Deterministic and probabilistic evaluation of raw and post-processing monthly precipitation forecasts: a case study of China

Yujie Li, Bin Xu, Dong Wang, QJ Wang, Xiongwei Zheng, Jiliang Xu, Fen Zhou, Huaping Huang and Yueping Xu

ABSTRACT

Monthly Precipitation Forecasts (MPF) play a critical role in drought monitoring, hydrological forecasting and water resources management. In this study, we applied two advanced Machine Learning Models (MLM) and latest General Circulation Models (GCM) to generate deterministic MPFs with a resolution of 0.5° across China. Then the Bayesian Joint Probability (BJP) modeling approach is employed to calibrate and generate corresponding ensemble MPFs. Raw and post-processing MPFs were put against gridded observations over the period of 1981–2015. The results indicated that: (1) for deterministic evaluation, the forecasting performance of MLMs was more inclined to generate random forecasts around the mean value, while the GCMs could reflect the increasing or decreasing trend of precipitation to some degree; (2) for probabilistic evaluation, the four BJP calibrated ensemble MPFs were unbiased and reliable. Compared to climatology, reliability and sharpness were all significantly improved. However, in terms of overall accuracy metric, the ensemble MPFs generated from MLMs were similar to climatology. In contrast, the ensemble MPFs generated from GCMs achieved better forecasting skill and were not dependent on forecasting regions and months. Moreover, the post-processing method is necessary to achieve not only bias-free but also reliable as well as skillful ensemble MPFs.

Key words | Bayesian joint probability, general circulation model, machine learning model, monthly precipitation forecast, post-processing

HIGHLIGHTS

- Two advanced Machine Learning Models are employed to generate monthly precipitation forecasts with a resolution of 0.5 degree.
- The latest seasonal forecasts of ECMWF are evaluated with the same forecasting grid cells and lead time.
- The BJP modeling approach is used to calibrate above four raw forecasts.
- A comprehensive comparison is achieved for the raw and post-processing forecasts.

Yujie Li
Yueping Xu (corresponding author)
Institute of Hydrology and Water Resources,
Zhejiang University,
Hangzhou 310058,
China
E-mail: *yuepingxu@zju.edu.cn*

Yujie Li
Zhejiang Design Institute of Water Conservancy and Hydroelectric Power,
Hangzhou, Zhejiang 310002,
China

Bin Xu
Hangzhou Design Institute of Water Conservancy and Hydropower,
Hangzhou, Zhejiang 310016,
China

Dong Wang
Bureau of Hydrology,
Changjiang Water Resources Commission,
Wuhan 430010,
China

QJ Wang
Department of Infrastructure Engineering,
the University of Melbourne,
Melbourne, VIC 3010,
Australia

Xiongwei Zheng
Jiliang Xu
Fen Zhou
Zhejiang Design Institute of Water Conservancy and Hydroelectric Power,
Hangzhou, Zhejiang 310002,
China

Huaping Huang
China Water Resources Pearl River Planning Surveying & Designing Co.,Ltd,
Guangzhou, Guangdong 510610,
China

doi: 10.2166/hydro.2021.176

INTRODUCTION

Monthly Precipitation Forecasts (MPF) play a critical role in drought monitoring, hydrological forecasting and water resources management (Schepen *et al.* 2012; Yuan & Wood 2012; Wang *et al.* 2017). With the intensification of climate change in the low and middle latitudes of the Northern Hemisphere, the frequency of extreme rainfall and extreme drought in China is continuously increasing, which further enhances the requirements for accurate and reliable MPF (Peng *et al.* 2014b; Schepen *et al.* 2018; Wang *et al.* 2019a). Theoretically, recent methods to generate MPF can be broadly divided into two main approaches: Machine Learning Model (MLM) and General Circulation Model (GCM) (Yuan *et al.* 2016; Zhao *et al.* 2016; Shen 2018).

For the first approach, MLM implements the MPF by exploring the relationship between past precipitation and climate indices (Bazile *et al.* 2017). Operationally, MLM mainly consists of two components, climate predictors and a regression algorithm (Peng *et al.* 2014a). Here, climate predictors mean a number of large-scale climate indices, such as sea surface temperature, oscillation index, etc. In order to facilitate the research, the China Meteorological Administration (CMA) has established a climate predictors dataset including 130 types of climate indices (referred to hereafter as CI, see Appendix) and has been widely used to drive the regression algorithm (He *et al.* 2018). The regression algorithm is essentially a black box system and is capable of estimating the MPF without fully understanding the effects of climate change and human activities on the precipitation-streamflow-evaporation circle in a certain catchment. Although the regression algorithm contains a large number of different mathematical forms, it can still be roughly divided into the following categories (applied in the hydrological community): (1) Linear regression with regularization, such as Lasso (Jeon *et al.* 2016), Ridge (Jeong *et al.* 2012), and Elastic Network (Park & Mazer 2018); (2) Support Vector Regression with different kernel functions (Chen *et al.* 2010; Liang *et al.* 2018); (3) Deep Learning, which includes but is not limited to Convolutional Neural Network (Qiu *et al.* 2017), Long Short-Term Memory (Zhang *et al.* 2018), and Deep Belief Network (Bai *et al.* 2016); (4) Bagging strategy, which is famous for Random Forest (Wang *et al.* 2015); (5) Boosting strategy, in the forms of Adaptive Boosting (Liu *et al.* 2014), Gradient Boosting Decision Tree (Ma *et al.* 2018), and eXtreme Gradient Boosting (Fan *et al.* 2018). Among the above regression algorithms, the Boosting strategy is greatly developed and widely used in MPF due to the characteristics of solid mathematical theory, low computational cost, and easy implementation.

For the second approach, GCM as a category of climate models can directly generate precipitation forecasts with different lead times. GCM generally consists of numerous and complicated atmosphere, ocean, land components and corresponding dynamic exchange interfaces (Street 2016; Zhao *et al.* 2017). With the constant improvements of model configuration (e.g. structure and resolution), model initialization (e.g. operational analysis system), and ensemble generation (e.g. initial condition perturbations and stochastic perturbations), the GCM approach has become the main tool to implement MPF by most authoritative meteorological research centers (Molteni *et al.* 2011; Yuan *et al.* 2011; Schepen *et al.* 2014; Johnson *et al.* 2019), and has shown advantages in extending the lead time, improving the resolution, and quantifying the uncertainty. For instance, considering the interactions between the atmosphere and ocean allows GCM to simulate long-term phenomena such as phases of the El Niño–Southern Oscillation cycle which is a significant large-scale climate predictor deeply affecting the precipitation in the middle latitude of the Northern Hemisphere (Cao *et al.* 2014). Moreover, National Centers for Environmental Prediction (NCEP) of USA has transitioned to operationally use the Climate Forecast System version 2 (CFSv2) since 2011, to provide real-time forecasts with 24 ensemble members, maximum lead time of nine months, and resolution of 1° (Saha *et al.* 2014; Liu *et al.* 2019). The European Centre for Medium-Range Weather Forecasts (ECMWF) also upgraded the Seasonal Forecast System 5 (SEAS5) in 2017 to replace its predecessor System 4 (SEAS4) which had been operational since 2011 (Johnson *et al.* 2019).

Nonetheless, both MLM and GCM have their own shortcomings that cannot be ignored in practical application. In the matter of MLM, one of the most obvious drawbacks is that the MPF generated is always a deterministic forecast (i.e. single-valued forecast), which is generally considered to be outdated compared with the probabilistic forecast (i.e. ensemble forecast) especially in the sub-seasonal (1–3 months) scale (Bennett *et al.* 2016; Li *et al.* 2019). Duan *et al.* (2019) concluded that the traditional deterministic MPF is inadequate to meet different needs of emergency and water resources managers and that the emerging ensemble forecasting approach is the way forward. One immediate way to obtain ensemble forecasts is to input different scenarios of climate predictors or employ different regression algorithms. However, for a suitable combination of climate predictors and regression algorithms, the forecasts usually show a small variation. As for GCM, although it produces multi-member ensemble forecasts, its incomplete system patterns and unsuitable modeling parametrizations always lead to systematic bias and cannot be used directly by the end-users (Li *et al.* 2017; Shen 2018).

Therefore, in response to the above disadvantages, the post-processing method becomes a necessary step not only to generate an ensemble forecast but also to quantify and reduce the uncertainty. Moreover, the objectives of the post-processing method also contain: (1) correct bias; (2) preserve the raw predictive skills; (3) ensure the ensemble members have a reliable time-space relationship (Clark *et al.* 2004; Li *et al.* 2017). Recently, a Bayesian Joint Probability modeling approach (referred to hereafter as BJP), developed by Wang *et al.* (2009), as a conditional distribution-based statistical post-processing method, has been successfully applied to calibrate the MPF from GCM (Peng *et al.* 2014a; Bennett *et al.* 2016; Zhao *et al.* 2019a). In addition, BJP also has the ability to generate ensemble forecasts based on certain deterministic MPF, which provides an available method to overcome the shortcomings from the MLMs (Zhao *et al.* 2015), and also supports a worthwhile way to compare two MPFs from different sources within the framework of both deterministic and probabilistic forecasting.

In summary, the specific objectives this study aims to realize are as follows: (1) use the climate predictors from CI and two advanced Boosting models of MLMs (namely XGB and LGB) to generate raw MPFs with the resolution of 0.5° (3824 grid cells) covering the majority of the Chinese mainland over the period of 1981–2015; (2) assess the prediction performance of two GCMs from ECMWF (namely SEAS5 and SEAS4) by ensemble means and compare with the above two MLMs in terms of bias, accuracy, and correlation; (3) calibrate the above four raw deterministic MPFs by BJP modeling approach to generate corresponding probabilistic forecasts, and evaluate the four post-processing ensemble MPFs in terms of bias, accuracy, reliability, and sharpness. The paper is organized as follows: the next describes three categories of the dataset (observation, CI, and GCM forecasts) followed by a section providing a brief overview of XGB, LGB, BJP models as well as evaluation measures and metrics. This is followed by the results, discussion and conclusions, respectively.

DATASET

Gridded monthly precipitation observation

The observed precipitation is provided by the CMA, with a resolution of 0.5°, a total of 3824 grid cells, and the period of 1981–2015, covering the majority of the Chinese mainland. The original data is daily and comes from 2472 national meteorological stations that have been uniformly distributed throughout China since 1961. The Thin Plate Spline (TPS) spatial interpolation method is employed by CMA to generate gridded observations from sites and has been verified to represent a desirable accuracy. The specific introduction, establishment and assessment can be found in Zhao *et al.* (2014, 2018). Figure 1 shows the monthly mean precipitation that displays an enormous spatial and temporal variability. This increases the difficulty of accurate forecast. Due to the three-steps-distributed feature of Chinese terrain and altitude, the precipitation generally decreases from southeast to northwest, with several distinguishable boundaries. Besides, since China is mainly characterized by a continental monsoon climate, which often represents a co-occurrence of rain and hot summer, the precipitation from April to September is also significantly higher than that in other months.

