Multimodel Subseasonal Precipitation Forecasts over the Contiguous United States: Skill Assessment and Statistical Postprocessing

YANZHONG LI,[a,b] DI TIAN,[b] AND HANOI MEDINA[b]

[a] *School of Hydrology and Water Resources, Nanjing University of Information Science and Technology, Nanjing, China*
[b] *Department of Crop, Soil and Environmental Sciences, Auburn University, Auburn, Alabama*

(Manuscript received 12 February 2021, in final form 20 July 2021)

ABSTRACT: This study assessed multimodel subseasonal precipitation forecasts (SPFs) from eight subseasonal experiment (SubX) models over the contiguous United States (CONUS) and explored the generalized extreme value distribution (GEV)-based ensemble model output statistics (EMOS) framework for postprocessing multimodel ensemble SPF. The results showed that the SubX SPF skill varied by location and season, and the skill was relatively high in the western coastal region, north-central region, and Florida peninsula. The forecast skill was higher during winter than summer seasons, especially for lead week 3 in the northwest region. While no individual model consistently outperformed the others, the simple multimodel ensemble (MME) demonstrated a higher skill than any individual model. The GEV-based EMOS approach dramatically improved the MME subseasonal precipitation forecast skill at long lead times. The continuous ranked probability score (CRPS) was improved by approximately 20% in week 3 and 43% in lead week 4; the 5-mm Brier skill score (BSS) was improved by 59.2% in lead week 3 and 50.9% in lead week 4, with the largest improvements occurring in the northwestern, north-central, and southeastern CONUS. Regarding the relative contributions of the individual SubX model to the predictive skill, the NCEP model was given the highest weight at the shortest lead time, but the weight decreased dramatically with the increase in lead time, while the CESM, EMC, NCEP, and GMAO models were given approximately equal weights for lead weeks 2–4. The presence of active MJO conditions notably increased the forecast skill in the north-central region during weeks 3–4, while the ENSO phases influenced the skill mostly in the southern regions.

KEYWORDS: Ensembles; Forecast verification/skill; Probabilistic Quantitative Precipitation Forecasting (PQPF); Postprocessing

1. Introduction

Precipitation is arguably the most crucial component in the hydrologic cycle and atmospheric circulation. Its spatiotemporal variations directly affect the global water resources and energy balance (Donat et al. 2016; Q. Sun et al. 2018). Many natural disaster predictions can greatly benefit from accurate precipitation forecasting, such as early warnings of flash droughts and heavy rains, debris flows, and landslide hazards, and can inform decision-making processes for water resource management (Borga et al. 2014; Zhu et al. 2019; Feng and Chao 2020). However, accurate precipitation forecasting remains a challenging task because of its strong spatiotemporal heterogeneity. With advances in data assimilation, model physics, and ensemble forecasting techniques, numerical weather prediction and seasonal climate forecast have made great progress in recent decades, resulting in skillful precipitation forecasting with a 1–2-week lead time (Bauer et al. 2015) and improved seasonal predictions with a 1–12-month lead time (Yuan and Wood 2012). Weather forecasting is determined mostly by the initial atmospheric conditions (Vitart et al. 2008), whereas seasonal climate prediction is strongly influenced by atmosphere–ocean coupling processes and boundary conditions, such as sea surface temperature (Troccoli 2010), soil moisture (Koster et al. 2004), and stratosphere–troposphere interactions (Scaife et al. 2016). Subseasonal precipitation forecasts (SPF) are particularly challenging because the subseasonal time scale is too short to be favored by initial conditions and too long to be influenced by boundary conditions. However, SPF covers the critical time horizon of many hazardous events (White et al. 2015; Mo and Lettenmaier 2016), such as flash droughts, heat waves, and dry and wet spells. A skillful SPF will provide valuable information for early warnings and decision-making regarding natural hazard preparedness, risk reduction, and mitigation of socioeconomic loss.

To bridge this forecast gap, two subseasonal retrospective forecast (reforecast) datasets have been developed to date. One is the Subseasonal to Seasonal Prediction Project (S2S) launched by the World Weather Research Programme/World Climate Research Programme (Vitart et al. 2017), and the other is the newly launched Subseasonal Experiment (SubX), which was initiated by the U.S. National Oceanic and Atmospheric Administration (NOAA) climate testbed project (Pegion et al. 2019). Both datasets include multiple state-of-the-art numerical models that are devoted to advancing subseasonal forecasting (White et al. 2017; Wang and Robertson 2018). The multimodel ensemble has a higher skill level than most individual models since it is able to increase the forecast confidence by reducing system errors (Hagedorn et al. 2005; Medina et al. 2018). The SubX datasets provide not only subseasonal reforecast archives but also real-time forecasts from eight cutting-edge numerical weather prediction models. While SubX's ability to forecast subseasonal temperatures and precipitation has been investigated globally (Pegion et al. 2019),

Corresponding author: Di Tian, tiandi@auburn.edu

DOI: 10.1175/JHM-D-21-0029.1

SubX models and their multimodel ensemble have not been comprehensively evaluated in different climate regions over the contiguous United States (CONUS).

Since all models are a simplification of reality, systematic biases and uncertainties are inevitably introduced into model predictions (Murphy et al. 2004). Thus, postprocessing (sometimes also referred to calibration) of model forecasts has been widely adopted to address this issue (Kirtman et al. 2014; Li et al. 2017; Duan and Phillips 2010). Many statistical postprocessing methods have been developed to postprocess short- to medium-range precipitation forecasts by reducing systematic biases while improving the inadequate representation of forecasting uncertainties (e.g., Li et al. 2017; Scheuerer and Hamill 2015). In recent years, several statistical methods, such as Bayesian approaches, have been employed to postprocess subseasonal precipitation forecasts and have achieved improved performance (Specq and Batté 2020; Li et al. 2020). However, these techniques mostly focus on postprocessing single-model subseasonal forecasts, and they take the ensemble mean of subseasonal forecasts without fully considering their spread, which can reduce systematic errors to some extent but may not fully improve the adequate representation of forecasting uncertainties. In contrast, the generalized extreme value distribution (GEV)-based ensemble model output statistics (EMOS) postprocessing scheme (Scheuerer 2014) has the advantage of being able to efficiently postprocess multimodel ensemble (MME) forecasts considering both the mean and the spread of multimodel ensemble forecasts. Although short- to medium-range predictions have been successfully postprocessed by EMOS (Scheuerer and Hamill 2015; Vigaud et al. 2017, 2018, 2020; Medina and Tian 2020), a study focusing on postprocessing both the mean and spread of multimodel ensemble subseasonal forecasts by EMOS is still missing in the current literature.

El Niño–Southern Oscillation (ENSO) and the Madden–Julian oscillation (MJO) are large-scale phenomena that provide sources of subseasonal predictability with great impacts on subseasonal precipitation over different regions and seasons (viz., teleconnections) (Jones et al. 2011; Li and Robertson 2015; Tian et al. 2017; DelSole et al. 2017; Kim et al. 2019; Pan et al. 2019). For example, most S2S models showed high prediction skill levels in Southern California during El Niño years, while during La Niña years, most models showed better subseasonal skill levels in Oregon. Active MJO events can change the spatial pattern of precipitation (Pan et al. 2019). Over the Maritime Continent area, the MJO can modulate precipitation when ENSO is in a neutral year, while this effect can be weakened by strong ENSO events (Li and Robertson 2015). Given the teleconnections from MJO and ENSO, studies have detected their notable impacts on subseasonal precipitation forecast skill levels, focusing either on evaluating a single model (Tian et al. 2017) or on a specific region on the coast of the western U.S. (Pan et al. 2019). As a step forward, this study further examines the importance of the MJO and ENSO to the skill level of the multimodel ensemble (MME) subseasonal precipitation forecast (SPF) over the CONUS.

The aims of this paper are 1) to comprehensively evaluate the SPF skill levels of individual and MME forecasts from SubX in different climatic regions over the CONUS, 2) to explore the potential of using EMOS method for postprocessing the MME SPFs, and 3) to examine the effects of ENSO and MJO on the MME SPFs skill level over the CONUS. This work aims to postprocess multimodel ensemble subseasonal precipitation forecasts, but also provides a systematic skill evaluation as well as examining the effects of ENSO and MJO on the multimodel ensemble subseasonal precipitation forecast skill over the CONUS.

2. Data and methodology

a. Study area

Our study area covers nine climate regions across the CONUS (Fig. 1). These nine climate regions were defined by the National Centers for Environmental Information (NCEI) (Karl and Koss 1984) as follows: Northwest (NW), West (W), West North Central (WNC), Southwest (SW), Upper Midwest (UMW), Central (C), South (S), Northeast (NE), and Southeast (SE) regions. The topography and climate conditions vary significantly among these regions.

b. Dataset

In this study, we used the SubX precipitation retrospective forecast (reforecast) dataset at $1° \times 1°$ over the CONUS (Pegion et al. 2019), which is available at http://iridl.ldeo.columbia.edu/SOURCES/. Table 1 shows the basic information of the SubX data, which includes reforecast runs from eight state-of-the-art subseasonal numerical forecast models. Considering the availability of data, we chose the study period of 1999–2014 and a lead time of 1–4 weeks. The weekly mean of daily precipitation reforecast was produced by averaging daily precipitation over each week. Since the eight SubX models had different initialization dates and forecast lead times (Table 1), we constructed multimodel ensembles by combining all the forecasts initialized on different but closest dates over the course of 1 week, and the multimodel ensembles were verified for the same week. This type of processing is called a lagged average ensemble and has been widely applied in weather and seasonal forecasting (DelSole et al. 2017). Figure 2 provides a demonstration of the lagged average ensemble, taking the SubX CESE-30LCESM1, EMC-GEFS, and NRL-NESM datasets as an example. For the CESE-30LCESM1 dataset, since the forecast is initiated every week, matching with observations, the weekly mean of daily precipitation forecast is obtained by averaging daily precipitation over all seven days for each week. For EMC-GEFS and NRL-NESM, since the initialization time is different, there are less than seven days that fall within the same time period as CESE-30LCESM1. We search out the forecast initialization time that is closest to CESE-30LCESM1, and then obtain the weekly mean of daily precipitation over the same time period as CESE-30LCESM1. In this case, the weekly value is made up of less than seven daily values, since the model run does not cover the full week. While this might slightly influence the evaluation results, this allows us to obtain consistent subseasonal forecast from SubX model datasets for the follow-up multimodel ensemble

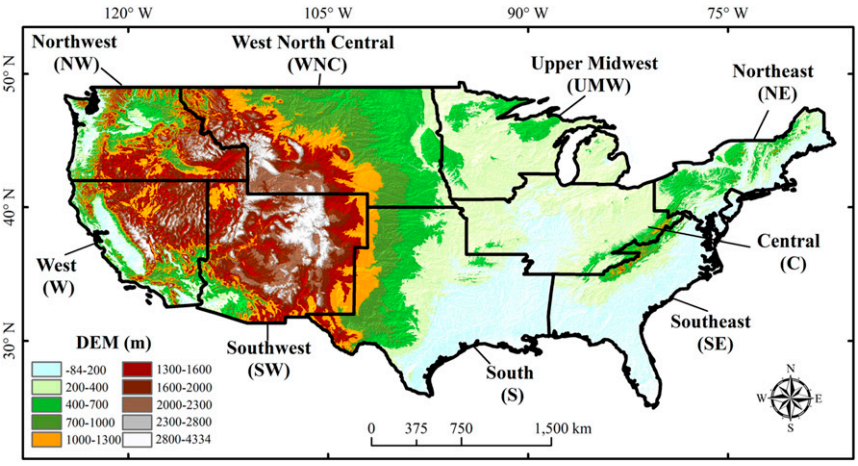

FIG. 1. Nine climatic regions over the contiguous United States.

forecasting and EMOS postprocessing. The multimodel ensemble (MME) forecast at the weekly time scale was generated by simply combining all the ensemble mean forecasts of each SubX model, resulting in eight-member ensemble forecasts. Simple MME is a common approach that has been widely employed in previous studies (e.g., Pincus et al. 2008; Wanders and Wood 2016; Ma et al. 2018; Pegion et al. 2019)

The observation-based daily precipitation of the Global Precipitation Climatology Project (GPCP 1dd) version 1.3 at a 1° grid during the 1999–2014 period was used as the observational data for calibration and verification. This database is constructed by obtaining instantaneous precipitation from the threshold matched precipitation index (TMPI) for 40°N–40°S threshold matched geo-IR dataset from the IR brightness temperatures and rescaling the TOVS precipitation estimates at higher latitudes (Huffman et al. 2001). Until now, it was reported that this dataset presented an excellent performance level that was able to meet the initialization requirement of numerical models, to drive land surface models, to resolve the advance and retreat of precipitation, and to validate model forecasts (Vigaud et al. 2017; S. Sun et al. 2018).

