Report 353
June 2021

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

Xiang Gao and Shray Mathur

MIT Joint Program on the Science and Policy of Global Change combines cutting-edge scientific research with independent policy analysis to provide a solid foundation for the public and private decisions needed to mitigate and adapt to unavoidable global environmental changes. Being data-driven, the Joint Program uses extensive Earth system and economic data and models to produce quantitative analysis and predictions of the risks of climate change and the challenges of limiting human influence on the environment—essential knowledge for the international dialogue toward a global response to climate change.

To this end, the Joint Program brings together an interdisciplinary group from two established MIT research centers: the **Center for Global Change Science (CGCS)** and the **Center for Energy and Environmental Policy Research (CEEPR)**. These two centers—along with collaborators from the Marine Biology Laboratory (MBL) at Woods Hole and short- and long-term visitors—provide the united vision needed to solve global challenges.

At the heart of much of the program's work lies MIT's Integrated Global System Model. Through this integrated model, the program seeks to discover new interactions among natural and human climate system components; objectively assess uncertainty in economic and climate projections; critically and quantitatively analyze environmental management and policy proposals; understand complex connections among the many forces that will shape our future; and improve methods to model, monitor and verify greenhouse gas emissions and climatic impacts.

This report is intended to communicate research results and improve public understanding of global environment and energy challenges, thereby contributing to informed debate about climate change and the economic and social implications of policy alternatives.

—Ronald G. Prinn,
Joint Program Director

MIT Joint Program on the Science and Policy of Global Change

Massachusetts Institute of Technology
77 Massachusetts Ave., E19-411
Cambridge MA 02139-4307 (USA)

T (617) 253-7492 **F** (617) 253-9845
globalchange@mit.edu
http://globalchange.mit.edu/

GLOBAL CHANGE

JUNE 2021

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

Xiang Gao[1,2] and Shray Mathur[3]

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 (PCCA) and the summer (JJA) of 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 extreme precipitation occurrence at a daily scale and its interannual variability over both regions. The 14-year (2006-2019) independent forecast shows Gilbert Skill Scores (GSS) in PCCA range from 0.06 to 0.32 across 24 CNN schemes and from 0.16 to 0.26 across 4 analogue schemes, in contrast to those from 0.1 to 0.24 and from 0.1 to 0.14 in MWST. Overall, CNN seems 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 multi-variate analogue scheme in PCCA and more than half of CNN schemes in MWST. Nevertheless, both methods highlight 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.

1 Corresponding author (Email: xgao304@mit.edu)

2 Joint Program on the Science and Policy of Global Change, Massachusetts Institute of Technology, MA, USA

3 Department of Computer Science, Birla Institute of Technology and Science, Pilani, India

1. Introduction

Extreme precipitation can lead to severe socio-economic 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 *et al.*, 2018). 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, 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 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 one (Hohenegger and Schar, 2007; Schumacher and Davis, 2010). However, to what extent an extreme precipitation event is predictable based on LSMPs is not 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 forecasting skill. Li and Wang (2017) 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 focus on two regions of the United States in our previous studies (Gao *et al.*, 2014; Gao *et al.*, 2017), where extreme precipitation regime presents its 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 that *1)* extreme precipitation (99^{th} percentile) is analyzed instead of heavy precipitation (95^{th} percentile). This leads to highly imbalanced dataset (dominance of non-extreme 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 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. and *5)* 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 summary and discussions in section 5.

2. Datasets

Daily precipitation observations, spanning from 1948 to 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° x 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 don't 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°x2° resolution via area averaging as suggested by Chen and Knutson (2008). 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 1980–2005 and then assessed

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 in bold are used to construct different analogue schemes, while all the variables, separately or in combination, are assessed for CNN (See text for details).

Variable (Symbol)	Atmosphere Levels	Analogue Schemes	CNN Schemes
Zonal wind speed ($\mathbf{u}_{500}$, u_{10m}, u_{2m})	500-hPa, 10-meter, 2-meter	$(uvw)_{500}tpw$	Oversampling
Meridional wind speed ($\mathbf{v}_{500}$, v_{10m}, v_{2m})	500-hPa, 10-meter, 2-meter	$(uvw)_{500}q_{2m}$	Individual variable (all)
Air temperature (t_{500}, t_{10m}, t_{2m})	500-hPa, 10-meter, 2-meter	$(uq)(vq)w_{500}tpw$	$(uq)(vq)$
Specific humidity (q_{500}, q_{10m}, $\mathbf{q}_{2m}$)	500-hPa, 10-meter, **2-meter**	$(uq)(vq)w_{500}q_{2m}$	$(uq)(vq)tpw$
Relative humidity (rh_{500}, rh_{700})	500-hPa, 700-hPa		$(uq)(vq)q_{2m}$
Vertical pressure velocity ($\mathbf{w}_{500}$)	500-hPa		$(uq)(vq)w_{500}tpw$
Precipitable water (**tpw**)	Total column		$(uq)(vq)w_{500}q_{2m}$
Vertically integrated zonal moisture flux (**uq**)	Total column		No-balance
Vertically integrated meridional moisture flux (**vq**)	Total column		Individual variable (all)
Dew-point Temperature ($tdew_{2m}$)	2-meter		

with blind prediction using the data from 2006–2019. The end of the training period is chosen to be consistent with the end of the Coupled Model Intercomparison Project Phase 5 (CMIP5) historical experiment (1850–2005) primarily for evaluation of climate models' prediction skill based on model-simulated LSMPs (not in this study). At each grid cell, we convert the meteorological fields of entire period (1980–2019) to normalized anomalies based on their respective seasonal climatological means and standard deviations of 26-yr training period (1980–2005). A precipitation event is a daily amount above 1 mm day^{-1} at one 2.5°x2° observational or model grid. An extreme precipitation event is defined when 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 and 123.75°–118.75°W at °x2° resolution) in winter season [December–February (DJF)] and the Midwestern United States (MWST) (39°–45° N sand 98.75° –88.75° W) in summer season [June–August (JJA)]. The extreme winter precipitation along the west coast in association with 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 characteristically 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 *et al.*, (2013) found that the Midwest is one of the poorly represented regions over the United States with either false extremes or underrepresentation of extreme events by three reanalysis 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. Methods

3.1 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 *et al.*, 2019) evaluated two analogue schemes ($uvw_{500}q_{2m}$ and $uvw_{500}tpw$) based on 500-hPa horizontal and vertical winds (uvw_{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 ($[(uq)(vq)w_{500}q_{2m}]$ and $[(uq)(vq)w_{500}tpw]$) that are constructed with moisture flux [(uq) and (vq)] in replace of 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 "vq") 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" 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 by recording the number of individual events whose standardized anomalies have consistent sign with the 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) which 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 a "multi-variate condition" to refer to any of such selections and use "case" to refer to any analogue scheme under any multi-variate 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 (2006–2019), daily MERRA-2 LSMPs will be evaluated against the "*criteria of detection*" established in the training phase. Any day that meets the "*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.

3.2 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 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 this study, we use a CNN to explore the potential predictability of extreme precipitation occurrence (binary classification of an extreme versus a non-extreme 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 a more balanced training data (the test data is left untouched). The model is trained batch-wise with each batch (~ 200 samples) of the oversampled dataset maintaining the same ratio of extreme to non-extreme 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 3.3) 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.

We implement a 2D CNN within Keras (Chollet, 2015), which uses an input layer and a series of hidden (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 and 8° ~ 66° N for PCCA and 120°~72.5° W and 18° ~ 58° N for MWST and applied, individually or in combination, as an input layer. The hidden layers consist of a set of convolutional and max pooling layers and 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 creates 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. 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 keep the maximum value within each cell. The pooling layer improves the efficiency of the network by reducing the number of parameters to learn and also avoids overfitting. There are usually several rounds of convolution and pooling: feature maps are filtered with new kernels, new feature maps is 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, pass through an activation function (logistic function in our case), and produces a new output vector (size 1 in our case). 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 3x3 filters and a step of 1 with the Rectified Linear Unit (ReLU) activation function. Each max pooling layer uses a 2x2 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, earlystopping is employed to prevent the network from overfitting. The network is evaluated with the Adam optimizer using binary cross-entropy as the objective function. 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.

3.3 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 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. Threat score (*TS*) measures the fraction of observed and/or forecast events that were correctly predicted. The TS is more complete than the H and FAR because it 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 has a tendency to underforecast or overforecast events. The forecast system is unbiased if B is equal to 1. Values higher than 1 indicate overforecasting (too frequently) and less than 1 indicate underforecasting (not frequent enough). The B is not a true verification measure as it does not measure how well the prediction corresponds to the observation. Skill Score (SS) uses 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, Heidke SS (HSS) and Gilbert SS (GSS, or equitable TS (ETS)), are employed in our study. The HSS measures the fraction of correct forecasts (events and non-events, $a+d$ in Table 2) after eliminating those forecasts which 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. In this study, the fraction of correct forecasts $[(a + d)/(a+b+c+d)]$ is not separately assessed because it becomes dominated by the number of correct non-events (d) and its value may be very large for rare events and thus obscure the forecast accuracy. Both the HSS and GSS allow fairer comparison across different regimes. However, they share some drawbacks with many other scores that 1) they tend to go to small values near 0

Table 2. Different prediction skill measures used in this study.

