Received: 19 May 2019 | Revised: 28 April 2020 | Accepted: 10 June 2020

DOI: 10.1002/met.1935

RESEARCH ARTICLE

Meteorological Applications RMetS

Intercomparison of multiple statistical methods in post-processing ensemble precipitation and temperature forecasts

Xiangquan Li[1] | Jie Chen[1,2] | Chong-Yu Xu[3] | Hua Chen[1] | Shenglian Guo[1]

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

[2]Hubei Provincial Key Lab of Water System Science for Sponge City Construction, Wuhan University, Wuhan, China

[3]Department of Geosciences, University of Oslo, Oslo, Norway

Correspondence
Jie Chen, State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China.
Email: jiechen@whu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 51779176, 51539009; the Thousand Youth Talents Plan from the Organization Department of CCP Central Committee; the Overseas Expertise Introduction Project for Discipline Innovation (111 Project) funded by the Ministry of Education and State Administration of Foreign Experts Affairs P.R. China, Grant/Award Number: B18037

Abstract

Ensemble weather forecasting generally suffers from bias and under-dispersion, which limit its predictive power. Several post-processing methods have been developed to overcome these limitations, and an intercomparison is needed to understand their performance. Four state-of-the-art methods are compared in post-processing precipitation and air temperature of the Global Ensemble Forecasting System (GEFS) reforecasts using a simple bias correction (BC) method as a reference. These methods include extended logistic regression (ExLR), generator-based post-processing (GPP), Bayesian model averaging (BMA) and affine kernel dressing (AKD). All these methods are tested over 659 national standard meteorological stations in China. The research concerns are the influence of region and forecast date and the role of BC on ensemble weather forecasting. It was found that: (1) the deterministic methods (GPP and ExLR) are more skilful than the probabilistic methods (BMA and AKD) in obtaining the well-calibrated and skilful ensemble forecasts; (2) the forecast skill of the post-processed ensemble weather forecasts is comparably high in the northern arid areas for precipitation, while the forecast skill for air temperature is only low in the Qinghai-Tibetan Plateau area; (3) the skill difference of the post-processed forecasts on different forecast date is only evident for air temperature, while not apparent for precipitation; and (4) only correcting bias for the ensemble weather forecasts can achieve about 0–70% (for precipitation) and 30–100% (for air temperature) forecast skill improvement for deterministic methods.

KEYWORDS

ensemble weather forecasting, Global Ensemble Forecasting System (GEFS), post-processing, method comparison, Bayesian model averaging

1 | INTRODUCTION

Ensemble weather forecasting (EWF) has been a growing field of numerical weather prediction (NWP) since the 1990s due to the fast-increasing computation resources (Gneiting and Raftery, 2005). Using EWF to generate ensemble forecasts involves running the NWP model multiple times with the perturbations added to the initial state and the model physics process (Bauer *et al.*, 2015). Many studies have been conducted to verify the advantages of ensemble weather forecasts, such as improving weather forecast predictability and providing forecast uncertainty information (Zhu *et al.*, 2002; Zhu, 2005; Leutbecher and Palmer, 2008). For example, Zhu (2005) concluded that the ensemble weather forecasts are better than the deterministic weather forecasts in several aspects, including maintaining a high forecast skill after three to five days in global modelling applications, extending the forecast lead time to about eight days and, most importantly, providing forecast uncertainty information. However, when comparing the ensemble forecasts with the corresponding observations, raw ensemble weather forecasts are typically unreliable, which results from insufficient model resolution, less-than-optimal initial conditions, suboptimal treatment of model uncertainty and sampling errors (Gneiting and Raftery, 2005). Furthermore, the skill of the ensemble forecasts is also influenced by systematic bias, insufficient representation of forecast uncertainty, and mismatched spatial scale between gridded forecasts and station-based observations (Hagedorn *et al.*, 2008; Hamill *et al.*, 2008; Scheuerer and Hamill, 2015).

Realizing the full potential of the ensemble forecasts requires statistical post-processing techniques that are used to remove the bias and reconstruct the proper ensemble spread. Various post-processing methods have been proposed and used for this purpose. These can be divided into two types according to the form of output, namely, probabilistic methods and deterministic methods.

Probabilistic methods seek to obtain the probabilistic forecasts calibrated from the raw ensembles. Some probabilistic methods build the relationships between the observed relative frequencies for the specified events and the probabilities derived from the raw ensemble weather forecasts. These relationships are usually based on the verification statistics used to evaluate the ensemble forecasts, such as rank histograms (Hamill and Colucci, 1997, 1998; Eckel and Walters, 1998), reliability diagrams (Atger, 2003) and the spread–skill relationship (Atger, 1999). Some probabilistic methods build the forecast model based on the raw forecasts to predict the probability of the specified event. For example, Hamill *et al.* (2004) used the logistic regression (LR) method and chose the ensemble mean as a predictor for precipitation and the ensemble mean anomaly as a predictor for temperature. The proposed method was evaluated for the improvement of the medium-range precipitation and air temperature forecast skill in the United States. Their results showed that the generated probabilistic forecasts are more skilful and reliable than the raw ensemble weather forecasts. Wilks (2009) further extended logistic regression (ExLR) to yield coherent probabilistic forecasts for multiple events simultaneously. Wilks (2006) used ensemble model output statistics (EMOS)—which is also called non-Gaussian regression (NGR) in some studies—proposed by Gneiting *et al.* (2005) to make a probabilistic forecast for the specified event. The EMOS assumes that the model mean can be predicted by the ensemble mean, and the model variance is a linear function of the ensemble variance. Other probabilistic methods are based on estimating the predictive probability distribution function (PDF) for generating probabilistic forecasts. Bayesian model averaging (BMA) is a statistical method used to combine forecasts from different sources. When the BMA is used to post-process ensemble forecasts, the predictive PDF is obtained by weighted averaging all PDFs centred on the individual bias-corrected forecasts. The weight assigned to the forecast is equal to the posterior probability of the model generating the forecasts and reflects the model's relative contribution to predictive skill over the training period. Raftery *et al.* (2005) used the BMA to improve the 48 hr temperature forecasts from the Pacific Northwest using the University of Washington's fifth-generation Pennsylvania State University–NCAR (National Center for Atmospheric Research) Mesoscale Model (MM5) ensemble. The results showed that the predictive PDF from the BMA is well calibrated and sharp compared with the raw ensemble forecasts. Sloughter *et al.* (2007) modified Raftery *et al.*'s (2005) BMA for post-processing precipitation forecasts. The revision is that the PDF for the individual member is replaced from a normal distribution by a discrete-continuous distribution, that is, a discrete probability for precipitation occurrence and a continuous gamma distribution for precipitation amount. The modified BMA was evaluated in the North America Pacific Northwest using the University of Washington's mesoscale ensemble. The results showed that the predictive PDF is well calibrated and sharp, the probability of precipitation (PoP) forecasts is much better calibrated than those based on the raw ensemble, and the estimates of high-precipitation amount probability are better than the results using the LR. Ensemble dressing is another strategy to obtain the forecast PDF by dressing the original forecast using historical error statistics or the variable's statistical properties (Roulston and Smith, 2003). Affine kernel dressing (AKD) (Bröcker and Smith, 2008) is a representative ensemble dressing strategy. In it, every member is represented by a kernel distribution with a common set of parameters linked to the ensemble, and the predictive PDF is obtained by equally weighted averaging the all-member PDF. The EMOS can also be extended to provide the predictive PDF for different weather variables, such as surface temperature

(Hagedorn *et al.*, 2008), precipitation (Scheuerer and Hamill, 2015; Baran and Nemoda, 2016) and surface wind speed (Taillardat *et al.*, 2016).

The deterministic methods use the forecast information from the ensemble weather forecasts for sampling from the historical observations, and the final output is the discrete regenerated ensemble forecasts. Roulin and Vannitsem (2012) proposed an inversing algorithm in Wilks's ExLR for generating the post-processed weather forecast ensembles by sampling the historical observations. This method was evaluated in two small catchments in Belgium using the European Centers for Medium-Range Weather Forecasts (ECMWF) forecasts. The results showed that the proposed methods improve reforecasts in terms of mean error and probabilistic performance. Chen *et al.* (2014) proposed a generator-based post-processing method (GPP) for post-processing precipitation and air temperature forecasts. The forecast generator is fitted using the historical observations which are selected based on the forecast information of the ensemble weather forecasts. The GPP was evaluated using the reforecasts from the Global Ensemble Forecasting System (GEFS) over two Canadian watersheds. The results showed that the GPP could increase the predictive power of the ensemble forecasts for one to seven lead days.

