Wanxian-

134 TGR dam-site. Their watershed areas are 76,900, 22,900, and 27,600 km^2 respectively. Meanwhile, there are 45, 38, and 60

135 gauged rainfall stations in these three sub-regions, respectively.

136

137   **Table 1** List of flow propagation time for hydrological control stations to TGR dam-site

Rivers	Hydrological control stations	To TGR dam flow propagation time (h)
Jinsha	Pingshan	48-66
Min	Gaochang	48-66
Jialing	Beibei	24-42
Tuo	Fushun	42-60
Wu	Wulong	15-30

138

139 **2.2 Study materials**

140   This study collects 6h observed flow discharges at TGR dam-site and five hydrological stations (Table 1), and 6h
141 observed rainfall in the interval-basin during the 2010-2021 flood season (May-September). The Tyson polygon method is
142 used to calculate areal average rainfall using rainfall station data for each sub-basin area. Meanwhile, this study collects the
143 forecasted precipitation data issued by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the
144 Hydrology Bureau of the Yangtze River Water Resources Commission (HBYRWRC) for the 2017-2021 flood season in the
145 three sub-basins. Their forecast time starts at 8:00, with the 6-168h forecast horizons and the 6h forecast interval. The spatial
146 resolution of each grid for the ECMWF forecasted precipitation is 0.125°×0.125°. The HBYRWRC forecasted precipitation
147 is the areal average forecasted precipitation data.

148   The training period is from 2010 to 2016, and the validation period is from 2017 to 2021. Since the precipitation
149 forecast starts at 8:00 a.m., the forecasted flow for the 6-168h forecast horizons is also calculated from the daily 8:00 a.m. in
150 the validation period.

151 **3 Methods**

152 **3.1 Proposed CHUP-BMA method**

153 **3.1.1 Bayesian model averaging (BMA)**

154   Bayesian model averaging (BMA) method's principle is as follows.

$$p(Q_o|Q_{f,1}, Q_{f,2}, \ldots, Q_{f,k}) = \sum_{i=1}^{k} w_i \cdot p(Q_o|Q_{f,i}) \tag{1}$$

155 where, where, $p(\cdot)$ denotes the probability density function. Q_o denotes the observed flow corresponding to the forecast
156 moment (target value). k is the number of ensemble members. Q_f denotes the forecasted flow of ensemble members. w_i
157 denotes the weight of the i-th model. $p(Q_o/Q_{f,i})$ denotes the conditional probability density of Q_o conditional on $Q_{f,i}$, which is
158 assumed to approximately obey a normal distribution with the expectation of $\mu_i = a_i + b_i \cdot Q_{f,i}$ and variance of σ_i. a_i and b_i are the
159 bias correction coefficients obtained by linear fitting of $Q_{f,i}$ to Q_o.

160   Therefore, Eq. (1) can be rewritten as follows.

$$p(Q_o|Q_{f,1}, Q_{f,2}, \ldots, Q_{f,k}) = \sum_{i=1}^{k} w_i \cdot N(Q_o|\mu_i, \sigma_i) \tag{2}$$

161   From Eq. (2), it can be seen that the BMA method does not consider the influence of the initial state (the actual
162 observed flow at the start of the forecast) on the posterior distribution. When the member forecasts at different times are the
163 same, the posterior probability distribution generated by the BMA is also the same, which lacks logical rationality.

3.1.2 Hydrological uncertainty processor (HUP)

Based on the assumption that the precipitation uncertainty is zero, under the condition that the i-th ensemble member forecasts ($Q_{f,i}$) and the observed flow at the start of the forecast (Q_b), the posterior distribution of Q_o derived by the HUP method is as follows.

$$p\left(Q_o\middle|Q_{f,i},Q_b\right) = \frac{p\left(Q_{f,i}\middle|Q_o,Q_b\right)\cdot p(Q_o|Q_b)}{\int_{-\infty}^{+\infty} p\left(Q_{f,i}\middle|Q_o,Q_b\right)\cdot p(Q_o|Q_b)dQ_o} \tag{3}$$

where, $p(Q_o/Q_b)$ is the prior density function, $p(Q_{f,i}/Q_o,Q_b)$ is the likelihood density function. $p(Q_o/Q_{f,i},Q_b)$ is the posterior density function.

The HUP method assumes 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, which 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).

$$\hat{Q}_o = N^{-1}\left(P(Q_o)\right), \hat{Q}_{f,i} = N^{-1}\left(P(Q_{f,i})\right) \tag{4}$$

where, $P(\cdot)$ denotes the probability distribution function. $N^{-1}(\cdot)$ denotes the inverse function of the standard normal distribution. $\hat{Q}_o$ and $\hat{Q}_{f,i}$ are the observed and forecasted flow transformed to the normal space, respectively.

The HUP method also assumes that the observed flow obeys the strictly stationary first-order Markov process (Krzysztofowicz and Kelly, 2000), i.e., the flows between adjacent forecast horizons obey the linear constraint after the normal transformation.

$$\hat{Q}_{o,t} = c_t \times \hat{Q}_{o,t-1} + \varepsilon_t \tag{5}$$

where, $\hat{Q}_{o,t}$ is the observed flow corresponding to the t-th forecast horizon. c is the regression coefficient. ε is the residual, obeying N(0,1-c_t^2).

The prior density function expressions are as follows.

$$p(\hat{Q}_{o,t}|\hat{Q}_b) = \frac{1}{(1-C_t^2)^{0.5}}n\left\{\frac{\hat{Q}_{o,t}-C_t\times\hat{Q}_b}{(1-C_t^2)^{0.5}}\right\}, C_t = \prod_{i=1}^{t}c_i \tag{6}$$

where, $n(\cdot)$ denotes standard normal density function; $\hat{Q}_b$ is the observed flow at the start of the forecast transformed to the normal space.

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

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

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

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

188   The posterior density function under normal space can be derived by substituting Eqs. (6) and (7) into Eq. (3).

$$p(\hat{Q}_{o,t}|\hat{Q}_{f,i,t},\hat{Q}_b) = \frac{1}{Y_t}n\left\{\frac{\hat{Q}_{o,t} - (A_t \times \hat{Q}_{f,i,t} + D_t \times \hat{Q}_b + B_t)}{Y_t}\right\},$$

$$A_t = \frac{a_t y_t^2}{a_t^2 y_t^2 + \sigma_t^2}, B_t = \frac{-a_t b_t y_t^2}{a_t^2 y_t^2 + \sigma_t^2}, D_t = \frac{C_t \sigma_t^2 - a_t d_t y_t^2}{a_t^2 y_t^2 + \sigma_t^2}, Y_t = \left(\frac{y_t^2 \sigma_t^2}{a_t^2 y_t^2 + \sigma_t^2}\right)^{0.5}, y_t^2 = 1 - C_t^2 \tag{8}$$

189   The posterior distribution function under the normal space can be converted to the original space by Jacobian
190 transformation (Liu et al., 2016). The posterior density function of $Q_{o,t}$ under $Q_{f,i,t}$ and Q_b conditions is as follows.

$$p(Q_{o,t}|Q_{f,i,t},Q_b) = \frac{J(Q_{o,t})}{Y_t}n\left\{\frac{N^{-1}\left(P(Q_{o,t})\right) - A_t N^{-1}\left(P(Q_{f,i,t})\right) - D_t N^{-1}(P(Q_b)) - B_t}{Y_t}\right\},$$

$$J(Q_{o,t}) = \frac{p(Q_{o,t})}{n\left(N^{-1}\left(P(Q_{o,t})\right)\right)} \tag{9}$$

191   where, $J(\cdot)$ is the Jacobian transformation function.

3.1.3 HUP-BMA method

193   Darbandsari and Coulibaly et al. (2021) applied the hydrological uncertainty processor (HUP) to the ensemble forecast
194 members, substituted the posterior density function obtained by the HUP method (Eq. (9)) into the BMA framework (Eq.
195 (2)), and then obtained the posterior distribution function of the target flow based on the initial state and the forecasted flow
196 of the ensemble member. Therefore, the expression of the HUP-BMA method is as follows.

$$p(Q_o|Q_{f,1}, Q_{f,2}, \ldots, Q_{f,k}, Q_b) = \sum_{i=1}^{k} w_i \cdot \frac{J(Q_{o,t})}{Y_t} n\left\{\frac{N^{-1}\left(P(Q_{o,t})\right) - A_t N^{-1}\left(P(Q_{f,i,t})\right) - D_t N^{-1}(P(Q_b)) - B_t}{Y_t}\right\} \tag{10}$$

3.1.4 Copula-based HUP-BMA (CHUP-BMA) method

198   **(1) Copula-based HUP**

199   According to Sklar's theorem (Sklar, 1959), the joint distribution of m variables is as follows.

$$P(x_1, x_2, \ldots\ldots, x_m) = C_m(P(x_1), P(x_2), \ldots\ldots, P(x_m)) \tag{11}$$

200 where, $C_m(\cdot)$ denotes the m-dimensional copula distribution.

201   The copula-based HUP method (CHUP) was proposed by Liu et al. (2018), which can avoid the normal quantile
202 transformation process of the flow series in the standard HUP method. With the help of the copula function, the prior density
203 function in Eq. (3) can be derived as follows.

$$p(Q_o|Q_b) = \frac{\partial^2 C_2(P(Q_o), P(Q_b))}{\partial P(Q_o)\partial P(Q_b)} \cdot \frac{dP(Q_o)}{dQ_o} = c_2(P(Q_o), P(Q_b)) \cdot p(Q_o) \tag{12}$$

204   where, $c_m(\cdot)$ denotes the m-dimensional copula density function. *m denotes the dimension.*

205   The likelihood density function in Eq. (3) can be derived as follows.

$$p(Q_{f,i}|Q_o,Q_b) = \frac{\frac{\partial^3 C_3(P(Q_o),P(Q_{f,i}),P(Q_b))}{\partial P(Q_o)\cdot\partial P(Q_{f,i})\cdot\partial P(Q_b)}}{\frac{\partial^2 C_2(P(Q_o),P(Q_b))}{\partial P(Q_o)\cdot\partial P(Q_b)}} \cdot \frac{dP(Q_{f,i})}{dQ_{f,i}} = \frac{c_3(P(Q_o),P(Q_{f,i}),P(Q_b))}{c_2(P(Q_o),P(Q_b))}\cdot p(Q_{f,i}) \tag{13}$$

206   The posterior density function in Eq. (3) can be derived as follows.

$$p(Q_o|Q_{f,i},Q_b) = \frac{c_3(P(Q_o),P(Q_{f,i}),P(Q_b))}{\int_0^1 c_3(P(Q_o),P(Q_{f,i}),P(Q_b))dP(Q_o)}\cdot p(Q_o) \tag{14}$$

207   **(2) Copula-based HUP-BMA method**

208   Applying CHUP to the i-th ensemble member, the posterior probability distribution function $p(Q_o|Q_{f,i},Q_b)$ of Q_o based
209   on $Q_{f,i}$ and Q_b can be obtained. Coupling $p(Q_o|Q_{f,i},Q_b)$ into the BMA framework, the copula-based HUP-BMA (CHUP-
210   BMA) method can be constructed, and Eq. (2) can become as follows.

$$p(Q_o|Q_{f,1},Q_{f,2},\dots,Q_{f,k},Q_b) = \sum_{i=1}^{k} w_i \cdot \frac{c_3(P(Q_o),P(Q_{f,i}),P(Q_b))}{\int_0^1 c_3(P(Q_o),P(Q_{f,i}),P(Q_b))dP(Q_o)}\cdot p(Q_o) \tag{15}$$

211   The forecast uncertainty is quantified by the forecast interval with a 90% confidence level. Before constructing the
212   copula, selecting the marginal distribution and the copula type is usually necessary. This study intends to select the
213   appropriate marginal distribution and copula function from five common distribution functions, such as Pearson type III (P-
214   III), Gamma, Normal, Lognormal, and Weibull, and five common copula functions, such as Gumbel-Hougaard, Frank,
215   Clayton, Student-t (Student) and Gaussian copula, according to the root mean square error (RMSE) minimization criterion,
216   respectively. The definition and mathematical expressions of copula functions can be referred to Liu et al. (2018) and Chen
217   and Guo (2019).

