Contents lists available at ScienceDirect

Atmospheric Research

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

Improving Subseasonal-to-Seasonal forecasts in predicting the occurrence of extreme precipitation events over the contiguous U.S. using machine learning models

Lujun Zhang [a], Tiantian Yang [a,*], Shang Gao [a], Yang Hong [a], Qin Zhang [b], Xin Wen [c], Chuntian Cheng [d,e]

[a] School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK, United States
[b] NOAA Climate Prediction Center, College Park, MD 20740, USA
[c] College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China
[d] Institute of Hydropower and Hydroinformatics, Dalian University of Technology, Dalian 116024, China
[e] Key Laboratory of Ocean Energy Utilization and Energy Conservation of Ministry of Education, Dalian University of Technology, Dalian 116024, China

ARTICLE INFO

Keywords:
S2S
Extreme
Precipitation
Ensemble forecast
Machine learning
CONUS

ABSTRACT

The precipitation forecasts at the Subseasonal-to-Seasonal (S2S) scale are valuable information to assist water resources planning and decision-making at an extended range. But the raw S2S precipitation forecasts are rather limited regarding their predictive skills, which hinders further hydrological applications. Many previous studies were carried out to validate the S2S precipitation forecasts from the perspective of the ensemble "mean", while existing study, which evaluates and improves the ability of S2S forecasts in predicting extreme precipitation events with their ensemble spreads, is rarely reported. This study aims to improve the S2S forecasts in predicting the occurrence of weekly extreme precipitation events above 99% over the contiguous United States (CONUS). The Random Forest Classifiers (RF) were employed and additional forecasts variables (i.e., surface air temperature, geopotential heights at 500 hPa and 850 hPa) were included to post-process the raw S2S forecasts from the NASA's Goddard Earth Observation System model version five (GEOS5). Different RF training inputs and RF hyperparameter sensitivity analysis are examined. We found that ***(1)*** using S2S precipitation forecast as the only inputs to RF, the forecast quality is improved significantly only at week 1; (2) When additional forecasts variables are included in RF training, the forecast skill tend to improve slightly at longer lead times after weeks 2; (3) The tunning of the maximum tree depth of RFs combine with the inclusion of additional forecasts variables as inputs to RF can improve the forecasts skill at all lead times over CONUS. In short, this study demonstrated the effectiveness of the application of RF as well as the effectiveness of additional forecast variables in improving the S2S extreme precipitation forecasts, which could be potentially useful for flood and river ensemble forecasting. Multiple statistical metrics, including the ensemble probability of detections (EPOD), ensemble false alarm ratios (EFAR), ensemble critical success index (ECSI), Brier Sill Score (BSS), and Area Under the Receiver Operating Characteristics Curves (AUROC) are employed for a comprehensive evaluation of the predictive performances of S2S extreme precipitation forecasts under different experiment scenarios.

1. Introduction

Riverine flooding typically comes after heavy precipitation, causing impacts on human socio-economic activities, including loss of human lives (Begum et al., 2007; Singh and Kumar, 2013), water contamination (Taylor et al., 2011; Yang et al., 2022), agricultural damage (Bremond et al., 2013), and disruption of transportation systems (Suarez et al., 2005). Reliable precipitation forecasts are critical for decision-makers to adaptively change their strategies to mitigate the impacts of such natural disasters (Sorooshian et al., 2011).

Advanced weather forecasting and climate models can generate precipitation forecasts at different timescales for water resources planning. At lead times within a medium range (i.e., 2 weeks into the future), precipitation forecasts generated from Numerical Weather Prediction

[*] Corresponding author.
E-mail address: tiantian.yang@ou.edu (T. Yang).

https://doi.org/10.1016/j.atmosres.2022.106502
Received 12 June 2022; Received in revised form 25 September 2022; Accepted 31 October 2022
Available online 4 November 2022
0169-8095/© 2022 Elsevier B.V. All rights reserved.

(NWP) models are the most widely adopted products (Kuligowski and Barros, 1998; Robertson et al., 2013), which simulate the propagation and evolution of atmospheres. At longer lead times, the General Circulation Models (GCMs) coupled with dynamic components of the atmosphere, ocean, and land surface are more suitable tools as compared to the NWP models in producing seasonal, annual, or even decadal precipitation outlooks (Vitart, 2017; Xiang et al., 2019). Both NWP and GCM-generated precipitation forecasts are widely studied and applied by various mission agencies and research communities (Clark et al., 2017; Wood and Lettenmaier, 2006; Yang et al., 2018).

However, a forecast gap lies in the Subseasonal-to-Seasonal (S2S) timescale, defined explicitly as the transitional period of 10 to 30 days between weather predictions and seasonal outlooks (White et al., 2017). Previous studies found that the predictability of weather is either lost already (i.e., initial conditions of the atmosphere) or has yet to start dominating local weather (i.e., ocean-weather interactions) at S2S timescales (Vitart et al., 2017). Such a lack of predictability sources makes it extremely challenging to provide accurate S2S forecasts (White et al., 2017).

Many efforts have been made to advance precipitation forecasts by identifying additional predictability sources of the weather at the S2S timescale. Series of initial conditions of land surface and atmosphere are found to be associated with the variation of sub-seasonal weather (Asoka and Mishra, 2015; Chelton and Wentz, 2005; Cohen et al., 2010; Guo et al., 2011; Stockdale et al., 2015; Thomas et al., 2016). More recently, multiple atmospheric variation modes over oceans are also found to be significantly dominating the distribution and magnitude of subseasonal precipitation events (Dai and Wigley, 2000; Hsiao et al., 2020; Mariotti et al., 2020; Yang et al., 2017a; Zhang and Ling, 2017).

Additional efforts have been made by mission agencies all over the globe to provide experiment S2S forecast datasets through coupled GCM models (Kirtman et al., 2014; Pegion et al., 2019; Vitart, 2014; Vitart et al., 2017; Yang et al., 2018). Available S2S forecast datasets include European Center for Medium-Range Forecasts (ECMWF), the S2S project by the World Weather Research Program, and the North America Multi-Model Ensemble Phase II (NMME-2), etc. Although existing S2S forecast products offer a promising opportunity for seamless hydrologic predictions, it is commonly agreed that S2S precipitation forecasts suffer from a substantial amount of forecasts biases and a marginal level of predictive skills (Baker et al., 2019; de Andrade et al., 2019; de Andrade et al., 2021; King et al., 2020; Tian et al., 2017; Vigaud et al., 2017; Wang and Robertson, 2019; Tao et al., 2018; Yang et al., 2015).

More importantly, existing studies were mainly focused on investigating the accuracy of ensemble means or medians of S2S precipitation forecasts, while the ability of ensemble S2S forecasts to predict extreme events with their ensemble spreads is rarely reported. Nevertheless, from a practical perspective, extreme information is critically needed as precipitation forecasts generally cannot be applied deterministically at S2S ranges in hydrology (Day, 1985). From the standpoint of water infrastructure operation (i.e., reservoirs and long-distance water transferring systems), decision-makers would desire the model-generated ensemble forecasts to envelop and include future extreme precipitation events as much as possible so that the potential risks of associated natural disasters could be considered beforehand and avoid infrastructure failure (VanBuskirk et al., 2021; Wu et al., 2020; Yuan et al., 2015; Yang et al., 2020).

As we mentioned earlier, there is only a limited number of studies focused on S2S extreme precipitation forecasts over CONUS. Among some recent studies, Cao et al. (2021) examined the hydrologic performances of the ensemble means of S2S forecasts in forecasting flood events at three watersheds in the western U.S. In addition, Zhang et al. (2021) studied and examined the capability of S2S precipitation forecasts from NMME-2 in capturing the pattern of extreme rainfall with the entire spreads of the ensembles. However, according to Zhang et al. (2021), the skills of the S2S precipitation forecasts in predicting the occurrence of weekly extreme precipitation events are marginal after two weeks into the future.

The poor quality of S2S precipitation forecasts in predicting extreme events hinders a broader application of S2S forecasts in flood predictions and sustainable water infrastructure operations. Various statistical-based post-processing techniques are available, which can help correct the forecast bias and improve S2S precipitation forecast accuracy. When using these statistical-based post-processing tools, though the forecast biases can often be removed nicely, many studies found that the forecast skill tends to remain either unchanged or even deteriorate after removing the biases (Baker et al., 2020; Li et al., 2017; Manzanas et al., 2018; Zhao et al., 2017). This is because S2S precipitation forecast skill is affected by multiple factors, including coarse spatial resolutions, imperfect/unrealistic precipitation parameterization schemes, and the computational errors originated in resolving the partial differential equations of the atmosphere when generating rainfall forecast in the physical models (Davis and Goadrich, 2006; Ebert and McBride, 2000). Due to these reasons, existing statistical-based post-processing approaches are somewhat limited in improving the predictive skills of S2S precipitation forecasts. Therefore, more advanced tools are still needed, especially for extreme precipitation events at the S2S timescale.

As an alternative to the existing statistical-based post-processing approaches, Machine Learning and Data Mining (ML&DM) techniques are promising tools for improving the S2S extreme precipitation forecasts skill scores. The ML&DM techniques have the flexibility to include an arbitrary set of input variables in the classification and regression process (Yang et al., 2017a and 2017b). They can effectively identify the complex relationships between selected input variables and target variables, which may not be directly related to each other. Many researchers have successfully applied ML&DM for precipitation forecast adaptations. For example, Miao et al. (2019), Pan et al. (2021), Pan et al. (2019), Wang et al. (2021) applied ML&DM to post-process precipitation forecasts over different study regions and reported overall improvements in forecast skills. However, some other studies reported that ML&DM tends to underestimate and limit the reproduction of extreme values (Akbari Asanjan et al., 2018; Baño-Medina et al., 2020; Kim et al., 2022; Sadeghi et al., 2020). Nevertheless, given the sensitive dynamics of extreme precipitation events (Faridzad et al., 2018; Nie et al., 2020; Pendergrass, 2020; Srinivas et al., 2018), we expect that the triggering of extreme events should be easier to identify when it is compared to regular precipitation events. Therefore, a popular ML&DM model, i.e., the Random Forest or RF classifier, is applied in this study to predict the occurrence of extreme precipitation events over different regions of CONUS. The RF is an ensemble-based tree algorithm. It can handle correlated conditional variables and is robust against overfitting with the presence of high-level noise in the training data (Breiman, 2001, Strobl and Zeileis, 2008). As one of the most popular ML algorithms, the RF has been widely applied in a variety of hydrometeorological studies, including statistical downscaling (He et al., 2016, Tao et al., 2018), reservoir release predictions (Yang et al. 2015, 2020, and 2021), and the post-processing of precipitation forecasts (Herman and Schumacher, 2018a,b, Loken et al., 2019).

On top of using advanced ML techniques, the inclusion of additional forecast variables into the ML training may further increase the predictive skill of precipitation forecasts. Precipitation forecasts from dynamic models are generated through a process called "parameterization." Instead of computing numerical values by resolving the partial differential equations in forecast models, this "parameterization" process generates quantitative precipitation forecasts empirically based on the importance of other explicit atmospheric variables (i.e., variables computed directly by resolving physical equations, such as temperature, pressure, etc.). The reason for adopting such "parameterization" schemes in dynamic models is because precipitation is formed through complex micro-physical and chemical processes that surpass the resolution and capability of current models (Stensrud, 2009). Such "parameterization" schemes are generally considered of high uncertainty and less reliable compared to the computations of explicit atmospheric variables (Bader and Roach, 1977; Best et al., 2011; Betts et al., 1998; Bukovsky and Karoly, 2007; Tian et al., 2017; Vitart, 2004; Yang et al., 2017a, 2017b). Thus, including additional forecast variables together with a robust ML algorithm may lead to a higher chance of identifying the dynamic, varying patterns of precipitation events over space and time.

Therefore, in this study, multiple S2S forecast variables are included in the training of RF to predict the occurrence of weekly extreme precipitation events over CONUS. The input forecast variables include each individual and a combination of (1) precipitation, (2) surface air temperature, (3) 500 hPa geopotential heights, and (4) 850 hPa geopotential heights. The surface air temperature, 500 hPa, and 850 hPa geopotential heights are related to the formation of precipitation events. These meteorological variables have been commonly used in previous studies to downscale and post-process precipitation forecasts (Li et al., 2022; Miao et al., 2019; Pan et al., 2019). In this study, we chose 99% as the threshold to identify weekly extreme precipitation events without a universal standard.

Previously, Zhang et al. (2021) reported that the capability of NMME-2 S2S forecast in capturing the occurrence of extreme precipitation events is marginal after week 2. Following Zhang et al. (2021)'s work, this study develops a prototype model to improve the capability of S2S forecasts in predicting the occurrence of extreme precipitation events. The experiment S2S forecast dataset used in this study is from one contributing forecast model of NMME-2, i.e., the NASA Goddard Earth Observing System version 5 model (GEOS5). Building off of many existing studies that focused on the ensemble means, we further consider the entire ensemble of S2S forecasts. A few existing statistical metrics were modified for the evaluation of ensemble forecasts. The contribution of this study also includes sensitivity tests on one RF hyperparameter, and different model input scenarios are also examined for comparison.

In summary of our research goals, we intend to address research questions: 1) Can the predictions on the occurrence of extreme precipitation events be improved at the S2S ranges through ML techniques over CONUS? 2) Can additional atmospheric forecast variables improve S2S extreme precipitation forecasts over CONUS? 3) Does tunning of ML hyperparameter significantly affect the quality of extreme precipitation forecasts over CONUS? And 4) Do the ML-enabled forecasts perform consistently in different regions over CONUS?

The rest of this paper is organized as follows: In section 2, we present data and study region. The experiment settings and evaluation metrics are described in Section 3. Section 4 presents the results. The discussions and main conclusions are presented and summarized in Section 5 and Section 6, respectively.

2. Data and study region

The datasets used in this study include (1) the daily S2S forecast variables from GEOS5 and (2) a ground truth daily precipitation observation dataset for validation, the AN81d daily precipitation dataset from the Parameter-elevation Regressions on Independent Slopes Model (PRISM). Both forecast and reference datasets are collected from 01/01/1982 to 12/31/2011 to cover a 30-year study period.