Various post-processing methods need to be compared to provide the guideline for method selection and insights for method improvement. Wilks (2006) summarized and compared eight probabilistic methods for post-processing ensemble weather forecasts in the Lorenz '96 settings, including: (1) early and ad-hoc approaches (direct model output, rank-histogram recalibration and multiple implementations of single-integration MOS equations); (2) ensemble dressing methods; (3) regression methods (LR and NGR); and (4) Bayesian methods (forecast assimilation and BMA). Finally, the study concluded that the LR, ensemble dressing and NGR were promising methods. Wilks and Hamill (2007) further compared the above three promising methods for post-processing the GEFS ensemble precipitation and temperature forecasts. They found that no single method was consistently better than the other two methods. Schmeits and Kok (2010) also found that the performance between the BMA and LR was not statistically significant when post-processing ECMWF ensemble precipitation forecasts. The above comparison studies mainly focus on the comparison between the probabilistic post-processing methods, while less attention is paid to the comparison with the deterministic post-processing methods. Vannitsem and Hagedorn (2011) compared the deterministic method error-in-variable model output statistics (EVMOS) and the probabilistic method NGR to improve the ECMWF temperature forecast performance over Belgium. The EVMOS is mainly used to correct the systematic bias and provide little improvement for the ensemble spread, while the NGR considers improving the ensemble spread. The results showed that the EVMOS could produce the ensemble consistent with the observations of the NGR, and even outperforms the NGR when the raw ensemble is highly skewed, or the extreme event occurred. Therefore, both deterministic and probabilistic post-processing methods need to be compared for a better understanding of the advantages and disadvantages of these methods.

The performance of the post-processing methods may be influenced by multiple factors, such as the region and date when the forecasts are made (Atger, 2003; Hagedorn *et al.*, 2008; Scheuerer and Hamill, 2015). Therefore, the study will provide some useful insights for method selection if the influence of these factors can be considered in the methods comparison.

For many post-processing methods, addressing the issue of bias and under-dispersion requires two associated procedures, bias correction (BC) and calibrating the PDF. The post-processing method differs in the way it calibrates the PDF, but the role of the BC in the post-processing methods has received less attention (Hagedorn *et al.*, 2008; Schmeits and Kok, 2010). It was also found that implementing the BC before weighing each member in using a BMA would bring an overweighting of climatology, finally resulting in an increase in the mean squared error (Erickson *et al.*, 2012; Hodyss *et al.*, 2016).

The study evaluated and compared four state-of-the-art post-processing methods: the BMA, AKD, weather GPP and ExLR, in order to post-process precipitation and air temperature over a large study area covering different climates and topographies. Their performances were also compared with a reference method: the BC. The following scientific questions will be addressed:

- How do the selected four methods perform in post-processing precipitation and air temperature ensemble forecasts?
- Is there any performance difference between the deterministic methods and probabilistic methods?
- How does the performance of the four methods vary across space and time?
- To what extent is the forecast skill of the ensemble weather forecasts better than the reference method?

2 | DATA AND STUDY AREA

2.1 | Study area

The post-processing methods were evaluated over the mainland of China, between 16–52° N and 75–133° E.

The study area is a vast territory with various climate categories and complex topographic conditions, which allow a consideration of the influence of region selection over the post-processing methods. According to the climate region division in Wang and Li (2007), fig. 3), China is classified into seven climate regions: Northeast (NE), North (N), Northwest (NW), East (E), Southwest (SW), South (S) and Qinghai-Tibetan Plateau (QT). The weather stations chosen from the Chinese mainland must have had at least 30 year available observations for evaluating the post-processing methods; 659 national standard meteorological stations were therefore selected (Figure 1).

2.2 | Data

The post-processing methods were used to post-process two different weather variables: daily precipitation and daily mean air temperature. The observed precipitation and mean air temperature were obtained from the China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn) covering the period 1961–2014. Ensemble precipitation and air temperature forecasts were taken from the second version of the Global Ensemble Forecasting System (GEFS) reforecasts (http://portal.nersc.gov/project/refcst/v2/) (Hamill *et al.*, 2013). Eleven-member forecasts of up to 16 days were provided from December 1984, and this data set was archived with a global grid of 1° for latitude and longitude.

For the study, GEFS forecasts and observations were subsampled using the common period 1985–2014. The GEFS forecast data were interpolated to the surface observation locations by the inverse distance weighting (IDW) method using the nearby four grids. Since one week is the maximum lead time for skilful precipitation forecasts (Liu and Coulibaly, 2011; Chen *et al.*, 2014), the GEFS reforecasts with seven lead days were used in the study.

3 | METHODOLOGY

The post-processing methods used included two deterministic methods and two probabilistic methods. The deterministic methods included the GPP and ExLR; the probabilistic methods used the BMA and AKD. For both deterministic and probabilistic methods, the final post-processed forecasts were presented in the form of a multi-member ensemble. For the probabilistic methods, the final PDF was transformed into an ensemble by random sampling, whose frequency was defined as 1,000 to represent the property of the PDF.

Two different weather variables, precipitation and air temperature, were used for the evaluation of the post-processing methods. The distinct statistical property of the two variables offers an excellent opportunity to evaluate these post-processing methods comprehensively. Specifically, the air temperature forecasts are normally distributed, while the precipitation forecasts are skewed. Generally, the post-processing of air temperature with a

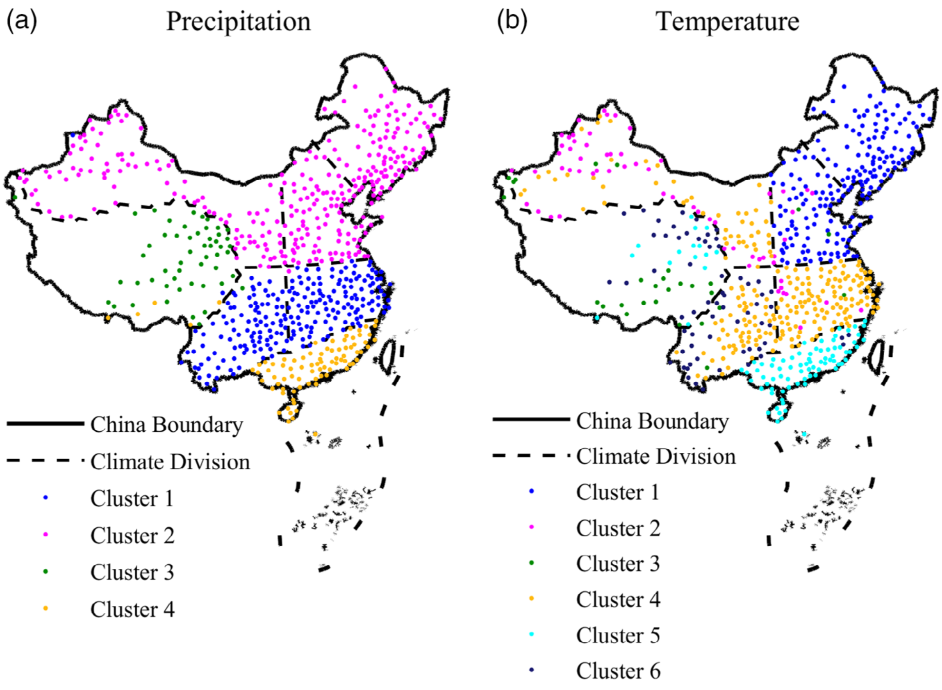


FIGURE 1 Clustered weather stations using different colours for precipitation (a) and air temperature (b). The dashed line separates different climate divisions in China

normally distributed variable is relatively easy, while the post-processing of precipitation has significant challenges. For example, Scheuerer and Hamill (2015) listed the three difficulties for post-processing precipitation: (1) the precipitation forecasts are hard to depicted because of their mixed discrete/continuous nature; (2) forecast uncertainty is generally greater for a large precipitation amount; and (3) the high precipitation amount occurs very infrequently. Thus, the calibration of precipitation forecasts requires much more extensive training data than the air temperature forecasts to cover a possible rare event (Hagedorn *et al.*, 2008).

When simulating the highly skewed distributed precipitation amounts using the methods listed above, it was found that the precipitation amounts are poorly fitted, especially for extreme values. Using the power-transformed precipitation amounts can give an especially good fit (Sloughter *et al.*, 2007; Hagedorn *et al.*, 2008). Settings the exponent as $\leq 1/3$ gives the best fitting performance when different exponent values (1, 1/2, 1/3, 1/4 and 1/5) were tested to fit the power-transformed non-zero precipitation amounts to the skewed two-parameter Gamma distribution. Using the power-transformed precipitation amounts is also recommended by Sloughter *et al.* (2007) and Roulin and Vannitsem (2012), the authors of the BMA and ExLR methods. Therefore, the cubic root of the precipitation amounts was used for the four post-processing methods and BC.