218   Darbandsari and Coulibaly (2021) demonstrated that the HUP-BMA method could improve the probabilistic forecasting
219   performance of the HUP and BMA methods in the short forecast horizons. Therefore, this paper focuses on analyzing and
220   comparing the performance of the HUP-BMA and CHUP-BMA methods. The HUP-BMA and CHUP-BMA methods only
221   calibrate the ensemble members' weights through the Expectation-Maximization (EM) algorithm (Darbandsari and
222   Coulibaly, 2021). Meanwhile, since the forecast accuracy of ensemble members may change with time due to seasonality
223   and other factors (Zhong et al., 2020), the sliding window approach is used to update the weighting parameters. Parrish et al.
224   (2012) and Darbandsari and Coulibaly (2019) have shown that the BMA method with the sliding window can obtain better
225   probabilistic forecast performance compared to the method without the sliding window.

3.2 Ensemble forecasting scheme

An ensemble forecast scheme containing multi-source uncertainties in the model input, the model structure, and the parameter is constructed using a multi-member approach consisting of two forecasted precipitation, two models, and two objective functions used to calibrate parameters, as shown in Fig. 2.

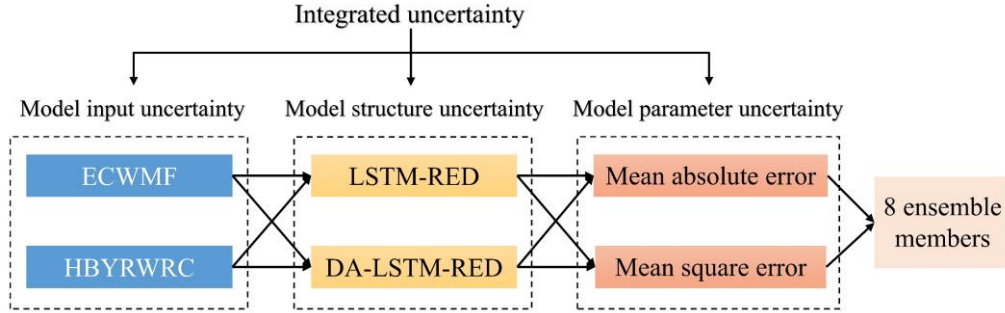



Fig. 2 The TGR's flood ensemble forecast scheme

3.2.1 Model input uncertainty

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. This study collected the forecasted precipitation data from the European Centre for Medium-Range Weather Forecasts (ECMWF) and HBCWRC in these three sub-basins. Since the rainfall data is more diverse and has relatively large uncertainty, the forecast rainfall input variable is used to explore the impact of forecast rainfall uncertainty on the TGR inflow forecasts. The TGR is a river-type reservoir, so building a river confluence model for flood forecasting is necessary. The observed and forecasted precipitations are converted to effective precipitation, i.e., the precipitation after the loss of plant retention, infiltration, evaporation, etc., to consider the runoff yield in the three sub-basin areas. The rainfall-runoff relationship (Fedora and Beschta, 1989) commonly used in the Yangtze River basin to 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).

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

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

248 where P_a denotes the antecedent precipitation index, P_t is the daily precipitation, I_m is the water storage capacity of the basin,

249 k denotes evaporation reduction index.

250 The values of k and I_m for these three sub-basins are listed in Table 2, which are obtained from the Hydrological Bureau

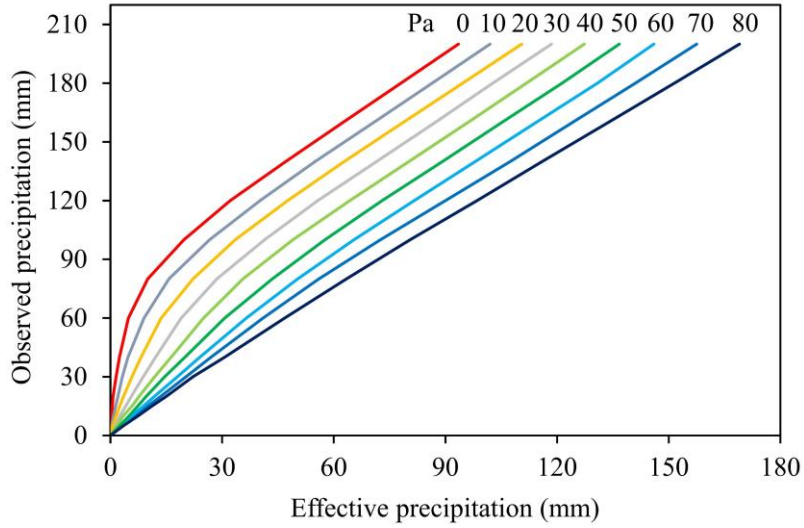
251 of the Changjiang Water Resources Commission (HBCWRC). Since the rainfall-runoff relationship graph method have been

252 widely used for runoff generation calculation in the Yangtze River basin, the rainfall-runoff relationship between Xiangjiaba

253 and Three Gorges Dam-site uncontrolled interval basin are established and plotted in Fig. 3, which is used to calculate the

254 effective precipitation based on the antecedent precipitation index (P_a) and observed (or forecasted) precipitation for these

255 three sub-basins.

256

257
258 **Fig. 3** Rainfall-runoff relationship between Xiangjiaba and Three Gorges Dam-site uncontrolled interval basin

259

260 **Table 2** The k and I_m values for the three sub-basin areas

Sub-basin	k	Im
Pingshan-Cuntan	0.90	50
Cuntan-Wanxian	0.95	80
Wanxian-TGR dam-site	0.95	80

261

262 After obtaining the daily antecedent precipitation index at 8:00, the antecedent precipitation index for the 6-h time scale

263 is calculated as follows.

$$P_{a,t,m} = (P_{a,t} + \sum P_{t,n}) \times k^{\frac{h}{24}} \tag{18}$$

264 where, $P_{a,t,m}$ denotes the antecedent precipitation index at m:00 on the t-th day. $\sum P_{t,n}$ denotes the cumulative observed

265 precipitation from 8:00 to m:00 on the t-th day. h denotes the time gap from 8:00 to m:00 on the t-th day.

3.2.2 Model structure uncertainty

267 The TGR inflow forecasting is influenced by the upstream mainstream and tributary reservoir scheduling decisions, the

268 rainfall intensity and distribution in the interval basin, and the changes in the subsurface characteristics, which is challenging

269 to establish complex and physical-based hydrological models (Yang et al., 2019; Cho et al., 2022; Hauswirth et al., 2023).

270 The simulation or forecast accuracy in this interval-basin needs to be improved to meet the needs of the work. Therefore, two

271 advanced data-driven models for obtaining multi-step-ahead flood processes forecasting, namely the long short-term

272 memory (LSTM-RED) model based on an encoder-decoder structure and the coupled dual attention LSTM-RED (DA-

273 LSTM-RED) model, are used for confluence calculations as a way to consider the uncertainty in the model structure. Since

274 the forecast data series at the outlets of tributaries are inconsistent, the observed flow at the outlets of five large tributaries

275 are used to train and validate the proposed models.

(1) Long short-term memory model based on encoder-decoder structure

277 The structure of LSTM neural network includes forgetting gate, input gate, updating the state of the memory unit, and

278 output gate (Hochreiter and Schmidhuber, 1997). The forgetting gate can select the relatively important information in the

279 previous memory unit. The input gate can select useful information from the input variables at the current moment. The

280 memory unit state can store relatively important information extracted from historical moments, which is updated under the

281 control of the forgetting gate and the input gate. The output gate selects and outputs useful information from the memory cell

282 state. More detailed procedures of the LSTM neural network formulation have been described by Kratzert et al. (2018).

283 This study nests an LSTM neural network into a recursive encoder-decoder (RED) structure that can be obtained for

284 forecasting flood processes to build an LSTM-RED model. Among them, the RED structure is similar to that of Kao et al.

285 (2020). The description of the LSTM neural network can be found in Cui et al. (2022). The encoding process of the RED

286 structure is used to extract the critical information (C_t) of the input (Xiang et al., 2020). In the decoding process, forecast

287 information of the same category as the encoding process is input to neural network of the latter moment, besides the C_t and

288 the output of the hidden layer at the previous moment.