The GEOS5 produces 10-member ensemble forecasts on the first day of each month during the study period. Each member of the GEOS5 dataset provides daily forecasts of multiple hydrometeorological variables with a lead time of up to 274 days. The 10-member ensemble forecasts are generated with perturbed initial conditions (Borovikov et al., 2019). Raw GEOS5 forecasts have spatial resolutions of 1° (∼100 km). Four GEOS5 forecast variables are used in this study, including precipitation (P), surface air temperature (T), and geopotential height at 500 hPa (G500) and 850 hPa (G850).

The PRISM is a gridded dataset that covers the entire CONUS with a spatial resolution of ∼0.04° (4 km). The PRISM combines a digital elevation model with both surface and Radar precipitation measurements (Daly and Bryant, 2013). It is a reliable reference for validating satellite precipitation estimation and precipitation forecast products (Gowan et al., 2018; Mizukami and Smith, 2012; Prat and Nelson, 2015).

The collected S2S and PRISM datasets are pre-processed as follows. All collected S2S forecasts are truncated to 28 days (4 weeks) to focus on the S2S timescale specifically. The truncated forecasts are labeled with individual lead times (1–28 days). Both S2S forecasts and PRISM were then resampled to 0.25-degree pixels to match with each other. Specifically, the S2S forecasts were downscaled using the nearest neighbor approach. The PRISM was upscaled through aggregations (i.e., areal averages of all origin PRISM pixels fall in the corresponding 0.25-degree pixels) in order to match the coarse spatial resolution of GEOS5 data.

To validate the experiment results, we conducted both pixel-based and regional evaluations of the raw and improved forecasts to better observe spatial patterns over CONUS. The regional assessment is performed based on the nine climate regions defined by the National Centers for Environmental Information (NCEI) (Karl and Koss, 1984) shown in Fig. 1. A few NCEI regions include the mountainous terrains. Specifically, the Rocky Mountains span the Northwest, West, West North Central, and Southeast regions; the Appalachian Mountainous regions are included in parts of the Northeast and Southeast regions.

3. Experiment Settings and Evaluation Metrics

3.1. Experiment Settings

In this study, the RF classifiers are individually trained at each 0.25-degree pixel over CONUS to identify weekly extreme precipitation events beyond 99% using the 10-member ensemble S2S forecasts. The employment of RF is based on the open-source package in Python. The "leave-10-years-out" model across-validation strategy was adopted to avoid overfitting of the ML model over the entire study period. This strategy was commonly applied in the fields of hydrometeorology and climate change for forecast correction and statistical downscaling (Li et al., 2019; Manzanas et al., 2018). To perform the "leave-10-years-out" cross-validation, the whole study period was divided into three 10-year periods (i.e., 1982–1991, 1992–2001, and 2002–2011). When a particular 10 years are selected as the targeting period for forecast corrections, the remaining 20 years' data will be used to train the RF model.

The training of RF was carried out at different lead times as well. Taking week 1 forecasts as an example, the daily S2S forecasts with lead times smaller than 1 week (lead times 1 to 7 days) from the reference period are used as the inputs to train RF. The model training target is set to be the weekly extreme events observed in the PRISM dataset. The trained RF model is further used to produce categorial predictions upon extreme events at the same lead time for the target period. We repeated this training process at each 0.25-degree pixel over CONUS for week 2 (day 8 to 14), week 3 (day 15 to 21), and week 4 (day 22 to 28).

We expect that the S2S extreme precipitation forecasts can be improved via 1) the inclusion of additional atmospheric forecast variables as inputs to RF, and 2) the tunning of the max tree depth of RF. To test our hypothesis, a total of ten experiment scenarios were designed and shown in Table 1.

Among all scenarios presented in Table 1, we first benchmark the performances of the raw S2S precipitation forecasts in predicting the occurrence of extreme events (E1). The weekly extreme events over CONUS are identified by aggregating PRISM into weekly-averaged values and sorting out weeks with averaged values above the 99% threshold. Similarly, for the raw precipitation forecasts for all ensemble members of GEOS5, we first aggregate them into weekly values. Then, the positive predictions from GEOS5 upon the occurrence of extreme precipitation events are identified by sorting out weeks with averaged forecast values above 99% according to their own statistics.

For experiment scenarios 2–5 (E2-E5) (Table 1), the hyperparameters of RF are controlled and remain unchanged, whereas the

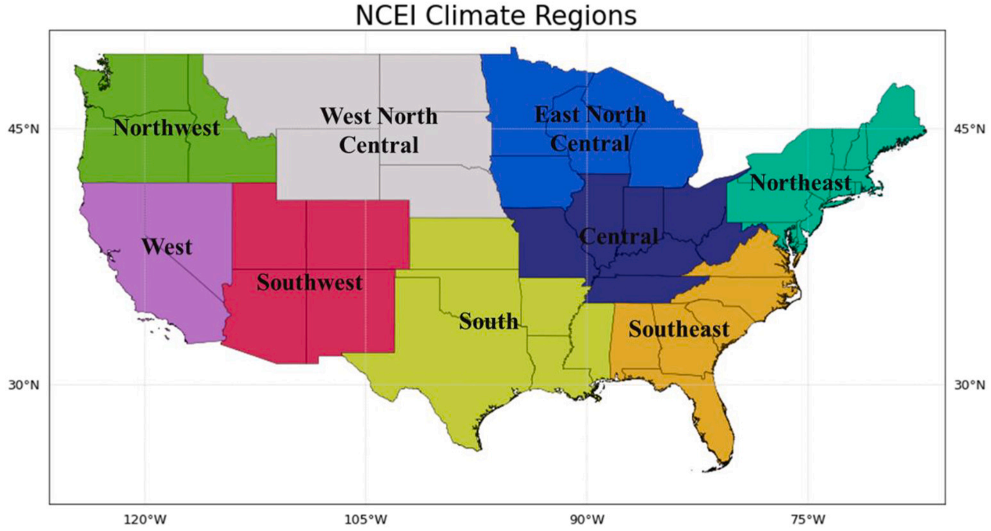

Fig. 1. nine NCEI climate regions.

Table 1
Designed experiment scenarios. "P" refers to Precipitation Forecasts; "T" refers to Temperature Forecasts; and "G" refers to Geopotential Heights.

Experiment Scenarios	E1 (Raw)	E2	E3	E4	E5	E6	E7	E8	E9	E10
Input variable(s)	N/A	P	P T	P G500	P G850	P T G500 G850	P T G500 G850	P T G500 G850	P T G500 G850	P T G500 G850
Maximum RF Tree depth	N/A	3	3	3	3	3	6	9	12	15

"max_depth" parameter is set to be 3, "max_features" is set to be 0.6, "n_estimators" is set to 150, and other hyperparameters remain as default for the RF model. The scenarios E2-E5 only differ in the variables used as inputs to RF, where the same P, P and T, P and G500, or P and G850 are used as the input variables under different scenarios E2-E5, respectively.

For experiment scenarios 6–10 (E6-E10), the input variables used to train RF are controlled and remain unchanged, whereas all collected S2S forecast variables are used as inputs to RF. However, under E6-E10, the "max_depth" of RF is set to be 3, 5, 7, 9, 12, and 15, respectively, while other hyperparameters remain unchanged (i.e., max_features = 0.6, n_estimators = 150).

3.2. Pixel-based and Event-based Forecast Evaluation Metrics

This study used three sets of evaluation metrics to compare the corrected forecast results. The first set of metrics is pixel-based evaluation metrics, including three popular categorical evaluation metrics of Probability of Detection (POD), False Alarm Ratio (FAR), and Critical Success Index (CSI). In the second set of metrics, an event-based probabilistic evaluation metric termed the Brier Skill Score (BSS), is employed. The last set of metrics is called the Area Under the Receiver Operating Characteristics Curves (AUROC), which is a graphical index used for spatial skill score evaluation for different climate regions over the CONUS.

To apply the POD, FAR, and CSI, we further modified these metrics to consider the entire ensemble of S2S forecasts and named the modified metrics as Ensemble Probability of Detection (EPOD), Ensemble False Alarm Ratio (EFAR), and Ensemble Critical Success Index (ECSI); The BSS and the AUROC are two widely applied metrics to quantify the model performances in the field of Hydrometeorology (Wilks, 2011) and Machine Learning (Davis and Goadrich, 2006), respectively. In this study, we choose BSS and AUROC to verify the performances of RF in addition to EPOD, EFAR, and ECSI. The following sections introduce the logic and detailed calculation of the employed metrics.

3.2.1. Mathematical definitions of EPOD, EFAR, and ECSI

Table 2 describes a so-called contingency table, which illustrates the categorical relationship between the forecast and reference datasets. In this contingency table, the occurrence of extreme precipitation events is represented in the form of binary events with the possibility of two scenarios, either "*True*" or "*False*". The "H" in the contingency table means "hits," and it refers to the number of successful predictions upon the occurrence of extreme events. Similarly, "M" means "misses," which refers to the number of non-forecasted extreme events. "F" means "false alarms," and it refers to the number of false positive forecasts upon the occurrence of extreme events. Based on the contingency table, three evaluation metrics of POD, FAR, and CSI are computed as follows: POD measures the fraction of accurate detection of events, and its calculation equation is H/(H + M); FAR measures the fraction of erroneous detection of events and its equation becomes F/(F + H), and CSI determines the fraction of correct event detection after ignoring the correct negative events, and the equation of CSI is H/(H + F + M).

However, precipitation forecasts generally cannot be applied deterministically on sub-seasonal timescales due to their uncertainties, and it

Table 2
Contingency table of all possible outcomes for categorical forecasts of binary events.

Events	O	NO
P	H	F
NP	M	CR

O = observed; NO = not observed; P = predicted; NP = not predicted; H = hit; F = false alarm; M = miss; CR = correct rejection.

is not informative to use the original POD, FAR, and CSI metrics to evaluate S2S precipitation forecasts as they do not consider the entire ensemble spreads from different model outputs. Therefore, the use of EPOD, EFAR, and ECSI as the evaluation metrics will be more realistic and inclusive. The definitions of EPOD, EFAR, and ECSI generally follow the definitions of traditional POD, FAR, and CSI but consider all ensemble members of forecasts. By adopting modified EPOD, EFAR, and ECSI, we would like to quantify if extreme precipitation events are well-enveloped by the spreads of S2S forecasts.

To compute EPOD, EFAR, and ECSI, we define that successful categorical prediction made by any ensemble members upon extreme events will be counted as a “hit.” Similarly, an unsuccessful categorical prediction made by all ensemble members will be counted as a “miss.” Following the same logic, if none of the ten ensemble members has forecasted an extreme event above 99% while such event did not happen, it will be counted as a “correct negative”. Finally, if all ensemble members have forecasted an extreme event above 99% while such an event did not happen, it will be counted as a “false alarm”. With such modified definitions of “hit,” “miss,” and “false alarm,” the EPOD, EFAR, and ECSI can be computed as follows:

$$EPOD = \frac{\sum_{i=1}^{n} I\left[\bigwedge_{j=1}^{m}\left(x_{ij}\ AND\ y_i\right)\right]}{\sum_{i=1}^{n} I(y_i)} \tag{1}$$

$$EFAR = \frac{\sum_{i=1}^{n} I\left\{\bigvee_{j=1}^{m}\left[x_{ij}\ AND\ (NOT\ y_i)\right]\right\}}{\sum_{i=1}^{n} I\left[\bigwedge_{j=1}^{m}\left(x_{ij}\right)\right]} \tag{2}$$

$$ECSI = \frac{\sum_{i=1}^{n} I\left[\bigvee_{j=1}^{m}\left(x_{ij}\ AND\ y_i\right)\right]}{\sum_{i=1}^{n} I(y_i) + \sum_{i=1}^{n} I\left\{\bigvee_{j=1}^{m}\left[x_{ij}\ AND\ (NOT\ y_i)\right]\right\}} \tag{3}$$

In the equations above, n is the total number of categorical forecasts with the same lead time; and m quantifies the full ensemble members of the GEOS5 (i.e., 10); x_{ij} is the categorical prediction (i.e., either "*True*" or "*False*") made by the jth ensemble member of GEOS5 for a particular week. Similarly, y_i is the reference categorical indicator of whether an extreme event above 9the 9% threshold has happened or not at a particular week. The symbol ^ and ∨ represent large logical operations of *OR* and *AND* respectively. I is the indicator function where $I(True) = 1$ and $I(False) = 0$.

3.2.2. Event-based Brier Skill Scores (BSS)

The Brier Skill Scores (BSS) is a probabilistic evaluation metric that has been widely applied in the field of meteorology and atmospheric science. The BSS describes the quality of categorical probabilistic forecasts (Wilks, 2011) and it quantifies the extent to which a forecast strategy improves predictions with respect to a reference forecast. The BSS is defined by the following Eq. (4):

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{4}$$

where BS and BS_{ref} are the Brier Scores (Brier, 1950) of GEOS5 (i.e., E1-E10) and reference forecast for a sample of n binary events, respectively. In this study, we use climatology as the reference forecast. The BS and BS_{ref} are computed with the following Eq. (5):

$$BS = \frac{1}{n}\sum_{t=1}^{n}\left(f_t - O_t\right)^2 \tag{5}$$

where f_t is the predicted probability of the event at time t, and O_t is equal to 1 or 0, depending on whether the extreme event subsequently occurred or not. However, it has been reported that the BSS could falsely inflate the skill when adopted for evaluations upon extreme events (Wilks, 2011). This is because extreme events are natrually rare in data records and too many negative predictions upon the occurrence of extreme events could lead to favourable statistics. We therefore assigned different weights to categorical events to derive a more realistic skill socre of the forecasts. The BS and BS_{ref} are computed following Eq. (6):

$$BS = \frac{1}{n}\left(\sum_{t=1}^{n} W_{O_t}{*}\left(f_t - O_t\right)^2\right) \tag{6}$$

In Eq. (6), $W_{O_t} = 0.99$ when O_t is equal to 1 and $W_{O_t} = 0.01$ when O_t is equal to 0, based on their probabilities in climatology. In this study, when computing the BS of GEOS5 forecast following Eq. (6), f_t is computed based on number of the positive categorical predictions among all 10 ensemble forecast members at time t. For example, when six out of ten GEOS5 members gave positive predictions upon extreme events, f_t would equal to 0.6. On the other hand, the f_t of reference forecast would constantly be equal to 0.01, given the climatology probability threshold of extreme events is 99%.

