Journal of Hydrology 603 (2021) 127058

Contents lists available at ScienceDirect

Journal of Hydrology

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

Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous U.S.

Lujun Zhang [a], Taereem Kim [a], Tiantian Yang [a,*], Yang Hong [a], Qian Zhu [b]

[a] The School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK, USA
[b] School of Civil Engineering, Southeast University, Nanjing, China

ARTICLE INFO

Keywords:
NMME-2
Subseasonal-to-seasonal Precipitation Forecast
CONUS
Forecast Validation
Forecast Bias
Extreme Precipitation

ABSTRACT

The second phase of the North America Multi-Model Ensemble (NMME-2) provides globally available Subseasonal-to-Seasonal (S2S) precipitation forecasts with a daily resolution. The S2S precipitation forecasts are getting increasing attention for their potentials in providing hydrometeorological forcing information for water resources planning at an extended range. However, the forecast skills of many existing S2S forecast products will significantly decrease when the lead time increases, hindering their applicability for watershed-scale hydrologic modeling. Therefore, forecast validation and large-scale evaluation are of great importance for water resources planning and hydrological applications. In this study, we comprehensively evaluate the S2S precipitation forecasts from the NMME-2 dataset over the contiguous United States (CONUS) and during the study period from 1982 to 2011. Three aspects of precipitation forecast capabilities are compared and analyzed: bias, skill scores, and the ability to predict extreme precipitation events. The Parameter-elevation Regressions on Independent Slopes Model (PRISM) is used as ground truth reference. Differs from other regional forecast validation study, we further examined and analyzed the dependences of NMME-2 precipitation forecast skills according to different seasonality, geographical locations, and lead times. Results show that the forecast biases are not sensitive to lead times but are seasonally dependent of all NMME-2 models. Overestimations are found in the Western U.S. in cooler seasons while underestimations are observed in the central regions of the U.S. in warmer seasons. The forecast skill of all individual NMME-2 models generally decreases as increases of lead times. The simple model averaging (SMA) of five NMME-2 models demonstrates a higher forecast skill than any individual NMME-2 models. Spatially, the highest forecast skill scores are observed at coastal areas in the Western U.S. with an one-week lead time. As compared to the historical resampled forecasts, NMME-2 also shows better performance in predicting extreme precipitation events above 99% percentiles and below 1% percentiles with higher probability of detections and lower false alarm ratios. The obtained results suggest the great potentials of NMME-2 precipitation forecasts in assisting ensemble hydrologic forecasts at the S2S scale over the CONUS.

1. Introduction

Precipitation is one of the most important components in the hydrologic cycle (Sorooshian et al., 2011). Accurate and reliable precipitation forecasts with certain lead times could be beneficial in planning and managing social economic activities, preventing financial and life losses from water-related disasters (Akbari Asanjan et al., 2018, Palmer 2002). Different precipitation forecast products can be categorized by the available lead times, such as short-, medium- and long-range forecast products. At the short- and/or medium-ranges (i.e., 2–3 days and 7–10 days, respectively), Numerical Weather Prediction (NWP) models can provide reliable and skillful forecasts globally (Bauer et al., 2015). Especially, at the short- and/or medium ranges, NWP models could generate skillful precipitation forecasts by taking advantages of the high predictability of rainfall from initial atmospheric states, various types of observations, and advanced data assimilation techniques. However, one common critique on the NWP model is that its forecast skill decreases rapidly and the associated forecast uncertainty increases dramatically, when the forecast lead time extends two weeks and beyond (Alley et al., 2019, Berner et al., 2011, Hamill and Juras 2006, Lin et al., 2005,

* Corresponding author at: 202 W. Boyd St., Room 334, Norman, OK 73019, USA.
E-mail address: tiantian.yang@ou.edu (T. Yang).

https://doi.org/10.1016/j.jhydrol.2021.127058
Received 16 August 2021; Received in revised form 28 September 2021; Accepted 7 October 2021
Available online 15 October 2021

Palmer et al., 2004, Ritter and Geleyn 1992, Shrestha et al., 2013, Sun et al., 2014a). This is because NWP model heavily relies on the initial states of atmosphere, and the predictability coming from the initial states dissipates rapidly over lead time. At long-ranges (i.e., months, seasons, years, and even decades), Earth System Models (ESMs) and General Circulation Models (GCMs) coupled with dynamic oceanic and land surface components are reliable alternatives. Both ESMs and GCMs produce more skillful and informative climate outlooks than the NWP models at longer forecast lead times, because by design, they are able to incorporate both the local land surface conditions and sea surface temperature (SST) circulations into the computation for future weather and climate predictions (Vitart et al., 2017).

A forecast gap lies in the transitional period between the medium-range weather forecasts and longer-range seasonal climate outlooks. This transitional period is also referred to as the Subseasonal-to-Seasonal (S2S) timescale, which defines a specific time range beyond 10 days and up to 30 days into the future (White et al., 2017). At the S2S time range, the forecast lead time is sufficiently long that most of the predictability from the initial conditions would be lost but it is also too short for the variability of the ocean to have a strong influence upon local weather (Vitart et al., 2017). This unique physical feature of weather predictability made precipitation forecast at the S2S range notoriously challenging and also makes S2S forecast often considered as a "desert of predictability" (White et al., 2017).

The S2S hydrometeorological forecasts are important information and have a great potential in providing seamless streamflow and flood forecasts at the S2S range. Traditionally, river forecast centers and weather service centers over the globe issue probabilistic seasonal streamflow forecasts forced by seasonal and/or monthly climate outlooks (Wood and Lettenmaier 2006). However, this kind of seasonal streamflow forecasts can only reflect an increased or decreased risks of flooding but do not have the ability to predict floods at the S2S range. Meanwhile, the S2S hydrometeorological forecasts provide an opportunity in assisting streamflow forecasts, which not only reflects the flooding risks but also provides additional information regarding the timing, frequency, or severity of potential floods within seasons (White et al., 2015). Accurate S2S hydrometeorological forecasts could also help the operation of reservoirs in scheduling optimal water supplies and hydropower generations given foreseeable dry and wet water conditions (Sankarasubramanian et al., 2009, Yang et al., 2020; 2021). Despite the potential benefits of S2S ensemble forecasts in water-energy system operation (Ding et al., 2021), it also appears to be a new research area to extend our existing knowledge about weather and climate forecasts in different space and time (Vitart and Robertson 2018).

Previous studies concluded that the predictability of S2S forecast comes from several sources, including initial atmospheric conditions (Cohen et al., 2010, Stockdale et al., 2015), initial land surface soil moisture (Asoka and Mishra 2015, Guo et al., 2011), initial snow conditions (Thomas et al., 2016), and initial sea surface temperatures (Chelton and Wentz 2005). In some recent studies, the planetary-scale oceanic patterns are found to be the main predictability source of S2S forecasts, and these interconnection climate indices include the El Niño–Southern Oscillation (ENSO), Madden–Julian oscillation (MJO), quasi-biennial oscillation (QBO) (Nardi et al., 2020, Pan et al., 2019b, Yang et al., 2017). There are many existing efforts and programs that focus on the hydrometeorological forecasts at the S2S timescales, such as the European Center for Medium Range Weather Forecasts (ECMWF), the Environment Canada (EC), the Japan Meteorological Agency (JMA), the China Meteorological Administration (CMA), etc. Each of those agencies uses different coupled GCMs and ESMs to obtain the precipitation forecasts at the S2S range with different temporal and spatial resolutions.

The North America Multi Model Ensemble (NMME) is a multi-agency initiated and collaborative program that provides a variety of hydrometeorological forecasts at the S2S timescale (Kirtman et al., 2014a). The NMME consists of the outputs from multiple coupled GCMs and ESMs, each providing independent retrospective forecasts (hereafter referred to as "hindcasts") and real-time forecasts. The NMME phase 1 project (NMME-1) was initially launched in 2014 and further transitioned into Phase 2 (NMME-2) in more recent years. Both NMME-1 and NMME-2 datasets provide monthly initialized hindcasts and forecasts with lead-time up to 12 months. The major advancement of NMME-2 over the NMME-1 dataset is the provision of dynamically downscaled forecasts, in which the new dataset provides daily precipitation forecasts at the S2S range, while the outputs from the NMME-1 dataset only provides forecasts with monthly resolution (Kirtman et al., 2014b).

There is a good number of existing research that investigated the quality and accuracy of the hydrometeorological forecasts from NMME-1 dataset. For example, Becker et al. (2014) and Krakauer (2019) evaluated the skill of precipitation and temperature forecasts from NMME-1 globally. And they found that the simple model averaging (SMA) of NMME-1 models shows better forecast skill than any individual NMME-1 models and the forecast skills vary depending on geographical regions and seasons. Similar evaluation studies upon NMME-1 monthly hydrometeorological forecasts have been carried out in different regions. For example, Cash et al. (2019) observed significant systematic error in both precipitation and temperature forecasts from NMME-1 in two Southern Asia regions and found the highest forecast skills are observed at the shortest lead times. Shukla et al. (2019) found the skill of precipitation forecasts from NMME-1 is higher during ENSO years over East Africa. Slater et al. (2019) evaluated the skill of precipitation and temperature forecasts from NMME-1 over seven geological regions of the continental United States. Slater et al. (2019) found the highest forecast skill is generally observed at the shortest lead time and the performances of NMME-1 forecast are spatially and seasonally dependent. The study from Slater et al. (2019) also consistently shows that higher forecast skills can be gained by averaging multiple NMME-1 models and the skill of hydrometeorological forecast from NMME-1 quickly declines to marginal levels as lead time increases. To address the low forecast accuracies issue associated with the NMME-1 dataset, many follow-on studies focused on improving precipitation forecast from NMME-1 dataset. For example, Slater et al. (2017) deployed different multi-model weighting techniques to improve the skill of NMME-1 monthly precipitation forecast across Europe. Xu et al. (2019) applied several machine learning and wavelet approaches to bias-correct and downscale the monthly precipitation forecast from the NMME-1 dataset over China. Khajehei et al. (2018) developed a Bayesian ensemble approach based on a Copula function to bias-correct the NMME-1 monthly precipitation forecast over the CONUS.

As compared to the studies on NMME-1 dataset, there is fewer studies that investigate the quality of S2S precipitation forecasts from the newer NMME-2 dataset. Among a limited number of studies, Wanders and Wood (2016) evaluated the precipitation forecast from NMME-2 globally on a bi-weekly basis. By aggregating the evaluation result into three global regions, including the tropics, extratropic and northern latitudes, they found the forecasts skill decreases over lead time as well as performance discrepancies between different NMME-2 models. Zhou and Kim (2018) evaluated the ability of NMME-2 in predicting the wintertime atmospheric rivers (AR) and moisture flux over the Northeast Pacific in response to ENSO. They found NMME-2 dataset has significant regional biases in anomalous landfalling AR frequency which underlining challenges in forecasting regional precipitation events. More recently, Baker et al. (2019) studied the precipitation forecasts of Climate Forecast System version 2 (CFSv2) from the NMME-2 dataset over the entire CONUS on a bi-weekly basis, and they found the forecast skill of CFSv2 decreases over lead time rapidly, but the forecast biases are insensitive to lead times. Guo and Nie (2020) evaluated the daily precipitation forecasts of CFSv2 over east China. Their result indicates the raw precipitation forecasts of CFSv2 are substantially biased and the extreme precipitation events over east China have been underestimated by CFSv2. Becker et al. (2020) studied the performances of precipitation forecasts from NMME-2 globally. However, Becker et al. (2020)

aggregated the daily forecast values from NMME-2 into monthly values since their study focus on seasonal scale and is more focused on proving the improvement of forecast quality corresponding to the iteration of NMME models. To summarize the existing studies focusing on precipitation forecasts from NMME-2, existing research either (1) targeted the evaluation of NMME-2 at a spatial scale that is too large to provide a useful reference for regional hydrologic studies, or (2) are only focusing on monthly forecast values and overlook the S2S forecasts from NMME-2, or (3) have only included a certain member of NMME-2 into the study and did not comprehensively evaluate all NMME-2 members as a whole with the consideration of lead times, seasonality and precipitation's geographical characteristics.

Thus, more inclusive and comprehensive evaluation of the S2S precipitation forecast from the NMME-2 data across CONUS is still critically needed. According to the conclusion of many existing studies, although the evaluation results of the seasonal precipitation forecast from NMME over CONUS may have some consistency and similarities, it is still unknown how exactly does precipitation forecast from different NMME-2 models perform at the S2S range over the entire CONUS. In addition, the merit of the S2S precipitation forecast from NMME-2 in forecasting extreme events has not been verified at large scales. And these missing pieces of research, in return, limits further hydrologic applications of the NMME-2 dataset, since potential maximum streamflow prediction is one of the most important and desired outcomes of hydrologic forecasts at the S2S range (Day 1985, Gobena and Gan 2010).

To fill the gap, as well as to provide valuable reference information and knowledge for future hydrologic research on NMME-2, this study aims to answer the following research questions: 1) How does the S2S precipitation forecast from NMME-2 perform over the entire CONUS? 2) What are the differences between S2S precipitation forecasts from different NMME-2 models in terms of their performances over CONUS? 3) What is the forecast skill of the individual NMME-2 models at different regions over the CONUS? 4) Do certain NMME-2 models outperform others with the consideration of certain regions, lead times, and seasons? and 5) what are the NMME-2 model's performances in predicting extreme precipitation events over the CONUS?

To answer these research questions, in this study, we evaluated the S2S precipitation forecasts from five NMME-2 models and their grand ensemble (i.e., all five NMME-2 models as a whole set) is collected and analyzed over the CONUS. The study period is from 1982 to 2011. All five NMME-2 models selected in this study provide daily S2S precipitation forecast, except for the CFSv2, which was already studied by Baker et al. (2019). The AN81d dataset generated from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) is used as the reference dataset. The forecast bias and forecast skill are examined since forecast bias and poor forecast skill are two major obstacles in applying precipitation forecasts to hydrologic simulations (Zalachori et al., 2012). In addition, the ability of S2S precipitation forecasts from NMME-2 in predicting extreme precipitation events are further evaluated and compared to the benchmark performances of the historical resampled precipitation forecasts. During our evaluation study, the forecast lead time is considered on a weekly basis (i.e., from week 1 to week 4 to cover the whole S2S range). Comparison and analysis on the forecast data quality are further conducted over nine National Centers for Environmental Information (NCEI) climate regions and four seasons, which are more inclusive and detailed as compared to the existing NMME-2 evaluation studies mentioned above. For example, this study extends the existing studies from Baker et al. (2019) and Wanders and Wood (2016). Specifically, our study conducts a more representative validation of five NMME-2 models, emphasizing on the forecast performance evaluation at the weekly scale and over the entire CONUS. In other words, our study provides a temporally-finer and spatially larger evaluation as compared to that from Baker et al. (2019) and Wanders and Wood (2016). Lastly, besides the traditional evaluation of forecast biases and skill scores, we also included extensive validation experiments focusing on the extreme rainfall performance and compared the forecast accuracy on different percentile thresholds of dry and wet extremes.

The rest of this paper is organized as follows: In section 2, we present data and study regions. Section 3 describes the evaluation metrics and methodologies. Sections 4 and 5 present the results and discussions, respectively. The main conclusions and findings are summarized in Section 6.