Contingence Table	Observation (Y)		Observation (N)	
	Prediction (Y)	Prediction (N)	Prediction (Y)	Prediction (N)
	a (hits)	c (misses)	b (false alarms)	d (correct non-events)

Skill Measures			
	Formula	**Range**	**Characteristics**
Hit Rate (*H*)	a/(a+c)	[0, 1] (poor – good)	ignore false alarms (b), artificially improved by overforecast
False Alarm Ratio (*FAR*)	b/(a+b)	[0, 1] (good – poor)	ignore misses (c), artificially improved by underforecast
Threat Score (*TS*)	a/(a+b+c)	[0, 1] (poor – good)	ignore correctly-predicted non-events (d)
Frequency Bias (*B*)	(a+b)/(a+c)	0 ~ ∞	<1 (underpredict); >1 (overpredict) Not a true skill measure
Gilbert Skill Score (*GSS*)	a_{ref} = (a+b)*(a+c)/(a+b+c+d) GSS = (a-a_{ref})/(a-a_{ref}+b+c)	[-1/3, 1] (random if = 0)	based on TS corrected for hits due to random chance
Heidke Skill Score (*HSS*)	2*(ad-bc)/ ((a+c)*(c+d)+(a+b)*(b+d))	[-1, 1] (random if = 0)	based on accuracy [(a+d)/(a+b+c+d)] corrected for hits due to random chance
Symmetric Extremal Dependence Index (*SEDI*)	F = b/(b+d) x= *log*F – *log*H – *log*(1– F) + *log*(1– H) y= *log*F + *log*H + *log*(1– F) + *log*(1– H) SEDI = x/y	[-1, 1] (random if =0)	Nondegnerating, base-rate independent, asymptotically equitable, harder to hedge

as the event becomes rarer, which leads to a misperception that rare events cannot be skillfully forecast or a difficulty to track improvements in the forecast performance; and 2) they are susceptible to hedging by overforecasting. In light of these, 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). So far, SEDI has been largely used for the verification of precipitation forecasts from Numerical Weather Prediction (NWP) centers (North *et al.*,, 2013; Rodwell *et al.*, 2015; Haiden and Duffy, 2016). SEDI possesses more desirable properties in comparison with other scores and its family predecessors, including nondegenerating, base-rate independent, asymptotically equitable, harder to hedge, symmetric and asymmetric relative operating characteristic curves (Ferro and Stephenson, 2011). SEDI value of zero demarcates forecasts better (> 0) and worse (< 0) than random forecast (= 0). Ferro and Stephenson (2011) suggest that the SEDI metric only be calculated for recalibrated forecasts (B = 1) in order to avoid the effect of any model bias on the prediction of extremes. However, it is not always possible or feasible to recalibrate forecasts. Although an uncalibrated SEDI score may not be effective at comparing how well a forecast matches reality, it may still provide useful information due to its resistance to hedging, particularly if presented along with the forecast bias. For this reason, SEDI is included here for evaluating the ability of different statistical models against each other in predicting the occurrence of extreme precipitation event.

There are many other metrics used in weather forecast verification and implementing all of them is not feasible. Different metrics measure different aspects of forecast quality and the use of several permits these different aspects or attributes to be assessed. 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 and HSS measure "skill". They together provide complementary assessments of forecast performance. SEDI measures the association between forecast and observed rare events. It is just beginning to be used in weather verification activities, its evaluation is therefore essential to determine what new perspectives it may provide on forecast skill. 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

4.1 Precipitation characteristics

Figure 1 compares the 99th percentile daily precipitation and mean daily precipitation intensity at 0.25° grid for DJF and

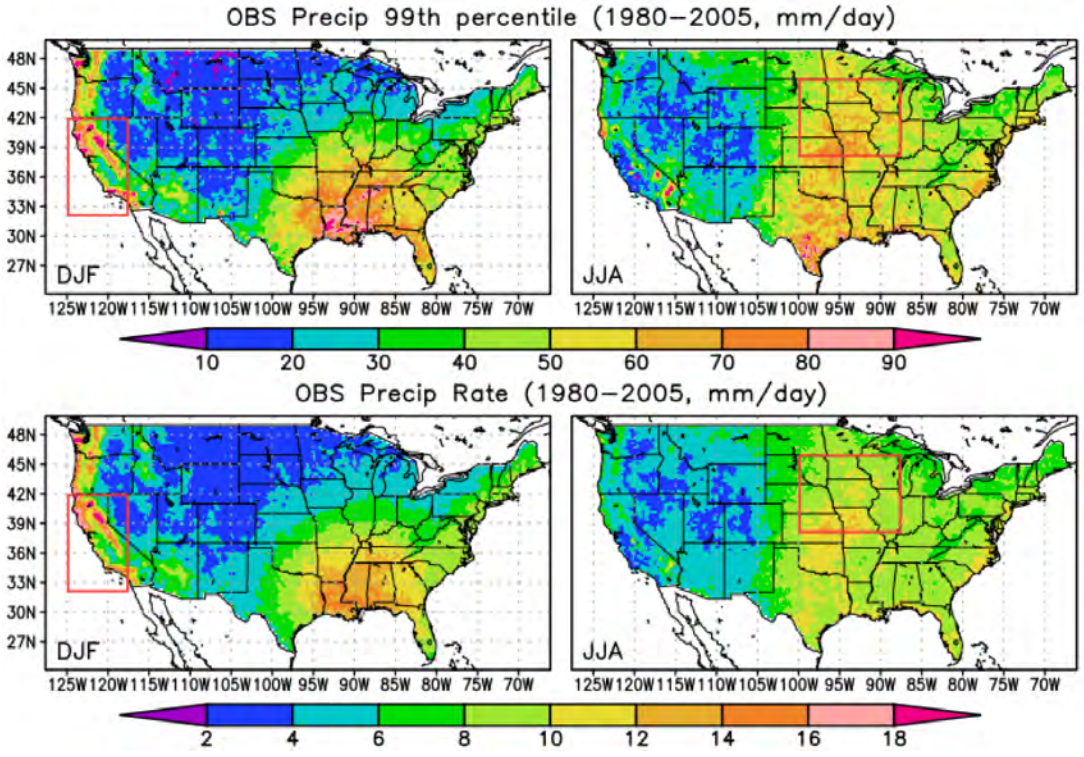

Figure 1. Observed 99th percentile daily precipitation (top) and mean daily precipitation intensity (bottom) at 0.25° grid for DJF (left) and JJA (right) of 1980-2005 in mm/day. Both quantities are calculated with dry days (precipitation < 1 mm/day) excluded. The red rectangles denote our study regions: 30*40 grids and 50*32 grids at 0.25° grid for PCCA and MWST, respectively.

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 U.S. receives a large amount of precipitation, with the 99th percentile possibly reaching up to 130 mm/day. Little precipitation falls in summer with the mean precipitation generally less than 6 mm/day and extreme precipitation less than 30 mm/day 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 seasons, while 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 in summer and 10 to 30 mm/day in winter across the region. Note that the south central U.S. 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°x2° grid our analysis is performed on, the magnitude of extreme precipitation (99th percentile) is systematically underestimated by a factor of two (not shown). However, its large scale pattern across 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.

4.2 Composites for analogue schemes

We extract 41 and 163 extreme precipitation events from the observation of 1980–2005 at 2.5° x 2° for the winter of PCCA and summer of MWST, respectively. **Figure 2** 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 cut-off 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

