Predictability of U.S. Regional Extreme Precipitation Occurrence Based on Large-Scale Meteorological Patterns (LSMPs)

XIANG GAOa AND SHRAY MATHURb

a Joint Program on the Science and Policy of Global Change, Massachusetts Institute of Technology, Cambridge, Massachusetts
b Department of Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani, India

(Manuscript received 16 February 2021, in final form 22 April 2021)

ABSTRACT: In this study, we use analogue method and convolutional neural networks (CNNs) to assess the potential predictability of extreme precipitation occurrence based on large-scale meteorological patterns (LSMPs) for the winter (DJF) of “Pacific Coast California” region (PCCA) and the summer (JJA) of the midwestern United States (MWST). We evaluate the LSMPs constructed with a large set of variables at multiple atmospheric levels and quantify the prediction skill with a variety of complementary performance measures. Our results suggest that LSMPs provide useful predictability of daily extreme precipitation occurrence and its interannual variability over both regions. The 14-yr (2006–19) independent forecast shows Gilbert skill scores (GSS) in PCCA ranging from 0.06 to 0.32 across 24 CNN schemes and from 0.16 to 0.26 across four analogue schemes, in contrast from 0.1 to 0.24 and from 0.1 to 0.14 in MWST. Overall, CNN is shown to be more powerful in extracting the relevant features associated with extreme precipitation from the LSMPs than analogue method, with several single-variate CNN schemes achieving more skillful prediction than the best multivariate analogue scheme in PCCA and more than half of CNN schemes in MWST. Nevertheless, both methods highlight that the integrated vapor transport (IVT, or its zonal and meridional components) enables higher skills than other atmospheric variables over both regions. Warm-season extreme precipitation in MWST presents a forecast challenge with overall lower prediction skill than in PCCA, attributed to the weak synoptic-scale forcing in summer.

KEYWORDS: North America; Atmospheric circulation; Extreme events; Downscaling; Neural networks; Pattern recognition; Precipitation; Deep learning

1. Introduction

Extreme precipitation can lead to severe socioeconomic impacts and is also expected to change in severity, frequency, and duration as a result of anthropogenic global warming (Min et al. 2011; Kharin et al. 2013; Sillmann et al. 2013). However, skill is often limited for global climate models to capture these localized extremes due to the lack of the ability to resolve the relevant local terrain and mesoscale systems at their typical model resolutions (DeAngelis et al. 2013; Gao et al. 2014). Regional climate models are capable of providing more realistic representation of topography and mesoscale processes, but limited by their computational feasibility and high sensitivity to the chosen physical parameterizations and lateral boundary conditions (Christensen et al. 2007; Wehner 2013; Gao and Schlosser 2019). Nevertheless, it has been shown that large-scale meteorological patterns (LSMPs) accompanying extreme precipitation are well resolved in both weather and climate models (DeAngelis et al. 2013; Kawazoe and Gutowski 2013) and can thus provide a great potential for predictability via statistical downscaling (Hewitson and Crane 2006; Gao et al. 2017; Farnham et al. 2018).

LSMPs typically refer to synoptic-scale meteorological variables that have an understandable physical relationship to and a primary influence on a specific phenomenon (e.g., extreme precipitation), including those characterizing primary circulation, thermodynamics, and water vapor attributes at surface level and different levels of atmosphere. LSMPs establish a favorable environment for triggering and/or enhancing mesoscale processes to promote the occurrence of the phenomenon. There exist a range of methods for identifying LSMPs associated with extremes, including composites (Milrad et al. 2014; Gao et al. 2014), regression, and empirical orthogonal function (EOF) or principal component analysis (PCA; Reusch et al. 2005; Jewson 2020), as well as automated pattern-extraction methods such as cluster analysis (Casola and Wallace 2007; Agel et al. 2018) and self-organizing maps (SOMs; Lennard and Hegerl 2015; Loikith et al. 2017). LSMPs have been employed to evaluate model fidelity in producing synoptic conditions associated with extreme precipitation and understand the physical mechanisms conducive to these events (DeAngelis et al. 2013; Kawazoe and Gutowski 2013), as well as to assess future changes in these conditions (Hope 2006; Lennard and Hegerl 2015).

Although physical causes of extreme precipitation have been well explored, its prediction remains a great challenge due to its infrequent and irregular occurrence as well as different types of weather systems involved. It is widely recognized that synoptic-scale forcing in general has greater predictability than small-scale forcing (Hohenegger and Schar 2007; Schumacher and Davis 2010). However, to what extent an extreme precipitation event is predictable based on LSMPs has not been sufficiently assessed. Lu et al. (2016) investigated the predictability of 30-day extreme precipitation occurrence using a logistic principal component regression on time-lagged sea surface temperature (SST) and sea level pressure (SLP) and further identified several regions across the world with potential...

DOI: 10.1175/JCLI-D-21-0137.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
forecasting skill. Li and Wang (2018) found significant skill in prediction of summer extreme precipitation days over eastern China using the stepwise regression models on large-scale lower boundary anomalies. Knighton et al. (2019) used a convolutional neutral network (CNN) to predict seasonal archetypes of regional precipitation and discharge extremes in the eastern United States based on a suite of synoptic-scale climate variables and found that all the employed variables yielded reliable predictions with some differences by season and region. Barlow et al. (2019) reviewed the current state of knowledge regarding LSMPs associated with short-duration extreme precipitation events over North America from the perspectives of meteorological systems, dynamical mechanisms, model representation, and climate change trends. They stated that most of the studies naturally focused on analyzing LSMPs occurring with extreme precipitation, with less emphasis on testing the causal nature of the identified relationships (i.e., examining to what degree the identified features are necessary and/or sufficient conditions for extreme precipitation).

In this study, we assess the prediction skill of regional extreme precipitation occurrence based on LSMPs at a daily scale. We focused on two regions of the United States in our previous studies (Gao et al. 2014; Gao et al. 2017), where extreme precipitation regimes presented distinct seasonality and atmospheric circulation patterns (Schlef et al. 2019). This current study differs from our most previous ones (Gao et al. 2014; Gao et al. 2017) in several ways. 1) Extreme precipitation (99th percentile) is analyzed instead of heavy precipitation (95th percentile). This leads to highly imbalanced dataset (dominance of nonextreme precipitation days) and thus pose a considerable challenge to train our classification predictive model. 2) We compare a relatively simple analogue method developed in our previous studies with a more sophisticated CNN approach to predict extreme precipitation occurrence. Both methods do not require making assumptions about the normality, linearity, or continuity of the data sample. 3) We examine a larger set of meteorological variables from different atmospheric levels to characterize the LSMPs. Our objective is to understand which features of the large-scale circulation are most relevant for predicting regional extreme events of our interest and how this varies by season and region. 4) The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2) is used to characterize LSMPs instead of MERRA. 5) Last, a different set of performance measures is employed. This work could provide a basis for evaluating climate models’ skill in prediction of historical and future extreme precipitation occurrence based on model-simulated LSMPs. The remainder of the paper is organized as follows. Section 2 presents the data used and study region. Section 3 describes two statistical methods employed to quantify the potential predictability of extreme precipitation occurrence. A quantitative evaluation of the prediction skill is presented in section 4 followed by a summary and discussion in section 5.

2. Data

Daily precipitation observations, spanning from 1948 to the present and confined to the continental United States land areas, are obtained from the NOAA Climate Prediction Center (CPC) unified rain gauge-based analysis (Higgins et al. 2000). These observations are gridded to a 0.25° × 0.25° resolution from three sources of station rain gauge reports using an optimal interpolation scheme. The analysis went through rigorous quality control procedures and was shown to be reliable for studies of fluctuations in daily precipitation (Higgins et al. 2007).