2. Data and study regions

There is a total of seven different models available in the NMME-2 dataset. Among them, we select five NMME-2 models that provide daily precipitation forecasts covering the S2S range. Table 1 presents the basic information of the selected five NMME-2 models, including the Canadian Coupled Climate Model version 3 and 4 (CanCM3, CanCM4) from the Canada's Climate Model Center (CMC), the Community Climate System Model 4.0 (CCSM4) from the National Center for Atmospheric Research (NCAR), the Forecast-oriented Low Ocean Resolution model using parameter set B (FLORB01) from the Geophysical Fluid Dynamics Laboratory (GFDL), and the Goddard Earth Observing System version 5 model (GEOS5) from the National Astronautics and Space Administration (NASA).

Each NMME-2 model generates ensemble forecasts through perturbed physics strategy and/or under different initial conditions: The CanCM3, CanCM4, CCSM4, and GEOS5 models consist of 10 ensemble members and the FLORB01 model consists of 12 ensemble members. The study period was set from 01/01/1982 to 12/31/2011 (30 years), which overlaps with the hindcast/forecast period for all five NMME-2 models. All NMME-2 hindcast/forecast datasets are available at: https://www.earthsystemgrid.org/search.html?Project=NMME.

The daily precipitation dataset AN81d generated from the Parameter-elevation Regressions on Independent Slopes Model (hereafter referred to as PRISM) is used as a reference precipitation dataset in this study. The PRISM data is available from 1981 to near-present in gridded format with a spatial resolution of 4 km (~0.04°) across the CONUS. The PRISM data combines surface observations with a digital elevation model to account for the orographic enhancement of precipitation. In addition to rain-gauge records, the PRISM data also incorporates the Radar measurement into account when producing corrected data over the central and eastern U.S. regions (Daly and Bryant 2013). Since the PRISM dataset does not incorporate assimilated information from numerical weather forecasting models or meteorological reanalysis, it represents an independent dataset suitable for hydrologic studies (Radcliffe and Mukundan 2017). Numerous hydrologic studies have used PRISM precipitation data as a reliable reference for model evaluation, bias-correction for remotely sensed precipitation estimation products, and forecast verification studies (Ashfaq et al., 2016, Oubeidillah et al., 2014, Prat and Nelson 2015, Widmann and Bretherton 2000).

Figure 1 shows nine different climate regions across the CONUS, which are defined by the National Centers for Environmental Information (NCEI) (Karl and Koss 1984). These nine climatic regions separate the CONUS into Northwest, West North Central, East North Central, Northeast, Central, West, Southwest, South, and Southeast regions. Within the NECI climate regions, the Sierra Nevada Mountain and the Rocky Mountain are across the Northwest, West, Southeast, and West North Central regions; and the Appalachian Mountain covers parts of the Northeast and Southeast regions. In this study, we evaluate the precipitation forecast of the NMME-2 at each pixel across the CONUS and obtain the spatially averaged results over these nine climate regions for regional analysis. In this study, we analyze the results based on these nine climatic regions, because rainfall presents different physical and dynamical features and patterns over different regions over the CONUS, i.e., orographic elevation induced rainfall, frontal precipitation, and/or, convective systems.

Table 1
A list of selected NMME-2 models.

Model	Data period	Temporal resolution	Spatial resolution	Lead time (Days)	Ensemble members	Reference
CanCM3	01/1981-08/2012	Daily	1°×1°	up to 365	10	Merryfield et al. (2013)
CanCM4	01/1979-08/2012	Daily	1°×1°	up to 365	10	Merryfield et al. (2013)
CCSM4	01/1982-12/2016	3-hourly/Daily	1°×1°	up to 365	10	Vertenstein et al. (2010)
FLORB01	01/1980-07/2014	Daily	0.5°×0.625°	up to 365	12	Delworth et al. (2012)
GEOS5	01/1982-12/2012	Daily	1°×1°	up to 274	10	Vernieres et al. (2012)

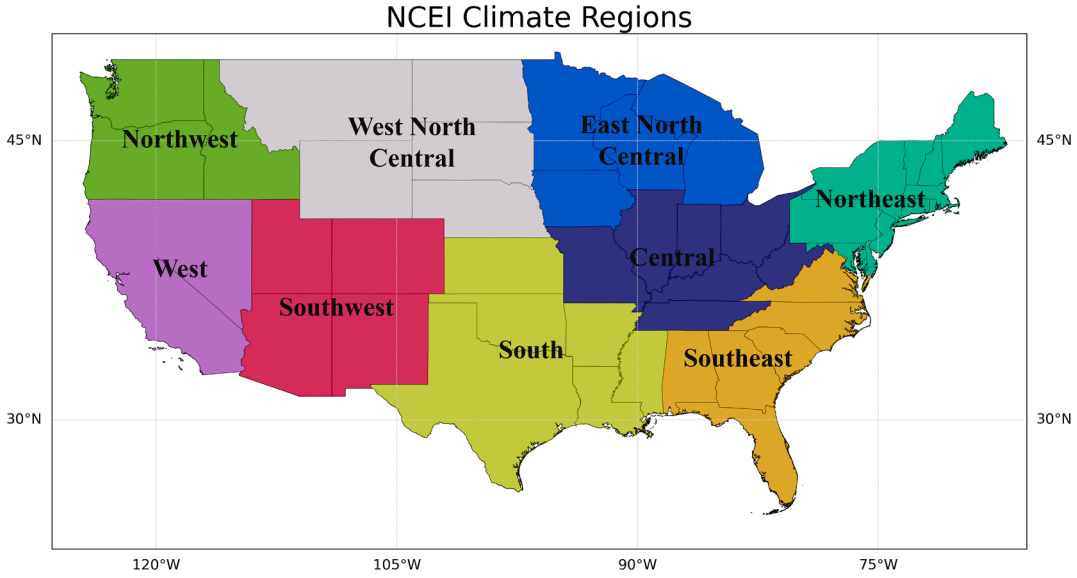

Fig. 1. NECI Climate Regions across the CONUS.

3. Methodology, procedures, and evaluation metrics

In this study, we first collected the NMME-2 precipitation forecasts and PRISM dataset and then conducted initial data pre-processing. The NMME-2 precipitation forecasts are produced at the beginning of each month and the forecast lead time are up to 1 year (365 or 366 days) into the future. We truncated all collected NMME-2 precipitation forecasts to 28 days (4 weeks) and then aggregated them into weekly forecast values (e.g., day 1 to 7 as of week 1; day 8 to 14 as of week 2; day 15 to 21 as of week 3; and day 22 to 28 as of week 4). Both NMME-2 and PRISM datasets were re-gridded into 0.25° resolution using the same nearest neighbor method for consistency. The grand ensemble of all selected NMME-2 models (i.e., a total of 52 different realizations in Table 1) was also constructed after the data pre-processing. We also re-organized the collected precipitation forecasts from both individual NMME-2 models and the 52-member grand ensemble into different seasons, i.e., December-January-February (DJF), March-April-May (MAM), June-July-August (JJA), and September-October-November (SON). In other words, the evaluation experiment comprehensively considers the forecast skill and model bias by different climate regions, forecast lead times, as well as the seasonality over the CONUS.

In this study, we use the commonly accepted approaches of pixel-based and spatial forecast evaluation metrics. Four evaluation metrics are included, i.e., the percentage bias (PBIAS), anomaly correlation coefficient (ACC), quantile probability of detection (QPOD), and quantile false alarm ratios (QFAR). The PBIAS and ACC of (i) the ensemble means of individual NMME-2 models and (ii) the SMA of the grand ensemble, are computed to evaluate forecast bias and forecast skill quantitatively. The QPOD and QFAR metrics are used to evaluate the capabilities of individual NMME-2 model in predicting extreme precipitation events at weekly scales. The evaluation of extreme precipitation is as important as forecast bias and forecast skill, because the S2S precipitation forecasts potentially serve as inputs to the ensemble streamflow prediction (ESP) approach, which are the official approach used in each National Weather Service's River Forecast Centers for estimating river stages and potential floods over the CONUS. Within the ESP framework, the extreme streamflow values associated with extreme precipitation events are one of the most important outcomes regarding flood predictions (Day 1985, Gobena and Gan 2010). With this understanding, in this study, the ensemble spreads of individual NMME-2 models and the grand ensemble of five NMME-2 models are employed to compute the QPOD and QFAR for extreme precipitation evaluation. While there is no golden standard for defining extreme precipitation events, we chose 99% and 95% percentiles, and 5% and 1% percentiles as the thresholds of extreme precipitation events corresponding to flood and drought events, respectively. Note that we only present the QPOD and QFAR results for extreme events above 99% and below 1% in the main article for conciseness, and the 95% and 5% events results are included in the supplementary material for interested readers. Detailed descriptions for the four employed evaluation metrics are presented as follows.

3.1. Percentage bias (PBIAS)

The PBIAS measurement reflects the degree of the under- and/or over- estimations of precipitation forecast that are vital for potential future hydrologic applications. To quantify the bias pattern of NMME-2 precipitation forecasts over the CONUS, the PBIAS of the ensemble means of single NMME-2 models and the SMA of the grand ensemble are computed with Equation (1).

$$PBIAS = \frac{\overline{x} - \overline{y}}{\overline{y}} \times 100\% \quad (1)$$

Where $\overline{x}$ is long-term mean value of precipitation forecasts at a certain

location over CONUS, $\bar{y}$ is the long-term mean reference precipitation at a certain location over CONUS. And PBIAS is the percentage differences between mean forecast values and mean reference values over the study period. Positive PBIAS values indicate overestimations by forecasts, while negative PBIAS values indicate underestimations by forecasts.

3.2. Anomaly correlation coefficient (ACC)

The ACC is a widely used metric in the climate prediction community. It measures the degree of association between forecast and observed deviation from the climatology. The advantage of ACC over some other metrics is that ACC can separate effects due to the existence of forecast bias in evaluating forecast skill. The ACC score of 1 indicates that the forecast provides perfect information and a score of zero means the forecast contains no information at all. The ACC skill scores of the ensemble mean of each individual NMME-2 model and the SMA of the grand ensemble are calculated following Equation (2) (Murphy and Epstein 1989).

$$ACC = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} - \sqrt{n\sum y^2 - (\sum y)^2}} \quad (2)$$

where x is the forecast/hindcast precipitation anomalies at a certain lead time and y is reference precipitation anomalies at the same lead time, n is the total number of hindcasts/forecast values made for that lead time, and ACC is the anomaly correlation coefficient skill score for the forecasts/hindcasts.

3.3. Quantile probability of detection (QPOD), quantile false alarm ratio (QFAR)

3.3.1. Quantile probability of detection (QPOD)

The QPOD is a statistical evaluation measurement, which is defined as the probability of detection (POD) above a certain quantile threshold (AghaKouchak et al., 2011, Wilks 2011). In this study, we used the whole ensemble spreads of NMME-2 models to compute the QPOD. Taking the CanCM3 model and extreme events above 99% percentiles as an example: CanCM3 model produces ten forecast values at a certain time step, and if any one of the ten forecast values has successfully forecasted an extreme event exceeded 99% quantile according to its own model output statistics, it will be count as a "hit". The QPOD has the advantage of ignoring the effect of forecast bias as it is a quantile-based evaluation metric compared to the conventional probability of detection (POD) (AghaKouchak et al., 2011, Wilks 2011). The QPOD ranges from 0 to 1 and the value of 1 is ideal, indicating all extreme events above/below a certain percentile threshold are successfully forecasted. The QPOD of single NMME-2 models and the grand ensemble of five NMME-2 models considering their whole ensemble spreads can be calculated with the following Equation (3):

$$QPOD = \frac{\sum_{i=1}^{n} \cup_{j=1}^{m} I(x_{ij} > x_{j-99}) \cap I(y_i > y_{99})}{\sum_{i=1}^{n} I(y_i > y_{99})} \quad (3)$$

Where n is the length of the forecast time series, and m is the ensemble size of a certain forecast model. x_{ij} is the forecast value by jth ensemble member of a model at time step i, and x_{j-99} is the 99% threshold of the jth ensemble member according to its own statistics. Similarly, y_i is the reference value at time step i, and y_{99} is the 99% threshold of the reference precipitation according to its own statistics. I is the indicator function (e.g., $I(\text{true}) = 1, I(\text{false}) = 0$), $\cap$ and $\cup$ represent set operations (e.g., $1 \cap 1 = 1$, $0 \cup 1 = 1$).

3.3.2. Quantile false alarm ratio (QFAR)

The QFAR is a categorical evaluation measurement, which is defined as the False Alarm ratio (FAR) above a certain quantile threshold (Mehran and AghaKouchak 2014). In this study, we used the whole ensemble spreads of NMME-2 models to compute the QFAR. Again, taking the CanCM3 model as an example, which has ten ensemble members: the CanCM3 model produces ten forecast values at a certain time step, and if any one of the ten forecast values has made a forecast exceeded 99% quantile according to its own model output statistics while there's no extreme event happened according to the reference, it will be count as a "false alarm". Since adopting quantile thresholds, the QFAR also has the advantage of ignoring the effect of forecast bias compared to the conventional false alarm ratio (FAR) (AghaKouchak et al., 2011, Wilks 2011). The QFAR ranges from 0 to 1 and the value of 0 is ideal, indicating there's no "false alarm" at all. The QFAR of single NMME-2 models and the grand ensemble of five NMME-2 models considering their whole ensemble spreads can be calculated with the following Equation (4).

$$QFAR = \frac{\sum_{i=1}^{n} \cup_{j=1}^{m} I(x_{ij} > x_{j-99}) \cap I(y_i < y_{99})}{\sum_{i=1}^{n} I(y_i < y_{99})} \quad (4)$$

Where n is the length of the forecast time series, and$m$ is the ensemble size of a certain forecast model. x_{ij} is the forecast value by jth ensemble member of a model at time step i, and x_{j-99} is the 99% threshold of the jth ensemble member according to its own statistics. Similarly, y_i is the reference value at time step i, and y_{99} is the 99% threshold of the reference precipitation according to its own statistics. I is the indicator function (e.g., $I(\text{true}) = 1, I(\text{false}) = 0$), $\cap$ and $\cup$ represent set operations (e.g., $1 \cap 1 = 1$, $0 \cup 1 = 1$).

3.3.3. Benchmarking QPOD and QFAR

In this study, we benchmark the QPOD and QFAR values of historical resampled precipitation forecasts in predicting extreme precipitation events. The historical resampled forecasts are commonly used as hydrometeorological inputs to the ESP framework for hydrologic forecasts at S2S range. Assuming historical resampled precipitation forecast with an ensemble size of m, if any single randomly drawn forecast values out of m forecasts values have successfully predicted an extreme event above/below a certain threshold according to the historical statistics, it will be counted as a "hit". Similarly, if any single randomly drawn forecast values out of m forecasts values contain a value above/below a certain threshold while there's no such extreme events happened according to the reference, it will be counted as a "false alarm".

