Journal of Hydrology 593 (2021) 125862

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

On the use of convolutional Gaussian processes to improve the seasonal forecasting of precipitation and temperature

Chao Wang [a,b], Wei Zhang [b], Gabriele Villarini [b,c,*]

[a] Department of Industrial and Systems Engineering, The University of Iowa, Iowa City, IA, USA
[b] IIHR—Hydroscience & Engineering, The University of Iowa, Iowa City, IA, USA
[c] Department of Civil and Environmental Engineering, The University of Iowa, Iowa City, IA, USA

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief

Keywords:
Convolutional Gaussian process
NMME
Seasonal forecasting
Machine learning

ABSTRACT

This study examines the potential improvement in seasonal predictability of monthly precipitation and temperature using a novel machine learning approach, the convolutional Gaussian process (CGP). This approach allows us to take into account multiple quantities and their interdependencies simultaneously. We use one global climate model (FLORb01) part of the North American Multi-Model Ensemble (NMME) project and quantify its skill in reproducing precipitation and temperature in March and July across Iowa (central United States) for lead times from one month to one year. As a first step we train the CGP over the 1985–2005 period, and then apply it out of sample from 2006 to 2019. Over the validation period, our results indicate that the CGP is able to increase the skill (i.e., increased correlation coefficient and reduced root mean squared error) in predicting precipitation and temperature compared to both the raw outputs and after standard bias correction. These statements are consistent across different lead times and target month (i.e., March or July). These encouraging findings provide a new potential path towards improved predictability of the regional climate at the seasonal scale.

1. Introduction

Issues related to the prediction of climate-related variables with a lead time from several weeks to several months have been the subject of growing research interest. Examples of these heightened interest are represented by the operational European Seasonal-to-Interannual Prediction (EuroSIP; Vitart et al., 2007) and the North American Multi-Model Ensemble (NMME; Kirtman et al., 2014) and its related collection of papers (Archambault et al., 2019). Several studies have examined the performance of these models in forecasting different quantities of interest, including precipitation and temperature (e.g., Cash et al., 2019; Krakauer, 2019; Slater et al., 2017, 2019a; Zhang et al., 2019; Vittal et al., 2020), streamflow and coastal flooding (e.g., Khouakhi et al., 2019; Slater et al., 2019b; Slater and Villarini, 2018), drought (e.g., Ma et al., 2019), sea surface temperature (e.g., Hervieux et al., 2019; Shin and Huang, 2019), the El Niño Southern Oscillation (ENSO; Zhang et al., 2017; Kang and Lee, 2019; Tippett et al., 2019), and tropical cyclone activity (e.g., Harnos et al., 2019; Manganello et al., 2019; Villarini et al., 2019).

Many of these studies highlighted the strength associated with these projects in terms of skill at the shortest lead times; this is particularly true for precipitation, temperature and drought (e.g., Mo and Lyon, 2015; Roundy et al., 2015; Ma et al., 2016) for areas that are under significant influence by ENSO. For instance, Infanti and Kirtman (2016) showed higher correlation between forecasted and observed precipitation/temperature in the southeastern, southwestern and northwestern North America during Eastern Pacific El Niño events. Some of the weaknesses of these models are tied to their limited capability in reproducing the precipitation and temperature at seasonal to annual lead times (e.g., Slater et al., 2017), or in forecasting extended period of above/below average conditions for the right reason (Kam et al., 2014).

Different approaches have been developed to improve the forecasting skill of these models, including bias correction and ensemble averaging (e.g., Hagedorn et al., 2005; DelSole and Tippett, 2014; Thober et al., 2015; Rodrigues et al., 2019; Slater et al., 2017). In terms of bias correction, most of the work has generally focused on the comparison of the observed and forecasted anomalies (e.g., Kirtman et al., 2014; Becker et al., 2014), with other approaches that have been applied to try to improve the predictive skill (e.g., Tian et al., 2014; Khajehei et al., 2018; Singh et al., 2017). For instance, Singh et al. (2017) explored the performance of standardization, quantile mapping, and nonlinear transformation in bias correcting the precipitation forecasts

* Corresponding author at: IIHR—Hydroscience & Engineering, The University of Iowa, 107C C. Maxwell Stanley Hydraulics Laboratory, Iowa City 52242, IA, USA.
E-mail address: gabriele-villarini@uiowa.edu (G. Villarini).

https://doi.org/10.1016/j.jhydrol.2020.125862
Received 9 April 2020; Received in revised form 2 December 2020; Accepted 4 December 2020
Available online 9 December 2020

over India, and found that the nonlinear transformation led the largest improvements. Narapusetty et al. (2018) proposed to first apply quantile mapping to bias-correct the temperature forecast, and then use the observed relationship between temperature and precipitation to forecast precipitation; they applied this methodology to the summer across the continental United States and found a reduction in root mean squared error (RMSE). Khajehei et al. (2018) compared and contrasted the performance of different post-processing approaches, finding a good performance for a copula-based ensemble post-processing. Slater et al. (2017) applied a simple Bayesian updating approach for temperature and precipitation forecasts across Europe, and found that it performed well in correcting the models' forecasts for conditional and unconditional biases (see also Zhang et al. (2017) for its application to ENSO).

However, most of the existing methods correct the biases in temperature and precipitation separately, neglecting the significant correlation between temperature and precipitation and their biases; in other words, the bias information in temperature would be informative to infer the bias in precipitation, and vice versa. As a result, the joint/simultaneous modeling and correction of temperature and precipitation biases would be a promising way to improve the overall prediction performance. However, this is not a trivial task. The major limitations that complicate the joint bias-correction of temperature and precipitation are: 1) the conventional methods (e.g., quantile mapping) are applied to single quantities after data fitting, thus they are not general and flexible enough to represent the information in both temperature and precipitation; and 2) there lacks an efficient way to explicitly characterize the relationship between temperature and precipitation (and their biases).

Machine learning techniques represent a good alternative, offering great flexibility in data mapping. For instance, Xu et al. (2019) applied machine learning to bias correct and statistically downscale precipitation forecasts over China. More specifically, they used wavelet support vector machine and wavelet random forest, and highlighted their better performance with respect to the more traditional quantile mapping. However, their method also only focuses on a single quantity, i.e., precipitation. Building explicit relationships between temperature and precipitation (and their biases) within the machine learning context is still an unsolved task. Moreover, few studies have examined the use of machine learning techniques to increase the predictive skill of these models simultaneously. More generally, Cohen et al. (2019) argued that "new statistical techniques mostly developed outside the field of climate science, collectively referred to as machine learning, can be adopted by climate forecasters to increase the accuracy of S2S [Subseasonal-to-Seasonal] predictions" and suggested that "S2S prediction will be most beneficial to the public by incorporating mixed or a hybrid of dynamical forecasts and updated statistical techniques such as machine learning."

Here we propose to use a novel machine learning approach, the convolutional Gaussian processes (CGP), to jointly improve the predictive skill of precipitation and temperature over Iowa (central United States) for two target months (i.e., March and July) and lead times from one month to one year. The basic idea of this method is to use a specially designed CGP to fit the relationship between the forecast and bias values for temperature and precipitation. The feature of the proposed method is that it not only provides a flexible representation of the bias in temperature and precipitation, but also builds an explicit relationship between the bias in these two quantities. This feature is expected to outperform existing bias-correction methods in terms of both accuracy and flexibility.

This study is organized as follows. In the next section, we will describe the data and the methodology, while in Section 3 we present the results. Section 4 summarizes the main points and concludes the paper.