3.1 | Generator-based post-processing (GPP)

The GPP proposed by Chen *et al.* (2014) uses a post-processing generator with parameters linked to the ensemble forecasts. The generated ensemble forecasts are proved to be fully coherent with the ensemble forecasts. A brief introduction to the GPP now follows.

3.1.1 | Precipitation

For post-processing the precipitation forecasts, the first step is to define precipitation classes according to precipitation intensity. For each season, precipitation intensity is divided into several precipitation classes according to the ensemble mean precipitation amounts. In the study, 10 precipitation classes were defined using 11 quantiles (0.0, 0.1, ..., 1.0) of the non-zero ensemble-mean precipitation amounts for all forecasts in the same season during all historical periods. The GPP is calibrated using the relationship between the forecasted precipitation classes and the probability of observed occurrence or observed precipitation amounts. Specifically, for each class, the probability of the observed precipitation occurrence corresponding to this class is used as the probability of precipitation (PoP). The cubic root of the non-zero precipitation amount is supposed to follow a two-parameter gamma distribution. For any given day, ensemble forecasts with an arbitrary size (1,000 in the study) can be generated using the following steps: (1) determine the precipitation class according to the ensemble mean precipitation; (2) 1,000 random numbers generated from the 0–1 uniform distribution are used to represent 1,000 possible precipitation probabilities; and (c) the 1,000 members whose random number is ≤ PoP are deemed as wet, and the precipitation amount for these wet members is generated using the fitted gamma distribution.

3.1.2 | Temperature

Using the GPP to post-process air temperature forecasts consists of two associated procedures: the BC and reconstructing the ensemble spread. The ensemble mean air temperature forecasts are corrected using the linear regression method. A linear equation is calibrated for each day by fitting between the observed and ensemble mean air temperature anomalies using a neighbouring 15 day window. The observed and ensemble mean air temperature anomalies are obtained by subtracting the long-term daily mean observed air temperature from the corresponding observed and ensemble mean air temperature. The ensemble spread is generated using a two-parameter normal distribution. The corrected ensemble mean air temperature is used as the mean of the normal distribution, and the standard deviation of the normal distribution is calibrated on the seasonal scale by using an iterative method proposed by Chen *et al.* (2014). For a given day, the post-processed air temperature ensembles are generated by repeatedly multiplying the optimized standard deviation by a normally distributed random number and adding to the bias-corrected ensemble mean air temperature.

3.2 | Extended logistic regression (ExLR)

The ExLR associates the probability that weather quantity y (e.g. daily precipitation amount or air temperature) is less than or equal to the threshold q to the predictor X (e.g. the ensemble mean or ensemble spread) and the threshold q itself:

$$P(y \leq q) = H[f(X) + g(q)] \tag{1}$$

where $H(\cdot)$ is the logit function with the form of $H(t) = [1 + \exp(-t)]^{-1}$; $f(X)$ is a linear combination of the predictors X; and $g(q)$ is a non-decreasing function concerning the threshold q. For precipitation $g(q)$ takes the form $\beta_g \times \sqrt{q}$, the same as in Wilks (2009) and Roulin and Vannitsem (2012). For temperature, $g(q) = \beta_g \times q$, where the air temperature can be negative; and β_g is the parameter in $g(q)$.

The selection of predictors X is essential for building the LR model. The critical information for the ensemble is found to be the ensemble mean and ensemble spread (Wilks, 2009). However, Roulin and Vannitsem (2012) found that choosing the ensemble spread offers marginal improvement for precipitation. Therefore, for air temperature, both the ensemble mean and ensemble spread are selected as predictors. For precipitation, the cubic root of the ensemble mean is used.

Roulin and Vannitsem developed a method to generate forecasts by inverting the logistic function in Equation 2:

$$y = g^{-1}\left[\frac{\ln\left(\frac{1-P}{P}\right) - f(X)}{\beta_g}\right] \tag{2}$$

where P is a random number drawn from the uniform distribution.

For any given day forecast, ensemble forecasts with 1,000 members are generated using Equation 2.

3.3 | Bayesian model averaging (BMA)

3.3.1 | Precipitation

The forecast PDF for the cubic root of precipitation accumulation y is defined as:

$$p(y|f_1,\dots,f_K) = \sum_{k=1}^{K} W_k\{P(y=0|f_k)\cdot I(y=0) + P(y>0|f_k)\cdot g_k(y|f_k)\cdot I(y>0)\} \tag{3}$$

where f_k is the cube root of precipitation amount for member k; K is the number of members; W_k is the posterior probability of ensemble member k to be selected; I[...] is unity if the condition in brackets holds, and 0 otherwise; $P(y = 0|f_k)$ and $P(y > 0|f_k)$ are the probabilities of non-precipitation and precipitation given the forecast f_k, respectively; $g_k(y|f_k)$ is the conditional PDF of the cube root precipitation amount y given that y is positive for member k, and $g_k(y|f_k)$ takes the form of a gamma distribution.

For parameter estimation, $P(y = 0|f_k)$ and $P(y > 0|f_k)$ are estimated by an LR model. The model uses f_k and the member precipitation state indicator as predictors and is associated with three member-specific parameters. $G_k(y|f_k)$ is a two-parameter gamma distribution; and the gamma mean and variance are assumed to have a linear relationship with the member value and the ensemble variance, respectively. The two parameters of the Gamma mean correction model are also member specific. The weights W_k and two correction parameters for the gamma variance correction model need to be optimized by the expectation–maximization (EM) technique. For more details about parameter estimation, see Sloughter *et al.* (2007).

3.3.2 | Temperature

The forecast PDF of daily air temperature y is specified by:

$$p(y|f_1,\dots,f_K) = \sum_{k=1}^{K} W_k l_k(y|f_k) \tag{4}$$

where f_k is the air temperature forecast for member k; K is the size of the ensemble; W_k is the posterior probability of ensemble member k to be the best one; and $l_k(y|f_k)$ is the conditional distribution of y for the k-th member, and here it takes the form of a two-parameter normal distribution.

For parameters estimation, the Normal mean of $l_k(y|f_k)$ is assumed to have a linear relationship with the member values and contains two member-specific parameters. The Normal variance and the weights W_k are optimized using the EM technique. For more details about parameter estimation, see Raftery *et al.* (2005).

For any given day, ensemble forecasts with 1,000 members are generated by randomly sampling the built PDF from Equations 4 or 5.

3.4 | Affine kernel dressing (AKD)

The forecast PDF for the weather variable is a combination of K kernel distributions, as specified by:

$$p(y|f_1,\dots,f_K;\theta) = \frac{1}{K\sigma}\sum_{k=1}^{K} l\left(\frac{y-z_k}{\sigma}\right) \tag{5}$$

where y is the forecast variable; f_k is the ensemble forecast for the member k; K is the number of members; $l(.)$ is the kernel distribution (a Normal distribution is selected here); and z_k and σ are the mean and bandwidth for the kernel distribution, as defined by:

$$z_k = a \times f_k + r_2 \times m(F) + r_1 \tag{6}$$

$$\sigma^2 = h_s^2 (s_1 + s_2 \times \mathrm{var}(Z)) \tag{7}$$

where $m(F)$ is the ensemble mean; $h_s = 0.5 \cdot (4/[3 \cdot K])^{1/5}$ is Silverman's factor; var(Z) is the variance of the affine ensemble; a is the scaling parameter for the ensemble; and r_1, r_2, s_1 and s_2 are optimized by a sequence of constrained quadratic optimization algorithms. For more details about the parameter estimation, see Bröcker and Smith (2008).

For any given day forecast, ensemble forecasts with 1,000 members are generated by randomly sampling the PDF from Equation 5.

3.5 | Bias correction (BC)

The BC proposed by Chen *et al.* (2014) is used as a reference to evaluate the above post-processing methods. Here it only corrects the bias and is served as a reference to show the additional improvement of the post-processing methods, which address bias and under-dispersion simultaneously.

3.5.1 | Precipitation

The linear equation with the form $y = ax$ (where a is the regression co-efficient) is used to correct precipitation forecasts. Dropping the intercept out of the linear equation can avoid the meaningless negative precipitation values and has a negligible influence on the effect of reducing the bias. The correction equations were fitted using the neighbouring observed and ensemble mean precipitation amounts during the training period, all using the cubic root-transformed values. The above procedure is repeated for 365 days of the year to form 365 bias-correction equations (for simplicity, February 29 shares the similar correction equation with February 28). The fitted correction equations are then used to correct all ensemble members.

3.5.2 | Temperature

The linear correction equation with the form $y = ax + b$ (where a and b are two regression co-efficients) is used to correct air temperature forecasts. The correction equations are fitted using the neighbouring 15 day observed and forecasted air temperature anomalies. The air temperature anomalies are obtained by subtracting the long-term mean observed air temperature. Similar to precipitation, the 365 well-calibrated equations are used for all ensemble members.