Since historical resampled forecasts are randomly drawn values from historical records, they are totally independent of the actual weather happened in the real world. Thus, theoretically, for any true positive and/or true negative events, the QPOD and/or QFAR of historical resampled forecasts are the same and can be computed with Equation (5):

$$QPOD_{resampledforecasts} = QFAR_{resampledforecasts} = 1 - 0.99^m \quad (5)$$

According to Equation (5), the $QPOD_{resampledforecasts}$ and $QFAR_{resampledforecasts}$ with an ensemble size of 10 (CanCM3, CanCM4, GEOS5, CCSM4), 12 (FLORB01), and 52 (the grand ensemble of NMME-2) are 0.10, 0.11, 0.40, respectively, for the extreme precipitation events above 99% or below 1% percentiles. Since the historical resampled forecasts randomly draw values from historical records, their performances are only affected by the ensemble size but are independent of forecast lead times.

4. Results

4.1. Forecast bias

Figure 2 presents the overall PBIAS of NMME-2 precipitation hindcasts over CONUS. It consists of the results from five individual NMME-2 models (CanCM3, CanCM4, FLORB01, GEOS5, and CCSM4) and the SMA of the grand ensemble of five NMME-2 models. In Fig. 2, the positive bias in cooler colors (blue) is associated with overestimation and

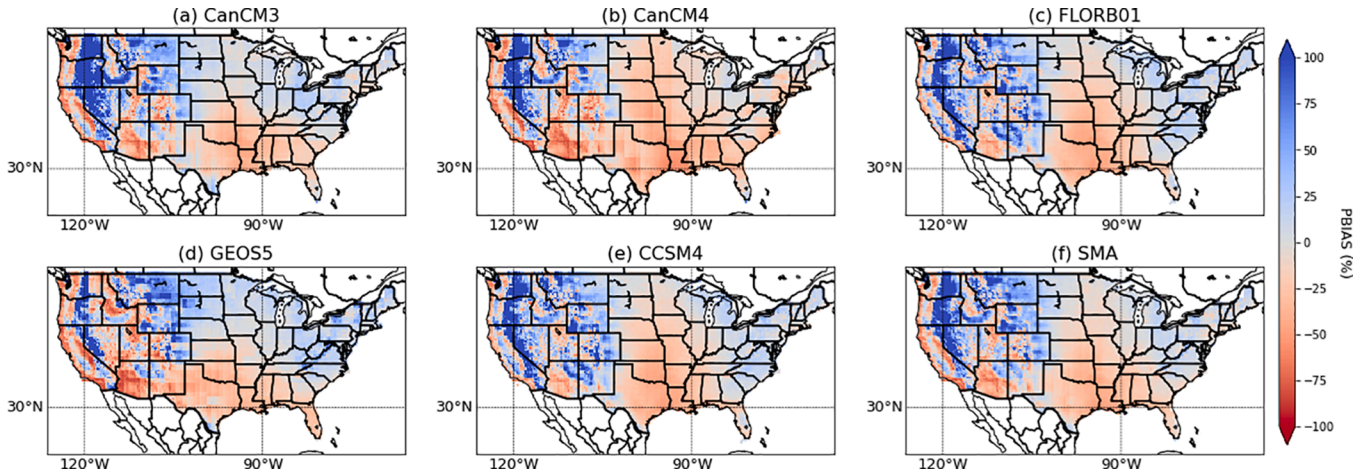

Fig. 2. The overall pattern of PBIAS for NMME-2 precipitation hindcasts over the CONUS.

the negative bias in warmer colors (red) is associated with underestimation.

According to Fig. 2, the highest level of model PBIAS of the NMME-2 dataset appears in central and western U.S., and the forecast biases are relatively lower in eastern regions than that over the western regions. This similar spatial variability of overall PBIAS can be observed across individual NMME-2 models and the SMA. In addition, we notice that all five NMME-2 models demonstrate both positive and negative PBIAS with a "mosaic-like" pattern over the Northwest, West, West North Central, and Southwest regions (Fig. 2a–e), where the Sierra Nevada Mountains and the Rocky Mountains ranges. On the contrary, across the Northwest and Southeast regions where Appalachian Mountain is located, the GEOS5 model (Fig. 2d) demonstrates a significant level of overestimation while other NMME-2 models show a relatively smaller level of biases. In the South region, all five NMME-2 models are consistently underestimating precipitation. The SMA (Fig. 2f) has a very similar pattern of spatial variability in the overall pattern of PBIAS compared to the five individual single NMME-2 models.

The following Fig. 3 presents the PBIAS of NMME-2 precipitation hindcasts across CONUS at different lead times (i.e., week 1 to week 4).

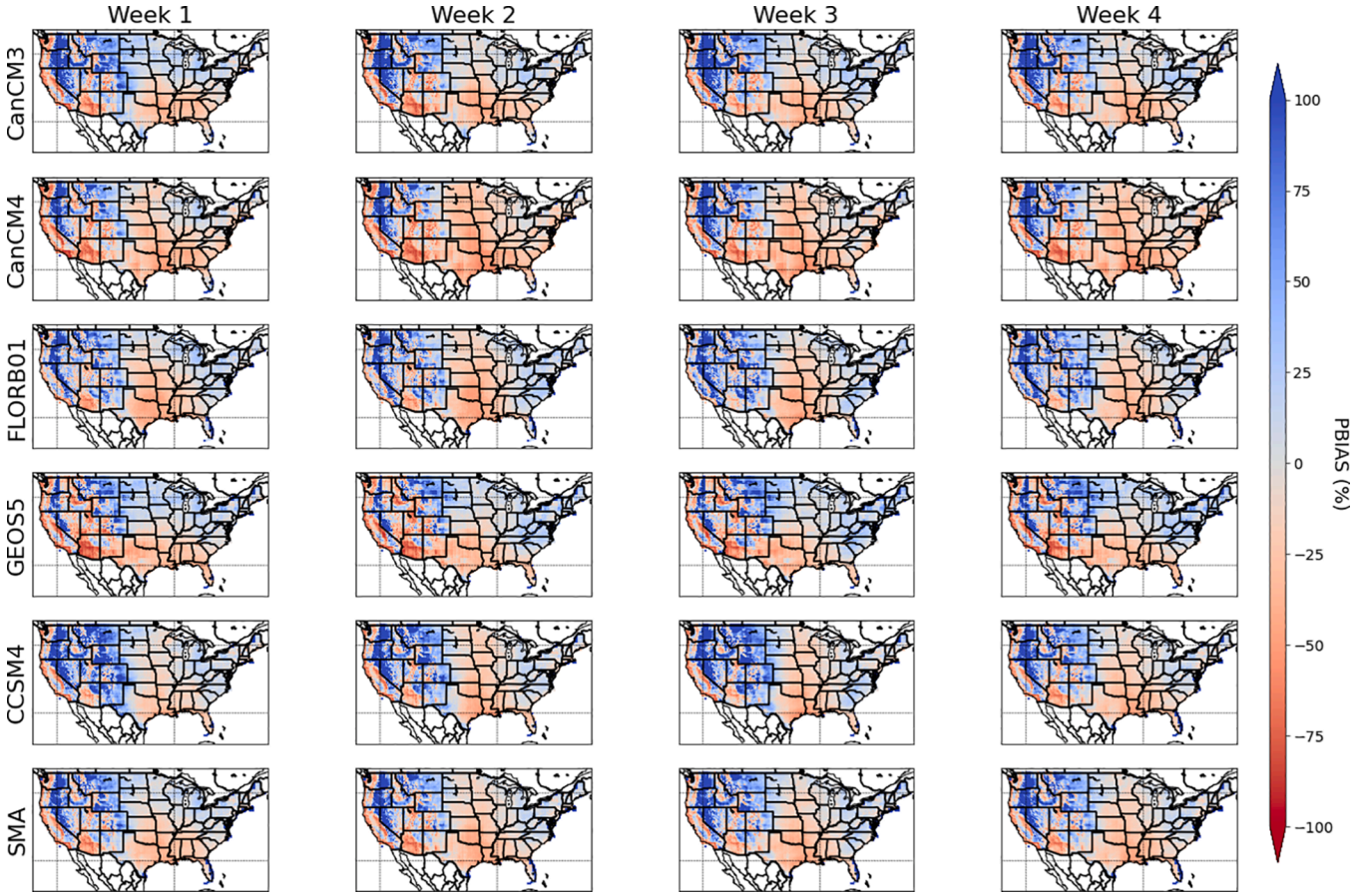

Fig. 3. The pattern of PBIAS for NMME-2 precipitation hindcasts over the CONUS at different lead times (i.e., day 1 to 7: week 1, day 8 to 14: week 2, day 15 to 21: week 3, and day 22 to 28: week 4).

According to Fig. 3, only some minor differences are observed in different climate regions and no obvious lead-time dependences of PBIAS are observed for all NMME-2 members. At all lead times, the spatial pattern of weekly PBIAS is very similar to the overall pattern of PBIAS observed in Fig. 2. The SMA also shows similar behavior to the five individual NMME-2 models at all lead times. Combining the results from both Figs. 2 and 3, we observe that the overall model biases are not sensitive to lead times, though individual model could be associated with different levels of PBIAS, and the model errors are also spatially varying across the CONUS.

In the following Fig. 4, we present the seasonal PBIAS of NMME-2 precipitation hindcasts over CONUS. According to Fig. 4, both the individual NMME-2 model and the SMA exhibit a strong PBIAS variation in different seasons over the CONUS, and the individual model biases may demonstrate significant changes from positive to negative or vice-versa when the season changes. Specifically, in DJF, all NMME-2 models show significant overestimation at most parts of the Northwest, West, West North Central, and Southwest regions. In MAM, the general behavior of the model overestimation is improved and the level of overestimation of all individual models decreased comparing to that in DJF. In MAM, we also notice that the spatial patterns of the obtained PBIAS are very similar to the overall patterns of PBIAS as shown in prior Fig. 2, suggesting the MAM is the most representative season to show model's error variability in space. In JJA, all NMME-2 models show a significant level of underestimation at the Northwest, West, Southwest, South, and West North Central regions. In SON, the level of the underestimation by all models are smaller than that in JJA, while the level of overestimation increased in the Northwest region.

The results in Fig. 4 indicate that there are strong seasonal dependences of model PBIAS at different regions across CONUS, especially in DJF and JJA. Some noticeable model discrepancies are observed in different seasons. For example, in SON, the CanCM4 model exhibits a more significant level of underestimation than other NMME-2 models in the Southeast, South, Southwest, and West North Central regions. In MAM, the GEOS5 model demonstrates a higher level of overestimation than other NMME-2 models in the West North Central region, while a higher level of underestimation is observed with the same GEOS5 model comparing to other models in the Southwest, South, and Southeast regions.

To better illustrate the large-scale spatial patterns of model biases with a joint consideration of seasonality and lead time dependences, we further computed and presented the regional averaged PBIAS values of NMME-2 over nine NCEI climate regions for different seasons (DJF, MAM, JJA, and SON) and lead times (weeks 1–4) in the following Fig. 5. In Fig. 5, the cooler colors indicate overestimation, while the warmer colors indicate underestimation. The numbers in each cell of Fig. 5 indicate the actual PBIAS values obtained for each model, lead time, and season. According to the heat map of Fig. 5, the largest overestimations made by all NMME-2 models occur in the West North Central region in DJF. And significant underestimations are observed in the warmer seasons of SON and JJA at most of the regions over CONUS except for the Northwest. The layout pattern of nine NCEI climate regions in Fig. 5 (from left to right) is consistent with the geographic patterns of those regions in the real-world layout (from west to east). According to the quantitative values presented in Fig. 5 (from left to right), there is an obvious decreasing trend of PBIAS from west to east. This numerical trend and decreasing pattern are more obvious than that of different seasonality and lead times. Specifically, the overall overestimation in the Northwest, West, and West North Central regions observed in Fig. 2 appears to be mainly contributed by the overestimations that occurred in DJF. And the underestimation observed in Fig. 2 appears to be mainly contributed by the underestimations that occurred in warmer seasons of

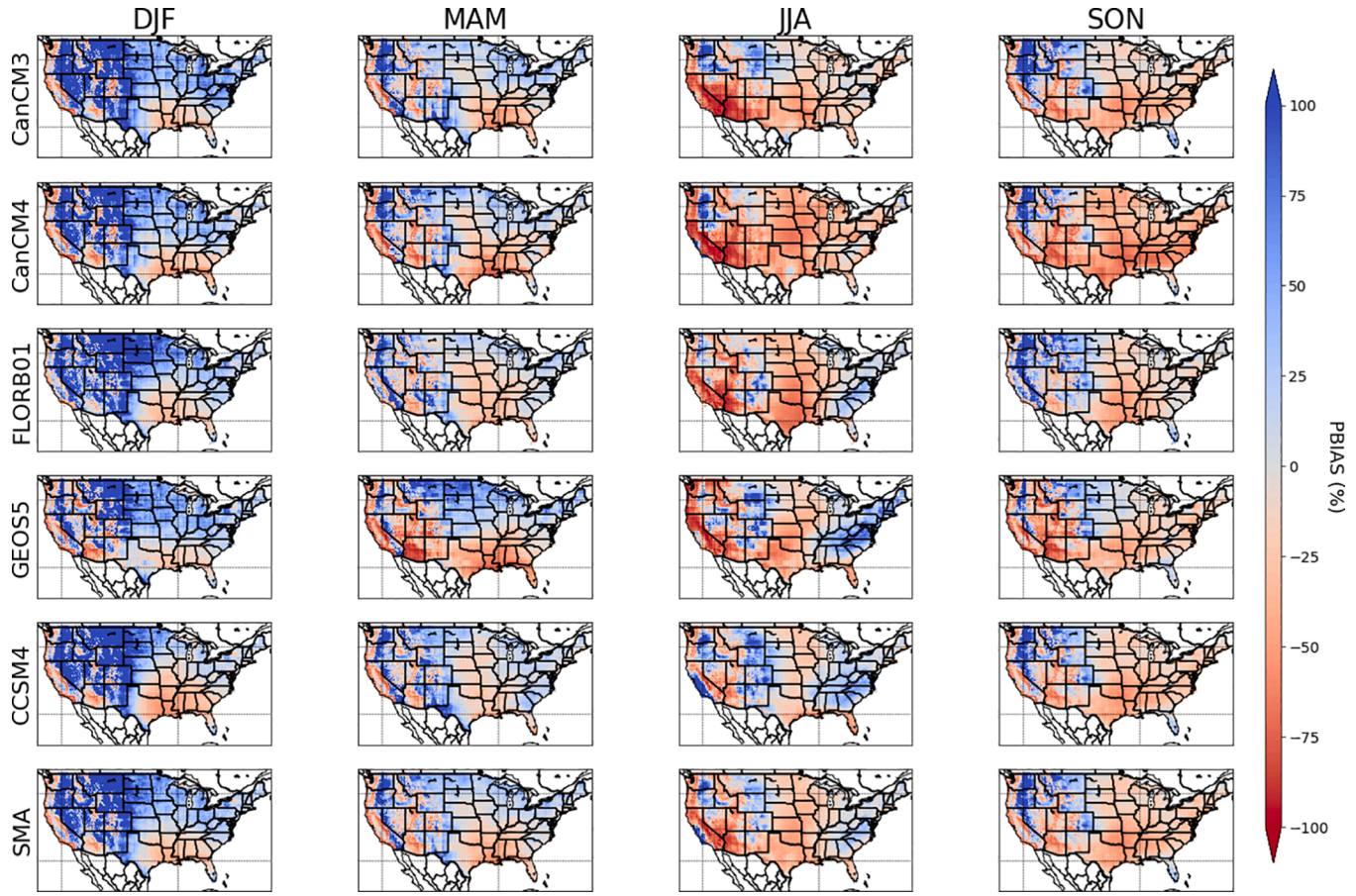

Fig. 4. The PBIAS of NMME-2 precipitation hindcasts within different seasons (i.e., DJF, MAM, JJA, SON) across CONUS.