2. Data and methodology

We focus on a domain centered over Iowa, within the central United States. This is an area of the country that has been experiencing several hydrometeorological extremes, with the Missouri-Mississippi flood event of Spring 2019 being one of the most recent examples. This is a highly agricultural region, with most of the land used for corn and soybean production. Therefore, our improved capability in predicting weather conditions with long lead times can have significant impacts for our preparation, response and mitigation efforts.

Here we focus on the evaluation of March and July precipitation and temperature for lead times ranging from one month to one year. We have selected these two months as representative of two different hydrometeorological conditions, one tied to the early spring snowmelt, and one to summertime convection. The reference data for temperature and precipitation are based on the monthly Parameter-elevation Regressions on Independent Slopes Model (PRISM) dataset during 1981–2018 at a spatial resolution of 4 km across the continental United States (Daly et al., 2008). We focus on the seasonal forecasts performed by the Geophysical Fluid Dynamics Laboratory (GFDL) Forecast-oriented Low Ocean Resolution version of the CM2.5 (GFDL FLOR) Version B01 available from the NMME project (Vecchi et al., 2014). It has 12 members, but here we consider the ensemble average. This climate model has shown relatively promising skill in forecasting precipitation and temperature across the continental United States (Jia et al., 2014).

We compare the skill in forecasting temperature and precipitation produced by this model against what can be obtained through standard bias-correction approach (i.e., anomaly correlation; we use "BC" in referring to it in the rest of the paper) and the machine learning method introduced next. We jointly formulate the relationship between forecast and bias values of temperature and precipitation as follows:

$$\boldsymbol{e} = \boldsymbol{y}^O - \boldsymbol{y}^F = \boldsymbol{f}\left(\boldsymbol{y}^F\right) + \epsilon = \begin{bmatrix} f_T\left(\boldsymbol{y}^F\right) \\ f_P\left(\boldsymbol{y}^F\right) \end{bmatrix} + \epsilon \tag{1}$$

where $\boldsymbol{y}^O = \left[y_T^O, y_P^O\right]^T$ is a vector of observed values of temperature (y_T^O) and precipitation (y_P^O) at a specific location (grid); $\boldsymbol{y}^F = \left[y_T^F, y_P^F\right]^T$ is the vector of forecast values of temperature (y_T^F) and precipitation (y_P^F) at the same location (grid); $\boldsymbol{e} = \left[y_T^O - y_T^F, y_P^O - y_P^F\right]^T$ is the vector of the bias values, while ϵ is the vector of random noise described by an identically and independent distributed multivariate Normal distribution with mean vector 0 and covariance matrix $\sigma^2\mathbf{I}_2$ ($\mathbf{I}_2$ is a 2×2 identity matrix); $\boldsymbol{f}\left(\boldsymbol{y}^F\right) = \left[f_T\left(\boldsymbol{y}^F\right), f_P\left(\boldsymbol{y}^F\right)\right]^T$ is the target transformation vector that builds the relationship between the forecast ($\boldsymbol{y}^F$) and the bias ($\boldsymbol{e}$) vectors. Once we have this relationship $\boldsymbol{f}(\cdot)$, we can apply it to any new forecast value vector ($\tilde{\boldsymbol{y}}^F$) to obtain the corresponding bias value vector $\boldsymbol{f}\left(\tilde{\boldsymbol{y}}^F\right)$. Then the predicted observation vector would be $\tilde{\boldsymbol{y}}^F + \boldsymbol{f}\left(\tilde{\boldsymbol{y}}^F\right)$, which represents the temperature and precipitation values after correction.

The key to the accuracy and flexibility of this model is the way we design and construct $\boldsymbol{f}(\cdot)$. As mentioned in the Introduction, $\boldsymbol{f}(\cdot)$ should be: 1) flexible enough to capture the bias feature in temperature and precipitation; and 2) capable of capturing the implicit relationship between temperature and precipitation biases. We propose to use the convolutional based Gaussian process to construct $\boldsymbol{f}(\cdot)$ to satisfy these requirements.

The Gaussian process is a flexible nonparametric tool that provides a distribution over the infinite space of functions by treating all observations as random variables with a joint Gaussian distribution (Quinonero-Candela and Rasmussen, 2005). The convolution process is an effective way to construct a Gaussian process by convolving a Gaussian white noise process with a smoothing kernel (Boyle and Frean, 2005). We propose to generalize the univariate Gaussian process to the multivariate case to satisfy our requirements:

$$f_i\left(\boldsymbol{y}^F\right) = \sum_q \int_{-\infty}^{+\infty} K_{qi}\left(\boldsymbol{y}^F - \boldsymbol{t}\right) X_q(\boldsymbol{t}) d\boldsymbol{t} \tag{2}$$

$$\mathrm{cov}\left(f_i(\boldsymbol{y}^F), f_j(\boldsymbol{y}'^F)\right) = \sum_q \int_{-\infty}^{+\infty} K_{qi}(\boldsymbol{y}^F - \boldsymbol{t}) K_{qj}(\boldsymbol{y}'^F - \boldsymbol{t}) d\boldsymbol{t} \quad (3)$$

where $X_q(\cdot)$ is the qth Gaussian white noise process (latent function), $K_{qi}(\cdot)$ is the absolutely integrable kernel applied to the ith ($i = 1$ represents the temperature bias, $i = 2$ represents the precipitation bias) climate-metric bias based on the q^{th} latent function, and the $\mathrm{cov}\left(f_i(\boldsymbol{y}^F), f_j(\boldsymbol{y}'^F)\right)$ represents the covariance (inter-relationship) between the ith metric bias at input (forecast) value $\boldsymbol{y}^F$ and the jth metric bias at another input (forecast) value $\boldsymbol{y}'^F$. The univariate Gaussian process only allows single response metric (Higdon, 1998), while our proposed structure allows multiple response metrics and can characterize their interactions through Eq. (3). Please note the proposed framework in Eqs. (2) and (3) is not limited to two response metrics, and it can be easily extended to model inter-relationship among multiple response metrics. We also want to point out that different kernels (Rasmussen, 2003), such as Gaussian kernel, rational quadratic kernels, and periodic kernels, can be applied to the CGP to handle different input-output relationships, making the CGP flexible and adaptive to the forecast data $\boldsymbol{y}^F$ from different climatological models.

The proposed CGP structure is graphically demonstrated in Fig. 1a, where each surface is a Gaussian process representing a climate metric bias, and the black asterisks represent bias data (Eq. (1)) sampled at various input (forecast) values $\boldsymbol{y}^F = [y_T^F, y_P^F]$. In this structure, the $X_0(\cdot)$ is shared between two bias metrics to represent the metric similarity (interaction), while each bias has also its own latent function $X_i(\cdot)$ and kernel $K_{ii}(\cdot)$ to capture its uniqueness. The parameters in various kernels can be estimated from the data points (black asterisks in Fig. 1a) based on the maximum likelihood estimation method (Ying, 1991).