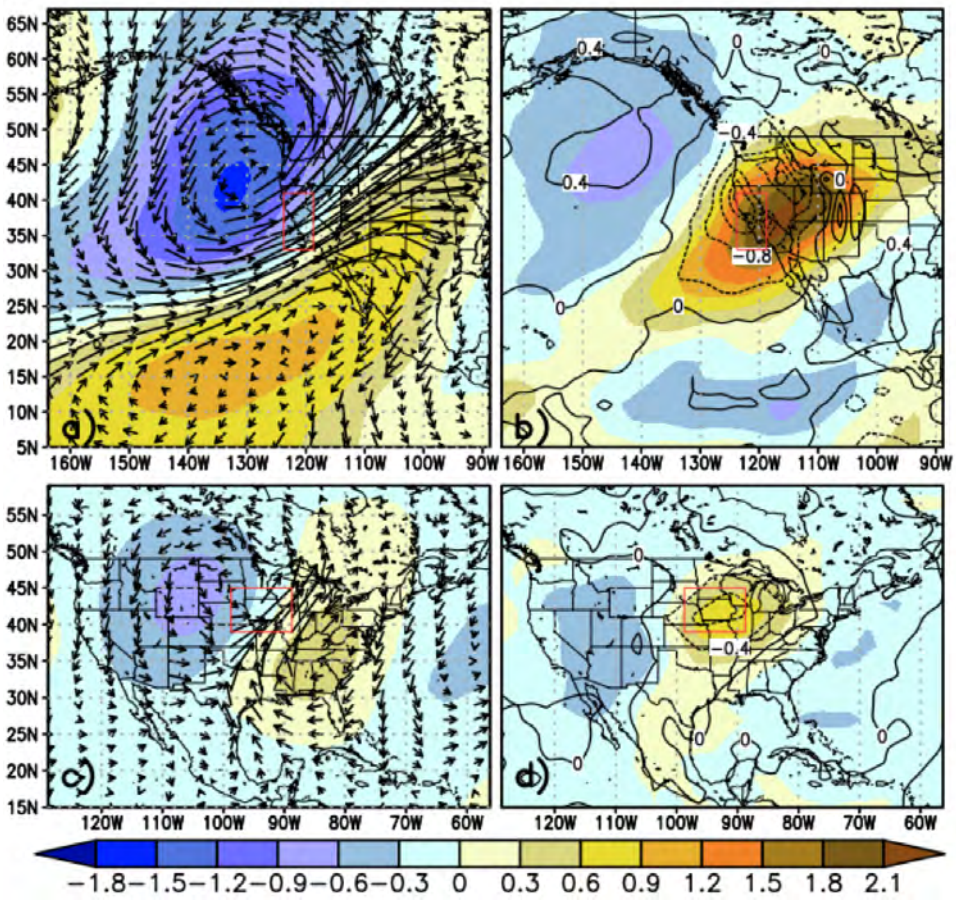

Figure 2. Composite normalized anomalies of (a) 500-hPa geopotential height (shaded, h_{500}) and the vertical integrated moisture flux vector (arrow, *uq,vq*), (b) 500-hPa vertical velocity (contour, w_{500}) and total precipitable water (*tpw*; shaded) for the Pacific Coast California (PCCA) in DJF based on 41 extreme precipitation events identified from the precipitation observation of 1980-2005 at 2.5°x2° grid. (c),(d) as in (a),(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.

towards the West coast of the United States (Fig. 2a). The region is also characterized by stronger large-scale upward motion and higher amount of water vapor (Fig. 2b). 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. 2c). 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. 2c). This fetch of Caribbean moisture links into the Great Plains low-level jet (Dirmeyer and Kinter, 2009), 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. 2d) 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.

4.3 Prediction skill of analogue schemes

Different multi-variate 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 multi-variate conditions are really small, mostly on an order of a hundredth. Here we only present the results based on the multi-variate condition that gives the best GSS during the calibration period. GSS is selected as a reference because it might be the one used most frequently among the variety of performance measures to evaluate skill of deterministic precipitation forecasts (Wang 2014; Boluwade *et al.*, 2017; Chen *et al.*, 2018).

4.3.1 PCCA

Figure 3 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 ET (0.28), GSS (0.27) and HSS (0.42) among all the schemes. This is likely attributed to the tradeoff between high H and high FAR. How much will these scores (ET, GSS, and HSS) be affected by a bias is not entirely clear, i.e. what value of the bias will help or hurt these scores? Nevertheless, an improvement in the skill of MERRA2_Pc over MERRA2_P is evident, with consistently higher ET (0.36), GSS (0.35) and HSS (0.52) but lower FAR (0.5). All the analogue schemes outperform MERRA2_Pc with higher H (0.54 ~ 0.63), ET (0.37~ 0.46), GSS (0.36 ~ 0.46), and HSS (0.53 ~ 0.63), but lower FAR (0.37 ~ 0.46) values. Among the four analogue schemes, the new group constructed with uq and vq [$(uq)(vq)w_{500}tpw$ and $(uq)(vq)w_{500}q_{2m}$] is superior to that constructed with u_{500} and v_{500} [$(uvw)_{500}tpw$ and $(uvw)_{500}q_{2m}$] 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}. These results suggest that the better skill of the $(uq)(vq)$ group is not an artifact of hedging by increasing the bias (B=1 for all the analogue schemes) or

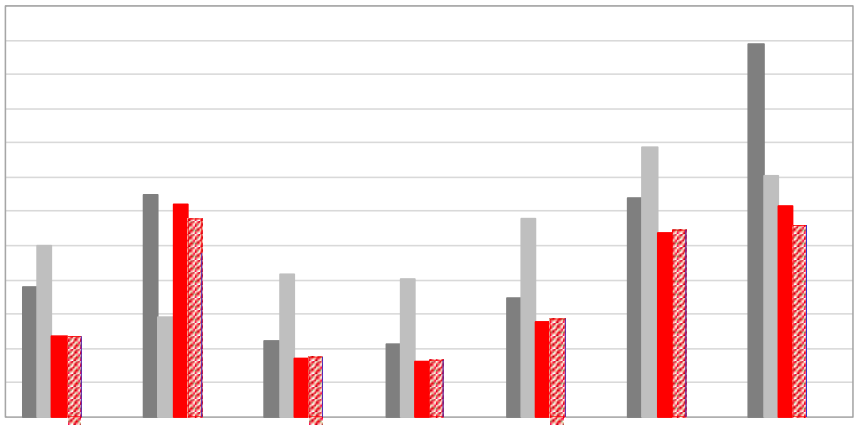

Figure 3. Performance measures of MERRA2 precipitation and various analogue schemes for DJF of PCCA during the a) calibration and b) validation (blind prediction) periods. The number of observed extreme precipitation events is 41 and 34 for two periods, respectively. The numbers in the parentheses of the legend represent the extreme precipitation events detected by each scheme during two periods (separated by slash), respectively. The numbers on the bar indicate the frequency bias of MERRA2_P.

the choice of verification metrics. Among the variety of performance measures, ET, GSS, and HSS exhibit similar behavior except that GSS values are just slightly lower than those of ET due to the number expected correct by chance and that HSS values are the highest. The small differences between ET 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). SEDI has higher magnitudes than all the other verification metrics, but doesn't present much difference among various schemes (two MERRA2 precipitation and four analogue).

Performances are generally worse during the validation than the calibration period, in particular for all the analogue schemes which exhibit large reductions (-0.19 ~ -0.33) in H, ET, GSS, HSS, and SEDI but large increases in FAR (0.11 ~ 0.18) values. MERRA2_Pc is the only case with slight increases (~ 0.06) in ET, GSS and HSS values. 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 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 biases (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, a small decrease in H but a large decrease in FAR for MERRA_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 versus 0.38), ET (0.42 versus 0.22), GSS (0.41 versus 0.21), HSS (0.58 versus 0.35), SEDI (0.79 versus 0.64), and lower FAR (0.20 versus 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 u_{500} and v_{500} 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_{500}) 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_{500}) 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 4 presents the performances of various analogue schemes in depicting the interannual variations of PCCA winter extreme precipitation frequency from 1980 to 2005 (calibration) and 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

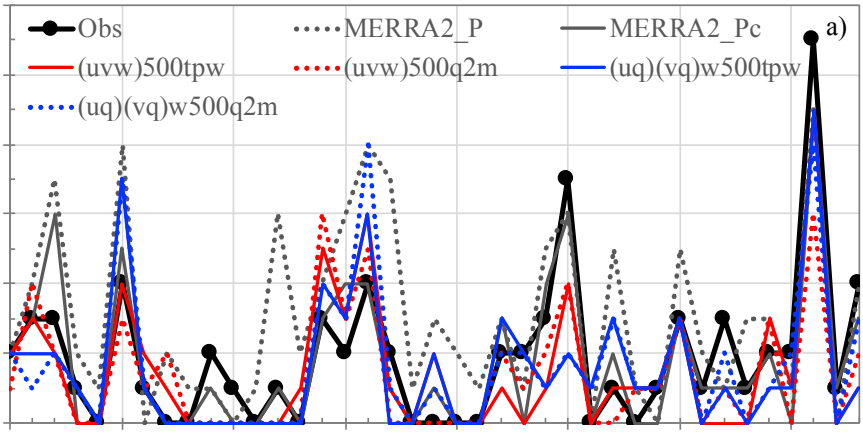

Figure 4. a) Interannual variations of PCCA winter extreme precipitation frequency obtained from various analogue schemes, MERRA2 precipitation, and the observation (obs) during the calibration (1980–2005) and validation (2006–19) periods. b) RMSE (bar) and temporal correlations (scatter, aligned with the corresponding bar for each scheme) between various schemes and observation during two periods. All the correlations are significant at the 0.01 level.

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 (or group of) analogue scheme demonstrating consistently superior performance across the two periods, with RMSEs ranging from 1.3 to 1.8 days. We don't 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).

4.3.2 MWST

Immediately evident are poorer performances in MWST than in PCCA during both periods, in particular for the analogue schemes and MERRA2_Pc (**Figure 5**). There are large differences in biases. 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 ET (~ 0.25), GSS (~ 0.2), HSS (~ 0.35), and SEDI (~ 0.6) values, except for a slightly higher SEDI (~ 0.7) of MERRA2_P. 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), ET (~ 0.15), GSS (~ 0.1), HSS (~ 0.2), and SEDI (~ 0.35) values, but higher FAR (~ 0.65). We also see the group of $(uq)(vq)$ analogue schemes have a slight edge over that of (uv_{500}) by all the performance measures, but differences are likely not significant. Despite the contrasting H and FAR values between two MERRA2 precipitation products, they have comparable skill measures (ET, GSS, HSS, and SEDI) and both outperform all the analogue schemes. However, it is not evident whether the superior performance of MERRA2_P is an artifact of the strong overforecasting because skill measures could behave differently within the wide range of frequency bias. Nevertheless, our results suggest that warm-season extreme precipitation (in our case of MWST), which often occurs with weak synoptic-scale forcing, presents a great forecast challenge, consistent with what previous studies revealed (Carbone *et al.*, 2002; Schumacher and Davis, 2010).