MERRA-2 provides data beginning in 1980 at a spatial resolution of 0.625° × 0.5° (Bosilovich et al. 2016). In comparison with the original MERRA dataset, MERRA-2 represents the advances made in both the Goddard Earth Observing System Model, version 5 (GEOS-5) (Molod et al. 2015) and the Global Statistical Interpolation (GSI) assimilation system that enable assimilation of modern hyperspectral radiance and microwave observations, along with GPS-Radio Occultation datasets. MERRA-2 is the first long-term global reanalysis to assimilate space-based observations of aerosols and represent their interactions with other physical processes in the climate system.

We assemble a set of daily meteorological variables at different levels from MERRA-2 to characterize the LSMPs (Table 1). These variables have been widely used for statistical downscaling of precipitation in various studies as summarized by Anandhi et al. (2008) and Sachindra et al. (2014). We do not include the commonly used SLP and geopotential height because Gao et al. (2017) showed that the overall increasing trend of geopotential height associated with climate warming makes the use of geopotential height anomalies problematic within the analogue approach for future climates. Variables at 850 hPa are also not examined due to regions of high orography.

We analyze two precipitation estimates from MERRA-2: 1) the precipitation generated within the cycling data assimilation system, hereinafter referred to as MERRA2_P [M2AGCM in Reichle et al. (2017)], and 2) the corrected precipitation that is seen by the land surface and that modulates aerosol wet deposition over land and ocean, hereinafter referred to as MERRA2_Pc [M2CORR in Reichle et al. (2017)]. The daily precipitation from observation and MERRA-2 as well as daily meteorological fields are all regridded to 2.5° × 2° resolution via area averaging as suggested by Chen and Knutson (2008). The use of 2.5° × 2° resolution is primarily for assessing climate models’ skill in prediction of extreme precipitation occurrence based on model-simulated LSMPs [i.e., from phases 5 and 6 of the Coupled Model Intercomparison Project (CMIP5 and CMIP6)], which is a topic for a future study. The overlap between the CPC observation (1948–present) and MERRA-2 (1980–present) is 1 January 1980–31 December 2019. The constructed statistical models for identifying the daily occurrence of extreme precipitation event are first trained with the data from 1980 to 2005 and then assessed with blind prediction using the data from 2006 to 2019. The end of the training period is chosen to be consistent with the end of the CMIP5 historical experiment (1850–2005). At each grid cell, we convert each meteorological variable of entire period (1980–2019) to its standardized anomaly, defined as the anomaly from its seasonal climatological mean divided by the standard deviation of the 26-yr training period (1980–2005). A precipitation event is a daily amount above 1 mm day−1 at one 2.5° × 2° observational or model grid. An extreme precipitation event is defined when
Table 1. Large-scale meteorological variables at different levels from MERRA-2 (left two columns) and two group of statistical schemes (right two columns) examined in this study. The variables with asterisks are used to construct different analogue schemes, while all the variables, separately or in combination, are assessed for CNN (see text for details).

<table>
<thead>
<tr>
<th>Variable (symbol)</th>
<th>Atmosphere levels</th>
<th>Analogue schemes</th>
<th>CNN schemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zonal wind speed ($u_{500}$, $u_{10m}$, $u_{2m}$)</td>
<td>500 hPa, 10 m, 2 m</td>
<td>($u_{500}$)$_{500}$tpw</td>
<td>Oversampling</td>
</tr>
<tr>
<td>Meridional wind speed ($v_{500}$, $v_{10m}$, $v_{2m}$)</td>
<td>500 hPa, 10 m, 2 m</td>
<td>($u_{500}$)$_{500}$tpw</td>
<td>Individual variable (all)</td>
</tr>
<tr>
<td>Air temperature ($t_{500}$, $t_{10m}$, $t_{2m}$)</td>
<td>500 hPa, 10 m, 2 m</td>
<td>($uq$)$_{w500}$tpw</td>
<td>($uq$)$_{w500}$tpw</td>
</tr>
<tr>
<td>Specific humidity ($q_{500}$, $q_{10m}$, $q_{2m}$)</td>
<td>500 hPa, 10 m, 2 m</td>
<td>($uq$)$_{w500}$tpw</td>
<td>($uq$)$_{w500}$tpw</td>
</tr>
<tr>
<td>Relative humidity ($rh_{500}$, $rh_{2m}$)</td>
<td>500 hPa, 700 hPa</td>
<td>($uq$)$_{w500}$tpw</td>
<td>($uq$)$_{w500}$tpw</td>
</tr>
<tr>
<td>Vertical pressure velocity ($w_{500}$*)</td>
<td>500 hPa</td>
<td>($uq$)$_{w500}$tpw</td>
<td>($uq$)$_{w500}$tpw</td>
</tr>
<tr>
<td>Precipitable water (tpw*)</td>
<td>Total column</td>
<td>No-balance</td>
<td>Individual variable (all)</td>
</tr>
<tr>
<td>Vertically integrated zonal moisture flux ($uq$*)</td>
<td>Total column</td>
<td>No-balance</td>
<td>Individual variable (all)</td>
</tr>
<tr>
<td>Vertically integrated meridional moisture flux ($uq$*)</td>
<td>Total column</td>
<td>No-balance</td>
<td>Individual variable (all)</td>
</tr>
<tr>
<td>Dewpoint temperature ($tdew_{2m}$)</td>
<td>2 m</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

the daily amount at any grid cell exceeds its 99th percentile, which is derived from the cumulative distribution of all the observed precipitation events at this grid cell across a particular season of the entire training period. We then pool such events at all grid cells within the regions of our interest for the observation and MERRA-2.

Our analyses focus on the same two regions in our previous studies (Gao et al. 2014; Gao et al. 2017) where extreme precipitation regimes present distinct seasonality and circulation patterns: the “Pacific Coast California” (PCCA) region (33°–41°N, 123.75°–118.75°W at 2.5° × 2° resolution) in winter [December–February (DJF)] and the midwestern United States (MWST) (39°–45°N, 98.75°–88.75°W) in summer [June–August (JJA)]. The extreme winter precipitation along the west coast in association with atmospheric rivers (ARs) have been widely studied (Ralph et al. 2006; Leung and Qian 2009; Lamjiri et al. 2017; Gershunov et al. 2019) and it was shown that ARs can be used to skillfully predict the occurrence of extreme precipitation events at a daily scale (Chen et al. 2018). However, it is well recognized that the forecast skill of summertime precipitation variability is characterized weak, attributable to deficiencies in small-scale cumulus convection parameterization, which plays a larger role in summer than in winter when synoptically driven systems dominate (Sukovich et al. 2014; Wehner et al. 2014). In particular, Bosilovich (2013) found that the Midwest is one of the most poorly represented regions over the United States, with either false extremes or underrepresentation of extreme events by three reanalyses examined, mainly due to the increased dependence of summer precipitation in this region on the boundary layer parameterization. Therefore, our analysis of these two regions (and seasons) based on different statistical methods could provide us a general insight into the predictability limit of daily extreme precipitation occurrence based on LSMPs.

3. Methodology

a. Analogue method

The analogue method employs “composites” to identify prevailing LSMPs associated with the observed extreme precipitation events at a local scale, through the joint analyses of precipitation-gauge observations and atmospheric reanalysis. Our previous studies (Gao et al. 2017; Gao and Schlosser 2019) evaluated two analogue schemes [(($u_{500}$)$_{500}$tpw and ($u_{500}$)$_{500}$tpw)] based on 500-hPa horizontal and vertical winds ($u_{500}$) and each of two moisture variables, namely, near-surface specific humidity ($q_{2m}$) and total-column precipitable water (tpw). Here we examine two additional analogue schemes [(($uq$)$_{w500}$tpw and ($uq$)$_{w500}$tpw)] that are constructed with moisture flux ($uq$) and its place in ($u_{500}$) and ($v_{500}$) respectively (Table 1). There may exist some degree of collinearity between variables used in these two new schemes. Ideally, the variables selected for construction of any prediction model should generally be independent and a relatively small number of variables should be used in order to avoid problems with overfitting and collinearity. However, we still test these schemes in order to understand the trade-off between extra information added by these “new” variables ($uq$ and $q$) and their collinearity with other moisture variables ($q_{2m}$ or tpw) and how prediction skill will be affected.