With the estimated parameters, it is straightforward to predict the bias response through Eq. (2) (Quinonero-Candela and Rasmussen, 2005). The prediction interval at a specific input (forecast) value $\boldsymbol{y}^F$, representing the uncertainties associated with our estimates, can also be obtained from Eqs. (2) and (3) (Quinonero-Candela and Rasmussen, 2005), which gives the prediction variance at a specific input (forecast) value $\boldsymbol{y}^F$. Please note that we consider two types of uncertainties: the observation errors and the prediction uncertainty. The observation error represents the measurement noise resulting from data collection, which is denoted as ϵ in Eq. (1). The prediction uncertainty represents the confidence of the prediction results. More specifically, if a prediction were generated at an input location near the training data, then the uncertainty for this prediction would be lower than that far away from the training data. We characterize this type of uncertainty as the covariance between difference Gaussian processes $f(\boldsymbol{y}^F)$, which corresponds to Eq. (3). As a result, the prediction variance comprehensively quantifies data uncertainties and the effect of the interactions among responses. The prediction results are demonstrated in Fig. 1b through a cross section view conditioning on a forecast temperature value (y_T^F). When we compare the prediction results between the proposed method and the standard bias correction methods (Maraun, 2016), we find that the proposed method leads to better prediction accuracy, demonstrating that we are able to obtain an accurate prediction (solid line). We can also observe that the proposed method can successfully quantify the uncertainties (95% confidence interval) around the predicted bias values, and all the data are within the intervals. These intervals reveal the 'normal' range for the bias, and they can also be incorporated into the anomaly detection of climate metric observations. The good performance of the proposed CGP structure is because we consider both the flexibility of individual metric bias and the interaction between metric biases, which brings additional information and facilitates a better constraining of the predictions.

The structure in Fig. 1 provides a flexible formulation of the climate metric biases and their interactions, and also facilitates uncertainty quantification. To apply the structure in Fig. 1 to correct the bias for new forecast temperature and precipitation, denoted as $\tilde{\boldsymbol{y}}^F$, we can first collect various forecast and observation pairs to construct the CGP $f(\cdot)$ based on Eq. (1)-(3). The $f(\cdot)$ builds the relationship between forecast values and the bias values. Once we have $f(\cdot)$, we can apply it to the new forecast value vector ($\tilde{\boldsymbol{y}}^F$) to obtain the corresponding bias value vector $f\left(\tilde{\boldsymbol{y}}^F\right)$. Then the predicted observation vector would be $\tilde{\boldsymbol{y}}^F + f\left(\tilde{\boldsymbol{y}}^F\right)$, which is the temperature and precipitation values after correction.

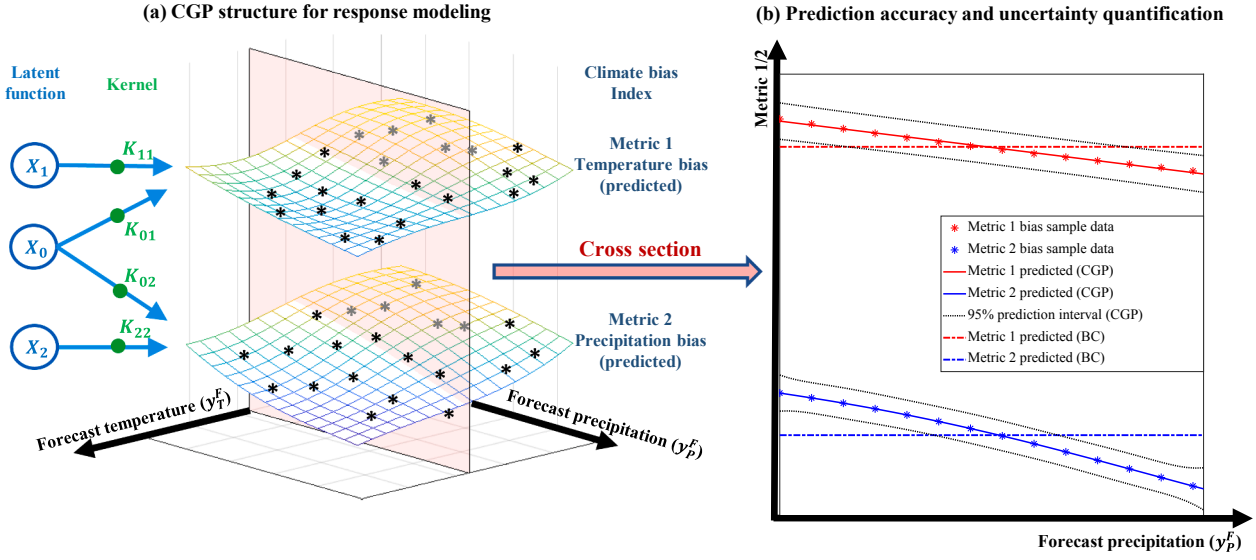

Fig. 1. The proposed CGP structure and the prediction results. (a) Schematic highlighting the relationship between model inputs (i.e.,y_T^F and y_P^F) and outputs (i.e., bias values of precipitation and temperature). Notice that neither surface is built in isolation, but their interdependencies are captured by X_0. (b) Predicted metrics (blue and red lines) and associated uncertainties across different values of precipitation for a cross-section obtained by model conditioning on a value of y_T^F. The results for the CGP (solid and dotted lines) are compared against what would be obtained using a simple bias correction method (dot-dashed horizontal lines). This highlights the better performance by CGP in reproducing the observations (asterisks) and in quantifying uncertainties. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Here we evaluate the forecast skill of the raw model outputs and after correction using standard BC and the CGP in terms of correlation coefficient and RMSE. We divide the sample into calibration and validation periods: the overall bias used for the BC approach and the parameters of the CGP approach are estimated based on the 1985–2005 period and evaluated on an out-of-sample validation period from 2006 to 2019. We estimate these parameters for every pixel within the study area, every target month and lead time.

3. Results

We will start by providing an example of the model performance for one pixel within the domain, target month March, with the forecast initialized in July of the previous year (Fig. 2). During the calibration period (1985–2005), the GFDL FLOR exhibits biases in both precipitation and temperature with respect to the observations (compare the black and orange lines). Moreover, the interannual variability is not captured either. Bias correction (BC) can help with the biases, even though it just represents a translation of the forecast and cannot help with the variability: at such a long lead time (i.e., 9-month lead time), much of the year-to-year variability is lost, and the time series tends to oscillate around the mean values (see also Villarini et al. (2019) for similar findings in terms of seasonal forecasting of hurricane activity). The use of the CGP approach, however, leads to a much improved forecast skill both in terms of magnitude and variability. These positive features are tied to the fact that we have a flexible modeling framework that is able to leverage the information shared between precipitation and temperature. The excellent performance exhibited in the calibration period is maintained in the validation period (2006–2019). The raw model outputs cannot represent well the observational record, and the situation is only mitigated once we use the BC correction based on the training period. On the other hand, the CGP approach is able to significantly improve the overall performance, closely mimicking the observational record. These statements are valid for both precipitation and temperature.

These features are clear once we examine the scatterplots in the right-hand panels of Fig. 2. They highlight the forecast model's lack of skill in reproducing the observations, with a correlation coefficient between raw (or BC) forecasts and observations that is either marginally positive (0.19 for temperature) or even negative (−0.28 for precipitation). The use of the CGP approach leads to a significant improvement in performance in terms of magnitude and variability. This is particularly true for precipitation, as highlighted by the scatterplot (Fig. 2, bottom-right panel): the correlation coefficient increases from −0.28 for the raw/BC forecast to 0.80 for the CGP method. The correlation coefficient for temperature is high and equal to 0.88, strongly suggesting that our proposed approach can improve the skill of not only precipitation, but also of temperature because of the way that the relationship between biases in these quantities is handled concurrently.