Figure 6 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 MER-

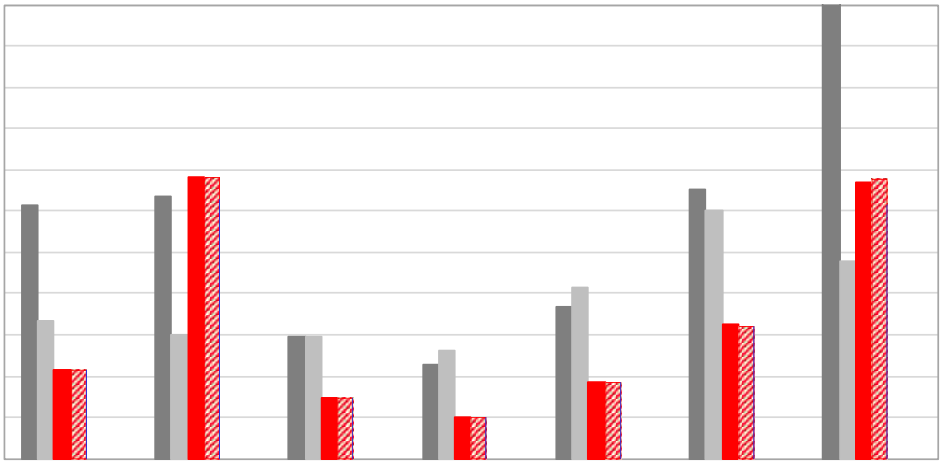

Figure 5. As in Figure 3, but for JJA of MWST. The number of observed extreme precipitation events is 163 and 140 for two periods, respectively.

RA2 precipitation and $(uq)(vq)w_{500}q_{2m}$ are significant at the 0.01 level and that of $(uq)(vq)w_{500}tpw$ significant at the 0.05 level. The group of (uv_{500}) analogue schemes does not capture well these temporal variations with low correlations (~ 0.35). However, the RMSEs of various schemes are approximately one to four 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 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 >= 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 one 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 9 major floods) than during the calibration period (34 events in 2 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 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, 2006 to 2019, and 2014 to 2019 (not shown) and find that the western ridge and moisture transport into the study region has slightly displaced eastward in 2006-2019 and even further in 2014-2019 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.

4.4 Prediction skill of CNN schemes

4.4.1 Oversampling in PCCA

Figure 7 shows performance measures of different CNN schemes trained with the oversampled data set 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 (postive 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. ET, GSS,

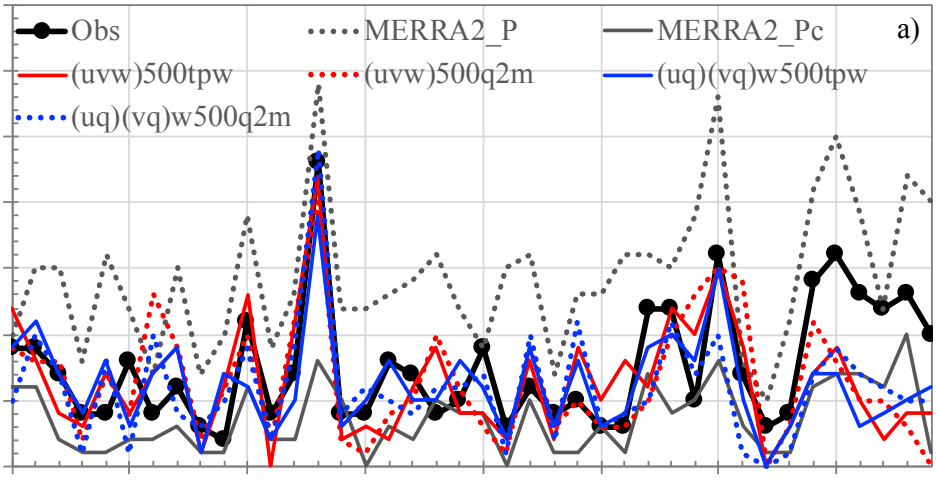

Figure 6. As in Figure 4, but for MWST summer extreme precipitation frequency. All the correlations during the calibration period are significant at the 0.01 level. During the validation period, correlations of $(uvw)_{500}tpw$ and $(uvw)_{500}q2m$ are not significant at the 0.05 level. Correlation of $(uq)(vq)w_{500}tpw$ is significant at the 0.05 level (not the 0.01 level), while correlation of $(uq)(vq)w_{500}q_{2m}$ is significant at the 0.01 level.

and HSS follow a similar pattern of variations, with HSS presenting the largest magnitudes and little difference in those of ET and GSS (hits associated with random chance are negligible). The variations of SEDI largely mimic those of H. Among different schemes, three single-variate (rh_{500}, q_{2m}, and $tdew_{2m}$) and four multi-variate schemes (except $(uq)(vq)w_{500}tpw$) 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 tpw, whose rankings are 18 and 24 in the calibration, but 1 and 7 in the validation, respectively. Such inconsistences are also observed for the analogue schemes, but the exact reason is unknown.

A majority of schemes exhibit consistently poor performances (ranking in the bottom15) between two periods, including the horizontal wind speed, temperature, and specific humidity at all levels (except for t_{500} and q_{2m}) and rh_{700}. Several schemes demonstrate consistently good performances, such as (vq), $tdew_{2m}$, w_{500}, and four multi-variate schemes (except $(uq)(vq)w_{500}tpw$). In addition, multi-variate 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, ET, GSS, HSS, and SEDI 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 rh_{500}, the variations in GSS values are fairly small, ranging from 0.32 ((uq)) to 0.13 (rh_{700}) 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 MERR2_P. Analogue schemes slightly outperforms their CNN counterparts {$(uq)(vq)w_{500}tpw$ and $(uq)(vq)w_{500}q_{2m}$} with higher GSS values (0.256 versus 0.207 and 0.234 versus 0.228, respectively).

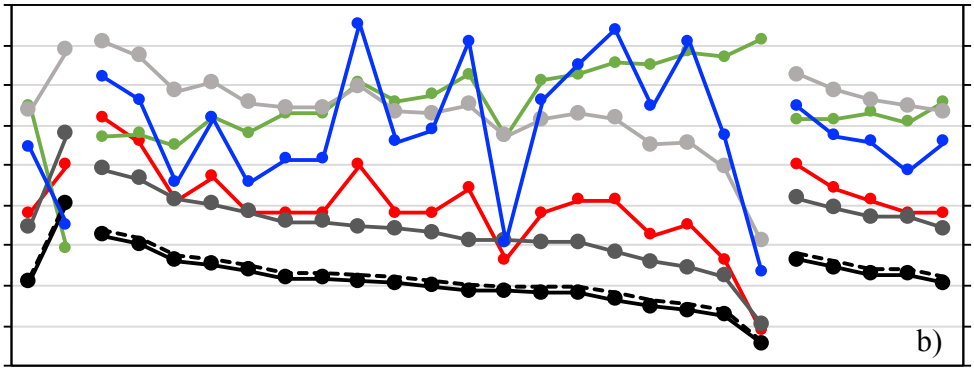

Figure 7. Performance measures of MERRA2 precipitation and various CNN schemes trained with the oversampled data set for DJF of PCCA during the a) calibration and b) validation (blind prediction) periods.

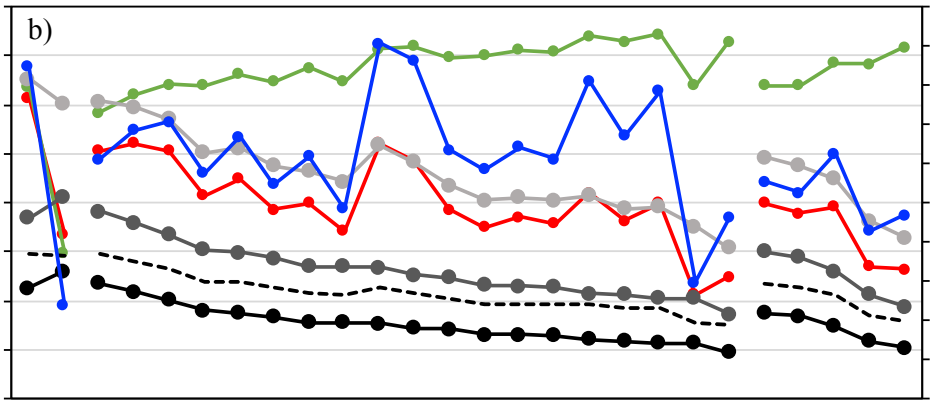

Figure 8. As in Figure 7, but for JJA of MWST.

4.4.2 Oversampling in MWST

Figure 8 shows performance measures of different CNN schemes trained with the oversampled data set 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 u10m which is nearly unbiased (B ~ 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 (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 ET 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 ET, GSS and HSS values, a strong resemblance between the fluctuations of H and SEDI, 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. w_{500} and q_{500} (ranking 11 and 17 based on GSS) show poor calibration performances but good prediction skills (ranking 3 and 4), while the opposite occurs to *q2m* and $(uq)(vq)w_{500}q_{2m}$ with their rankings dropping from 2 to 15 and 6 to 20 between two periods, respectively. We see that LSMPs 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 LSMPs based on vertically-integrated variables and their combinations {(uq), (vq), tpw, $(uq)(vq)$, $(uq)(vq)w_{500}tpw$} 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 could be prevented. However, it is not ensured that such a multi-variate scheme will consistently outperform all the single-variate ones. Both regions share some common variables that provide high skills in prediction of extreme event occurrence, including (uq), (vq), $(uq)(vq)$, w_{500}, and tpw. The specific variables 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 individual or combined variables 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 is particularly true 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 the similar biases.