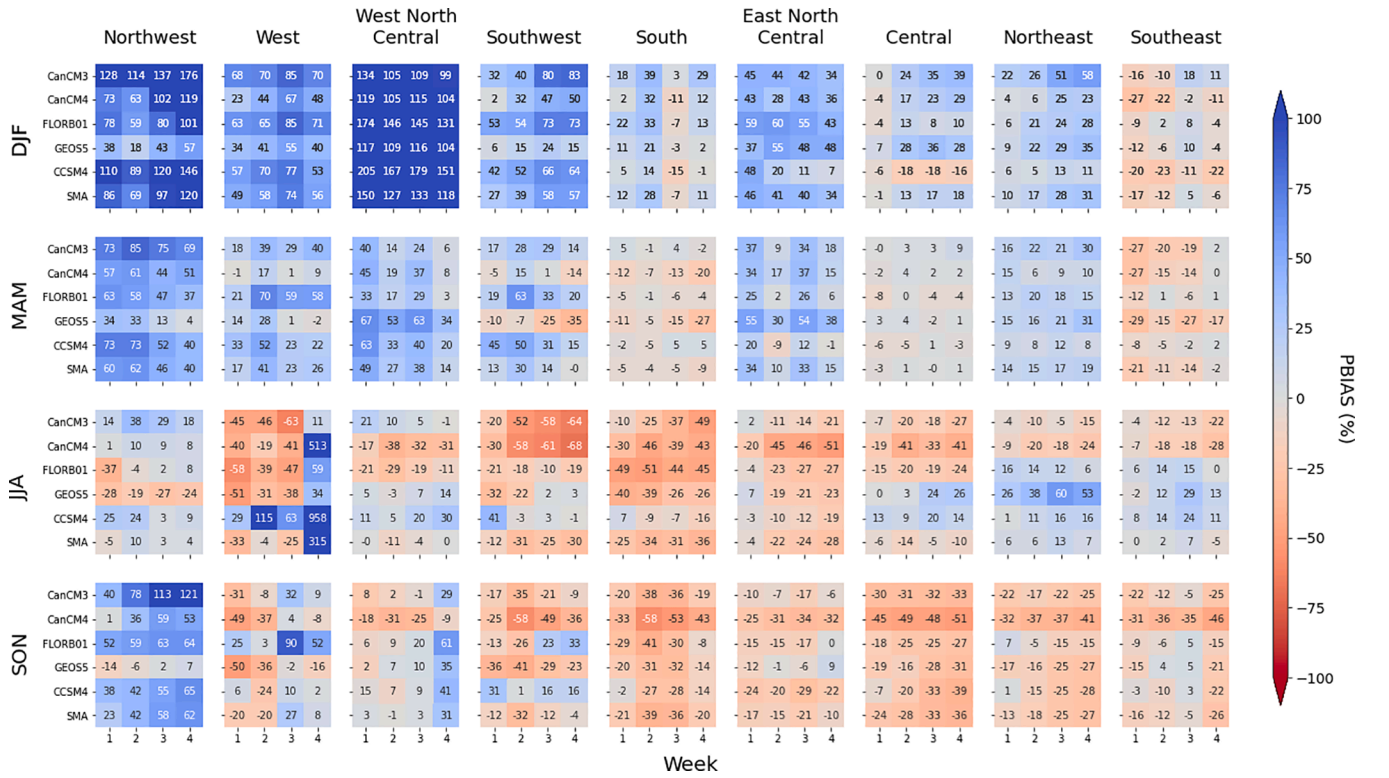

Fig. 5. Regional averaged PBIAS of NMME-2 precipitation hindcasts from the five NMME-2 models and an SMA model for each season (DJF, MAM, JJA, and SON) and lead time (week 1 to 4) at the nine NCEI climate regions.

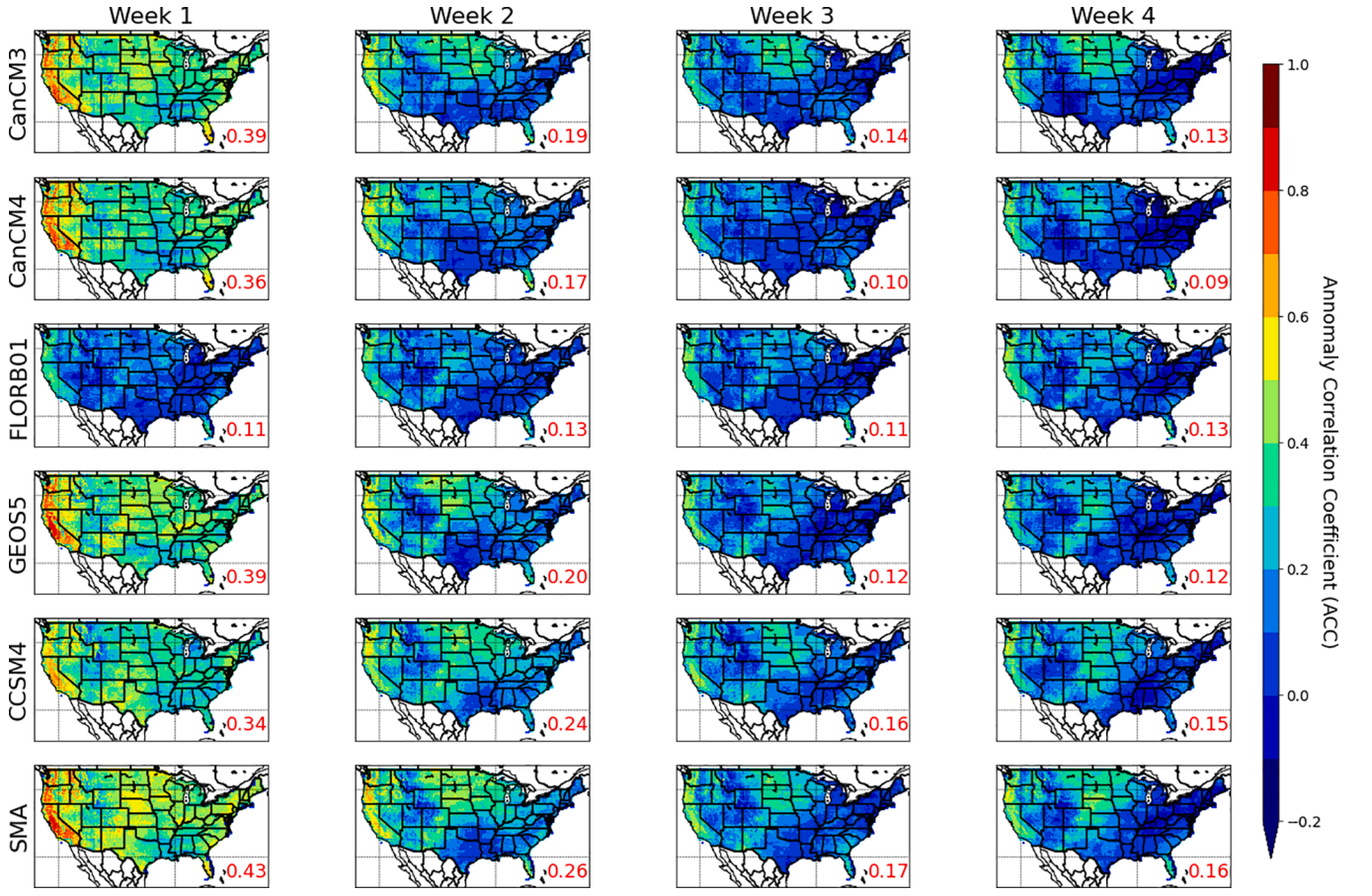

Fig. 6. The pattern of ACC for NMME-2 precipitation hindcasts over the CONUS at different lead times (weeks 1 to 4).

JJA and SON. In terms of the different lead times, we found that the PBIAS values are not sensitive to the lead times since the colors of grid boxes in Fig. 5 remain rather consistent along the x-axis within most of the subplots. In JJA and SON, only CanCM4, FLORB01, CCSM4, and SMA models show some noticeable variations of PBIAS over lead times in the West, the West North Central, and the Southwest regions.

4.2. Forecast skill

To evaluate the forecast skills of NMME-2 precipitation hindcasts at different lead times (week 1 to week 4) across CONUS, we calculated the ACC values of the ensemble means of five NMME-2 models and the SMA of the grand ensemble of NMME-2. The results are presented in Fig. 6. In Fig. 6, warmer colors indicate higher ACC scores and cooler colors indicate lower ACC scores. The red-colored numbers on the lower-right corners of each subplot in Fig. 6 are the spatially averaged ACC value over the entire CONUS.

According to the results in Fig. 6, except for the FLORB01, the forecast skill of all other NMME-2 models shows a decreasing trend over lead times as their CONUS-averaged ACC values decrease rapidly from week 1 (0.34 to 0.43) to week 2 (0.19 to 0.26), and the skill score remains at a marginal level at week 3 (0.10 to 0.17) and week 4 (0.09 to 0.10). The CONUS-averaged ACC values of FLORB01 are consistently lower than other NMME-2 models, especially at week 1 and week 2. The SMA shows a higher forecast skill than any other individual NMME-2 models at all lead times over the entire CONUS.

In addition, the results in Fig. 6 also indicate that spatially, all NMME-2 models have the highest forecast skill over the coastal areas in the Northwest and West regions as compared to other regions in CONUS. Higher ACC values can be observed in the Northwest, West, West North Central, and some part of the Southwest Central regions, as compared with the CONUS-averaged ACC values. However, the ACC values in the South, Central, some parts of the Southwest and Northeast regions are consistently lower than in other regions, especially at the longer lead times (i.e., week 2 to week 4).

In the following Fig. 7, we present the ACC values in different seasons of (a)DJF, (b)MAM, (c)JJA, and (d)SON at different lead times (week 1 to week 4). The difference between Fig. 7 and Fig. 6 is that the prior Fig. 6 differentiated the forecast lead times but did not separate the forecasts made in different seasons. In Fig. 7, we further grouped the forecasts into different seasons and lead times for evaluation. The warmer colors in Fig. 7 indicate higher ACC scores and cooler colors in Fig. 7 indicate lower ACC scores, and the red-colored numbers on the lower-right corners of each subplot in Fig. 7 are the spatially averaged ACC value over the entire CONUS.

In Fig. 7, the weekly pattern of the ACC values in all seasons generally follows similar behaviors as that shown in Fig. 6. The ACC values of the NMME-2 models decrease rapidly from week 1 to week 2 and show marginal forecast skill in week 3 and week 4 in all seasons. The FLORB01 is still an outlier compared with the other NMME-2 models as it continuously presents lower CONUS-averaged ACC values. The SMA of NMME-2 shows the highest CONUS-averaged ACC values at almost all lead times in all seasons. It is also found that there are seasonal dependences in the forecast skills according to the lead times. Taking SMA results as an example, the highest ACC values at week 1 and week 3 are observed in DJF and MAM, respectively. Although overall marginal

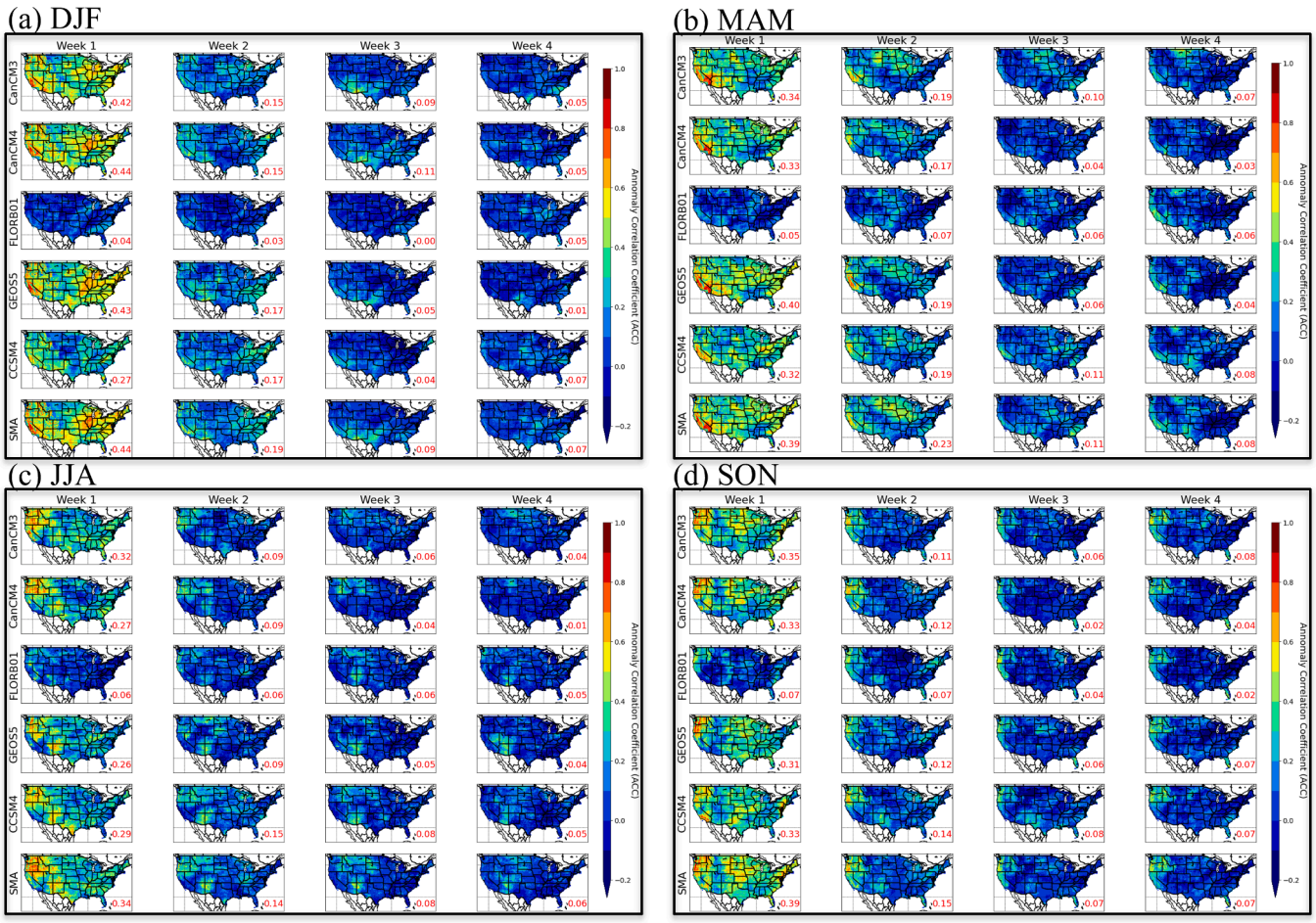

Fig. 7. The pattern of ACC for NMME-2 precipitation hindcasts within different seasons (i.e., DJF, MAM, JJA, SON) at different lead times (weeks 1 to 4) over the CONUS.

forecast skill is observed at week 3 and week 4 in all seasons, a moderate forecast skill (0.3 to 0.4) can be observed at some areas within different seasons. In addition, at week 3 and week 4, moderate forecast skills are observed (i) over the South, Southwest, some parts of the West, the West North Central, and the Southeast in DJF; (ii) over the North Central, the South, and the coastal area of the West regions in MAM; (iii) over the Southwest and the West North Central regions in JJA; and (iv) over the Northwest, and the coastal areas of the West regions in SON, respectively.