We follow the same procedure as described in Gao et al. (2017) to calibrate the analogue schemes and will briefly state it here. Two metrics, the “hotspot” metric and the spatial anomaly correlation coefficient (SACC), are employed to characterize the degree of consistency between daily MERRA-2 LSMPs and the composites. The hotspot metric diagnoses the extent to which each atmospheric variable of the composite represents that of identified individual event. It involves the calculation of sign count at each grid cell for each variable by recording the number of identified events when standardized anomaly of each variable has a consistent sign with its respective composite. Hotspots are identified as the grid cells where the events used to construct the composites exhibit strong sign consistency with the composite (i.e., the larger sign counts). SACC is calculated between the daily MERRA-2 LSMPs and the corresponding composites over the region that captures the coherent structures of the composites. Ten SACC thresholds are assessed from 0.0 to 1.0 with an interval of 0.1.

We experiment selections of different number of variables (out of four variables in each analogue scheme) that have consistent signs with the corresponding composites over the selected hotspot grid cells and have SACC larger than the designated thresholds. Theoretically, there are 16 selections in total.
Hereinafter we use the term “multivariate condition” to refer to any of such selections and “case” to refer to any analogue scheme under any multivariate condition. During the training phase, we perform automatic calibration to simultaneously determine the optimal cut-off values for the number of hotspots and SACC of all relevant variables for each case. The procedure is conducted by running different combinations of the number of hotspots and SACC thresholds across all relevant variables in each case. The combination that produces the observed number of extreme precipitation events with the best Gilbert skill score (GSS; described later) will be denoted as the “criteria of detection” for that case. During the blind prediction (validation) period (2006–19), daily MERRA-2 LSMPs will be evaluated against the criteria of detection established in the training phase. Any day that meets these criteria of detection is considered as an extreme precipitation event. In other words, we use the occurrence of composite-type LSMPs to predict the occurrence of extreme precipitation events.

b. CNN

CNN is a class of deep neural networks and commonly applied to image-related recognition, classification, and analysis. A major advantage of CNN is that it requires little a priori knowledge of the underlying data structure and input–output relationships, and enables assembling more complex, hierarchical patterns in high-dimensional space using smaller and simpler patterns. Like other neural networks (NNs), CNNs can theoretically approximate any function of input–output relationships to an arbitrary degree of precision, although sometimes at the expense of interpretability of such relationships. In climate science, CNNs have been shown as a powerful tool to predict clustered weather patterns (Chattopadhyay et al. 2020) and identify various extreme weather events in large climate datasets, including tropical depressions, tropical cyclones, extratropical cyclones, weather fronts, and atmospheric rivers (Liu et al. 2016; Racah et al. 2017). In this study, we use a CNN to explore the potential predictability of extreme precipitation occurrence (binary classification of an extreme versus a nonextreme event) as compared to a relatively simple analogue method.

Machine learning algorithms for classification are usually designed to perform well when the number of samples in each class is about equal. Extreme weather event prediction in our case is an imbalanced classification where the distribution of samples across the classes is biased or skewed. Imbalanced classification poses a challenge because it often causes models developed using conventional machine learning algorithms to have poor predictive performance, specifically for the minority class. This is a problem because the minority class is typically more important than the majority class. It has been shown that class imbalance can affect both convergence during the training phase and generalization of a predictive model on the test dataset for traditional classifiers (Japkowicz and Stephen 2002; Mazurowski et al. 2008).

In this study, we employ oversampling, the method most commonly applied in deep learning to address class imbalance (Buda et al. 2018). Oversampling simply replicates randomly selected samples from the minority class (extreme precipitation days in our case) to achieve more balanced training data (the test data are left untouched). The model is trained batch-wise with each batch (~200 samples) of the oversampled dataset maintaining the same ratio of extreme to nonextreme event days. There is no simple rule of thumb for an optimal oversampling ratio. We have experimented different ratios and select 1:4 and 2:3 for PCCA and MWST, respectively, which give the best Gilbert skill scores (GSS; described in section 3c) during both calibration and validation periods. We also compare classification performance of a CNN trained based on the oversampled (hereinafter referred to as “oversampling”) and the original imbalanced dataset (hereinafter referred to as “no-balance”) to examine the effectiveness of oversampling.

The architecture of our CNN algorithm is shown in Fig. 1. We implement a 2D CNN within Keras (https://keras.io/), which uses an input layer and a series of intermediate layers to produce an output of binary classification (an output layer). Each synoptic-scale atmospheric field is extracted over the spatial domain of 172.5°–90°W, 8°–66°N for PCCA (29 × 33 grids) and 120°–72.5°W, and 18°–58°N for MWST (20 × 19 grids) and applied, individually or in combination, as an input layer. The intermediate layers consist of a set of convolutional layers, Rectified Linear Unit (ReLU) layers, and max pooling layers as well as one flatten layer. The convolutional layer serves as a feature detector (also referred to as a “filter” or a “kernel”) over the previous layer (i.e., the input layer for the first convolutional layer) and produces feature maps that provide an insight into where a certain feature is found. This is done by sliding the filter over the layer received as input and computing the dot product (or “convolution filter”). The higher the value is in a feature map, the more the corresponding place resembles the feature. In the ReLU layer, the ReLU activation function, \( f(x) = \max(0, x) \), is applied to the feature maps to introduce nonlinearity and prevent saturation at larger positive inputs. The pooling layer is often placed between two layers of convolution and applied to reduce the dimensions of feature maps generated by a convolutional layer while preserving the most important characteristics of each feature. This is achieved by cutting the feature map into regular cells and keeping the maximum value within each cell. The pooling layer improves the computational efficiency of the network by reducing the number of parameters to learn, controls overfitting, and achieves translational and scale invariance. There are usually several rounds of convolution and pooling: feature maps are filtered with new kernels, new feature maps are further resized and filtered again, and so on. The flatten layer converts the last feature maps into a vector. The output layer applies weights to the input vector via matrix multiplication, passes it through an activation function (sigmoid function in our case), and produces a new output vector (size 1 in our case). The element of the new output vector is the probability of the minority class (extreme event) between 0 and 1. A threshold value of 0.5 is used to label the predicted class, above which an extreme event is considered to occur and below which not to occur.

Our CNN is composed of two convolutional layers and two max pooling layers. Each convolutional layer use sixteen \( 3 \times 3 \) filters and a step of 1. Each max pooling layer uses a \( 2 \times 2 \) square cell and a step of 2. We apply L2 regularization
with a weight of 0.01 and 0.001 to the first and second convolutional layer, respectively. In addition, early stopping is employed to prevent the network from overfitting. The network is evaluated with the Adam optimizer using binary cross-entropy as the objective function and the network weights are learned by backpropagation. We assess one meteorological variable at a time to understand the relative importance of each for prediction of extreme precipitation occurrence. We also explore some combinations of meteorological variables to see if any additional predictive power may be added. These combinations are tested largely based on the knowledge learned from our experience with analogue schemes. All the "oversampling" schemes, based on an individual or combined meteorological variables (Table 1), are trained with the same set of hyperparameters described above. All the "no-balance" schemes follow the similar CNN structure but are trained with a different set of hyperparameters.