4.4.3 Interannual Variability

In this section, we only demonstrate the performances of those common variables described above in depicting the interannual variability of extreme event frequency (**Figure 9 and 10**), 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 (Figure 9b). 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 (Figure 9a), we see 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 $(uq)(vq)w_{500}q_{2m}$ outperforms all the others with the lowest RMSEs in both periods (1.58 and 1.3 days), which are comparable to its analogue counterpart (1.6 and 1.3 days). The w_{500} is also superior to most of the schemes with the RMSEs of 1.8 and 1.6 days for two periods, respectively. The $(uq)(vq)w_{500}tpw$ has RMSEs of 2.3 days and 1.6 days, which are worse than its analogue counterpart (1.5 days), particularly in the calibration period. (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

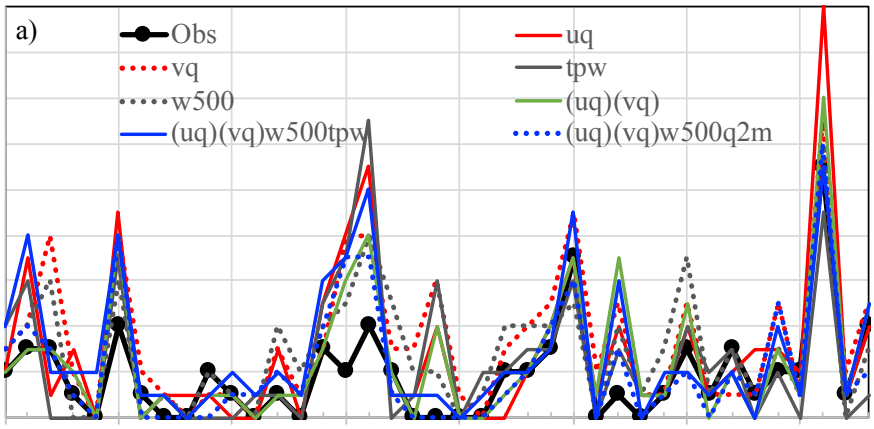

Figure 9. a) Interannual variations of PCCA winter extreme precipitation frequency obtained from selected CNN schemes trained with the oversampled data set and the observation (obs) during the calibration (1980–2005) and validation (2006–19) periods. b) RMSE (bar) and temporal correlations (scatter, aligned with the corresponding bar for each scheme) between various CNN schemes and observation during two periods. MERRA2_Pc is included for a reference.

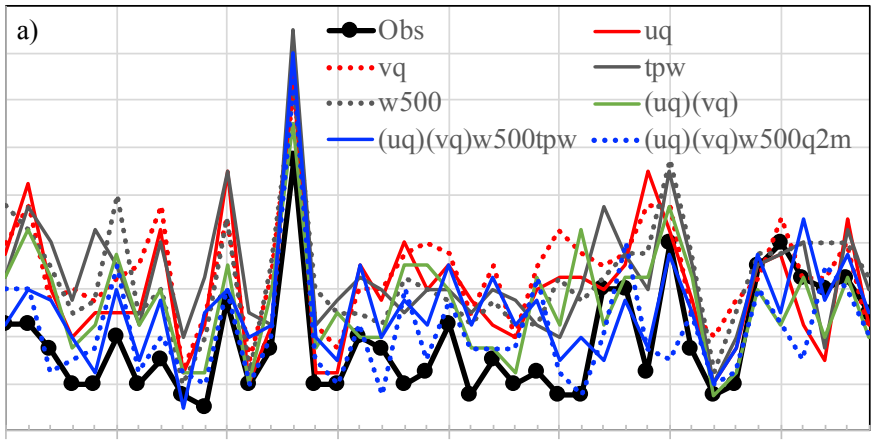

Figure 10. As in Figure 9, but for MWST summer extreme precipitation frequency.

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 9 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 rh_{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 $(uq)(vq)w_{500}q_{2m}$ 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 (Figure 10a). 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 multi-variate 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 versus 2.7 for $(uq)(vq)w_{500}tpw$ and 3.5 versus 3.1 for $(uq)(vq)w_{500}q_{2m}$) in the calibration but lower RMSEs in the validation periods (3.3 and 4.3 versus 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 (Figure 6a).

4.4.4 No-balance

Figure 11 shows performance measures of 19 single-variate CNN schemes trained with the original data set (no-balance) for PCCA and MWST in the validation period. The result of rh_{500}

is not shown for PCCA due to its zero hit and negative skill scores. One distinct difference from the oversampling

Table 3. Correlations and RMSEs of MERRA2 precipitation and CNN schemes trained with the oversampled data set for PCCA and MWST during two periods. "Cal" and "Val" represent "calibration" and "validation", respectively. The normal font in the "Correlation" column indicates that the correlations are significant at the 0.01 level. The bold font in the "Correlation" column indicates that the correlations are **not** significant at the 0.05 level. "#" in the "Correlation" column indicates that the correlations are significant at the 0.05 level, but not at the 0.01 level. "*" in the "RMSE" column indicates that the RMSEs are better than MERRA2_Pc.

	PCCA				MWST			
	Correlation		RMSE		Correlation		RMSE	
	Cal	Val	Cal	Val	Cal	Val	Cal	Val
MERRA2_P	0.68	0.8	2.43	1.71	0.73	0.71	7.39	8.17
MERRA2_Pc	0.87	0.95	1	1.07	0.7	0.79	4.49	5.93
u_{2m}	0.56	0.78	3.51	1.96	0.69	**0.13**	7.52	5.13*
u_{10m}	0.71	0.73	2.32	1.9	0.66	**0.24**	3.08*	6.12
u_{500}	0.64	0.84	2.82	3.67	0.75	**0.22**	12.22	9.83
v_{2m}	0.66	0.67 #	2.9	2.72	0.58	**-0.04**	6.41	5.81*
v_{10m}	0.77	0.67 #	2.67	2.96	0.72	**0.09**	7.53	5.95
v_{500}	0.51	0.58 #	3.55	3	**0.37**	**0.31**	11.03	6.6
$tdew_{2m}$	0.65	0.79	1.99	1.82	0.68	0.72	11.41	8.05
t_{2m}	0.76	0.7	3.06	4.52	0.69	0.72	11.68	9.43
t_{10m}	0.82	0.7	2.59	2.73	0.78	**0.13**	9.69	6.63
t_{500}	0.61	0.83	1.99	1.73	0.77	**0.52**	5.7	4.72*
q_{2m}	0.79	0.77	1.86	2	0.93	0.66 #	5.45	4.44*
q_{10m}	0.61	0.67 #	2.5	2.35	0.84	0.69	10.12	9.87
q_{500}	0.75	0.81	2.02	1.9	0.68	**0.36**	8.53	5.26*
rh_{700}	0.69	0.8	2.15	1.78	0.53	**0.41**	8.98	6.52
rh_{500}	0.71	0.85	1.21	1.9	0.68	**-0.11**	9.56	6.18
tpw	0.6	0.87	2.75	1.41	0.87	0.68	7.33	4.96*
w_{500}	0.77	0.88	1.84	1.64	0.84	0.74	6.59	5.15*
uq	0.7	0.95	2.56	2.27	0.81	**0.29**	6.67	5.97
vq	0.77	0.88	2.51	1.86	0.77	**0.42**	7.66	5.76*
(uq)(vq)	0.79	0.87	1.65	2.09	0.82	**0.3**	5.08	5.27*
(uq)(vq)tpw	0.82	0.89	1.96	1.36	0.76	0.72	5.51	4.06*
$(uq)(vq)q_{2m}$	0.83	0.86	1.97	1.59	0.39 #	0.64 #	6.57	3.56*
$(uq)(vq)w_{500}tpw$	0.81	0.85	2.28	1.64	0.89	0.71	4.62	3.31*
$(uq)(vq)w_{500}q_{2m}$	0.8	0.91	1.58	1.3	0.87	**0.53**	3.5*	4.29*

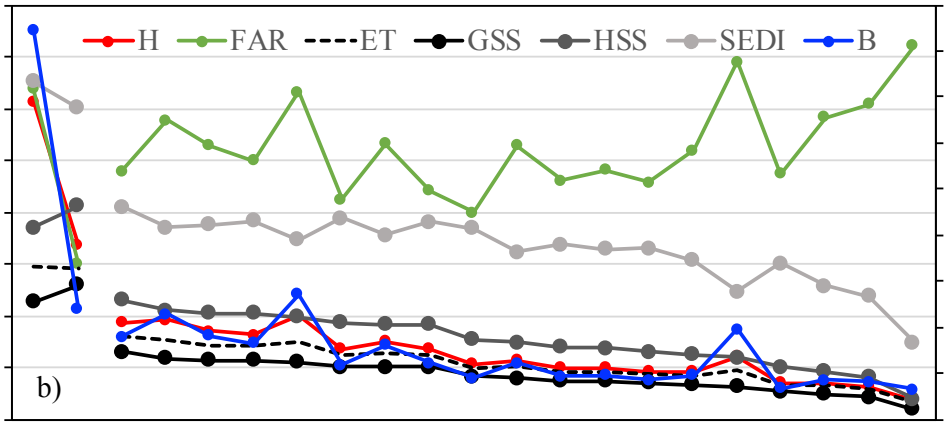

Figure 11. Performance measures of MERRA2 precipitation and various CNN schemes trained with the original dataset during the validation period for a) DJF of PCCA and b) JJA of MWST.