130 types of large-scale climate indices

CI is a continuously improved large-scale climate predictor dataset which has been provided by CMA since 1970. CI contains 88 types of atmospheric circulation (e.g. Pacific Subtropical High of area/intensity/position

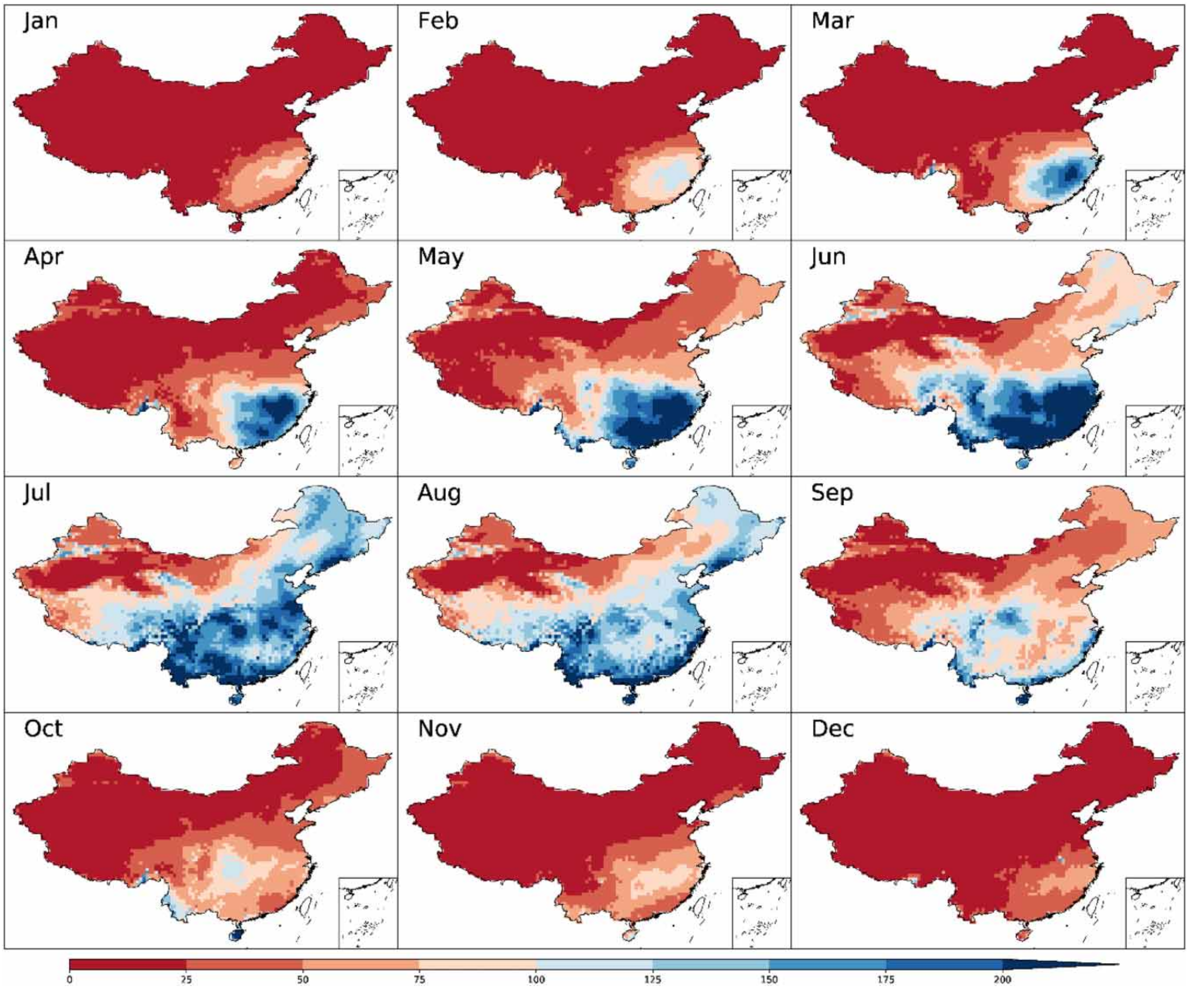

Figure 1 | Monthly mean precipitation (mm) over Chinese mainland during the period 1981–2015.

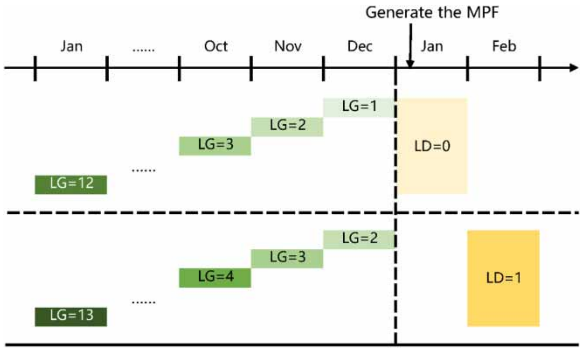

Figure 2 | The timeline of the lead time of predictand and the lag time of predictors.

index), 26 types of sea surface temperature (e.g. Niño anomaly of 1 + 2/3.4), and 16 types of other indices (e.g. landing typhoon, sunspot), which have demonstrated good adaptability and flexibility as predictors to establish monthly, seasonal, even yearly forecasts. The specific description can be found at https://cmdp.ncc-cma.net/Monitoring/cn_index_130.php. Here, we take the forecasting timeline shown in Figure 2 to illustrate the lead time (shown as LD) of predictand and the lag time (shown as LG) of predictors in detail. First, we assume that the lag periods of CI are 12 months, i.e. when the predictand is the forecasts in January, the predictor with the longest duration is in January of the previous year. Then, when we intend to generate MPF of January and February on any day in January (usually at the beginning), the corresponding lead times are 0 and 1, respectively. In this study, we focus on the forecast scenery with LD = 1, i.e. LG from 2 to 13 (as shown in the lower half of Figure 2).

Seasonal forecast system products

ECMWF has been operating the SEAS4 since November 2011 and it was upgraded to the latest version, SEAS5, in November 2017. SEAS5 is a substantially changed and coupled ocean-atmosphere-land dynamical forecast system. Compared with the predecessor (i.e. SEAS4), SEAS5 has improvement in the following main aspects: (1) The atmosphere model is the ECMWF IFS (Integrated Forecast System) from version 36r4 to 43r1, with a higher resolution from 0.8 to 0.36°; (2) The ocean model is the NEMO (Nucleus for European Modelling of the Ocean) from version 3.0 to 3.4, with a higher resolution from 1 to 0.25° as well as the number of vertical levels from 42 to 75 levels; (3) A sea-ice model named LIM2 has been added; (4) Wave model resolution has improved from 1 to 0.5°; (5) The real-time forecasts still comprise 51 ensemble members but the re-forecasts (also known as hindcasts) members have increased from 15 to 25. Details of the comparisons can be found in Johnson *et al.* (2019). In this study, since the precipitation resolutions of SEAS4 and SEAS5 are 0.75 and 0.4° degrees respectively, Bilinear Interpolation, as an extension of linear interpolation and widely used interpolating function on a rectilinear 2D grid, is used to process the calculation of downscaling (for SEAS4) or upscaling (for SEAS5) to coordinate and unify the spatial resolution issue.

METHODOLOGY

To assist the reader through the description of the components of this research, a conceptual representation of the calculation process is shown in Figure 3. First, we obtained the raw deterministic MPFs from the XGB and LGB models, as well as obtaining raw ensemble GCMs from SEAS4 and SEAS5. Then a deterministic evaluation was applied to analyse the quality of the raw MPFs. Second, the BJP model was used to calibrate the raw MPFs which contains both MLMs and GCMs approaches to obtain the calibrated MPFs. Third, a probabilistic evaluation was applied to analyse the quality of the calibrated MPFs.

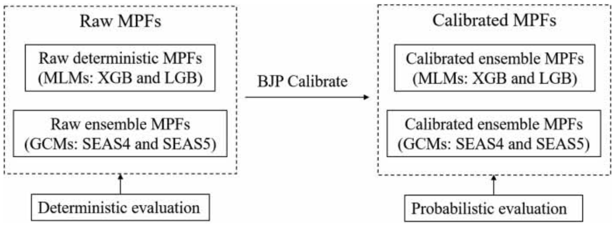
Raw MPFs
Raw deterministic MPFs (MLMs: XGB and LGB)
Raw ensemble MPFs (GCMs: SEAS4 and SEAS5)
BJP Calibrate
Calibrated MPFs
Calibrated ensemble MPFs (MLMs: XGB and LGB)
Calibrated ensemble MPFs (GCMs: SEAS4 and SEAS5)
Deterministic evaluation
Probabilistic evaluation

Figure 3 | The flowchart of the research process.

Extreme gradient boosting

Extreme Gradient Boosting (referred to hereafter as XGB) is presented by Chen & Guestrin (2016) and Chen *et al.* (2015) and is widely used in classification and regression issues as an enhanced implementation of the Gradient Boosting Decision Tree (Friedman 2002). The core idea of XGB is to combine a large number of weak learners (e.g. classification and regression tree) into strong learners through continuously fitting residuals. This is obtained by adding regularization into the loss function that not only minimizes the computational cost but also prevents over-fitting. The loss function of XGB can be defined as:

$$\begin{aligned} L &= \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \\ \Omega(f) &= \gamma J + \frac{1}{2}\lambda\|\omega\|^2 \end{aligned} \tag{1}$$

where l is a differentiable convex loss function which measures the error of observed y_i and simulated $\hat{y}_i$, ω is the fraction of leaf nodes, λ is the regularization parameter, γ is the minimum loss required to further divide the leaf nodes and J is the quantity of the leaf nodes. In addition, XGB adopts the viewpoint of parallel computing in Random Forest (Breiman 2001), which obviously improves the calculation efficiency. The detailed information of the XGB model can be found in Chen & Guestrin (2016) and Chen *et al.* (2015).

Light gradient boosting machine

Light Gradient Boosting Machine (referred to hereafter as LGB) is a fast, distributed and high-performance framework based on GBDT (Ke *et al.* 2017). Compared with XGB, several obvious differences are as follows. First, LGB further studies the accelerated calculation method based on the histogram algorithm. Second, XGB adopts a level-wise growth strategy, while LGB adopts a more efficient leaf-wise growth strategy with depth restriction on leaf growth. Third, LGB proposes Gradient-based One-Side Sampling (GOSS) so that the training is accelerated under the condition that the precision is hardly affected. Fourth, LGB proposes Exclusive Feature Bundling (EFB) for the optimization of column sampling. Detailed information of the LGB model can be found in Ke *et al.* (2017). Theoretically, these differences play an important role in optimizing the computational performance, but there is no research on MPF showing that the forecast skills of LGB will be better or worse than XGB (Ukkonen & Mäkelä 2019).