The MJO index produced by the Australian Bureau of Meteorology from 1999 to 2014 was used to assess the impact of this phenomenon on the predictions (http://www.bom.gov.au/climate/mjo/graphics/rmm.74toRealtime.txt). This index is derived from the two major principal components of the combined fields of near-equatorially averaged 850-hPa zonal wind, 200-hPa zonal wind, and outgoing longwave radiation (OLR) from an empirical orthogonal function analysis (Wheeler and Hendon 2004). The OLR is a proxy for convection and can effectively capture the baroclinic, convectively driven circulation in the equatorial plane of the MJO (Madden and Julian 1972). The OLR has been carefully validated and improved for inclusion in the latest version of ACCESS (Bi et al. 2013; Sun et al. 2013). These two leading principal components at a daily time step, namely, real-time multivariate MJO series 1 (RMM1) and 2 (RMM2), defined eight MJO phases and an MJO amplitude. We first evaluated the impact of MJO on prediction skill with RMM threshold of 1.0, but there was no significant improvement for most areas. After testing different RMM thresholds ranging from 1.0 to 1.5, we found that the threshold value 1.5, indicating strongest MJO events, is appropriate to substantially detect the impact of MJO on prediction skill. Therefore, we finally defined RMM amplitudes exceeding (below) 1.5 as active (no-active) phases. In total, for the study period of 834 weeks, 269 weeks are belonging to active MJO.

The ENSO index provided by the NOAA Climate Prediction Center (https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php) was derived from the monthly

TABLE 1. Basic information of the eight subseasonal datasets. Note: ins means that model will be initialized four times every 7 days.

ID	Model names	Ensemble members	Initial intervals	Lead time (days)	Timespan	References
1	CESM-30LCESM1	10	7 days	45	1 Jan 1999 to 30 Dec 2015	Hurrell et al. (2013)
2	EMC-GEFS	11	7 days	35	6 Jan 1999 to 30 Dec 2015	Zhou et al. (2017), Wei et al.(2008)
3	ESRL-FIM	4	7 days	32	6 Jan 1999 to 30 Dec 2015	Bleck et al. (2015)
4	ECCC-GEM	4	7 days	32	1 Jan 1999 to 28 Dec 2014	Lin et al. (2016)
5	GMAO-GEOS	4	5 days	45	1 Jan 1999 to 30 Dec 2015	Reichle and Liu (2014)
6	NCEP-CFSv2	4	1 days	45	1 Jan 1999 to 29 Dec 2015	Tian et al. (2017)
7	NRL-NESM	1	4 ins/7 days	45	2 Jan 1999 to 29 Dec 2015	Hogan et al. (2014)
8	RSMAS-CCSM4	3	2 ins/7 days	45	6 Jan 1999 to 30 Dec 2015	Gent et al. (2011)

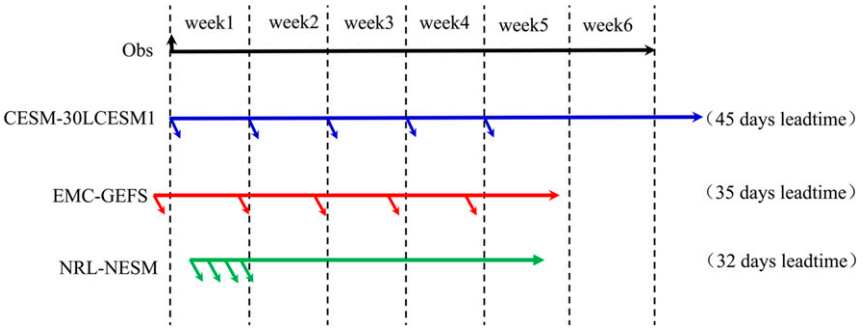

FIG. 2. A demonstration of the lagged average ensemble, taking the SubX CESE-30LCESM1, EMC-GEFS, and NRL-NESM datasets as an example.

sea surface temperature (SST) anomalies in Niño-3.4 in the area of 170°–120°W, 5°S–5°N. The ENSO data series was divided into three categories, including the El Niño phase, La Niña phase, and ENSO neutral, as defined by a standard threshold of ±0.5°C. As the initial days of SubX models are fixed, we divided the initial time lists into three categories by determining whether they fall into a specific phase of ENSO. There are 143, 392, and 299 sample days in El Niño, La Niña, and ENSO neutral, respectively.

c. *EMOS postprocessing methods*

It is more difficult to formulate a suitable probabilistic model for quantitative precipitation compared to other variables because the precipitation is nonnegative, may be equal to zero with positive probability, and positively skewed for a nonzero component. To satisfy these particular characteristics of precipitation, the generalized extreme distribution (GEV)-based EMOS method was developed by Scheuerer (2014). This novel approach can be considered as an ideal candidate for depicting predictive distributions for quantitative precipitation thanks to its reasonable choices of location, scale, and shape parameters. Therefore, in this study, we used the GEV-based EMOS method (Scheuerer 2014) to postprocess the MME subseasonal precipitation forecasts at a weekly time scale.

As the MME only provides finite ensembles, it cannot provide a full predictive density function (PDF), while the EMOS method can make use of discrete forecast members to produce a predictive PDF. Since precipitation has a special distribution with different features, i.e., nonnegative, equal to zero with positive probability, and positive skew, in the GEV EMOS framework, the PDF is expressed as a left-censored GEV distribution, i.e., $\mathrm{GEV}_0(m, \sigma, \xi)$, where m is the mean, σ is the scale factor, and ξ is the shape factor. The GEV_0 model's mean m is parameterized as follows:

$$m = a + b_1 X_1 + \cdots + b_m X_m + s p_0 . \tag{1}$$

The scale parameter σ is parameterized as follows:

$$\sigma = c + d \times \mathrm{MD}(X_1, \ldots, X_m). \tag{2}$$

In Eqs. (1) and (2), $X_1, X_2, \ldots, X_m$ denote the means of each model's ensemble forecast for a given location and time, p_0 is the fraction of zero precipitation members, $\mathrm{MD}(X_1, \ldots, X_m)$ is Gini's ensemble mean difference, and $a, b_1, \ldots, b_m, s, c, d$, and ξ are fitted to optimize the mean continuous ranked probability score (CRPS) between subseasonal precipitation forecasts and observations over the specified training period, where $b_1, \ldots, b_m$, are coefficients representing the weights for each model's ensemble mean forecast. Using the ensemble mean of each model $(X_1, X_2, \ldots, X_m)$ helps reveal the weights of each model $(b_1, \ldots . b_m)$, and therefore, it helps us understand their contributions to the predictive skill.

The GEV has the cumulative density function (CDF) is as follows:

$$G(y) := \begin{cases} \exp\left\{ -\left[1 + \xi\left(\dfrac{y-\mu}{\sigma}\right)\right]^{-1/\xi} \right\} & \text{for} \quad \xi \neq 0 \\ \exp\left[-\exp\left(\dfrac{y-\mu}{\sigma}\right) \right] & \text{for} \quad \xi = 0 \end{cases}, \tag{3}$$

where parameters μ, σ, and ξ characterize the location, scale, and shape of the GEV. As in Scheuerer (2014), when $\xi \in (-0.278, 1)$, the GEV has a positive skew, and its mean is equal to the following:

$$m = \begin{cases} \mu + \sigma \dfrac{\Gamma(1-\xi) - 1}{\xi} & \text{for} \quad \xi \neq 0 \\ \mu + \sigma\gamma & \text{for} \quad \xi = 0 \end{cases} \tag{4}$$

where Γ denotes the gamma function and $\gamma \approx 0.5772$ is the Euler–Mascheroni constant. For modeling precipitation amounts, GEV is considered to be left-censored at zero, namely, a value below zero is assigned to a zero value. The predictive CDF then becomes the following:

$$\tilde{G}(y) := \begin{cases} G(y) & \text{for} \quad y \geq 0 \\ 0 & \text{for} \quad y < 0 \end{cases} . \tag{5}$$

If either $\xi \leq 0$ or $\xi > 0$ and $\mu < (\sigma/\xi)$, this distribution is nonnegative and exactly zero with a positive probability. This left-censored GEV distribution (GEV_0) permits modeling precipitation on the original scale without prior transformation of the data. Specifically, in this study, the ensembleMOS package in "R" was used to implement the EMOS method considering the GEV_0 probability distribution as a predictive function

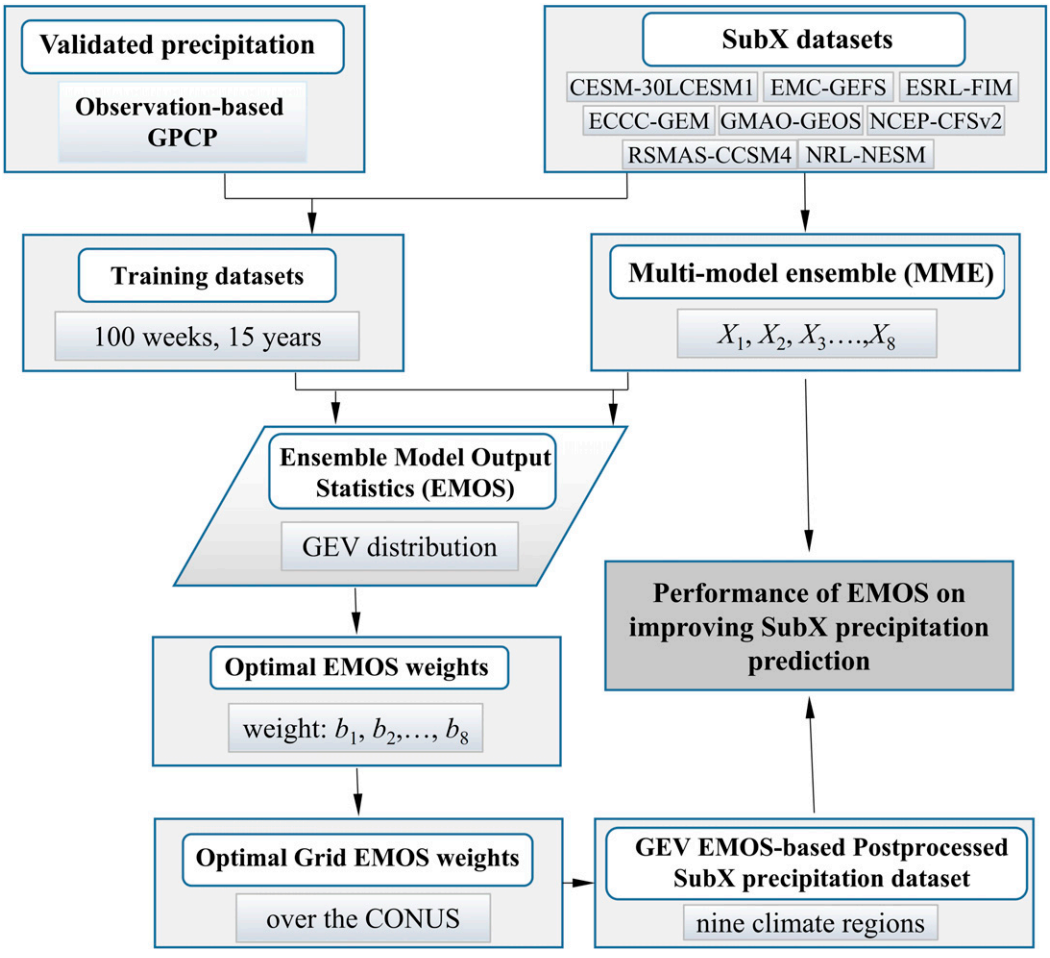

FIG. 3. Schematic diagram of the generalized extreme distribution (GEV) based EMOS approach for postprocessing SubX precipitation over the CONUS.

(Baran and Nemoda 2016). The detailed schematic of the GEV-based EMOS approach for postprocessing eight SubX precipitation datasets over the CONUS is shown in Fig. 3.

d. Forecast evaluation

We evaluated both deterministic and probabilistic forecasts of each SubX model, MME, and EMOS postprocessed forecasts in nine climate regions across the CONUS. To evaluate the deterministic forecast skill, two metrics, the Pearson correlation coefficient (CORR) and Kling–Gupta efficiency (KGE), were employed. The CORR, with a range from 0 to 1, is a common metric describing the strength and direction of the linear relationship between prediction and observation. The KGE is an integrated performance statistic incorporating linear correlation, relative bias, and variability (Gupta et al. 2009), and it spans from $-\infty$ to 1, with an expected value of 1 indicating perfect performance. A KGE value greater than -0.41 indicates that using a model forecast is better than using the mean value (Knoben et al. 2019). The equations of the two statistical metrics are shown as follows:

$$\mathrm{CORR} = \frac{\sum_{i=1}^{N}(y_{\mathrm{obs},i} - \overline{y_{\mathrm{obs}}})(y_{\mathrm{pred},i} - \overline{y_{\mathrm{pred}}})}{\sqrt{\sum_{i=1}^{N}(y_{\mathrm{obs},i} - \overline{y_{\mathrm{obs}}})^2 \sum_{i=1}^{N}(y_{\mathrm{pred},i} - \overline{y_{\mathrm{pred}}})^2}} \tag{6}$$

$$\mathrm{KGE} = 1 - \sqrt{(1-\mathrm{CORR})^2 + \left(1 - \frac{\sigma_{\mathrm{pred}}}{\sigma_{\mathrm{obs}}}\right)^2 + \left(1 - \frac{\mu_{\mathrm{pred}}}{\mu_{\mathrm{obs}}}\right)^2} \tag{7}$$

where $y_{\mathrm{obs},i}$ and $y_{\mathrm{pred},i}$ refer to the observation and prediction at the ith week, $\overline{y_{\mathrm{obs}}}$ and $\overline{y_{\mathrm{pred}}}$ refer to the mean value of all weekly precipitation, and N is the number of total weeks from 1999 to 2014. The terms σ_{obs} and μ_{obs} denote the standard deviation and mean of observational precipitation, respectively, and σ_{pred} and μ_{pred} denote the standard deviation and mean of the predicted weekly precipitation, respectively.