(2) LSTM-RED neural network coupled dual attention mechanism

290 The LSTM-RED model based on the dual attention mechanism (DA-LSTM-RED) is established by adding the feature-

291 temporal dual attention mechanism to the LSTM-RED model, which can enable the model to highlight effective information

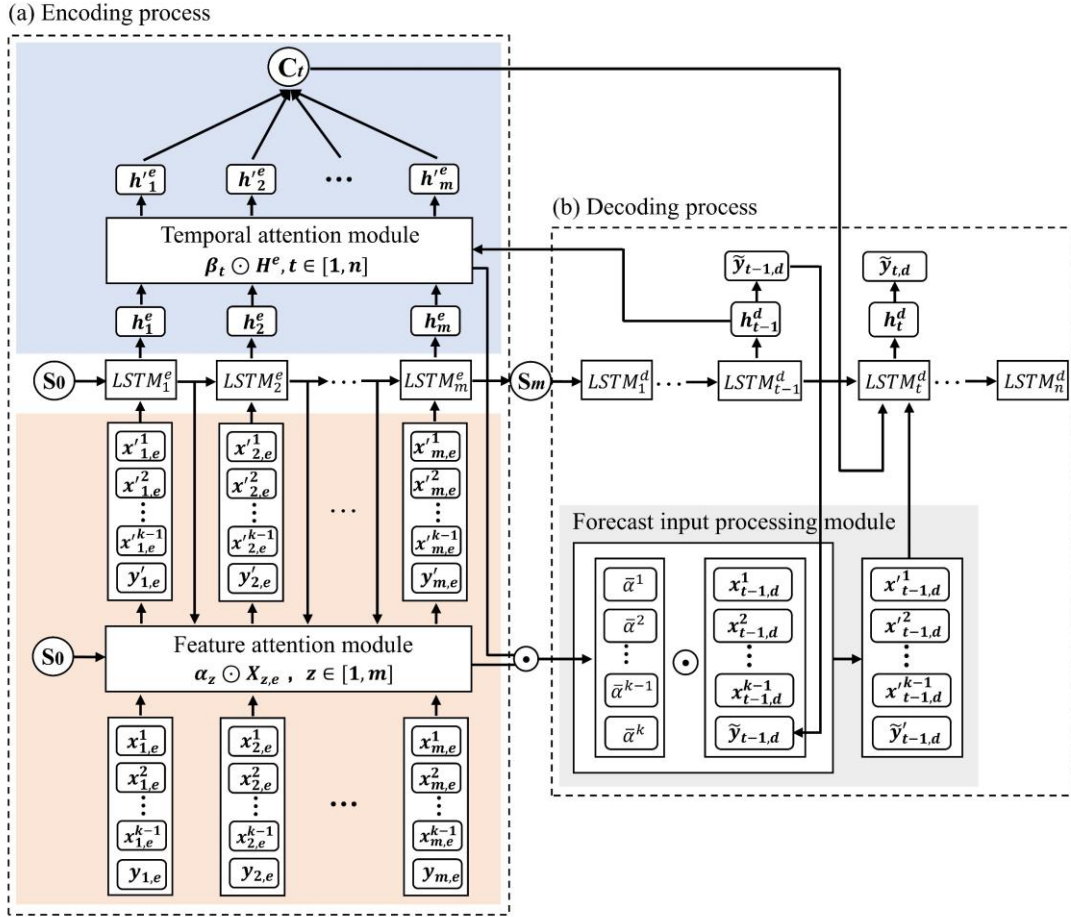
292 in different types and moments of the input. The DA mechanism (Fig. 4) consists of the feature attention module, the

293 temporal attention module, and the forecast input processing module.

294 The feature attention module can adaptively highlight the critical input types by assigning feature weights to the input

295 of the encoding process (Qin et al., 2017). The temporal attention module can highlight the information (hidden layer states)

296 extracted at critical time step by assigning temporal weights to the information extracted at all time step in the encoding

297 process (Ding et al., 2020). Meanwhile, the feature weights are averaged based on temporal weights and applied them to the
298 forecast information inputted in the decoding process, thus highlighting the key forecast input variables. The principle of the
299 DA-LSTM-RED model can be found in Cui et al. (2023).
300

301
302 **Fig. 4** Schematic diagram of the DA-LSTM-RED model. e and d are the encoding and decoding processes, respectively. k is the number of
303 input types. $X_{z,e}$ is the input variables of encoding process, $X_{z,e}=\{x_{z,e}^1, x_{z,e}^2, ..., x_{z,e}^{k-1}, y_{z,e}\}$. α_z denotes the weights of the input variables,
304 $\alpha_z=\{\alpha_z^1, \alpha_z^2, ..., \alpha_z^k\}$. m is the input time-steps in the encoding process. S is the hidden layer output. n is the maximum forecast horizon. H^e
305 is the hidden layer state, $H^e=\{h_1^e, h_2^e, ..., h_m^e\}$. β_t denotes the weights of the hidden layer states of the encoding process, $\beta_t=\{\beta_t^1, \beta_t^2, ..., \beta_t^m\}$.
306 C denotes the key information highlighted by the temporal attention. $\bar{\alpha}$ denotes the forecast input weights.
307

308 **(3) Model input and hyperparameter selection**
309   In this study, the input types for the encoding process include effective precipitation in the three sub-basins, flow
310 discharge in the mainstream and tributaries (i.e., five hydrological stations in Table 1), and previously observed inflow to the

311 TGR for a total of nine types data. In order to make the model learn comprehensive information, input variables with the last

312 11-time steps (66h) are inputted to the encoding process according to the flow propagation times from the hydrological

313 stations to the TGR dam site in Table 1.

314 The forecasted effective precipitation, the forecasted flow of the mainstream and tributaries, and the forecasted flow for

315 the previous forecast horizon are used as inputs of the decoding process. Among them, the forecasted effective precipitation

316 is calculated by the observed precipitation during the training period and by the forecast precipitation during the validation

317 period. The forecasted flow of the upstream mainstream and tributaries is replaced by the observed flow during the training

318 and validation periods. The TGR's observed inflow for the 6-168h forecast horizons is the target output, needed for practical

319 forecasting.

320 The input and output data are handled by the normalization method. Moreover, the trial-and-error method is used for

321 debugging the network hyperparameters. The model is trained by the Adam method (Kingma& Ba, 2014).

322 **3.2.3 Model parameter uncertainty**

323 Different parameter-optimization objective functions may focus on different forecast results (Zhong et al., 2020). For

324 example, the mean absolute error function focuses on the magnitude of the error mean. The mean square error function

325 usually is sensitive to outliers with large errors, which may make the model parameters with different objective functions

326 produce forecast results with different focus points (Duan et al., 2007). Therefore, it is necessary to consider the uncertainty

327 of the model parameters. Neural network models usually train model parameters (such as model internal weights and bias

328 values, etc.) based on loss functions, so this paper uses two common loss functions, namely the mean absolute error and the

329 mean square error, to train the model as a way to consider the uncertainty of model parameters.

330 **3.3 Evaluation metrics**

331 **3.3.1 Deterministic forecast evaluation metrics**

332 The accuracy of deterministic forecast is evaluated by three metrics: the Nash-Sutcliffe efficiency (Nash and Sutcliffe,

333 1970) (NSE), the mean absolute error (MAE) and the relative error of total runoff (RE).

$$NSE = 1 - \frac{\sum_{i=1}^{N}(Q_{o,i} - Q_{f,i})^2}{\sum_{i=1}^{N}(Q_{o,i} - \overline{Q_o})^2} \tag{19}$$

$$RE = \frac{\sum_{i=1}^{N} Q_{f,i} - \sum_{i=1}^{N} Q_{o,i}}{\sum_{i=1}^{N} Q_{o,i}} \times 100\% \tag{20}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Q_{o,i} - Q_{f,i}| \tag{21}$$

334 where, N is the sample number. $\overline{Q_o}$ and $\overline{Q_f}$ are the average of the observed and forecasted flow, respectively.

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.

3.3.2 Probabilistic forecast evaluation metrics

(1) Forecast interval evaluation metrics

The forecast interval is evaluated by three metrics: the average coverage cate (CR), average interval width (IW), and the percentage of observations bracketed by the unit confidence Interval (PUCI) (Li et al., 2011).

$$CR = \frac{n_c}{N} \tag{22}$$

$$IW = \frac{1}{N}\sum_{i=1}^{N}(Q_{u,i} - Q_{l,i}) \tag{23}$$

$$PUCI = \frac{CR}{\frac{1}{N}\sum_{i=1}^{N}(\frac{Q_{u,i} - Q_{l,i}}{Q_{o,i}})} \tag{24}$$

where, n_c denotes the number of Q_o located in the forecast interval. Q_u and Q_l are the upper and lower boundaries of the forecast interval with a 90% confidence level, respectively.

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.

(2) Probabilistic forecast evaluation metrics

The probabilistic forecast is evaluated by three metrics: the α_index (Renard et al., 2010), the ignorance score (IGS) (Gneiting et al., 2005), and continuous ranked probability score (CRPS) (Raftery et al., 2005).

$$\alpha\_index = 1 - \frac{2}{N}\sum_{i=1}^{N}|q_{e,i} - q_{th,i}| \tag{25}$$

$$IGS = -\frac{1}{N}\sum_{i=1}^{N}ln(p(Q_{o,i})) \tag{26}$$

$$CRPS = \frac{1}{N} \sum_{i=1}^{N} \int_0^{+\infty} \left(P_i(r) - I(r - Q_{o,i}) \right)^2 dr,$$

$$I(r - Q_{o,i}) = \begin{cases} 1 & r \geq Q_{o,i} \\ 0 & r < Q_{o,i} \end{cases}$$

(27)

356   where, $q_{e,i}$ and $q_{th,i}$ denote observed and theoretical p-values of $Q_{o,i}$, respectively. p-value denotes the posterior probability
357   distribution value of the $Q_{o,i}$ (Renard et al., 2010). $I(\cdot)$ denotes the indicator function. r denotes the flow variable.

358   The α_index metric can quantitatively assess the reliability of ensemble probabilistic forecasts from the perspective of
359   distribution function values (Renard et al., 2010). The closer the α_index value is to 1, the more reliable the probabilistic
360   forecast is. The IGS and CRPS metrics can reflect the reliability and sharpness of the probabilistic forecast. The former can
361   quantify the forecast probability density at the observation, while the latter can indicate the fit performance between the
362   posterior probabilistic distribution and the actual probabilistic distribution of Q_o (Raftery et al., 2005). Both CRPS and IGS
363   are negative scores, i.e., the smaller the value, the better. The IGS imposes severe penalties for particularly poor probabilistic
364   predictions and may be extremely sensitive to outliers and extreme events, yet also lacks robustness (Raftery et al., 2005).

365   **4 Results evaluation**

366   **4.1 Deterministic forecast results of ensemble member**

367   Since the study focuses on the differences in ensemble forecast performance between the HUP-BMA and CHUP-BMA
368   methods, the overall forecast accuracy of members is analysed (Fig. 5), and the differences in forecast accuracy between
369   members are not explicitly analysed. As shown in Fig. 5, using the observed values as input during the training period, high
370   forecast accuracy can be acquired in different forecast horizons, with the NSE values exceeding 0.95 and the MAE values
371   below 1400 m³/s, and the absolute value of RE within 4%.

372   After combining the forecasted precipitation during the validation period, the NSE values show a decreasing trend, and
373   the MAE and RE values show an increasing trend with the increase of the forecast horizon. Taking the NSE metrics of the 1-
374   7d forecast horizons as an example (Table 3), the average value of the NSE metric decreases from 0.97 to 0.89, which
375   indicates that the forecast accuracy gradually decreases. Meanwhile, the range of evaluation metrics gradually increases with
376   the increase of the forecast horizon. It can be seen from Table 3 that the difference between the maximum and minimum
377   values of NSE indicators for the 1d forecast horizon is only 0.01. In contrast, the difference for the 7d forecast horizon is as
378   high as 0.05, which indicates that the difference in forecast accuracy of members is also more significant, and the forecast
379   uncertainty gradually increases. Overall, the NSE values of the forecast members in the 6-168h forecast horizons are higher
380   than 0.88, and the absolute values of the RE metrics are within 7%. Hence, the forecast accuracy of members is relatively
381   high, and the forecast error is low, which can be used for flood ensemble forecasting.