Bayesian joint probability post-processing modeling approach

The BJP post-processing modeling approach is originally presented to generate ensemble streamflow forecasts (Wang *et al.* 2009; Wang & Robertson 2011; Robertson *et al.* 2013) and has been successfully applied to post-process precipitation (Shrestha *et al.* 2015; Bennett *et al.* 2016) and other hydrometeorological predictions (Zhao *et al.* 2019a, 2019b). The BJP post-processing modeling approach begins with the predictand y (in this case observed precipitation) and predictor x (in this case raw MPF), and then the log-sinh transformation (Wang *et al.* 2012) is used to normalize the variables and homogenize their variances. Mathematically, the log-sinh transformations are given by:

$$\begin{aligned}\hat{x} &= \frac{1}{\beta_x}\ln\left[\sinh\left(\alpha_x + \beta_x x\right)\right]\\ \hat{y} &= \frac{1}{\beta_y}\ln\left[\sinh\left(\alpha_y + \beta_y y\right)\right]\end{aligned} \tag{2}$$

where $\hat{x}$ and $\hat{y}$ are transformed variables of x and y, respectively, and α and β are the transformation parameters. Then, the transformed variables are assumed to follow a bivariate normal distribution:

$$p(\hat{x}, \hat{y}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3}$$

where $\boldsymbol{\mu} = \begin{bmatrix}\mu_{\hat{x}}\\ \mu_{\hat{y}}\end{bmatrix}$ and $\Sigma = \begin{bmatrix}\sigma_{\hat{x}}^2 & \rho\sigma_{\hat{x}}\sigma_{\hat{y}}\\ \rho\sigma_{\hat{x}}\sigma_{\hat{y}} & \sigma_{\hat{y}}^2\end{bmatrix}$. μ and σ are the mean and standard deviation for each predictand and predictor; ρ is the correlation coefficient. Then the forecast function could be defined as:

$$p(\hat{y}|\hat{x}) \sim N\left(\mu_{\hat{y}} + \rho\frac{\sigma_{\hat{y}}}{\sigma_{\hat{x}}}(\hat{x} - \mu_{\hat{x}}),\ (1-\rho^2)\sigma_{\hat{y}}^2\right) \tag{4}$$

From Equation (4) above, it can be concluded that the BJP model can further estimate the uncertainty of prediction while correcting the systematic bias. Here we obtain a set of nine parameters, $\boldsymbol{\theta} = [\alpha_x, \beta_x, \mu_{\hat{x}}, \sigma_{\hat{x}}, \sigma_y, \beta_y, \mu_{\hat{y}}, \sigma_{\hat{y}}, \rho]$. The posterior distribution function of $\boldsymbol{\theta}$ is as follows:

$$\begin{aligned} p = (\theta|\mathbf{y}^N, \mathbf{y}^{N-1}, \ldots \mathbf{y}^1) &\propto P(\boldsymbol{\theta})P(\mathbf{y}^N, \mathbf{y}^{N-1}, \ldots \mathbf{y}^1|\boldsymbol{\theta})\\ &= P(\theta)\prod_{n=1}^{N} p(\mathbf{y}^n|\theta)\end{aligned} \tag{5}$$

where N is the length of training period and $p(\mathbf{y}^N, \mathbf{y}^{N-1}, \ldots \mathbf{y}^1|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are likelihood function and prior distribution of $\boldsymbol{\theta}$, respectively. The Gibbs sampling based on Markov chain Monte Carlo is the core of BJP to establish the Bayesian parameters inference and generate climatological reference distribution (Zhao *et al.* 2019b). Detailed information on the BJP model can be found in Wang *et al.* (2009) and Wang & Robertson (2011).

Evaluation measures and metrics

We employ a leave-one-month-out-cross-validation (lomocv) procedure to evaluate the performance of raw deterministic MPFs, calibrated deterministic MPFs and calibrated ensemble MPFs. Specifically speaking, when we drive XGB and LGB models to generate the gridded MPFs, all available datasets except one month are used to train the parameters and the prediction for the leave-out month is compared to the corresponding observation. Since we have 35 years of observation, the procedure will be repeated 35 times and the corresponding cross-validated predictions will be obtained for 12 months. Similarly, lomocv is also used to infer parameters and generate the ensemble forecast in the BJP post-processing model.

A detailed assessment usually requires several performance metrics to analyze different aspects of forecast quality attributes which normally contain Bias, Accuracy, Correlation, Reliability, Sharpness and Skill.

(1) Bias measures the difference between average forecast and average observation. Here we use the relative bias

(RB) to measure the Bias. The perfect value of RB is 0 and could be defined as:

$$RB(\%) = \frac{\sum_{t=1}^{T} y^t_{fct} - \sum_{t=1}^{T} y^t_{obs}}{\sum_{t=1}^{T} y^T_{obs}} \times 100 \tag{6}$$

where y_{obs} is the observation, y_{fct} is the raw deterministic MPF or the mean of calibrated ensemble MPF, and T is the length of verification period (in this case 35).

(2) Accuracy measures the average distance between forecast and observation. Here we use the well-known Continuous Ranked Probability Score (CRPS) for ensemble MPF (Hersbach 2000). The perfect value of CRPS is 0 and can be defined as:

$$CRPS = \frac{1}{T}\sum_{t=1}^{T} \int [F_{fct}(Y^t_{fct}) - H(Y^t_{fct}) - H(y^t_{fct} - y^t_{obs})]^2 dy^t \tag{7}$$

where $F_{fct}(y^t_{fct})$ is the cumulative distribution function (CDF) of the forecasts and H is the Heaviside step function and is defined as:

$$H(y^t_{fct} - y^t_{obs}) = \begin{cases} 0 & y^t_{fct} < y^t_{obs} \\ 1 & y^t_{fct} \geq y^t_{obs} \end{cases} \tag{8}$$

CRPS corresponds to the Mean Absolute Error (MAE) for deterministic MPF. The perfect value of MAE is 0 and can be defined as:

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |y^t_{fct} - y^t_{obs}| \tag{9}$$

(3) Association reflects the linear relationship forecast and observation. Here we use the Pearson Correlation Coefficient (PCC) which can be expressed as:

$$PCC = \frac{\sum_{t=1}^{T} (y^t_{fct} - \bar{y}_{fct})(y^t_{obs} - \bar{y}_{obs})}{\sqrt{\sum_{t=1}^{T} (y^t_{fct} - \bar{y}_{fct})^2}\sqrt{\sum_{t=1}^{T} (y^t_{obs} - \bar{y}_{obs})^2}} \tag{10}$$

(4) Reliability describes how well the forecast agrees with the observation when a specific forecast is issued and here a probability integral transform (PIT) histogram is employed to assess the reliability of ensemble forecasts. PIT is the CDF of the forecast $F_{fct}(y^t_{fct})$ evaluated at observation y^t_{obs} and is given by:

$$PIT^t = F_{fct}(y^t_{obs}) \tag{11}$$

when the ensemble forecast reliably captures the distribution of observation, the observation y^t_{obs} can statistically be regarded as random samples drawn from $F_{fct}(y^t_{obs})$. Therefore, the reliability of ensemble spread is shown by the uniformity of the PIT histogram (Laio & Tamea 2007; Duan *et al.* 2019). In addition, in order to compare the reliability with climatology, the PIT Area is also used (Renard *et al.* 2010; Schepen *et al.* 2018) and is defined by:

$$PIT\,Area = \frac{2}{T}\sum_{t=1}^{T} \left| PIT^t_* - \frac{t}{T+1} \right| \tag{12}$$

where PIT^t_* is the sorted PIT^t in increasing order. The PIT Area represents the total deviation of PIT^t_* from the corresponding uniform quantile (i.e. the tendency to deviate from the 1:1 line in PIT diagrams). The PIT Area ranges from 0 (perfect reliability) to 1 (worst reliability).

(5) Sharpness describes the concentration of the predicted distribution and indicates the distribution of the ensemble members. Here the 90% interquartile range (IQR) is used to evaluate the sharpness (Crochemore *et al.* 2016) and can be defined as follows:

$$IQR = \frac{1}{T}\sum_{t=1}^{T} (Q^t(95\%) - Q^t(5\%)) \tag{13}$$

where $Q(95\%)$ and $Q(5\%)$ are the 95 and 5% percentiles of the forecast distribution. The final IQR score is the average of the whole interquartile range at a certain period. The narrower the IQR, the sharper the ensemble forecasts.

(6) Skill describes the accuracy of a forecast relative to a reference forecast or benchmark (Duan *et al.* 2019). When the skill score is superior (inferior) to zero, this means that the forecast is more (less) skillful than the reference. When it is equal to zero, the forecasts and the reference have equivalent skill (Crochemore *et al.* 2016). Here CRPS Skill Score (CRPSS), PIT Area Skill Score (PITSS), and IQR Skill Score (IQRSS) are computed for the probabilistic comparison. The three skill scores can be calculated by:

$$\begin{aligned} CRPSS(\%) &= \frac{CRPS_{ref} - CRPS_{fct}}{CRPS_{ref}} \times 100 \\ PITSS(\%) &= \frac{PITArea_{ref} - PITArea_{fct}}{PITArea_{ref}} \times 100 \\ IQRSS(\%) &= \frac{IQR_{ref} - IQR_{fct}}{IQR_{ref}} \times 100 \end{aligned} \quad (14)$$

Here we refer to climatology as the reference and the climatology is calculated by the BJP method (Wang *et al.* 2019b).

RESULTS

Quality of raw deterministic MPFs

Bias in raw deterministic MPFs is analyzed by RB using Figure 4. Each row corresponds to a month and each column corresponds to a model. In general, the MLMs overall outperform the GCMs, representing a smaller RB throughout 12 months. From February to October, almost half of the grid cells are slightly underestimated (shown in light red), and others are slightly overestimated (show in light blue). From November to January, some unideal dark blue pixels occur in the middle and northwest, but with no obvious aggregation. RB in XGB is the smallest overall, varying from –20 to 20%, and is distributed evenly throughout the mainland, slightly better than LGB. In contrast to MLMs, the MPFs from SEAS4 and SEAS5 are badly overestimated (show dark blue) in most month and grid cells, especially during January–April. For April–November, both of them show positive and negative alternation in other regions except for the negative RB of less than –40% in the northwest region. Meanwhile, the distribution of RB always shows an obvious aggregation. These systematic errors are most likely caused by the incomplete simplification of hydrological cycle mechanisms and unsuitable modeling parametrizations.

The accuracy of raw deterministic MPFs is analyzed by MAE in Figure 5, which shows a clear dividing line conforming to the geographic elevation. The distribution of MAE exhibits certain common temporal and spatial distribution characteristics of MAE among four models, particularly for November–April. The MAE values of these months are mainly distributed in three intervals. Most of them vary from 0 to 10 in the northwest; some are 10–30 in the middle, and the rest are less than 50 in the southeast. In terms of May–October, the MAE values are worse than other months, while the worst results are detected in July and August. Some MAE values even exceed 70, which means an unacceptable accuracy. It can be found that the MAE generally increases from northwest to southeast, matching the distribution of precipitation. Meanwhile, it is also difficult to tell the best model in terms of MAE. The performance of accuracy seems to depend on the magnitude of precipitation rather than the difference in models, which confirms the results in Liu *et al.* (2017) and Tian *et al.* (2017).