To evaluate the probabilistic forecast skill, the CRPS and the Brier skill score (BSS) were used in this study (Wilks 2011). The CRPS is a standard measure for evaluating the reliability and accuracy of probabilistic forecasts. It describes the integrals of

the squared difference between the cumulative distribution function (CDF) of the probabilistic forecast from MMEs and the CDF of the observed values (Wang and Robertson 2018; Mishra et al. 2018). The mean CRPS ($\overline{\text{CRPS}}$) for each grid point is calculated using the following equation,

$$\overline{\text{CRPS}} = \frac{1}{N}\text{CRPS} = \frac{1}{N}\int_R [F_{\text{pred}}(x) - F_{\text{obs}}(x)]^2\, dx \approx \frac{1}{N}\int_R [\hat{F}_{\text{pred}}(x) - F_{\text{obs}}(x)]^2\, dx, \quad (8)$$

where

$$F_{\text{obs}}(x) = \begin{cases} 0, x < \text{observed value} \\ 1, x > \text{observed value} \end{cases}, \quad (9)$$

where N is the total number of weeks, and F_{obs} and F_{pred} are the CDFs of the observed precipitation and the predicted precipitation, respectively. The term $\hat{F}_{\text{pred}}$ is an estimated F_{pred} produced by MME or EMOS. The CRPS is a negative orientation metric, where better performance is indicated by smaller values. If the predictive distribution is centered on the observed values, it indicates an accurate forecast.

The Brier score (BS) was used to assess the probabilistic forecast skill level for different precipitation events (Wilks 2011; De Felice et al. 2015; Bliefernicht et al. 2019). The BS is defined as follows:

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n} (f_i - o_i)^2, \quad (10)$$

where n is the total number of forecast-event pairs for a given tercile-based precipitation category; f_i is the ith forecast probability, which is calculated as the percentage of forecast members exceeding a certain threshold (e.g., 5-mm precipitation); and o_i is the ith observation probability and is a dichotomous event (0 or 1), i.e., if the observed value is greater than a certain threshold, $o_i = 1$; otherwise, $o_i = 0$. The BS is between 0 and 1 with a smaller value denoting better performance.

The BSS was used to assess the forecasting skill relative to climatology. The BSS can be defined as follows:

$$\text{BSS} = \frac{\text{BS} - \text{BS}_{\text{clm}}}{0 - \text{BS}_{\text{clm}}} = 1 - \frac{\text{BS}}{\text{BS}_{\text{clm}}}, \quad (11)$$

where BS_{clm} is the BS of the climatology and the BSS is between $-\infty$ and 1. BSS < 0, BSS $= 0$, and BSS > 0 indicate worse, the same, and better predictive skills than the climatology, respectively (Ma et al. 2017).

A "leave-one-year-out cross-validation" procedure was used to evaluate the skill of EMOS postprocessing (Wilks 2011). This method was robust in evaluating the skill of weather and climate forecasts and was widely used by researchers (Barnston and van den Dool 1993; Monhart et al. 2018). For example, there were 16 years of available forecasts in our study. When forecasting precipitation for a given week of a year, all other years of data with a window of $\pm n$ weeks, i.e., $2n$ (weeks yr^{-1}) $\times$ 15 (years) were used as the training dataset, and the forecast target was used as the evaluation data. This process was repeated from the first week of the year to the last data point. The value of n was the training period, which was determined by comparing the performance values of different periods from 30 to 130 weeks using the data from three randomly selected grid points in each climate region (total 27 sample grid points). In addition, the effects of MJO or ENSO were assessed by comparing the forecast skill during different MJO and ENSO phases.

e. *Significance test*

The significance level of prediction skill differences among the different SubX models, MME and EMOS for each of the nine climate regions, and the differences between the active and inactive MJO and the El Niño phase and La Niña phase were tested by using the one-way analysis of variance (ANOVA), the least significant difference (LSD) test, and the F test ($p < 0.05$).

3. Results and discussion

a. *Comparison of the SubX weekly precipitation forecasts with observations*

The intercomparison of SubX model forecasts and observations could reveal the discrepancies and consistencies of model forecasts in depicting the spatial–temporal distribution of subseasonal precipitation (Figs. 4 and 5). The spatial distribution of the weekly predicted precipitation of eight SubX models is shown in Fig. 4, and the statistical percentage of bias relative to observational precipitation is presented in Fig. 5a. While the average observational weekly precipitation is approximately 15 mm, it varies from more than 23.9 mm in the SE, with a subtropical humid climate, to less than 6.4 mm in the SW, with a plateau mountain climate (Fig. 4a9). In general, the SubX precipitation forecasts from weeks 1 to 4 followed a similar spatial distribution as that of the observations, all showing a decreasing trend from the southeastern to western CONUS, with the precipitation especially low in the west and southwest regions, but relatively higher on the west coast of the northwest region (Fig. 4). However, the eight models exhibited great discrepancies in the weekly precipitation forecasts across the CONUS. For week 1 (Figs. 4a0–a8 and 5), the EMC, ECCC, NCEP, and NRL overpredicted precipitation compared to the observations, with weekly mean values of 15.97, 16.36, 17.49, and 16.75 mm and relative biases of 3.8%, 4.9%, 16.1%, and 9.4%, respectively. The CESM, ESRL, GMAO, and RSMAS underpredicted precipitation, with mean values of 12.65, 15.24, 13.07, and 14.46 mm, and relative biases of −13.9%, −3.1%, −16.2%, and −0.5%, respectively. It is worth noting that the MME had the lowest bias (0.2%), indicating that the simple average of all the models notably decreased the bias compared to the individual model, which confirmed that the multimodel performed well with a smaller bias than any individual forecast model. For week 2, the EMC, ESRL, ECCC, and NCEP models overpredicted the precipitation; the CESM, GMAO, NRL, and RSMAS underpredicted the precipitation, with all except the NRL having bias in week 1. Compared to weeks 1–2, the precipitation for weeks 3–4 was highly overpredicted by most models, especially for the EMC (Fig. 4d2), ESRL (Fig. 4d3), and NCEP (Fig. 4d6). These large

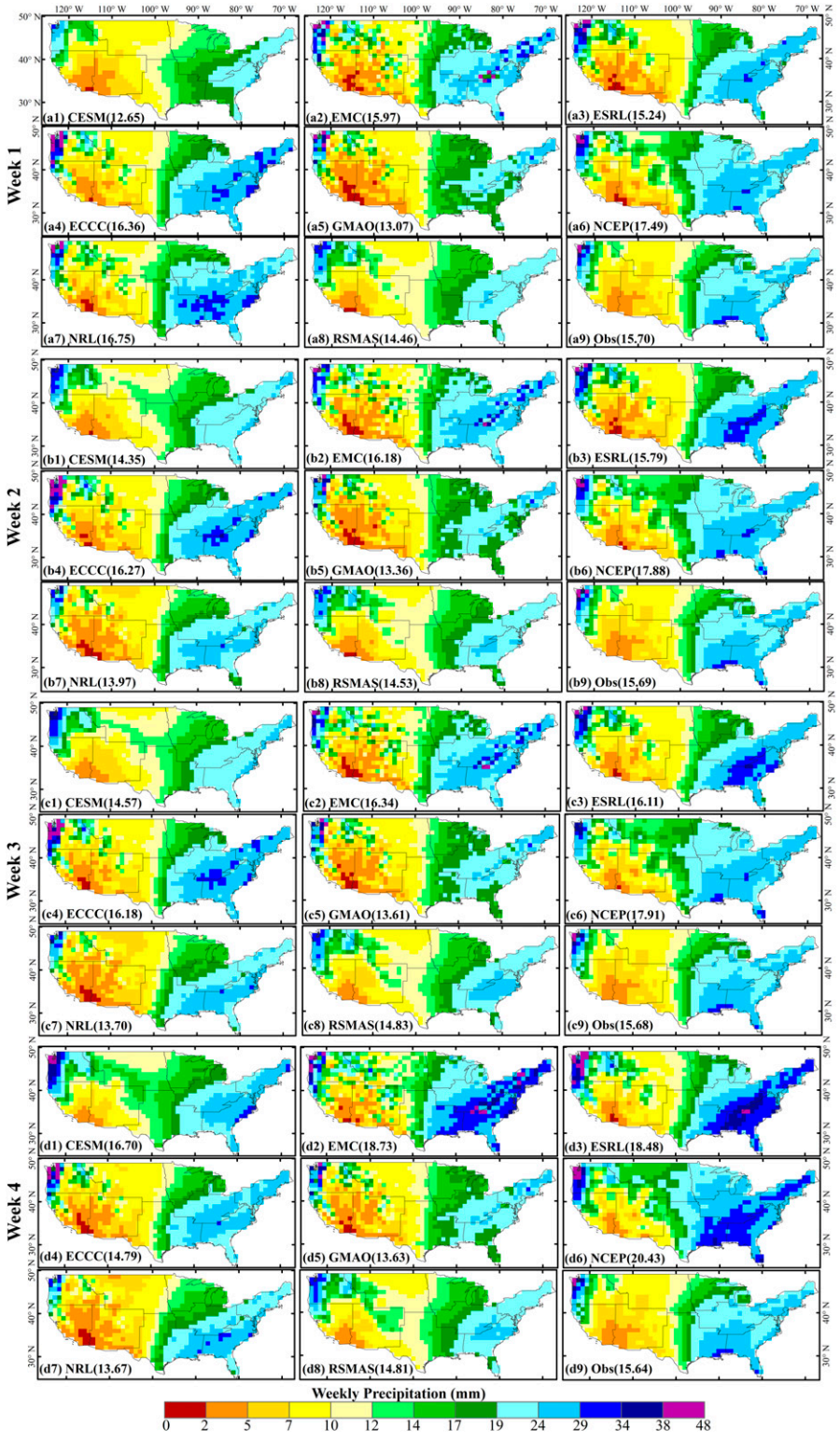

FIG. 4. Spatial pattern of weekly mean precipitation reforecasts for different lead times (weeks 1–4) from eight SubX models over the United States. The numbers in the parentheses are the mean weekly precipitation values over the United States.