instance is that all the schemes significantly underpredict the extreme frequency for both regions, generally less than half as often as it occurs. The frequency biases range from 0.06 to 0.44 in PCCA and from 0.13 to 0.54 in MWST (only $tdew_{2m}$ exceeds 0.5). The resulting H values are low in both regions, ranging from 0.03 to 0.32 in PCCA and from 0.04 to 0.2 in MWST. However, the FAR values are quite high. In PCCA, only one scheme has the value below 0.4, while 11 out of 19 schemes have the values exceeding 0.6 with the largest as high as 0.8. In MWST, the FAR values vary from 0.4 to 0.72. As compared to the oversampling instance, the GSS values drop remarkably. In PCCA, GSS values range from 0.02 to 0.28 with only one scheme larger than 0.2 and 11 schemes around or below 0.05. Instead, only one out of 19 schemes has the GSS value below 0.1 in the oversampling case. In MWST, GSS values vary from 0.02 to 0.13 versus from 0.1 to 0.24 in the oversampling instance.

Nevertheless, two instances do share some commonalities. In PCCA, (uq), (vq), tpw, w_{500}, and $tdew_{2m}$ are among the top-performing schemes in terms of the GSS value with (uq) and (vq) the two best, while relative humidity and v_{500} generally have poor performances with rh_{500} the worst.

MWST presents the similar top-ranking schemes to PCCA except for $tdew_{2m}$, but the relative humidity and zonal winds at 10m and 500hPa give the overall poor performances. Since the performances of all the schemes in the no-balance case are much poorer than their oversampling counterparts, we will not discuss it further.

5. Summary and Discussions

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 general, synoptic-scale atmospheric dynamics and thermodynamics associated with extreme precipitation are more reliably simulated in general circulation models (GCMs) than mesoscale precipitation, and therefore have shown great promise for predictability via statistical downscaling. In this study, we demonstrate the use of 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) and the summer season of the Midwestern United States (MWST). The potential predictability is explored using two machine learning approaches of different complexity. One is a relatively simple analogue method which has been successfully applied to detect the occurrence of heavy precipitation (95th percentile events) in these two regions (Gao *et al.*, 2014, 2017; Gao and Schlosser, 2018). The other is Convolutional Neural Networks (CNNs), one of the widely used deep learning algorithms. We evaluate the LSMPs constructed with a large set of variables at multiple atmospheric levels in order to understand the relative importance of each variable for predicting extreme precipitation occurrence and how it varies by season and region. The prediction skill of various schemes is quantified using a variety of complementary performance measures.

Our study demonstrates that LSMPs provide useful predictability of extreme precipitation occurrence at a daily scale (only the results of the validation period is summarized). 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. 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. The CNN schemes trained with the oversampled data present more skillful predictions than those with the original data. For the winter extreme precipitation event in PCCA, 15 (20) out of 19 single-variate (all 24) schemes present GSS values around or above 0.2 (maximum 0.32) in the oversampling case with only 2 schemes in the no-balance case. Analogue schemes exhibit comparable prediction skills to those of the oversampled data with GSS values of 0.17 for the $(uv)_{500}$ group and 0.25 for the (uq)(vq) group. Although the performances of both analogue and CNN schemes are not as good as that of MERRA2_Pc (0.41), they are at least comparable or superior to MERRA2_P (0.21), particularly in consideration of MERRA2 precipitation being observation-assimilated. All the analogue schemes and a majority of CNN schemes trained with the oversampled data reproduce the observed interannual variations of extreme frequency reasonably well with the temporal correlations significant at the 0.01 level. MERRA2_Pc has the lowest RMSE (~ 1 day), while analogue schemes (1.3 ~ 1.8 days) are comparable to MERRA2_P (1.7 days). CNN schemes tend to overestimate the frequency in most years with the largest RMSEs (1.3 ~ 4.5 days). For the summer extreme precipitation event in MWST, the prediction skills by all the schemes and MERRA2 precipitation are lower than in PCCA, attributed to the weaker synoptic-scale forcing in the warm season. GSS values of the CNN schemes based on the oversampled data range from 0.1 to 0.237, while those based on the original data range from 0.02 to 0.13 with 11 schemes less than 0.1. The GSS values of analogue schemes range from 0.1 to 0.14, comparable to their two CNN counterparts. MERRA2_P and MERRA2_Pc retain similar prediction skills (GSS values 0.23 and 0.26, respectively) and both outperform all the other schemes except for the one best scheme trained with the oversampled data. More than half of the CNN schemes based on the oversampled data do not track the observed year-to-year variations of extreme events well with correlation coefficients not significant at the 0.05 level. Only the (uq)(vq) group of analogue schemes presents the correlations significant at the 0.05 level, while both MERRA2 precipitation have the correlations significant

at the 0.01 level. Analogue schemes and most CNN schemes based on the oversampled data show comparable RMSEs to or lower RMSEs than MERRA2_Pc (5.9 days) and much lower RMSEs than MERRA2_P (8.2 days).

Regardless of the regions (seasons) examined here, there is no single scheme in any of two methods that will perform consistently the best in detecting extreme precipitation occurrence at a daily scale and its interannual variation during both the calibration and validation periods. Nevertheless, 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 based on the oversampled data, (uq) and (vq) unanimously perform the best in predicting daily extreme precipitation occurrence followed by w_{500} and tpw over both regions. Surface temperature and horizontal wind speeds, relative humidity at the lower (700hpa) and middle (500hPa) 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 (uq)(vq) group usually outperforming the $(uv)_{500}$ group over both regions. Previous studies have also documented Integrated Vapor Transport (IVT, equivalent to the magnitude of "uq" and "vq") as a very important ingredient for extreme precipitation production (Nakamura *et al.*, 2013; Agel *et al.*, 2019) 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 collinear ones, seems to help improve prediction skill. Although it is not ensured that such multi-variate 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 has so far been based solely on the GSS values without taking into account the frequency bias of each scheme. There are large differences in the frequency bias of various CNN schemes, which somewhat 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 than if the two forecasts have the same bias. However, it is not obvious how much and in what way GSS values are affected by frequency bias. For example, what specific value of frequency bias may help or hurt GSS value of particular scheme? The H value could be increased by forecasting the event more often, but the FAR value might also rise. The actual impact on GSS value will depend on their relative increases and may result in the difference in the benefit from the elevated bias. Although the identified top-performing CNN schemes tend to overforecast ($B > 1$), their frequency biases are moderate ($1.2 \sim 1.4$) among all. In addition, the analogue schemes tend to underforecast ($B < 1$), but have comparable GSS values to their CNN counterparts. These results imply that the high prediction skills of top-performing CNN schemes cannot be solely attributed to the overprediction artifact. In terms of other prediction accuracy measures, ET and HSS follow the similar pattern of variations to GSS with the differences in magnitude, while SEDI largely resembles H.

For the two methods employed in this study, the analogue method mainly relies on characterizing the similarity of daily LSMPs to their corresponding composites averaged from the ensemble of observed extreme events, while CNNs utilizes multiple layers of artificial neurons and filters to extract the relevant features from complex, hierarchical high-dimensional data. Both methods require no assumptions about the normality, linearity or continuity of the data sample. However, each method 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 from different schemes (combinations of variables) 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. This could be particularly useful when there is a slight shift in the relevant features of LSMPs associated with extreme precipitation occurrence. However, there could be a large difference in the frequency bias of various CNN schemes trained with the same dataset, which somewhat complicates the evaluation of their relative skills. CNN schemes may also be prone to overfitting or suffer from the issue of physical interpretability. Nevertheless, some arbitrariness is inevitable in determining either hyperparameters in CNN (i.e. number of hidden layers, filter size, etc.) or criteria in analogue method (i.e. spatial domain to calculate spatial correlation, specific multi-variate condition, etc.). Such choice could lead to minor differences in the resulting skills and often there is no simple rule of thumb to obtain their optimal values. In addition, non-stationarity of the relationships between the LSMPs and extreme precipitation occurrence has been a common major challenge for both methods.

Prediction of extreme precipitation events at a regional scale is of great significance due to their severe impacts on society. One immediate question is: is there any true skill in forecasting these rare events? If so, how much? Our results suggest the answer is yes and quite a bit. Overall, CNN seems 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 multi-variate analogue scheme in PCCA and more than half of CNN schemes in MWST. It is possible that there

is more predictability to be harvested from the LSMPs than we realized from this exercise. It should be noted that the specific details of the results of this investigation are almost certainly dependent on the choices of many elements, such as the definition of extreme precipitation events, study region, reanalysis data, predictors, hyperparameters of CNN, criteria of analogue method, etc. Few of the studies are directly comparable because these elements and employed approaches are quite varied. To improve prediction skill, oversampling strategy could be explored for the analogue method as well. The application of CNN in our study 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 no enough training data (e.g. extreme weather prediction). Future works will focus on knowledge-guided machine learning (KGML) models, which stand a better chance in safeguarding against non-generalizable features by integrating scientific knowledge (explainable physical theories) and data synergistically.