3.2.3. Area under the receiver operating characteristics curves (AUROC)

The last set of evaluation metric, i.e., AUROC, is defined as the Area Under a Receiver Operating Characteristics (ROC) curve. Note that due to the huge amount of computation at all 0.25-degree pixels across the entire CONUS, we cannot plot and present ROC curves at each pixel and with different probability thresholds of RF. Instead, we pooled the categorical predictions obtained from all ensemble forecasts under different experiment scenarios at different NCEI climate regions to calculate the area based AUROC. In other words, the ROC curves and their corresponding AUROC were plotted and computed at nine NCEI climate regions with the default probability threshold of RF of 50%.

The POD and POFD are needed to plot ROC curves and to compute AUROC. According to the contingency Table 2, the calculation of POD follows H/(H + M) and calculation of POFD follows F/(F + CR). For example, assuming a classifier has made series of predictions upon a certain “yes/no” type of event and resulted in POD of 0.8 and POFD of 0.3, then, such a classifier's corresponding ROC curve can be plotted as

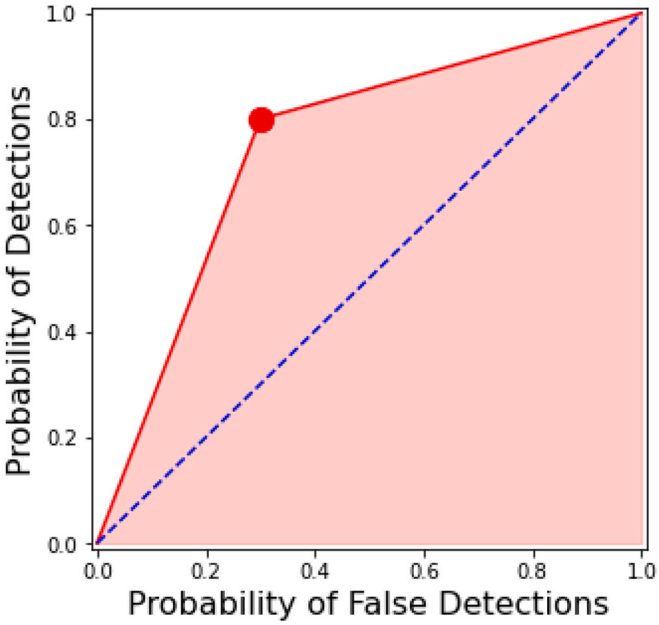

Fig. 2. An Illustrating figure of ROC and AUROC. Red line in the figure is the ROC curve obtained by a conceptual RF classifier with POD = 0.8 and POFD = 0.3. The red shaded area is the ROC curve's corresponding AUROC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2. In Fig. 2, the red line is the ROC curve. The blue dotted diagonal line indicates the skill of a classifier constantly making random guesses. The red-shaded area below the ROC curve is defined as the AUROC. The ideal value of AUROC would be equal to one (i.e., POD = 1 and POFD = 0). In this study, we only present the NCEI regional AUROC for conciseness. The individual ROC curves at each NCEI climate region are presented in the supplementary materials.

4. Results

4.1. Model's sensitivity on inputs (Comparison of Scenarios E1-E6)

In this section 4.1, we first present model's sensitivity analysis on different combination of inputs based on the results obtained from scenarios E1-E6. Then, in section 4.2, we further analyze how model's correction performance changes over different ML hyperparameters based on the results from scenarios E1, and E6-E10. Here, E1 benchmarks the performances of raw precipitation forecasts in predicting the occurrences of weekly extreme events above 99%. E2-E6 are the RF-generated categorial predictions with different input variables, and E6-E10 differs on the “max-depth” parameter of the RF model with same input combination of using all meteorological predictors. The evaluations and results are summarized with the order of previously defined EPOD, EFAR, ECSI, BSS and AUROC metrics for both sensitivity analysis cases. The individual ROC curves at NCEI climate region are provided in the supplementary information section.

Fig. 3 presents the EPOD under E1-E6 with different forecast lead times. The ideal value of EPOD is 1, indicating all extreme events have been enveloped by the spreads of ensemble forecasts. In Fig. 3, darker red colors indicate higher EPOD. The average EPOD over the entire CONUS is computed and presented in red in each subplot for comparison.

Compared to the raw forecast performances (E1), the RF-generated forecasts with P as the only input (E2) show significantly higher EPOD at the shortest lead time of week 1 and slightly higher EPOD after week 1 (Fig. 3). The inclusion of additional forecast variables of T, G500, or G850 (E3-E5) can further increase the EPOD slightly at week 2, week 3 and week 4, as it is compared to the case where only P is used to train RF (E2). When all forecast variables of P, T, G500, G850 are included together as input to train RF model (E6), additional improvements of EPOD at weeks 2–4 can be observed when comparing to the rest of the cases where only using one or two forecast variables as inputs to train RF (E2-E5). However, without additional tunning of the hyperparameter of the RF model, the inclusion of additional forecast variables (E2- E5) has slightly decreased the EPOD after week 1 compared the raw GEOS5 forecasts (E1).

Spatially, the raw precipitation forecasts (E1) deliver higher EPOD values at the West Coastal regions of CONUS. But this advantage decreases rapidly over lead times. Different spatial patterns are also observed between experiment scenarios using P, P and T, P and G500, or P and G850 as the only input to the RF model (E2-E5). While using P as the only input to the RF model (E2), higher EPOD can be observed at West Coastal regions and Appalachian Mountainous regions. While including T as an additional input to the RF model (E3), higher EPOD can be observed at most of the NCEI climate regions except for the South at week 2, week 3, and week 4. As a contrast, when including G500 or G850 as an additional input to the RF model (E4 and E5), noticeable higher EPOD values are only observed in the NCEI climate regions of East North Central, Central, as well as Southeast at week 2, week 3, and week 4. Finally, when using all available S2S forecast variables as inputs

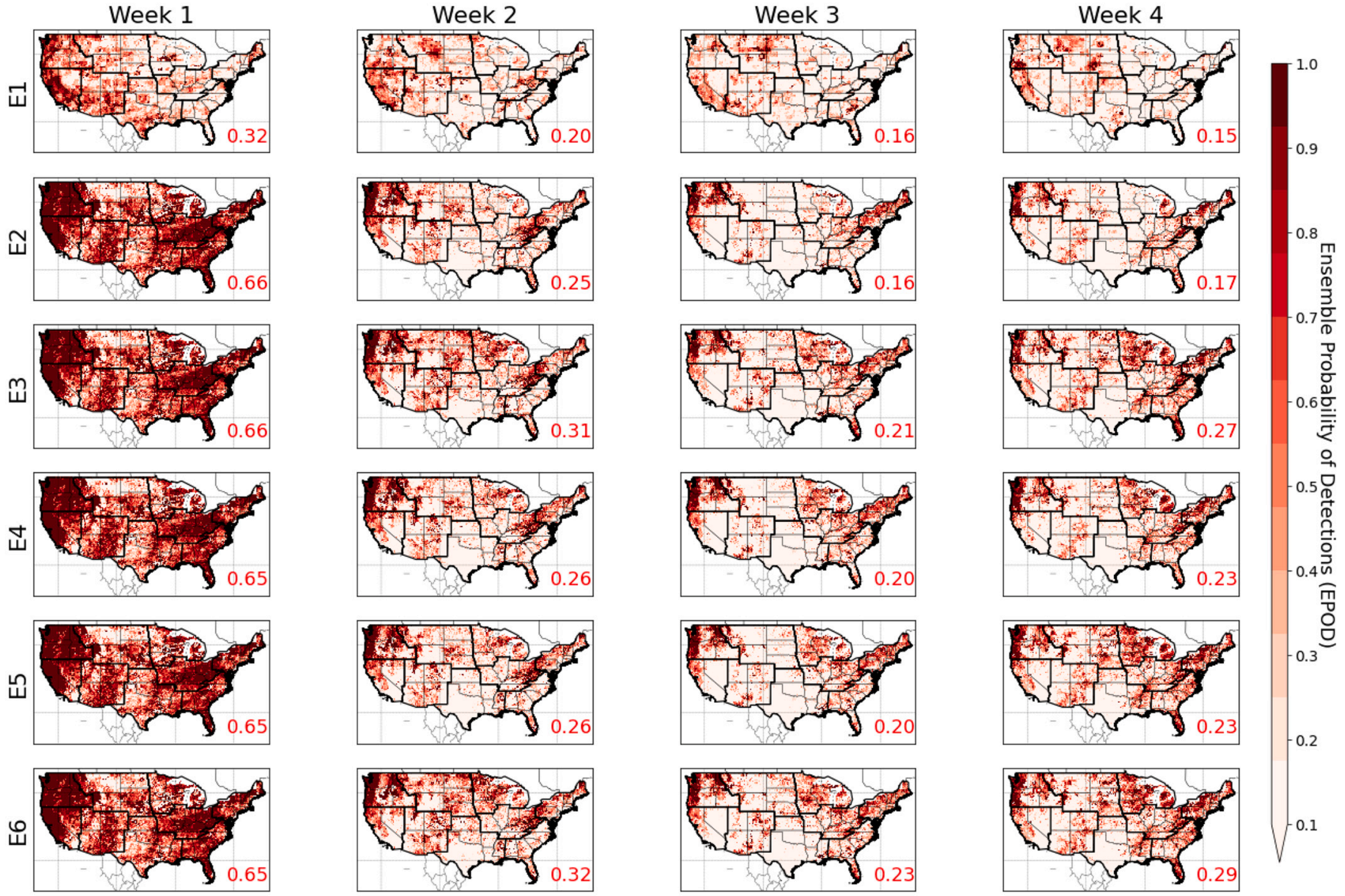

Fig. 3. The pattern of EPOD from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (weeks 1 to 4).

to RF classifiers (E6), the obtained EPOD values are increased at most of the locations over CONUS at week 2, week 3, and week 4, compared to the cases of using only one or two variables (E2-E5).

Fig. 4 presents the EFAR under E1-E6 at different forecast lead times. The ideal value of EFAR is 0, indicating the ensemble spreads of forecasts are not under-dispersed and lead to ensemble false alarms. In Fig. 4, darker blue colors indicate higher EFAR. The average EFAR over the entire CONUS is computed and presented in red in each subplot for comparison.

According to Fig. 4, the overall small EFAR values (< 0.1) of raw forecast seem evenly presented over CONUS without any noticeable high-value spots. Further, the utilization of RF model to correct extreme precipitation events (E2-E5) has further reduced the overall EFAR values over CONUS. However, the application of the RF model seems to increase EFAR at some exceptional locations over CONUS. For example, when including T as an additional input to the RF model (E3), a few bluish pixels can be observed in the Rocky Mountain regions and in the Florida Peninsula. This pattern has become more apparent when utilizing all available forecast variables as inputs to the RF model (E6), as more blue-colored pixels appeared in the Northwestern regions of CONUS, Rocky Mountainous regions and in the Florida Peninsula.

Fig. 5 presents the ECSI under E1-E6 at different forecast lead times. The ECSI considers the number of ensemble detections and ensemble false alarms at the same time. The ideal value of ECSI is 1, indicating all forecasts have been enveloped by the spreads of ensemble forecasts while no false alarms were issued due to the under-dispersion of ensemble forecasts. In Fig. 5, darker green colors indicate higher ECSI. The average ECSI over entire CONUS is computed and presented in red in each subplot for comparison.

Comparing different experiment scenarios E1-E6 across entire CONUS, it is obvious to notice that when using P as the only input to RF (E2), significantly higher skill over the raw forecasts (E1) can be observed at week 1 lead time (Fig. 5). However, such advantages become marginal at longer lead times of weeks 3 and 4. When including additional variables of T, G500, or G850 as input to the RF model (E3-E5), we found that higher ECSI values can be observed at week 2, week 3, and week 4 lead times. However, the inclusion of additional forecast variables (E3-E5) leads to a marginal level of decrease of ECSI at the shortest lead time of week 1, compared to the case where only P is used as input the RF model (E2). When all forecast variables are included as inputs to the RF model (E6), further skill improvements at longer lead times of weeks 2–4 are observed over CONUS. But the inclusion of all variables still leads to a marginal level of ECSI decrease at lead time of week 1 as it is compared to cases of using less input variables (E2-E5).

Fig. 6 presents the BSS under E1-E6 at different forecast lead times. The BSS measures the agreement between forecast probability and measured events. The ideal value of BSS is 1. In Fig. 6, darker purple colors indicate higher BSS. The average BSS values over the entire CONUS is computed and presented in red in each subplot for comparison.