382

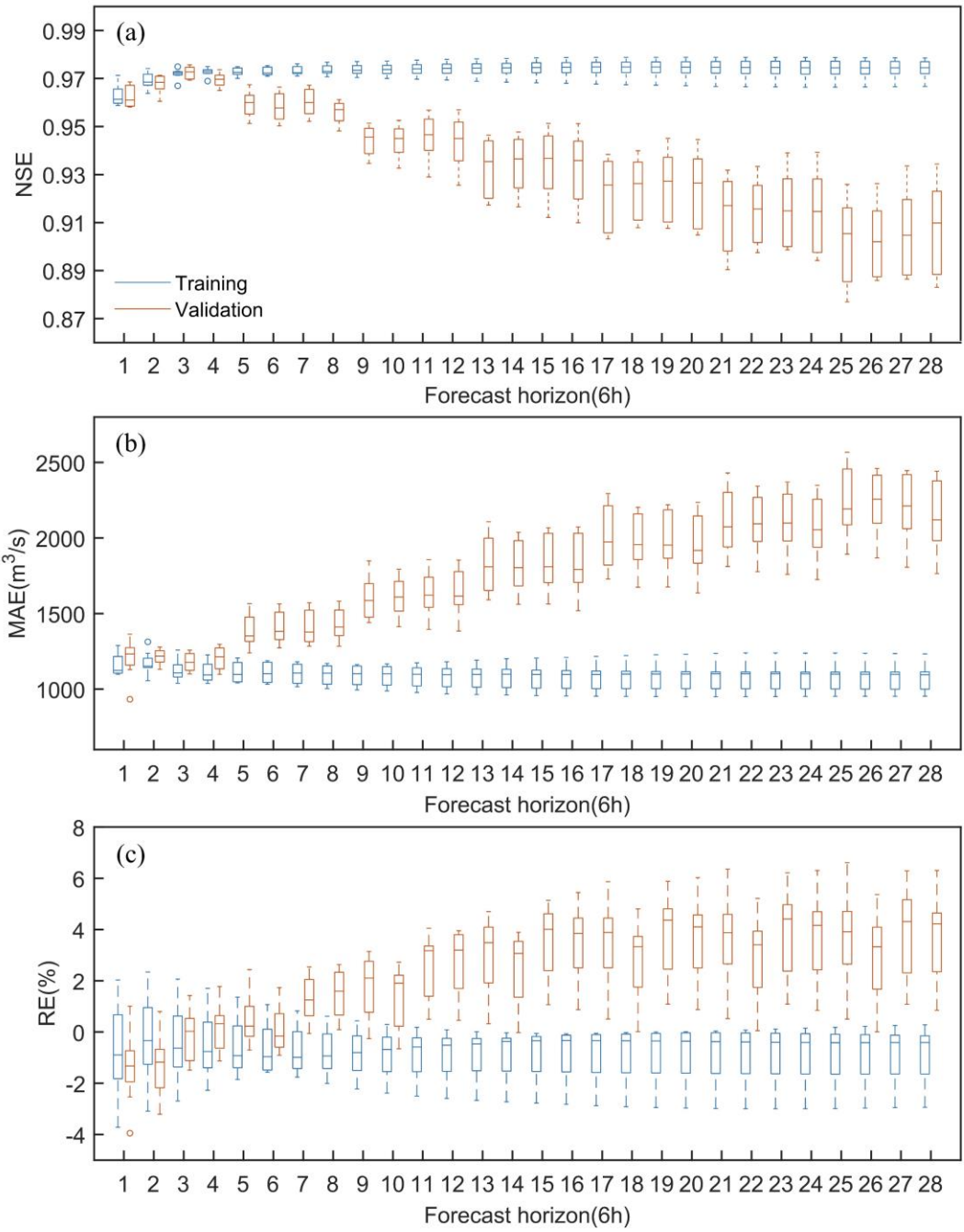



Fig. 5 Statistical chart of evaluation metrics of 8 ensemble members

Table 3 Mean, minimum, and maximum values of NSE metrics for 8 ensemble members in the validation period

Forecast horizon (h)	Mean	Max	Min	Forecast horizon (h)	Mean	Max	Min
6	0.96	0.97	0.96	90	0.93	0.95	0.91
12	0.97	0.97	0.96	96	0.93	0.95	0.91
18	0.97	0.98	0.97	102	0.92	0.94	0.90
24	0.97	0.97	0.97	108	0.92	0.94	0.91
30	0.96	0.97	0.95	114	0.93	0.95	0.91
36	0.96	0.97	0.95	120	0.92	0.94	0.90
42	0.96	0.97	0.95	126	0.91	0.93	0.89
48	0.96	0.96	0.95	132	0.91	0.93	0.90
54	0.94	0.95	0.93	138	0.92	0.94	0.90
60	0.94	0.95	0.93	144	0.91	0.94	0.89
66	0.95	0.96	0.93	150	0.90	0.93	0.88
72	0.94	0.96	0.93	156	0.90	0.93	0.89
78	0.93	0.95	0.92	162	0.91	0.93	0.89
84	0.93	0.95	0.92	168	0.91	0.93	0.88

389

390   **4.2 Ensemble forecast results**

391   **4.2.1 Marginal distribution and copula function selection**

392   It is necessary first to fit the marginal distributions of the observed flow and the forecasted flow of the 6~168h forecast

393   horizons. The Q_o and Q_b obey the same distribution. The RMSE criterion is used to select the marginal distribution type. In

394   each forecast horizon, the RMSE values of the 8 members are averaged to obtain the marginal distribution suitable for the

395   forecasted flow intuitively. Meanwhile, according to Eq.(14), the three-dimensional joint distribution of Q_o, Q_b, and Q_f needs

396   to be constructed. The RMSE criterion is used to select the copula function. Similarly, the RMSE values for the 8 members

397   of each forecast horizon were averaged.

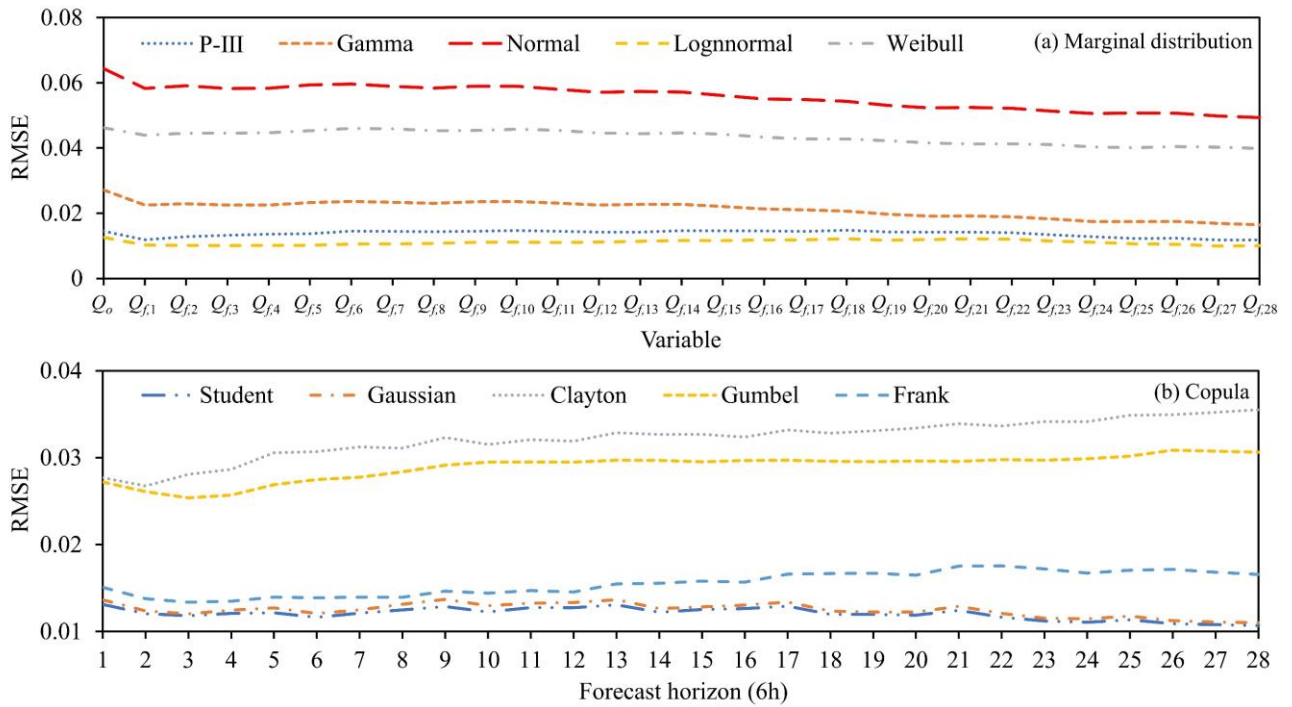
398   Figs.6(a) and (b) show the RMSE values generated by fitting the marginal distribution and copula function, respectively.

399   It can be seen from Fig. 6(a) that the Lognormal distribution has the lowest RMSE value among the five alternative marginal

400   distributions and is chosen as the sequence marginal distribution type. As shown in Fig.6(b), the student copula has the

401   lowest RMSE value in the 6-168h forecast horizons and is chosen to construct the three-dimensional joint distribution

402   function of Q_o, Q_b, and Q_f.

403

404

405 **Fig. 6** The RMSE values of Q_o, Q_b and Q_f sequence marginal distributions and copula functions. 1, 2, … ,28 denote 6h, 12h, …, 168h

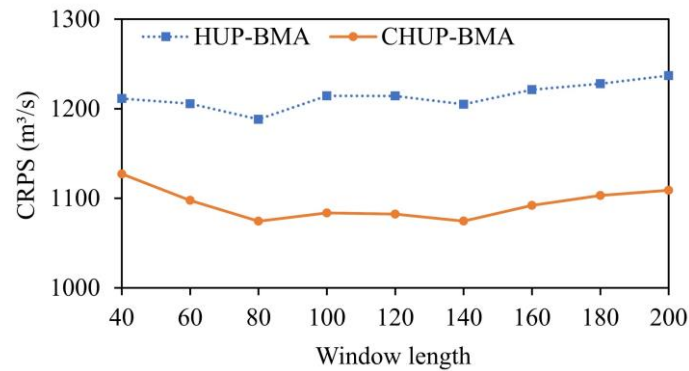
406 forecast horizons, respectively.

407 **4.2.2 Sliding window length selection**

408 Since there is no specific method or rule to calculate the sliding window length, this study adopts the CRPS metric as

409 the objective function and the trial-and-error method to select the sliding window length. The range of window lengths is [40,

410 200].

