Objective Methods for Thinning the Frequency of Reforecasts while Meeting Postprocessing and Model Validation Needs

Sergey Kravtsov,[a] Paul Roebber,[a] Thomas M. Hamill,[b] and James Brown[c]

[a] University of Wisconsin–Milwaukee, Milwaukee, Wisconsin
[b] NOAA/Physical Sciences Laboratory, Boulder, Colorado
[c] Hydrologic Solutions Limited, Southgate Chambers, Winchester, Southampton, United Kingdom

(Manuscript received 30 September 2021, in final form 28 February 2022)

ABSTRACT: This paper utilizes statistical and statistical–dynamical methodologies to select, from the full observational record, a minimal subset of dates that would provide representative sampling of local precipitation distributions across the contiguous United States (CONUS). The CONUS region is characterized by a great diversity of precipitation-producing systems, mechanisms, and large-scale meteorological patterns (LSMPs), which can provide favorable environment for local precipitation extremes. This diversity is unlikely to be adequately captured in methodologies that rely on grossly reducing the dimensionality of the data—by representing it in terms of a few patterns evolving in time—and thus requires data thinning techniques based on high-dimensional dynamical or statistical data modeling. We have built a novel high-dimensional empirical model of temperature and precipitation capable of producing statistically accurate surrogate realizations of the observed 1979–99 (training period) evolution of these fields. This model also provides skillful hindcasts of precipitation over the 2000–20 (validation) period. We devised a subsampling strategy based on the relative entropy of the empirical model's precipitation (ensemble) forecasts over CONUS and demonstrated that it generates a set of dates that captures a majority of high-impact precipitation events, while substantially reducing a heavy-precipitation bias inherent in an alternative methodology based on the direct identification of large precipitation events in the Global Ensemble Forecast System (GEFS), version 12 reforecasts. The impacts of data thinning on the accuracy of precipitation statistical postprocessing, as well as on the calibration and validation of the Hydrologic Ensemble Forecast Service (HEFS) reforecasts are yet to be established.

SIGNIFICANCE STATEMENT: High-impact weather events are usually associated with extreme precipitation, which is notoriously difficult to predict even using highly resolved state-of-the-art numerical weather prediction models based on first physical principles. The same is true for statistical models that use past data to anticipate the future behavior likely to stem from an observed initial state. Here we use both types of models to identify the occurrences of the states, over the historical climate record, which are likely to lead to extreme precipitation events. We show that the overall statistics of precipitation over the contiguous United States is encapsulated in a greatly reduced set of such states, which could substantially alleviate the computational burden associated with testing of hydrological forecast models used for decision support.

KEYWORDS: Precipitation; Numerical weather prediction/forecasting; Probabilistic Quantitative Precipitation Forecasting (PQPF); Statistical forecasting

1. Introduction

The statistical postprocessing of weather forecasts has been shown to be extremely useful for ameliorating model biases and extracting usable forecast signal amidst the noise due to chaotic error growth and sampling due to limited ensemble size (Hamill and Whitaker 2006; Hamill et al. 2006, 2013, 2015; Scheuerer and Hamill 2015). Postprocessed forecasts are typically more skillful and reliable, rendering them useful for automated decision support. Large sample sizes of reforecasts are particularly helpful in two particular situations: (i) the postprocessing of rare events; and (ii) the postprocessing of longer-lead events, where usable signal is small, noise is large, and forecasts are for time-averaged quantities. While the production of a long, complete time series of reforecasts is desirable for such situations, the computational expense of reforecasting scales linearly with the reforecast sample size. Objective methods that can indicate what subset of dates are the most important to generate reforecasts are greatly desired. Given the national forecast responsibilities of the National Weather Service (NWS), that subset of dates should ideally be large enough to provide the necessary training and validation data over the contiguous United States (CONUS). In this paper, we develop a novel, numerically efficient predictive statistical model of precipitation over CONUS that is able to isolate weather states leading to a likely extreme precipitation event in the future. The occurrences of these states form the basis of our desired thinned subsample of reforecast dates.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-21-0162.s1.