3.6 | Verification metrics

The rank histogram is first used to evaluate the calibration performance of ensemble forecasts. Calibration refers to the statistical consistency between the ensemble forecasts and the observations (Gneiting *et al.*, 2008). An asymmetric rank histogram indicates the consistent bias in the ensemble forecasts. The concave (convex) rank histogram suggests that the ensemble forecasts are under-dispersive (over-dispersive). However, a flat rank histogram is not sufficient to guarantee the reliability of the ensemble, and it only measures whether the observed probability distribution is well represented by the ensemble (Hamill, 2001). It was found that the rank histogram is strongly influenced by many factors, including the variance within each ensemble member, the correlation between ensemble members, and the correlation between observations and forecasts (Marzban *et al.*, 2011; Wilks, 2011). The reliability index (Δ) is used to quantify the deviation from uniformity in a rank histogram, which is defined by:

$$\Delta = \sum_{k=1}^{K+1} \left| P_k - \frac{1}{K+1} \right| \tag{8}$$

where P_k is the observed relative frequency of rank k.

Two verification metrics from the ensemble verification system (EVS) by Brown *et al.* (2010) are used to evaluate the ensemble forecasts, including the deterministic metric of the mean absolute error (MAE), and the probabilistic metric of the continuous ranked probability skill score (CRPSS). The MAE measures the difference between the ensemble mean forecasts and the observations, and a small MAE close to zero is preferred. The CRPSS measures the performance of the ensemble weather forecasts relative to climatology (the mean observations) in terms of the continuous ranked probability score (CRPS), where CRPS is the mean-squared difference between the distribution of ensemble forecasts and corresponding distributions of observations. The CRPSS is positively oriented, with a value of 1 being perfect.

Fractional improvement (FR) proposed by Hagedorn *et al.* (2008) is a metric to measure the FR of the BC for the post-processing method:

$$FR=\frac{CRPSS_{BC}-CRPSS_{raw}}{CRPSS_{pp}-CRPSS_{raw}} \quad (9)$$

where $\mathrm{CRPSS}_{\mathrm{raw}}$, $\mathrm{CRPSS}_{\mathrm{BC}}$ and $\mathrm{CRPSS}_{\mathrm{pp}}$ denote the CRPSS of the raw, bias-corrected and calibrated (by the GPP, BMA, AKD and ExLR) forecasts, respectively. A larger FR indicates a smaller improvement in probabilistic performance compared with the reference BC method. A FR > 1 denotes that the post-processing method is inferior to the BC; a FR < 0 denotes that the raw ensemble forecasts are not improved by the BC. A total of five classes are thus divided based on the FR, including < 0.0, 0.0–0.3, 0.3–0.7, 0.7–1.0 and > 1.0.

To understand better the regional differences of the ensemble weather forecasts before and after post-processing, it is necessary to adopt a specific cluster method in order to classify these stations into different groups. The key to using the cluster analysis is to choose a similarity measure. For example, Lerch and Baran (2017) found using the distribution of forecast errors as the similarity measure augments the training data, which helps to improve the predictive performance of the post-processing methods. Diaz *et al*. (2019) pointed out that the distance measure should include the station climatology and ensemble forecast errors. Therefore, when using the *K*-means method for classifying these stations, the similarity between two stations is defined by the forecast performance (including calibration metric Δ, forecast error metric MAE and forecast skill metric CRPSS) and climate division. The cluster analysis was based on the *K*-means package in MATLAB, and the Silhouette value is used to evaluate the performance of clustering. Three steps are involved in using the *K*-means method:

1. The number of clusters is predetermined using one fixed *K*-means settings. Specifically, the cluster numbers from two to eight are evaluated based on the default *K*-means settings provided in the MATLAB package. The cluster number with the highest Silhouette value is chosen.
2. The various *K*-means settings are evaluated for the cluster number determined in Step (1). Specifically, the study tested different distance metrics (e.g. squared Euclidean distance, sum of absolute difference), different ways to obtain the initial cluster centroid, different *K*-means clustering algorithms, and so on. If the above factors contribute to improving the clustering performance measured in the Silhouette value, the *K*-means algorithm settings will be updated.
3. The updated *K*-means algorithm settings will be used to replace the default settings in Step (1) and to verify whether the choice of the cluster number is reliable. If the new cluster number using the updated *K*-means settings is the same as the old run, the number of clusters is determined. If not, Steps (2) to (3) will be repeated until the cluster number is unchanged.

3.7 | Description of the experiment

Most studies have found that using a large training data set can consistently improve the performance of the post-processing methods (Hagedorn *et al*., 2008; Hamill *et al*., 2008; Scheuerer and Hamill, 2015). In order to make the best use of the available data, the study chose cross-validation to implement the post-processing methods. Given 30 years of available forecasts and observations, when making forecasts for a particular year, the remaining 29 years were used as training data.

4 | RESULTS

4.1 | Classification of the stations

The study defines the similarity between two stations as the performances of the one lead day GEFS forecasts measured by the Δ, MAE and CRPSS and the climate region, and tests the cluster numbers from two to eight to determine the proper cluster number. A total of four clusters are identified for precipitation and six clusters for temperature (Figure 1 and see Supporting Information Table S1). The above clustering schemes are based on the results of the Silhouette value over different cluster numbers in Figure 2.

The cluster divisions for precipitation have a good match to the existing climate divisions. For example, cluster 1 includes SW and E; cluster 2 is distributed in the northern region, including NE, N and NW; cluster 3 is in the QT region; and cluster 4 is mainly located in S and some QT regions. The MAE for clusters 2 and 3 (1.68 and 1.47, respectively) is better than the MAE for clusters 1 and 4 (4.04 and 5.19, respectively). In terms of the CRPSS, cluster 3 is characterized by a poorer CRPSS (0.13) and a more significant variation (0.53) when compared with other clusters. For air temperature, the cluster divisions are partly consistent with the climate divisions. For example, cluster 1 is in northeastern China, including NE and N; cluster 2 is mainly distributed in Xinjiang province in NW; cluster 3 is mainly spread in QT; cluster 4 is in SE, E and NW; cluster 5 is in S and some QT regions; and cluster 6 is distributed in some areas of QT and SW. Also, cluster 3 owns the worst MAE (7.51) and CRPSS (−0.76), while clusters 1, 4 and 5 have the best MAE and CRPSS.

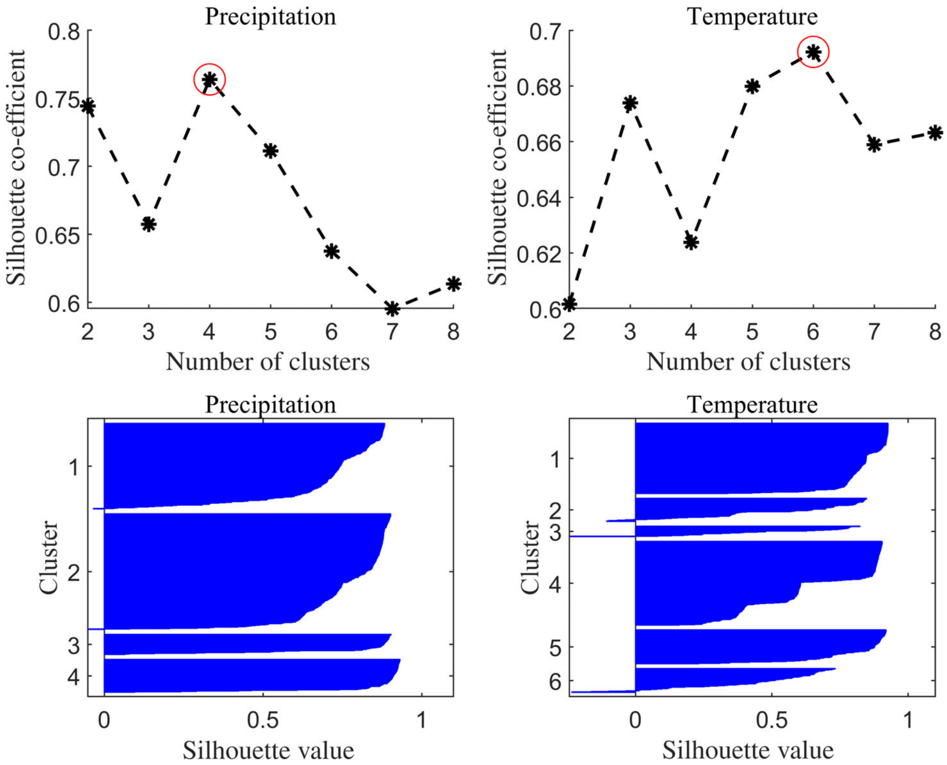


FIGURE 2 (top row) Silhouette co-efficient against the number of clusters from two to eight. Stars with a red circle indicate the optimal cluster number setting. (bottom row) Silhouette plot using the optimal cluster number. The left column is for precipitation, the right column is for air temperature