**c. Measures of prediction skill**

We compare the occurrence of extreme precipitation events estimated from various analogue and CNN schemes with that identified from the observation and two MERRA-2 precipitation products at 2.5° × 2° resolution during both calibration and blind prediction periods. Several performance measures are adopted that are used extensively by the National Weather Service for deterministic categorical forecast evaluation (Table 2). The hit rate (H) measures the fraction of observed events that is correctly predicted and is sensitive only to missed events. The false alarm ratio (FAR) measures fraction of predicted events that actually did not occur (i.e., were false alarms) and is sensitive only to false alarms. The threat score (TS) measures the fraction of observed and/or forecast events that were correctly predicted and is sensitive to both missed events and false alarms. Frequency bias (B) measures the relative frequency of forecast events to observed events and indicates whether the forecast system is unbiased (B = 1) or has a tendency to underforecast (B < 1) or overforecast (B > 1) events. Skill scores (SSs) use a single value to summarize forecast accuracy relative to a reference forecast and essentially represents fractional improvement over the reference forecast. SS values larger (smaller) than 0 indicate more (less) skillful than reference forecast, with higher SS values denoting more skillful predictions. Two SSs, the Heidke SS (HSS) and Gilbert SS [GSS, or equitable threat score (ETS)], are examined in our study. The HSS measures the fraction of correct forecasts (events and nonevents; a + d in Table 2) after eliminating those forecasts that would be correct due purely to random chance, while GSS measures the fraction of observed and/or forecast events (a + b + c in Table 2) that were correctly predicted, adjusted for hits associated with random chance. The reference forecasts for the HSS and GSS are the proportion correct and the TS expected by random chance, respectively.
Symmetric extremal Gilbert skill score (GSS)

False alarm ratio (FAR)

Threat score (TS)

Frequency bias (B)

Gilbert skill score (GSS)

Heidke skill score (HSS)

Symmetric extremal dependence index (SEDI)

We also evaluate a quite new score, the symmetric extremal dependence index (SEDI), which is designed for forecast verification of extreme binary events (Ferro and Stephenson 2011; North et al. 2013).

There are many other metrics used in weather forecast verification and implementing all of them is not feasible. Various metrics provide complementary assessments of different aspects of forecast quality. In our case, the H, FAR, and TS all measure “accuracy” (the level of agreement between forecasts and observations) in slightly different ways, while GSS, HSS, and SEDI measure “skill” (the accuracy relative to a reference forecast). In the following analyses, we do not present the results of HSS and SEDI. They show similar behaviors to GSS, but with higher magnitudes. Interested readers can refer to Gao and Mathur (2021) for more details. GSS is retained because it might be the one used most frequently among a variety of performance measures to evaluate skill of deterministic precipitation forecasts (Wang 2014; Boluwade et al. 2017; Chen et al. 2018). We also assess the performance of different statistical schemes in depicting the interannual variability of extreme precipitation occurrence against the observation with temporal correlation (CORR) and the root-mean-square error (RMSE).

4. Results

a. Precipitation characteristics

Figure 2 compares the 99th percentile daily precipitation and mean daily precipitation intensity at 0.25° grid for DJF and JJA based on the observation of 1980–2005. The immediately evident is the strong seasonality exhibited by these precipitation quantities over two study regions. In the winter season, the coastal mountain ranges in the western United States receives a large amount of precipitation, with the 99th percentile possibly reaching up to 130 mm day\(^{-1}\). Little precipitation falls in summer with the mean precipitation generally less than 6 mm day\(^{-1}\) and extreme precipitation less than 30 mm day\(^{-1}\) over much of the study region. In the upper U.S. Midwest, summer is the wettest season. In addition, more heavy rainstorms occur in summer than in any other season, and the least number occur in winter (Huff and Angel 1992; Gao et al. 2014). The 99th percentile generally ranges from 50 to 80 mm day\(^{-1}\) in summer and from 10 to 30 mm day\(^{-1}\) in winter across the region. Note that the south central United States is active in terms of rainfall and extreme precipitation during both seasons. Regardless of the regions (seasons), the 99th percentile or extreme precipitation is about 4 times higher than the mean precipitation intensity. At the 2.5° × 2° grid our analysis is performed on, the magnitude of extreme precipitation (99th percentile) is systematically underestimated by a factor of 2 (not shown). However, its large-scale patterns differ among different regions and seasons is well preserved. The specific resolution and associated criteria chosen to estimate the observed extreme precipitation events represent one source of uncertainty but will not be discussed in this study.

b. Composites for analogue schemes

We extract 41 and 163 extreme precipitation events from the observation of 1980–2005 at 2.5° grid for the winter of PCCA
and summer of MWST, respectively. Figure 3 shows various composite synoptic atmospheric conditions as standardized anomalies for two regions, produced by averaging the MERRA-2 Reanalysis across the observed event days. PCCA is a region where both large-scale circulations and orographic enhancement play important roles in the generation of extreme precipitation. LSMPs are dominated by a cutoff low to the west-northwest and a ridge to the southwest of the study region, promoting strong southwesterly flow of moist air from central Pacific toward the west coast of the United States (Fig. 3a). The region is also characterized by stronger large-scale upward motion and higher amount of water vapor (Fig. 3b). As expected, composite anomalies of synoptic fields are weaker in summer (MWST) than in winter (PCCA). Nevertheless, an anomalous trough to the west and a ridge to the east of the study region is evident (Fig. 3c). A key ingredient for heavy precipitation in the region is strong southerly winds and sustained advection of warm air and low-level moisture from the tropical Atlantic Ocean, through the Caribbean Sea, turning northward through the Gulf of Mexico, and then northeastward into the U.S. Midwest (Fig. 3c). This fetch of Caribbean moisture links into the Great Plains low-level jet (Dirmeyer and Kinter 2010), creating ARs similar to those associated with the western United States. The synoptic patterns promote the development of strong upward motion and positive precipitable water anomalies centered over the study region (Fig. 3d) as well as enhanced moisture flux around the periphery of the subtropical high. These elements intersect a quasi-stationary baroclinic zone and support the development of frequent mesoscale convective systems.

c. Prediction skill of analogue schemes

Different multivariate conditions indicate that no particular one will lead to consistently best skill scores across all the analogue schemes during both calibration and validation periods. Nevertheless, the differences in all the skill scores among various multivariate conditions are really small, mostly on an order of a hundredth. Here we only present the results based on the multivariate condition that gives the best GSS during the calibration period.

1) PCCA

Figure 4 shows performance measures of two MERRA-2 precipitation products and four analogue schemes for DJF of PCCA during the calibration and validation periods. GSS is typically analyzed in conjunction with the bias because higher scores can be achieved by increasing the bias above unity. During the calibration period, MERRA2_P strongly overforecasts the number of extreme precipitation events by approximately 110% ($B = 2.1$), while MERRA2_Pc significantly reduces the bias with a slight overforecast by approximately 10% ($B = 1.1$). All the analogue schemes are deliberately calibrated to be unbiased ($B = 1$). As a result, MERRA2_P presents the highest $H$ (0.68), but at the expense of the highest FAR (0.56).
as well. However, the expected benefit of an elevated bias in MERRA2_P is not reflected in other measures, with the lowest GSS (0.27) among all the schemes. This is likely attributed to the trade-off between high $H$ and high FAR. How much will a GSS score be affected by a bias is not entirely clear: what value of the bias will help or hurt the score? Nevertheless, an improvement in the skill of MERRA2_Pc over MERRA2_P is evident, with consistently higher GSS (0.35) but lower FAR (0.5). All the analogue schemes outperform MERRA2_Pc with higher $H$ (0.54–0.63), TS (0.37–0.46), and GSS (0.36–0.46), but lower FAR (0.37–0.46) values. Among the four analogue schemes, the new group constructed with $uq$ and $yq$ [($(uq)vq)_{w500}$tpw and $(uq)vq)_{w500q2m}$] is superior to that constructed with $u_{500}$ and $v_{500}$ [$(uw)_{500}$tpw and $(uw)_{500q2m}$] consistently across all the performance measures. In terms of the choice between two moisture variables, the analogue schemes based on tpw generally yield marginally better performances than those based on $q_{2m}$. Among a variety of performance measures, TS and GSS exhibit similar behavior except that GSS values are just slightly lower than those of TS due to the number expected correct by chance. The small differences between TS and GSS are not unexpected because the number correct by random guessing would be small for rare events (it is more difficult to randomly guess rare events than common events).