Figure 8 presents the regionally averaged ACC scores of NMME-2 precipitation hindcasts in different seasons (DJF, MAM, JJA, and SON) and at different lead times (week 1 to week 4) over nine NCEI climate regions. In Fig. 8, the lighter colors indicate lower ACC values, and the darker colors indicate higher ACC values. The numbers in the figure are the actual ACC scores. Similar to the presentation of Fig. 5, Fig. 8 follows the pattern that the layout of nine NCEI climate regions from left to right in Fig. 8 corresponds to the real-world geography from western to eastern U.S. per the real-world layout.

The results in Fig. 8 indicate a spatially varying pattern of the precipitation forecast skill of NMME-2 models, when the seasonality and lead time dependences are jointly considered. Specifically, in Fig. 8, there is a significant decreasing trend of ACC values from the west to the east, especially in MAM and JJA. In addition, we observe that the ACC values are relatively higher at the shortest lead time (i.e., week 1), as compared to that at longer lead times (i.e., week 2 to week 4) across all regions and seasons. According to Fig. 8, in general, the highest overall ACC value is likely to appear in the winter season (i.e., DJF). On the contrary, the NMME-2 models tend to produce the lowest forecast skill in the summer season (i.e., JJA) except for the Northwest and Southwest regions. Except for the FLORB01 model, all other NMME-2 models show similar behaviors in most of the NECI climate regions and the different models show similar performances over most of the lead times. The SMA shows the highest ACC values than any single NMME-2 models under most of the scenarios, which is expected as the SMA approach can eliminate the outliners and produce a more conservative forecast as compared to each single NMME-2 model.

As a summary of this section 3.2, the following three important findings are evidenced by the presented results: (1) the NMME-2 precipitation forecast skill consistently decreases over the lead times across the CONUS in all seasons; (2) The raw forecast skill of NMME-2 dataset is generally higher in cooler seasons than that in warmer seasons; And (3) spatially, the raw forecast skill of all five employed NMME-2 models tends to be higher in western regions than in eastern regions across the CONUS, and the performance of different models is similar to each other, except for the FLORB01 model.

4.3. Capability in predicting extreme precipitation events

We evaluate the NMME-2 models' capabilities in predicting extreme precipitation events using the QPOD and QFAR. Here we obtain and analyze the NMME-2 models' capabilities in predicting extreme precipitation events above 99% and below 1% percentiles, respectively. The results for extreme precipitation events above 95% and below 5% percentiles are available in the Supplementary Materials for interested readers.

4.3.1. Extreme events above 99% threshold

The following Fig. 9 presents the QPOD of NMME-2 precipitation hindcasts in predicting extreme precipitation events above 99% threshold over CONUS at different lead times. In Fig. 9, the warmer colors indicate higher QPOD values, and red-colored numbers at the lower-right corners of each subplot is the spatially averaged QPOD values over the entire CONUS.

The results from Fig. 9 indicate that the QPOD patterns of different NMME-2 models are generally similar to each other with only minor discrepancies at forecast lead time of week 1. Both the GEOS5 and CCSM4 model generally show higher QPOD values than other NMME-2 models at week 1. According to the numerical QPOD values on each

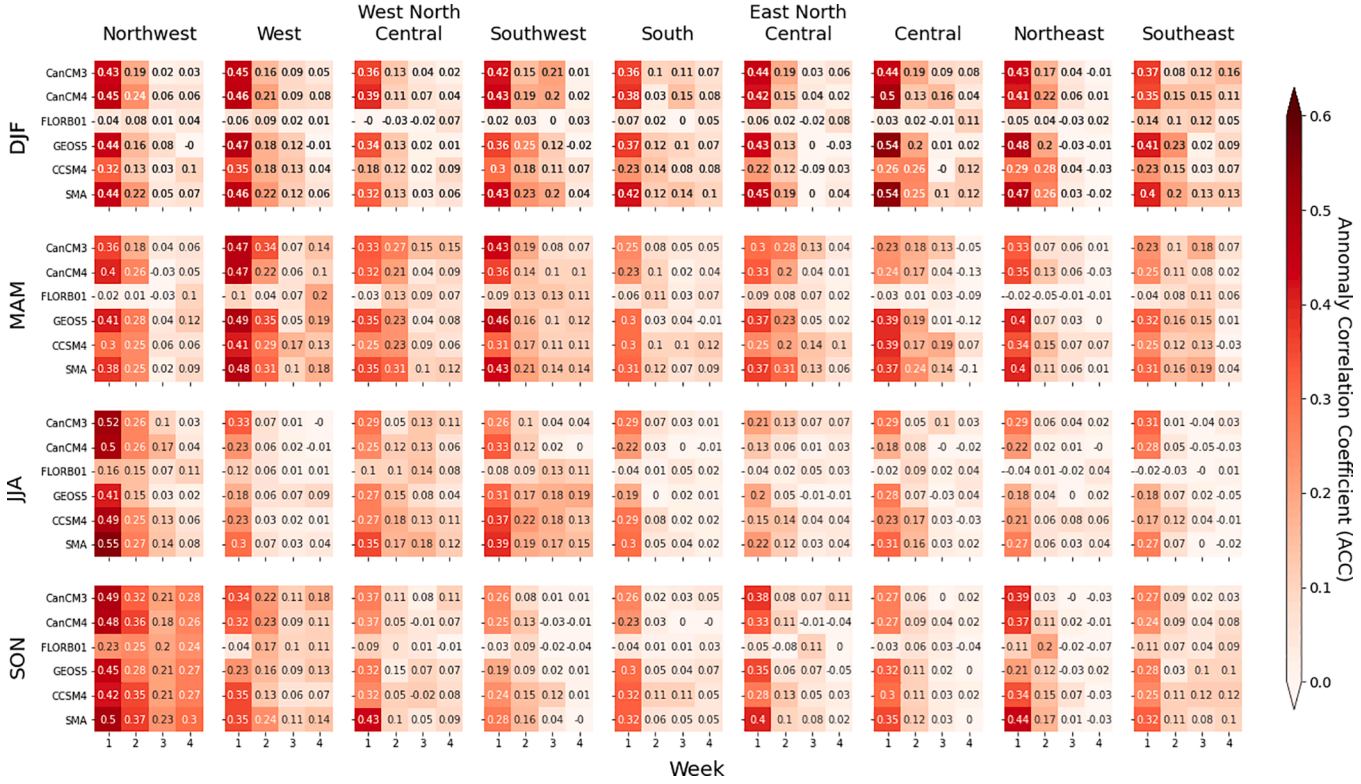

Fig. 8. Regional averaged ACC of NMME-2 precipitation hindcasts from the five NMME-2 models and an SMA model for each season (DJF, MAM, JJA, and SON) and lead time (week 1 to 4) at the nine NCEI climate regions.

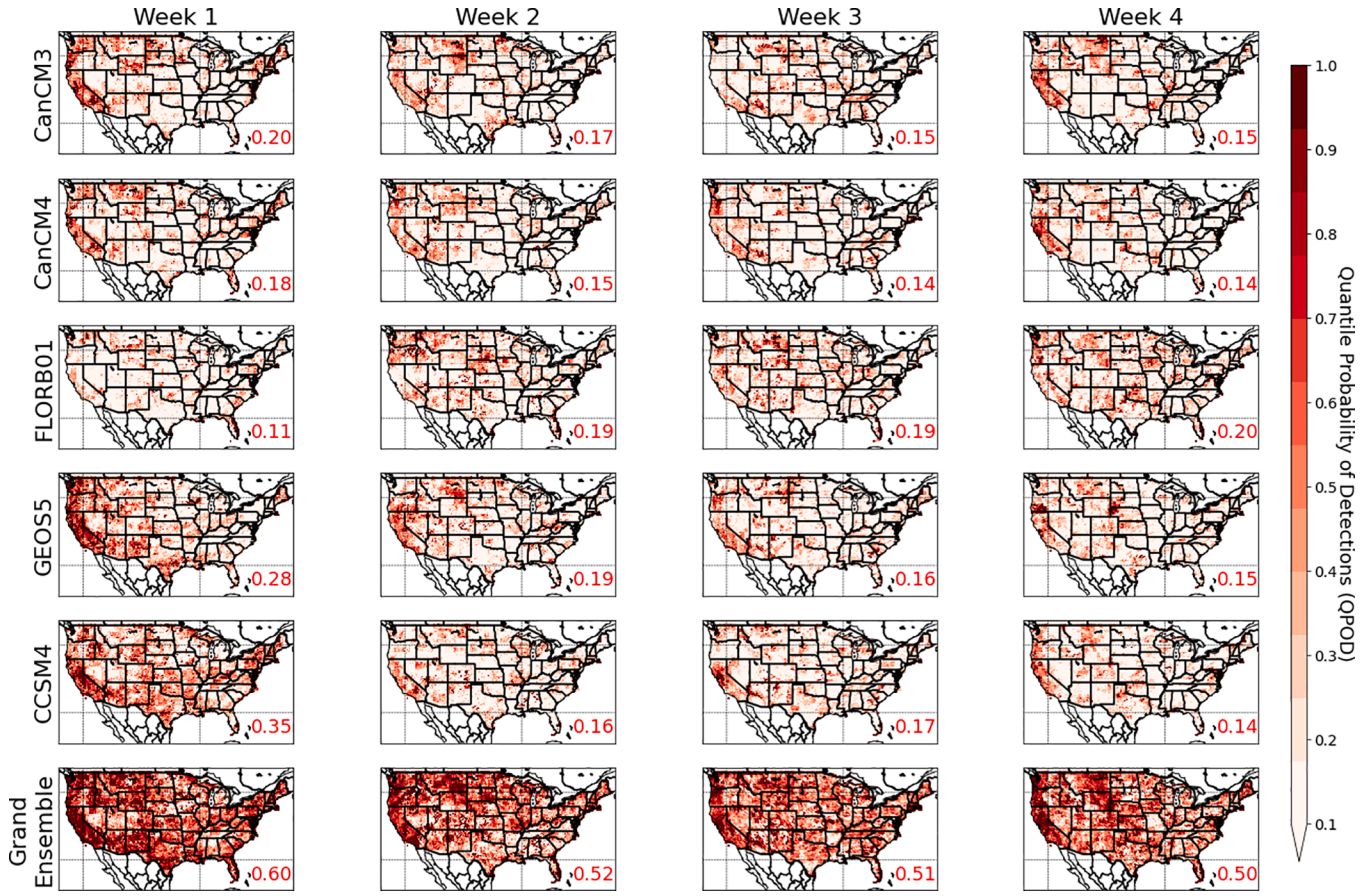

Fig. 9. The QPOD (99% threshold) for NMME-2 precipitation hindcasts over the CONUS at different lead times (weeks 1 to 4).

subplot rows, all NMME-2 models, except for the FLORB01, exhibit a decreasing trend of QPOD value over the lead times. The FLORB01 model produces the lowest QPOD value in week 1, but the QPOD value increases from 0.11 to 0.20 over lead time. Spatially, the QPOD values at the coastal areas of the Northwest and the West regions are higher than those over other regions at all lead times. This result is consistent with the ACC evaluation result presented in the previous section 3.2. The QPOD values of the grand ensemble of 5 NMME-2 models (a total of 52 ensemble members) are presented in the last row in Fig. 9. The ensemble results show that there is a significant increase in the QPOD values of the grand ensemble compared to individual NMME-2 models with smaller ensemble sizes.

Figure 10 presents the QFAR of NMME-2 precipitation hindcasts in predicting extreme precipitation events above 99% percentiles over CONUS at different lead times. In Fig. 10, the cooler colors indicate higher QFAR values, and red-colored numbers at the lower-right corners of each subplot is the spatially averaged QFAR values over the entire CONUS. Note that in contrary to QPOD, lower QFAR values indicate more superior performance in terms of predicting extremes.

According to Fig. 10, All single NMME-2 models' QFAR values are around or less than 0.1 across the CONUS. The GEOS5 and CCSM4 model, which shows slightly higher QPOD values than other NMME-2 models, also demonstrate relatively higher QFAR values. Comparing to the spatially-averaged QFAR values within different rows in Fig. 10, all NMME-2 models exhibit an increasing trend of QFAR value over the lead times. Spatially, QFAR values are lower than average over the coastal areas of the Northwest and the West regions at all lead times. This result is consistent with the ACC evaluation result presented in the previous section 3.2. The QFAR values of the grand ensemble of five NMME-2 models (a total of 52 ensemble members) are presented in the last row in Fig. 10. The result shows that with a larger ensemble size, the false alarm ratios of the grand ensemble of NMME-2 are significantly larger than any single NMME-2 models.

We also compare the QPOD and QFAR of NMME-2 models (five NMME-2 models and Grand Ensemble) with the benchmark values from resampled forecasts as presented in Table 2.

By comparing the NMME-2 QPOD and QFAR values with the benchmark values of the historical resampled forecasts, we can see all individual NMME-2 models and the grand ensemble of NMME-2 have shown superior performances in predicting extreme events exceeded 99% percentiles. All NMME-2 models have shown higher QPOD and lower QFAR than the benchmarks. Although larger ensemble sizes may bring higher QPOD values, it also increases QFAR values. Higher QPOD values indicate higher chances of coverages of extreme precipitation events above 99% percentiles by the ensemble spreads of precipitation forecasts (NMME-2 or historical resampled forecasts). Lower QFAR values indicate that when the ensemble spreads of precipitation forecasts reached above 99% percentiles, there are higher chances of such extreme events eventually happen.

4.3.2. *Extreme events below 1% threshold*

Figure 11 presents the QPOD of NMME-2 precipitation hindcasts in predicting extreme precipitation events below 1% percentiles over CONUS at different lead times. In Fig. 11, the warmer colors indicate higher QPOD values, and red-colored numbers at the lower-right corners of each subplot is the spatially averaged QPOD values over the entire CONUS.