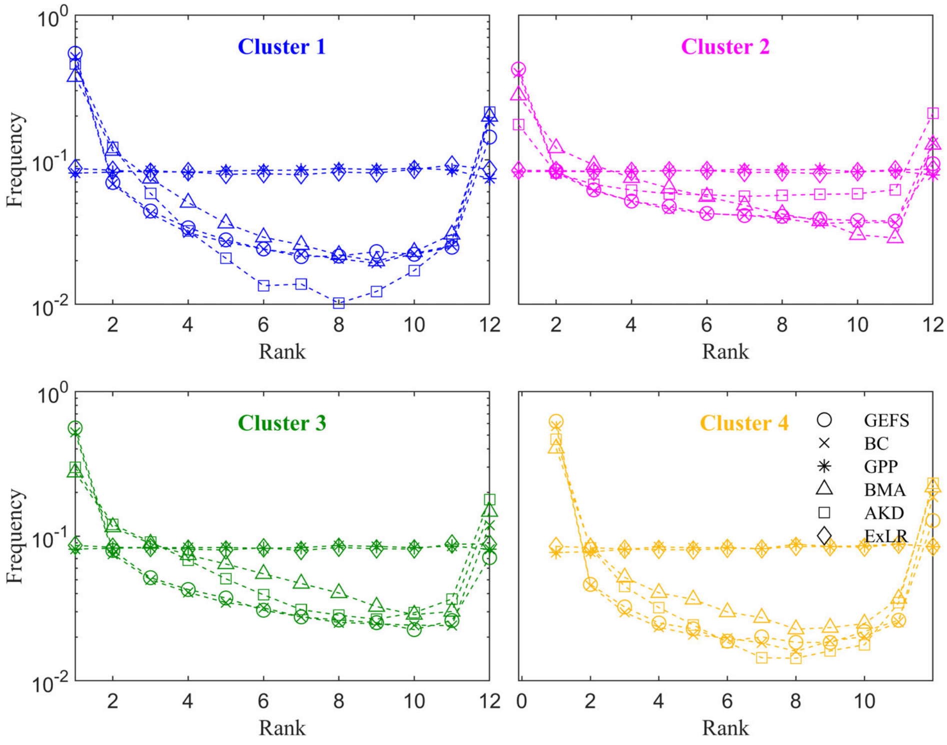


FIGURE 3 Line plot for the rank histogram value of the one lead day precipitation forecasts in four clusters

4.2 | Calibration performance

Figures 3 and 4 plot the verification rank value of the one lead day precipitation and temperature forecasts for the GEFS, BC, GPP, ExLR, BMA and AKD over different clusters. The formation of the rank histogram requires an equal number of ensemble members. The 1,000-member post-processed ensemble forecasts and the 11-member GEFS forecasts and BC-corrected ensemble forecast are thus not matched. Therefore, 11 members are randomly selected from the 1,000 members. For precipitation and air temperature, a considerable number of observations fall outside the range of GEFS ensemble forecasts (see the lowest and highest ranks), forming an apparent

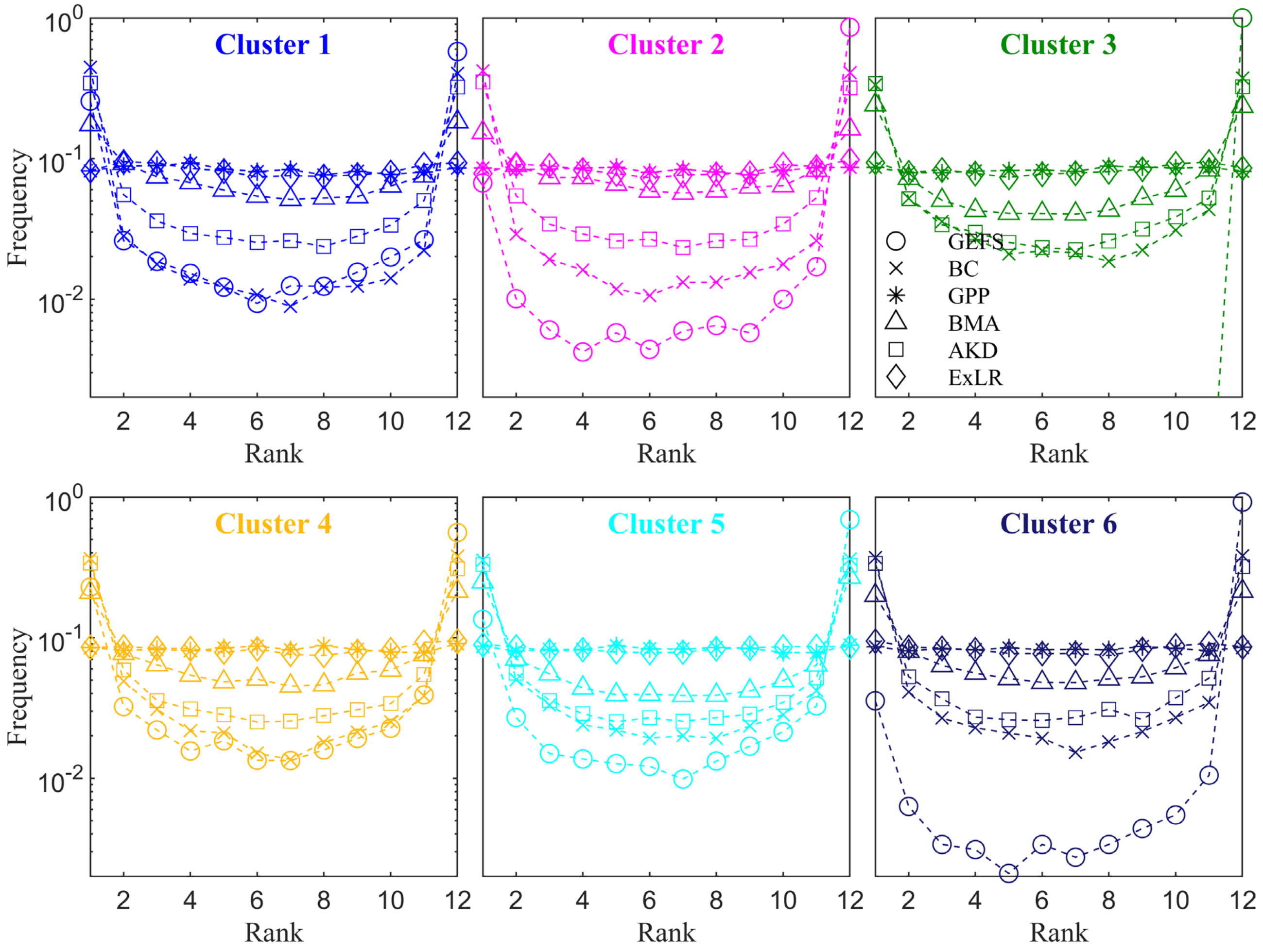


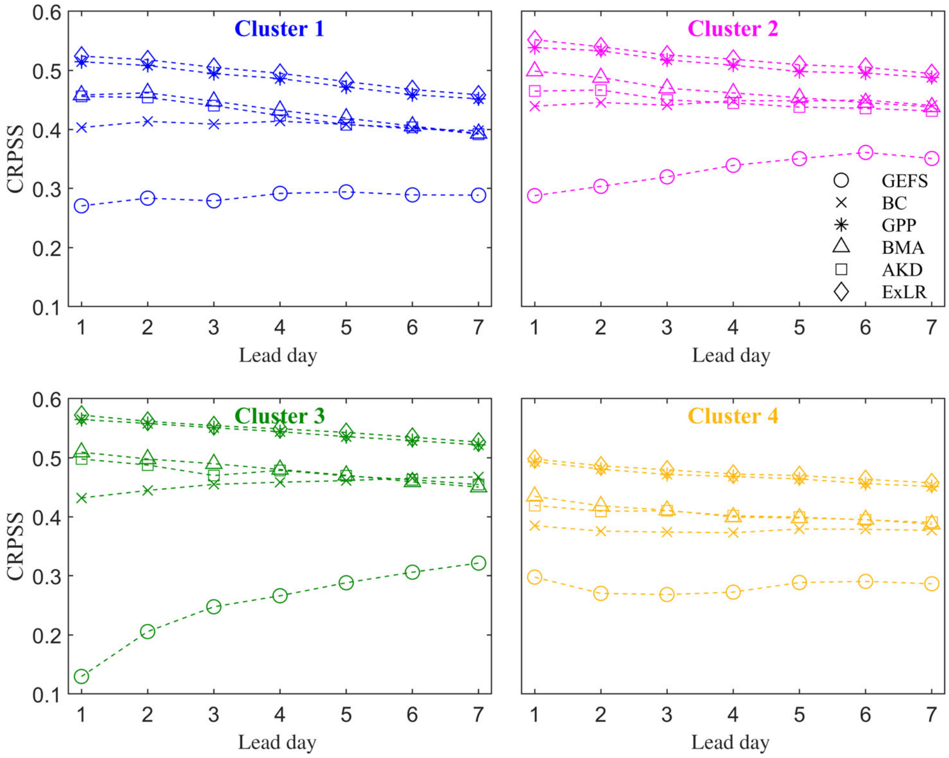
FIGURE 4 Line plot for the rank histogram value of the one lead day air temperature forecasts in six clusters