Performances are generally worse during the validation than the calibration period, in particular for all the analogue schemes, which exhibit large reductions (from −0.19 to −0.33) in $H$ and GSS but large increases in FAR (0.11–0.18). MERRA2_Pc is the only case with slight increase (~0.06) in GSS. This is expected because analogue schemes are evaluated based on an independent dataset from the dataset used for calibration, while MERRA2 precipitation products adopt the same assimilation system to ingest observations during both periods (their performance should be independent of periods). MERRA2_P

**FIG. 3.** Composite standardized anomalies of (a) 500-hPa geopotential height ($h_{500}$; shaded) and the vertical integrated moisture flux vector ($uq$, $vq$; arrow), (b) 500-hPa vertical velocity ($w_{500}$; contour) and total precipitable water (tpw; shaded) for the Pacific Coast California (PCCA) in DJF based on 41 extreme precipitation events identified from the precipitation observation of 1980–2005 at 2.5° × 2° grid. (c),(d) As in (a) and (b), but for the midwestern United States (MWST) in JJA based on 163 extreme precipitation events. The red rectangles denote our study regions: 15 (only 8 valid grids) and 20 grids for PCCA and MWST, respectively.
slightly overpredicts the occurrence of extreme precipitation events by approximately 10%, while MERRA2_Pc and all the analogue schemes have a tendency to underpredict the occurrence to various extents (30% for MERRA2_Pc and 20%–45% for analogue schemes). It is worth noting that the decreases in bias (relative to the calibration period) lead to very different outcomes—we see a large decrease in \(H\) but small decrease in FAR for MERRA2_P, and a small decrease in \(H\) but a large decrease in FAR for MERRA2_Pc, as well as large decreases in \(H\) and large increases in FAR for all the analogue schemes. An advantage of overforecasting (the bias beyond one) in MERRA2_P is not seen in its performance metrics. MERRA2_Pc significantly improves the skill over MERRA2_P with much higher \(H\) (0.5 vs 0.38) and GSS (0.41 vs 0.21) but lower FAR (0.20 vs 0.65) values. MERRA2_Pc also outperforms all the analogue schemes in terms of all the performance measures. The best skill of MERRA2_Pc is consistent with the changes in \(H\) and FAR values described above. The performances of analogue schemes are mixed. The group of \(uq\) and \(vq\) schemes performs better than MERRA2_P consistently across all the performance measures, while the group of \(uv\) 500hPa and \(uv\) 850hPa performs consistently worse than MERRA2_P. Because all the analogue schemes are calibrated to be unbiased for blind prediction, the improvement in skill of the \((uq)(vq)\) group over the \((uv)850\) group can be considered genuine. Chen et al. (2018) showed the use of ARs to predict the occurrence of extreme precipitation events in western U.S. watersheds at a daily scale, with GSS values of 0.05–0.2 based on different AR-tracking algorithms. Our analyses indicate GSS values of 0.25 and 0.17 for the \((uq)(vq)\) and \((uv)\) 500hPa analogue groups, respectively, which is in agreement with Chen et al. (2018). The validation results further suggest that it is not obvious how much and in what way a bias will affect various performance measures and there is no simple and direct interpretation.

Figure 5 presents the performances of various analogue schemes in depicting the interannual variations of PCCA winter extreme precipitation frequency from 1980 to 2005 (calibration) and from 2006 to 2019 (validation) as compared to the observations and MERRA2 precipitation. The number of extreme precipitation events for each “year” is computed based on the numbers in December of the current year and the numbers in January and February of the subsequent year (thus, the numbers in January and February of 1980 and in December of 2019 are not included). The year is labeled based on December of that year. All the schemes reproduce the observed interannual variations of winter extreme precipitation frequencies reasonably well with the temporal correlations generally above 0.65 and significant at the 0.01 level during both periods. The MERRA2_Pc performs best with the highest correlations (>0.85) but the lowest root-mean-square errors (RMSEs) of ~1 day, while the MERRA2_P performs worst with RMSEs of 2.4 and 1.7 days for two periods, respectively.
There is no particular analogue scheme (or group of schemes) demonstrating consistently superior performance across the two periods, with RMSEs ranging from 1.3 to 1.8 days. We do not see an apparent performance degradation in the validation as opposed to the calibration period, particularly in terms of correlations, which are surprisingly much higher. Also evident is that MERRA2_P tends to strongly overestimate the observed number of extreme precipitation events in most years, but does capture well big peaks of 2005 and 2016. MERRA2_Pc closely adheres to the observed year-to-year variations except for a slight overestimation in 1982. All the analogue schemes capture the largest peak in 2016, but strongly underestimate the second largest peak in 2005 and overestimate the peak in 1994 and 1996 to various extents. In addition, they are able to depict the conditions where no extreme precipitation event is observed (zero event).

2) MWST

Immediately evident are poorer performances in MWST than in PCCA during both periods, in particular for the analogue schemes and MERRA2_Pc (Fig. 6). There are large differences in bias. MERRA2_P significantly overforecasts the number of extreme precipitation events by 110% ($B = 2.1$) during the calibration and 70% ($B = 1.7$) during the validation period, while MERRA2_Pc identifies only a little fewer than half as often as the events occur during both periods (76 forecasts versus 163 occurrences and 67 forecasts versus 140 occurrences, respectively). Various analogue schemes also tend to underpredict the event frequency by approximately 30%–45% during the validation period. During the calibration period, the highest $H$ ($-0.6$) is achieved by MERRA2_P at the cost of the highest FAR ($-0.7$). In contrast, the lowest $H$ ($-0.3$) of MERRA2_Pc corresponds to its lowest FAR ($-0.4$). Analogue schemes present moderate $H$ ($-0.4$) and FAR ($-0.6$). However, there is little difference in the performances of all the schemes in terms of TS ($-0.25$) and GSS ($-0.2$) values. During the validation period, there is an apparent degradation in the performances of analogue schemes as compared to the calibration period, with much lower $H$ ($-0.2$), TS ($-0.15$), and GSS ($-0.1$) values, but higher FAR ($-0.65$). We also see that the group of $(uq)(uq)500tpw$ and $(uq)(uq)2m$ are not significant at the 0.05 level. Correlation of $(uq)(uq)500tpw$ is significant at the 0.05 level (not the 0.01 level), while correlation of $(uq)(uq)5002m$ is significant at the 0.01 level.