Corresponding author: Sergey Kravtsov, kravtsov@uwm.edu

DOI: 10.1175/WAF-D-21-0162.1

There are several challenges to be anticipated with designing a procedure for reforecast subsampling. One challenge of sub-selecting past dates is that they will be less useful for training if the dates are based on the existence of *observed* high-impact weather such as heavy precipitation. In such a case, the training data are biased toward the existence of high-impact events, and postprocessed guidance will likely overforecast them. Accordingly, we seek methodologies for deciding on which dates to use that avoid the use of validating observations but instead use only information such as the initial condition state or the existence of conditions related to severe weather at a similar date noted in previous reforecasts.

Yet another challenge could be the undersampling of more commonplace events. Were such a reforecast subsampling procedure designed for a very limited geographic area, dry weather or light/moderate precipitation could be drastically undersampled, leading to poor-quality guidance of more common weather events. However, suppose a methodology is developed to identify past cases with high-impact weather separately for multiple regions across the CONUS. We would anticipate that high-impact weather in one region would coincide with more commonplace weather in other regions, thereby avoiding undersampling of more commonplace events when forming the overall sample. Thus, reforecasts conducted from a union of the identified dates, we hypothesize, should be adequate for training of both common and uncommon weather forecast postprocessing.

In subsampling, and thereby reducing, the number of historical dates on which reforecasting is conducted, the “thinned” reforecasts must eventually facilitate end-user applications, such as hydrologic forecasting, watch/warning operations, and decision support. These applications include, for example, the Hydrologic Ensemble Forecast Service (HEFS; Demargne et al. 2014), which is used by the 13 River Forecast Centers (RFCs) of the NWS to produce reliable and skillful hydrologic forecasts for, among other things, informing flood forecasting operations and managing water resources. In this study, we will evaluate the importance of a case based on the forecasts of precipitation exclusively. While hydrologic predictions can be sensitive to other weather variables such as temperature and melting level, these are likely to be second-order effects which will be ignored here to generate a benchmark solution. Furthermore, this paper will only deal with the construction of an optimal thinned sample based solely on the meteorological information; the hydrologic forecasting and validation is a separate subject that will be reported on in a future publication.

Our solution of the optimal subsampling problem posed above will build on the construction of a novel high-dimensional statistical model of precipitation able to incorporate, implicitly, a multitude of precipitation-producing systems over CONUS in a seamless way. Section 2 provides scientific background that illustrates our thought process in developing this novel statistical methodology to model precipitation and introduces our proposed case selection techniques. Our methodologies are described in detail and their performance is evaluated in sections 3 and 4, respectively. Section 5 contains a summary of the paper, as well as some discussion and outlook. Some of the more technical figures are placed in the supplemental material, which also includes a link to the datasets used or generated in this study. Finally, for readers’ convenience, the appendix provides the list of abbreviations used in the paper.

2. Background and proposed methodologies

a. Statistical downscaling and prediction of precipitation

Statistical prediction and downscaling methods for precipitation are based on the extensively studied association between extreme precipitation and recurrent large-scale meteorological patterns (LSMP), which provide favorable environment for smaller-scale processes often underlying the extreme precipitation events (although not all such events are tied to LSMP). Barlow et al. (2019) reviewed, among other things, the types of meteorological synoptic systems and mechanisms for extreme precipitation LSMPs for the North America region and found a great diversity of LSMPs depending on the geographical location and season. LSMPs are distinct from teleconnection patterns in that the LSMPs are conditioned on the occurrence of a specific event—for example, extreme precipitation—whereas classical teleconnections are not. The most intuitive way of defining the LSMP is through compositing, although a variety of other methods are available, including regression-based and cluster-analysis methods (Grotjahn et al. 2016). For example, Robertson et al. (2016) used K-means cluster analysis (Robertson and Ghil 1999) of the reanalysis wind data over North America to identify seven distinct large-scale circulation types and tie some of them to enhanced probability of springtime flooding events in the Midwest of the United States. We note here that while identifying a small subset of large-scale recurrent patterns to classify precipitation-producing weather states is an attractive methodology, it is at odds with the extreme-precipitation LSMPs’ great diversity mentioned above. Hence, the practical utility of such methodologies to downscale precipitation over a large domain such as CONUS is likely to be limited.