According to Fig. 11, for all individual NMME-2 models, marginal QPOD values are observed at most regions of CONUS. Relatively higher QPOD values are only observed for certain models in certain regions and at certain lead times. For example, CanCM4 model shows higher QPOD values in the Southwest region at all lead times. GEOS-5 also presents

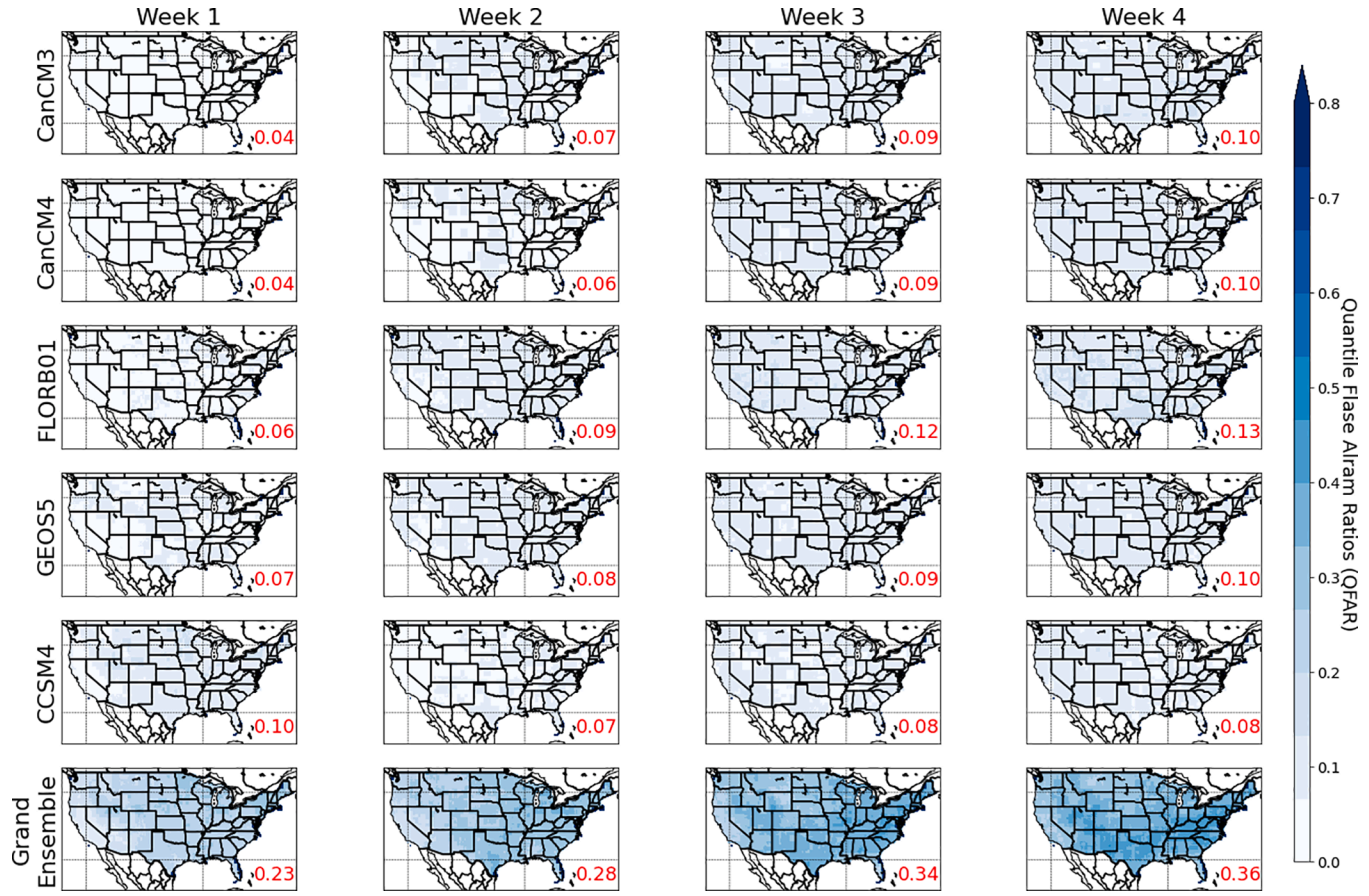

Fig. 10. The QFAR (99% threshold) for NMME-2 precipitation hindcasts over the CONUS at different lead times (weeks 1 to 4).

Table 2
NMME-2 QPOD and QFAR of extreme events above 99% percentile.

	QPOD				Benchmark	QFAR				Benchmark
	Week 1	Week 2	Week 3	Week 4		Week 1	Week 2	Week 3	Week 4	
CanCM3	0.20	0.17	0.15	0.15	0.10	0.04	0.07	0.09	0.10	0.10
CanCM4	0.18	0.15	0.14	0.14	0.10	0.04	0.06	0.09	0.10	0.10
FLORB01	0.11	0.19	0.19	0.20	0.11	0.06	0.09	0.12	0.13	0.11
CCSM4	0.28	0.19	0.16	0.15	0.10	0.07	0.08	0.09	0.10	0.10
GEOS5	0.35	0.16	0.17	0.14	0.10	0.10	0.07	0.08	0.08	0.10
Grand Ensemble	0.60	0.52	0.51	0.50	0.42	0.23	0.28	0.34	0.36	0.42

slightly higher QPOD values at some locations in the Southwest, Southeast, and Northeast regions at week 1 and week 2. The QPOD of the grand ensemble of NMME-2 presented at the last row of Fig. 11 shows significantly higher QPOD compared to single NMME-2 models over the entire CONUS at all lead times.

Figure 12 below presents the QFAR with a 1% percentile threshold of NMME-2 precipitation hindcasts over CONUS at different lead times. In Fig. 12, the cooler colors indicate higher QFAR values, and red-colored numbers at the lower-right corners of each subplot is the spatially averaged QFAR values over the entire CONUS.

The overall QFAR patterns are similar to the overall QPOD patterns shown in Fig. 11. Marginal QFAR values are observed at most of regions in CONUS for all individual NMME-2 models. Higher QFAR values are only found for certain models and at certain regions, mirroring Fig. 11. CanCM4 and GEOS5 show higher QFAR values at some locations in the Southwest region at all lead times. The QFAR value of the grand ensemble of NMME-2, which is presented at the last row of Fig. 11, is significantly higher than that of each individual NMME-2 models over the CONUS.

We also compare the QPOD and QFAR of NMME-2 models (five NMME-2 models and Grand Ensemble) in predicting extreme events below 1% percentile thresholds with the benchmark values from resampled forecasts as presented in Table 3.

Comparing the NMME-2 QPOD and QFAR values with the benchmark values of the historical resampled forecasts from Table 3, all individual NMME-2 models show higher QPOD values than the benchmark. However, the QPOD values of the grand ensemble of NMME-2 are slightly lower than but still comparable to the benchmarks. Higher QPOD values indicate higher chances of coverages of extreme precipitation events below 1% percentiles by the ensemble spreads of precipitation forecasts (NMME-2 or historical resampled forecasts) and vice versa. Lower QFAR values indicated that when the ensemble spreads of precipitation forecasts fall below 1% percentiles, that there are higher chances that such extreme events may eventually happen. Considering both QPOD and QFAR values, it is reasonable to say that the NMME-2 still presents overall better performances in predicting extreme events below 1% percentiles than the benchmark does.

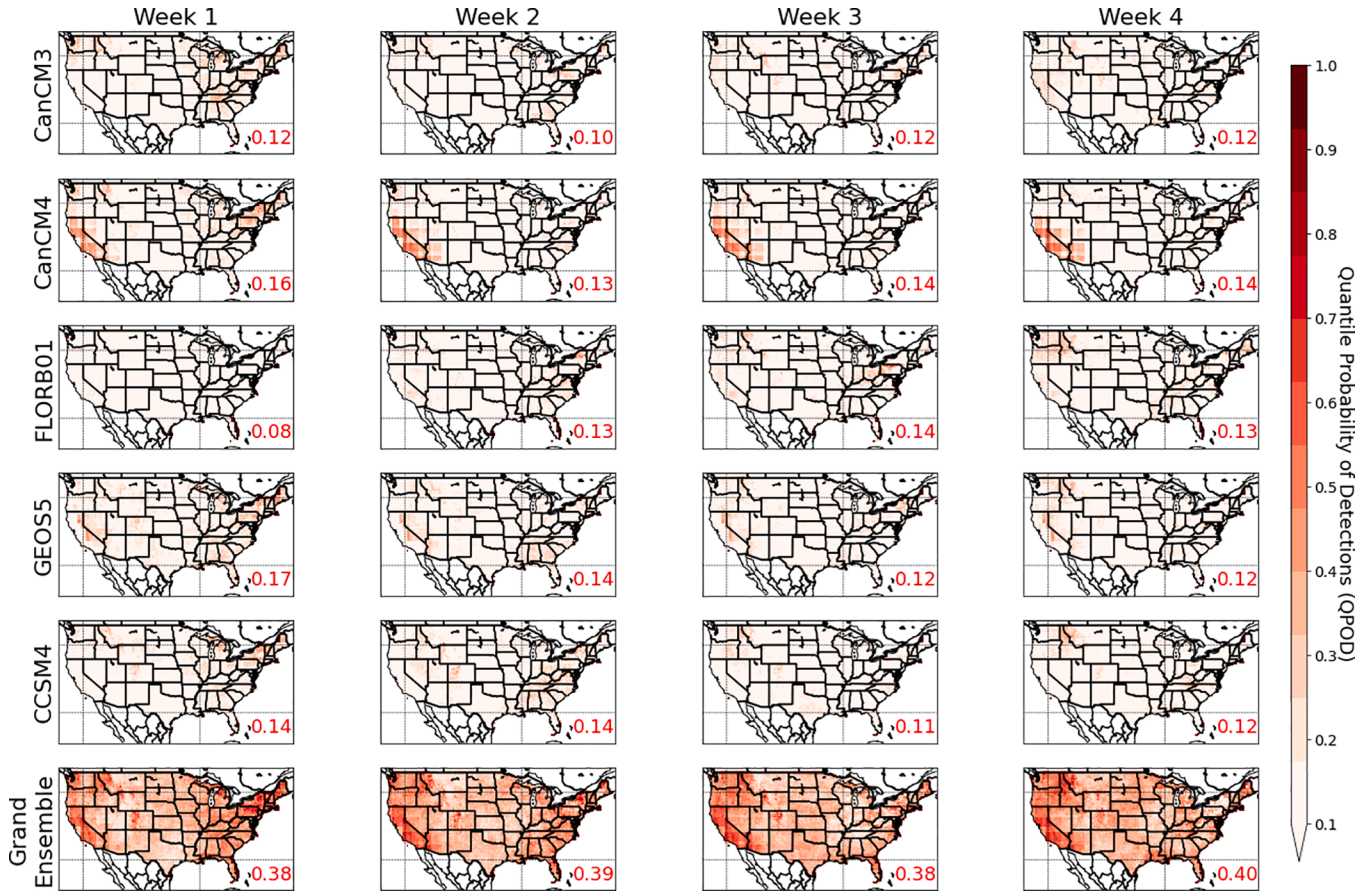

Fig. 11. The QPOD (1% threshold) for NMME-2 precipitation hindcasts over the CONUS at different lead times (weeks 1 to 4).

5. Discussion

The PBIAS result shown in prior Sections 3.1 and 3.2 can be explained along with the geographical characteristics and rainfall mechanisms of the CONUS. The overall pattern of PBIAS in the winter DJF season showed significant overestimations in the Northwest, the West, and West North Central regions. For these regions, precipitation events are dominated by synoptic-scale extratropical cyclones (ECs) and atmospheric rivers (ARs) related weathers in cooler seasons (Zhang et al., 2019). A number of previous studies have reported that GCMs tend to produce too much low-volume precipitation in comparison with reference (also known as the "drizzle effect") when simulating synoptic precipitation events (Hill 1993, Maraun 2013). Thus, the authors suspect the observed overestimations at the above-mentioned regions in cooler seasons are largely due to the "drizzle effect". However, other studies also suggest GCMs tend to underestimate high-volume precipitations related to convective weather systems in the Western US (e.g. Norris et al., 2021). In this regard, our evaluation is limited as it only reflects aggregated biases of precipitation forecasts from NMME-2, which entangles all potential affecting factors. Thus, more detailed examinations of NMME-2 at regions in the Western U.S. should be carried out to attribute the sources of bias in a more rigorous way. Although significant overestimation was generally observed in the DJF, there were also underestimations in some areas within the same regions in DJF. These mixed behaviors of over- and under- estimations could be attributed to the complex terrains brought by the Sierra Nevada Mountain and the Rocky Mountain, which are likely to trigger orographic precipitations with the nearby climatic regions. This also suggests that the raw coarse spatial resolution of NMME-2 should be considered through proper bias corrections and downscaling before any further hydrologic applications. In contrast the PBIAS patterns in DJF, an overall underestimation of precipitation from NMME-2 was observed in the South, the Southwest and part of the West North Central regions in warmer seasons, especially in JJA. These regions lie in the middle of the continent with nearly half of the precipitation contributed by mesoscale convective systems (MCSs) in warmer seasons (Easterling et al., 2017, Fritsch et al., 1986, Nesbitt et al., 2006). However, the convection systems are neither perfectly parameterized nor resolved for most of the GCMs (Moncrieff 2019). Moreover, MCSs with even smaller spatial scales normally operate within typical GCM grids, which surpass the capability of GCMs and ESMs (Eden et al., 2012). As a result, it is reasonable to suspect the observed underestimations at the South, the Southwest and part of the West North Central regions in warmer seasons are associated with sub-grid convective precipitation events.

Regarding the obtained forecast skill results, one major discrepancy was identified between FLORB01 and other NMME-2 models, as shown in Fig. 6. The FLORB01 model showed consistently lower forecast skill throughout all lead times, while other NMME-2 models showed higher forecast skill at week 1 and rapidly decreased to marginal levels at week 3 and week 4. This result agrees with the previous evaluation study upon monthly NMME-1 precipitation forecasts by Slater et al. (2019), in which the authors found the FLORB01 model sometimes does not display higher skill at the shortest lead time (one month). The obtained result in our study further shown that within the one-month lead time (i.e., week 1 to week 4), the FLORB01 model still presents poorer performances compared with other NMME models. Our obtained results (Fig. 9) also exhibit that the application of the grand ensemble with larger ensemble sizes generally shows better statistics than individual NMME-2 models with smaller ensemble sizes. This observation strongly supports the hypothesis that creating a grand model ensemble with a larger ensemble size through multiple techniques (e.g., multi-model, perturbed physics, and perturbed initial conditions etc.) will