As mentioned above, the PCC between the raw MPF and the observation is a key parameter used by BJP in the calibration process. For this reason, Figure 6 represents the PCC of four models and only contains the significantly positive correlations at a 90% confidence level (according to a t-test). Contrary to the results in Figures 4 and 5, the GCMs overall outperform the MLMs. The number of PCC grid cells satisfying the above conditions in XGB and LGB are sparse over 12 months and far lower than those in SEAS4 and SEAS5. In addition, the PCC values in MLMs are also detected to be much smaller than in GCMs and mainly vary between 0.3 and 0.6. In contrast, PCC values of SEAS4 and SEAS5 basically cover the mainland except for some northwest grid cells, mainly varying between 0.4 and 0.9. It can be concluded that although MLMs display better performance in term of bias and accuracy, the MPFs tend to be a phenomenon of random prediction due to poor PCC values. Meanwhile, although the RB and MAE of GCM are not ideal, both SEAS4 and

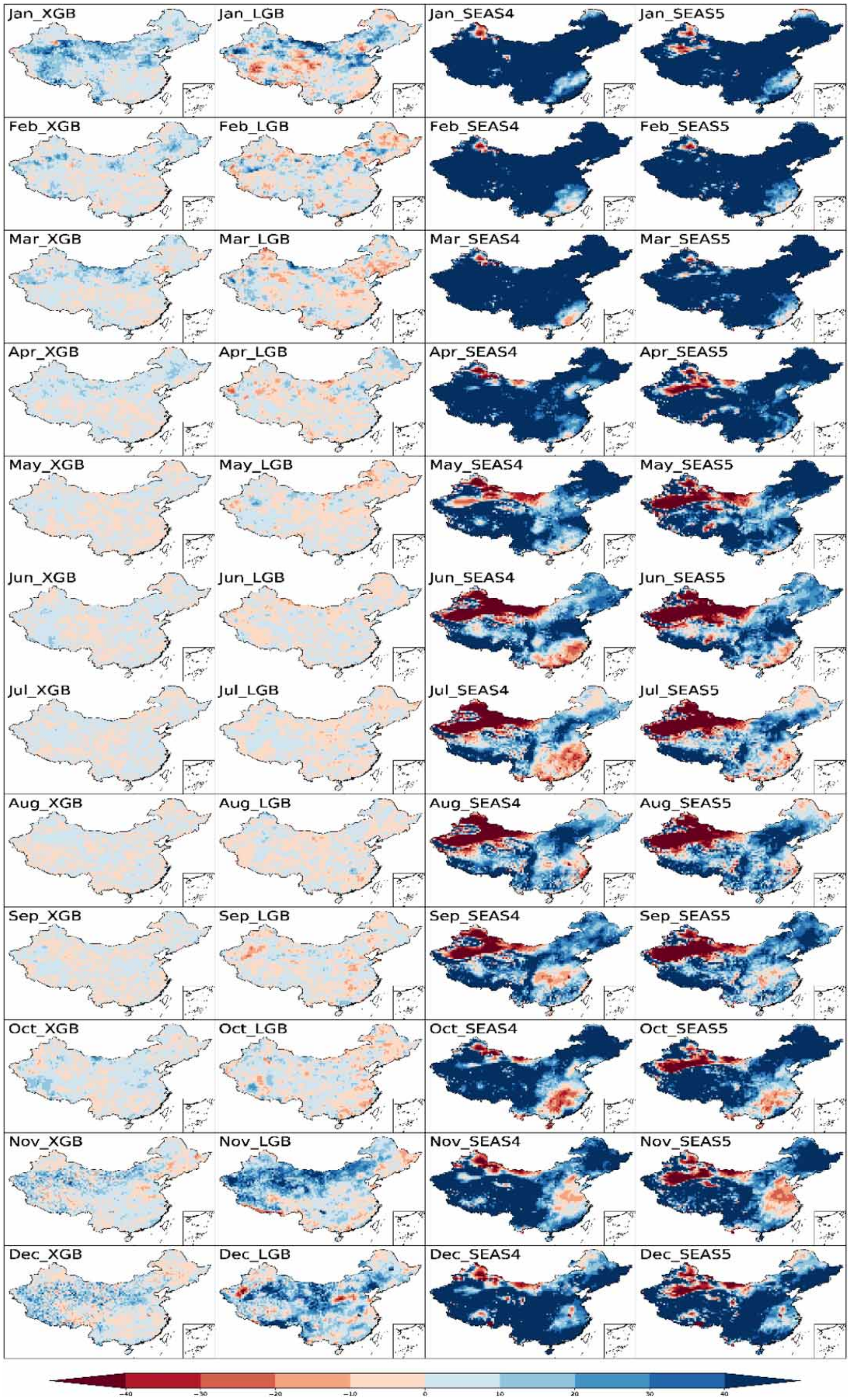

Figure 4 | RB (%) between four raw deterministic MPFs (namely XGB, LGB, the ensemble means of SEAS4 and SEAS5) and observations throughout 12 months for the period of 1981–2015. Blue areas represent a tendency of overpredicting precipitation, and red areas represent a tendency of underpredicting precipitation.

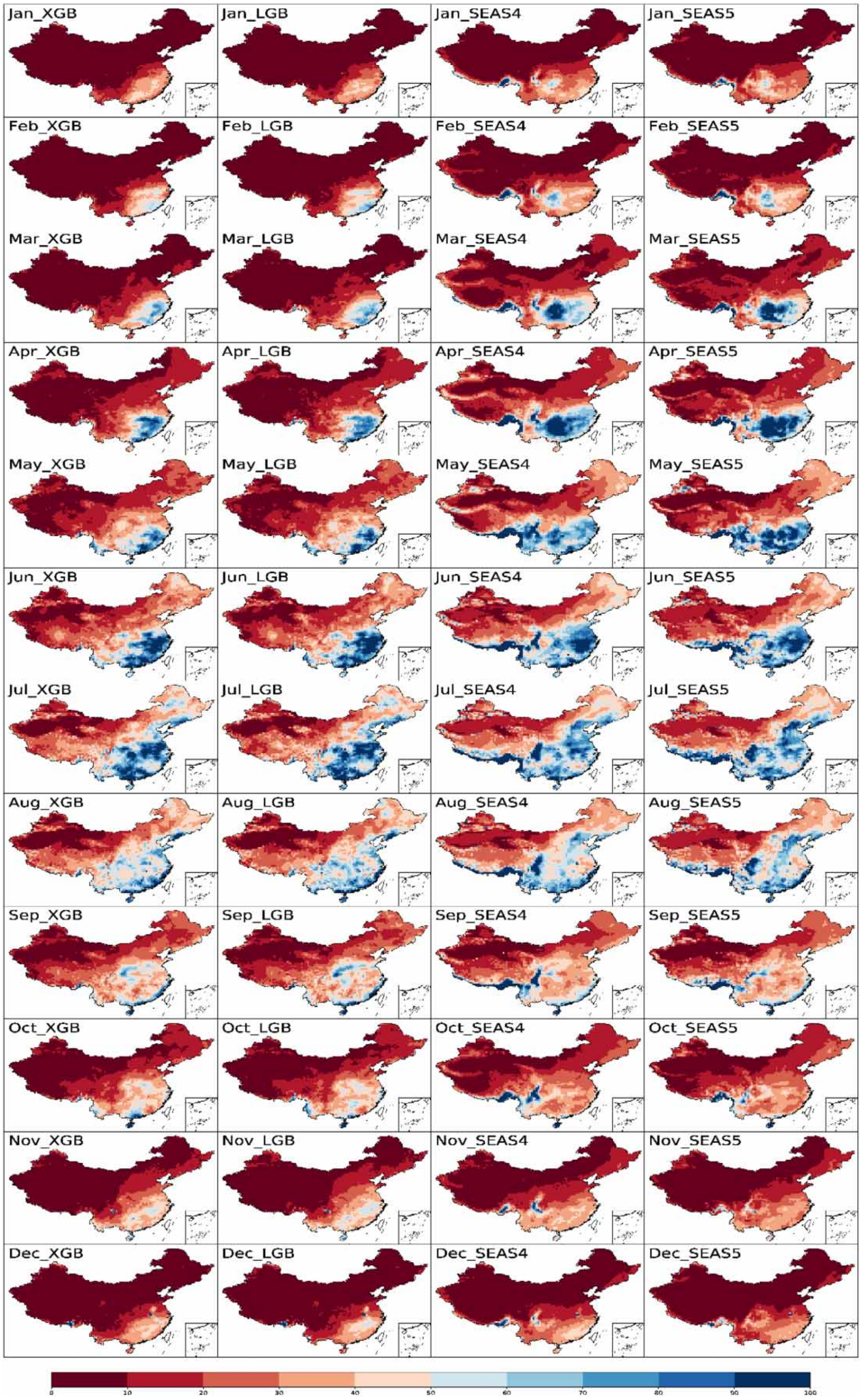

Figure 5 | MAE between four raw deterministic MPFs (namely XGB, LGB, the ensemble means of SEAS4 and SEAS5) and observations throughout 12 months for the period of 1981–2015. The best value of MAE is zero. The larger the values, the worse the forecast performance.