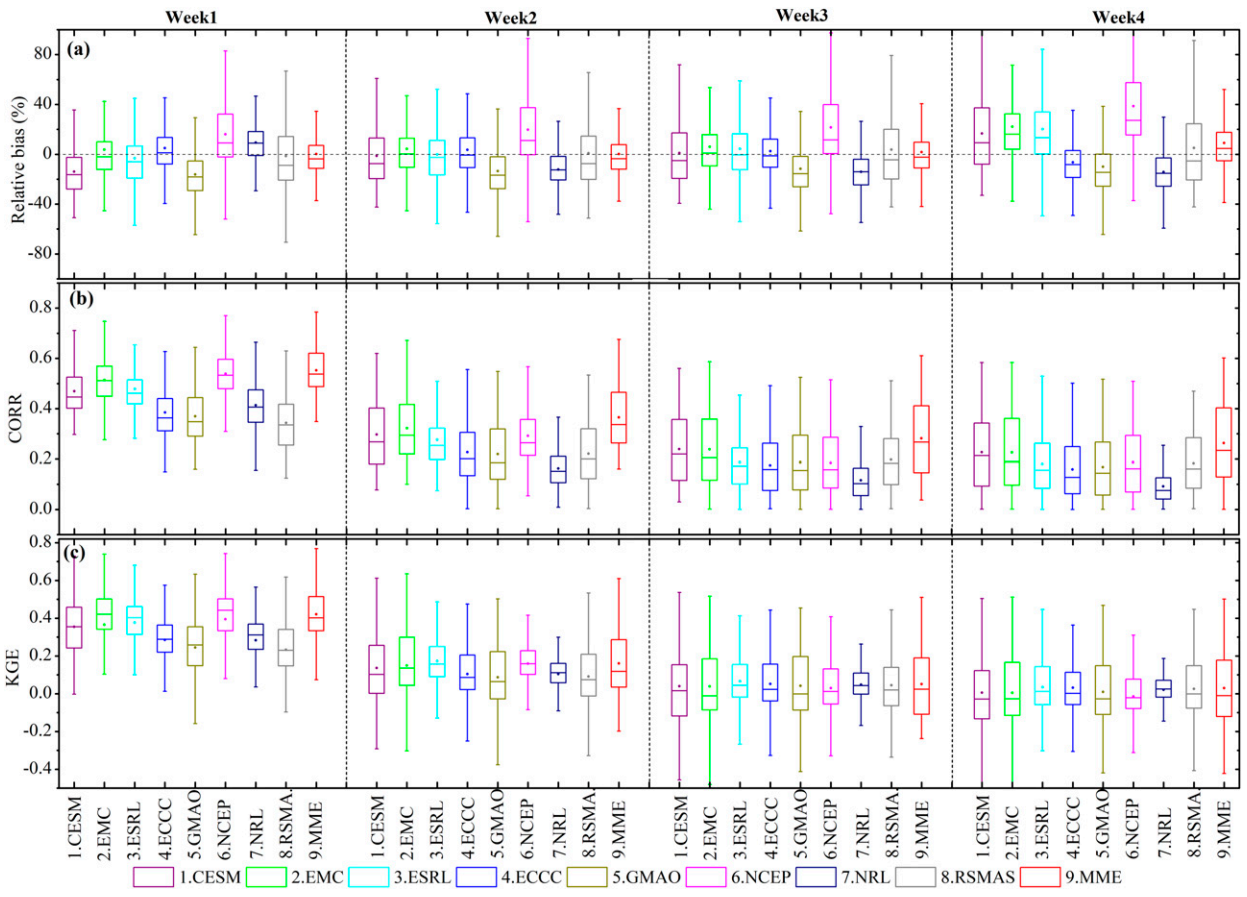

FIG. 5. Distribution of SubX precipitation reforecast: (a) relative bias, (b) correlation, and (c) KGE for four lead times (weeks 1–4). MME accounts for the average precipitation among the eight SubX models and in all seasons.

differences mainly occurred in the eastern United States, specifically in the southeast and northeast regions. Overall, the MME notably reduced the forecast biases compared to the individual model.

b. Deterministic forecast skill of individual model and MME forecasts from SubX

The SPF skills may vary by models, seasons, regions, and phases and magnitudes of large-scale circulation patterns. To compare the deterministic forecast skill of the SubX models, we evaluate the correlation of each SubX model with the observational precipitation for weeks 3–4 over different climate regions during winter and summer seasons (Fig. 6). The correlation skill in winter for week 3 was higher than that in summer, especially for CESM, EMC, and ESRL. These three models also showed similar skills in week 4, while the performance was much lower than that in week 3. This result was consistent with the findings of previous studies, which reported that SPF had a better performance in winter than in summer seasons (Tian et al. 2017; Vigaud et al. 2017).

Figure 6 also demonstrates that the prediction skill of the SubX models varied significantly by region. Most of the models showed satisfactory skill in the NW region in winter, especially the CESM model, while moderate skill was presented in the UMW, NE, S, and C regions. The high prediction skill in the NW region is probably because it has distinct synoptic and precipitation regimes than the eastern regions and is adjacent to the Pacific Ocean. The precipitation variation in this region is strongly affected by the ENSO or MJO through their influences on the Aleutian low, subtropical jets, and propagation of extratropical wave trains (Schonher and Nicholson 1989; Mo and Higgins 1998). A high prediction skill in the NW region was also found in Pan et al. (2019), who showed skillful precipitation forecasts beyond week 2 in this region, especially when the MJO was active.

Figures 7 and 5b show the correlations between the SPF and the observations during all seasons. For week 1 (Figs. 7a1–a8), the NCEP model showed the best performance among the eight models, with a mean CORR 0.54, followed by the EMC (CORR 0.51), ESRL (CORR 0.48), and CESM (CORR 0.47), with the NW and WNC regions showing higher skills in the SW and S. In contrast, the RSMAS, ECCC, and GMAO models did not represent precipitation variation well, especially in the SE region. However, the MME (Fig. 7a9) of all SubX models showed a better skill (mean CORR 0.55) than any individual model. For week 2, the EMC, CDSM, NCEP, and ESRL

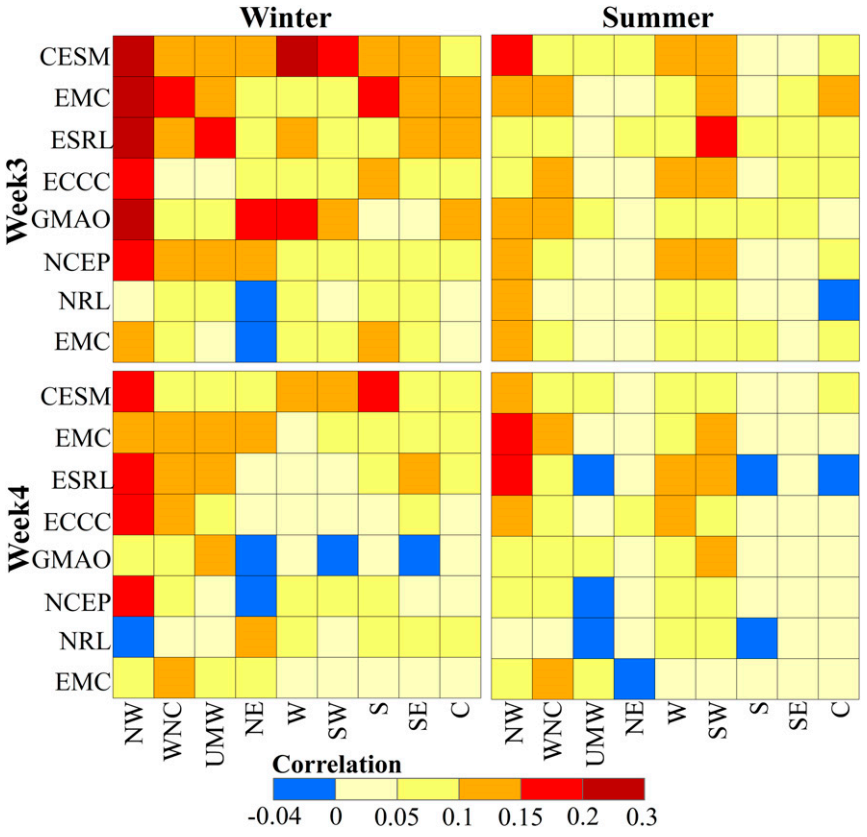

FIG. 6. Correlation of eight SubX models with observations in NCEI climate regions for lead weeks 3–4 in winter and summer.

models performed well, but the performance was greatly decreased compared to week 1; the MME had a CORR of 0.37, indicating increased performance compared to the individual model. For weeks 3–4, the CESM and EMC showed the best skill with a higher CORR in the W, NW, and WNC. The NRL performed the worst with a few negative CORRs found in the eastern United States (Figs. 7c7,d7). Overall, the CORR decreased with an increasing lead time, while the MME consistently improved the forecast performance during weeks 1–4.

The model errors can be directly related to the variability, bias, and correlation between the observations and the SubX forecasts. KGE considers all these aspects and is considered a comprehensive metric for assessing the model performance (Gupta et al. 2009). Figure 8 shows the KGE of SPF from each SubX model. For week 1, all models matched well with the observations, with the NCEP, CESM, ESRL, EMC, and CESM showing better performance compared to the others (KGE equals 0.39, 0.38, 0.37, and 0.35, respectively) (Figs. 8a1–a8). However, many areas showed negative values in week 2, especially in the southern CONUS (Figs. 8b1–b8). For weeks 3–4, approximately half of the grid points showed negative KGE values, indicating poor performance in these areas, such as NE, C, SW, and S (Figs. 8c1–c8,d1–d8). Relatively better skills were found in the West Coast region and most parts of the WNC. Compared with the individual model, the MME can improve the prediction skill of weeks 1–4 (Figs. 8a9,b9,c9,d9). Taking week 1 as an example, the KGE is generally improved by the MME in most regions, especially in the W, NW, and WNC regions (Fig. 8a9). Compared with RSMAS, the MME resulted in a significant improvement of 82.6% ($p < 0.05$). However, with an increase in lead time, the prediction skill of the MME decreased rapidly from 0.42 in week 1 to 0.16, 0.05, and 0.03 in weeks 2–4, respectively. According to the discussion above, we concluded that no individual model consistently outperformed the others for all regions and weeks, while the MME showed better skill than any individual model; thus, suitable postprocessing techniques should combine the strengths and weaknesses of each SubX model.

c. *Skill of EMOS postprocessed forecasts*

1) DETERMINING THE OPTIMAL TRAINING PERIOD FOR EMOS POSTPROCESSING

Several factors influenced the performance of EMOS postprocessing, and the most important two factors were the training period and the individual model skill (Ma et al. 2018; Scheuerer 2014). For determining the optimal training period, previous studies suggested that, if possible, larger training samples and longer training periods were favorable to optimize and achieve a more stable parameter estimation for each verification

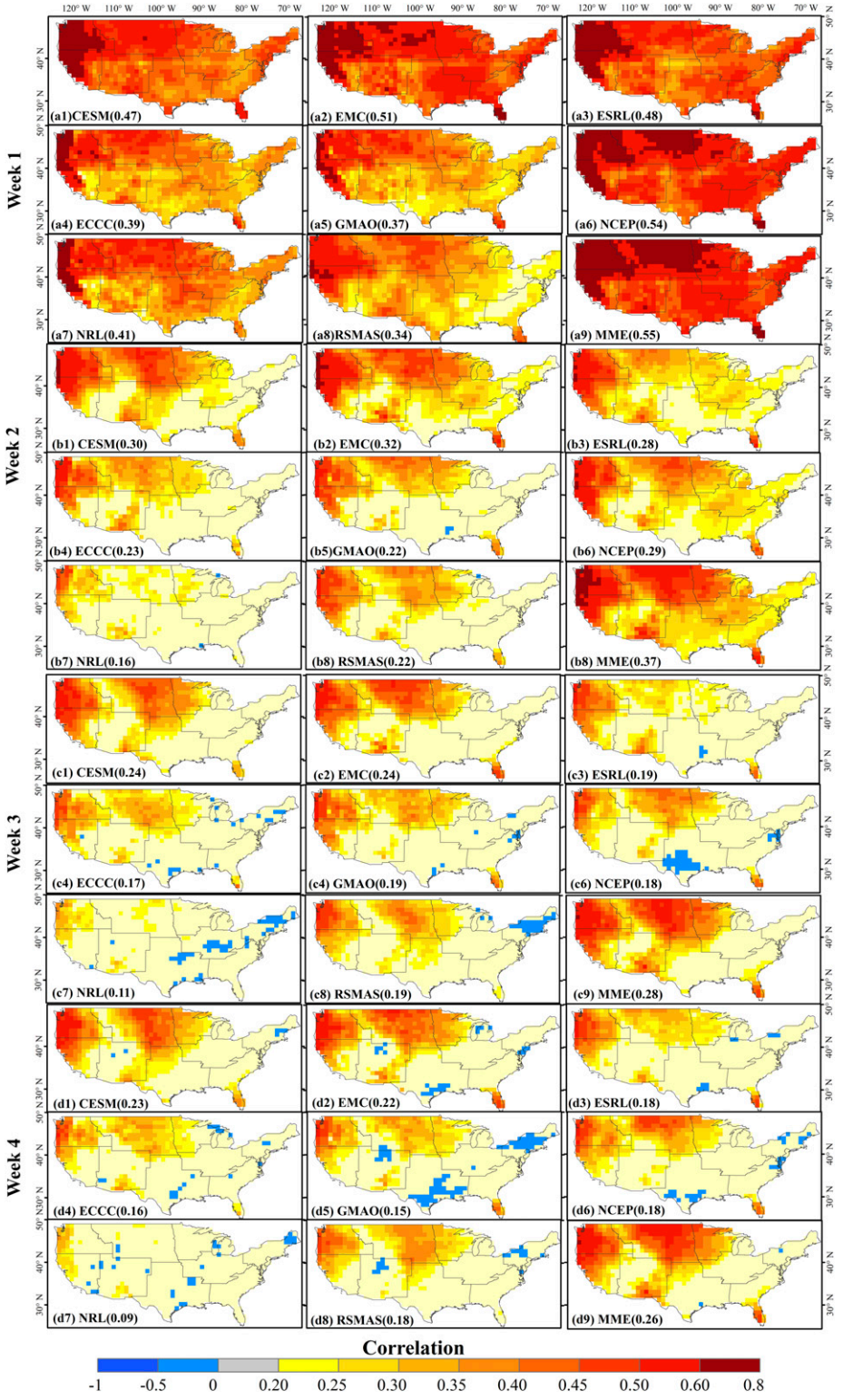

FIG. 7. Maps of the correlation between the model [including the mean of multimodel ensemble (MME)] forecasts of weekly precipitation and the observations in all seasons.