concave rank histogram, which indicates that the GEFS ensembles are under-dispersive. Also, the rank histogram plot for precipitation in clusters 2 and 3 is inclined to the right, showing the observations are more distributed between the smaller intervals than the larger intervals, which indicates that the GEFS forecasts are biased. When comparing these post-processing methods, the well-calibrated results are achieved by the deterministic methods (GPP and ExLR) by a flat rank histogram plot. While for the BC and probabilistic methods (BMA and AKD) the calibration results show improvement compared with the GEFS results, but the ensemble forecasts are still under-dispersive for both precipitation and temperature and biased for precipitation.

4.3 | Performance evaluation

Figure 5 gives the precipitation results (CRPSS) of different methods over four clusters against the lead time. For detailed results for precipitation, see Supporting Information Tables S2 and S3. The biased and unskilful GEFS forecasts, as expected, have a poor MAE and CRPSS performances, and tend to become even worse for longer lead days. One exceptional condition is that the GEFS seems to become better for longer lead days in all clusters. Similar results were also found in Chen *et al.* (2014, fig. 8). Because the ensemble spread (forecast uncertainty) for shorter lead days is generally smaller than for the longer lead days, the amplified ensemble spread for the longer lead day may contribute to the false improvement of the CRPSS. When comparing with other post-processing methods, the BC achieves a desirable performance in decreasing forecast bias (MAE) and has comparable CRPSS performance with the probabilistic methods (BMA and AKD) after about four lead days. The deterministic methods (GPP and ExLR) share similar performances, both outperforming the probabilistic methods (BMA and AKD) in terms of the CRPSS. When considering different clusters, the most significant improvement of the CRPSS happens in cluster 3, distributed in the QT region. The skill of ensemble forecasts

FIGURE 5 Probabilistic performance (continuous ranked probability skill score—CRPSS) of the precipitation forecasts over different lead times in four clusters



after post-processing is generally higher in the dry northern regions than in the humid southern regions.

Figure 6 presents the CRPSS results for air temperature; for the detailed results, including the MAE and CRPSS, see Supporting Information Tables S4 and S5. The GEFS forecasts are typically biased and unskilful, especially for clusters 2, 3 and 6, where the CRPSS is smaller than the lower limit of the plot. The above clusters are distributed in the western dry and cold regions. The BC is shown to be effective in improving the ensemble forecasts, as shown by a comparable MAE performance and a slightly worse CRPSS performance (mainly for precipitation) compared with the post-processing method. When comparing the post-processing methods, only the GPP, ExLR and BMA tend to outperform the BC for all lead times consistently. The AKD is only useful in less than four lead days when compared with the BC. After post-processing, all the stations except those located in the QT region can achieve a comparable CRPSS performance.

Figure 7 and Supporting Information Tables S6 and S7 show the performance of the one lead day ensemble forecast against the forecast date. Only two clusters from northern and southern China are selected for display. The skill of the GEFS forecasts depends on the date when and region where the forecast is made. Specifically, for the northern region, the GEFS tends to be more skilful in the warm season (April–September) than in the cold season (December–March), while for the southern region, the performance of the GEFS tends to be opposite to the northern region. A simple BC method can well improve the ensemble forecasts for all forecast dates, but the improvement can be further improved when using the post-processing methods. The GPP, ExLR and BMA have a comparable CRPSS performance, and the AKD is found to be consistently lower than the other methods. After post-processing, the skill difference for the forecasts made in different seasons is evident for air temperature, while it is not evident for precipitation.

4.4 | Role of the BC

The above results have shown the additional improvement of using the complicated post-processing methods compared with solely correcting bias using the BC. A further quantitative result for displaying the role of the BC method is shown in Figure 8. The results are based on the one lead day ensemble forecasts for two representative methods: the GPP and BMA. For precipitation, the FR of the BC in the GPP is about 0–70% for most stations, and in the BMA it is about 30–100% for most stations. The GPP and BMA differ in more green points in S and E, and more blue points in the northern and western regions. The results show that the GPP can achieve additional improvement compared with the BMA over two cases: the rainy region and the semi-arid region. For temperature, the FR for the GPP and BMA is about 30–100%

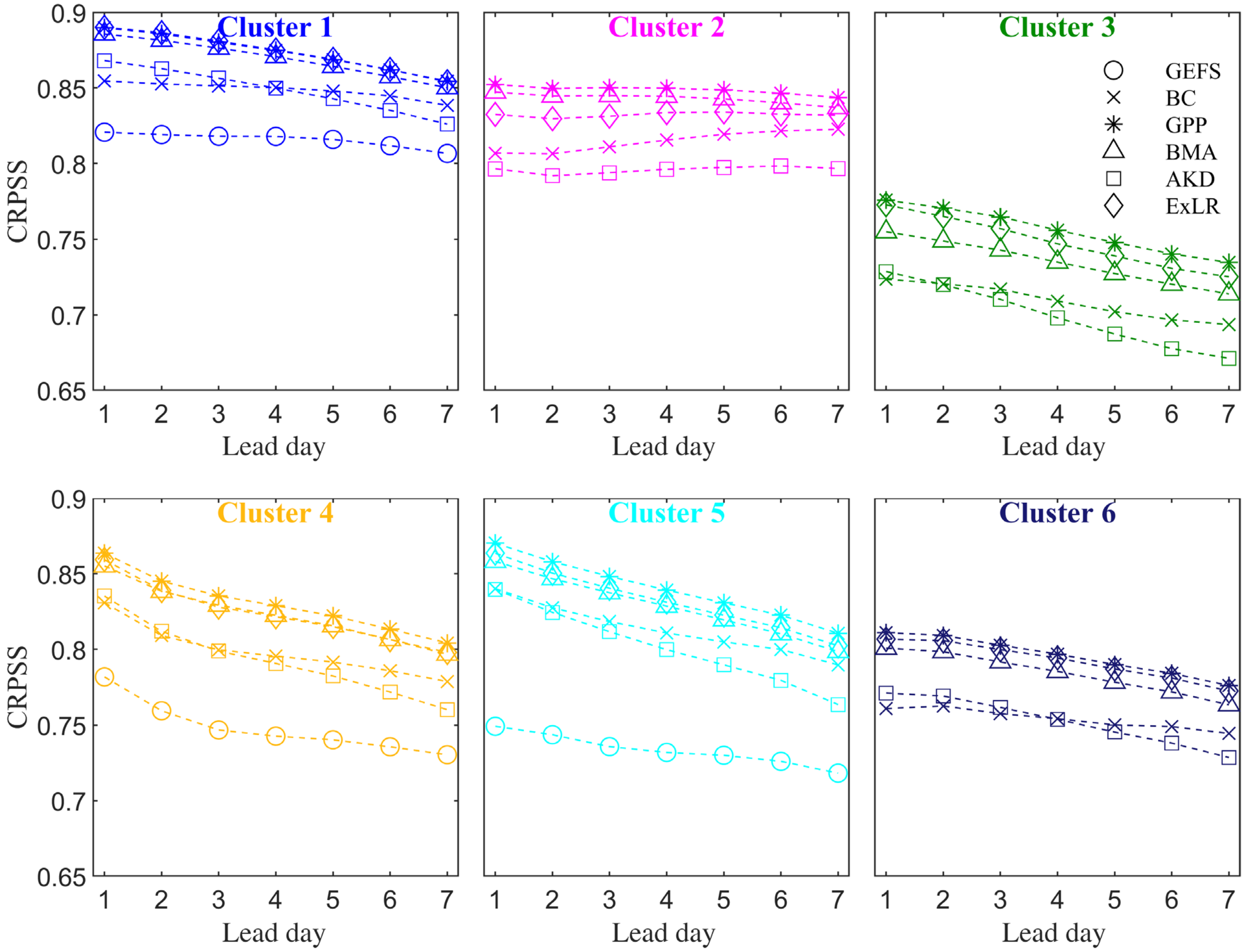


FIGURE 6 Probabilistic performance (continuous ranked probability skill score—CRPSS) of the air temperature forecasts over different lead times in six clusters

for most stations, and the two methods share a small difference. Both methods have a large proportion of red points, distributed in the warm southern region with small yearly fluctuations and cold western region with rapid yearly fluctuations. The results show that for this region, only correcting the bias is enough to achieve the comparable CRPSS skill compared with the post-processing methods.