We can extend the results in Fig. 2 to the entire domain, both target months and all lead times. We start with the evaluation of the skill in forecasting precipitation for the target month March for the validation

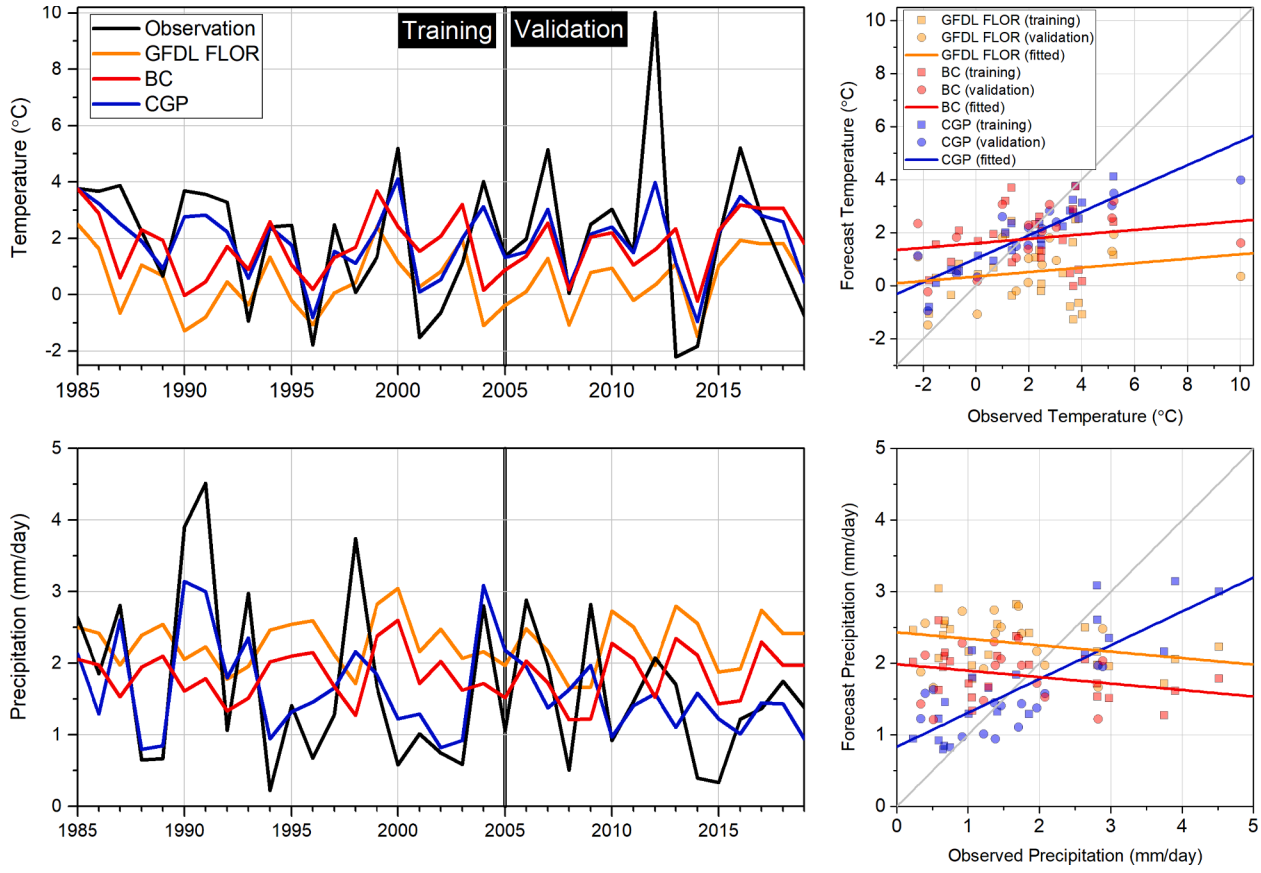

Fig. 2. Left panels: time series of temperature (top-left panel) and precipitation (bottom-left panel). The black lines represent the observations; the orange and red lines represent the GFDL FLOR outputs before and after BC, respectively, while the blue lines show the results after the application of the CGP method. The training period ranges from 1985 to 2005, while the validation period ranges from 2006 to 2019 (the vertical black lines separate the two periods). Right panels: scatterplots of observed and forecast temperatures (top-right panel) and precipitation (bottom-right panel). The orange and red symbols represent the GFDL FLOR outputs before and after BC, respectively, while the blue symbols show the results after the application of the CGP method. The squares and circles represent the training and validation periods, respectively, while the lines are based on linear regression results over the entire period. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

period 2006–2019 (Figs. 3 and 4). The correlation coefficient between the raw GFDL FLOR outputs and the observations tends to be relatively high at the shortest lead time, and to generally decrease as we increase the lead time. For instance, the February-initialized forecast tends to have negative correlation, and the situation does not change as we increase the lead time, with the correlation coefficients that vary between marginally positive to negative. For this metric, the BC approach does not help as the value of the correlation coefficient does not change if we shift all the values by a same amount. On the other hand, the situation significantly improves after using the CGP approach. For the shortest lead time we are able to preserve the areas with higher skill and to increase the values in the parts of the domain with lower skill. These improvements become even clearer as we increase the forecast horizon: the use of the CGP method leads to an overall increase in the correlation coefficient between observations and forecasts, with the majority of the domain exhibiting positive correlations. Another important element is that this heightened skill tends to be preserved across lead times.

This overall improvement is also clear when we consider the RMSE as the performance metric (Fig. 4). When we consider the raw GFDL FLOR output, the higher RMSE values tend to be located in the eastern part of the domain, with an overall tendency to increase as we move from the northwestern to the southeastern part of the domain. These results are consistent across the different lead times, with a tendency towards increasing values for longer lead times. When we perform BC, the situation tends to improve, with the RMSE that tends to increase from north to south. The use of the CGP method leads to a better performance compared to the BC: the RMSE values are smaller, with a better performance that now encompasses the entire study area across the different forecast horizons.

The advantage of using the CGP method is also clear when we consider March temperature (Figs. 3 and 5). The correlation coefficient between observations and the GFDL FLOR model is marginally positive across the entire domain for the March- to December-initialization; as we increase the lead time, there are large swings between overall negative (e.g., forecasts initialized in November, October, August and May) and positive (e.g., forecasts initialized in September and July), without a strong dependence on lead time (Fig. 3). The CGP approach, on the other hand, is able to increase the skill across all lead times (Fig. 3); the values of the correlation coefficient are generally large across the entire domain; the exceptions are represented by the November- and June-initialized forecasts, and to a lesser extent for May, even though the CGP method still leads to a much better performance compared to the raw model outputs. The improved performance of our proposed approach is also evident when we consider the RMSE as evaluation metric (Fig. 5). The larger values tend to be concentrated in the northern part of the study region, and to move east as we increase the lead time. The BC approach generally helps in reducing these values, even though the regional variability observed in raw outputs remains largely unchanged. The use of the CGP, on the other hand, leads to a significant performance improvement, with RMSE values that are much smaller than the original GFDL FLOR outputs both before and after bias correction. Based on these results, it is clear that the CGP method leads to a better performance in forecasting March precipitation and temperature.

We expand our analyses to the forecasting of July precipitation and temperature (Figs. 6–8). The correlation coefficient between the GFDL FLOR forecast and the observations is generally low to negative, with no regional consistency or dependence on lead time (Fig. 6); as we apply the proposed approach, there is a marked increase in skill, with large areas of the study region that exhibit high correlation with the observations regardless of the forecast horizon. The results related to the RMSE (Fig. 7) highlight how the BC approach can reduce it compared to the raw model outputs, and how the CGP approach reduces the RMSE even further.