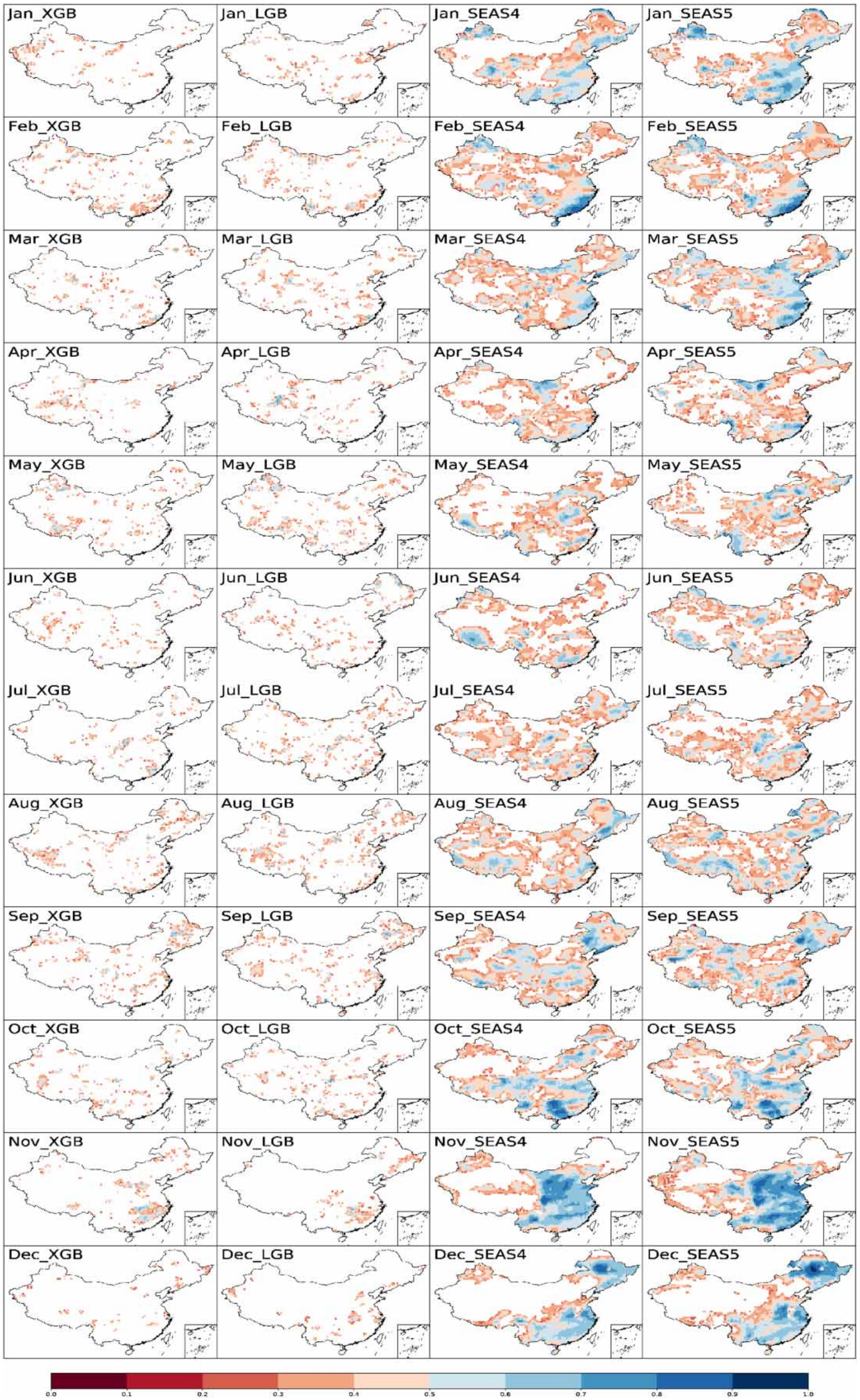

Figure 6 | PCC between four raw deterministic MPFs (namely XGB, LGB, the ensemble means of SEAS4 and SEAS5) and observations throughout 12 months for the period of 1981–2015. Only significant correlations (90% confidence level, according to a t-test) are shown. The best value of PCC is 1. The larger the values, the better the forecast performance.

SEAS5 can more effectively reflect the increasing or decreasing trend of precipitation than XGB and LGB.

Quality of BJP calibrated ensemble MPFs

As introduced above, we employed the BJP post-processing modeling approach to calibrate the four raw deterministic MPFs, which aims to correct bias and generate corresponding 1000-member ensemble MPFs. Figure 7 illustrates the RB between four calibrated ensemble MPFs and observation. As expected, in contrast to the results of Figure 4, post-processed MPFs are less biased than raw MPFs. The BJP method is effective at reducing RB of MLM-based and GCM-based MPFs. However, for November and December on the Qinghai-Tibet Plateau, the MPFs suffer negative RB, the magnitude of which varies between –30 and 0%. This is mostly caused by inaccurate precipitation observation or extreme precipitation events. Due to the sophisticated underlying surface environment, the error of precipitation observation may be magnified in dry months. For the southeast coast with dense observation sites, the RB for calibrated MPFs ranges from approximately 0 to 10%. This suggests that BJP is a useful method to generate bias-free ensemble MPFs. Meanwhile, it is also evident that post-processing is a necessary step before using the MLM-based and GCM-based MPFs in hydrological forecasting.

The accuracy of calibrated ensemble MPFs is analyzed by CRPSS using Figure 8. The climatology as the reference is also calculated by the BJP method (Wang *et al.* 2019b). Here we consider that white pixels (CRPSS between −5 and 5%) show neutrally (little or no) skillful forecasts (Zhao *et al.* 2017). Blue pixels (>5%) show positively skillful forecasts and red pixels (<− 5%) show negatively skillful forecasts. Overall, there is no great difference between XGB and LGB, with approximately 90% values of CRPSS located in white pixels throughout all months. The other 10% values mainly vary from 5 to 15, scattering sporadically on all the grid cells. For the SEAS4 and SEAS5, except that about 40% values of CRPSS are similar with climatology, the other 60% values tend to be positive, varying between 5 and 35, sometimes more than 35. Although there are several red pixels in the southwest and northwest of China, the proportion is very small. Moreover, the predictive performance of summer (June–August) is not as good as the other three seasons. Basically, SEAS5 shows a slightly better performance than SEAS4 and represents the most optimal forecasting skill in four models.

Reliability is analyzed in terms of PIT histograms (Figure 9) and PITSS plots (Figure 10). The PIT histograms visualize the distribution of PIT values to reflect the property of reliability. As shown in Figure 9, the PIT histograms exhibit inconsiderable variability between different models and different months. The red dotted line means an average level. Since there are 3824 grid cells and the period of loocv is 35 years, the average level for each month in this study is $3824 \times 35/10 = 13384$. The values of the frequency corresponding to 10 bins in the x-axis float around the red line (i.e. the PIT histograms concentrate uniform probability density in all ten bins), which presents good reliability. Figure 10 clearly illustrates that four calibrated ensemble MPFs are more reliable than climatology, with the emergence of a large proportion of blue pixels. It should be noted that in terms of reliability, the predictive performance is too close to distinguish.

Sharpness is analyzed using Figure 11. It should be noted that without reliability, a sharp forecast is misleading (Crochemore *et al.* 2016). Four calibrated ensemble MPFs are overall sharper than climatology in the large majority of grid cells. Some exceptions appear for December, and especially in the edge of the Qinghai-Tibet Plateau of west China. The IQRSS of SEAS4 and SEAS5 is much better than that of XGB and LGB, confirmed by a double proof based on a larger area and darker blue. Moreover, the obvious aggregation pixels, especially in the southeast coastal area, reflect the remarkable improvement of calibrated ensemble MPFs on becoming sharper and gaining skill.

DISCUSSION

As illustrated in the introduction section, massive MLMs have been widely implemented in the MPFs, and the prediction performance of these deterministic forecasts has been verified around the world. However, most previous studies always focus on small catchments, with not many precipitation stations. This kind of small-scale evaluation is unconvincing to some extent. As a key process of applying

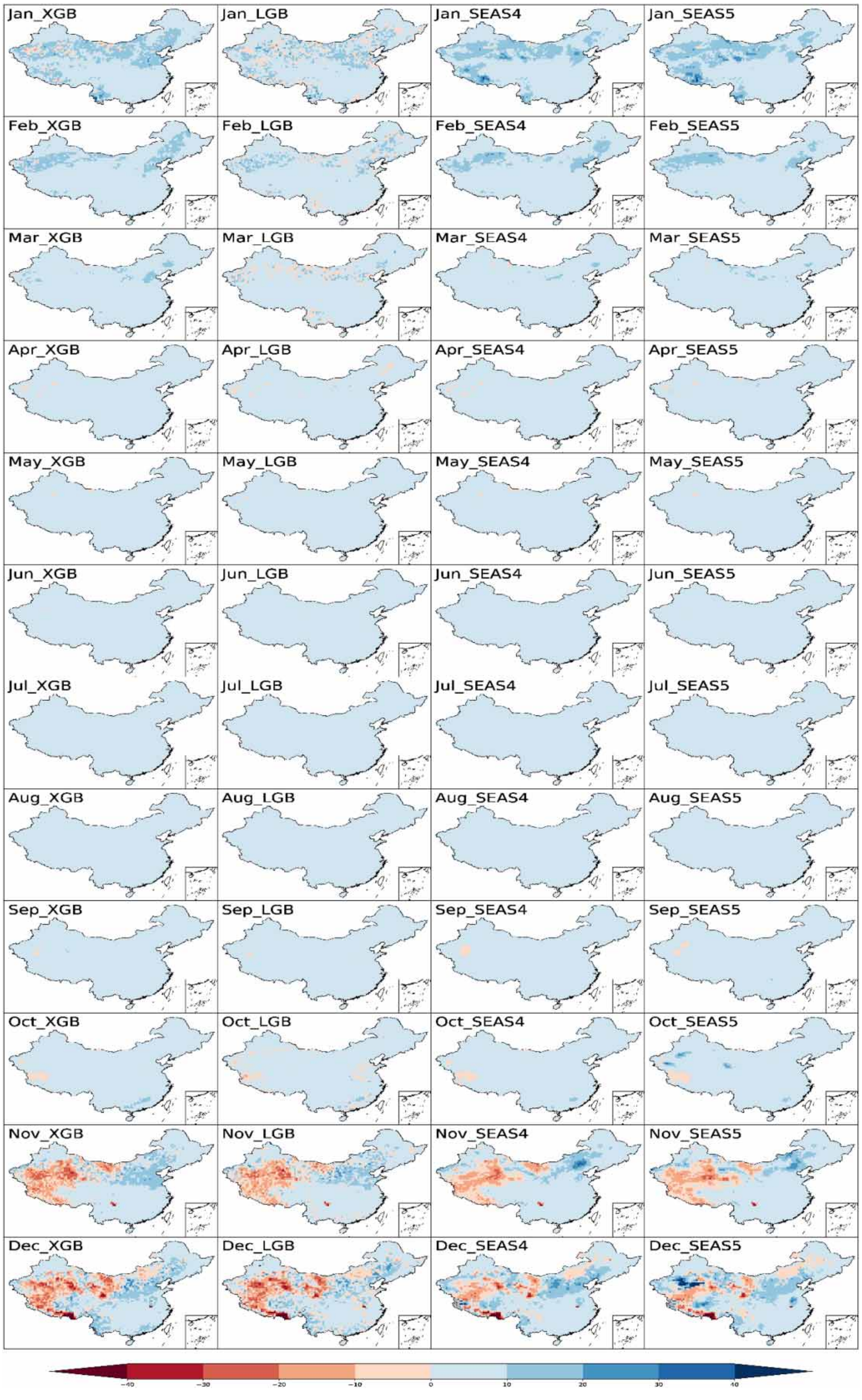

Figure 7 | RB (%) between ensemble mean of four calibrated ensemble MPFs (namely XGB, LGB, SEAS4, and SEAS5) and observations throughout 12 months for the period of 1981–2015. Blue areas represent a tendency of overpredicting precipitation, and red areas represent a tendency of underpredicting precipitation.