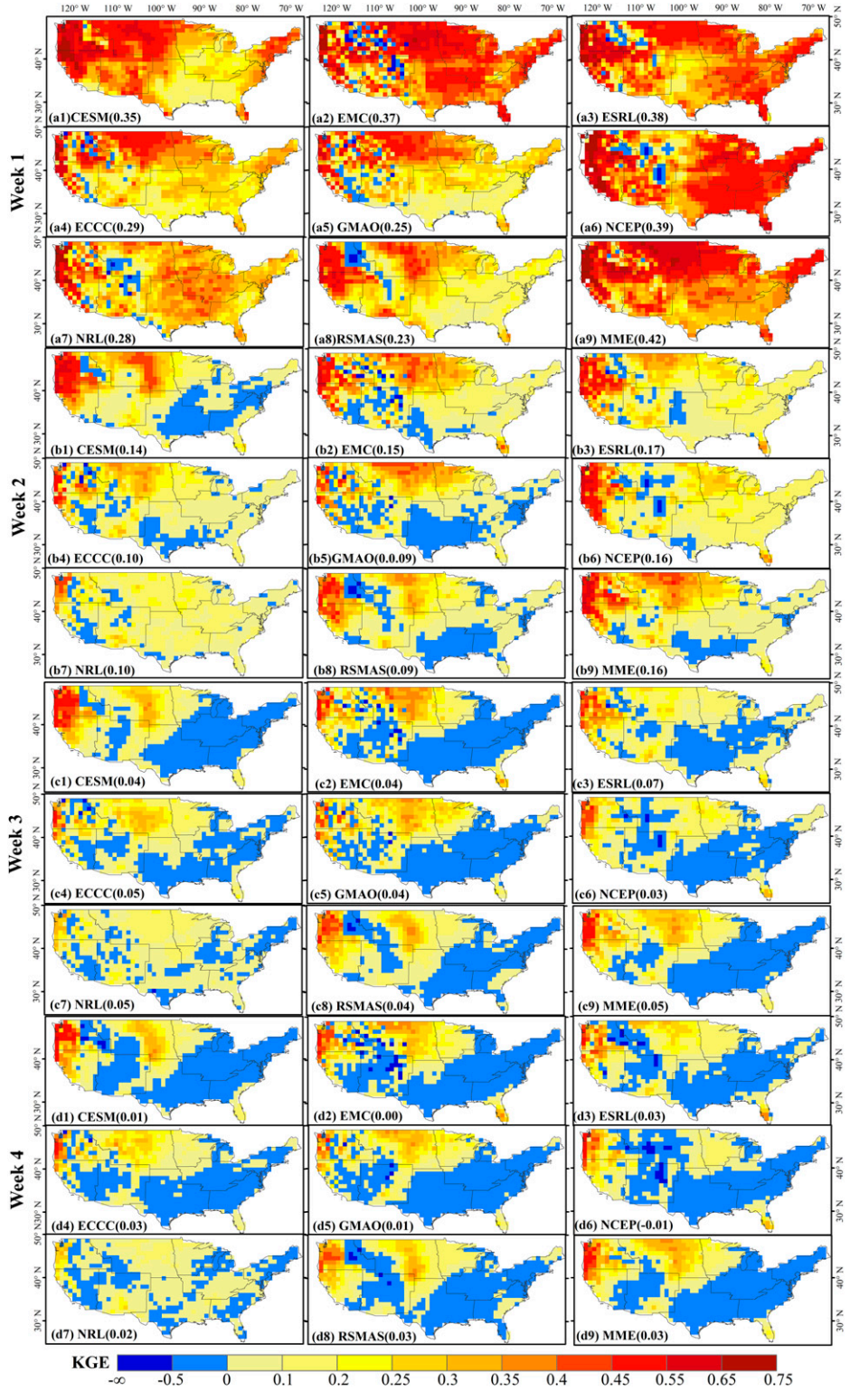

FIG. 8. The same analysis as in Fig. 7, but for KGE.

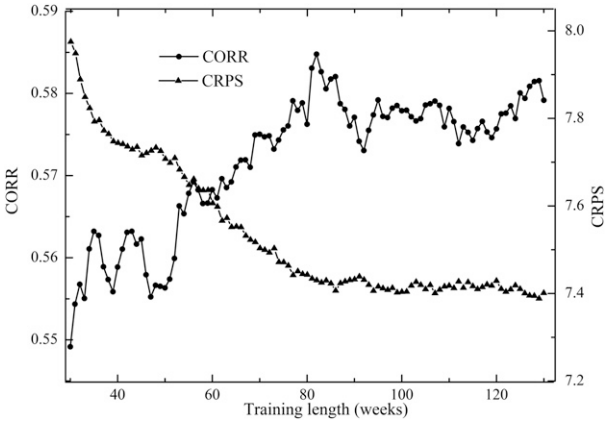

FIG. 9. Impact of training weeks on the correlation (CORR) and continuous ranked probability score (CRPS) averaged over random locations in each climate region.

dataset (Ma et al. 2018; Wilson et al. 2007). On the other hand, the shorter training period could improve the computational efficiency while achieving similar performance to that of the long training period. To evaluate the impact of the training period on the postprocessed subseasonal forecast skill, 3 grid points in each climate region, for a total of 27 grid points, were randomly chosen for analysis in this study. Figure 9 shows the changes in CORR and CPRS with different training periods ranging from 30 to 130 weeks. The results clearly show that the CORR and CRPS generally increased from 30 weeks to approximately 80 weeks but tended to stabilize after 80 weeks. Therefore, by considering the effectiveness and computational efficiency of the EMOS method, training weeks of 100 were employed in the SubX models across the CONUS.

2) Deterministic forecast skill of EMOS postprocessed forecasts

In this section, we compared the skill of EMOS postprocessed forecasts with simple MME forecasts. Figure 10 shows the difference between the EMOS and MME correlation skills. Across the CONUS, the EMOS showed better performance than the MME over most of the area but with large spatial variations (Figs. 10c1–c4). On average, the prediction skills in nine climate regions were significantly ($p < 0.05$) improved by 12% for weeks 1–2 and 20% for weeks 3–4 compared with the MME forecast skill. For week 1, which lies in the weather forecast horizon, the MME CORR values in all regions were already high. Although the EMOS postprocessing method improves the prediction skills in 98.1% of the areas, the increase is relatively small, with an average increase of 9.6% (Figs. 10c1,d). For week 2, EMOS improved prediction skill for more than 95% of the grid points, with an average increase of 14.9%. Significant improvements were found in the NE, UMW, and S regions, with values of 27.7%, 19.3%, and 16.5%, respectively ($p < 0.05$), while lower improvements occurred mainly in the NW (7%, $p > 0.05$) (Fig. 10a2). Overall, significant improvements were found across the CONUS, with

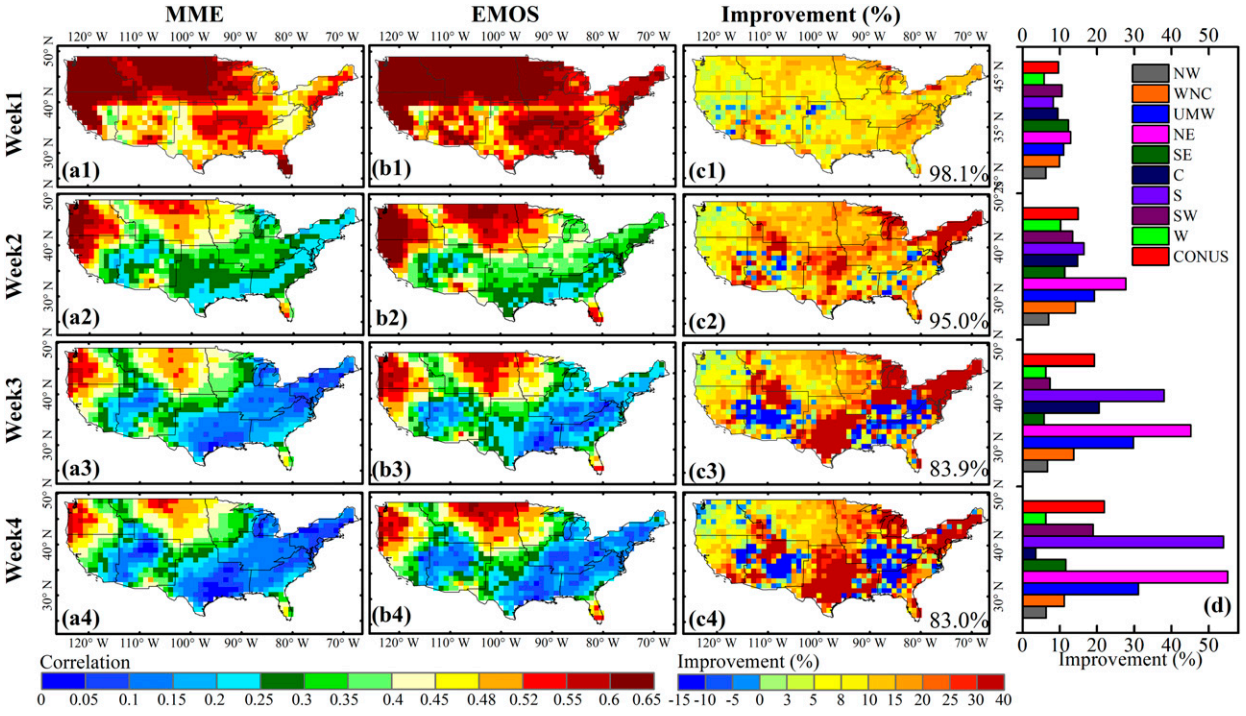

FIG. 10. Spatial distribution of correlations for (a1)–(a4) multimodel ensembles (MME), (b1)–(b4) ensemble model output statistics (EMOS), (c1)–(c4) percentage improvement by EMOS against MME, and (d) summary improvement in nine climate regions and the CONUS. The number in the bottom right in the third column is the positive area as a percentage of the CONUS.

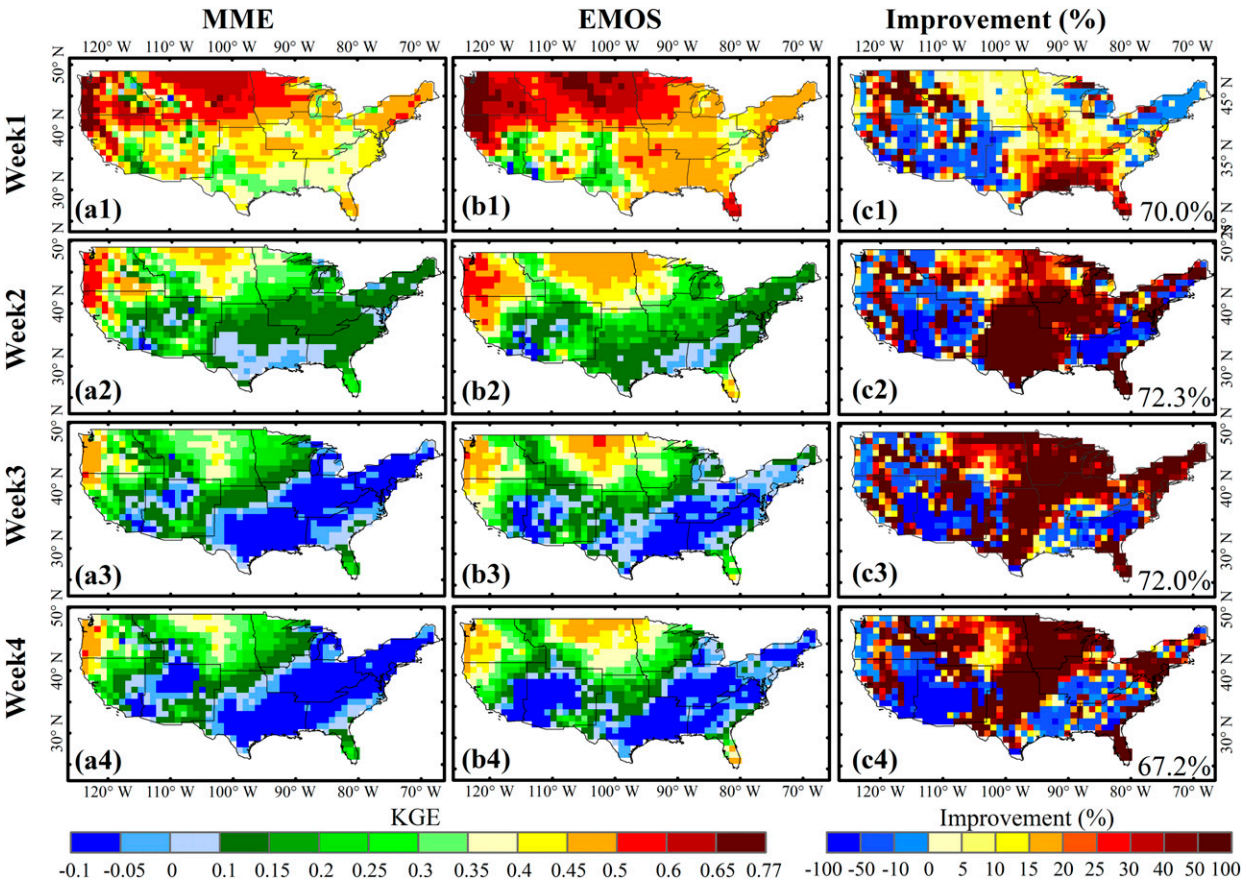

FIG. 11. Includes the same analysis as that in Fig. 10, but for KGE.

values of 19.3% and 22.1% for week 3 and 4, respectively ($p < 0.05$). Similar to weeks 1–2, a large improvement in CORR was found in the NE, S, and UMW regions for weeks 3–4 but the value was much larger than that in the first two weeks. However, ~17% areas showed skill reduction, which mainly occurred in the SW and SE regions, indicating that the prediction skill of MME was better than that of EMOS in those areas (Fig. 10).