Figure 7 shows the interannual variations of MWST summer extreme precipitation frequency from 1980 to 2019 estimated by observation, MERRA2 precipitation, and analogue schemes. In terms of tracking the observed year-to-year variations of extreme events, the correlations of all the schemes are significant.
at the 0.01 level during the calibration period. During the validation period, only the correlations of MERRA2 precipitation and $(uq)(wq)_{500p2m}$ are significant at the 0.01 level and that of $(uq)(wq)_{500tpw}$ significant at the 0.05 level. The group of $(wq)_{500}$ analogue schemes does not capture well these temporal variations with low correlations (~0.35). However, the RMSEs of various schemes are approximately 1–4 times larger than those in PCCA. MERRA2_P has the highest RMSEs of about 7.4 days and 8.2 days for the calibration and validation periods, respectively. The analogue schemes have the overall lowest RMSEs of 2.7–3.3 days and 5.4–6.3 days for two periods, respectively. RMSEs of MERRA2_Pc are 4.5 days for the calibration and 5.9 days for the validation, which are more than quadruple and quintuple those in PCCA, respectively. There is an apparent performance degradation in terms of RMSEs for all the schemes during the validation (as compared to the calibration). We see that MERRA2_P strongly overestimates the observed number of extreme precipitation events in nearly all the years, while MERRA2_Pc persistently underestimates the event frequency, particularly for the major floods (e.g., frequency per year $\geq 10$). MERRA2_Pc predicts only one-third as often as the events occurred in 1993 and 2008 historical floods, about half in 2010 and 2014–2017, and 1 out of 10 observed events in 2019. The larger RMSE of MERRA2_Pc during the validation period is likely attributed to major floods occurring more often (118 events in nine major floods) than during the calibration period (34 events in two major floods). The contrasting features between two MERRA2 precipitation products are well consistent with their frequency biases. Analogue schemes capture several major floods reasonably well (e.g., 1990, 1993, 2008, and 2010), but with their frequency biases. Analogue schemes capture several major floods reasonably well (e.g., 1990, 1993, 2008, and 2010), but significantly underestimate those from 2014 to 2019, which is the main cause for the large increases in RMSEs during the validation period. This constant underestimation is likely attributed to the lack of enough major flood events in the calibration period to adequately train analogue schemes for capturing their complete characteristics. Another possibility is the slight shift in the relevant features of LSMPs associated with extreme precipitation events between two periods, which the calibrated analogue schemes fail to capture. We examine the composites of LSMPs from 1980 to 2005, from 2006 to 2019, and from 2014 to 2019 (not shown) and find that the western ridge and moisture transport into the study region have slightly displaced eastward in 2006–19 and even further in 2014–19 as compared to 1980–2005. The centers of maximum anomalies of total precipitable water and upward motion have also shifted southeastward slightly. However, the exact reason for this constant underestimation is worthy of further study.

d. Prediction skill of CNN schemes

1) Oversampling in PCCA

Figure 8 shows performance measures of different CNN schemes trained with the oversampled dataset for PCCA during both periods. Because machine learning aims at developing algorithms that can automatically make accurate predictions, we present CNN schemes in descending order based on their GSS values in the validation period. Nearly all the CNN schemes tend to overpredict the frequency of extreme occurrence ($B > 1$), in particular during the calibration period. One probable consequence of overprediction (positive frequency bias relative to $B = 1$) is to increase the $H$, but the FAR might also rise. There are large differences in the frequency bias from different schemes, ranging from 1 to 3 in the calibration and from 0.5 to 1.7 in the validation, respectively. Only 5 out of 24 schemes indicate underprediction during the validation period. Such a large difference poses a challenge in making a fair comparison of various schemes’ performances. TS and GSS follow a similar pattern of variations with little difference in their magnitudes (hits associated with random chance are negligible). Among different schemes, three single-variate ($rh_{500}$, $q_{2m}$, and $tdw_{2m}$) and four multivariate schemes [except $(uq)(wq)_{500tpw}$] demonstrate relatively better performances in the calibration with GSS values ranging from 0.46 to 0.6. These schemes generally have relatively higher $H$ (>0.83) but lower FAR (<0.5) values, except for $rh_{500}$, which has moderate $H$ (0.76) but the lowest FAR (0.25). The differences in the GSS values of the remaining 17 schemes are not large, ranging from 0.26 to 0.38. The overall performances of CNN schemes are comparable to that of MERRA2_Pc, with half of the schemes having higher GSS values.

We find strong inconsistencies in some schemes’ performances (based on GSS) between two periods. For example, $rh_{500}$ performs best in the calibration but the worst in the validation. The ranking of $q_{2m}$ also drops dramatically from 2 to 17. The opposite trend is seen for $uq$ and $twp$, whose rankings are 18 and 24 in the calibration, but 1 and 7 in the validation, respectively. Such inconsistencies are also observed for the analogue schemes, but the exact reason is unknown. A majority

![Figure 8](image-url)
of schemes exhibit consistently poor performances (ranking in the bottom 15) between two periods, including the horizontal wind speeds, temperature, and specific humidity at all levels (except for \(v_{500}\) and \(q_{2m}\) and \(r_{h700}\)). Several schemes demonstrate consistently good performances, such as \(uq\), \(tdew_{2m}\), \(w_{500}\), and four multivariate schemes [except \((uq)(w)q_{500}tpw\)]. In addition, multivariate schemes do not necessarily outperform single-variate schemes, but their overall performances are robust. Regardless of the scheme, all the performance measures degrade remarkably during the validation period with lower \(H\), TS, and GSS values but higher FAR values. The \(H\) values range from 0.09 to 0.62, while the FAR values vary between 0.55 and 0.81. During the calibration period, however, \(H\) values never drop below 0.54, while the FAR values rarely exceed 0.7. Except for the \(r_{h700}\), the variations in GSS values are fairly small, ranging from 0.32 (\(uq\)) to 0.13 (\(r_{h700}\)) across 23 schemes. The decreasing GSS values in the sequence of schemes correspond well to their overall decreasing \(H\) and increasing FAR values. In summary, MERRA2_Pc outperforms all the CNN schemes, while about half of the CNN schemes are superior to MERRA2_P.

Analogue schemes slightly outperform their CNN counterparts \([uq](w)q_{500}tpw\) and \((uq)(w)w_{500}q_{2m}\) with higher GSS values (0.256 vs 0.207 and 0.234 vs 0.228, respectively).

2) OVERSAMPLING IN MWST

Figure 9 shows performance measures of different CNN schemes trained with the oversampled dataset for the MWST. During the calibration period, all the CNN schemes predict extreme precipitation event more often than the observation with the frequency bias ranging from 1.4 to 2.6, except for \(u_{10m}\), which is nearly unbiased \((B \sim 1)\). During the validation period, a majority of CNN schemes (19) also tend to overpredict the event frequency, but to a lesser extent (maximal \(B\) is 1.8). The resulting \(H\) values are much lower in the validation period (0.21–0.52) than in the calibration period (0.31–0.88). In comparison with PCCA, a distinct difference is that the hits due to random chance are not negligible, particularly during the validation period. The resulting differences between TS and GSS values are around 0.04 and 0.06 for the calibration and validation periods, respectively. Two regions do share some common features, such as the same pattern of variations in TS and GSS values, and performance degradation in the validation as opposed to the calibration period. Overall, the performances of CNN schemes in MWST are poorer than in PCCA, with GSS values of 0.09–0.43 and 0.1–0.24 for two periods, respectively. PCCA has corresponding GSS values of 0.26–0.6 and 0.06–0.32 (0.13–0.32 if the lowest one is excluded), respectively. Such regional differences in the performances are consistent with the results of analogue schemes. Inconsistent performances between two periods are also observed, but for different variables from in PCCA. The schemes \(w_{500}\) and \(q_{500}\) show poor calibration performances (ranking 11 and 17 based on GSS) but good prediction skills (ranking 3 and 4), while the opposite occurs to \(q_{2m}\) and \((uq)(w)w_{500}q_{2m}\) with their rankings dropping from 2 to 15 and from 6 to 20 between the two periods, respectively. We see that schemes based on the horizontal wind speed, temperature, and relative humidity at all levels, \(tdew_{2m}\), and \(q_{10m}\) generally exhibit limited skills in detecting extreme precipitation events in both periods. In contrast, the skills of schemes based on vertically integrated variables and their combinations \([uq], (uq)(w)q_{500}tpw\) are fairly robust. In comparison with MERRA2 precipitation, more than half of the schemes (13) perform better than both precipitation products in terms of GSS values during the calibration period, but only one scheme, \(uq\), performs better than MERRA2_P and none better than MERRA2_Pc during the validation period. However, CNN schemes generally have higher GSS values than their analogue counterparts in both periods.