The application of the RF model with P as the only input (E2) has significantly improved BSS over CONUS compared to the raw GEOS5 forecast (E1) at week 1 lead time (Fig. 6). However, the BSS got decreased over CONUS at weeks 2 to 4 with the applications of the RF model. The inclusion of T as an additional input variable to the RF model (E3) could slightly increase the BSS over CONUS at lead times of weeks 2 to 4, as it is compared to the case of using P as the only input to the RF model (E2). The inclusion of G500 or G850 as an additional input variable to the RF model (E4 and E5) could only improve the BSS over CONUS marginally at week 2 lead time, as it is compared to the case of

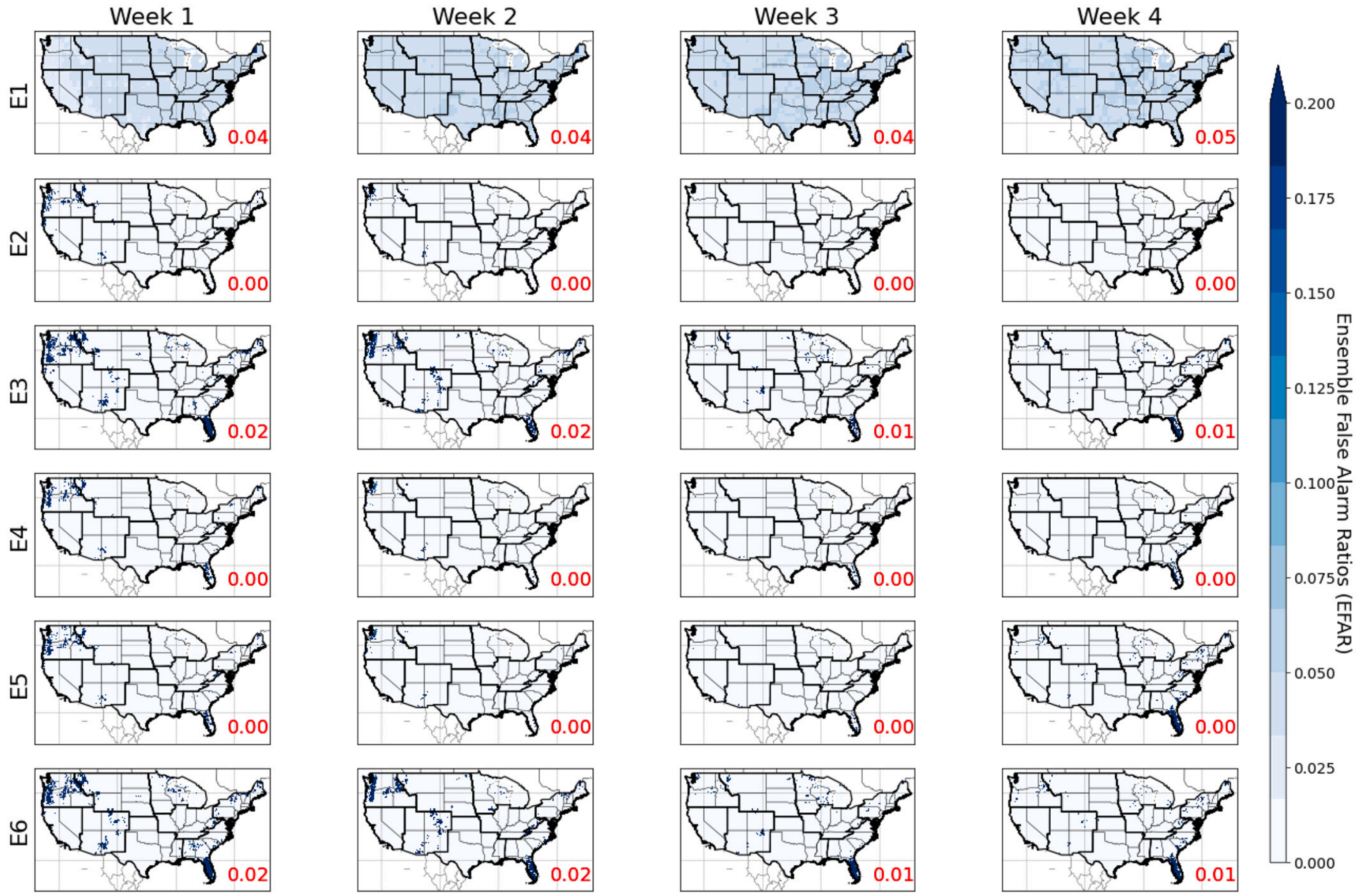

Fig. 4. The pattern of EFAR from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (weeks 1 to 4).

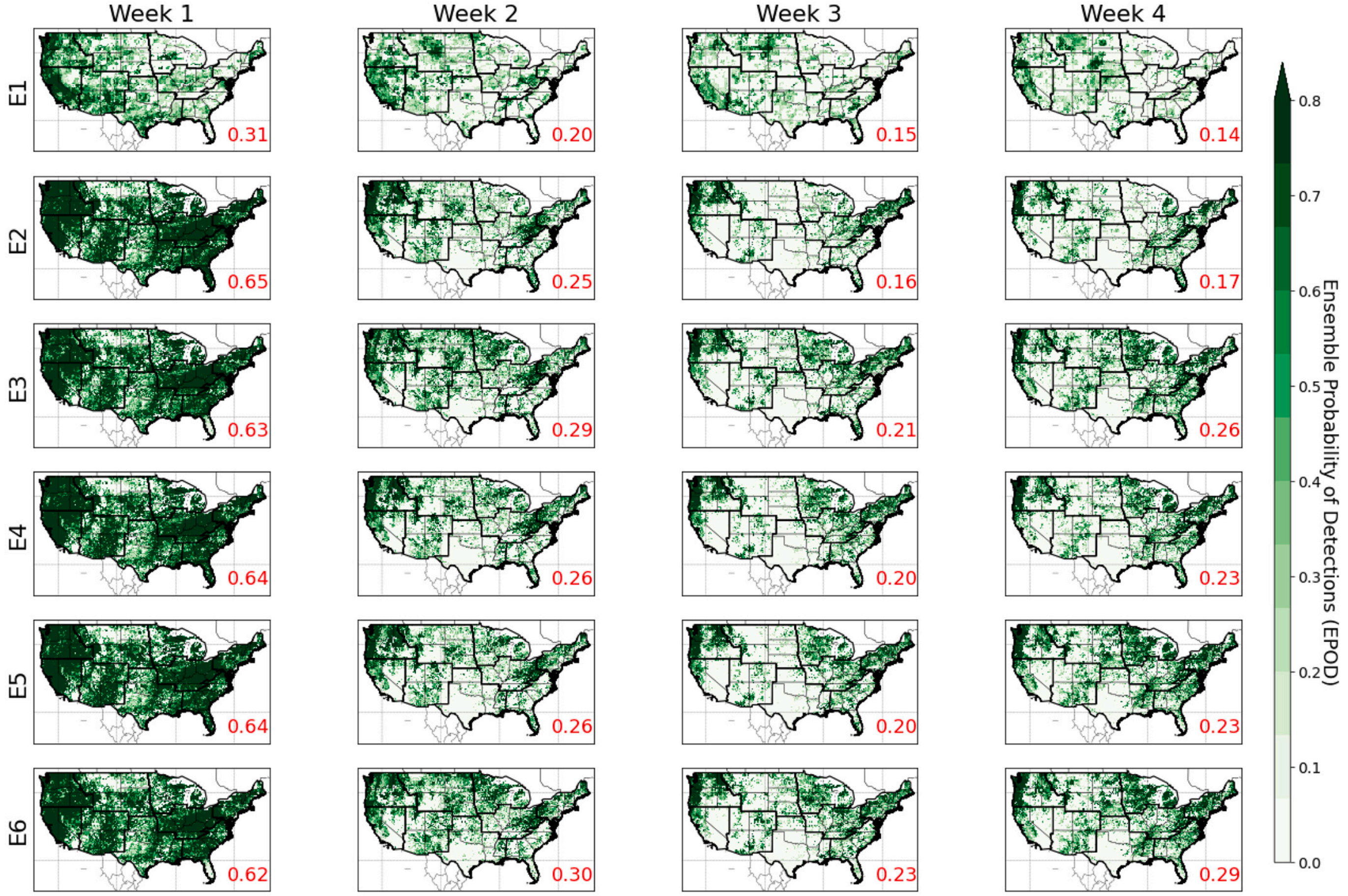

Fig. 5. The pattern of ECSI from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (weeks 1 to 4).

using P as the only input to the RF model (E2). When all available forecast variables are used as inputs to the RF model, only a marginal level of additional improvement of BSS can be observed at lead times of week 2 and week 4, as it is compared to the case of using P as the only input to the RF model (E2).

The spatial patterns of BSS over CONUS appear to be similar among all experiment scenarios from E1 to E6. At lead time of week 1, higher BSS values are observed along the western coastal regions over CONUS. At lead time of week 2, higher BSS values are observed along the Rocky mountainous regions and the Appalachian mountainous regions over CONUS. At week 3 and week 4, higher BSS values are only observed at some scattered spots in the NCEI climate regions of Southwest, West North Central, Southeast, and Northeast, given an overall marginal level improvement of BSS values over CONUS.

Fig. 7 presents the regional AUROC under E1 to E6 at different lead times. The ideal value of AUROC is 1, indicating all ensemble forecast members have performed perfectly with all extreme events detected while making no false detections at all. In Fig. 7, darker red colors represent higher AUROC. The fractional values that appeared in each colored box are the computed AUROC values. Note that the “left-to-right” sequence of 9 NCEI regions in Fig. 7 corresponds to the real-world geographical layout of these climate regions in the U.S., i.e., the columns from left to right of Fig. 9 correspond to the western coast to the eastern coast of CONUS.

A major difference of AUROC is observed between the raw forecast (E1) and the remaining experiment scenarios with the applications of the RF model (E2-E6) (Fig. 7). For example, using P as the only input to the RF model can bring noticeable improvement of AUROC at week 1 lead time at all NCEI climate regions over CONUS. However, such improvement becomes marginal at longer lead times after week 2, especially in the central regions of CONUS.

Compared to only using P as input to the RF model (E2), the inclusion of additional forecast variable brings marginal improvement of AUROC at longer lead times after week 2. The inclusion of T as input to the RF model (E3) has slightly increased the AUROC at longer lead times after week 2 at NCEI climate regions of Northwest, East North Central, Central, Northeast, and Southeast. The inclusion of G500 or G850 as an additional variable to the RF model has increased the AUROC after week 2 lead time at NCEI climate regions of Northwest, East North Central, Central, and Northeast. However, with the default hyperparameter settings of RF, the inclusion of all available forecast variables (E6) does not show overall superior AUROC values over the entire CONUS. Compare to using P as the only input to the RF model (E2), the inclusion of all available forecasts variables (E6) is able to generate higher AUROC values at lead times after week 2 over NCEI climate regions of Northwest, West North Central, East North Central, Northeast, and Southeast.

4.2. *Model's sensitivity on hyperparameter (comparison of scenarios E1 and E6-E10)*

In this section 4.2, we present the results from E1, and E6-E10, where E1 evaluates the raw precipitation forecast in predicting the occurrence of weekly extreme events above 99% and scenarios E6-E10 evaluate RF-generated categorial prediction upon weekly extreme precipitation events above 99%. In addition to EPOD, EFAR, and ECSI, we also present the regional AUROC over CONUS. The individual ROC curves at NCEI climate region are provided in the supplementary information section.

Fig. 8 presents the EPOD values under E1, and E6-E10 at different forecast lead times. The ideal value of EPOD is 1, indicating all extreme events have been enveloped by the spreads of ensemble forecasts. In

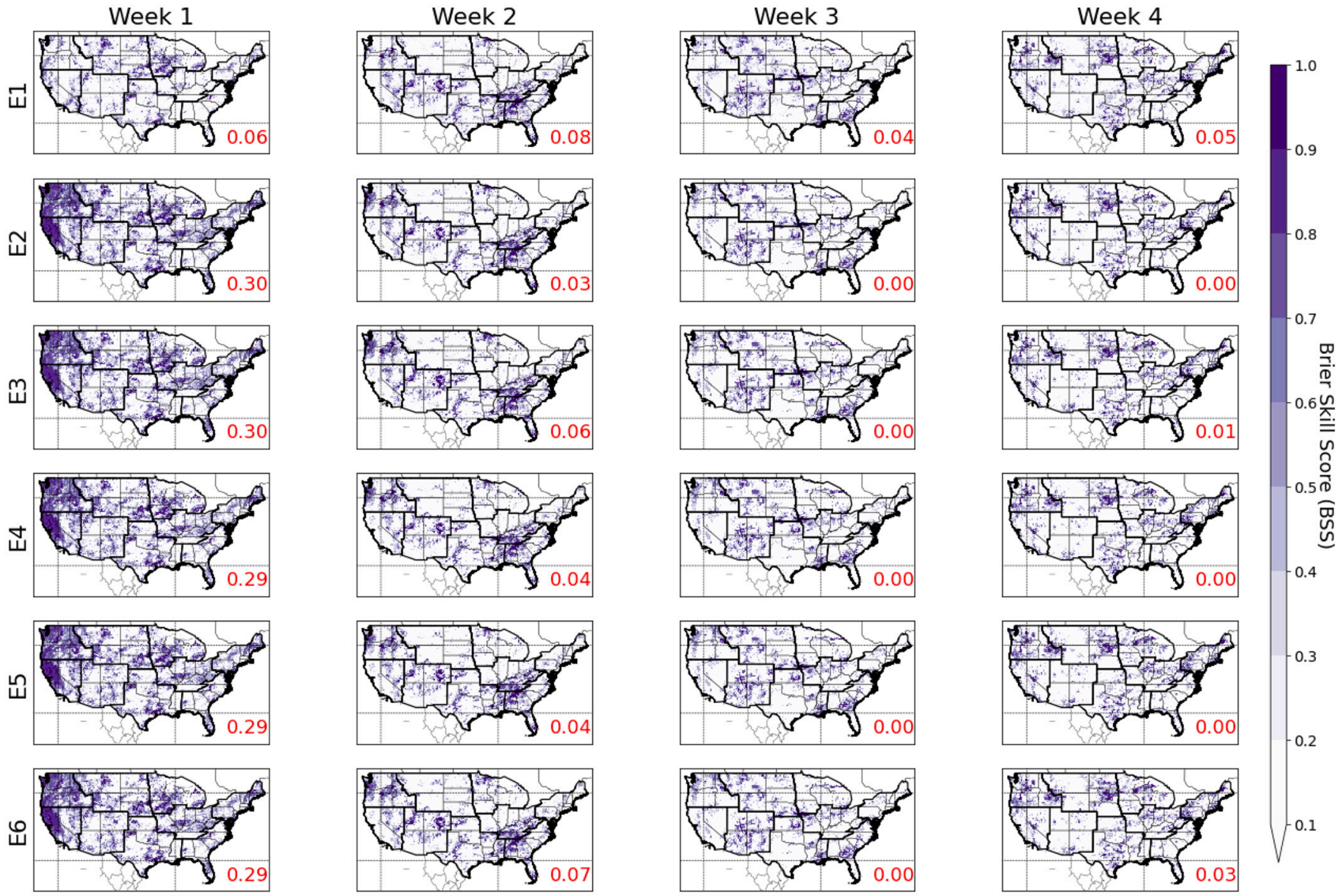

Fig. 6. The pattern of BSS from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (weeks 1 to 4).