411

412

413 **Fig. 7** The average CRPS values of the CHUP-BMA and HUP-BMA methods with different window lengths

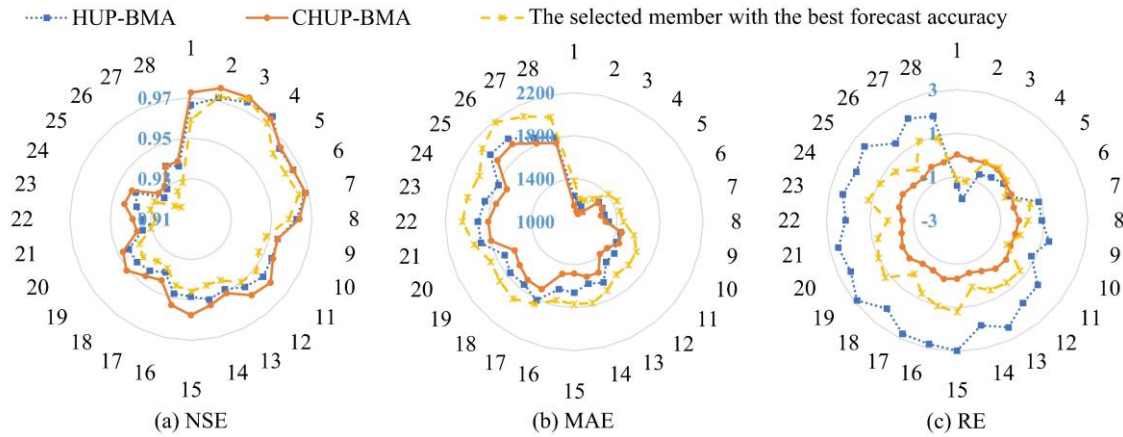
414

415   To facilitate the selection of the sliding window lengths, Fig. 7 shows the average CRPS values of the HUP-BMA and
416 CHUP-BMA methods for all forecast horizons with different window lengths. It can be seen from Fig. 7 that the HUP-BMA
417 and CHUP-BMA methods all have the lowest CRPS values at the sliding window length of 80. Therefore, 80 is the optimal
418 window length for the ensemble forecasting study.

419 **4.2.3 Deterministic forecast results of ensemble forecast**

420   The HUP-BMA and CHUP-BMA methods use expectation values of ensemble forecasts as deterministic forecast
421 results. In order to analyze the deterministic forecast performance of ensemble forecasts, one member with the best forecast
422 accuracy is selected for comparative analysis based on the criteria of the relatively low RE and MAE values and relatively
423 high NSE values, which is composed of the forecast rainfall from ECWMF, the DA-LSTM-RED model, and the
424 objective function with mean square error to optimize the parameters.

425

(a) NSE   (b) MAE   (c) RE

426
427 **Fig. 8** Deterministic forecast evaluation metrics for the HUP-BMA, the CHUP-BMA, and the selected member with the best forecast
428 accuracy

429

430   Fig.8(a), (b), and (c) show the NSE, MAE, and RE metrics of three deterministic forecast results, respectively. It can be
431 seen that the NSE metrics show a decreasing trend, and the MAE metrics show an increasing trend as the forecast horizon
432 increases, indicating a gradual decrease in forecast accuracy.

433   As shown in Fig. 8(a), the NSE metrics of three forecast results are at least 0.92 during the 6-168h forecast horizons.
434 The difference between the two is small, not more than 0.02. Among them, the CHUP-BMA method has the best NSE
435 metrics. However, the advantage value gradually decreases as the forecast horizon increases. The NSE metrics of the HUP-
436 BMA method are better than those of the selected forecast member in most forecast horizons. From Fig.8(b), the maximum
437 and mean values of MAE are 1923 and 1513 m³/s for the CHUP-BMA method, 1999 and 1582 m³/s for the HUP-BMA
438 method, and 2179 and 1719 m³/s for the selected forecast member, respectively. The CHUP-BMA method has the best MAE

metric, with the maximum and average reduction of 10.69% and 4.36% relative to the HUP-BMA method, respectively. Meanwhile, the MAE values of two ensemble forecasting methods are lower than those of the selected forecast members. As shown in Fig.8(c), the maximum and mean of the RE metric are 0.02% and -0.27% for the CHUP-BMA method, 2.97% and 1.36% for the HUP-BMA method, and 1.20% and 0.34% for the selected forecast member, respectively. The CHUP-BMA method can reduce the RE metrics of the selected forecast member in most forecast horizons, while the HUP-BMA method has no advantage in the RE metric. Overall, ensemble forecast methods can somewhat improve the selected best member forecast accuracy. The CHUP-BMA method's expectation forecast has the best accuracy, which indicates that the copula-based CHUP-BMA method can improve the performance of the HUP-BMA method in correcting errors.

To further analyze the accuracy of ensemble forecast methods, seven floods with peaks exceeding 50,000 m³/s during the 24 and 168h forecast horizons in the validation period (2017-2021) are selected for analyzing. The average relative error metric of peak (PRE) (Cui et al., 2022) is added to analyze the forecasting performance for flood peaks. Table 4 demonstrates the forecast evaluation metrics for the seven flood events. With the increase in the forecast horizon, the NSE metric shows a decreasing trend, and the RE and MAE metrics show an increasing trend, indicating a gradual decrease in forecasting performance. It can be seen from Table 4 that (1) in the 24h forecast horizon, the forecast accuracy of the two methods is similar for most flood events and quality metrics, (2) in the 168h forecast horizon, the forecast accuracy of the CHUP-BMA method is better than HUP-BMA method in most flood events and quality metrics. The average values of the NSE, RE, MAE, and PRE are 0.88, -0.63%, 2980m³/s, and -4.55% for CHUP-BMA, and 0.84, -2.38%, 3188m³/s, and -6.46% for HUP-BMA, respectively, indicating an overall improvement of CHUP-BMA over HUP-BMA in forecasting accuracy.

Table 4 Evaluation metrics for forecast flood events for 24 and 168h forecast horizons

Flood event	Method	Forecast horizon (h)	Evaluation metric			
			NSE	RE (%)	MAE (m³/s)	PRE (%)
20180703 (2018/7/1-7/7)	HUP-BMA	24	0.93	1.95	1697	-3.29
		168	0.80	1.69	2709	-8.60
	CHUP-BMA	24	0.94	3.63	1667	1.64
		168	0.78	1.30	2988	-6.26
20180714 (2018/7/11-7/17)	HUP-BMA	24	0.85	-1.38	2768	-8.04
		168	0.97	0.11	1101	0.88
	CHUP-BMA	24	0.84	-1.97	2874	-7.70
		168	0.95	-2.37	1587	-1.23
20200717 (2020/7/14-7/20)	HUP-BMA	24	0.91	-7.02	3094	-10.02
		168	0.64	-11.67	5965	-19.00
	CHUP-BMA	24	0.91	-4.75	3211	-8.80
		168	0.75	-7.45	5255	-13.58

20200727 (2020/7/25-7/31)	HUP-BMA	24	0.97	-0.22	1371	0.02
		168	0.84	-4.73	3044	-13.47
	CHUP-BMA	24	0.94	4.40	1819	3.62
		168	0.88	0.04	3155	-7.79
20200815 (2020/8/12-8/17)	HUP-BMA	24	0.93	-1.31	2714	-8.21
		168	0.94	-1.96	2259	-9.25
	CHUP-BMA	24	0.96	2.06	2062	-3.53
		168	0.95	3.05	2167	-3.82
20200820 (2020/8/18-8/24)	HUP-BMA	24	0.95	-0.79	2772	0.22
		168	0.92	5.74	3509	11.72
	CHUP-BMA	24	0.96	2.58	2125	2.60
		168	0.96	4.08	2816	6.06
20210907 (2021/9/4-9/10)	HUP-BMA	24	0.94	-3.26	2231	-7.43
		168	0.87	-4.66	3042	-13.15
	CHUP-BMA	24	0.97	-0.64	1722	-4.07
		168	0.94	-0.99	2016	-6.82

459

460   To further demonstrate the accuracy of flood process forecasting and applicability of the two methods, four relatively
461   large flood events are selected for comparative analysis for 168h forecast horizon (Fig.9). In the 20180703-flood event (Fig.
462   9a), the two methods have similar forecast performance, underestimating the peak and rising water processes and
463   overestimating the receding water process. The CHUP-BMA method has relatively low PRE values and total runoff error.
464   The HUP-BMA method accurately forecasts the peak present time. In the 20200815-flood event (Fig.9b), two methods
465   underestimate the flood peak and overestimate the receding water process. The HUP-BMA method has a larger flood peak
466   error, and the CHUP-BMA method has a better fitting performance. In the 20200820-flood event (Fig.9c), two methods
467   overestimate the observed flood process, with the CHUP-BMA method having the lower peak and total runoff error than the
468   HUP-BMA method. In the 20210907-flood event (Fig.9d), the CHUP-BMA and HUP-BMA methods underestimate the
469   flood peak and delay the forecast peak occurring time. The former has smaller peak and water volume error.
470

(a) 20180703 (2018/7/1-7/7)

(b) 20200815 (2020/8/12-8/17)

(c) 20200820 (2020/8/18-8/24)

(d) 20210907 (2021/9/4-9/10)

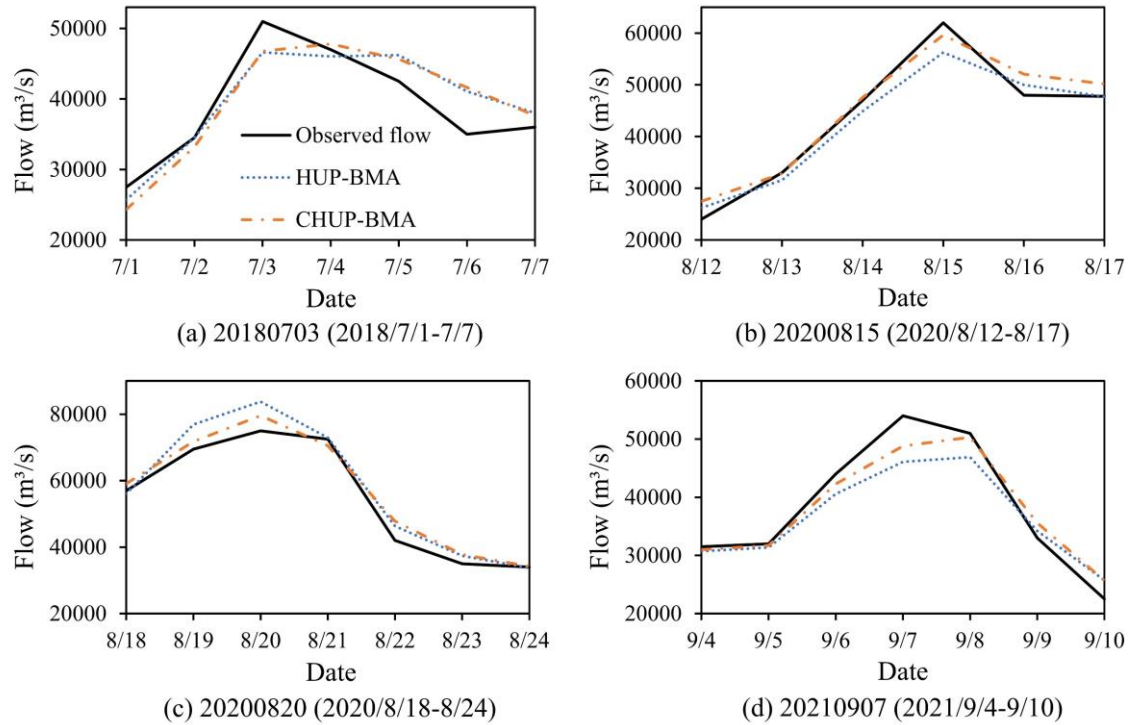



Fig. 9 Forecasted flood events during 168h forecast horizon for the HUP-BMA and the CHUP-BMA methods

4.2.4 Probabilistic forecast results of ensemble forecast

(1) Evaluation of forecast interval

Figs. 10(a), (b), and (c) show the CR, IW, and PUCI metrics for the forecast interval with a 90% confidence level, respectively. Fig.10(a) shows that during the 6-168h forecasting period, the maximum, minimum, and mean of the CR metric for the forecast interval of the CHUP-BMA method are 0.92, 0.88, and 0.89, respectively, and 0.93, 0.88, and 0.91 for the HUP-BMA method, respectively. The CR values of the two methods' forecast intervals are close to or exceed the 90% confidence level, indicating that the forecast intervals are reliable.