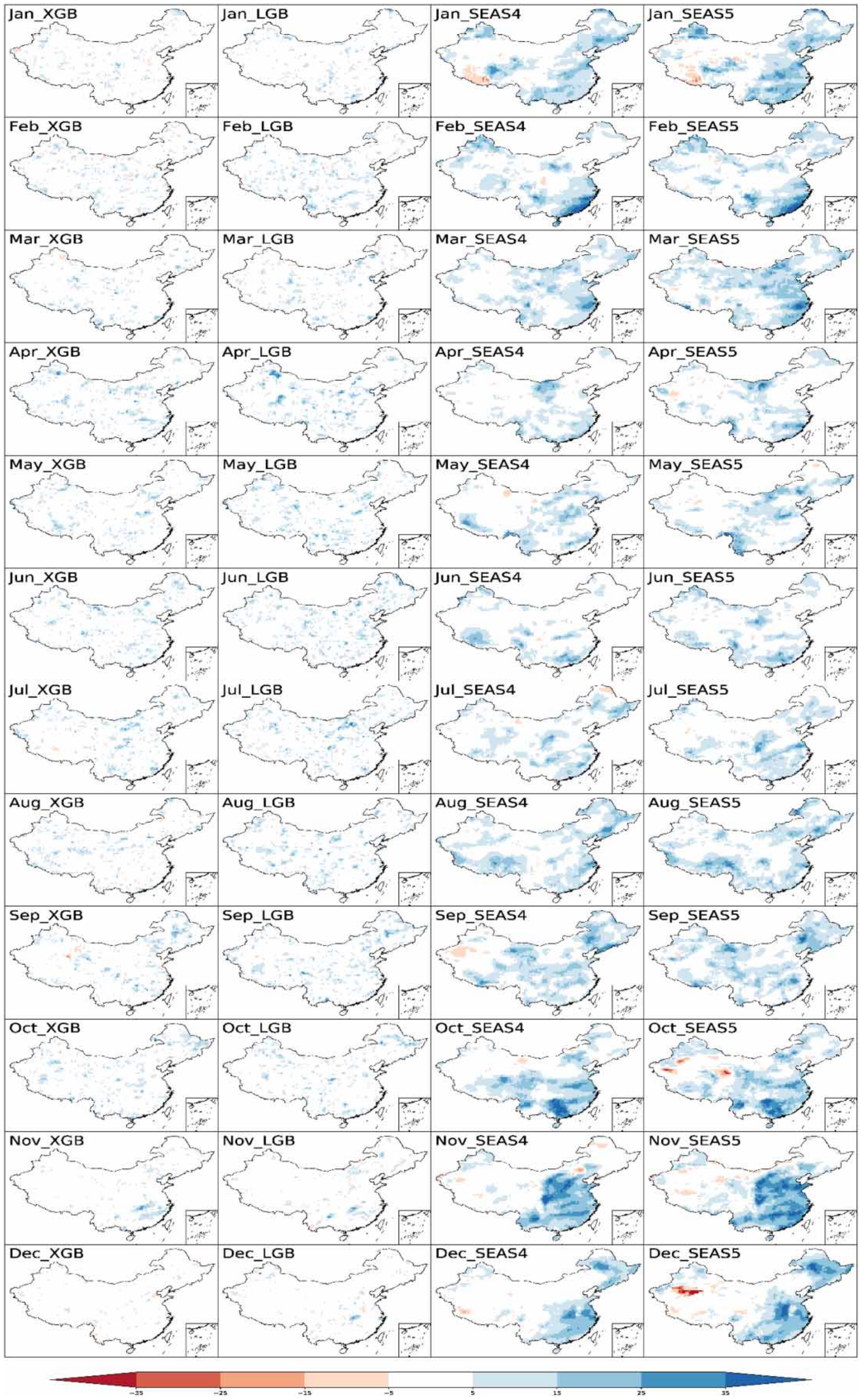

Figure 8 | CRPSS between four calibrated ensemble MPFs (namely XGB, LGB, SEAS4, and SEAS5) and reference (namely climatology) throughout 12 months for the period of 1981–2015. Blue pixels (>5%) show positively skillful forecasts; white pixels (−5 to 5%) show neutrally (little or no) skillful forecasts; red pixels (<− 5%) show negatively skillful forecasts.

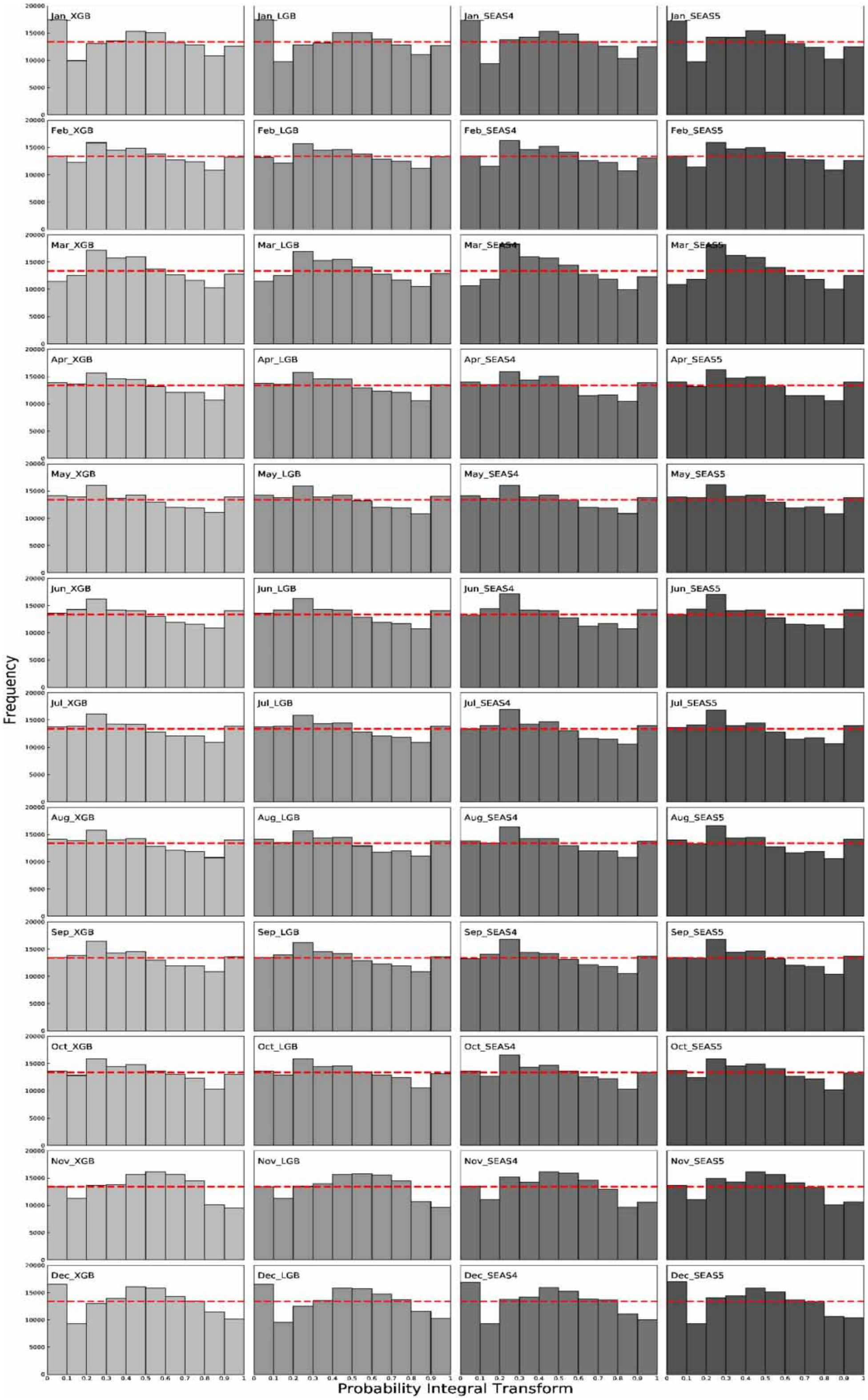

Figure 9 | PIT histograms of four BJP calibrated ensemble MPFs (namely XGB, LGB, SEAS4, and SEAS5) throughout 12 months for the period of 1981–2015. Red dotted line means an average level (in this case 13,384).

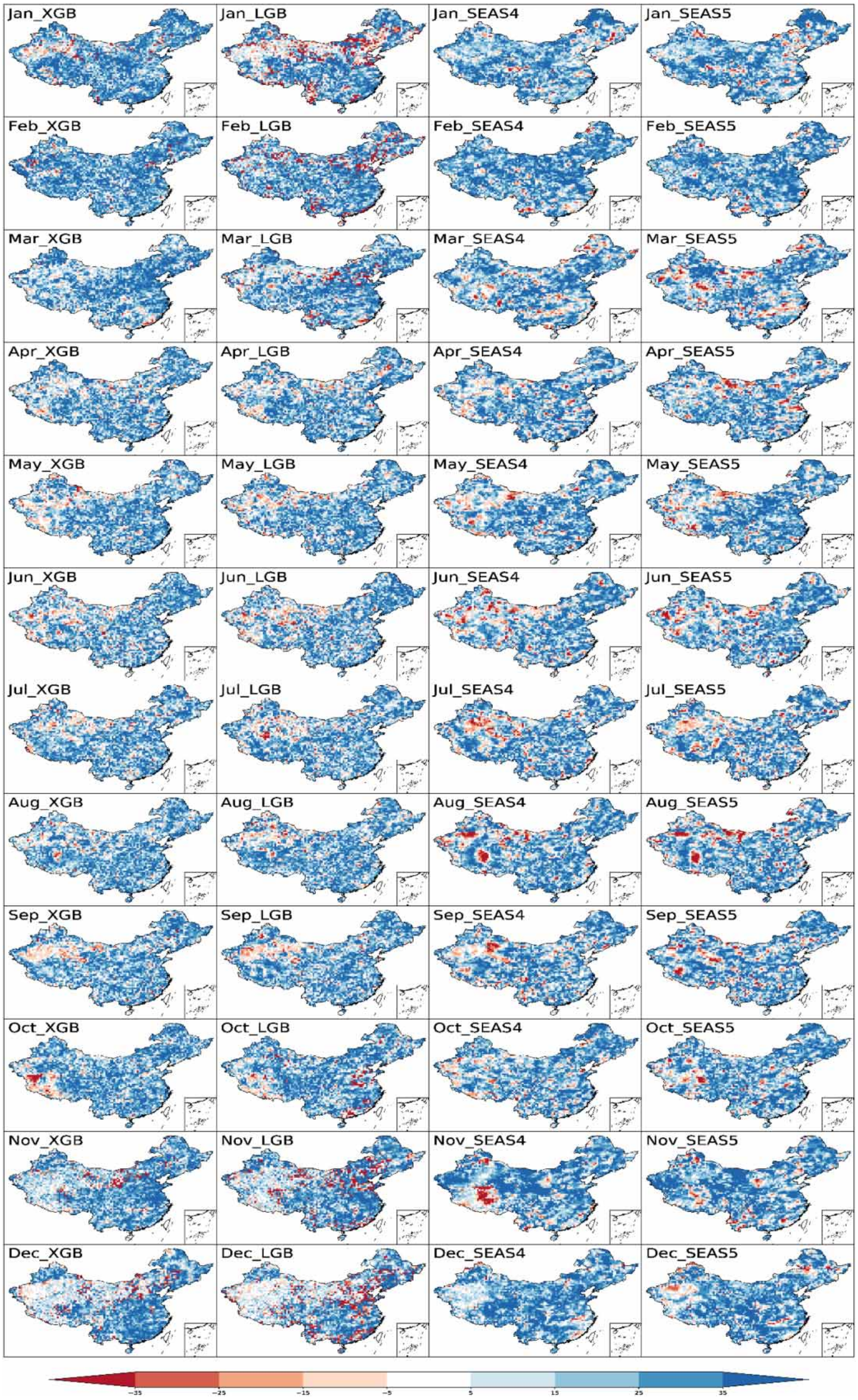

Figure 10 | PITSS between four calibrated ensemble MPFs (namely XGB, LGB, SEAS4, and SEAS5) and reference (namely climatology) throughout 12 months for the period of 1981–2015. Blue pixels (>5%) show positively skillful forecasts; white pixels (−5 to 5%) show neutrally (little or no) skillful forecasts; red pixels (<− 5%) show negatively skillful forecasts.