To summarize over both regions, an appropriate combination of multiple variables may help achieve a high predictive skill if overfitting can be prevented. However, it is not ensured that such a multivariate scheme will consistently outperform all the single-variate ones. Both regions share some common schemes that provide high prediction skills, including \(uq\), \(vq\), \((uq)(w)q_{500}tpw\), \(w_{500}\), and \(tpw\). The specific schemes are \(tdew_{2m}\) to PCCA and \(q_{500}\) to MWST, respectively. These interpretations are based solely on the GSS values during the validation period without taking into account the large differences in the frequency bias of each scheme. It is well known that forecasts with a larger bias tend to have a higher GSS, which complicates the direct comparison of various schemes’ performances. However, further examination indicates that these top-performing schemes have moderate frequency biases (1.1–1.4) among all, implying that their high prediction skills are not simply attributed to the overprediction artifact \((B > 1)\). The complication arises in particular when we compare the same CNN and analogue schemes in MWST, in which lower GSS values of analogue schemes are accompanied by their lower frequency biases \((B < 1)\). It would be preferable to remove the effect of bias in overprediction and underprediction with a performance measure corresponding to unit bias, or to compare competing schemes that have similar biases.
3) INTERANNUAL VARIABILITY

In this section, we only demonstrate the performances of those common variables described above in depicting the inter-annual variability of extreme event frequency (Figs. 10 and 11), with the relevant statistics of all the schemes summarized in Table 3. In PCCA, the selected CNN schemes depict the observed interannual variation of extreme event frequency fairly well with correlation coefficients larger than 0.6 and 0.85 in the calibration and validation periods, respectively (Fig. 10b). The correlations of all the CNN schemes are significant at the 0.01 level (Table 3), except for those of the meridional wind at all levels and $q_{10m}$ in the validation period which are only significant at the 0.05 level. However, CNN schemes tend to overestimate the frequency, with the RMSEs ranging from 1 to 3.5 days in the calibration and from 1 to 4.5 days in the validation (Table 3). Among the selected variables (Fig. 10a), we see that large peaks in 2005, 2016, and 2018 are successfully captured by several schemes, but the estimated frequencies in 1985 and 1996 by all the schemes are at least double or triple the observation. Persistent overestimation also occurs to the years with low frequency, including 1981–82, 1995, 1999, and 2007. The ($uq$)($vq$)$w_{500}$ outperforms all the others with the lowest RMSEs in both periods (1.58 and 1.3 days), which are comparable to its analogue counterpart (1.5 days), particularly in the calibration period. Notably, $uq$ performs the worst among all with the overall largest RMSEs in both periods (2.6 and 2.3 days). These results also suggest that the scheme with the best GSS value is not necessarily the one that performs best in depicting the interannual variability (i.e., lowest RMSE).

As expected, the performances of all the CNN schemes are poorer in MWST than in PCCA. During the validation period, only 7 out of 24 schemes have correlation coefficients significant at the 0.01 level and only 11 significant at the 0.05 level, also with negative correlations for $v_{2m}$ and $r_{500}$. This implies that most of the schemes fail to reproduce the observed interannual variability of summertime extreme event frequency in the validation period over the MWST. Among the selected schemes, only tpw, $w_{500}$, and ($uq$)($vq$)$w_{500}$tpw have correlations significant at the 0.01 level and ($uq$)($vq$)$w_{500}$2m significant at the 0.05 level. The RMSEs are much larger than in PCCA, ranging from 3 to 12 days and from 3 to 10 days for two periods, respectively. It is immediately evident that all the selected schemes constantly overestimate the extreme event frequency to various extents (Fig. 11a). The overestimation is particularly strong by four single-variate schemes during the calibration period, with the RMSEs ranging from 6.6 to 7.7 days. Overall, three multivariate schemes are superior to the single-variate schemes with lower RMSEs in both periods. In comparison with their analogue counterparts, two CNN schemes have larger RMSEs [4.6 vs 2.7 for ($uq$)($vq$)tpw and 3.5 versus 3.1 for ($uq$)($vq$) $w_{500}$tpw] in the calibration but lower RMSEs in the validation periods (3.3 and 4.3 vs 5.4). The larger RMSEs of analogue schemes in the validation period are likely attributed to their significant underestimation of extreme frequency from 2014 to 2019 (Fig. 7a).
4. NO-BALANCE

Table 3. Correlations and RMSEs of MERRA2 precipitation and CNN schemes trained with the oversampled dataset for PCCA and MWST during two periods. “Cal” and “Val” indicate calibration and validation, respectively. Normal font in the Correlation columns indicates that the correlations are significant at the 0.01 level. Bold font in the Correlation columns indicates that the correlations are not significant at the 0.05 level. A pound sign (#) in the Correlation column indicates that the correlations are significant at the 0.05 level, but not at the 0.01 level. An asterisk (*) in the RMSE column indicates that the RMSEs are better than MERRA2_Pc.

<table>
<thead>
<tr>
<th>PCCA</th>
<th>MERRA2_P</th>
<th>MERRA2_Pc</th>
<th>$v_{2m}$</th>
<th>$v_{10m}$</th>
<th>$v_{500}$</th>
<th>$v_{2m}$</th>
<th>$v_{10m}$</th>
<th>$v_{500}$</th>
<th>$v_{500}$</th>
<th>$\text{rd}_{2m}$</th>
<th>$\text{rd}_{500}$</th>
<th>$\text{tpw}$</th>
<th>$w_{500}$</th>
<th>$uq$</th>
<th>$(u)(w)$</th>
<th>$(u)(w)\text{tpw}$</th>
<th>$(u)(w)\text{w}_{500}\text{tpw}$</th>
<th>$(u)(w)\text{w}<em>{500}\text{f}</em>{2m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cal</td>
<td>0.68</td>
<td>0.85</td>
<td>0.76</td>
<td>0.76</td>
<td>0.61</td>
<td>0.79</td>
<td>0.75</td>
<td>0.62</td>
<td>0.78</td>
<td>0.69</td>
<td>0.71</td>
<td>0.65</td>
<td>0.77</td>
<td>0.78</td>
<td>0.79</td>
<td>0.82</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td>Val</td>
<td>0.83</td>
<td>0.95</td>
<td>0.76</td>
<td>0.79</td>
<td>0.64</td>
<td>0.77</td>
<td>0.77</td>
<td>0.85</td>
<td>0.78</td>
<td>0.69</td>
<td>0.72</td>
<td>0.76</td>
<td>0.81</td>
<td>0.8</td>
<td>0.8</td>
<td>0.83</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td>Cal</td>
<td>2.43</td>
<td>1.67</td>
<td>3.06</td>
<td>2.59</td>
<td>1.98</td>
<td>2.5</td>
<td>2.02</td>
<td>2.75</td>
<td>1.84</td>
<td>2.56</td>
<td>1.84</td>
<td>2.75</td>
<td>1.84</td>
<td>1.84</td>
<td>1.84</td>
<td>1.84</td>
<td>1.84</td>
<td>1.84</td>
</tr>
<tr>
<td>Val</td>
<td>1.71</td>
<td>1.07</td>
<td>4.52</td>
<td>2.73</td>
<td>1.73</td>
<td>2.35</td>
<td>1.98</td>
<td>1.41</td>
<td>1.64</td>
<td>2.27</td>
<td>1.64</td>
<td>1.41</td>
<td>1.64</td>
<td>1.64</td>
<td>1.64</td>
<td>1.64</td>
<td>1.64</td>
<td>1.64</td>
</tr>
<tr>
<td>Cal</td>
<td>0.73</td>
<td>0.71</td>
<td>0.69</td>
<td>0.77</td>
<td>0.78</td>
<td>0.84</td>
<td>0.68</td>
<td>0.87</td>
<td>0.78</td>
<td>0.81</td>
<td>0.77</td>
<td>0.84</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.79</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td>Val</td>
<td>0.71</td>
<td>0.75</td>
<td>1.12</td>
<td>0.84</td>
<td>0.95</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
<td>0.74</td>
<td>0.72</td>
<td>0.42</td>
<td>0.84</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>0.79</td>
<td>0.72</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table 3. Correlations and RMSEs of MERRA2 precipitation and CNN schemes trained with the oversampled dataset for PCCA and MWST during two periods. “Cal” and “Val” indicate calibration and validation, respectively. Normal font in the Correlation columns indicates that the correlations are significant at the 0.01 level. Bold font in the Correlation columns indicates that the correlations are not significant at the 0.05 level. A pound sign (#) in the Correlation column indicates that the correlations are significant at the 0.05 level, but not at the 0.01 level. An asterisk (*) in the RMSE column indicates that the RMSEs are better than MERRA2_Pc.