It is obvious from Fig.10(b) that the forecast interval width tends to increase with the increase of the forecast horizon, indicating that the forecast uncertainty gradually increases. The maximum, minimum, and mean of the IW metrics for the forecast interval of the CHUP-BMA method are 7820, 3337, and 6257 m³/s, respectively, and 8888, 4662, and 7345 m³/s for the HUP-BMA method, respectively. The forecast intervals of the CHUP-BMA method are significantly narrower than those of the HUP-BMA method, with the maximum and average reduction of 28.42% and 15.32%, respectively, which indicates that the CHUP-BMA method can effectively reduce the interval width and forecast uncertainty.

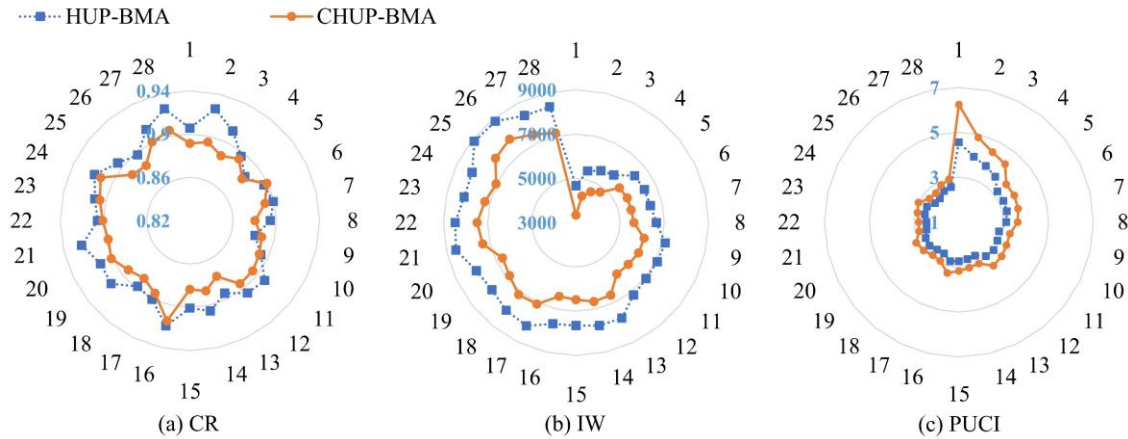



Fig. 10 Evaluation metrics of forecast intervals with the 90% confidence level of the HUP-BMA and CHUP-BMA methods

From Fig.10(c), the maximum, minimum, and mean of the PUCI metric for the forecast interval of the CHUP-BMA method are 6.24, 2.65, and 3.48, respectively, and 4.55, 2.35, and 2.95 for the HUP-BMA method, respectively. The CHUP-BMA method has the higher PUCI values, indicating that the forecast interval of the CHUP-BMA method reflects the forecast uncertainty relatively well.

In summary, the CHUP-BMA outperforms the HUP-BMA method under the premise that the CR values are close to or exceed the 90% confidence level. The CHUP-BMA method has narrower forecast intervals and better performance in quantifying forecast uncertainty. Although the HUP-BMA method has a higher CR value, its IW value is larger, and the PUCI value is smaller for the long forecast horizon, indicating that the forecast interval is too conservative to estimate the uncertainty range reasonably.

In order to visually analyze the ability of the CHUP-BMA method to quantify forecast uncertainty, the forecast intervals with a 90% confidence level of the HUP-BMA and CHUP-BMA methods for 168h forecast horizon in the 2020 flood season are compared. It can be seen from Fig.11 that the forecast intervals of the two ensemble forecasts can cover most of the observed flows and always cover the annual maximum flood peak, indicating that the forecast intervals are reliable. Meanwhile, the forecast intervals of the CHUP-BMA method are remarkably narrower than those of the HUP-BMA method, indicating that the forecast uncertainty of the former is relatively low, which can provide more reasonable risk information for TGR flood control decisions.

(2) Evaluation of overall probabilistic forecast

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.

513

514

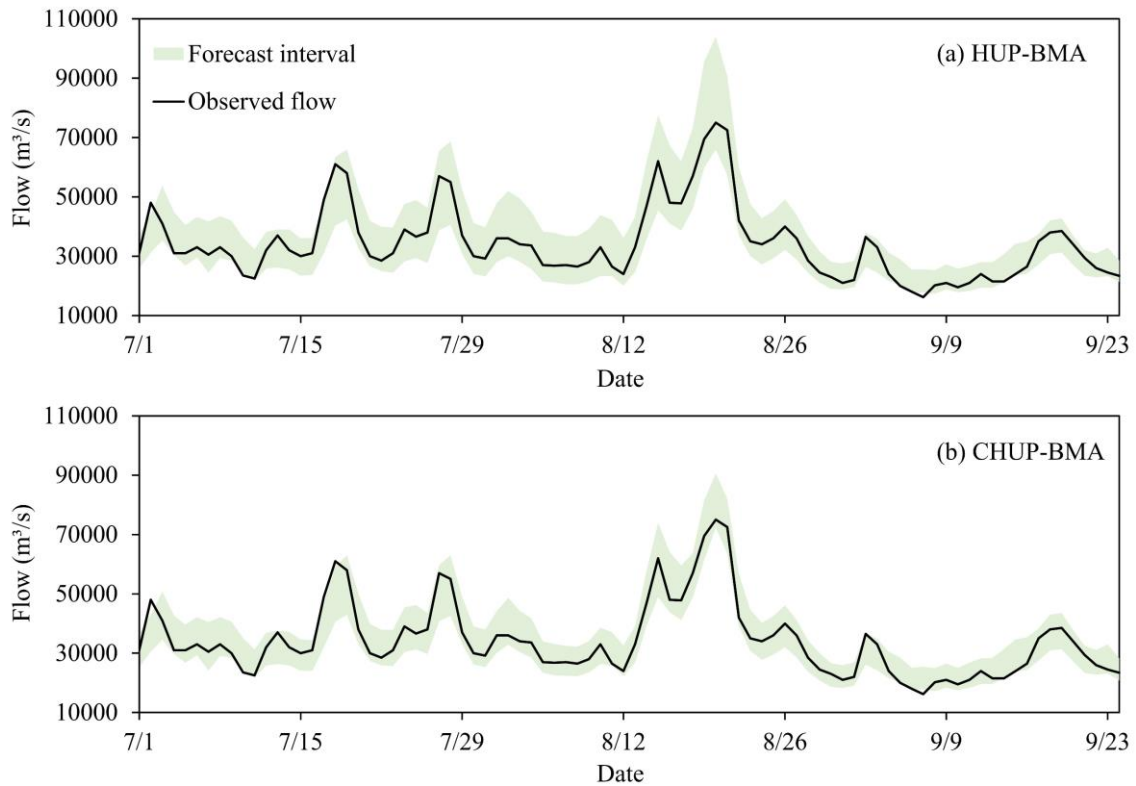



Fig. 11 Forecast intervals with the 90% confidence level for the HUP-BMA and CHUP-BMA methods from 2020/7/1 8:00 to 9/24 8:00

516

Meanwhile, Fig.13(a), (b), and (c) show the evaluation metrics of α_index, IGS, and CRPS metrics for the two ensemble probabilistic forecasts, respectively. It can be seen from Fig. 13(a) that the α_index metrics of the CHUP-BMA method-based probabilistic forecasts are significantly higher than those of the HUP-BMA method in the 6-168h forecast horizons. Among them, the maximum, minimum, and mean of the α_index metric for CHUP-BMA method-based probabilistic forecasts are 0.98, 0.93, and 0.97, respectively, and 0.95, 0.88, and 0.93 for the HUP-BMA method, respectively. The α_index metric of the CHUP-BMA method-based probabilistic forecast is closer to the perfect value of 1, indicating that its probability forecast is the more reliable.

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.

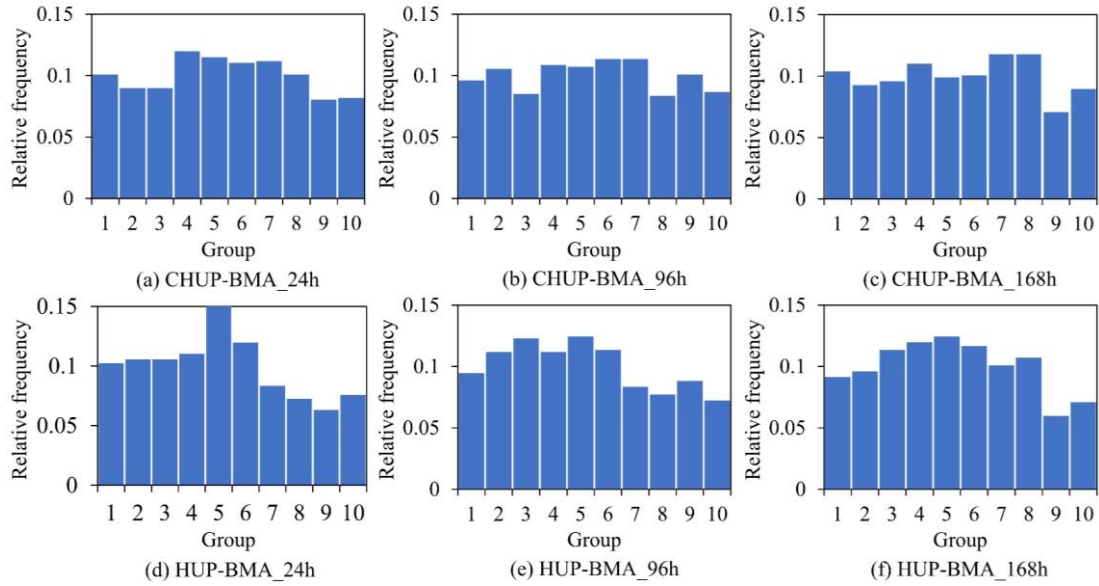



Fig. 12 The probability integral transform (PIT) histograms of the HUP-BMA and CHUP-BMA methods for the ensemble forecasts of the 24, 96, and 168h forecast horizons.