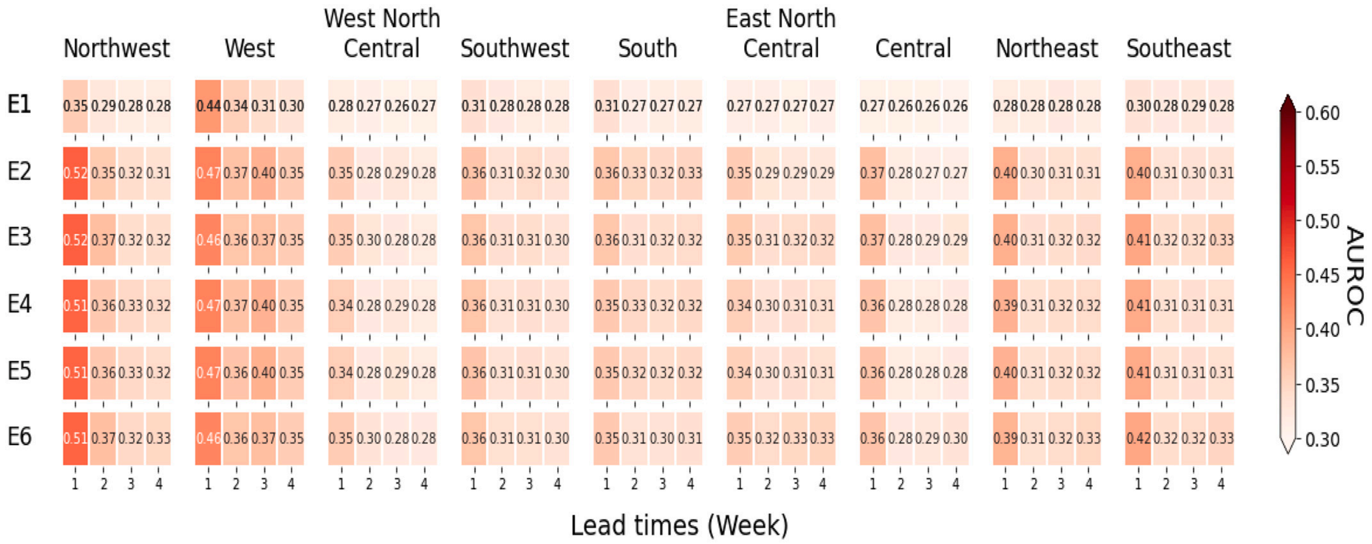

Fig. 7. Spatially-averaged AUROC values at 9 NCEI climate regions from experiment scenarios E1-E6, over CONUS and at different lead times (weeks 1 to 4).

Fig. 8, darker red colors indicate higher EPOD. The average EPOD over the entire CONUS are computed and presented in red in each subplot for comparison.

Drastic differences are observed between the performances of raw forecast (E1) and RF-generated forecasts from other experiment scenarios (E6- E10) (Fig. 8). This result agrees with previous EPOD values presented in Fig. 3, indicating that the application of RF to incorporate additional forecast variables can significantly improve the EPOD values over CONUS.

Spatial differences are observed between experiment scenarios E6-E10, in which the maximum tree depth of RF varies from 3 to 15, respectively. The EPOD values steadily increase over CONUS with a larger max tree depth used in the RF model (Fig. 8). But after the tree depth is over 9 (E8), the improvement of EPOD becomes marginal.

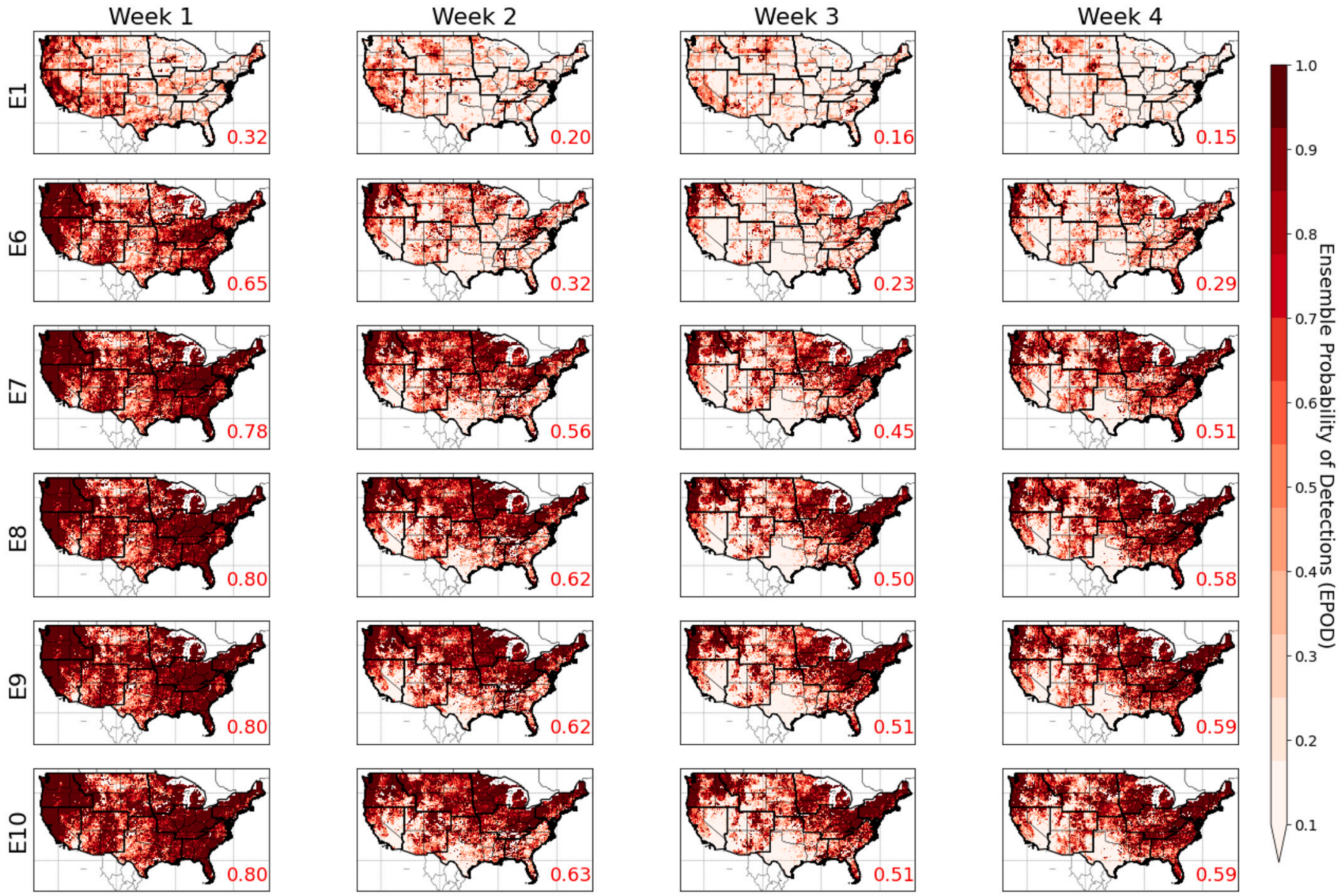

Fig. 8. The pattern of EPOD from experiment scenarios E1, and E6-E10, in predicting extreme precipitation events over the CONUS at different lead times (weeks 1 to 4).

Spatially, although a larger max tree depth leads to higher EPOD over CONUS in general, the EPOD values in some regions remain unchanged. For example, at lead time of weeks 2–4, the EPOD over Southern California and Texas is not improved at all with a larger number of max tree depth.

Fig. 9 presents the EFAR under E1, and E6-E10 at different forecast lead times. The ideal value of EFAR is 0, indicating that the ensemble spreads of forecasts are not under-dispersed and lead to ensemble false alarms. In Fig. 9, darker blue colors indicate higher EFAR. The average EFAR over entire CONUS are computed and presented in red in each subplot for comparison.

The EFAR of the raw forecasts are spatially uniform over CONUS, and they are without any noticeable high-value spots at all lead times, when compared with other experiment scenarios (E6- E10) (Fig. 9). The utilization of RF with a max tree depth of 3 (E6) reduced the EFAR values at most of the regions over CONUS. However, higher EFAR values are observed in Northwestern regions and Florida Peninsula at all lead times from E6, when it is compared to the raw forecasts. We also notice that if the tree depth is further increased (E7- E10), the EFAR will become larger at the Northwestern regions of CONUS, the Florida Peninsula, as well as the Appalachian Mountainous regions. Similarly, according to Fig. 9, the increase of tree depth seems to only increase the EFAR at certain regions (i.e., Northwestern regions, Appalachian Mountainous regions, Florida Peninsula, as well as some scattered locations lies in the Central regions of CONUS), while the EFAR over other regions seems to be not sensitive to the increase of tree depth.

Fig. 10 presents the ECSI under E1, and E6-E10 at different forecast lead times. The ideal value of ECSI is 1, indicating all forecasts have been enveloped by the spreads of ensemble forecasts while ensemble forecasts are not under-dispersed and lead to false alarms. In Fig. 10, darker green colors indicate higher ECSI. The average ECSI over entire CONUS is computed and presented in red in each subplot for comparison.

Results from Fig. 10 show an overall similar pattern to that of Fig. 8. The major difference resides in between the ECSI plots of raw forecast (E1) and remaining experiments E6-E10. According to the CONUS-averaged ECSI, the application of RF with multiple S2S forecast variables (E6-E10) has significantly improved the overall skill in predicting the occurrence of extreme precipitation events at all lead times. The results obtained under scenarios E6-E10 also show some degree of differences, which are mainly due to the increase of max tree depth used in the RF model. It is obvious that as the max tree depth increases, the overall ECSI over CONUS steadily increases. However, when tree depth exceeds 9 (E8- E10), the improvement of ECSI becomes marginal over CONUS. Among all experiment scenarios, the improvements of ECSI values appear to be the most significant when the tree depth is set to 9 (E8). Compared to the baseline E1 of raw precipitation forecasts, the relative ECSI improvements under Scenarios E6–10 are 120% (0.30 to 0.66), 155% (0.20 to 0.51), 187% (0.15 to 0.43), and 250% (0.14 to 0.49) at lead times of weeks 1–4, respectively. Spatially, we observe that a larger max tree depth will likely result in larger ECSI values over most of the regions across CONUS. But no improvements of ECSI can be observed at some specific locations, e.g., Southern California, parts of Nevada, and Texas.

Fig. 11 presents the BSS under E1 and E6-E10 at different forecast lead times. The BSS measures the agreement between forecast probability and recorded extreme events. The ideal value of BSS is 1. In Fig. 11, darker purple colors indicate higher BSS. The average BSS values over the entire CONUS are computed and presented in red in each

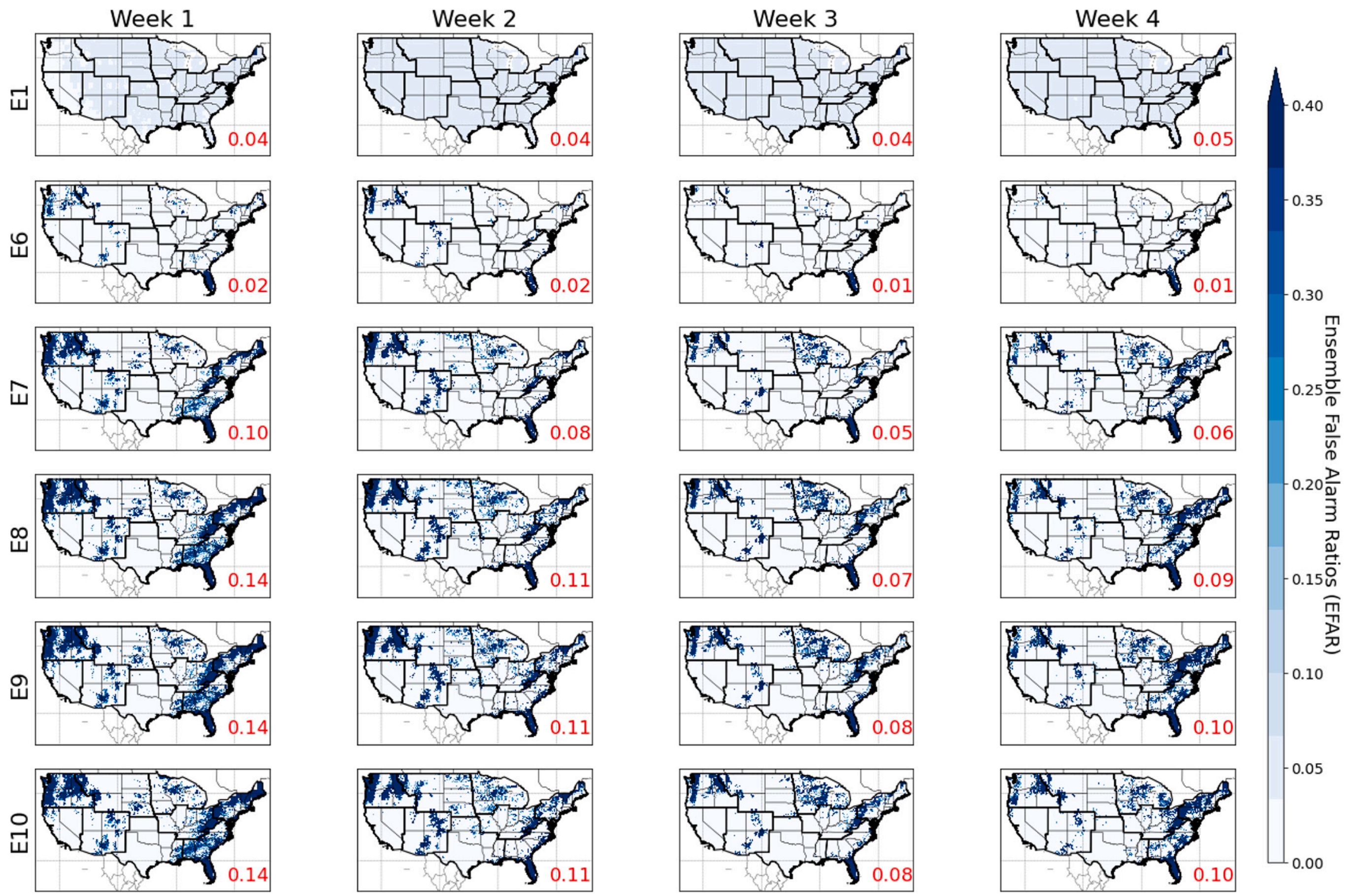

Fig. 9. The pattern of EFAR from experiment scenarios E1, and E6–10, in predicting extreme precipitation events over the CONUS at different lead times (weeks 1 to 4).

subplot for comparison.