5. Summary and discussion

Prediction of extreme precipitation event has long been a challenge due to its infrequent and irregular occurrence as well as different types of weather systems involved. In this study, we use two machine learning approaches of different complexity to examine the LSMPs as predictors of extreme precipitation (99th percentile event) occurrence in two regions of the United States where extreme precipitation regimes exhibit distinct seasonality and circulation patterns, namely, the winter season.
of the “Pacific Coast California” (PCCA) region and the summer season of the midwestern United States (MWST). Our study demonstrates that LSMPs provide useful predictability of extreme precipitation occurrence at a daily scale. However, the prediction skill is strongly affected by the region/season, the choice of a meteorological variable or combination of variables, and the employed method.

Warm-season extreme precipitation in MWST presents lower prediction skill than cold-season extreme precipitation in PCCA, attributed to the weak synoptic-scale forcing in summer. The 14-yr (2006–19) independent forecast shows GSS values in PCCA range from 0.06 to 0.32 across 24 CNN schemes and from 0.16 to 0.26 across four analogue schemes. In MWST, GSS values range from 0.1 to 0.24 and from 0.1 to 0.14 across CNN and analogue schemes, respectively. All the analogue schemes and a majority of CNN schemes reproduce the observed interannual variations of extreme precipitation frequency reasonably well in PCCA with the temporal correlations significant at the 0.01 level and RMSEs less than 2 days. However, more than half of the analogue and CNN schemes in MWST have correlation coefficients not significant at the 0.05 level and RMSEs larger than 5.5 days. The prediction skills (in terms of GSS values) of both analogue and CNN schemes are generally worse than MERRA2_Pc in both regions, but occasionally better than MERRA2_P in PCCA.

Regardless of the regions (seasons) examined and methods employed, there is no single scheme that will consistently perform best in reproducing daily extreme precipitation occurrence and its interannual variation during both the calibration and validation periods. One notable finding is that vertically integrated variables generally provide higher prediction skill (in the validation) than those of a single level. Among the single-variate CNN schemes, $u_{w}q$ and $v_{w}q$ unanimously perform best in predicting daily extreme precipitation occurrence over both regions, followed by $w_{500}$ and $tpw$. Surface temperature and horizontal wind speeds, relative humidity at the lower ($700 \text{ hPa}$) and middle ($500 \text{ hPa}$) troposphere, and horizontal wind speeds at the middle troposphere usually offer relatively low prediction skill variably over both regions. The advantage of vertically integrated variables is also seen in the analogue schemes with the $(u_{w}q)(v_{w}q)$ group usually outperforming the $(w_{500})$ group over both regions. Previous studies have also documented integrated vapor transport (IVT; equivalent to the magnitude of “$u_{w}$” and “$v_{w}$”) as a very important ingredient for extreme precipitation production (Nakamura et al. 2013; Agel et al. 2018) and its high relevance in predicting ARs (Gao et al. 2015) and regional precipitation extremes (Knighton et al. 2019). Combination of multiple variables, even the colinear ones, seems to help improve prediction skill. Although it is not ensured that such multivariate schemes outperform the best single-variate one, their performances are fairly robust and generally superior to most single-variate schemes.

Our interpretation of prediction skill is largely based on the GSS values without taking into account the frequency bias of each scheme. There are large differences in the frequency bias of various schemes from different methods, which complicates the evaluation of their relative prediction performances. It is often recognized that the forecast with the larger bias (the wetter forecast) tends to have a higher GSS. However, it is not obvious how much and in what way GSS values are affected by frequency bias. In both regions, analogue schemes tend to underpredict the event. A majority of CNN schemes trained with the oversampled data significantly overpredict the event frequency, while all the CNN schemes trained with the original data (no-balance) strongly underpredict the event. However, the top-performing CNN schemes usually have moderately large frequency biases (1.2–1.4) among all, implying that their high prediction skills cannot be solely attributed to the over-prediction artifact.

Each of the two methods employed here has its own advantage and disadvantage. The analogue method is straightforward to implement. In particular, it can be calibrated to be unbiased ($B = 1$) before being used for the prediction and therefore prediction skills of different schemes are directly comparable. One limitation of analogue method is that only one distinct LSMP of each variable (composites) is considered to support the occurrence of extreme precipitation. The main advantage of CNN is that it enables capturing or learning location invariant features of different levels automatically. We see that several single-variate CNN schemes achieve more skillful prediction than the best multivariate analogue scheme in PCCA and more than half of CNN schemes in MWST. The power of CNN in pattern recognition is likely attributed to the specific design of its architecture. For example, the multiple convolutional layers, each learning multiple filters, allow for a hierarchical set of features to be extracted; translation and scale invariance achieved by max pooling layers enables the class label of the prediction not to change even if the image is translated/shifted by a small amount. However, there could
be a large difference in the frequency bias of various CNN schemes trained with the same dataset. CNN schemes may also be prone to overfitting or suffer from the issue of physical interpretability. Nevertheless, both methods involve a subjective selection of method-specific parameters, such as filter size and number of convolutional and max pooling layers for CNN, spatial domain to calculate spatial correlation and specific multivariate condition for analogue method. These choices could lead to marginal differences in the resulting skills and often there is no simple rule of thumb to obtain their optimal values. Nonstationarity of the relationships between the LSMPs and extreme precipitation occurrence has been a common major challenge as well.

Prediction of extreme precipitation event at a regional scale is of great significance due to its severe impacts on society. One immediate question is whether there exists any true skill in forecasting these rare events, and if so, how much. Our results suggest the answer is yes and quite a bit. It is possible that there is more predictability to be harvested from the LSMPs than we realized from this exercise. Our study shows that the CNN schemes trained with the oversampled data present much higher prediction skills than those with the original data. Therefore, oversampling strategy may be explored for the analogue method to see if the prediction skill could be further improved. For the CNN, its application is purely data driven and the constructed model may not be well generalizable beyond the data on which it is trained. This problem becomes worse when there is not enough training data (e.g., extreme weather prediction). Knowledge-guided machine learning (KGML) models that integrate scientific knowledge (explainable physical theories) and data synergistically would stand a better chance in safeguarding against such nongeneralizable features (Wang et al. 2017; Cao et al. 2018). It should be noted that the specific details of this exercise are almost certainly dependent on the choices of many elements, such as the definition of extreme precipitation events, study region, reanalysis data, predictors, and method-specific parameters. Few of the studies are directly comparable because these elements and employed approaches are quite varied. Nevertheless, the presented analyses demonstrate that machine learning techniques like CNN could serve as a promising tool to predict regional extreme precipitation occurrence. In principle, these techniques could be adapted to other types of extremes (i.e., heatwaves) under the supposition that large-scale atmospheric conditions play a role. Our future work will focus on applying the calibrated analogue and CNN schemes to climate model-simulated LSMPs (i.e., from CMIP6) for assessing models’ skill in prediction of historical extreme precipitation occurrence and its shift under climate warming.

Acknowledgments. This work was supported by the U.S. Department of Energy (DOE) under DE-FOA-0001968 and other government, industry, and foundation sponsors of the MIT Joint Program on the Science and Policy of Global Change. For a complete list of sponsors, see http://globalchange.mit.edu/sponsors/.

Data availability statement. All the data used in the study are freely available. The sources where the authors obtained the datasets are as follows. Precipitation observation data are from https://psl.noaa.gov/data/gridded/data.unified.daily.conus.html; MERRA2 reanalysis are from https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/.

REFERENCES