Classical regression approaches such as canonical correlation analysis (CCA: Wilks 2011) also have a limited applicability to short-term precipitation modeling due to non-Gaussian and intermittent nature of precipitation. However, they may be suitable and have been utilized for the prediction of *seasonal* rainfall both directly (Sinha et al. 2013) and as an auxiliary tool for selecting external predictors in conjunction with alternative methodologies (Holsclaw et al. 2016). The most widely used class of the latter alternative methods for statistical modeling, downscaling and prediction of precipitation involves, in one way or another, generalized linear models (GLM; McCullagh and Nelder 1989). GLMs are an extension of classical linear regression models to simulate the conditional expectation of a non-Gaussian distributed variable—such as precipitation—as a function of external predictors referred to as exogenous variables. The latter variables are associated with nonstationary forcing related to the climate variability external to the climate subsystem of interest or, of most relevance to the present discussion, with the occurrence of LSMPs. The GLM models are typically constructed to estimate probability of daily precipitation at a grid point or weather station level (e.g., Furrer and Katz 2007), although some generalizations to multiple stations accounting for spatial correlations between them are also available (Kenabatho et al. 2012). Manzanas et al. (2018) fitted separate GLM models to downscale daily precipitation occurrence and, separately, daily precipitation amount at each grid cell using upper-air predictors simulated by multimodel seasonal climate hindcasts over the Philippines. They showed that this methodology can yield a significant forecast skill improvement for seasonal precipitation prediction over that of raw forecasts in cases where the dynamical model predicts large-scale exogenous variables better than it predicts the precipitation itself.

An alternative approach to precipitation modeling over a spatially extended array of grid points or stations—a hidden Markov model (HMM) approach—assumes the existence of a few discrete "hidden" weather states that capture spatial dependencies of rainfall probabilities within the region considered. The Markovian daily transitions between these states are tied to exogenous predictors via GLM regression; in the latter case these models are referred to as nonhomogeneous hidden Markov models (NHMM; Robertson et al. 2004). Holsclaw et al. (2016) developed a combined HMM-GLM approach, in which a weather state HMM model is complemented by a GLM model that can modify individual hidden states at a station level in response to external predictors, rather than the probabilities of transitions between fixed states, as in the traditional NHMM. We speculate that this approach would also be challenging to adapt for faithful modeling of extreme precipitation over the entire CONUS, where, once again, the heaviest tails of local precipitation distributions are associated with a multitude of precipitation producing systems (Barlow et al. 2019), rather than with a small number of weather states and/or exogenous predictors.

b. *Present approaches*

To summarize the above discussion, neither classical linear regression-based methods nor clustering or HMM methods are directly suitable for statistical modeling and prediction of precipitation over the entirety of CONUS due to non-Gaussian and intermittent nature of precipitation and a great diversity of precipitation-producing systems/mechanisms in this region, respectively. GLM regression methods may work at a gridpoint level but will still require the choice of exogenous dynamical variables based on a subjective zoning of the area; these methods are also incompatible with automated linear regularization and predictor-selection techniques such as CCA or closely related partial least squares methods (PLS; Wold et al. 1984).

Here we address these difficulties via a new methodology based on statistical modeling of the so-called pseudo-precipitation (PP) field, which uses column integrated water vapor saturation deficit as a negative complement to precipitation (Yuan et al. 2019). Pseudo-precipitation is thus characterized by a more symmetric distribution than the actual precipitation, opening up a possibility of utilizing standard linear regression methods for its modeling. Furthermore, in contrast to classical precipitation field, pseudo-precipitation patterns provide, additionally, information on both the synoptic-scale and anisotropic mesoscale environment, including LSMPs, in which local precipitation occurs, making it ideally suited for linear inverse modeling (LIM; Penland 1996; Penland and Sardeshmukh 1995) and related data-driven modeling methodologies (Kravtsov et al. 2005, 2010, 2016, 2017). The LIMs exhibit subseasonal forecasts skill comparable to that of state-of-the-art numerical weather prediction (NWP) models (see e.g., Winkler et al. 2001) and, most importantly, are able to isolate initial states associated with useful predictability of its own, as well as of NWP-model based forecasts (Newman et al. 2003; Albers and Newman 2019). This property can be helpful for identifying potentially predictable high-impact precipitation events—the main focus of the present study. The proof-of-concept mesoscale-resolving regional inverse models of surface temperature over CONUS have been developed and tested before (Kravtsov et al. 2017); these models are complex enough (yet numerically efficient) to provide an overarching description and forecast utilization of LSMPs associated with local weather extremes. We expect the same statement to be true for the combined surface temperature/pseudo-precipitation modeling we propose here.