The results for temperature lead to the same conclusions for precipitation: the CGP approach is able to increase the correlation relative

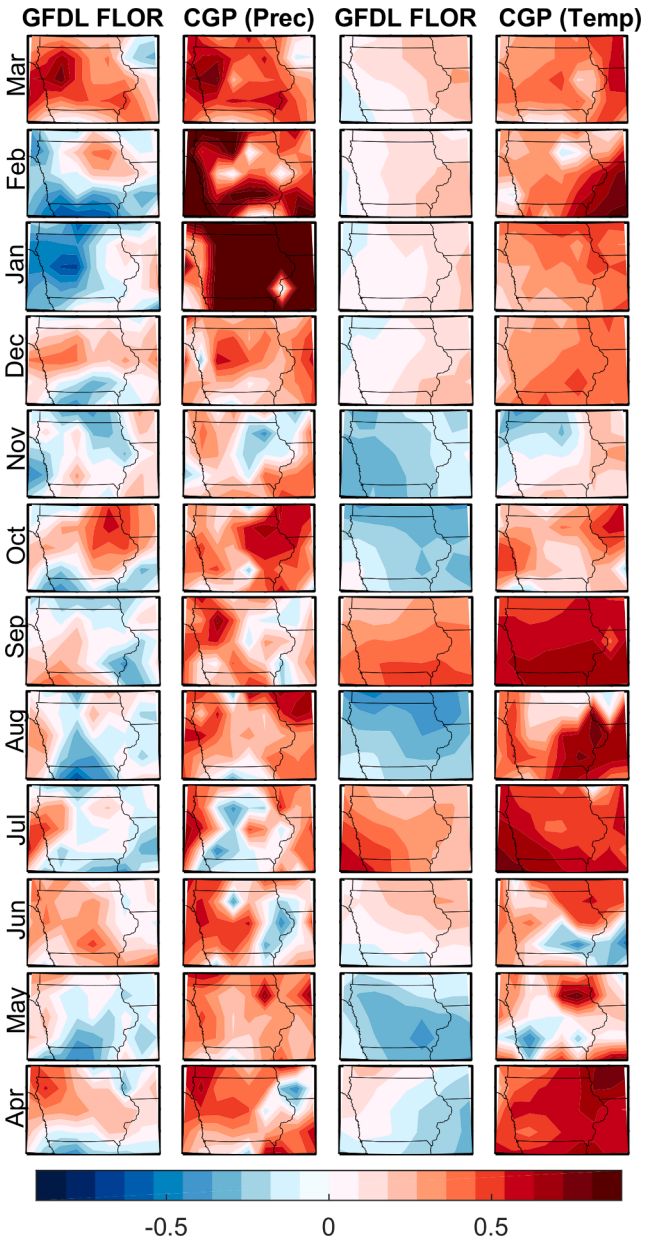

Fig. 3. Correlation coefficient between observed and forecast precipitation (first two columns) and temperature (last two columns) based on raw GFDL FLOR and after CGP for target month March. Moving from the top to the bottom row, the lead time increases from the shortest (i.e., March-target initialized at the beginning of March) to the longest (i.e., March-target initialized in April of the previous year). These results are for the validation period 2006–2019.

to the raw mode outputs, but also in an absolute sense, with values of the correlation coefficient of 0.4 and larger (Fig. 6). The improvements of our approach with respect to the raw and BC forecasts are even more apparent with considering RMSE as performance metric (Fig. 8): the largest values are for the raw model outputs, with the BC that leads to an improved performance. This is especially true for the shortest lead times, which suggests that the model outputs tend to have a warm bias, whose impacts can be generally mitigated by the BC. The application of the CGP approach leads to much smaller RMSE values, indicating that our proposed approach outperforms both the model and the standard bias correction approach.

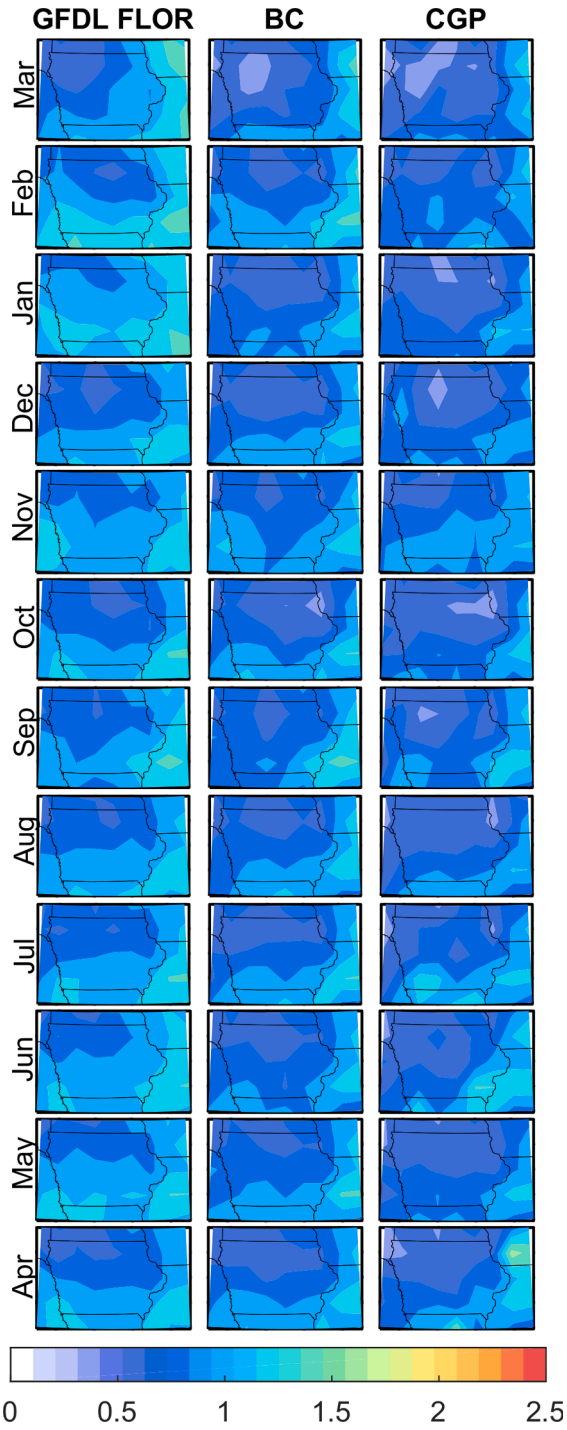

Fig. 4. Root mean square error (units: mm/day) between observed and forecast precipitation based on raw GFDL FLOR and after BC, and after CGP for target month March. Moving from the top to the bottom row, the lead time increases from the shortest (i.e., March-target initialized at the beginning of March) to the longest (i.e., March-target initialized in April of the previous year). These results are for the validation period 2006–2019.