For KGE, the improvement also varied by location and lead time (Fig. 11). There was a significant improvement in the SE and NC parts of the CONUS for week 1 ($p < 0.05$), accounting for 70% of the total area, while for ~30% of the areas, a negative improvement was found over the CONUS, which mainly occurred in the NE, SW, and W regions (Fig. 11c1, $p > 0.05$). For weeks 2–4, at least 67.2% of the total areas showed significant improvement for each week, and these sites were mainly located in the S, C, and UMW regions (Figs. 11c2–c4). In contrast, ~30% of the areas showed negative improvements, indicating that the simple MME is the better choice for these regions (Figs. 11c2–c4). In summary, our findings indicated that EMOS postprocessing can significantly improve SPF performance in most of the regions, particularly regarding the correlation in the NE and S regions and the KGE in the S, C, and UMW regions compared to the simple MME. However, it is important to remember that we selected the CRPS as the objective function to minimize the EMOS approach. Better performance can be achieved with this CRPS-based method compared to other metrics.

3) PROBABILISTIC FORECAST SKILL OF EMOS POSTPROCESSED FORECASTS

For the SPF, improving the probabilistic forecasting skill is the target for the EMOS optimization. To compare the probabilistic forecasting performance of EMOS with respect to MME, the probabilistic skill of CRPS for MME and EMOS and their percentages of differences in each location and climate region were analyzed (Fig. 12). The MME in weeks 1–4 showed lower CRPS values in the SW, W, and WNC regions (CRPS < 10) than in the SE and west coast regions of the CONUS (CRPS > 20; Figs. 12a1–a4), suggesting a better probabilistic forecast performance. With EMOS postprocessing, in terms of the CRPS, the probabilistic precipitation forecasts consistently improved across the CONUS for all lead times (Figs. 12c1–c4). This result was reasonable since the EMOS considered the CRPS as its objective function to be minimized (Scheuerer 2014). Notable regions with more than 50% improvement were found in the NW and WNC regions, while the C, SE, and S regions had relatively lower improvements, with values

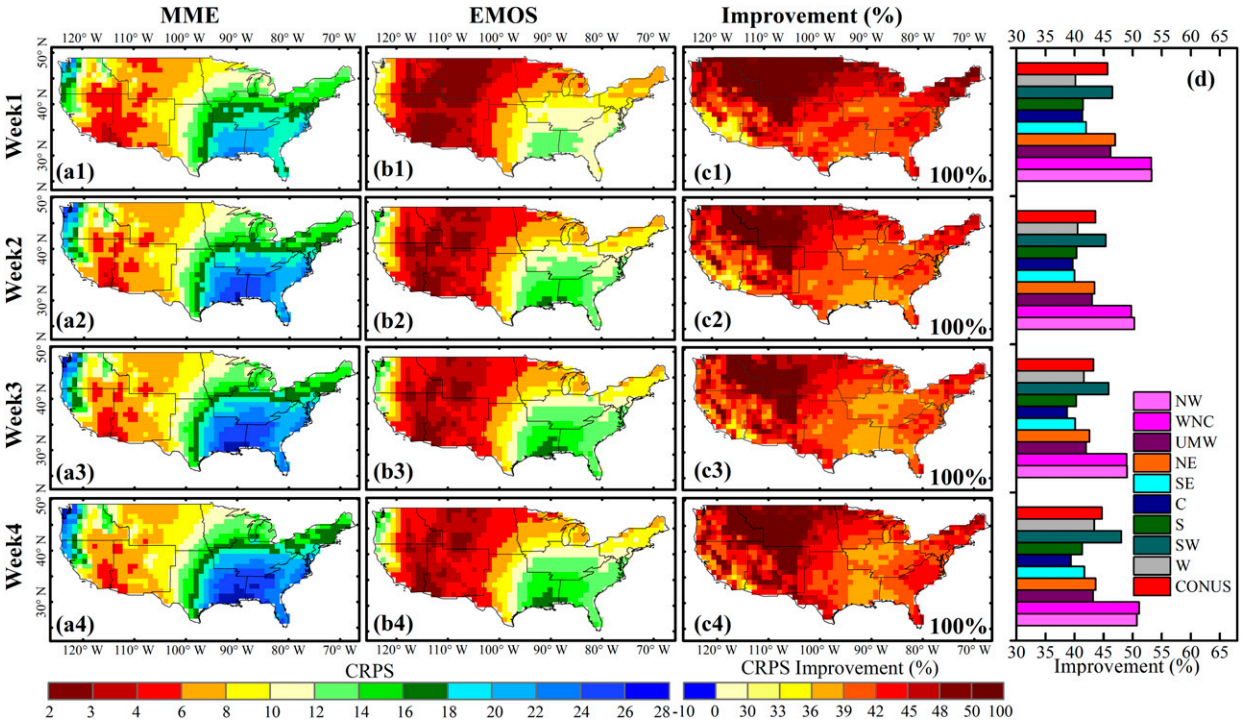

FIG. 12. Spatial distribution of the CRPS skill levels for (a1)–(a4) MME, (b1)–(b4) EMOS, (c1)–(c4) the percentage improvement by EMOS, and (d) the summary improvement in nine climate regions and the CONUS. The number in the bottom right in the third column is the positive area as a percentage of the CONUS.

of approximately 40% (Fig. 12d). For weeks 3 and 4, the CRPS improvement level across the CONUS were 43.2% and 44.7%, respectively, which were only slightly lower than the improvement for weeks 1–2.

Investigating the BSS at a certain threshold can help us understand the ability of the models to predict a certain type of precipitation event. As an example, we computed the BSS of MME and EMOS in forecasting rainy events with a threshold of 5 mm day^{-1} and investigated whether EMOS could improve the forecast performance. For week 1, the prediction skill of EMOS to 5 mm day^{-1} was significantly improved compared to MME in the West Coast, NC, and Florida regions (Fig. 13, $p < 0.05$). Most areas of the CONUS (83.6%) have positive values, and only a small fraction (16.4%) show negative values. Furthermore, for weeks 2–4, with the EMOS method, approximately twice as many areas showed a positive BSS compared to the MME results (Figs. 13a2–a4,b2–b4). The percentages of the areas showing positive improvements for weeks 2–4 were 78.7%, 59.2%, and 50.9%, respectively (Figs. 13c1–c4), and those areas covered the majority of the CONUS except the SW region. Overall, the improvement skill decreases sharply with an increasing lead time. Especially for week 4, only half of the areas have positive values for EMOS improvement, which indicates that the MME can also capture this event well over the remaining half of the areas. According to these results, we can conclude that the EMOS can significantly improve the 5 mm day^{-1} rainy event forecast, but with significant spatial heterogeneity. EMOS can be used as an efficient approach for improving the weekly probabilistic precipitation forecast in the most northern areas, while the MME is an alternative choice over the southern area.

4) EMOS WEIGHT COEFFICIENTS FOR INDIVIDUAL SUBX MODELS

One of the advantages of the EMOS method is that the weights assigned by EMOS can reflect the relative contributions of the subseasonal models to the predictive skill over a training period. The better performing model was prone to obtaining higher coefficients (Wilson et al. 2007), and the best combination of skillful models would maximize the performance of the EMOS. Thus, we examined the EMOS weight coefficients of each SubX model, which provided the performance-based contributions of each model in different climate regions. Figure 14 shows the averaged weight coefficients of b [Eq. (1)] over nine climate regions for different lead times and SubX models. For week 1, the NCEP model was given the highest weight coefficient in week 1 in all NCEI regions, which was approximately 3–4 times higher than the values of the other models, followed by CESM and EMC, which had coefficient values of approximately 0.1. For a longer lead time (weeks 2–4), however, the coefficient of NCEP decreased sharply, with six of the nine regions showing lower coefficients than the EMC. This result may be due to the inaccurate representation of MJO propagation in the NCEP model, which degraded the subseasonal forecast skill (Wang et al. 2014). Overall, the NCEP model was given the highest weight with the

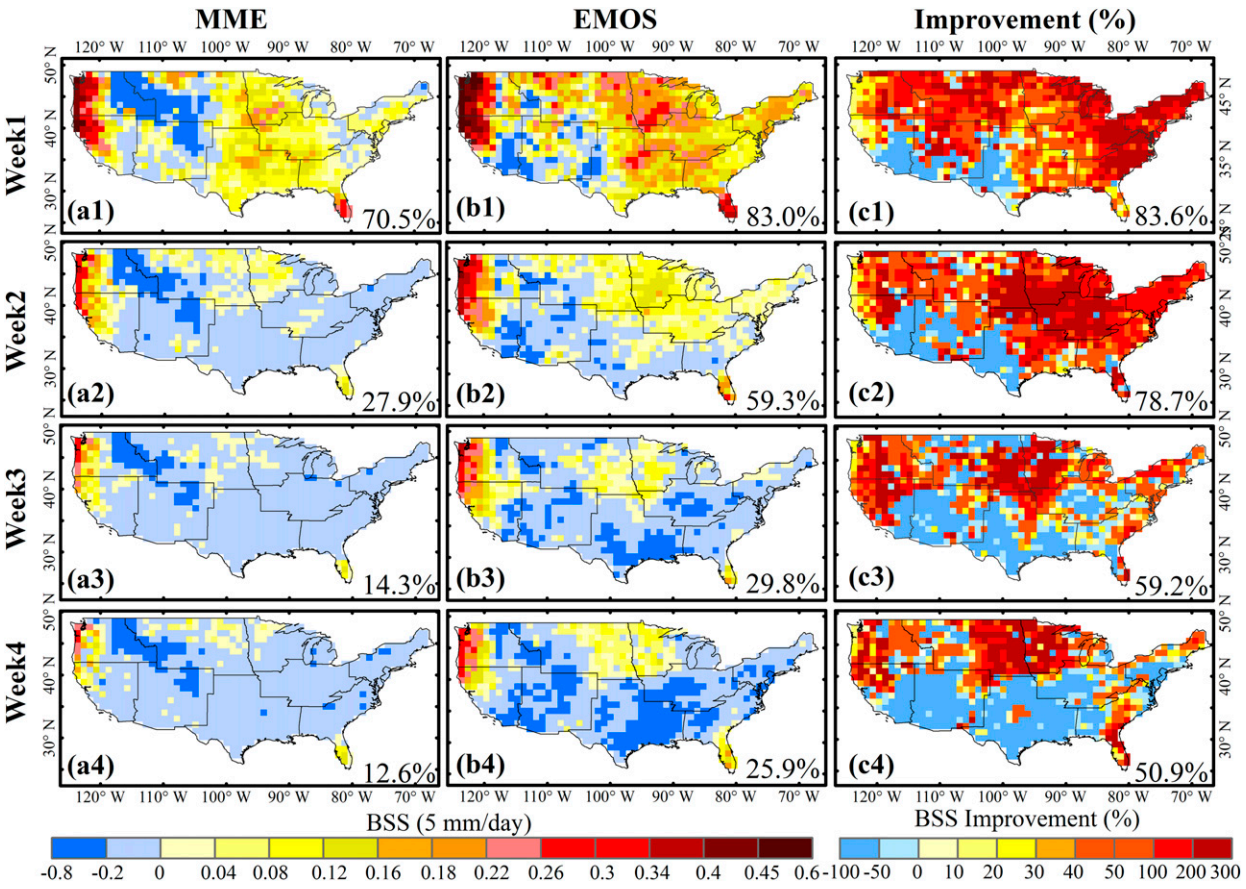

FIG. 13. Spatial distribution of the BSS skill level for 5 mm precipitation for (a1)–(a4) MME, (b1)–(b4) EMOS, and (c1)–(c4) the percentage improvement by EMOS. The numbers in the bottom right in the first and second column refer to the percentage of positive BSS over the CONUS, and those in the third column are the percentage of area showing positive improvement over the CONUS.

short lead time, but the weight decreased dramatically with the increase in lead time, while the CESM, EMC, NCEP, and GMAO were given approximately equal weights for weeks 2–4.

d. Effects of MJO and ENSO on SPF skill

To investigate the influence of MJO or ENSO, we examined the weekly precipitation forecasting skill levels considering the MJO or ENSO phases (Fig. 15). The daily threshold value for the active MJO was defined as 1.5, and then the MJO data series were divided into two parts: active and nonactive. These two parts of the MJO data are significantly different at the $p = 0.05$ level. The week was considered MJO active if that week had more than six days in which MJO was active. Then, the CORR and CRPS of the MME under active MJO and nonactive MJO were calculated and compared at each grid point (Figs. 15a1,a2,b1,b2). Comparing the MJO active versus nonactive periods, the CORR values in weeks 3–4 were higher in the WNC, S (approximately 30% improvement), and eastern SE regions (more than 80% at several locations) but the differences were not obvious in the other areas. The CRPS was significantly improved in most areas of the CONUS (except in the W, SW, and S regions), with a mean improved value of 5% over the CONUS. Significant improvement was found in the central area, with a value of more than 17% ($p < 0.05$). We also found that an active MJO improved the CORR and CRPS in the West Coast region for week 3 but showed only a small improvement for week 4. This result indicates that the MJO positively influenced the long lead time predictions for the northern and central areas of the CONUS. This result is consistent with Pan et al. (2019), who found that MJO events could systematically modulate the West Coast's precipitation distribution, and a better representation of the MJO in forecasting models could enhance the subseasonal precipitation forecast skill level under active MJO conditions.