In addition to the above purely statistical and numerically efficient methodology, we will also develop and test a procedure for selecting an optimal thinned subsample of representative dates by utilizing the GEFSv12 reforecasts of precipitation for the 2000–19 period. This procedure would allow one to conduct a greatly reduced number of hydrologic hindcasts to estimate the adequacy of the reduced sample for the postprocessing, validation and end-user needs. However, it is much more computationally demanding than the proposed purely data-driven methodology insofar as it still requires, in the first place, the full-blown meteorological reforecasts of the entire climate state to determine the thinned subsample, which somewhat defies the purpose of data thinning. Full, everyday reforecasts were available for the GEFS versions 10 (Hamill et al. 2013) and 12 (Guan et al. 2022), but such full records may not be available in the future to be subsampled. Yet, the present dynamical/statistical ad hoc algorithm based on the Global Ensemble Forecast System version 12 (GEFSv12; Hamill et al. 2022; Zhou et al. 2022) reforecasts can be considered a control against which to evaluate our main statistical modeling methodology, and, in what follows, we describe this algorithm first.

3. Datasets and methodological details

a. *Selecting reforecast case dates based on heavy precipitation in GEFSv12 reforecasts*

The GEFSv12 reforecasts are retrospective forecasts on a global 0.25° grid, spanning the period 2000–19; they are 5-member ensemble reforecasts generated once per day, from 0000 UTC initial conditions. We argue here that a metric of an event's extremeness should be based on precipitation magnitude as opposed to, say, the quantile of today's forecast relative to its climatological distribution (e.g., a 0.1-in. forecast in the desert may be an extreme event relative to the local climatology but still of marginal significance to hydrologic applications). In the present methodology, the importance of a case for potential selection was judged based on the 0–10-day total GEFSv12 ensemble-mean reforecast precipitation P_{10}, sampled daily over the 2000–19 period for each of the 18 enclosed CONUS regions associated with distinct 2-digit Hydrologic Unit Codes (HUC-2 units: https://nas.er.usgs.gov/hucs.aspx). Some case choices were based on large ensemble-mean precipitation averaged over the entire HUC-2 unit, while others were optimized on the top 20% of grid points inside that HUC-2 unit (at the 0.25° resolution) to emphasize smaller-scale impactful events. A small number of cases were also based on large CONUS-wide ensemble-mean precipitation. More specifically, the subjectively chosen breakdown of cases was as follows:

1) 30% of the total cases were optimized based on the maximum 10-day ensemble mean precipitation in that HUC-2 unit. After choosing a case day on this criterion, an ad hoc de-weighting of the day before and the day after was applied so they are less likely to be chosen. However, we find that the algorithm often chooses case days separated by at least 2 days (which can be easily adjusted if desired).
2) 60% of the total cases are optimized based on the maximum 10-day ensemble-mean precipitation at a smaller number of grid points within the HUC. Specifically, for each HUC, precipitation means are sorted at grid points within the HUC, lowest to highest. The mean of the highest 20 sorted points is calculated, and the case days are selected based on the dates with the highest mean values for those 20 points. As with 1 above, an ad hoc de-weighting factor is applied after the selection of a particular case day to surrounding days to make that date less likely to be chosen.
3) The remaining 10% are chosen based on maximal CONUS-averaged ensemble-mean precipitation.

These choices were admittedly arbitrary, but were intended to balance cases selected for impact across large river basins, for flooding in smaller sub-basins, and for overall impact that may cross basins. De-weighting of adjacent days was applied to guard against a tendency for case days to gather around a few significant events, thereby providing less independent samples.

In developing the above merged set of dates from across the subdomains, we chose the first case date from each subdomain unless it was a repeat. Then we proceeded to the second ordered case date in each subdomain, the third, and so forth, until we have reached n total cases, where n is an adjustable predetermined size of the thinned sample. The lists of presumed important cases were developed separately for the warm (April–September) and cool season (October–March), with $n = 520$, or 1 sample per week.