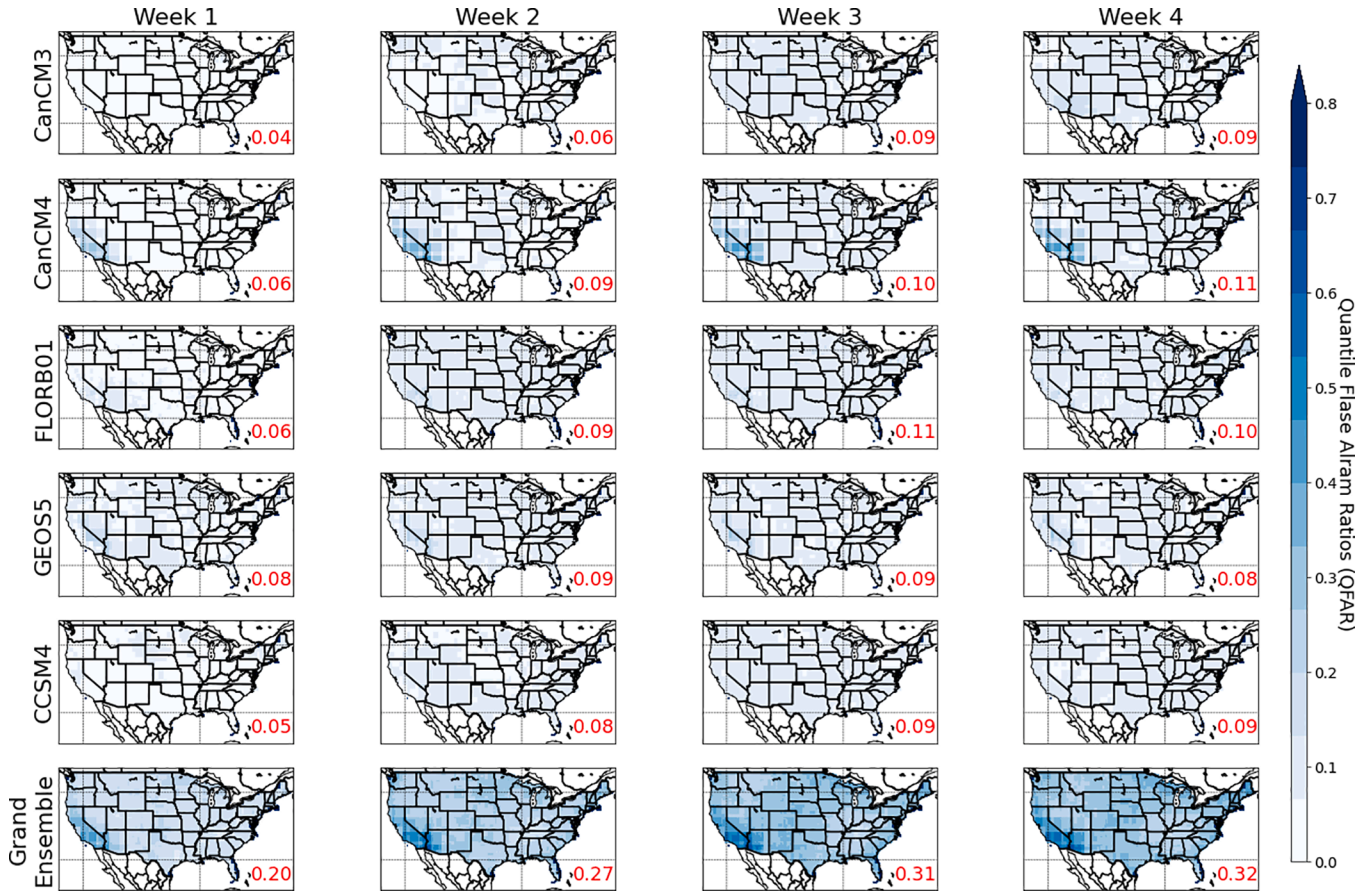

Fig. 12. The QFAR (1% threshold) for NMME-2 precipitation hindcasts over the CONUS at different lead times (weeks 1 to 4).

Table 3
NMME-2 QPOD and QFAR of extreme events below 1% percentile.

	QPOD				Benchmark	QFAR				Benchmark
	Week 1	Week 2	Week 3	Week 4		Week 1	Week 2	Week 3	Week 4	
CanCM3	0.12	0.10	0.12	0.12	0.10	0.04	0.06	0.09	0.09	0.10
CanCM4	0.16	0.13	0.14	0.14	0.10	0.06	0.09	0.10	0.11	0.10
FLORB01	0.08	0.13	0.14	0.13	0.11	0.06	0.09	0.11	0.10	0.11
GEOS5	0.17	0.14	0.12	0.12	0.10	0.08	0.09	0.09	0.08	0.10
CCSM4	0.14	0.14	0.11	0.12	0.10	0.05	0.08	0.09	0.09	0.10
Grand Ensemble	0.38	0.39	0.38	0.40	0.42	0.20	0.27	0.31	0.32	0.42

effectively increase precipitation forecast skill and reliability.

The spatial dependences of forecast skill of all employed NMME-2 model are evident. Higher ACC values were observed over the coastal areas of the Northwest, the West, and West North Central regions at all lead times. Within these regions, the winter precipitation is mainly associated with cyclonic (synoptic) scale weathers (Cayan and Roads 1984). Because of the life cycle of cyclones and their large spatial scales, the corresponding precipitation events are generally easier for GCMs to predict compared to other types of precipitation events (i.e., convective and orographic) (Kumar et al., 2011, Zhu et al., 2014). Even at longer lead times (week 2 to week 4), the ACC values at these regions are still higher than at other regions. It is also noticeable that the regions and seasons observed overestimations seem more likely to achieve higher forecast skills if cross observing Fig. 5 and Fig. 8.

Our results regarding the extreme events suggest that there is a good potential of NMME-2 forecasts to be used for hydrologic applications at the S2S range. At the S2S range, precipitation forecasts generally do not show reliable forecast skills and cannot be used deterministically. And for this reason, ensemble precipitation forecasts with the ability to cover the extreme events with their ensemble spreads are widely used. Thus, from a practical point of view, the abilities of ensemble precipitation forecasts in predicting extreme events become extremely important. According to Table 2 and Table 3, most NMME-2 models show superior QPOD and QFAR values compared to the benchmark when predicting extreme precipitation events above 99% percentiles and below 1% percentiles. NMME-2 show less dominant performances when predicting events below 1% percentiles compared to predicting events above 99% percentiles, which indicates that NMME-2 may be better at predicting floods than predicting droughts. But overall speaking, NMME-2 still appears to be a better option with generally higher probabilities of detections and lower false alarm ratios of predicting extreme events than the historical resampled forecasts. This finding suggests that NMME-2 may be a better fit to the ESP framework than historical resample forecasts in terms of hydrologic predictions at the S2S range. It is also noteworthy to mention that the grand ensemble of NMME-2 models should provide more information for water resources planning at the S2S range to mitigate the impacts from floods and droughts compared to individual NMME-2 models.

Nonetheless, many challenges remain when applying the NMME-2 S2S precipitation forecast on hydrologic modeling and real-world water resources planning. The first issue that needs to be addressed is the substantial forecast bias. Although our result has demonstrated good potentials of NMME-2 in predicting extreme events, the raw forecast values need to be bias corrected before assisting real-world flood predictions (Brown and Seo 2010, Tiwari et al., 2021). Currently, most of the popular bias correction approaches heavily rely on the "stationary assumption", where statistical moments are generalized from the historical records and will be used to correct the raw forecast values (Teutschbein and Seibert 2012). However, some recent studies have identified trends in not only frequencies but also magnitudes of extreme precipitation events in various regions over the globe (Madsen et al., 2014, Miao et al., 2015, Sun et al., 2014b), which likely undermines the efficacy of existing approaches in bias correcting extreme precipitation forecast values. And the authors believe properly considering the "non-stationary" in bias correction, especially for extreme values, might be a major challenge for future studies.

Great efforts have also been made to to improve the marginal forecast skill of the S2S precipitation forecasts by utilizing hydrometeorological forecasts from multiple sources and models to assist ensemble hydrologic forecasting. We believe that one promising direction is to construct a super NMME-2 ensemble, and further apply more sophisticated multi-model averaging techniques to improve the forecast data quality (Ji et al., 2020, Sloughter et al., 2007, Yang et al., 2018). Multi-model ensemble with proper data quality control could further increase the precipitation forecast skill across different temporal and spatial resolutions. Based on the obtained results in this study, we observe that in most of the cases, the SMA method can produce slightly higher forecast skills than any individual NMME-2 models at all lead times. Nevertheless, the improvements of SMA are still limited over individual models. Authors believe that advanced model ensemble technique needs to be used together with the fundamental enhancements of the GCMs and ESMs, which provide better precipitation predictability and less system and random errors in the S2S forecasts. For example, this could be done by advancing the current land surface components of GCMs (Dirmeyer et al., 2018, Zhou et al., 2020) or through better sub-grid convective parameterizations (Eden et al., 2012).

From the perspective of hydrology community, exploiting the benefits of deep learning techniques to post-process S2S precipitation forecasts might be another feasible alternative (Akbari Asanjan et al., 2018, Weyn et al., 2021), given the slow advances in fundamentally enhancing the physical dynamics of coupled GCMs or ESMs. Most of the previous studies have tried to correct the bias of GCM generated precipitation forecast separately, without considering the forecast skill. However, our study implies that the forecasts bias and forecast skill are somehow connected with each other. The performance of precipitation forecasts can essentially be attributed to different climate and weather patterns and mechanisms (Eden et al., 2012, Kirtman et al., 2014a). Only correcting bias may lead to degradation of forecast skills (Mendoza et al., 2017). Recently, it has been reported that taking the advantage of deep learning techniques to handle additional resolved atmospheric forecast variables to improve the S2S precipitation forecast skills first then correcting the bias may provide some major improvements (Miao et al., 2019, Pan et al., 2019a). Given the recent development of artificial intelligence and other statistical tools from computer science, the uses of machine learning methods to improve the NMME-2 S2S daily precipitation forecasts may provide a variable way leading to successful forecast adaptations for hydrologic applications. These new machine learning approaches allow auxiliary information to be considered during the bias-correction process, such as the forecast lead times, seasonality, regional factors, and relevant atmospheric forecast variables, which are all found to be important factors in improving the accuracy of S2S precipitation forecasts pertinent to the presented sensitivity analysis in our study over the CONUS.

Last but not least, it is also important for hydrologists to apply the improved and corrected S2S precipitation forecasts for hydrologic forecasts at watersheds with different hydrometeorological conditions, spatial and temporal resolutions (Cao et al., 2021, Li et al., 2019). After the necessary post-processing and bias corrections, we encourage practitioners to apply other popular metrics such as the Mean Squared Error, the Ranked Probabilistic Skill Score (RPSS), the Continuous RPSS (CRPSS), and other Categorical skill metrics, including the False Alarm Ratio (FAR), the Probability of Detection (POD), the Critical Success Index (CSI), and Equitable Threat Score (ETC) and etc. to conduct more detailed examinations of S2S precipitation forecasts on daily basis for regional studies. This would overcome the limitation of our study, in which only the NMME-2 S2S forecasts on a weekly basis are studied with a limited number of evaluation metrics. One of the major motivations for our study is to identify trends and quantify the data quality of the NMME-2 forecasts over the entire CONUS, and therefore, the streamflow simulation capabilities and forecast data improvements will be investigated in future studies.

To summary, the authors believe future studies on NMME-2 S2S precipitation forecasts may include (1) proper bias corrections and downscaling of the NMME-2 S2S precipitation forecasts at different regions across the CONUS; (2) application of deep learning approaches to provide more accurate and reliable NMME-2 precipitation forecasts, especially at longer lead times; (3) hydrologic applications of the NMME-2 S2S precipitation forecasts with ESP framework to further investigate the efficacy of NMME-2 S2S precipitation forecasts over traditional historical resample forecasts in streamflow forecasting.

6. Conclusion

In this study, the S2S precipitation forecasts of NMME-2 are comprehensively evaluated across CONUS and during a hindcast period of 1982–2011. Both deterministic evaluations of forecast bias (PBIAS) and forecast skill (ACC), and probabilistic evaluations aiming at the extreme precipitation events (QPOD, QFAR) have been conducted. The spatial, seasonal, and lead time dependence of the performances of NMME-2 daily precipitation forecasts have been analyzed over nine NCEI climate regions. The extreme precipitation performances of five NMME-2 models are also evaluated against benchmark resampled historical forecasts. Our results highlight the strengths and weaknesses of the NMME-2 S2S precipitation forecast and its potential for hydrologic applications. The major findings and conclusions of this study are summarized below.

1. The NMME-2 S2S precipitation forecasts show substantial biases across CONUS. The forecast biases also demonstrate strong spatial and seasonal dependences, but we found the biases are not sensitive to forecast lead times. Five individual NMME-2 models and the SMA of their grand ensemble show similar spatial bias patterns across the CONUS. Based on our spatial analysis, significant overestimations of the NMME-2 forecasts are observed in the western (Northwest, West, and West North Central) regions in DJF season. And significant model underestimations are identified in the South region of CONUS in warmer seasons of JJA and SON.
2. In terms of forecast skill, a major discrepancy exists between the FLORB01 and the rest of NMME-2 models. Specifically, the FLORB01 model consistently shows a lower precipitation forecast skill, especially at week-1 and week-2 lead times, as compared to other NMME-2 models. The forecast skill of the rest of NMME-2 models is the highest at week-1, tends to decrease rapidly from week-1 to week-2, and remains at a marginal level at week-3 and week-4 across all regions and in all seasons. Spatially, all NMME-2 models show higher forecast skills in the western areas (Northwest and West regions). The SMA of five NMME-2 members shows better deterministic forecast skill than any single NMME-2 model under most comparison scenarios.

3. The NMME-2 S2S precipitation forecasts also show better potentials in predicting extreme (above 99% and below 1%) precipitation events at all lead times compared to historical resampled forecasts. In addition, the formation of a grand ensemble of NMME-2 with a bigger ensemble size can further increase the performance of NMME-2 in predicting extreme events. Therefore, we believe the grand ensemble of NMME-2 S2S precipitation forecasts is a good alternative to the historical resampled forecast within the ESP framework for hydrologic applications.
4. Our study has presented more detailed evaluations of the precipitation forecasts from NMME-2 within one month (week 1 to week 4) compared to previous evaluations of NMME-1 monthly precipitation forecasts across the entire CONUS. And our evaluation results should be able to serve as an important reference for future hydrologic-related studies utilizing NMME-2 at watershed scales across CONUS.

CRediT authorship contribution statement

Lujun Zhang: Conceptualization, Methodology, Software, Writing – original draft. **Taereem Kim:** Methodology, Validation, Writing – review & editing. **Tiantian Yang:** Funding acquisition, Methodology, Software, Writing – review & editing. **Yang Hong:** Funding acquisition, Writing – review & editing. **Qian Zhu:** Funding acquisition, Writing – review & editing.

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.

Acknowledgments

The material is based upon work supported by the National Science Foundation under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3, and the National Science Foundation under Grant No. NSF1802872. This work is partially supported by the U.S. Department of Energy (DOE Prime Award # DE-IA0000018) and the Natural Science Foundation of Jiangsu Province, China (BK20180403).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2021.127058.

References

AghaKouchak, A., Behrangi, A., Sorooshian, S., Hsu, K., Amitai, E., 2011. Evaluation of satellite-retrieved extreme precipitation rates across the central United States. J. Geophys. Res.: Atmos. 116 (D2) https://doi.org/10.1029/2010JD014741.

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J. and Peng, Q. (2018) Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. Journal of Geophysical Research: Atmospheres 123(22), 12,543-512,563.

Alley, R.B., Emanuel, K.A., Zhang, F., 2019. Advances in weather prediction. Science 363 (6425), 342–344.

Ashfaq, M., Rastogi, D., Mei, R., Kao, S.-C., Gangrade, S., Naz, B.S., Touma, D., 2016. High-resolution ensemble projections of near-term regional climate over the continental United States. J. Geophys. Res.: Atmos. 121 (17), 9943–9963.

Asoka, A., Mishra, V., 2015. Prediction of vegetation anomalies to improve food security and water management in India. Geophys. Res. Lett. 42 (13), 5290–5298.

Baker, S.A., Wood, A.W., Rajagopalan, B., 2019. Developing subseasonal to seasonal climate forecast products for hydrology and water management. JAWRA J. Am. Water Resour. Assoc. 55 (4), 1024–1037.

Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nature 525 (7567), 47–55.

Becker, E., den Dool, H.v. and Zhang, Q. (2014) Predictability and forecast skill in NMME. J. Clim. 27(15), 5891-5906.

Becker, E., Kirtman, B.P. and Pegion, K. (2020) Evolution of the North American multi-model ensemble. Geophys. Res. Lett. 47(9), e2020GL087408.