Compared to the raw GEOS5 forecasts (E1), the utilization of the RF model with a max tree depth of 3 (E6) significantly increases the BSS at lead time of week 1 but slightly decreases the BSS at lead times of weeks 2 to 4 over the entire CONUS (Fig. 11). However, as the maximum tree depth increases (E7-E10), the post-processed forecasts present overall higher BSS at all lead times compared to the raw GEOS5 forecasts (E1) over CONUS. Among all experiment scenarios, the improvement of BSS is the most significant when the tree depth is set to 12 or 15 (E9 or E10).

For all RF-involved experiment scenarios (E6–10), the spatial patterns of BSS are similar to that of Fig. 6 (i.e., BSS from E1-E6). At lead time of week 1, the BSS values at parts of the NCEI regions of West, Southwest, South, and West North Central are consistently lower than at other regions. At week 2 lead time, higher BSS values are only observed at locations in the Rocky mountainous regions and Appalachian mountainous regions over CONUS. At lead times of week 3 and week 4, a few scattered spots show higher BSS values in the NCEI climate regions of Southwest, West North Central, Southeast, and Northeast. In general, the tunning of the maximum tree depth of RF does not improve BSS at certain locations and lead times over CONUS.

Fig. 12 presents the regional AUROC under E1, and E6-E10 at different lead times. The ideal value of AUROC is 1, indicating all ensemble forecast members have performed perfectly with all extreme events detected while making no false detections at all. In Fig. 12, darker red colors represent higher AUROC. The fractional values that appeared in each colored box are the computed AUROC values. Note that the “left-to-right” sequence of 9 NCEI regions in Fig. 12 corresponds to the real-world geographical layout of these climate regions in the U.S., i.e., the columns from left to right of Fig. 9 correspond to the western coast to the eastern coast of CONUS.

Similar to previous results, we also compare the AUROC values RF-generated forecasts from E6-E10 with the raw forecast performance (E1). It is apparent that the application of the RF model along with additional atmospheric variables (E6-E10) resulted in improved values of AUROC in all regions over CONUS and at all lead times (Fig. 12), which were similar to our previously presented results of EPOD, EFAR, and ECSI.

Fig. 12 also shows that the larger the tree depths of the RF, the higher AUROC over CONUS in general. However, some regions show less improvements of the AUROC as the max tree depth increases. These NCEI regions include West North Central, Southwest, and South (Fig. 1). Within these regions, the improvement of AUROC seems to be neglectable after the tree depth exceeds 9 (E8 to E10). Among all NCEI climate regions, the AUROC values are significantly higher in the western and eastern coastal regions of U.S. than that over other regions in the middle of the continent. The results of AUROC values indicate that the employed RF model performed the best in the NCEI region of Northwest (Fig. 1).

5. Discussion

The results from section 4.1 (E1-E6) show that without the tunning of the hyperparameter of the RF model, the post-processed S2S precipitation forecasts present mixed performances that are subject to different evaluation statistics. With the inclusion of additional forecast variables of T, G500, and/or G850, the evaluation metrics of EPOD, EFAR, and ECSI were improved in general, which were also consistent through all lead times over CONUS. But the BSS and AUROC only showed a

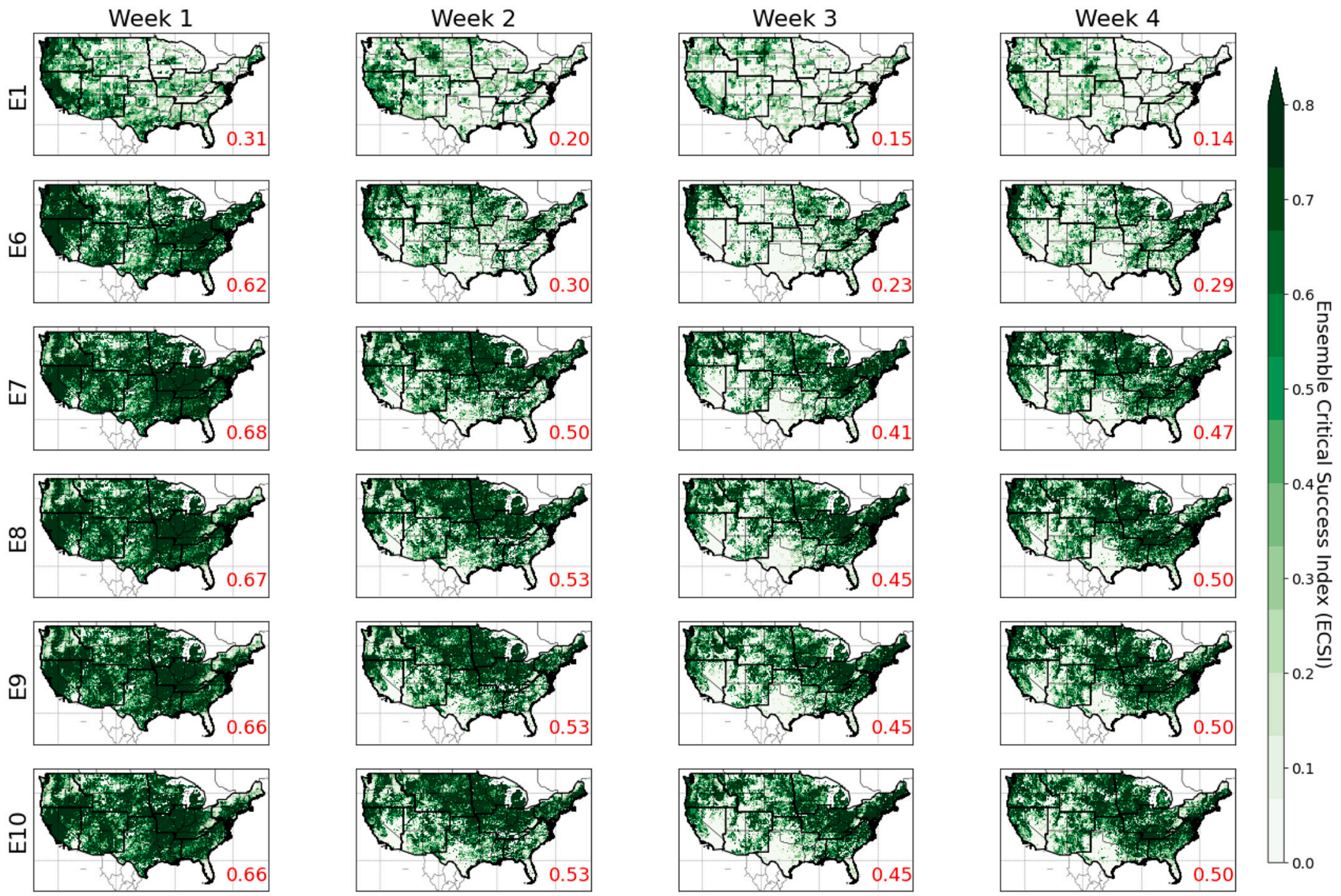

Fig. 10. The pattern of ECSI from experiment scenarios E1, and E6-E10, in predicting extreme precipitation events over the CONUS at different lead times (weeks 1 to 4).

marginal level of improvement after week 2 and sometimes slightly deteriorated at week 1 lead time. We reckon such mixed performances of the post-processed forecasts show that the inclusion of additional meteorological forecast variables is still informative, especially at longer lead times after week 2. However, the tunning of the hyperparameter of the RF model is critical when additional forecast variables are included as inputs to the RF model.

The BSS and the regional AUROC from section 4.2 (E6-E10) confirm the effectiveness of the inclusion of additional variables as well as our proposed RF model in post-processing S2S extreme precipitation forecast over CONUS. With a proper tunning scheme for the hyperparameter of the RF model, all evaluation statistics have shown noticeable improvements over CONUS. However, we also noticed that once the max maximum tree depth of the RF model exceeds 12, the overall improvements of S2S forecasts become neglectable and the EFAR values at some locations even got slightly deteriorated. We suspect this indicates an overfitting of the RF model. Although the RF model does not overfit when the number of trees of RF model is large enough (Breiman, 2001), we did not test it in our study to keep the study focus. On the other hand, when the depth of trees of RF model become too deep but train with relatively too few input variables, it is also possible that the RF model can no longer generalize over unseen points in the test dataset (Tang et al., 2018).

Nevertheless, combining all evaluation statistics from both section 4.1 and section 4.2, we reckon that the S2S extreme precipitation forecasts can be improved through the employment of RF when (1) additional atmospheric forecast variables other than precipitation are included as inputs to the RF model; and with (2) the hyperparameter of RF classifiers is manually tunned to allow the model better to capture the spatial and temporal patterns of the extreme precipitation events.

The combined use of ML model with the inclusion of additional forecast variables to improve S2S precipitation forecasts has a potential value in assisting flood predictions and river forecasts, especially at an extended range. Before the emergence of available S2S forecasts, the classical approach in predicting streamflow at the S2S timescale is to create an ensemble of multiple precipitation timeseries and by randomly resampling historical rainfall measurements. Although S2S forecasts provide an alternative to the classical approach, the accuracy and reliability of the raw S2S forecasts are rather limited as indicated by existing research, especially in predicting extreme precipitation events. Many practitioners argue that the traditional way of resampling hydrometeorological measurements is less computationally expensive and could easily gain higher probability of enveloping extreme events by creating a large ensemble size through resampling. But this study demonstrated that the employment of the RF model can improve S2S extreme precipitation forecasts without increasing the size of ensemble forecasts.

In this study, we demonstrated that additional atmospheric information, i.e., T, G500 and G850, could greatly benefit extreme precipitation forecasts at the S2S timescale. Given the fact that heavy precipitation events could be triggered through different mechanisms of convections, orographic lifts, and/or large-scale synoptic systems, it is reasonable to expect additional information could further improve the extreme precipitation forecasts at the S2S scale. We suspect that additional slope and DEM information might be helpful in identifying orographic precipitations. Additional atmospheric forecast variables (e.g., wind direction and speed, specific humidity, sea level pressure, etc.) might be able to improve the prediction upon stratiform precipitation events brought by synoptic systems. As for convective precipitation

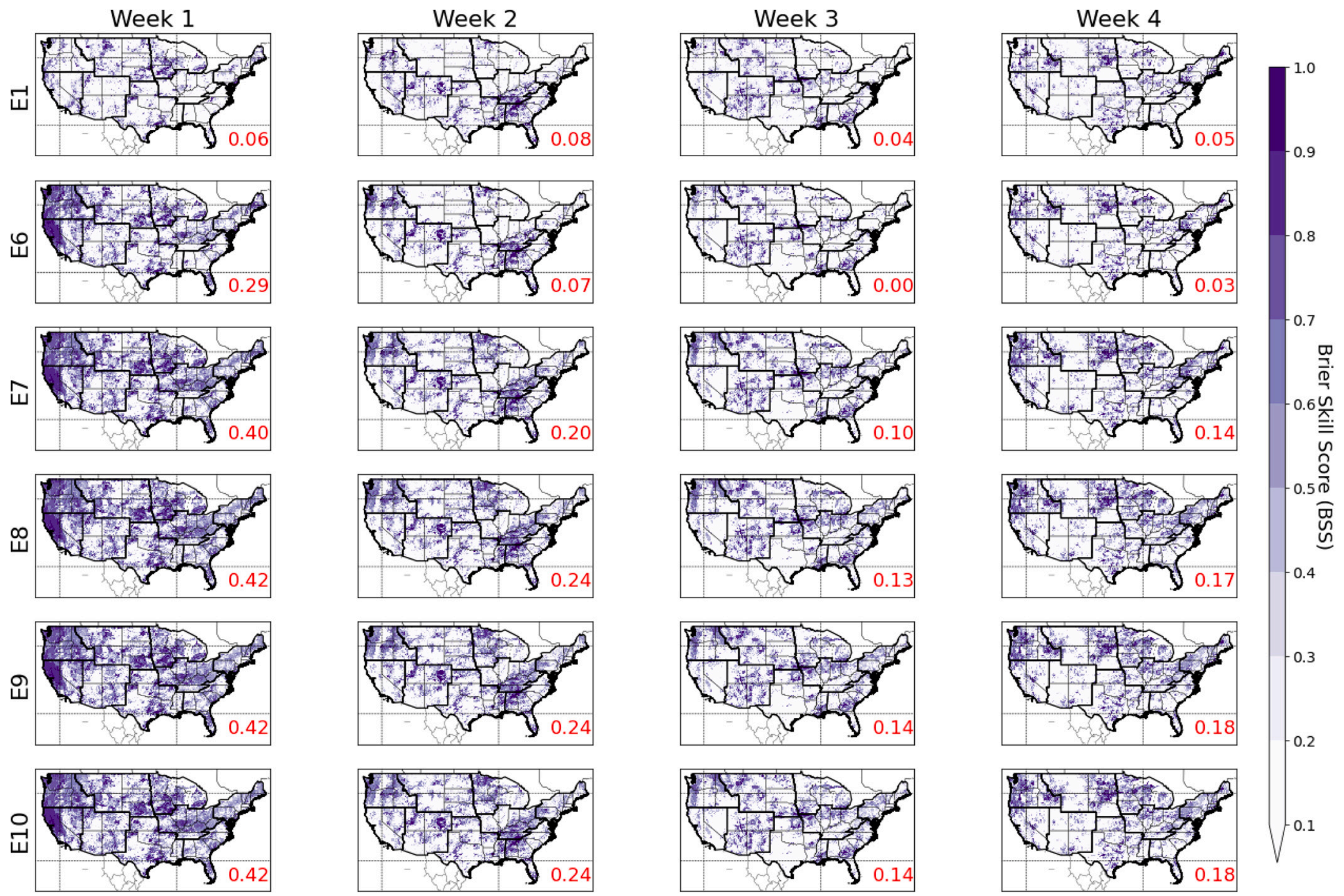

Fig. 11. The pattern of BSS from experiment scenarios E1, and E6-E10, in predicting extreme precipitation events over the CONUS at different lead times (weeks 1 to 4).