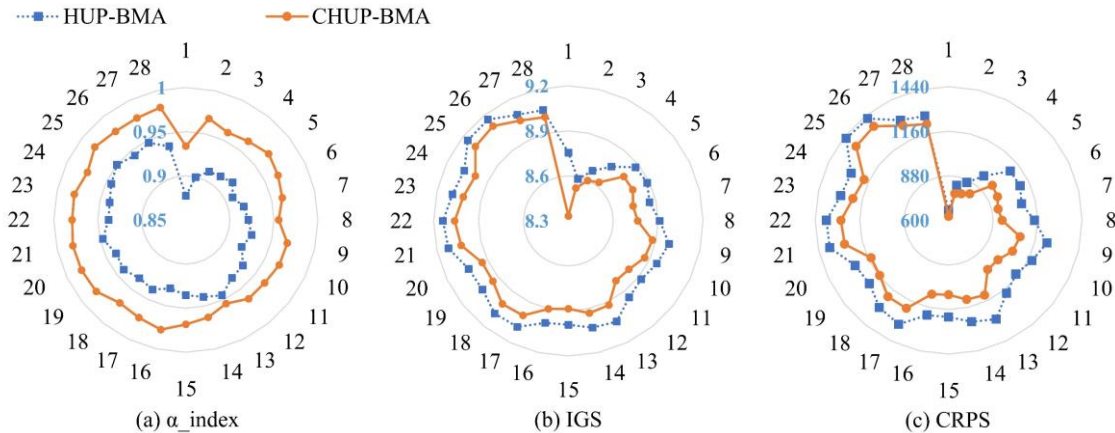



Fig. 13 Evaluation metrics of reliability, sharpness, and overall performance of two ensemble forecasts

As shown in Fig. 13(c), the CRPS values of the two methods are lower than the MAE values of the selected member (Fig. 8(b)), indicating that the probabilistic forecasts are effective and can fit the probabilistic distribution of the target values well. Meanwhile, during the 6-168h forecast horizons, the maximum, minimum, and mean of the CRPS metric for the CHUP-BMA method are 1356, 625, and 1074 m³/s, respectively, and 1425, 662, and 1188 m³/s for the HUP-BMA method, respectively. It can be seen that the CRPS values of the CHUP-BMA method are lower than those of the HUP-BMA method,

543 with a maximum and average reduction of 17.86% and 9.71%, respectively. It can be seen that the CHUP-BMA method can

544 better fit the posterior distribution of the actual values and effectively improve the probabilistic forecast performance of the

545 HUP-BMA method.

546   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

547 the threshold value, indicating that there is a significant difference between the scores of the CHUP-BMA and HUP-BMA

548 methods, i.e., the CHUP-BMA method is significantly better than the HUP-BMA method for ensemble forecasting metrics

549 and performance.

550

551 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	

552

553   In summary, the CHUP-BMA method considers the influence of the initial state on the ensemble forecast, bypasses the

554 normal quantile transformation of the HUP-BMA method, derives the posterior distribution of the target flow without

555 restrictions, and improves the probabilistic forecast performance of the HUP-BMA method. Therefore, the ensemble

556 forecasting by CHUP-BMA method can provide more reasonable and reliable risk information for the TGR.

557 **5 Conclusion and prospects**

558   In this study, we proposed a novel CHUP-BMA method that the copula-based HUP (CHUP) is coupled with the BMA

559 method, which not only can consider the influence of the initial state on the ensemble forecast, but also can avoid the

560 assumption of normal distribution in the HUP-BMA method and derive the posterior distribution function more accurately.

561 An ensemble forecast scheme that consists of two forecasted precipitation, two hydrological models, and two objective

562 functions of parameter calibration is established. The ensemble forecasting performance of the HUP-BMA and CHUP-BMA

563 methods is discussed from the perspective of deterministic and probabilistic forecasts. The flood ensemble forecasting

564 experiment with 6-168h forecast horizons is conducted in the Xiangjiaba-TGR dam-site interval basin. The main conclusions

565 are summarized as follows.

(1) The two ensemble forecasting methods can improve the members' forecast accuracy. The proposed CHUP-BMA method performs better than the HUP-BMA method, and the MAE metric value is reduced by a maximum of 10.69%.

(2) The coverage rate of the forecast interval of the CHUP-BMA method is close to or exceeds the specified 90% confidence level, and the forecast interval is significantly narrower than that of the HUP-BMA method, with a maximum reduction of 28.42%, which can effectively reduce the forecast uncertainty.

(3) The probabilistic forecast of the CHUP-BMA method has better reliability and sharpness, and its CRPS values are reduced by a maximum of 17.86% relative to the HUP-BMA method, which indicates that the CHUP-BMA method can better fit the posterior distribution of the actual values.

(4) The CHUP-BMA method can derive the posterior distribution of the target flow without restriction under the condition of considering the initial constraint, which makes the BMA method more towards perfection. Therefore, it is more suitable for the flood forecasting in the 6-168h forecast horizons and provides reliable risk information for reservoir scheduling decision-making.

The present study focuses on flood ensemble forecasting for the TGR's 6-168h forecast horizons. Future studies can explore the ensemble forecasting performance of the proposed CHUP-BMA method for longer forecast horizons and further validate the effectiveness of the proposed method in global basins. Meanwhile, 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. And the effective way or method of guiding reservoir scheduling based on ensemble forecasts can be further explored so that ensemble forecasts can be widely used in decision-making.

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

590

Code availability

592 The code used to support the findings of this study are available from the corresponding author upon request.

Data availability

594 The data generated and/or analyzed during the current study are not publicly available for legal/ethical reasons but are
595 available from the corresponding author on reasonable request.

Author contributions

597 Zhen Cui and Shenglian Guo conceived and designed the experiments; Zhen Cui performed the experiments and wrote the
598 manuscript draft; Zhen Cui, Shenglian Guo, Chong-Yu Xu, Hua Chen, Dedi Liu, and Yanlai Zhou reviewed and edited the
599 manuscript.

Competing interests

601 The authors declare that they have no conflict of interest.

Acknowledgments

603 This study was financially supported by the National Key Research and Development Program of China (2022YFC3202801,
604 2021YFC3200305), and China Three Gorges Cooperation (0799254).

References

606 Baran, S., Hemri, S. and El Ayari, M. Statistical post-processing of water level forecasts using Bayesian model averaging
607 with doubly-truncated normal components. Water Resour. Res. 55, 3997–4013. https://doi.org/10.1029/2018WR024028,
608 2019.

Biondi, D., and Todini, E. Comparing hydrological postprocessors including ensemble predictions into full predictive probability distribution of streamflow. Water Resour. Res., 54(12), 9860-9882, https://doi.org/10.1029/2017WR022432, 2018.

Chen, L., and Guo, S. Copulas and its application in hydrology and water resources. Springer Water, Springer Singapore, https://doi.org/10.1007/978-981-13-0574-0, 2019.

Cho, K., and Kim, Y. Improving streamflow prediction in the WRF-Hydro model with LSTM networks. J. Hydrol., 605, 127297. https://doi.org/10.1016/j.jhydrol.2021.127297, 2022.

Cloke, H. L., and Pappenberger, F. Ensemble flood forecasting: A review. J. Hydrol., 375(3-4), 613-626. https://doi.org/10.1016/j.jhydrol.2009.06.005, 2009.

Cui, Z., Guo, S., Zhou, Y., Wang, J. Exploration of dual-attention mechanism-based deep learning for multi-step-ahead flood probabilistic forecasting. J. Hydrol., 622, 129688. https://doi.org/10.1016/j.jhydrol.2023.129688, 2023.

Cui, Z., Zhou, Y., Guo, S., Wang, J., and Xu, C. Y. Effective improvement of multi-step-ahead flood forecasting accuracy through encoder-decoder with an exogenous input structure. J. Hydrol., 609, 127764. https://doi.org/10.1016/j.jhydrol.2022.127764, 2022.

Darbandsari, P., and Coulibaly, P. Introducing entropy-based Bayesian model averaging for streamflow forecast. J. Hydrol., 591, 125577. https://doi.org/10.1016/j.jhydrol.2020.125577, 2020.

Darbandsari, P., and Coulibaly, P. HUP-BMA: An Integration of Hydrologic Uncertainty Processor and Bayesian Model Averaging for Streamflow Forecasting. Water Resour. Res., 57(10), e2020WR029433. https://doi.org/10.1029/2020WR029433, 2021.

Ding, Y., Zhu, Y., Feng, J., Zhang, P., and Cheng, Z. Interpretable spatial-temporal attention LSTM model for flood forecasting. Neurocomputing, 403, 348-359. https://doi.org/10.1016/j.neucom.2020.04.110, 2020.

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

Fedora, M. A., and Beschta, R. L. Storm runoff simulation using an antecedent precipitation index (API) model. J. Hydrol., 112(1-2), 121-133. https://doi.org/10.1016/0022-1694(89)90184-4, 1989.

Ferretti, R., Lombardi, A., Tomassetti, B., Sangelantoni, L., Colaiuda, V., Mazzarella, V., Maiello, I., Verdecchia, M., and Redaelli, G.: A meteorological–hydrological regional ensemble forecast for an early-warning system over small Apennine catchments in Central Italy, Hydrol. Earth Syst. Sci., 24, 3135–3156, https://doi.org/10.5194/hess-24-3135-2020, 2020.

Gelfan, A., Moreydo, V., Motovilov, Y., and Solomatine, D. P. Long-term ensemble forecast of snowmelt inflow into the Cheboksary Reservoir under two different weather scenarios. Hydrol. Earth Syst. Sci., 22(4), 2073-2089. https://doi.org/10.5194/hess-22-2073-2018, 2018.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev., 133(5), 1098-1118. https://doi.org/10.1175/MWR2904.1, 2005.

643 Guo, Y., Yu, X., Xu, Y. P., Chen, H., Gu, H., and Xie, J. AI-based techniques for multi-step streamflow forecasts:
644 application for multi-objective reservoir operation optimization and performance assessment. Hydrol. Earth Syst. Sci.,
645 25(11), 5951-5979. https://doi.org/10.5194/hess-25-5951-2021, 2021.