The resulting procedure produces a list of dates with an irregular sampling in time, which is to be expected if there exist long periods with no hydrologically significant activity, which the algorithm aims to skip to provide more samples when there is strong forcing. The clustering around the largest storms from multiple initial conditions/issued date–times is controlled, to an extent, by our de-weighting procedure. This procedure involves a trade-off: on the one hand, we do not want a lot of shared information between samples; on the other hand, we do want to sample the largest events from several issued date–times and, hence, lead durations. Other adjustable parameters include the total number of cases n and the proportions of cases associated with each of the case categories (1, 2, and 3) above.

We will hereafter refer to the thinned sample produced by the above procedure as sample$_A$; illustrative examples from this sample will be presented alongside with the results from our alternative, purely data-driven methodology presented below.

b. Selecting reforecast cases using an EMR statistical model

1) Datasets and variables: Introducing pseudo-precipitation

We analyzed data from the National Centers for Environmental Prediction North American Regional Reanalysis (NARR) (http://www.esrl.noaa.gov/psd/data/gridded/data.narr.html); Mesinger et al. (2006), using daily "observations" on a 349 × 277 grid with nominal horizontal resolution of 32 km and 29 pressure levels, over the 1979–2020 period; about a third of these data are from locations over land, leading to ~30000 data points for each day throughout the 42-yr period and for a single-level field. The NARR dataset has been widely used in the climate downscaling community (see Zobel et al. 2018 and references therein). Bukovsky and Karoly (2007) found that NARR provides faithful estimates of the observed precipitation over CONUS, although some biases exist over Canada due to a relatively poor quality of the assimilated data there.

We utilized NARR datasets for the (daily) accumulated precipitation Pr and 2-m air temperature T_a. We also used the air temperature T and specific humidity Q data at all available pressure levels to compute the *air dryness D* related to the column-integrated water vapor saturation deficit (Yuan et al. 2019). In an air column of area δA, the mass of water vapor δm to be added to achieve saturation throughout the column is

$$\delta m = -\delta A\int(\rho_v - \rho_{v,s})dz = \delta A\int(\rho_v - \rho_{v,s})\frac{dp}{\rho g}$$
$$= \frac{\delta A}{g}\int(Q - Q_s)dp. \qquad (1)$$

Here z is the geometric height, p is the pressure, ρ and ρ_v are the dry air and water vapor densities, respectively, the subscript s denotes the quantities for saturated air and $g = 9.82$ m s^{-2} is the gravity acceleration. The specific humidity of saturated air Q_s can be computed as (Bolton 1980) follows:

$$Q_s = \varepsilon\frac{e_s}{p};\ e_s = 6.112\ \exp\left(\frac{17.67T}{T + 243.5}\right), \qquad (2)$$

where $\varepsilon = 0.62198$ is the ratio of the molecular weights of water and dry air, e_s is the saturation water vapor pressure, and air temperature T is expressed in degrees Celsius (°C).

Air dryness D is defined as the equivalent water depth associated with the quantity δm in (1):

$$D = -\frac{\delta m}{\rho_w \delta A} = -\frac{1}{\rho_w g}\int (Q - Q_s)dp, \qquad (3)$$

where $\rho_w = 1000\ \mathrm{kg\ m^{-3}}$ is water density. The air dryness in (3) can be thought of as a negative complement to precipitation and used to construct the so-called pseudo-precipitation field PP, which is, here, equal to the actual precipitation Pr if Pr > 0.001 m day^{-1} or to Pr + D (essentially, the air dryness D) otherwise.

The PP field incorporates the information about both precipitation, which can exhibit small-scale intermittent structures, and multiscale synoptic environment (see Fig. 1); it thus provides a promising, yet unexplored way to characterize and predict, statistically, wet and dry weather conditions. One of its attractive features is that the distribution of PP, unlike that of Pr, is a single-mode, two-tailed distribution, which makes PP more similar to other dynamical and thermodynamic variables describing atmospheric state. *This opens up a possibility for using standard methodologies developed previously for temperature and flow-field analysis and modeling (CCA, LIMs) to analyze and model pseudo-precipitation PP and, hence, the actual precipitation Pr, which equals pseudo-precipitation when PP > 0 and zero otherwise.*