In addition to MJO, ENSO is another source of subseasonal precipitation predictability (Vitart 2014). ENSO includes El Niño (positive phase), La Niña (negative phase), and a neutral phase. We first compared La Niña/El Niño versus neutral conditions. The results showed there was no significant improvement ($p > 0.05$) over most climate regions of the CONUS. When comparing La Niña versus El Niño conditions, there are substantial differences ($p < 0.05$) for most climate regions. Therefore, we compared the model performances (CORR and CRPS) for weeks 3–4 during the El Niño phase versus the La Niña phase (Figs. 15c1,c2,d1,d2) across the CONUS. The results showed that the forecast during the La

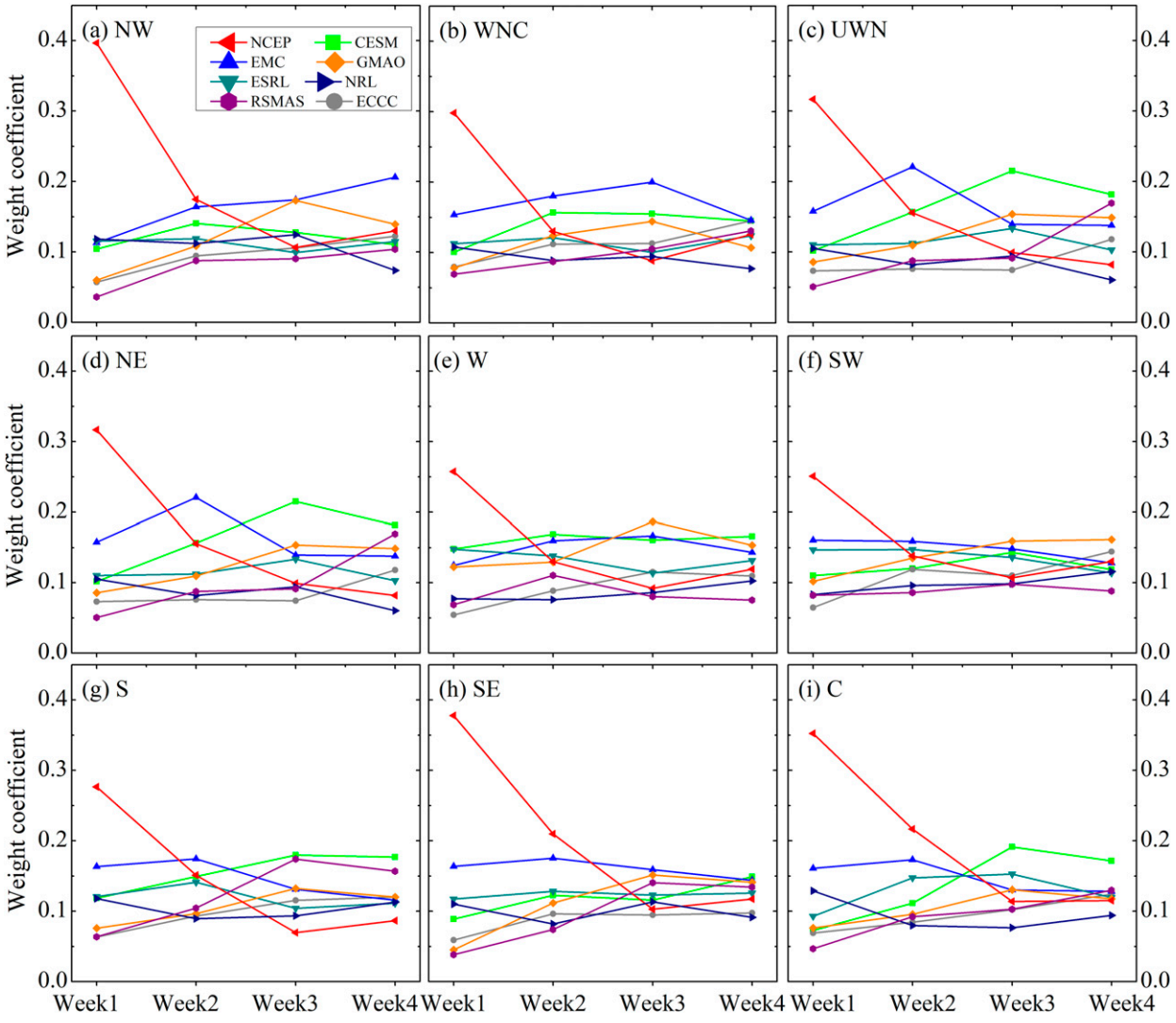

FIG. 14. Weight coefficient of the SubX models on EMOS for weeks 1–4.

Niña phase had a higher CORR in most of the CONUS compared to that in the El Niño phase, particularly in the eastern and southern CONUS for week 3 and in the western CONUS for week 4. During La Niña, the forecast showed better CRPS values in most southern and western areas, while the forecast was better in the northern area during El Niño. DelSole et al. (2017) also found that the most predictable components of winter precipitation in CFSv2 were highly correlated with ENSO. A high impact of ENSO on the SubX skill level was found in the southwestern CONUS (Figs. 15c,d), which was consistent with the findings of Vigaud et al. (2017), who found similar effects of ENSO on the ECMWF, NCEP, and CMA model forecasts.

5. Conclusions

The SPF represents a major challenge to bridge the gap between weather- and climate-scale predictions. In this study, we assessed the deterministic and probabilistic SPF skill from eight state-of-the-art models under different conditions (in terms of climate regions, lead times, and ENSO and MJO patterns), and explored the potential of the GEV-based EMOS framework to postprocess MME SPF skill. The skill of eight SPF models was highly dependent on models, lead times, regions, and MJO and ENSO conditions. All models showed high skill levels for the short lead time (weeks 1–2), but the skill decreased sharply for the longer lead time (weeks 3–4). For the deterministic forecast skill, the EMC, ERSL, CESM, and NCEP had relatively better performances than the other models, with particularly high skill in the West Coast and north-central CONUS. Overall, there was no individual model that achieved a higher skill than the others for all lead times and regions, while the MME obtained higher skill than any individual model. For the subseasonal time scale of weeks 3–4, the EMOS method can generally improve both the deterministic and probabilistic forecast skill levels compared to the

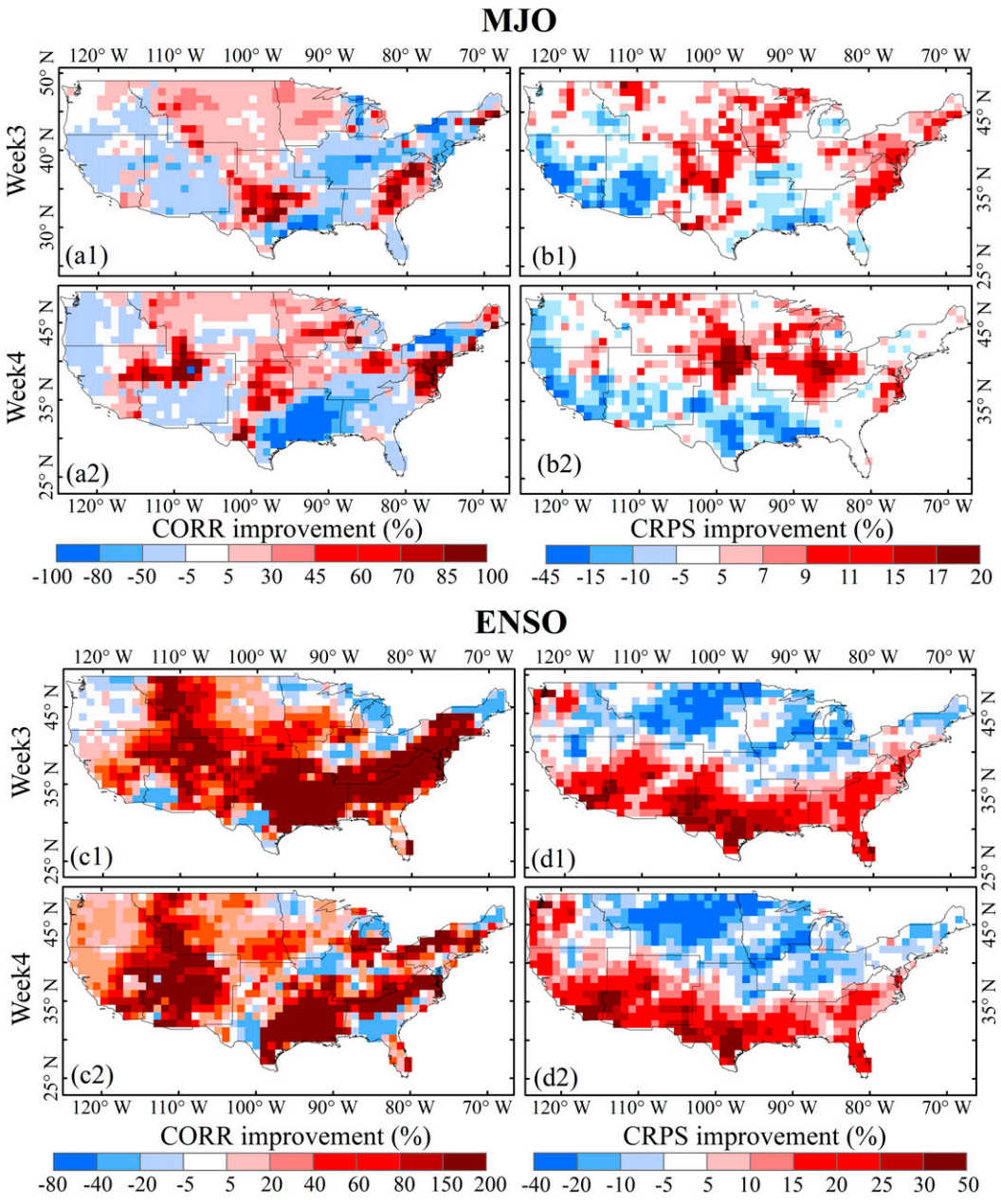

FIG. 15. Improvement of the deterministic (CORR) and probabilistic (CRPS) forecast skill levels for weeks 3–4 when comparing (a)–(b) the active MJO to the nonactive MJO and (c)–(d) the La Niña phases to the El Niño phase.

MME over most areas of the CONUS, with a mean percentage improvement of approximately 20% for CORR and 43% for CRPS ($p < 0.05$). The improvement of EMOS also presents regional differences, and better performance mainly occurs in the northern CONUS. In terms of the BSS at the 5-mm threshold, skill improvements due to EMOS were found in the West Coast and northern CONUS, where the original forecast skill was high. Regarding the relative contributions of the individual SubX model to the predictive skill, the NCEP model was given the highest weight at the shortest lead time, but the weight decreased dramatically with the increase in lead time, while the CESM, EMC, NCEP, and GMAO models were given approximately equal weights for lead weeks 2–4. While the performance of the EMOS postprocessed SPFs varied by the lead time, the length of the training period, and the model, it generated more accurate and reliable forecasts at subseasonal time scales compared to the MME forecasts and will be beneficial for real-world operations. The presence of active MJO events notably increased the forecast skill level in the north-central region

during weeks 3–4, while the ENSO phases influenced the skill level mostly in the southern regions.