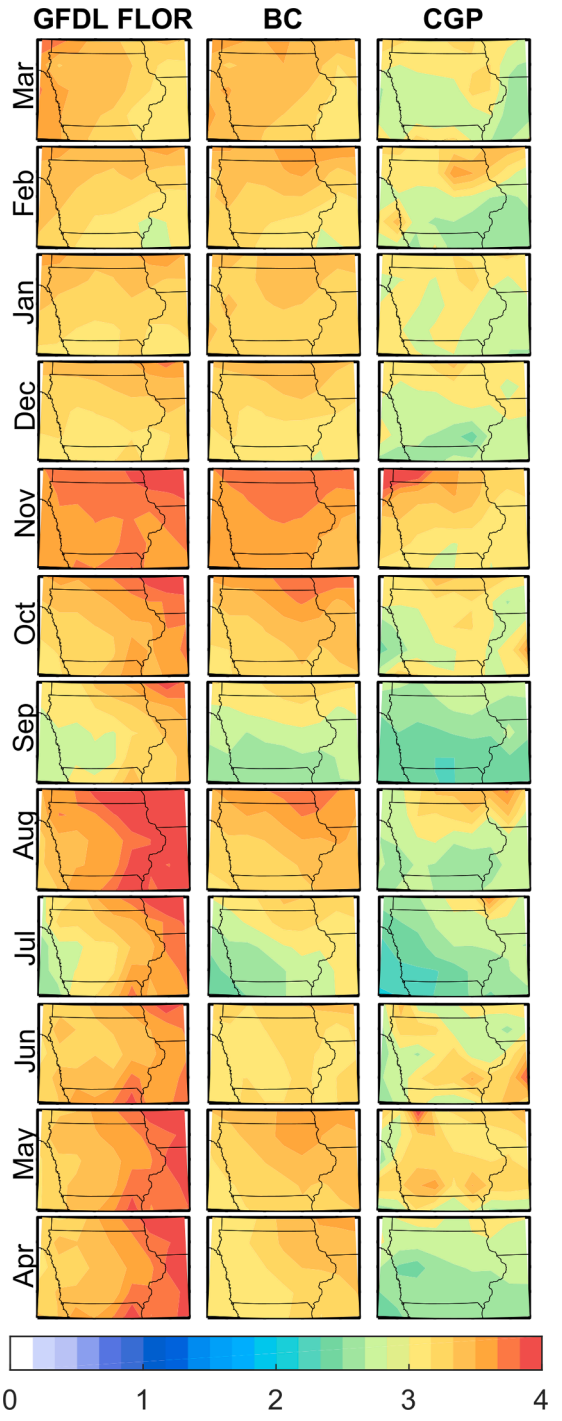

Fig. 5. Root mean square error (units: °C) between observed and forecast temperature based on raw GFDL FLOR and after BC, and after CGP for target month March. Moving from the top to the bottom row, the lead time increases from the shortest (i.e., March-target initialized at the beginning of March) to the longest (i.e., March-target initialized in April of the previous year). These results are for the validation period 2006–2019.

4. Conclusions

This study has examined the capability of a new machine learning technique, the convolutional Gaussian processes, in improving the performance of the GFDL FLOR model (part of the NMME project) in forecasting precipitation and temperature. Analyses have focused on Iowa (central United States), two target months (March and July) and forecast horizons from one month to one year. The main findings of this study can be summarized as follows:

- The raw GFDL FLOR model has biases, limited correlation with the observations, and large RMSE values. This statement is valid across lead times and target months. The BC approach can mitigate some of these issues (bias and RMSE) but not increase correlation. Despite

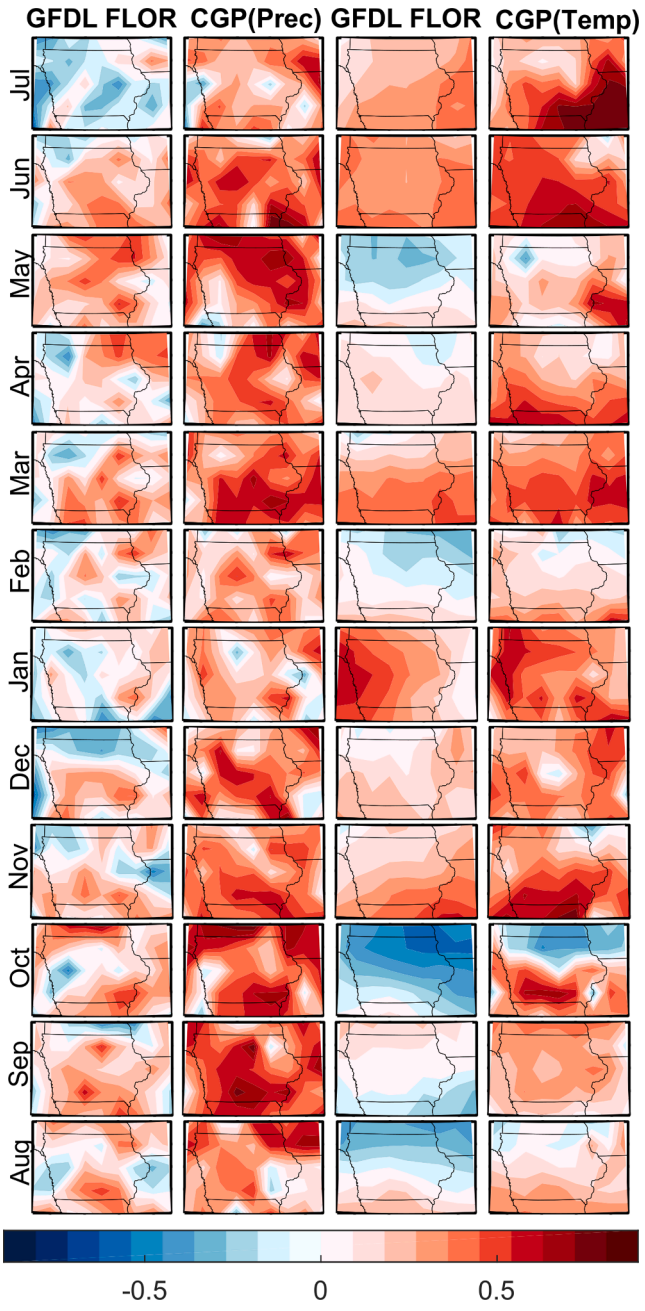

Fig. 6. Correlation coefficient between observed and forecast precipitation (first two columns) and temperature (last two columns) based on raw GFDL FLOR and after CGP for target month July. Moving from the top to the bottom row, the lead time increases from the shortest (i.e., July-target initialized at the beginning of July) to the longest (i.e., July-target initialized in August of the previous year). These results are for the validation period 2006–2019.

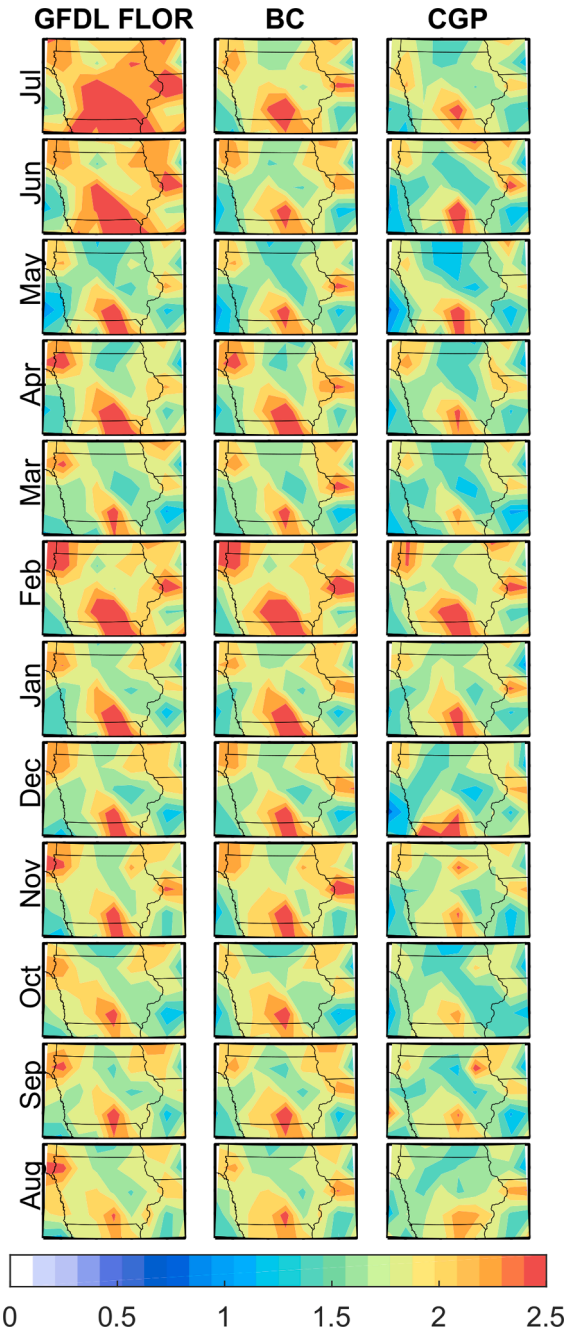

Fig. 7. Root mean square error (units: mm/day) between observed and forecast precipitation based on raw GFDL FLOR and after BC, and after CGP for target month July. Moving from the top to the bottom row, the lead time increases from the shortest (i.e., July-target initialized at the beginning of July) to the longest (i.e., July-target initialized in August of the previous year). These results are for the validation period 2006–2019.

these improvements, there are still residual errors that are not corrected for.