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/.

6. References

Agel, L., M. Barlow, S.B. Feldstein & W.J. Gutowski Jr. (2018). Identification of large-scale meteorological patterns associated with extreme precipitation in the US northeast. *Clim. Dyn.*, 50, 1819–1839, doi: 10.1007/s00382-017-3724-8

Anandhi, A., V.V. Srinivas, R.S. Nanjundiah & D. Nagesh Kumar (2008). Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. *Int. J. Climatol.*, 28: 401-420, doi: 10.1002/joc.1529

Barlow, M., W.J. Gutowski, J.R. Gyakum, *et al.* (2019). North American extreme precipitation events and related large-scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. *Clim. Dyn.*, 53, 6835–6875, doi: 10.1007/s00382-019-04958-z.

Boluwade, A., T. Stadnyk, V. Fortin and G. Roy (2017). Assimilation of precipitation Estimates from the Integrated Multisatellite Retrievals for GPM (IMERG, early Run) in the Canadian Precipitation Analysis (CaPA). *J. Hydrology: Regional Studies*. 14, 10-22.

Bosilovich, M.G. (2013). Regional climate and variability of NASA MERRA and recent reanalyses: US summertime precipitation and temperature. *J. Appl. Meteorol. Climatol.*, 52: 1939–1951. doi: 10.1175/JAMC-D-12-0291.1

Bosilovich, M.G., R. Lucchesi & M. Suarez (2016). MERRA-2: file specification. GMAO Office Note No. 9 (Version 1.1): p 73. http://gmao. gsfc.nasa.gov/pubs/office_notes

Buda M, A. Maki & M.A. Mazurowski (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249-259.

Carbone, R.E., J.D. Tuttle, D.A. Ahijevych & S.B. Trier (2002). Inferences of predictability associated with warm season precipitation episodes. *J. Atmos. Sci.*, 59, 2033–2056.

Casola, J.H. & J.M. Wallace (2007). Identifying weather regimes in the wintertime 500-hpa geopotential height field for the Pacific– North American sector using a limited-contour clustering technique. *J. Appl. Meteor.*, 46, 1619–1630, doi:10.1175/JAM2564.1.

Cavazos, T. (2000). Using self-organizing maps to investigate extreme climate events: An application to wintertime precipitation in the Balkans. *J. Climate*, 13, 1718–1732.

Chen, X., L.R. Leung, Y. Gao, *et al.* (2018). Predictability of extreme precipitation in western U.S. watersheds based on atmospheric river occurrence, intensity, and duration. *Geophys. Res. Lett.*, 45, 11,693– 11,701, doi: 10.1029/2018GL079831

Chen, C. & T. Knutson (2008). On the verification and compari- son of extreme rainfall indices from climate models. *J. Climate*, 21, 1605–1621, doi:10.1175/2007JCLI1494.1.

Chollet, F. (2015). Keras. https://keras.io.

Christensen J.H., T.R. Carter, M. Rummukainen & G. Amanatidis (2007). Evaluating the performance and utility of regional climate models: the PRUDENCE project. *Clim. Chang.*, 81(Suppl 1): 1–6, doi: 10.1007/s10584-006-9211-6

DeAngelis, A.M., A.J. Broccoli & S.G. Decker (2013). A comparison of CMIP3 simulations of precipitation over North America with observations: Daily statistics and circulation features accompanying extreme events. *J. Climate*, 26, 3209–3230, doi:10.1175/JCLI-D-12-00374.1.

Dirmeyer, P.A. & J.L. Kinter III (2010). Floods over the U.S. Midwest: A regional water cycle perspective. *J. Hydrometeor.*, 11, 1172–1181, doi:10.1175/2010JHM1196.1.

Farnham, D.J., J. Doss-Gollin & U. Lall, U. 2018: Regional extreme precipitation events: Robust inference from credibly simulated GCM variables. *Water Resour. Res.*, 54, 3809– 3824, doi: 10.1002/2017WR021318.

Ferro, C.A.T. & D.B. Stephenson (2011). Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events. *Weather Forecast.* 26: 699–713.

Gao, X. & C.A. Schlosser (2019). Mid-Western US heavy summer-precipitation in regional and global climate models: the impact on model skill and consensus through an analogue lens. *Clim. Dyn.*, 52, 1569–1582, doi: 10.1007/s00382-018-4209-0

Gao, X., C.A. Schlosser, P.A. O'Gorman, E. Monier & D. Entekhabi (2017). Twenty-first-century changes in US regional heavy precipitation frequency based on resolved atmospheric patterns. *J. Clim*. 30: 2501–2521.

Gao, X., C.A. Schlosser, P. Xie, E. Monier & D. Entekhabi (2014). An analogue approach to identify heavy precipitation events: Evaluation and application to CMIP5 climate models in the United States. *J. Climate*, 27, 5941–5963, doi:10.1175/JCLI-D-13-00598.1.

Gao, Y., J. Lu, L.R. Leung, *et al.* (2015). Dynamical and thermodynamical modulations on future changes of landfalling atmospheric rivers over western North America, *Geophys. Res. Lett.*, 42, 7179– 7186, doi:10.1002/2015GL065435.

Gershunov, A. *et al.*, (2019). Precipitation regime change in Western North America: The role of Atmospheric Rivers. *Sci. Rep.*, 9, 9944. https://doi.org/10.1038/s41598-019-46169-w

Haiden, T. and S. Duffy (2016). Use of high-density observations in precipitation verification.

ECMWF Newsletter No. 147.

Higgins, R.W., W. Shi, E. Yarosh & R. Joyce (2000). Improved United States Precipitation Quality Control System and Analysis. NCEP/ Climate Prediction Center Atlas No. 7. [Available online at http://www.cpc.ncep.noaa.gov/research_papers/ ncep_cpc_atlas/7/index.html.]

Higgins, R.W., V. Silva, W. Shi & J. Larson (2007). Relationships between climate variability and fluctuations in daily precipitation over the United States. *J. Climate*, 20, 3561–3579, doi:10.1175/JCLI4196.1.

Hohenegger, C. & C. Schar (2007). Atmospheric Predictability at Synoptic Versus Cloud-Resolving Scales. *Bull. Amer. Meteor. Soc.*, **88**, 1783–1794, doi: 10.1175/BAMS-88-11-1783.

Hope, P.K. (2006). Projected future changes in synoptic systems influencing southwest Western Australia. *Clim. Dyn.*, 26, 765–780, doi: 10.1007/s00382-006-0116-x.

Huff, F.A. & James R. Angel (1992). *Rainfall Frequency Atlas of the Midwest*. Illinois State Water Survey, Champaign, Bulletin 71.

Japkowicz, N. & S. Stephen (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449.

Jewson, S. (2020). An Alternative to PCA for Estimating Dominant Patterns of Climate Variability and Extremes, with Application to U.S. and China Seasonal Rainfall. *Atmosphere*, 11, 354, doi: 10.3390/atmos11040354.

Kawazoe, S. & W.J. Gutowski (2013). Regional, very heavy daily precipitation in CMIP5 simulations. *J. Hydrometeor.*, 14: 1228–1242.

Kharin, V.V., F.W. Zwiers & M. Wehner (2013). Changes in temperature and precipitation extremes in the CMIP5 ensemble. *Climatic Change*, 119, 345–357, doi:10.1007/s10584-013-0705-8.

Knighton, J., G. Pleiss, E. Carter, *et al.* (2019). Potential Predictability of Regional Precipitation and Discharge Extremes Using Synoptic-Scale Climate Information via Machine Learning: An Evaluation for the Eastern Continental United States. *J. Hydrometeor.*, **20**, 883–900, doi: 10.1175/JHM-D-18-0196.1.

Lamjiri, M.A., M.D. Dettinger, F.M. Ralph & B. Guan (2017). Hourly storm characteristics along the U.S. west coast: Role of atmospheric rivers in extreme precipitation. *Geophys. Res. Lett.*, 44, 7020–7028, doi: 10.1002/2017GL074193

Lennard, C. & G. Hegerl (2015). Relating changes in synoptic circulation to the surface rainfall response using self-organising maps. *Clim. Dyn.*, 44, 861–879, doi: 10.1007/s00382-014-2169-6.

Leung, L.R. & Y. Qian (2009). Atmospheric rivers induced heavy precipitation and flooding in the western U.S. simulated by the WRF regional climate model. *Geophys. Res. Lett.*, 36, L03820, doi: 10.1029/2008GL036445

Li, J. & B. Wang (2018). Predictability of summer extreme precipitation days over eastern China. *Clim. Dyn.*, 51, 4543–4554, doi: 10.1007/s00382-017-3848-x

Loikith, P.C., B.R. Lintner & A. Sweeney (2017). Characterizing Large-Scale Meteorological Patterns and Associated Temperature and Precipitation Extremes over the Northwestern United States Using Self-Organizing Maps. *J. Climate*, **30**, 2829–2847, doi: 10.1175/JCLI-D-16-0670.1.