646 Han, S., and Coulibaly, P. Bayesian flood forecasting methods: A review. J. Hydrol., 551, 340-351.
647 https://doi.org/10.1016/j.jhydrol.2017.06.004, 2017.

648 Hauswirth, S. M., Bierkens, M. F., Beijk, V., and Wanders, N. The suitability of a seasonal ensemble hybrid framework
649 including data-driven approaches for hydrological forecasting. Hydrol. Earth Syst. Sci., 27(2), 501-517.
650 https://doi.org/10.5194/hess-27-501-2023, 2023.

651 Hemri, S., Fundel, M. and Zappa, M. Simultaneous calibration of ensemble river flow predictions over an entire range of
652 lead times. Water Resour. Res. 49, 6744–6755. https://doi.org/10.1002/wrcr.20542, 2013.

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

655 Huang, H., Liang, Z., Li, B., Wang, D., Hu, Y., and Li, Y. Combination of multiple data-driven models for long-term
656 monthly runoff predictions based on Bayesian model averaging. Water Resour. Manag., 33, 3321-3338.
657 https://doi.org/10.1007/s11269-019-02305-9, 2019.

658 Kao, I. F., Zhou, Y., Chang, L. C., and Chang, F. J. Exploring a Long Short-Term Memory based Encoder-Decoder
659 framework for multi-step-ahead flood forecasting. J. Hydrol., 583, 124631. https://doi.org/10.1016/j.jhydrol.2020.124631,
660 2020.

661 Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, https://doi.org/10.48550/ar
662 Xiv.1412.6980, 2014.

663 Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M. Rainfall–runoff modelling using long short-term memory
664 (LSTM) networks. Hydrology and Earth System Sciences, 22(11), 6005-6022. https://doi.org/10.5194/hess-22-6005-2018,
665 2018.

666 Krzysztofowicz, R. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resour. Res.,
667 35(9), 2739-2750. https://doi.org/10.1029/1999WR900099, 1999.

668 Krzysztofowicz, R., and Kelly, K. S. Hydrologic uncertainty processor for probabilistic river stage forecasting. Water Resour.
669 Res., 36(11), 3265-3277. https://doi.org/10.1029/2000WR900108, 2000.

670 Kussul, N., Shelestov, A., and Skakun, S. Grid system for flood extent extraction from satellite images. Earth Sci. Inform., 1,
671 105-117. https://doi.org/10.1007/s12145-008-0014-3, 2008.

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

675 Li, L., Xu, C. Y., Xia, J., Engeland, K., and Reggiani, P. Uncertainty estimates by Bayesian method with likelihood of AR (1)
676 plus Normal model and AR (1) plus Multi-Normal model in different time-scales hydrological models. J. Hydrol., 406(1-2),
677 54-65. https://doi.org/10.1016/j.jhydrol.2011.05.052, 2011.

678 Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z. A review on statistical postprocessing methods for
679 hydrometeorological ensemble forecasting. Wiley Interdiscip. Rev.-Water, 4(6), e1246. https://doi.org/10.1002/wat2.1246,
680 2017.

681 Liu, J., Yuan, X., Zeng, J., Jiao, Y., Li, Y., Zhong, L., and Yao, L. Ensemble streamflow forecasting over a cascade reservoir
682 catchment with integrated hydrometeorological modeling and machine learning. Hydrol. Earth Syst. Sci., 26(2), 265-278.
683 https://doi.org/10.5194/hess-26-265-2022, 2022.

684 Liu, Z., Guo, S., Xiong, L., and Xu, C. Y. Hydrological uncertainty processor based on a copula function. Hydrol. Sci. J.-J.
685 Sci. Hydrol., 63(1), 74-86. https://doi.org/10.1080/02626667.2017.1410278, 2018.

686 Liu, Z., Guo, S., Zhang, H., Liu, D., and Yang, G. Comparative study of three updating procedures for real-time flood
687 forecasting. Water Resour. Manag., 30, 2111-2126. https://doi.org/10.1007/s11269-016-1275-0, 2016.

688 Madadgar, S., and Moradkhani, H. Improved Bayesian multi-modelling: Integration of copulas and Bayesian model
689 averaging. Water Resour. Res., 50(12), 9586-9603. https://doi.org/10.1002/2014WR015965, 2014.

690 Matthews, G., Barnard, C., Cloke, H., Dance, S. L., Jurlina, T., Mazzetti, C., and Prudhomme, C. Evaluating the impact of
691 post-processing medium-range ensemble streamflow forecasts from the European Flood Awareness System. Hydrol. Earth
692 Syst. Sci., 26(11), 2939-2968. https://doi.org/10.5194/hess-26-2939-2022, 2022.

693 Nash, J. E., and Sutcliffe, J. V. River flow forecasting through conceptual models: part I—A discussion of principles. J.
694 Hydrol., 10(3), 282-290. https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

695 Parrish, M. A., Moradkhani, H., and DeChant, C. M. Toward reduction of model uncertainty: Integration of Bayesian model
696 averaging and data assimilation. Water Resour. Res., 48(3). https://doi.org/10.1029/2011WR011116, 2012.

697 Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., and Cottrell, G. A dual-stage attention-based recurrent neural network for
698 time series prediction. arXiv preprint arXiv:1704.02971. https://doi.org/10.48550/arXiv.1704.02971, 2017.

699 Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. Using Bayesian model averaging to calibrate forecast
700 ensembles. Mon. Weather Rev., 133(5), 1155-1174. https://doi.org/10.1175/MWR2906.1, 2005.

701 Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W. Understanding predictive uncertainty in hydrologic
702 modeling: The challenge of identifying input and structural errors. Water Resour. Res., 46(5).
703 https://doi.org/10.1029/2009WR008328, 2010.

704 Shu, Z., Zhang, J., Wang, L., Jin, J., Cui, N., Wang, G., ..., and Liu, C. Evaluation of the impact of multi-source uncertainties
705 on meteorological and hydrological ensemble forecasting. Engineering. https://doi.org/10.1016/j.eng.2022.06.007, 2022.

706 Saleh, F., Ramaswamy, V., Georgas, N., Blumberg, A. F., and Pullen, J. A retrospective streamflow ensemble forecast for an
707 extreme hydrologic event: a case study of Hurricane Irene and on the Hudson River basin. Hydrol. Earth Syst. Sci., 20(7),
708 2649-2667. https://doi.org/10.5194/hess-20-2649-2016, 2016.

709 Sklar, M. Fonctions de repartition an dimensions et leurs marges. Publ. inst. statist. univ. Paris, 8, 229-231. 1959.

710 Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. Probabilistic quantitative precipitation forecasting using

711 Bayesian model averaging. Mon. Weather Rev. 135, 3209–3220. https://doi.org/10.1175/MWR3441.1, 2007

712 Sloughter, J. M., Gneiting, T., and Raftery, A. E. Probabilistic wind speed forecasting using ensembles and Bayesian model

713 averaging. J. Am. Stat. Assoc., 105(489), 25-35. https://doi.org/10.1198/jasa.2009.ap08615, 2010.

714 Todini, E. A model conditional processor to assess predictive uncertainty in flood forecasting. Int. J. River Basin Ma., 6(2),

715 123-137. https://doi.org/10.1080/15715124.2008.9635342, 2008.

716 Vegad, U., and Mishra, V. Ensemble streamflow prediction considering the influence of reservoirs in Narmada River Basin,

717 India. Hydrol. Earth Syst. Sci., 26(24), 6361-6378. https://doi.org/10.5194/hess-26-6361-2022, 2022.

718 Wei, L., Jiang, S., Dong, J., Ren, L., Liu, Y., Zhang, L., ..., and Duan, Z. Fusion of gauge-based, reanalysis, and satellite

719 precipitation products using Bayesian model averaging approach: Determination of the influence of different input sources. J.

720 Hydrol., 618, 129234. https://doi.org/10.1016/j.jhydrol.2023.129234, 2023.

721 Xiang, Z., Yan, J., and Demir, I. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. Water Resour.

722 Res., 56(1), e2019WR025326. https://doi.org/10.1029/2019WR025326, 2020.

723 Xiong, L., Wan, M. I. N., Wei, X., O'connor, K. M. Indices for assessing the prediction bounds of hydrological models and

724 application by generalised likelihood uncertainty estimation. Hydrological sciences journal, 54(5), 852-871.

725 https://doi.org/10.1623/hysj.54.5.852, 2009.

726 Xu, C., Zhong, P. A., Zhu, F., Yang, L., Wang, S., and Wang, Y. Real-time error correction for flood forecasting based on

727 machine learning ensemble method and its uncertainty assessment. Stoch. Environ. Res. Risk Assess., 1-21.

728 https://doi.org/10.1007/s00477-022-02336-6, 2022.

729 Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., ..., and Liu, C. Evaluation and machine learning improvement of

730 global hydrological model-based flood simulations. Environmental Research Letters, 14(11), 114027.

731 https://doi.org/10.1088/1748-9326/ab4d5e, 2019.

732 Zhang, B., Wang, S., Qing, Y., Zhu, J., Wang, D., and Liu, J. A vine copula-based polynomial chaos framework for

733 improving multi-model hydroclimatic projections at a multi-decadal convection-permitting scale. Water Resour. Res., 58(6),

734 e2022WR031954. https://doi.org/10.1029/2022WR031954, 2022.

735 Zhou, Y., Cui, Z., Lin, K., Sheng, S., Chen, H., Guo, S., and Xu, C. Y. Short-term flood probability density forecasting using

736 a conceptual hydrological model with machine learning techniques. J. Hydrol., 604, 127255.

737 https://doi.org/10.1016/j.jhydrol.2021.127255, 2022.

738 Zhou, Y., Guo, S., and Chang, F. J. Explore an evolutionary recurrent ANFIS for modelling multi-step-ahead flood forecasts.

739 J. Hydrol., 570, 343-355. https://doi.org/10.1016/j.jhydrol.2018.12.040, 2019.

740 Zhong, Y., Guo, S., Ba, H., Xiong, F., Chang, F. J., and Lin, K. Evaluation of the BMA probabilistic inflow forecasts using

741 TIGGE numeric precipitation predictions based on artificial neural network. Hydrol. Res., 49(5), 1417-1433.

742 https://doi.org/10.2166/nh.2018.177, 2018a.

743 Zhong, Y., Guo, S., Liu, Z., Wang, Y., and Yin, J. Quantifying differences between reservoir inflows and dam site floods
744 using frequency and risk analysis methods. Stoch. Environ. Res. Risk Assess., 32, 419-433. https://doi.org/10.1007/s00477-
745 017-1401-4, 2018b.

746 Zhong, Y., Guo, S., Xiong, F., Liu, D., Ba, H., and Wu, X. Probabilistic forecasting based on ensemble forecasts and EMOS
747 method for TGR inflow. Front. Earth Sci., 14, 188-200. https://doi.org/10.1007/s11707-019-0773-9, 2020.

748