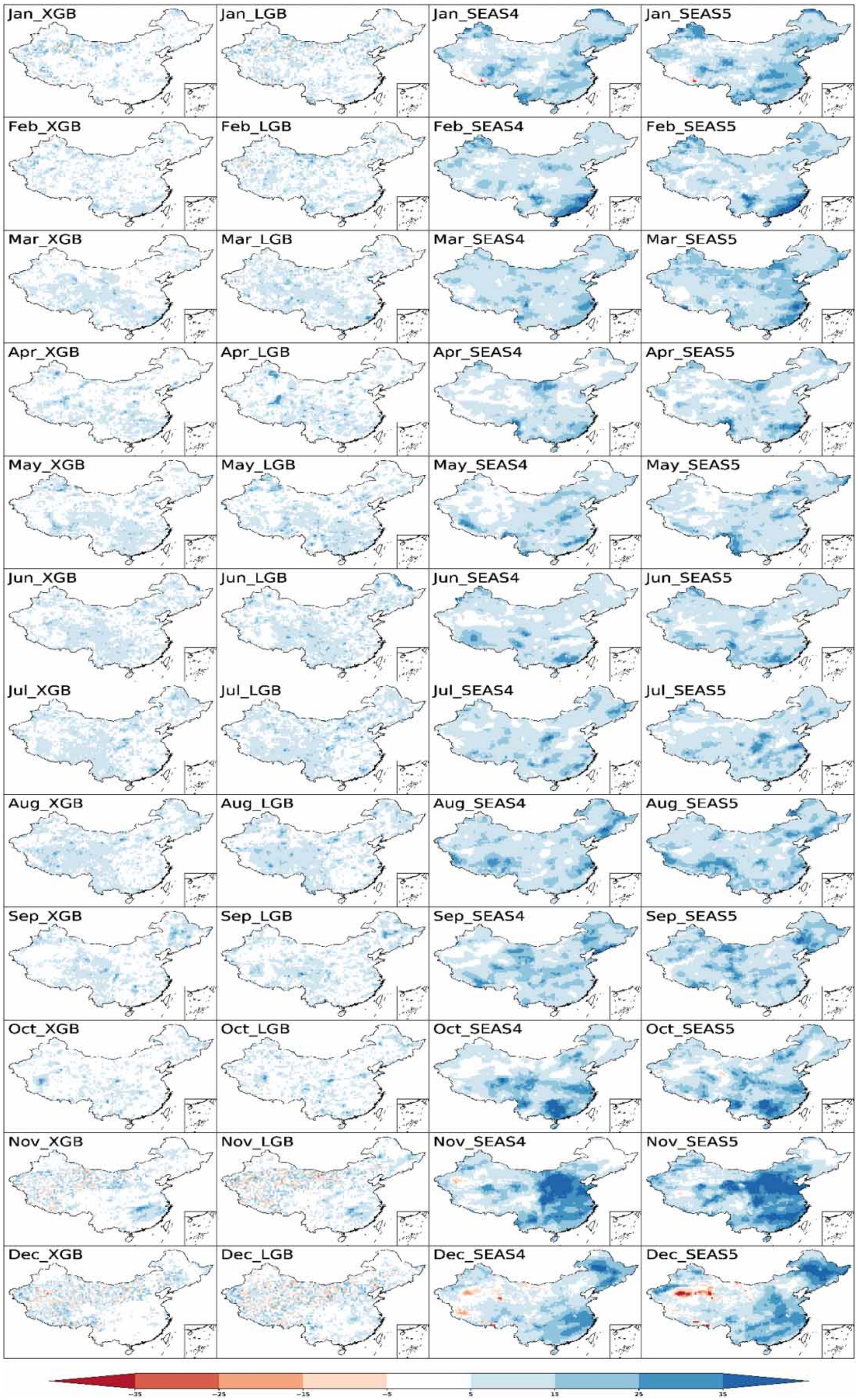

Figure 11 | IQRSS between four calibrated ensemble MPFs (namely XGB, LGB, SEAS4, and SEAS5) and reference (namely climatology) throughout 12 months for the period of 1981–2015. Blue pixels (>5%) show positively skillful forecasts; white pixels (−5 to 5%) show neutrally (little or no) skillful forecasts; red pixels (<− 5%) show negatively skillful forecasts.

MLMs to MPFs, hyperparameter optimization may be the most important factor affecting the forecasting skill. We have reason to believe that by excessively optimizing the hyperparameters, the small-scale predictions will be in good agreement with the observations.

Therefore, in order to comprehensively evaluate the deterministic forecasting performance, we have generated gridded MPFs with a resolution of 0.5° (3824 grid cells), almost spanning the entire Chinese mainland. In this case, it is almost impossible to repeatedly and accurately optimize the hyperparameters of each grid cell (although the continuous development of the Automatic Machine Learning method provides the possibility for the subsequent research of gridded hyperparameter optimization, it has not yet been applied to the hydrological community). The application of the same hyperparameters to all grid cells provides certain evidence for the general adaptation and performance of forecasting. Meanwhile, XGB and LGB have been proven to have advanced structures and weak sensitivity to hyperparameters (Chen *et al.* 2015; Chen & Guestrin 2016; Ke *et al.* 2017), which is also the basis of this prediction attempt. Moreover, due to the variety of large-scale climate predictors, as well as the complexity of the relationship between the precipitation and predictors, the main purpose of this study is to analyze the difference of MPFs between MLMs and GCMs in the above assumptions, rather than to explore what are the significant predictors that affect the precipitation in a specific region.

What calls for special attention is the evaluation of the second part. The forecasting skill of raw ensemble MPFs from SEAS4 and SEAS5 are much worse than climatology (not shown), this is the reason why we chose to compare calibrated GCMs with climatology instead of raw ensemble GCMs. Meanwhile, although only the ensemble mean values are used as the input of the calibration process, the BJP modeling approach has been proven to be a viable option for producing ensemble time-series MPFs (Hawthorne *et al.* 2013; Peng *et al.* 2014b; Schepen & Wang 2014; Schepen *et al.* 2014; Khan *et al.* 2015; Shrestha *et al.* 2015).

In addition, Molteni *et al.* (2011) and Johnson *et al.* (2019) have shown that the forecasting performance of SEAS4 and SEAS5 decreased (e.g. in terms of CRPSS) with the lead time increasing from 0 to one month. However, the 130 types of climate predictors provided by CMA have a one-month lag time, i.e. 0-month lead time forecasts cannot be generated by MLMs in this study (as shown in the upper half of Figure 2). For this reason, this study focuses on the comparison with the one-month lead time. In the future, we will continue the research to evaluate the forecasting performance between MLMs and GCMs with different lead times.

CONCLUSIONS

In this study, 130 types of large-scale climate predictors and two advanced MLMs (namely XGB and LGB) were applied to generate deterministic MPFs with a resolution of 0.5° across China. Meanwhile, the latest ECMWF's GCMs (namely Seasonal Forecast System 5 and its predecessor System 4) were used to compare with the above two MLMs for the same lead time and grid cells. Moreover, the BJP post-processing modeling approach was employed to calibrate the raw deterministic MPFs and to generate the corresponding ensemble MPFs. Raw and post-processing MPFs were put against gridded observations with different precipitation regimes for all months over the period of 1981–2015.

Deterministic evaluation of raw MPFs was evaluated by RB, MAE, and PCC in terms of bias, accuracy, and correlation. Compared with GCMs, the results indicated that the MLMs represented obviously smaller RB (Figure 4), similar MAE (Figure 5) as well as poorer PCC values (Figure 6). It can be concluded that the forecasting performance of MLMs was more inclined to generate random forecasts around the mean value. In contrast, due to the significant PCC values, the GCMs could reflect the increasing or decreasing trend of precipitation to some degree. However, the forecasting performance of raw deterministic MPFs was strongly dependent on forecasting regions and months.

Probabilistic evaluation of post-processing MPFs was evaluated by RB, CRPSS, PITSS, and IQRSS in terms of bias, accuracy, reliability, and sharpness. Since the BJP modeling approach had been proven to be effective in calibrating raw forecasts, we directly compared the post-processing ensemble MPFs with climatology that was also generated by BJP. The results indicated that the four post-processing ensemble MPFs were unbiased (Figure 7) and reliable (Figure 9). Meanwhile, in contrast to climatology, reliability (Figure 10) and sharpness (Figure 11) were all significantly improved. The results of the overall accuracy metric (Figure 8) showed that the ensemble MPFs generated from MLMs were similar to climatology, with no obvious difference. In contrast, the ensemble MPFs generated from GCMs achieved better forecasting skills and were not dependent on forecasting regions and months.

ACKNOWLEDGEMENTS

The study was financially supported by the Major Project of Zhejiang Natural Science Foundation, China (LZ20E090001), and the Science and Technology Project of Zhejiang Provincial Water Resources Department (RB2012). We gratefully acknowledge the anonymous editors and reviewers for their insightful and professional comments, which greatly improved this manuscript.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Bai, Y., Chen, Z., Xie, J. & Li, C. 2016 Daily reservoir inflow forecasting using multiscale deep feature learning with hybrid models. *Journal of Hydrology* **532**, 193–206.

Bazile, R., Boucher, M.-A., Perreault, L. & Leconte, R. 2017 Verification of ECMWF system 4 for seasonal hydrological forecasting in a northern climate. *Hydrology and Earth System Sciences* **21**, 5747–5762.