- The application of the CGP method leads to a consistent improved performance with respect to the BC approach, with no strong dependence on lead time. Therefore, our approach highlights the improvement due to considering simultaneously uncertainties in both precipitation and temperature, rather than in isolation from each other.
- While these results are very promising and provide a path forward towards reducing uncertainties in seasonal forecasting, there are a number of possible directions that could lead to an even improved performance. For example, here we considered one model and the average of its 12 members; future efforts could be devoted to the extension of the CGP approach towards ensemble averaging. Moreover, here we developed separate models for each pixel: in the future, we could develop a hierarchical approach to leverage information across neighboring pixels. Finally, this approach could be expanded to other target months and regions with different climates to examine the transferability of these results in space and time.

CRediT authorship contribution statement

Chao Wang: Conceptualization, Methodology, Software, Formal

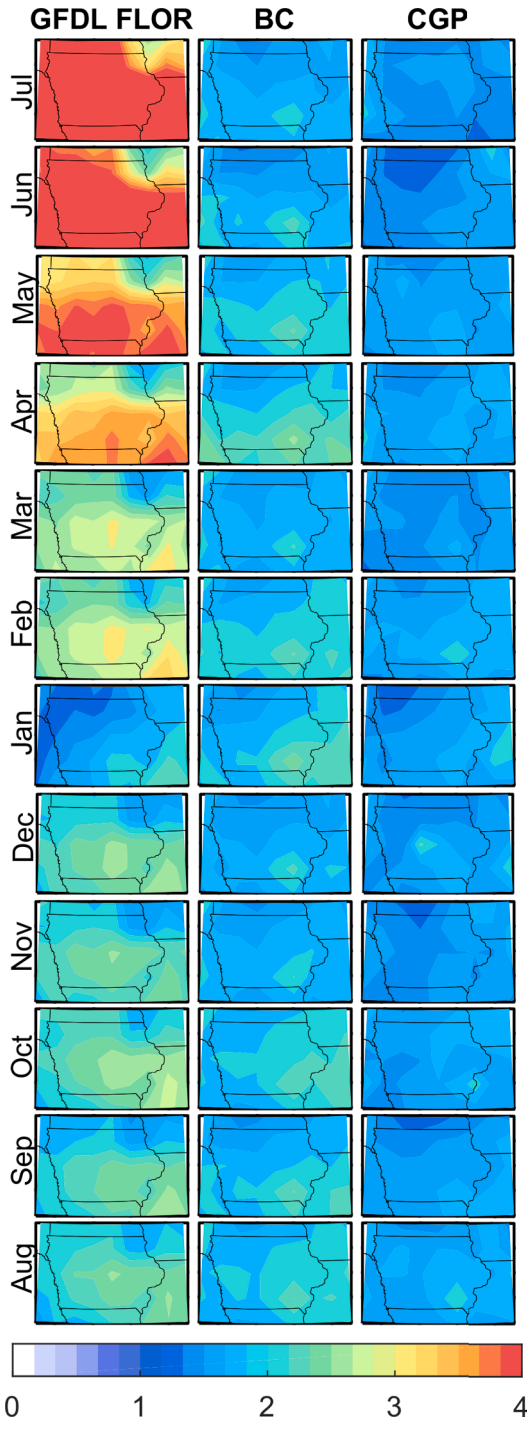

Fig. 8. Root mean square error (units: °C) between observed and forecast temperature based on raw GFDL FLOR and after BC, and after CGP for target month July. Moving from the top to the bottom row, the lead time increases from the shortest (i.e., July-target initialized at the beginning of July) to the longest (i.e., July-target initialized in August of the previous year). These results are for the validation period 2006–2019.

analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision. **Wei Zhang:** Software, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Gabriele Villarini:** Conceptualization, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This material is based in part upon work supported by the U.S. Army Engineer Institute for Water Resources (IWR) and IIHR—Hydroscience & Engineering.

References

Archambault, H., Huang, J., Kirtman, B., Mariotti, A., Villarini, G., 2019. Special issue: NMME. Clim. Dyn. 53 (12) https://doi.org/10.1007/s00382-019-05028-0, 7151-7151.

Becker, E., van den Dool, H., Zhang, Q., 2014. Predictability and forecast skill in NMME. J. Clim. 27 (15), 5891–5906. https://doi.org/10.1175/jcli-d-13-00597.1.

Boyle, P., Frean, M., 2005. Dependent Gaussian Processes. 217–224.

Cash, B.A., Manganello, J.V., Kinter, J.L., 2019. Evaluation of NMME temperature and precipitation bias and forecast skill for South Asia. Clim. Dyn. 53 (12), 7363–7380. https://doi.org/10.1007/s00382-017-3841-4.

Cohen, J., et al., 2019. S2S reboot: an argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. Wiley Interdisciplinary Reviews-Climate Change 10 (2). https://doi.org/10.1002/wcc.567.

Daly, C., et al., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol. 28 (15), 2031–2064. https://doi.org/10.1002/joc.1688.

DelSole, T., Tippett, M.K., 2014. Comparing forecast skill. Monthly Weather Rev. 142 (12), 4658–4678. https://doi.org/10.1175/mwr-d-14-00045.1.

Hagedorn, R., Doblas-Reyes, F.J., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. Tellus Series a-Dyn. Meteorol. Oceanogr. 57 (3), 219–233. https://doi.org/10.1111/j.1600-0870.2005.00103.x.

Harnos, D.S., Schemm, J.K.E., Wang, H., Finan, C.A., 2019. NMME-based hybrid prediction of Atlantic hurricane season activity. Clim. Dyn. 53 (12), 7267–7285. https://doi.org/10.1007/s00382-017-3891-7.

Hervieux, G., et al., 2019. More reliable coastal SST forecasts from the North American multimodel ensemble. Clim. Dyn. (12), 7153–7168. https://doi.org/10.1007/s00382-017-3652-7.

Higdon, D., 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environ. Ecol. Statistics 5 (2), 173–190. https://doi.org/10.1023/a:1009666805688.

Infanti, J.M., Kirtman, B., 2016. North American rainfall and temperature prediction response to the diversity of ENSO. Clim. Dyn. 46 (9–10), 3007–3023. https://doi.org/10.1007/s00382-015-2749-0.

Jia, L., et al., 2014. Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. J. Clim. 28 (5), 2044–2062. https://doi.org/10.1175/JCLI-D-14-00112.1.

Kam, J.H., Sheffield, J., Yuan, X., Wood, E.F., 2014. Did a skillful prediction of sea surface temperatures help or hinder forecasting of the 2012 Midwestern US drought? Environ. Res. Lett. 9 (3) https://doi.org/10.1088/1748-9326/9/3/034005.

Kang, D., Lee, M.I., 2019. ENSO influence on the dynamical seasonal prediction of the East Asian Winter Monsoon. Clim. Dyn. 53 (12), 7479–7495. https://doi.org/10.1007/s00382-017-3574-4.