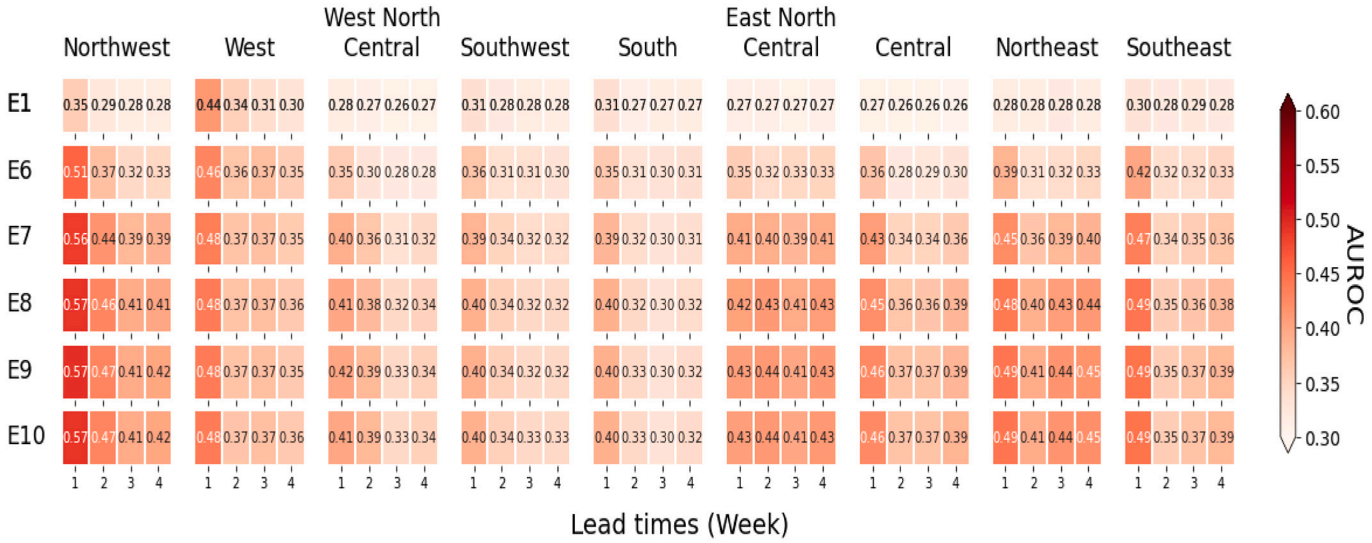

Fig. 12. Spatially-averaged AUROC values at 9 NCEI climate regions from experiment scenarios E1, and E6-E10, over CONUS and at different lead times (weeks 1 to 4).

events at the subseasonal timescales, they are sometimes considered as unpredictable given our currently limited understanding in the geographical distribution of their predictability (Moron and Robertson, 2020). Therefore, we could only hope that through the inclusion of multiple additional variables, ML&DM models may be able to identify of some patterns of convective precipitation given different local information (e.g., land cover types) as well as the projected atmospheric conditions at subseasonal timescales.

In this study, we also conducted a sensitivity test on the maximum tree depth of the RF model (section 4.2). The sensitivity test result shows that the occurrence of extreme precipitation events can be better predicted by slightly tunning one RF model hyperparameter. According to our results, a larger tree depth of the RF model can increase the overall predictive performance on the occurrence of weekly extreme events. However, such improvement becomes marginal after the tree depth reaches 9. Further increase of the tree depth does not bring significant improvement of the forecast statistical measures. For decision tree models, there are many other hyperparameters left un-tested in this study, such as the number of leaves, maximum data samples in one leaf, splitting criteria, etc. We also did not compare the predictive performance of RF with other popular ML models, such as Supportive Vector Machine, K-mean clusters, etc. There are some existing studies comparing popular ML models for interested readers in the field of hydrometeorology and water resources management (Hess and Boers, 2022; Moon et al., 2019; Nayak and Ghosh, 2013; Yang et al., 2016; Yang et al., 2021; Zhang et al., 2018). It will be a future effort to identify which ML is more capable of improving the S2S extreme precipitation forecasts together with a larger-scale model sensitivity study.

In this study, different spatial patterns of the categorial predictions are observed over CONUS. According to our regional AUROC results, as well as the obtained BSS values from section 4.2, we can confidently conclude that, in general, the extreme precipitation events that happened in the NCEI climate regions of Southwest, West North Central, and South tend to be harder to predict even with a larger tree depth of the RF model. This spatial pattern is probably because extreme precipitation events at different regions over CONUS are dominated by different meteorological mechanisms. Extreme precipitation events in the coastal regions of CONUS are largely due to or associated with synoptic-scale events, such as extratropical cyclones and atmospheric rivers (Chen et al., 2018; Kunkel et al., 2012; Mahoney et al., 2016). These types of extreme events have enormous spatial (horizontal length $>$ 100 km) and temporal (multiple days) extent thus might be easier for GCMs to consider with their coarse spatial resolutions. In contrast, mesoscale convective systems, such as thunderstorms that occur more frequently in the NCEI regions of West North Central, Southwest, South, and Central regions (Kunkel et al., 2012), are generally smaller in spatial (smaller than 100 km horizontally) and temporal (often sub-daily) scales. Thus, these convective systems are very difficult for GCMs to simulate and thus challenging for the ML&DM models to capture their spatial and temporal variabilities.

One limitation of this study is that only precipitation events above 99% threshold were examined. However, heavy precipitation events at other quantiles (e.g., above 95%, 90%, 75%, etc.) are also capable of causing floods, which should be further studied separately. Due to the limited length of this study, it will be a future effort to examine the available S2S precipitation forecasts in predicting heavy precipitation events at other quantiles.

Another follow-on study could be devoted to restoring the positive categorical predictions upon extreme events back to numerical values, i.e., rain rates. Afterall, categorical predictions cannot be directly used for hydrologic simulations to provide information on the magnitude of future streamflow. One simple way to resolve this issue is utilizing popular distribution-based approaches, such as quantile mapping (Cannon et al., 2015; Maraun, 2013), to conduct bias corrections upon S2S precipitation forecasts. Specifically, the RF-categorized extreme and non-extreme events will be corrected/restored to numerical values according to historical records and thus ready for next-step hydrologic forecasts. However, we expect that the treating of extreme values might be challenging, since a few recent studies identified that the extreme precipitation events over the globe exhibited significant changes in frequency and magnitudes under global climate change (Fan et al., 2021; Kunkel et al., 2003; Madsen et al., 2014; Sun et al., 2017). Thus, when bias-correcting the extreme precipitation forecasts from S2S models, the considerations on the “non-stationary” of climate and its impacts are also needed. (AghaKouchak et al., 2011; Cheng et al., 2014; Tao et al., 2018).

Finally, the application of S2S precipitation forecast products in hydrologic forecasts at watershed scales is still an ongoing effort due to the raw forecasts' coarse spatial resolutions. We encourage practitioners to further evaluate and apply S2S precipitation forecasts for riverine flood predictions at a daily time step (Cao et al., 2021; Li et al., 2017; McInerney et al., 2020; Quedi and Fan, 2020; Kim et al., 2021). This will help overcome the pitfall of the current study that only evaluates the S2S forecasts limited number of statistical metrics on a weekly basis. We encourage future research to be devoted to answering the questions of how post-processed S2S ensemble precipitation forecasts perform in predicting floods induced by extreme precipitation events.

In summary, future research could be devoted to (1) investigating how additional variables further help ML models to better post-process S2S forecasts and identify the intrinsic physical dynamics related to extreme precipitation over CONUS; (2) quantifying the influences of different ML hyperparameters on the predictive performance upon extreme precipitation events at the S2S timescale; (3) capturing the structural and random errors associated with different S2S forecast models at different spatial and temporal domains; and (4) linking the S2S forecasts to ensemble hydrologic forecasting and further investigating the usefulness of S2S forecasts in assisting ensemble streamflow forecasting of riverine floods.

6. Conclusion

In the current research, we conducted comprehensive evaluations regarding the performance of GEOS5 S2S forecasts with a specific focus on weekly extreme precipitation events over CONUS. We developed a proof-of-concept forecast adaptation framework using the Random Forest (RF) classifier to post-process the raw S2S forecasts at each 0.25-degree pixel over CONUS. Four S2S forecast variables are used, including precipitation, surface air temperature, and geopotential height at 500 hPa and 850 hPa. We examined the performance of RF with respect to different maximum tree depths over CONUS. A total of ten different experiment scenarios are created to identify the ideal input variable combination and maximum tree depth that leads to the best predictions upon extreme precipitation events.

To evaluate the skill and demonstrate the potential hydrologic effectiveness of S2S ensemble forecast, we employed modified categorical metrics of ensemble probability of detections (EPOD), ensemble false alarm ratios (EFAR), and ensemble critical success index (ECSI). The probabilistic evaluation metric of Brier Skill Score (BSS), as well as the regional Area Under the Receiver Operating Characteristics Curves (AUROC) are computed to confirm the improvements of S2S precipitation forecast. The improved S2S forecasts will be useful for future hydrologic studies in different spatial and temporal domains. Our research conclusions are listed as follows:

1. The application of RF can significantly improve S2S forecasts in terms of predicting the occurrence of weekly extreme precipitation events. The improvement is most significant at lead time of week 1 and deteriorates rapidly after week 2 lead time. We found out that by including additional S2S forecast variables as the RF inputs, the forecast performance is further improved at longer lead times (i.e., weeks 2–4). Compared to the raw forecasts, RF could improve ECSI up to 116% (0.31 to 0.67), 165% (0.20 to 0.53), 200% (0.15 to 0.45), and 257% (0.14 to 0.50), at weeks 1–4, respectively.
2. We also found that the extreme precipitation events in West North Central, Southwest, and South regions, per NCEI definition, are harder to predict than that in other regions. We speculate such differences are due to different dominating precipitation mechanisms, which may result in different spatial and temporal scales of extreme precipitation events.
3. Sensitivity analysis indicates that increasing the maximum tree depth of RF would result in overall better forecasts at all lead times over CONUS. However, the improvements become neglectable once

the tree depth exceeds nine in our study. We suspect this indicates an overfitting of the RF model. The obtained BSS and AUROC show consistent conclusions as indicated by EPOD, EFAR, and ECSI, which confirms the improvement of S2S forecast is not subject to the definitions of modified evaluation metrics.

Credit Author Statement.

Lujun Zhang: Conceptualization, Methodology, Software, Writing - original draft. **Tiantian Yang:** Methodology, Writing - review & editing, funding acquisition, supervision. **Shang Gao**: Writing - review & editing **Yang Hong:** Funding acquisition, supervision. **Qin Zhang:** Supervision. **Xin Wen:** Writing - review & editing. **Chuntian Cheng:** 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.

Data availability

Data will be made available on request.

Acknowledgments

The material is based upon work supported by the National Science Foundation (NSF) under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3, and the Oklahoma EPSCoR Research Seed Grand supported by NSF. This work is also partially supported by U.S. Department of Defense, Army Corps of Engineers (DOD-COR) Engineering With Nature (EWN) Program (Award No. W912HZ-21-2-0038); This research was also supported by a grant from the Research Council of the University of Oklahoma Norman Campus. The computing for this project was performed at the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosres.2022.106502.

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).

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

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.

Bader, M., Roach, W., 1977. Orographic rainfall in warm sectors of depressions. Q. J. R. Meteorol. Soc. 103 (436), 269–280.

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 Res. Assoc. 55 (4), 1024–1037.

Baker, S.A., Wood, A.W., Rajagopalan, B., 2020. Application of postprocessing to watershed-scale subseasonal climate forecasts over the contiguous United States. J. Hydrometeorol. 21 (5), 971–987.

Baño-Medina, J., Manzanas, R., Gutiérrez, J.M., 2020. Configuration and intercomparison of deep learning neural models for statistical downscaling. Geosci. Model Dev. 13 (4), 2109–2124.

Begum, S., Stive, M.J., Hall, J.W., 2007. Flood Risk Management in Europe: Innovation in Policy and Practice. Springer Science & Business Media.

Best, M., Pryor, M., Clark, D., Rooney, G., Essery, R., Ménard, C., Edwards, J., Hendry, M., Porson, A., Gedney, N., 2011. The Joint UK Land Environment Simulator (JULES), model description–part 1: energy and water fluxes. Geosci. Model Dev. 4 (3), 677–699.

Betts, A.K., Viterbo, P., Wood, E., 1998. Surface energy and water balance for the Arkansas–Red River basin from the ECMWF reanalysis. J. Clim. 11 (11), 2881–2897.

Borovikov, A., Cullather, R., Kovach, R., Marshak, J., Vernieres, G., Vikhliaev, Y., Zhao, B., Li, Z., 2019. GEOS-5 seasonal forecast system. Clim. Dyn. 53 (12), 7335–7361.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Bremond, P., Grelot, F., Agenais, A.-L., 2013. Economic evaluation of flood damage to agriculture–review and analysis of existing methods. Nat. Hazards Earth Syst. Sci. 13 (10), 2493–2512.

Brier, Glenn W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78 (1), 1–3.

Bukovsky, M.S., Karoly, D.J., 2007. A brief evaluation of precipitation from the north American Regional Reanalysis. J. Hydrometeorol. 8 (4), 837–846.

Cannon, A.J., Sobie, S.R., Murdock, T.Q., 2015. Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? J. Clim. 28 (17), 6938–6959.

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.

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.

Chen, X., Leung, L.R., Gao, Y., Liu, Y., Wigmosta, M., Richmond, M., 2018. Predictability of extreme precipitation in western US watersheds based on atmospheric river occurrence, intensity, and duration. Geophys. Res. Lett. 45 (21), 11,693-611,701.

Cheng, L., AghaKouchak, A., Gilleland, E., Katz, R.W., 2014. Non-stationary extreme value analysis in a changing climate. Clim. Chang. 127 (2), 353–369.