Bennett, J. C., Wang, Q., Li, M., Robertson, D. E. & Schepen, A. 2016 Reliable long-range ensemble streamflow forecasts: combining calibrated climate forecasts with a conceptual runoff model and a staged error model. *Water Resources Research* **52** (10), 8238–8259.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Cao, J., Yao, P., Wang, L. & Liu, K. 2014 Summer rainfall variability in low-latitude highlands of China and subtropical Indian Ocean dipole. *Journal of Climate* **27** (2), 880–892.

Chen, T. & Guestrin, C. 2016 Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, CA, pp. 785–794.

Chen, S.-T., Yu, P.-S. & Tang, Y.-H. 2010 Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology* **385** (1–4), 13–22.

Chen, T., He, T., Benesty, M., Khotilovich, V. & Tang, Y. 2015 Xgboost: extreme gradient boosting. R package version 0.4-2, pp. 1–4.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. & Wilby, R. 2004 The Schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology* **5** (1), 243–262.

Crochemore, L., Ramos, M.-H. & Pappenberger, F. 2016 Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences* **20** (9), 3601–3618.

Duan, Q., Pappenberger, F., Wood, A., Cloke, H. L. & Schaake, J. 2019 *Handbook of Hydrometeorological Ensemble Forecasting*. Springer, Berlin, Germany.

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X. & Xiang, Y. 2018 Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. *Energy Conversion and Management* **164**, 102–111.

Friedman, J. H. 2002 Stochastic gradient boosting. *Computational Statistics & Data Analysis* **38** (4), 367–378.

Hawthorne, S., Wang, Q., Schepen, A. & Robertson, D. 2013 Effective use of general circulation model outputs for forecasting monthly rainfalls to long lead times. *Water Resources Research* **49** (9), 5427–5436.

He, Q., Geng, F., Li, C., Mu, H., Zhou, G., Liu, X. & Cheng, T. 2018 Long-term variation of satellite-based PM2. 5 and influence factors over East China. *Scientific Reports* **8** (1), 1–10.

Hersbach, H. 2000 Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting* **15** (5), 559–570.

Jeon, J.-J., Sung, J. H. & Chung, E.-S. 2016 Abrupt change point detection of annual maximum precipitation using fused lasso. *Journal of Hydrology* **538**, 831–841.

Jeong, D., St-Hilaire, A., Ouarda, T. & Gachon, P. 2012 Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada. *Stochastic Environmental Research and Risk Assessment* **26** (5), 633–653.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L. & Monge-Sanz, B. M. 2019 SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development* **12** (3), 1087–1117.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. & Liu, T. Y. 2017 Lightgbm: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* **30**, 3146–3154.

Khan, M. Z. K., Sharma, A., Mehrotra, R., Schepen, A. & Wang, Q. 2015 Does improved SSTA prediction ensure better seasonal rainfall forecasts? *Water Resources Research* **51** (5), 3370–3383.

Laio, F. & Tamea, S. 2007 Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences* **11** (4), 1267–1277.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W. & Di, Z. 2017 A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water* **4** (6), e1246.

Li, W., Duan, Q., Ye, A. & Miao, C. 2019 An improved meta-Gaussian distribution model for post-processing of precipitation forecasts by censored maximum likelihood estimation. *Journal of Hydrology* **574**, 801–810.

Liang, Z., Li, Y., Hu, Y., Li, B. & Wang, J. 2018 A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework. *Theoretical and Applied Climatology* **133** (1–2), 137–149.

Liu, S., Xu, J., Zhao, J., Xie, X. & Zhang, W. 2014 Efficiency enhancement of a process-based rainfall–runoff model using a new modified AdaBoost. RT technique. *Applied Soft Computing* **23**, 521–529.

Liu, X., Wu, T., Yang, S., Li, T., Jie, W., Zhang, L. & Nie, S. 2017 MJO prediction using the sub-seasonal to seasonal forecast model of Beijing Climate Center. *Climate Dynamics* **48** (9–10), 3283–3307.

Liu, T., Chen, Y., Li, B., Hu, Y., Qiu, H. & Liang, Z. 2019 Long-term streamflow forecasting for the Cascade Reservoir System of Han River using SWAT with CFS output. *Hydrology Research* **50** (2), 655–671.

Ma, L., Zhang, G. & Lu, E. 2018 Using the gradient boosting decision tree to improve the delineation of hourly rain areas during the summer from advanced Himawari imager data. *Journal of Hydrometeorology* **19** (5), 761–776.

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L. & Vitart, F. 2011 *The new ECMWF Seasonal Forecast System (System 4)*, Vol. 49. European Centre for Medium-Range Weather Forecasts, Reading.

Park, I. W. & Mazer, S. J. 2018 Overlooked climate parameters best predict flowering onset: assessing phenological models using the elastic net. *Global Change Biology* **24** (12), 5972–5984.

Peng, Z., Wang, Q., Bennett, J. C., Pokhrel, P. & Wang, Z. 2014a Seasonal precipitation forecasts over China using monthly large-scale oceanic-atmospheric indices. *Journal of Hydrology* **519**, 792–802.

Peng, Z., Wang, Q. J., Bennett, J. C., Schepen, A., Pappenberger, F., Pokhrel, P. & Wang, Z. 2014b Statistical calibration and bridging of ECMWF system4 outputs for forecasting seasonal precipitation over China. *Journal of Geophysical Research: Atmospheres* **119** (12), 7116–7135.

Qiu, M., Zhao, P., Zhang, K., Huang, J., Shi, X., Wang, X. & Chu, W. 2017 A short-term rainfall prediction model using multi-task convolutional neural networks. In: *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, New Orleans, USA, pp. 395–404.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M. & Franks, S. W. 2010 Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. *Water Resources Research* **46** (5), 1–22.

Robertson, D., Shrestha, D. & Wang, Q. 2013 Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrology and Earth System Sciences* **17** (9), 3587–3603.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P. & Becker, E. 2014 The NCEP climate forecast system version 2. *Journal of Climate* **27** (6), 2185–2208.

Schepen, A. & Wang, Q. 2014 Ensemble forecasts of monthly catchment rainfall out to long lead times by post-processing coupled general circulation model output. *Journal of Hydrology* **519**, 2920–2931.

Schepen, A., Wang, Q. & Robertson, D. 2012 Evidence for using lagged climate indices to forecast Australian seasonal rainfall. *Journal of Climate* **25** (4), 1230–1246.

Schepen, A., Wang, Q. & Robertson, D. E. 2014 Seasonal forecasts of Australian rainfall through calibration and bridging of coupled GCM outputs. *Monthly Weather Review* **142** (5), 1758–1770.

Schepen, A., Zhao, T., Wang, Q. J. & Robertson, D. E. 2018 A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments. *Hydrology and Earth System Sciences* **22** (2), 1615–1628.

Shen, C. 2018 A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research* **54** (11), 8558–8593.

Shrestha, D. L., Robertson, D. E., Bennett, J. C. & Wang, Q. 2015 Improving precipitation forecasts by generating ensembles through postprocessing. *Monthly Weather Review* **143** (9), 3642–3663.

Street, R. B. 2016 Towards a leading role on climate services in Europe: a research and innovation roadmap. *Climate Services* **1**, 2–5.

Tian, D., Wood, E. F. & Yuan, X. 2017 CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States. *Hydrology and Earth System Sciences* **21** (3), 1477–1490.

Ukkonen, P. & Mäkelä, A. 2019 Evaluation of machine learning classifiers for predicting deep convection. *Journal of Advances in Modeling Earth Systems* **11** (6), 1784–1802.

Wang, Q. & Robertson, D. 2011 Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resources Research* **47** (2), 1–19.

Wang, Q., Robertson, D. & Chiew, F. 2009 A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resources Research* **45** (5), 1–18.

Wang, Q., Shrestha, D. L., Robertson, D. & Pokhrel, P. 2012 A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research* **48** (5), 1–7.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S. & Bai, X. 2015 Flood hazard risk assessment model based on random forest. *Journal of Hydrology* **527**, 1130–1141.

Wang, K., Yang, X., Liu, X. & Liu, C. 2017 A simple analytical infiltration model for short-duration rainfall. *Journal of Hydrology* **555**, 141–154.

Wang, K., Liu, X., Tian, W., Li, Y., Liang, K., Liu, C. & Yang, X. 2019a Pan coefficient sensitivity to environment variables across China. *Journal of Hydrology* **572**, 582–591.

Wang, Q., Zhao, T., Yang, Q. & Robertson, D. 2019b A seasonally coherent calibration (SCC) model for post-processing numerical weather predictions. *Monthly Weather Review* **147** (10), 3633–3647.

Yuan, X. & Wood, E. F. 2012 Downscaling precipitation or bias-correcting streamflow? some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resources Research* **48** (12), 1–7.

Yuan, X., Wood, E. F., Luo, L. & Pan, M. 2011 A first look at climate forecast system version 2 (CFSv2) for hydrological seasonal prediction. *Geophysical Research Letters* **38** (13), 1–7.

Yuan, X., Ma, F., Wang, L., Zheng, Z., Ma, Z., Ye, A. & Peng, S. 2016 An experimental seasonal hydrological forecasting system over the Yellow River basin – part 1: understanding the role of initial hydrological conditions. *Hydrology and Earth System Sciences* **20** (6), 2437–2451.

Zhang, J., Zhu, Y., Zhang, X., Ye, M. & Yang, J. 2018 Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *Journal of Hydrology* **561**, 918–929.

Zhao, Y., Zhu, J. & Xu, Y. 2014 Establishment and assessment of the grid precipitation datasets in China for recent 50 years. *Journal of Meteorological Sciences* **34** (4), 414–420.

Zhao, T., Wang, Q. J., Bennett, J. C., Robertson, D. E., Shao, Q. & Zhao, J. 2015 Quantifying predictive uncertainty of streamflow forecasts based on a Bayesian joint probability model. *Journal of Hydrology* **528**, 329–340.

Zhao, T., Schepen, A. & Wang, Q. 2016 Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach. *Journal of Hydrology* **541**, 839–849.

Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E. & Ramos, M. H. 2017 How suitable is quantile mapping for postprocessing GCM precipitation forecasts? *Journal of Climate* **30** (9), 3185–3196.

Zhao, Y., Duan, A. & Wu, G. 2018 Interannual variability of late-spring circulation and diabatic heating over the Tibetan Plateau associated with Indian Ocean forcing. *Advances in Atmospheric Sciences* **35** (8), 927–941.

Zhao, T., Wang, Q. J. & Schepen, A. 2019a A Bayesian modelling approach to forecasting short-term reference crop evapotranspiration from GCM outputs. *Agricultural and Forest Meteorology* **269**, 88–101.

Zhao, T., Wang, Q. J., Schepen, A. & Griffiths, M. 2019b Ensemble forecasting of monthly and seasonal reference crop evapotranspiration based on global climate model outputs. *Agricultural and Forest Meteorology* **264**, 114–124.

First received 24 November 2020; accepted in revised form 28 March 2021. Available online 11 May 2021