Khajehei, S., Ahmadalipour, A., Moradkhani, H., 2018. An effective post-processing of the North American multi-model ensemble (NMME) precipitation forecasts over the continental US. Clim. Dyn. 51 (1), 457–472. https://doi.org/10.1007/s00382-017-3934-0.

Khouakhi, A., Villarini, G., Zhang, W., Slater, L.J., 2019. Seasonal predictability of high sea level frequency using ENSO patterns along the US West Coast. Adv. Water Resour. 131 https://doi.org/10.1016/j.advwatres.2019.07.007.

Kirtman, B.P., et al., 2014. THE NORTH AMERICAN MULTIMODEL ENSEMBLE Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. Bull. Am. Meteorol. Soc. 95 (4), 585–601. https://doi.org/10.1175/bams-d-12-00050.1.

Krakauer, N.Y., 2019. Temperature trends and prediction skill in NMME seasonal forecasts. Clim. Dyn. 53 (12), 7201–7213. https://doi.org/10.1007/s00382-017-3657-2.

Ma, F., et al., 2016. Evaluating the skill of NMME seasonal precipitation ensemble predictions for 17 hydroclimatic regions in continental China. Int. J. Climatol. 36 (1), 132–144. https://doi.org/10.1002/joc.4333.

Ma, F., Ye, A.Z., Duan, Q.Y., 2019. Seasonal drought ensemble predictions based on multiple climate models in the upper Han River Basin, China. Clim. Dyn. 53 (12), 7447–7460. https://doi.org/10.1007/s00382-017-3577-1.

Manganello, J.V., Cash, B.A., Hodges, K.I., Kinter, J.L., 2019. Seasonal forecasts of North Atlantic tropical cyclone activity in the North American Multi-Model Ensemble. Clim. Dyn. 53 (12), 7169–7184. https://doi.org/10.1007/s00382-017-3670-5.

Maraun, D., 2016. Bias correcting climate change simulations - a critical review. Curr. Clim. Change Rep. 2 (4), 211–220. https://doi.org/10.1007/s40641-016-0050-x.

Mo, K.C., Lyon, B., 2015. Global meteorological drought prediction using the North American multi-model ensemble. J. Hydrometeorol. 16 (3), 1409–1424. https://doi.org/10.1175/jhm-d-14-0192.1.

Narapusetty, B., Collins, D.C., Murtugudde, R., Gottschalck, J., Peters-Lidard, C., 2018. Bias correction to improve the skill of summer precipitation forecasts over the contiguous United States by the North American multi-model ensemble system. Atmos. Sci. Lett. 19 (5), e818 https://doi.org/10.1002/asl.818.

Quinonero-Candela, J.Q., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. J. Mach. Learning Res. 6, 1939–1959.

Rasmussen, C.E., 2003. Gaussian processes in Machine Learning, Summer School on Machine Learning. Springer, Berlin, Heidelberg.

Rodrigues, L.R.L., Doblas-Reyes, F.J., Coelho, C.A.S., 2019. Calibration and combination of monthly near-surface temperature and precipitation predictions over Europe. Clim. Dyn. 53 (12), 7305–7320. https://doi.org/10.1007/s00382-018-4140-4.

Roundy, J.K., Yuan, X., Schaake, J., Wood, E.F., 2015. A framework for diagnosing seasonal prediction through canonical event analysis. Monthly Weather Rev. 143 (6), 2404–2418. https://doi.org/10.1175/mwr-d-14-00190.1.

Shin, C.S., Huang, B.H., 2019. A spurious warming trend in the NMME equatorial Pacific SST hindcasts. Clim. Dyn. 53 (12), 7287–7303. https://doi.org/10.1007/s00382-017-3777-8.

Singh, A., Sahoo, R.K., Nair, A., Mohanty, U.C., Rai, R.K., 2017. Assessing the performance of bias correction approaches for correcting monthly precipitation over India through coupled models. Meteorol. Appl. 24 (3), 326–337. https://doi.org/10.1002/met.1627.

Slater, L.J., Villarini, G., Bradley, A.A., 2017. Weighting of NMME temperature and precipitation forecasts across Europe. J. Hydrol. 552, 646–659. https://doi.org/10.1016/j.jhydrol.2017.07.029.

Slater, L.J., Villarini, G., 2018. Enhancing the predictability of seasonal streamflow with a statistical-dynamical approach. Geophys. Res. Lett. 45 (13), 6504–6513. https://doi.org/10.1029/2018gl077945.

Slater, L.J., Villarini, G., Bradley, A.A., 2019a. Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA. Clim. Dyn. 53 (12), 7381–7396. https://doi.org/10.1007/s00382-016-3286-1.

Slater, L.J., Villarini, G., Bradley, A.A., Vecchi, G.A., 2019b. A dynamical statistical framework for seasonal streamflow forecasting in an agricultural watershed. Clim. Dyn. 53 (12), 7429–7445. https://doi.org/10.1007/s00382-017-3794-7.

Thober, S., et al., 2015. Seasonal Soil Moisture Drought Prediction over Europe Using the North American Multi-Model Ensemble (NMME). J. Hydrometeorol. 16 (6), 2329–2344. https://doi.org/10.1175/jhm-d-15-0053.1.

Tian, D., Martinez, C.J., Graham, W.D., Hwang, S., 2014. Statistical downscaling multimodel forecasts for seasonal precipitation and surface temperature over the Southeastern United States. J. Clim. 27 (22), 8384–8411. https://doi.org/10.1175/Jcli-D-13-00481.1.

Tippett, M.K., Ranganathan, M., L'Heureux, M., Barnston, A.G., DelSole, T., 2019. Assessing probabilistic predictions of ENSO phase and intensity from the North American Multimodel Ensemble. Clim. Dyn. 53 (12), 7497–7518. https://doi.org/10.1007/s00382-017-3721-y.

Vecchi, G.A., et al., 2014. On the seasonal forecasting of regional tropical cyclone activity. J. Clim. 27 (21), 7994–8016. https://doi.org/10.1175/jcli-d-14-00158.1.

Villarini, G., Luitel, B., Vecchi, G.A., Ghosh, J., 2019. Multi-model ensemble forecasting of North Atlantic tropical cyclone activity. Clim. Dyn. 53 (12), 7461–7477. https://doi.org/10.1007/s00382-016-3369-z.

Vitart, F., et al., 2007. Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. Geophys. Res. Lett. 34 (16) https://doi.org/10.1029/2007gl030740.

Vittal, H., Villarini, G., Zhang, W., 2020. Early prediction of the Indian summer monsoon rainfall by the Atlantic Meridional Mode. Clim. Dyn. https://doi.org/10.1007/s00382-019-05117-0.

Xu, L., et al., 2019. Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning. Clim. Dyn. 53 (1–2), 601–615. https://doi.org/10.1007/s00382-018-04605-z.

Ying, Z., 1991. Asymptotic properties of a maximum-likelihood estimator with data from a Gaussian process. J. Multivariate Anal. 36 (2), 280–296. https://doi.org/10.1016/0047-259x(91)90062-7.

Zhang, W., Villarini, G., Slater, L., Vecchi, G.A., Bradley, A.A., 2017. Improved ENSO Forecasting Using Bayesian Updating and the North American Multimodel Ensemble (NMME). J. Clim. 30 (22), 9007–9025. https://doi.org/10.1175/jcli-d-17-0073.1.

Zhang, W., Villarini, G., Vecchi, G.A., 2019. Impacts of the Pacific meridional mode on rainfall over the maritime continent and Australia: potential for seasonal predictions. Clim. Dyn. 53 (12), 7185–7199. https://doi.org/10.1007/s00382-017-3968-3.