Berner, J., Ha, S.-Y., Hacker, J., Fournier, A., Snyder, C., 2011. Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. Monthly Weather Rev. 139 (6), 1972–1995.

Brown, J.D., Seo, D.-J., 2010. A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts. J. Hydrometeorol. 11 (3), 642–665.

Cao, Q., Shukla, S., DeFlorio, M.J., Ralph, F.M., Lettenmaier, D.P., 2021. Evaluation of the subseasonal forecast skill of floods associated with atmospheric rivers in Coastal Western US watersheds. J. Hydrometeorol. 22 (6), 1535–1552.

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.

Cayan, D.R., Roads, J.O., 1984. Local relationships between United States West Coast precipitation and monthly mean circulation parameters. Monthly Weather Review 112 (6), 1276–1282.

Chelton, D.B., Wentz, F.J., 2005. Global microwave satellite observations of sea surface temperature for numerical weather prediction and climate research. Bull. Am. Meteorol. Soc. 86 (8), 1097–1116.

Cohen, J., Foster, J., Barlow, M., Saito, K., Jones, J., 2010. Winter 2009–2010: a case study of an extreme Arctic Oscillation event. Geophys. Res. Lett. 37 (17).

Daly, C., Bryant, K., 2013. The PRISM Climate and Weather System—An Introduction. PRISM climate group, Corvallis, OR.

Day, G.N., 1985. Extended streamflow forecasting using NWSRFS. J. Water Resour. Plann. Manage. 111 (2), 157–170.

Ding, Z., Wen, X., Tan, Q., Yang, T., Fang, G., Lei, X., Zhang, Y., Wang, H., 2021. A forecast-driven decision-making model for long-term operation of a hydro-wind-photovoltaic hybrid system. Appl. Energy 291, 116820. https://doi.org/10.1016/j.apenergy.2021.116820.

Dirmeyer, P.A., Chen, L., Wu, J., Shin, C.-S., Huang, B., Cash, B.A., Bosilovich, M.G., Mahanama, S., Koster, R.D. and Santanello, J.A. (2018) Verification of land–atmosphere coupling in forecast models, reanalyses, and land surface models using flux site observations. J. Hydrometeorol. 19(2), 375-392.

Easterling, D.R., Arnold, J., Knutson, T., Kunkel, K., LeGrande, A., Leung, L.R., Vose, R., Waliser, D. and Wehner, M. (2017) Precipitation change in the United States.

Eden, J.M., Widmann, M., Grawe, D., Rast, S., 2012. Skill, correction, and downscaling of GCM-simulated precipitation. J. Clim. 25 (11), 3970–3984.

Fritsch, J.M., Kane, R.J., Chelius, C.R., 1986. The contribution of mesoscale convective weather systems to the warm-season precipitation in the United States. J. Appl. Meteorol. Climatol. 25 (10), 1333–1345.

Gobena, A.K., Gan, T.Y., 2010. Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system. J. Hydrol. 385 (1-4), 336–352.

Guo, Y., Nie, H., 2020. Summertime daily precipitation statistics over East China in CFSv2. Phys. Chem. Earth Parts A/B/C 115, 102841. https://doi.org/10.1016/j.pce:2020.102841.

Guo, Z., Dirmeyer, P.A., DelSole, T., 2011. Land surface impacts on subseasonal and seasonal predictability. Geophys. Res. Lett. 38 (24), n/a–n/a.

Hamill, T.M., Juras, J., 2006. Measuring forecast skill: Is it real skill or is it the varying climatology? Quart. J. Royal Meteorol. Soc.: J. Atmos. Sci. Appl. Meteorol. Phys. Oceanogr. 132 (621C), 2905–2923.

Hill, C.D., 1993. Forecast problems in the western region of the National Weather Service: an overview. Weather Forecasting 8 (2), 158–165.

Ji, L., Zhi, X., Simmer, C., Zhu, S., Ji, Y., 2020. Multimodel ensemble forecasts of precipitation based on an object-based diagnostic evaluation. Monthly Weather Rev. 148 (6), 2591–2606.

Karl, T. and Koss, W.J. (1984) Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983.

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 (1-2), 457–472.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M.P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M.K., Barnston, A.G., Li, S., Rosati, A., Schubert, S.D., Rienecker, M., Suarez, M., Li, Z.E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W.J., Denis, B., Wood, E.F., 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.

Krakauer, N.Y., 2019. Temperature trends and prediction skill in NMME seasonal forecasts. Clim. Dyn. 53 (12), 7201–7213.

Kumar, A., Chen, M., Wang, W., 2011. An analysis of prediction skill of monthly mean climate variability. Clim. Dyn. 37 (5-6), 1119–1131.

Li, W., Chen, J., Li, L., Chen, H., Liu, B., Xu, C.-Y., Li, X., 2019. Evaluation and bias correction of S2S precipitation for hydrological extremes. J. Hydrometeorol. 20 (9), 1887–1906.

Lin, C., Vasić, S., Kilambi, A., Turner, B., Zawadzki, I., 2005. Precipitation forecast skill of numerical weather prediction models and radar nowcasts. Geophys. Res. Lett. 32 (14).

Madsen, H., Lawrence, D., Lang, M., Martinkova, M., Kjeldsen, T.R., 2014. Review of trend analysis and climate change projections of extreme precipitation and floods in Europe. J. Hydrol. 519, 3634–3650.

Maraun, D., 2013. Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. J. Clim. 26 (6), 2137–2143.

Mehran, A., AghaKouchak, A., 2014. Capabilities of satellite precipitation datasets to estimate heavy precipitation rates at different temporal accumulations. Hydrol. Processes 28 (4), 2262–2270.

Mendoza, P.A., Wood, A.W., Clark, E., Rothwell, E., Clark, M.P., Nijssen, B., Brekke, L.D., Arnold, J.R., 2017. An intercomparison of approaches for improving operational seasonal streamflow forecasts. Hydrol. Earth Syst. Sci. 21 (7), 3915–3935.

Miao, C., Ashouri, H., Hsu, K.-L., Sorooshian, S., Duan, Q., 2015. Evaluation of the PERSIANN-CDR daily rainfall estimates in capturing the behavior of extreme precipitation events over China. J. Hydrometeorol. 16 (3), 1387–1396.
Miao, Q., Pan, B., Wang, H., Hsu, K., Sorooshian, S., 2019. Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network. Water 11 (5), 977.
Moncrieff, M.W., 2019. Toward a dynamical foundation for organized convection parameterization in GCMs. Geophys. Res. Lett. 46 (23), 14103–14108.
Murphy, A.H., Epstein, E.S., 1989. Skill scores and correlation coefficients in model verification. Monthly Weather Rev. 117 (3), 572–582.
Nardi, K.M., Baggett, C.F., Barnes, E.A., Maloney, E.D., Harnos, D.S., Ciasto, L.M., 2020. Skillful all-season S2S prediction of US precipitation using the MJO and QBO. Weather Forecasting 35 (5), 2179–2198.
Nesbitt, S.W., Cifelli, R., Rutledge, S.A., 2006. Storm morphology and rainfall characteristics of TRMM precipitation features. Monthly Weather Rev. 134 (10), 2702–2721.
Norris, J., Hall, A., Chen, D., Thackeray, C.W. and Madakumbura, G.D. (2021) Assessing the representation of synoptic variability associated with California extreme precipitation in CMIP6 models. J. Geophys. Res.: Atmos. 126(6), e2020JD033938.
Oubeidillah, A.A., Kao, S.-C., Ashfaq, M., Naz, B.S., Tootle, G., 2014. A large-scale, high-resolution hydrological model parameter data set for climate change impact assessment for the conterminous US. Hydrol. Earth Syst. Sci. 18 (1), 67–84.
Palmer, T., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delécluse, P., Déqué, M., Diez, E., Doblas-Reyes, F.J. and Feddersen, H. (2004) Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bull. Am. Meteorol. Soc. 85(6), 853-872.
Palmer, T.N., 2002. The economic value of ensemble forecasts as a tool for risk assessment: from days to decades. Quart. J. Royal Meteorol. Soc.: A J. Atmos. Sci. Appl. Meteorol. Phys. Oceanogr. 128 (581), 747–774.
Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S., 2019a. Improving precipitation estimation using convolutional neural network. Water Resour. Res. 55 (3), 2301–2321.
Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S., Higgins, W., 2019b. Precipitation prediction skill for the West Coast United States: from short to extended range. J. Clim. 32 (1), 161–182.
Prat, O.P., Nelson, B.R., 2015. Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). Hydrol. Earth Syst. Sci. 19 (4), 2037–2056.
Radcliffe, D.E., Mukundan, R., 2017. PRISM vs. CFSR precipitation data effects on calibration and validation of SWAT models. JAWRA. J. Am. Water Resour. Assoc. 53 (1), 89–100.
Ritter, B., Geleyn, J.-F., 1992. A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. Monthly Weather Rev. 120 (2), 303–325.
Sankarasubramanian, A., Lall, U., Souza Filho, F.A., Sharma, A., 2009. Improved water allocation utilizing probabilistic climate forecasts: short-term water contracts in a risk management framework. Water Resour. Res. 45 (11) https://doi.org/10.1029/2009WR007821.
Shrestha, D., Robertson, D., Wang, Q., Pagano, T., Hapuarachchi, H., 2013. Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose. Hydrol. Earth Syst. Sci. 17 (5), 1913–1931.
Shukla, S., Roberts, J., Hoell, A., Funk, C.C., Robertson, F., Kirtman, B., 2019. Assessing North American multimodel ensemble (NMME) seasonal forecast skill to assist in the early warning of anomalous hydrometeorological events over East Africa. Clim. Dyn. 53 (12), 7411–7427.
Slater, L.J., Villarini, G., Bradley, A.A., 2017. Weighting of NMME temperature and precipitation forecasts across Europe. J. Hydrol. 552, 646–659.
Slater, L.J., Villarini, G., Bradley, A.A., 2019. 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.
Sloughter, J.M.L., Raftery, A.E., Gneiting, T., Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Monthly Weather Rev. 135 (9), 3209–3220.
Sorooshian, S., AghaKouchak, A., Arkin, P., Eylander, J., Foufoula-Georgiou, E., Harmon, R., Hendrickx, J.M., Imam, B., Kuligowski, R. and Skahill, B. (2011) Advanced concepts on remote sensing of precipitation at multiple scales. Bull. Am. Meteorol. Soc. 92(10), 1353-1357.
Stockdale, T.N., Molteni, F., Ferranti, L., 2015. Atmospheric initial conditions and the predictability of the Arctic Oscillation. Geophys. Res. Lett. 42 (4), 1173–1179.
Sun, J., Xue, M., Wilson, J.W., Zawadzki, I., Ballard, S.P., Onvlee-Hooimeyer, J., Joe, P., Barker, D.M., Li, P.-W. and Golding, B. (2014a) Use of NWP for nowcasting convective precipitation: Recent progress and challenges. Bull. Am. Meteorol. Soc. 95(3), 409-426.
Sun, Q., Kong, D., Miao, C., Duan, Q., Yang, T., Ye, A., Di, Z., Gong, W., 2014b. Variations in global temperature and precipitation for the period of 1948 to 2010. Environ. Monitor. Assess. 186 (9), 5663–5679.
Teutschbein, C., Seibert, J., 2012. Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. J. Hydrol. 456-457, 12–29.
Thomas, J.A., Berg, A.A., Merryfield, W.J., 2016. Influence of snow and soil moisture initialization on sub-seasonal predictability and forecast skill in boreal spring. Clim. Dyn. 47 (1-2), 49–65.
Tiwari, A.D., Mukhopadhyay, P., Mishra, V., 2021. Influence of bias correction of meteorological and streamflow forecast on hydrological prediction in India. J. Hydrometeorol.
Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E. and Fuentes, M. (2017) The subseasonal to seasonal (S2S) prediction project database. Bull. Am. Meteorol. Soc. 98(1), 163-173.
Vitart, F., Robertson, A.W., 2018. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. NPJ Clim. Atmos. Sci. 1 (1), 1–7.
Wanders, N., Wood, E.F., 2016. Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. Environ. Res. Lett. 11 (9), 094007. https://doi.org/10.1088/1748-9326/11/9/094007.
Weyn, J.A., Durran, D.R., Caruana, R. and Cresswell-Clay, N. (2021) Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. arXiv preprint arXiv:2102.05107.
White, C., Franks, S., McEvoy, D., 2015. Using subseasonal-to-seasonal (S2S) extreme rainfall forecasts for extended-range flood prediction in Australia. Proc. Int. Assoc. Hydrol. Sci. 370, 229–234.
White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J.T., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A.P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N.J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T.J., Street, R., Jones, L., Remenyi, T.A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., Zebiak, S.E., 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorol. Appl. 24 (3), 315–325.
Widmann, M., Bretherton, C.S., 2000. Validation of mesoscale precipitation in the NCEP reanalysis using a new gridcell dataset for the northwestern United States. J. Clim. 13 (11), 1936–1950.
Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. Academic press.
Wood, A.W., Lettenmaier, D.P., 2006. A test bed for new seasonal hydrologic forecasting approaches in the western United States. Bull. Am. Meteorol. Soc. 87 (12), 1699–1712.
Xu, L., Chen, N., Zhang, X., Chen, Z., Hu, C., Wang, C., 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.
Yang, T., Asanjan, A.A., Welles, E., Gao, X., Sorooshian, S., Liu, X., 2017. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resour. Res. 53 (4), 2786–2812.
Yang, T., Tao, Y., Li, J., Zhu, Q., Su, L.u., He, X., Zhang, X., 2018. Multi-criterion model ensemble of CMIP5 surface air temperature over China. Theor. Appl. Climatol. 132 (3-4), 1057–1072.
Yang, T., Liu, X., Wang, L., Bai, P., Li, J., 2020. Simulating hydropower discharge using multiple decision tree methods and a dynamical model merging technique. J. Water Resour. Plann. Manage. 146 (2), 04019072. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001146.
Yang, T., Zhang, L., Kim, T., Hong, Y., Zhang, D., Peng, Q., 2021. A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region. J. Hydrol. 602, 126723.
Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., Gailhard, J., 2012. Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. Adv. Sci. Res. 8 (1), 135–141.
Zhang, Z., Ralph, F.M., Zheng, M., 2019. The relationship between extratropical cyclone strength and atmospheric river intensity and position. Geophys. Res. Lett. 46 (3), 1814–1823.
Zhou, T., Chen, Z., Zou, L., Chen, X., Yu, Y., Wang, B., Bao, Q., Bao, Y., Cao, J., He, B., Hu, S., Li, L., Li, J., Lin, Y., Ma, L., Qiao, F., Rong, X., Song, Z., Tang, Y., Wu, B., Wu, T., Xin, X., Zhang, H., Zhang, M., 2020. Development of climate and earth system models in China: Past achievements and new CMIP6 results. J. Meteorol. Res. 34 (1), 1–19.
Zhou, Y., Kim, H.-M., 2018. Prediction of atmospheric rivers over the North Pacific and its connection to ENSO in the North American multi-model ensemble (NMME). Clim. Dyn. 51 (5-6), 1623–1637.
Zhu, H., Wheeler, M.C., Sobel, A.H., Hudson, D., 2014. Seamless precipitation prediction skill in the tropics and extratropics from a global model. Monthly Weather Rev. 142 (4), 1556–1569.