5 | DISCUSSION AND CONCLUSIONS

The biased and under-dispersive ensemble forecasts cannot be directly used unless a specific post-processing method is applied. Various post-processing methods have thus been proposed in the last two decades. However, the choice of post-processing method is usually based on the user's preferences in practical applications. The previous comparison studies provided some properties about these methods, but there is a lack of guidance about how to choose and use the post-processing methods in practice. Therefore, the study included two variables: precipitation and air temperature, and focused on the probabilistic properties of the ensemble weather forecasts. With this focus, the study evaluated four post-processing methods in order to draw some guidance about the methods' abilities and influencing factors.

For a better comparison of the post-processing methods for the 659 national standard meteorological stations distributed over a large area, the study chose the *K*-means algorithm to classify the 659 stations into several groups. The clustering results using *K*-means are more informative than when using climate divisions. It process provides useful information about using the raw ensemble weather forecasts and selecting the proper post-processing methods for the Global Ensemble Forecasting System (GEFS) forecasts users in China.

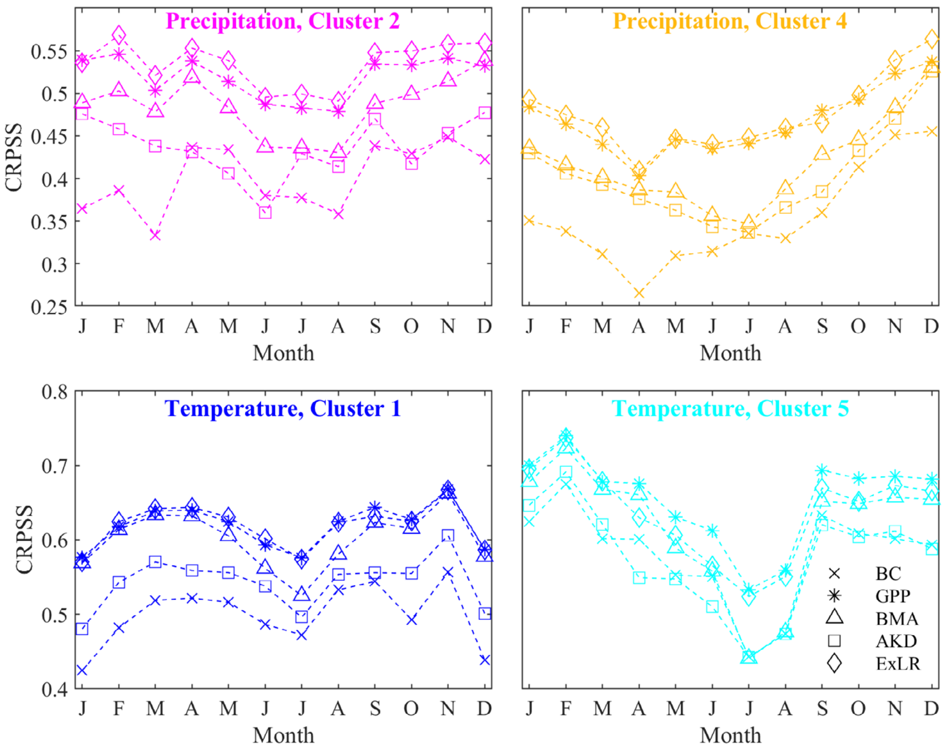


FIGURE 7 Probabilistic performance (continuous ranked probability skill score—CRPSS) of the precipitation forecasts (top row) and air temperature forecasts (bottom row) in different months. Only clusters 2 and 4 are chosen for precipitation, clusters 1 and 5 are chosen for air temperature

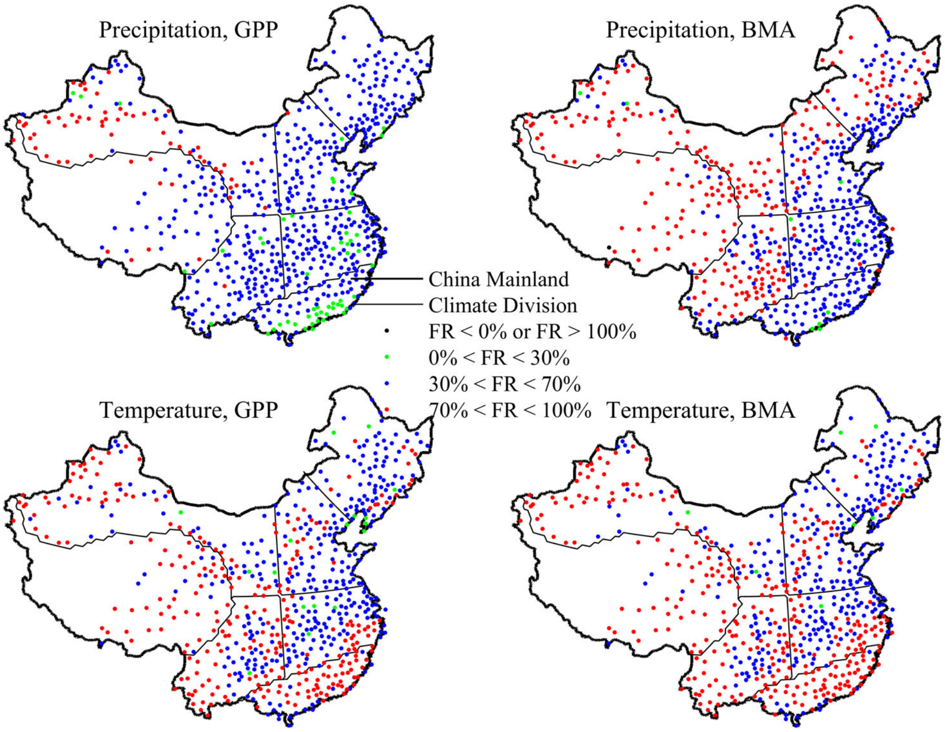


FIGURE 8 Fractional improvement (FR) of one lead day precipitation forecasts (top tow) and air temperature forecasts (bottom row) post-processed by generator-based post-processing (GPP) (left column) and Bayesian model averaging (BMA) (right column)

5.1 | Methods comparison

The study found that the deterministic methods—generator-based post-processing (GPP) and extended logistic regression (ExLR)—are consistently competitive in obtaining the well-calibrated and skilful post-processed ensemble forecasts compared with the probabilistic methods—Bayesian model averaging (BMA) and affine kernel dressing (AKD). For the deterministic methods, the ensemble spread is directly optimized from the historical observations and does not need to calibrate the relationship between the variance in the probability distribution function (PDF) and the ensemble spread. Therefore, the generated ensemble forecasts using the deterministic methods are guaranteed to share similar statistical properties with the observations, resulting in a better probabilistic performance. The AKD has nearly no contribution to improving the forecast skill after certain lead days when compared with bias correction (BC). For the AKD, the ensemble is transformed into a set of kernel distributions (Normal distribution) of the same size. If the number of well-estimated kernel distributions is adequate, the AKD can be used to simulate the forecast distribution of any kind. Therefore, the effectiveness of the AKD strongly relies on the number of kernels or the ensemble size.

5.2 | Influencing factors

When applying the post-processing methods, the influence of region/forecast date and the role of the BC should be considered. Previous studies have realized this issue when using the post-processing method. For example, Hagedorn *et al*. (2008) found that for air temperature, the probabilistic performance could be improved mainly in regions with complex terrain, where the forecast skill was initially lower. About 60–80% of the improvement from non-Gaussian regression (NGR) could be achieved by the simple BC method.

A thorough investigation is made on the influence of these factors in different post-processing methods. First, in terms of region, it is also found that the forecast skill of the post-processed ensemble weather forecasts is comparably high in the northern arid areas for precipitation. Moreover, the forecast skill for air temperature is only low in the Qinghai-Tibetan Plateau area. In terms of forecast date, the GEFS forecasts are more skilful when made in warm seasons (April–September) than in cold seasons (December–March) for northern regions. For southern regions the pattern is the opposite. After post-processing, the skill difference is only evident for air temperature, while it is not evident for precipitation. Besides, in regions with more precipitation events, the GPP can remarkably increase probabilistic performance, while the improvement of other post-processing methods is not apparent.

In terms of the role of the BC, a simple BC method can achieve about 0–70% (for precipitation) and 30–100% (for air temperature) forecast skill improvement of the best-performed deterministic methods.