Acknowledgments. This work was supported in part by the China Scholarship Council (201809040009), the USDA-NIFA Hatch project (Accession 1012578), USDA-NIFA AFRI program (Accession 1019690), NOAA (NA19NOS4510194), the National Natural Science Foundation of China (41701019), and the Startup Foundation for Introducing Talent of NUIST (2017r069).

Data availability statement. The data used in this study were all obtained from public data repositories, which are openly available at the locations cited in the reference section.

REFERENCES

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

Barnston, A. G., and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate*, **6**, 963–977, https://doi.org/10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2.

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

Bi, D., and Coauthors, 2013: The ACCESS coupled model: Description, control climate and evaluation. *Aust. Meteor. Oceanogr. J.*, **63**, 41–64, https://doi.org/10.22499/2.6301.004.

Bleck, R., and Coauthors, 2015: A vertically flow-following icosahedral grid model for medium-range and seasonal prediction. Part I: Model description. *Mon. Wea. Rev.*, **143**, 2386–2403, https://doi.org/10.1175/MWR-D-14-00300.1.

Bliefernicht, J., and Coauthors, 2019: Quality and value of seasonal precipitation forecasts issued by the West African regional climate outlook forum. *J. Appl. Meteor. Climatol.*, **58**, 621–642, https://doi.org/10.1175/JAMC-D-18-0066.1.

Borga, M., and Coauthors, 2014: Hydrogeomorphic response to extreme rainfall in headwater systems: Flash floods and debris flows. *J. Hydrol.*, **518**, 194–205, https://doi.org/10.1016/j.jhydrol.2014.05.022.

De Felice, M., A. Alessandri, and F. Catalano, 2015: Seasonal climate forecasts for medium-term electricity demand forecasting. *Appl. Energy*, **137**, 435–444, https://doi.org/10.1016/j.apenergy.2014.10.030.

DelSole, T., L. Trenary, M. K. Tippett, and K. Pegion, 2017: Predictability of week-3–4 average temperature and precipitation over the contiguous United States. *J. Climate*, **30**, 3499–3512, https://doi.org/10.1175/JCLI-D-16-0567.1.

Donat, M. G., A. L. Lowry, L. V. Alexander, P. A. O'Gorman, and N. Maher, 2016: More extreme precipitation in the world's dry and wet regions. *Nat. Climate Change*, **6**, 508–513, https://doi.org/10.1038/nclimate2941.

Duan, Q., and T. J. Phillips, 2010: Bayesian estimation of local signal and noise in multimodel simulations of climate change. *J. Geophys. Res.*, **115**, D18123, https://doi.org/10.1029/2009JD013654.

Feng, A., and Q. Chao, 2020: An overview of assessment methods and analysis for climate change risk in China. *Phys. Chem. Earth*, **117**, 102861, https://doi.org/10.1016/j.pce.2020.102861.

Gent, P. R., and Coauthors, 2011: The community climate system model version 4. *J. Climate*, **24**, 4973–4991, https://doi.org/10.1175/2011JCLI4083.1.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, **377**, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus*, **57A**, 219–233, https://doi.org/10.1111/j.1600-0870.2005.00103.x.

Hogan, T. F., and Coauthors, 2014: The Navy Global Environmental Model. *Oceanography*, **27**, 116–125, https://doi.org/10.5670/oceanog.2014.73.

Huffman, G. J., and Coauthors, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeor.*, **2**, 36–50, https://doi.org/10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2.

Hurrell, J. W., and Coauthors, 2013: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, **94**, 1339–1360, https://doi.org/10.1175/BAMS-D-12-00121.1.

Jones, C., L. M. V. Carvalho, J. Gottschalck, and W. Higgins, 2011: The Madden–Julian Oscillation and the relative value of deterministic forecasts of extreme precipitation in the contiguous United States. *J. Climate*, **24**, 2421–2428, https://doi.org/10.1175/2011JCLI-D-10-05002.1.

Karl, T., and W. J. Koss, 1984: Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983. Historical Climatology Series 4-3, National Climatic Data Center, 38 pp., https://repository.library.noaa.gov/view/noaa/10238.

Kim, H., M. A. Janiga, and K. Pegion, 2019: MJO propagation processes and mean biases in the SubX and S2S reforecasts. *J. Geophys. Res. Atmos.*, **124**, 9314–9331, https://doi.org/10.1029/2019JD031139.

Kirtman, B. P., and Coauthors, 2014: The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

Knoben, W. J. M., J. E. Freer, and R. A. Woods, 2019: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, *Hydrol. Earth Syst. Sci.*, **23**, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019.

Koster, R. D., and Coauthors, 2004: Regions of strong coupling between soil moisture and precipitation. *Science*, **305**, 1138–1140, https://doi.org/10.1126/science.1100217.

Li, S., and A. W. Robertson, 2015: Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Mon. Wea. Rev.*, **143**, 2871–2889, https://doi.org/10.1175/MWR-D-14-00277.1.

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

Li, Y., Z. Wu, H. He, Q. J. Wang, H. Xu, and G. Lu, 2020: Post-processing sub-seasonal precipitation forecasts at various spatiotemporal scales across China during boreal summer monsoon. *J. Hydrol.*, **598**, 125742, https://doi.org/10.1016/j.jhydrol.2020.125742.

Lin, H., N. Gagnon, S. Beauregard, R. Muncaster, M. Markovic, B. Denis, and M. Charron, 2016: GEPS-based monthly prediction

at the Canadian Meteorological Centre. *Mon. Wea. Rev.*, **144**, 4867–4883, https://doi.org/10.1175/MWR-D-16-0138.1.

Ma, F., A. Ye, and Q. Duan, 2017: Seasonal drought ensemble predictions based on multiple climate models in the upper Han River Basin, China. *Climate Dyn.*, **53**, 7447–7460, https://doi.org/10.1007/s00382-017-3577-1.

Ma, Y., and Coauthors, 2018: Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian Model Averaging scheme over the Tibetan Plateau. *J. Geophys. Res. Atmos.*, **123**, 814–834, https://doi.org/10.1002/2017JD026648.

Madden, R. A., and P. R. Julian, 1972: Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.*, **29**, 1109–1123, https://doi.org/10.1175/1520-0469(1972)029<1109:DOGSCC>2.0.CO;2.

Medina, H., and D. Tian, 2020: Comparison of probabilistic post-processing approaches for improving numerical weather prediction-based daily and weekly reference evapotranspiration forecasts. *Hydrol. Earth Syst. Sci.*, **24**, 1011–1030, https://doi.org/10.5194/hess-24-1011-2020.

——, ——, P. Srivastava, A. Pelosi, and G. B. Chirico, 2018: Medium-range reference evapotranspiration forecasts for the contiguous United States based on multi-model numerical weather predictions. *J. Hydrol.*, **562**, 502–517, https://doi.org/10.1016/j.jhydrol.2018.05.029.

Mishra, N., C. Prodhomme, and V. Guemas, 2018: Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe. *Climate Dyn.*, **52**, 4207–4225, https://doi.org/10.1007/s00382-018-4404-z.

Mo, K. C., and R. W. Higgins, 1998: Tropical convection and precipitation regimes in the western United States. *J. Climate*, **11**, 2404–2423, https://doi.org/10.1175/1520-0442(1998)011<2404:TCAPRI>2.0.CO;2.

——, and D. P. Lettenmaier, 2016: Precipitation deficit flash droughts over the United States. *J. Hydrometeor.*, **17**, 1169–1184, https://doi.org/10.1175/JHM-D-15-0158.1.

Monhart, S., and Coauthors, 2018: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *J. Geophys. Res. Atmos.*, **123**, 7999–8016, https://doi.org/10.1029/2017JD027923.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772, https://doi.org/10.1038/nature02771.

Pan, B., K. Hsu, A. AghaKouchak, S. Sorooshian, and W. Higgins, 2019: Precipitation prediction skill for the West Coast United States: From short to extended range. *J. Climate*, **32**, 161–182, https://doi.org/10.1175/JCLI-D-18-0355.1.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Glecker, 2008: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J. Geophys. Res.*, **113**, D14209, https://doi.org/10.1029/2007JD009334.

Reichle, R., and Q. Liu, 2014: Observation-corrected precipitation estimates in GEOS-5. NASA/TM-2014-104606, Vol. 35, 18 pp., https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150000725.pdf.

Scaife, A. A., and Coauthors, 2016: Seasonal winter forecasts and the stratosphere. *Atmos. Sci. Lett.*, **17**, 51–56, https://doi.org/10.1002/asl.598.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using Ensemble Model Output Statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, https://doi.org/10.1002/qj.2183.

——, and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1.

Schonher, T., and S. E. Nicholson, 1989: The relationship between California rainfall and ENSO events. *J. Climate*, **2**, 1258–1269, https://doi.org/10.1175/1520-0442(1989)002<1258:TRBCRA>2.0.CO;2.

Specq, D., and L. Batté, 2020: Improving subseasonal precipitation forecasts through a statistical-dynamical approach: Application to the southwest tropical Pacific. *Climate Dyn.*, **55**, 1913–1927, https://doi.org/10.1007/s00382-020-05355-7.

Sun, S., R. Bleck, S. G. Benjamin, B. W. Green, and G. A. Grell, 2018: Subseasonal forecasting with an icosahedral, vertically quasi-Lagrangian coupled model. Part I: Model overview and evaluation of systematic errors. *Mon. Wea. Rev.*, **146**, 1601–1617, https://doi.org/10.1175/MWR-D-18-0006.1.

Sun, Q., C. Miao, Q. Duan, H. Ashouri, S. Sorooshian, and K. L. Hsu, 2018: A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Rev. Geophys.*, **56**, 79–107, https://doi.org/10.1002/2017RG000574.

Sun, Z., and Coauthors, 2013: Modifications to atmospheric physical parameterisations aimed at improving SST simulations in the ACCESS coupled model. *Aust. Meteor. Oceanogr. J.*, **63**, 233–247, https://doi.org/10.22499/2.6301.015.

Tian, D., E. F. Wood, and X. Yuan, 2017: CFSv2-based subseasonal precipitation and temperature forecast skill over the contiguous United States. *Hydrol. Earth Syst. Sci.*, **21**, 1477–1490, https://doi.org/10.5194/hess-21-1477-2017.

Troccoli, A., 2010: Seasonal climate forecasting. *Meteor. Appl.*, **17**, 251–268, https://doi.org/10.1002/met.184.

Vigaud, N., A. W. Robertson, and M. K. Tippett, 2017: Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.*, **145**, 3913–3928, https://doi.org/10.1175/MWR-D-17-0092.1.

——, ——, and ——, 2018: Predictability of recurrent weather regimes over North America during winter from submonthly reforecasts. *Mon. Wea. Rev.*, **146**, 2559–2577, https://doi.org/10.1175/MWR-D-18-0058.1.

——, M. K. Tippett, J. Yuan, A. W. Robertson, and N. Acharya, 2020: Spatial correction of multimodel ensemble subseasonal precipitation forecasts over North America using local Laplacian eigenfunctions. *Mon. Wea. Rev.*, **148**, 523–539, https://doi.org/10.1175/MWR-D-19-0134.1.

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.*, **140**, 1889–1899, https://doi.org/10.1002/qj.2256.

——, and Coauthors, 2008: The new VarEPS-monthly forecasting system: A first step towards seamless prediction. *Quart. J. Roy. Meteor. Soc.*, **134**, 1789–1799, https://doi.org/10.1002/qj.322.

——, and Coauthors, 2017: The subseasonal to seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Wanders, N., and E. F. Wood, 2016: Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. *Environ. Res. Lett.*, **11**, 094007, https://doi.org/10.1088/1748-9326/11/9/094007.

Wang, L., and A. W. Robertson, 2018: Week 3–4 predictability over the United States assessed from two operational ensemble prediction systems. *Climate Dyn.*, **52**, 5861–5875, https://doi.org/10.1007/s00382-018-4484-9.

Wang, W., M.-P. Hung, S. J. Weaver, A. Kumar, and X. Fu, 2014: MJO prediction in the NCEP Climate Forecast System version 2. *Climate Dyn.*, **42**, 2509–2520, https://doi.org/10.1007/s00382-013-1806-9.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79, https://doi.org/10.1111/j.1600-0870.2007.00273.x.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

White, C. J., S. W. Franks, and D. McEvoy, 2015: Using subseasonal-to-seasonal (S2S) extreme rainfall forecasts for extended-range flood prediction in Australia. *Proc. IAHS*, **370**, 229–234.

——, and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging. *Mon. Wea. Rev.*, **135**, 1364–1385, https://doi.org/10.1175/MWR3347.1.

Yuan, X., and E. F. Wood, 2012: On the clustering of climate models in ensemble seasonal forecasting. *Geophys. Res. Lett.*, **39**, L18701, https://doi.org/10.1029/2012GL052735.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.

Zhu, E., X. Yuan, and A. W. Wood, 2019: Benchmark decadal forecast skill for terrestrial water storage estimated by an elasticity framework. *Nat. Commun.*, **10**, 1237, https://doi.org/10.1038/s41467-019-09245-3.