Clark, R.T., Bett, P.E., Thornton, H.E., Scaife, A.A., 2017. Skilful seasonal predictions for the European energy industry. Environ. Res. Lett. 12 (2), 024002.

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).

Dai, A., Wigley, T., 2000. Global patterns of ENSO-induced precipitation. Geophys. Res. Lett. 27 (9), 1283–1286.

Daly, C., Bryant, K., 2013. The PRISM Climate and Weather System—An Introduction. PRISM climate group, Corvallis, OR, p. 4.

Davis, J., Goadrich, M., 2006. The Relationship between Precision-Recall and ROC Curves, pp. 233–240.

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

de Andrade, F.M., Coelho, C.A., Cavalcanti, I.F., 2019. Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. Clim. Dyn. 52 (9), 5451–5475.

de Andrade, F.M., Young, M.P., MacLeod, D., Hirons, L.C., Woolnough, S.J., Black, E., 2021. Subseasonal precipitation prediction for Africa: Forecast evaluation and sources of predictability. Weather Forecast. 36 (1), 265–284.

Ebert, E., McBride, J., 2000. Verification of precipitation in weather systems: determination of systematic errors. J. Hydrol. 239 (1–4), 179–202.

Fan, X., Miao, C., Duan, Q., Shen, C., Wu, Y., 2021. Future climate change hotspots under different 21st century warming scenarios. Earth's. Future 9 (6), e2021EF002027.

Faridzad, M., Yang, T., Hsu, K., Sorooshian, S., Xiao, C., 2018. Rainfall frequency analysis for ungauged regions using remotely sensed precipitation information. J. Hydrol. 563, 123–142.

Gowan, T.M., Steenburgh, W.J., Schwartz, C.S., 2018. Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. Weather Forecast. 33 (3), 739–765.

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

He, Xiaogang, et al., 2016. Spatial downscaling of precipitation using adaptable random forests. Water Resour. Res. 52 (10), 8217–8237.

Herman, Gregory R., Schumacher, Russ S., 2018a. Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. Mon. Weather Rev. 146 (5), 1571–1600.

Herman, Gregory R., Schumacher, Russ S., 2018b. Dendrology in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. Mon. Weather Rev. 146 (6), 1785–1812.

Hess, P., Boers, N., 2022. Deep learning for improving numerical weather prediction of heavy rainfall. J. Adv. Model. Earth Syst. 14 (3), e2021MS002765.

Hsiao, W.T., Maloney, E.D., Barnes, E.A., 2020. Investigating recent changes in MJO precipitation and circulation in multiple reanalyses. Geophys. Res. Lett. 47 (22), e2020GL090139.

Karl, T., Koss, W.J., 1984. Regional and National Monthly, Seasonal, and Annual Temperature Weighted by Area, pp. 1895–1983.

Kim, T., Yang, T., Gao, S., Zhang, L., Ding, Z., Wen, X., Hong, Y., 2021. Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: a case study of four watersheds with different hydro-climatic regions across the CONUS. J. Hydrol. 598, 126423.

Kim, T., Yang, T., Zhang, L., Hong, Y., 2022. Near real-time hurricane rainfall forecasting using convolutional neural network models with Integrated multi-satellitE Retrievals for GPM (IMERG) product. Atmos. Res. 270, 106037.

King, A.D., Hudson, D., Lim, E.P., Marshall, A.G., Hendon, H.H., Lane, T.P., Alves, O., 2020. Sub-seasonal to seasonal prediction of rainfall extremes in Australia. Q. J. R. Meteorol. Soc. 146 (730), 2228–2249.

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., 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.

Kuligowski, R.J., Barros, A.P., 1998. Experiments in short-term precipitation forecasting using artificial neural networks. Mon. Weather Rev. 126 (2), 470–482.

Kunkel, K.E., Easterling, D.R., Redmond, K., Hubbard, K., 2003. Temporal variations of extreme precipitation events in the United States: 1895–2000. Geophys. Res. Lett. 30 (17).

Kunkel, K.E., Easterling, D.R., Kristovich, D.A., Gleason, B., Stoecker, L., Smith, R., 2012. Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous United States. J. Hydrometeorol. 13 (3), 1131–1141.

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

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.

Li, W., Pan, B., Xia, J., Duan, Q., 2022. Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. J. Hydrol. 605, 127301.

Loken, Eric D., et al., 2019. Postprocessing next-day ensemble probabilistic precipitation forecasts using random forests. Weather Forecast. 34 (6), 2017–2044.

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

Mahoney, K., Jackson, D.L., Neiman, P., Hughes, M., Darby, L., Wick, G., White, A., Sukovich, E., Cifelli, R., 2016. Understanding the role of atmospheric rivers in heavy precipitation in the Southeast United States. Mon. Weather Rev. 144 (4), 1617–1632.

Manzanas, R., Lucero, A., Weisheimer, A., Gutiérrez, J.M., 2018. Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? Clim. Dyn. 50 (3), 1161–1176.

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

Mariotti, A., Baggett, C., Barnes, E.A., Becker, E., Butler, A., Collins, D.C., Dirmeyer, P.A., Ferranti, L., Johnson, N.C., Jones, J., 2020. Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. Bull. Am. Meteorol. Soc. 101 (5), E608–E625.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., Kuczera, G., 2020. Multi-temporal hydrological residual error modeling for seamless subseasonal streamflow forecasting. Water Resour. Res. 56 (11), e2019WR026979.

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.

Mizukami, N., Smith, M.B., 2012. Analysis of inconsistencies in multi-year gridded quantitative precipitation estimate over complex terrain and its impact on hydrologic modeling. J. Hydrol. 428, 129–141.

Moon, S.-H., Kim, Y.-H., Lee, Y.H., Moon, B.-R., 2019. Application of machine learning to an early warning system for very short-term heavy rainfall. J. Hydrol. 568, 1042–1054.

Moron, Vincent, Robertson, Andrew W., 2020. Tropical rainfall subseasonal-to-seasonal predictability types. npj Clim. Atmos. 3 (1), 1–8.

Nayak, M.A., Ghosh, S., 2013. Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. Theor. Appl. Climatol. 114 (3), 583–603.

Nie, J., Dai, P., Sobel, A.H., 2020. Dry and moist dynamics shape regional patterns of extreme precipitation sensitivity. Proc. Natl. Acad. Sci. 117 (16), 8757–8763.

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S., 2019. Improving precipitation estimation using convolutional neural network. Water Resour. Res. 55 (3), 2301–2321.

Pan, B., Anderson, G.J., Goncalves, A., Lucas, D.D., Bonfils, C.J., Lee, J., Tian, Y., Ma, H. Y., 2021. Learning to correct climate projection biases. J. Adv. Model. Earth Syst. 13 (10), e2021MS002509.

Pegion, K., Kirtman, B.P., Becker, E., Collins, D.C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D., Zhu, Y., 2019. The Subseasonal Experiment (SubX): a multimodel subseasonal prediction experiment. Bull. Am. Meteorol. Soc. 100 (10), 2043–2060.

Pendergrass, A.G., 2020. Changing degree of convective organization as a mechanism for dynamic changes in extreme precipitation. Curr. Clim. Change Rep. 6 (2), 47–54.

Prat, O., Nelson, B., 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.

Quedi, E.S., Fan, F.M., 2020. Sub seasonal streamflow forecast assessment at large-scale basins. J. Hydrol. 584, 124635.

Robertson, D., Shrestha, D., Wang, Q., 2013. Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. Hydrol. Earth Syst. Sci. 17 (9), 3587–3603.

Sadeghi, M., Nguyen, P., Hsu, K., Sorooshian, S., 2020. Improving near real-time precipitation estimation using a U-Net convolutional neural network and geographical information. Environ. Model. Softw. 134, 104856.

Singh, O., Kumar, M., 2013. Flood events, fatalities and damages in India from 1978 to 2006. Nat. Hazards 69 (3), 1815–1834.

Sorooshian, S., AghaKouchak, A., Arkin, P., Eylander, J., Foufoula-Georgiou, E., Harmon, R., Hendrickx, J.M., Imam, B., Kuligowski, R., Skahill, B., 2011. Advancing the remote sensing of precipitation. Bull. Am. Meteorol. Soc. 92 (10), 1271–1272.

Srinivas, C., Yesubabu, V., Prasad, D.H., Prasad, K.H., Greeshma, M., Baskaran, R., Venkatraman, B., 2018. Simulation of an extreme heavy rainfall event over Chennai, India using WRF: Sensitivity to grid resolution and boundary layer physics. Atmos. Res. 210, 66–82.

Stensrud, D.J., 2009. Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models. Cambridge University Press.

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.

Strobl, C., Zeileis, A., 2008. Danger: High power!–exploring the statistical properties of a test for random forest variable importance.

Suarez, P., Anderson, W., Mahal, V., Lakshmanan, T., 2005. Impacts of flooding and climate change on urban transportation: a systemwide performance assessment of the Boston Metro Area. Transp. Res. Part D: Transp. Environ. 10 (3), 231–244.

Sun, Q., Miao, C., Qiao, Y., Duan, Q., 2017. The nonstationary impact of local temperature changes and ENSO on extreme precipitation at the global scale. Clim. Dyn. 49 (11), 4281–4292.

Tang, C., Garreau, D., von Luxburg, U., 2018. When do random forests fail? Adv. Neural Inf. Proces. Syst. 31.

Tao, Y., Yang, T., Faridzad, M., Jiang, L., He, X., Zhang, X., 2018. Non-stationary bias correction of monthly CMIP5 temperature projections over China using a residual-based bagging tree model. Int. J. Climatol. 38 (1), 467–482.

Taylor, J., Man Lai, K., Davies, M., Clifton, D., Ridley, I., Biddulph, P., 2011. Flood management: prediction of microbial contamination in large-scale floods in urban environments. Environ. Int. 37 (5), 1019–1029.

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), 49–65.

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

VanBuskirk, O., Ćwik, P., McPherson, R.A., Lazrus, H., Martin, E., Kuster, C., Mullens, E., 2021. Listening to stakeholders: initiating research on subseasonal-to-seasonal heavy precipitation events in the contiguous united states by first understanding what stakeholders need. Bull. Am. Meteorol. Soc. 102 (10), E1972–E1986.

Vigaud, N., Robertson, A.W., Tippett, M.K., 2017. Multimodel ensembling of subseasonal precipitation forecasts over North America. Mon. Weather Rev. 145 (10), 3913–3928.

Vitart, F., 2004. Monthly forecasting at ECMWF. Mon. Weather Rev. 132 (12), 2761–2779.

Vitart, F., 2014. Evolution of ECMWF sub-seasonal forecast skill scores. Q. J. R. Meteorol. Soc. 140 (683), 1889–1899.

Vitart, F., 2017. Madden—Julian Oscillation prediction and teleconnections in the S2S database. Q. J. R. Meteorol. Soc. 143 (706), 2210–2220.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., 2017. The subseasonal to seasonal (S2S) prediction project database. Bull. Am. Meteorol. Soc. 98 (1), 163–173.

Wang, L., Robertson, A.W., 2019. Week 3–4 predictability over the United States assessed from two operational ensemble prediction systems. Clim. Dyn. 52 (9), 5861–5875.

Wang, F., Tian, D., Lowe, L., Kalin, L., Lehrter, J., 2021. Deep learning for daily precipitation and temperature downscaling. Water Resour. Res. 57 (4), e2020WR029308.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J., Murray, V., 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorol. Appl. 24 (3), 315–325.

Wilks, Daniel S., 2011. Statistical Methods in the Atmospheric Sciences, 100. 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.

Wu, W., Emerton, R., Duan, Q., Wood, A.W., Wetterhall, F., Robertson, D.E., 2020. Ensemble flood forecasting: current status and future opportunities. Wiley Interdiscip. Rev. Water 7 (3), e1432.

Xiang, B., Lin, S.J., Zhao, M., Johnson, N.C., Yang, X., Jiang, X., 2019. Subseasonal week 3–5 surface air temperature prediction during boreal wintertime in a GFDL model. Geophys. Res. Lett. 46 (1), 416–425.

Yang, T., Gao, X., Sellars, S.L., Sorooshian, S., 2015. Improving the multi-objective evolutionary optimization algorithm for hydropower reservoir operations in the California Oroville–Thermalito complex. Environ. Model. Softw. 69, 262–279.

Yang, T., Gao, X., Sorooshian, S., Li, X., 2016. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. Water Resour. Res. 52 (3), 1626–1651.

Yang, T., Asanjan, A.A., Welles, E., Gao, X., Sorooshian, S., Liu, X., 2017a. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resour. Res. 53 (4), 2786–2812.

Yang, T., Asanjan, A.A., Faridzad, M., Hayatbini, N., Gao, X., Sorooshian, S., 2017b. An enhanced artificial neural network with a shuffled complex evolutionary global optimization with principal component analysis. Inf. Sci. 418, 302–316.

Yang, T., Tao, Y., Li, J., Zhu, Q., Su, L., He, X., Zhang, X., 2018. Multi-criterion model ensemble of CMIP5 surface air temperature over China. Theor. Appl. Climatol. 132 (3), 1057-1072.as.

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. Plan. Manag. 146 (2), 04019072.

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.

Yang, Z., Li, B., Xia, R., Ma, S., Jia, R., Ma, C., Wang, L., Chen, Y., Bin, L., 2022. Understanding China's industrialization driven water pollution stress in 2002–2015—A multi-pollutant based net gray water footprint analysis. J. Environ. Manag. 310, 114735.

Yuan, X., Wood, E.F., Ma, Z., 2015. A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. Wiley Interdiscip. Rev. Water 2 (5), 523–536.

Zhang, C., Ling, J., 2017. Barrier effect of the Indo-Pacific Maritime Continent on the MJO: Perspectives from tracking MJO precipitation. J. Clim. 30 (9), 3439–3459.

Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Sorooshian, S., Liu, X., Zhuang, J., 2018. Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. J. Hydrol. 565, 720–736.

Zhang, L., Kim, T., Yang, T., Hong, Y., Zhu, Q., 2021. Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the north American Multi-Model ensemble phase II (NMME-2) over the contiguous US. J. Hydrol. 603, 127058.

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