5.3 | Research prospects

There is an inherent autocorrelation structure of the time series and a dependence structure among different climate variables. For example, heavy rainfall that occurred on a previous day is likely to continue on the following day, and the temperature is generally cooler on wet days. Reproducing the autocorrelation structure of the time series and the dependence structure among different variables may have an essential influence on the impact studies. For instance, temperature and precipitation together determine the process of generating runoff. Therefore, one possible problem of using the probabilistic methods to generate ensemble forecasts by randomly sampling from the built PDF is that the generated time series may lack autocorrelation. However, the study is mainly concerned with the statistical performance of these methods over a long period. It may be interesting to consider autocorrelation and dependence structures in future studies.

ACKNOWLEDGMENTS

This study was partially supported by the National Natural Science Foundation of China (grant numbers 51779176, 51539009), the Thousand Youth Talents Plan from the Organization Department of the CCP Central Committee (Wuhan University, China), and the Overseas Expertise Introduction Project for Discipline Innovation (111 Project) funded by the Ministry of Education and State Administration of Foreign Experts Affairs P. R. China (grant number B18037). The authors acknowledge the GEFS Reforecast Version 2 (https://www.esrl.noaa.gov/psd/forecasts/reforecast2/), Boulder, Colorado, USA, for providing GEFS ensemble precipitation and temperature reforecasts. They also thank the China Meteorological Data Sharing Service System for providing the meteorological data throughout China. The authors declare that they have no conflict of interest.

ORCID

Jie Chen https://orcid.org/0000-0001-8260-3160

REFERENCES

Atger, F. (1999) The skill of ensemble prediction systems. *Monthly Weather Review*, 127, 1941–1953. http://doi.org/10.1175/1520-0493(1999)127<1941:TSOEPS>2.0.CO;2.

Atger, F. (2003) Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Monthly Weather Review*, 131, 1509–1523. http://doi.org/10.1175//1520-0493(2003)131<1509:saivot>2.0.co;2.

Baran, S. and Nemoda, D. (2016) Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27, 280–292. http://doi.org/10.1002/env.2391.

Bauer, P., Thorpe, A. and Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55. http://doi.org/10.1038/nature14956.

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

Brown, J.D., Demargne, J., Seo, D.-J. and Liu, Y. (2010) The ensemble verification system (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environmental Modelling and Software*, 25, 854–872. http://doi.org/10.1016/j.envsoft.2010.01.009.

Chen, J., Brissette, F.P. and Li, Z. (2014) Postprocessing of ensemble weather forecasts using a stochastic weather generator. *Monthly Weather Review*, 142, 1106–1124. http://doi.org/10.1175/MWR-D-13-00180.1.

Diaz, M., Nicolis, O., Marin, J.C., et al. (2019) Statistical post-processing of ensemble forecasts of temperature in Santiago de Chile. *Meteorological Applications*, 27, e1808.

Eckel, F.A. and Walters, M.K. (1998) Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Weather and Forecasting*, 13, 1132–1147. http://doi.org/10.1175/1520-0434(1998)013<1132:cpqpfb>2.0.co;2.

Erickson, M.J., Colle, B.A. and Charney, J.J. (2012) Impact of bias-correction type and conditional training on Bayesian model averaging over the Northeast United States. *Weather and Forecasting*, 27, 1449–1469. http://doi.org/10.1175/WAF-D-11-00149.1.

Gneiting, T. and Raftery, A.E. (2005) Atmospheric science: weather forecasting with ensemble methods. *Science*, 310, 248–249. http://doi.org/10.1126/science.1115255.

Gneiting, T., Raftery, A.E., Westveld, A.H., III and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118. http://doi.org/10.1175/MWR2904.1.

Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L. and Johnson, N. A. (2008) Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17, 211–235. http://doi.org/10.1007/s11749-008-0114-x.

Hagedorn, R., Hamill, T.M. and Whitaker, J.S. (2008) Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: two-meter temperatures. *Monthly Weather Review*, 136, 2608–2619. http://doi.org/10.1175/2007MWR2410.1.

Hamill, T.M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 550–560. http://doi.org/10.1175/1520-0493(2001)129<0550:iorhfv>2.0.co;2.

Hamill, T.M. and Colucci, S.J. (1997) Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125, 1312–1327. http://doi.org/10.1175/1520-0493(1997)125<1312:voersr>2.0.co;2.

Hamill, T.M. and Colucci, S.J. (1998) Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, 126, 711–724. http://doi.org/10.1175/1520-0493(1998)126<0711:eoerep>2.0.co;2.

Hamill, T.M., Whitaker, J.S. and Wei, X. (2004) Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, 132, 1434–1447. http://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

Hamill, T.M., Hagedorn, R. and Whitaker, J.S. (2008) Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Monthly Weather Review*, 136, 2620–2632. http://doi.org/10.1175/2007MWR2411.1.

Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau, T.J., Jr., Zhu, Y. and Lapenta, W. (2013) NOAA's second-generation global medium-range ensemble reforecast data set. *Bulletin of the American Meteorological Society*, 94, 1553–1565. http://doi.org/10.1175/BAMS-D-12-00014.1.

Hodyss, D., Satterfield, E., McLay, J., Hamill, T.M. and Scheuerer, M. (2016) Inaccuracies with multimodal post-processing methods involving weighted, regression-corrected forecasts. *Monthly Weather Review*, 144, 1649–1668. http://doi.org/10.1175/MWR-D-15-0204.1.

Lerch, S. and Baran, S. (2017) Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, 66, 29–51.

Leutbecher, M. and Palmer, T.N. (2008) Ensemble forecasting. *Journal of Computational Physics*, 227, 3515–3539. http://doi.org/10.1016/j.jcp.2007.02.014.

Liu, X. and Coulibaly, P. (2011) Downscaling ensemble weather predictions for improved week-2 hydrologic forecasting. *Journal of Hydrometeorology*, 12(6), 1564–1580. http://doi.org/10.1175/2011JHM1366.1.

Marzban, C., Wang, R., Kong, F. and Leyton, S. (2011) On the effect of correlations on rank histograms: reliability of temperature and wind speed forecasts from finescale ensemble reforecasts. *Monthly Weather Review*, 139, 295–310. http://doi.org/10.1175/2010MWR3129.1.

Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174. http://doi.org/10.1175/MWR2906.1.

Roulin, E. and Vannitsem, S. (2012) Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly Weather Review*, 140, 874–888. http://doi.org/10.1175/MWR-D-11-00062.1.

Roulston, M.S. and Smith, L.A. (2003) Combining dynamical and statistical ensembles. *Tellus A*, 55, 16–30. http://doi.org/10.1034/j.1600-0870.2003.201378.x.

Scheuerer, M. and Hamill, T. (2015) Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143, 4578–4596. http://doi.org/10.1175/MWR-D-15-0061.1.

Schmeits, M.J. and Kok, K.J. (2010) A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Monthly Weather Review*, 138, 4199–4211. http://doi.org/10.1175/2010MWR3285.1.

Sloughter, J.M.L., Raftery, A.E., Gneiting, T. and Fraley, C. (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, 135, 3209–3220. http://doi.org/10.1175/MWR3441.1.

Taillardat, M., Mestre, O., Zamo, M. and Naveau, P. (2016) Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144, 2375–2393. http://doi.org/10.1175/MWR-D-15-0260.1.

Vannitsem, S.P. and Hagedorn, R. (2011) Ensemble forecast post-processing over Belgium: comparison of deterministic-like and ensemble regression methods. *Meteorological Applications*, 18, 94–104. http://doi.org/10.1002/met.217.

Wang, S. and Li, W. (2007) *Climate of China*. Beijing: China Meteorological Press.

Wilks, D.S. (2006) Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, 13, 243–214. http://doi.org/10.1017/S1350482706002192.

Wilks, D.S. (2009) Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, 16, 361–368. http://doi.org/10.1017/S1350482706002192.

Wilks, D.S. (2011) On the reliability of the rank histogram. *Monthly Weather Review*, 139, 311–316. http://doi.org/10.1175/2010MWR3446.1.

Wilks, D.S. and Hamill, T.M. (2007) Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, 135, 2379–2390. http://doi.org/10.1175/MWR3402.1.

Zhu, Y. (2005) Ensemble forecast: a new approach to uncertainty and predictability. *Advances in Atmospheric Sciences*, 22, 781–788. http://doi.org/10.1007/BF02918678.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne, K. (2002) The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society*, 83, 73–83. http://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Li X, Chen J, Xu C-Y, Chen H, Guo S. Intercomparison of multiple statistical methods in post-processing ensemble precipitation and temperature forecasts. *Meteorol Appl*. 2020;27:e1935. https://doi.org/10.1002/met.1935