Lu, M., U. Lall, J. Kawale, S. Liess & V. Kumar (2016). Exploring the Predictability of 30-Day Extreme Precipitation Occurrence Using a Global SST–SLP Correlation Network. *J. Climate*, **29**, 1013–1029, doi: 10.1175/JCLI-D-14-00452.1.

Maraun, D. *et al.*, (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.*, 48, RG3003, doi:10.1029/2009RG000314.

Mazurowski, M.A., P.A. Habas, J.M. Zurada, J.Y. Lo, J.A. Baker & G.D. Tourassi. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2):427– 436.

Milrad, S.M., E.H. Atallah, J.R. Gyakum & G. Dookhie (2014). Synoptic typing and precursors of heavy warm-season precipitation events at Montreal, Quebec. *Weather Forecast*, 29, 419–444, doi: 10.1175/WAF-D-13-00030.1.

Min, S.K., X. Zhang, F.W. Zwiers & G.C. Hegerl (2011). Human contribution to more-intense precipitation extremes. *Nature*, 470, 378–381, doi:10.1038/nature09763.

Molod, A., L. Takacs, M. Suarez & J. Bacmeister (2015). Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA2. *Geosci. Model Dev.*, 8:1339–1356.

Nakamura, J., U. Lall, Y. Kushnir, *et al.* (2013). Dynamical Structure of Extreme Floods in the U.S. Midwest and the United Kingdom. *J. Hydrometeorology*, *14*(2): 485-504.

North, R., M. Trueman, M. Mittermaier & M.J. Rodwell (2013). An assessment of the SEEPS and SEDI metrics for the verification of 6 h forecast precipitation accumulations. *Met. Apps*, 20: 164-175, doi: 10.1002/met.1405

Ralph, F.M., P.J. Neiman, G.A. Wick, *et al.* (2006). Flooding on California's Russian River: Role of atmospheric rivers. *Geophys. Res. Lett.*, 33, L13801, doi: 10.1029/2006GL026689

Reichle R., Q. Liu, R. Koster, *et al.* (2017). Land Surface Precipitation in MERRA-2. *J. Clim.*, 30, 1643–1664.

Reusch, D.B., R.B. Alley & B.C. Hewitson (2005). Relative performance of self-organizing maps and principal component analysis in pattern extraction from synthetic climatological data. *Polar Geogr.*, 29, 188–212, doi:10.1080/789610199.

Rodwell, M.J. *et al.*, (2015). New developments in the diagnosis and verification of high-impact weather forecasts. *ECMWF Technical Memorandum* No. 759.

Sachindra, D.A., F. Huang, A. Barton & B.J.C. Perera (2014). Statistical downscaling of general circulation model outputs to catchment scale hydroclimatic variables: Issues, challenges and possible solutions. *J. Water Clim. Change*, 5(4): 496-525.

Schlef, K.E., H. Moradkhani & U. Lall (2019). Atmospheric circulation patterns associated with extreme united states floods identified via machine learning. *Sci. Rep.*, 9, 7171, doi: 10.1038/s41598-019-43496-w

Schumacher, R.S. & C.A. Davis (2010). Ensemble-Based Forecast Uncertainty Analysis of Diverse Heavy Rainfall Events. *Wea. Forecasting*, 25, 1103–1122, doi: 10.1175/2010WAF2222378.1.

Sillmann, J., V.V. Kharin, F.W. Zwiers, *et al.* (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *J. Geophys. Res. Atmos.*, 118, 2473–2493, doi: 10.1002/ jgrd.50188.

Sukovich, E.M., F.M. Ralph, F.E. Barthold, *et al.* (2014). Extreme quantitative precipitation forecast performance at the weather prediction center from 2001 to 2011. *Wea. Forecasting*, 29, 894-911.

Wang, C. (2014). On the calculation and correction of Equitable Threat Score for model quantitative precipitation forecasts for small verification areas: the example of Taiwan. *Weather and Forecasting*, *29*(4): 788-798.

Wehner, M.F. (2013). Very extreme seasonal precipitation in the NARCCAP ensemble: model performance and projections. *Clim. Dyn.* 40, 59–80, doi: 10.1007/s00382-012-1393-1

Wehner, M.F. *et al.*, (2014). The effect of horizontal resolution on simulation quality in the Community Atmospheric Model, CAM5.1. *J Model Earth Syst* 6: 980–997, doi: 10.1002/2013MS000276

Joint Program Report Series - Recent Articles

For limited quantities, Joint Program Reports are available free of charge. Contact the Joint Program Office to order.
Complete list: http://globalchange.mit.edu/publications

353. **Predictability of U.S. Regional Extreme Precipitation Occurrence Based on Large-Scale Meteorological Patterns (LSMPs).** *Gao & Mathur, Jun 2021*

352. **Toward Resilient Energy Infrastructure: Understanding the Effects of Changes in the Climate Mean and Extreme Events in the Northeastern United States.** *Komurcu & Paltsev, Jun 2021*

351. **Meeting Potential New U.S. Climate Goals.** *Yuan et al., Apr 2021*

350. **Hydroclimatic Analysis of Climate Change Risks to Global Corporate Assets in Support of Deep-Dive Valuation.** *Strzepek et al., Apr 2021*

349. **A Consistent Framework for Uncertainty in Coupled Human-Earth System Models.** *Morris et al., Mar 2021*

348. **Changing the Global Energy System: Temperature Implications of the Different Storylines in the 2021 Shell Energy Transformation Scenarios.** *Paltsev et al., Feb 2021*

347. **Representing Socio-Economic Uncertainty in Human System Models.** *Morris et al., Feb 2021*

346. **Renewable energy transition in the Turkish power sector: A techno-economic analysis with a high-resolution power expansion model, TR-Power.** *Kat, Feb 2021*

345. **The economics of bioenergy with carbon capture and storage (BECCS) deployment in a 1.5°C or 2°C world.** *Fajardy et al., Nov 2020*

344. **Future energy: In search of a scenario reflecting current and future pressures and trends.** *Morris et al., Nov 2020*

343. **Challenges in Simulating Economic Effects of Climate Change on Global Agricultural Markets.** *Reilly et al., Aug 2020*

342. **The Changing Nature of Hydroclimatic Risks across South Africa.** *Schlosser et al., Aug 2020*

341. **Emulation of Community Land Model Version 5 (CLM5) to Quantify Sensitivity of Soil Moisture to Uncertain Parameters.** *Gao et al., Feb 2020*

340. **Can a growing world be fed when the climate is changing?** *Dietz and Lanz, Feb 2020*

339. **MIT Scenarios for Assessing Climate-Related Financial Risk.** *Landry et al., Dec 2019*

338. **Deep Decarbonization of the U.S. Electricity Sector: Is There a Role for Nuclear Power?** *Tapia-Ahumada et al., Sep 2019*

337. **Health Co-Benefits of Sub-National Renewable Energy Policy in the U.S.** *Dimanchev et al., Jun 2019*

336. **Did the shale gas boom reduce US CO_2 emissions?** *Chen et al., Apr 2019*

335. **Designing Successful Greenhouse Gas Emission Reduction Policies: A Primer for Policymakers – The Perfect or the Good?** *Phillips & Reilly, Feb 2019*

334. **Implications of Updating the Input-output Database of a Computable General Equilibrium Model on Emissions Mitigation Policy Analyses.** *Hong et al., Feb 2019*

333. **Statistical Emulators of Irrigated Crop Yields and Irrigation Water Requirements.** *Blanc, Aug 2018*

332. **Turkish Energy Sector Development and the Paris Agreement Goals: A CGE Model Assessment.** *Kat et al., Jul 2018*

331. **The economic and emissions benefits of engineered wood products in a low-carbon future.** *Winchester & Reilly, Jun 2018*

330. **Meeting the Goals of the Paris Agreement: Temperature Implications of the Shell Sky Scenario.** *Paltsev et al., Mar 2018*

329. **Next Steps in Tax Reform.** *Jacoby et al., Mar 2018*

328. **The Economic, Energy, and Emissions Impacts of Climate Policy in South Korea.** *Winchester & Reilly, Mar 2018*

327. **Evaluating India's climate targets: the implications of economy-wide and sector specific policies.** *Singh et al., Mar 2018*

326. **MIT Climate Resilience Planning: Flood Vulnerability Study.** *Strzepek et al., Mar 2018*

325. **Description and Evaluation of the MIT Earth System Model (MESM).** *Sokolov et al., Feb 2018*

324. **Finding Itself in the Post-Paris World: Russia in the New Global Energy Landscape.** *Makarov et al., Dec 2017*

323. **The Economic Projection and Policy Analysis Model for Taiwan: A Global Computable General Equilibrium Analysis.** *Chai et al., Nov 2017*

322. **Mid-Western U.S. Heavy Summer-Precipitation in Regional and Global Climate Models: The Impact on Model Skill and Consensus Through an Analogue Lens.** *Gao & Schlosser, Oct 2017*

321. **New data for representing irrigated agriculture in economy-wide models.** *Ledvina et al., Oct 2017*

320. **Probabilistic projections of the future climate for the world and the continental USA.** *Sokolov et al., Sep 2017*

MIT Joint Program on the Science and Policy of Global Change

Massachusetts Institute of Technology
77 Massachusetts Ave., E19-411
Cambridge MA 02139-4307 (USA)

T (617) 253-7492 **F** (617) 253-9845
globalchange@mit.edu
http://globalchange.mit.edu/