2) EMR MODELING OF PRECIPITATION

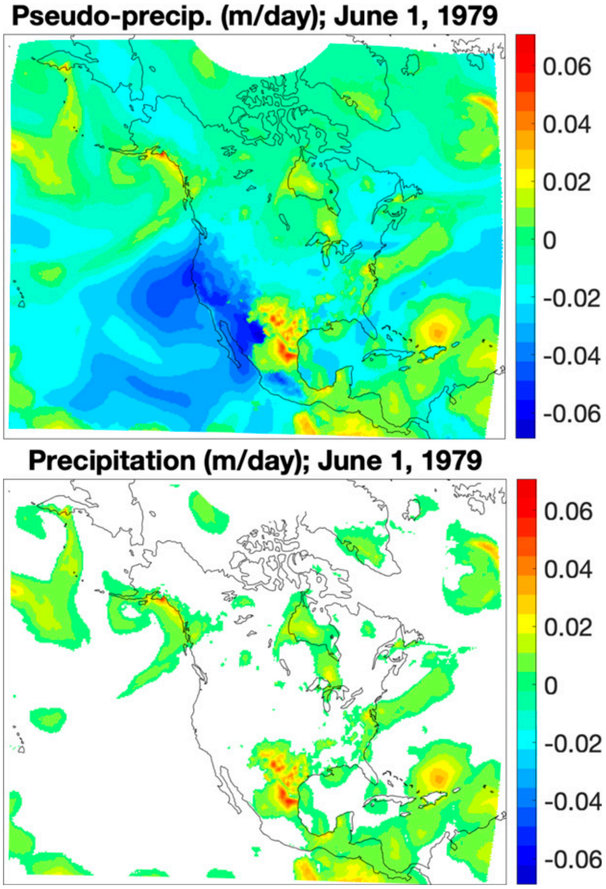
We here apply advanced methods for high-dimensional statistical data modeling to identify potentially predictable large/extreme precipitation events. This idea is rooted in the demonstrated ability of a subclass of such inverse models—LIM models (section 2b) to "forecast the forecast skill" (Albers and Newman 2019).

(i) General methodology

The empirical model reduction (EMR; Kravtsov et al. 2005, 2010, 2016, 2017) is a generalization of LIM data modeling methodology to incorporate memory effects in the postulated parametric form of this empirical model's evolution operator. The model construction usually takes place in a reduced phase space associated, for example, with L leading empirical orthogonal functions (EOFs) of the fields simulated, in which case the state of the system on a given day is described by the L-valued vector of PCs $\mathbf{x}$. The EMR emulator models the evolution of PCs using the following multilevel form:

$$\begin{aligned} d\mathbf{x} &= \mathbf{x}\cdot\mathbf{A}^{(1)} + \mathbf{r}^{(1)},\\ d\mathbf{r}^{(1)} &= \left[\mathbf{r}^{(1)}\ \mathbf{x}\right]\cdot\mathbf{A}^{(2)} + \mathbf{r}^{(2)},\\ d\mathbf{r}^{(2)} &= \left[\mathbf{r}^{(2)}\ \mathbf{r}^{(1)}\ \mathbf{x}\right]\cdot\mathbf{A}^{(3)} + \mathbf{r}^{(3)}, \end{aligned} \qquad (4)$$

where the differentials on the left-hand side denote the daily increments of the corresponding variables. The first model level in isolation, with the residual $\mathbf{r}^{(1)}$ represented, at the simulation stage (see below), by the spatially correlated white noise, would make up a classical LIM model; for example, its 1D analog would be the AR-1 red-noise model widely used to test for statistical significance of spectral peaks in a time series. Instead, in the EMR modeling, daily increments of the first-level residual $d\mathbf{r}^{(1)}$ are in turn modeled as a linear function of the extended predictor vector $[\mathbf{r}^{(1)}\ \mathbf{x}]$ to form the second level of the multilevel regression model (4). In the same way, the third level connects the daily increments of the second-level residual $d\mathbf{r}^{(2)}$ and the extended predictor vector $[\mathbf{r}^{(2)}\ \mathbf{r}^{(1)}\ \mathbf{x}]$ involving the variables from the previous two model levels.


Pseudo-precip. (m/day); June 1, 1979

Precipitation (m/day); June 1, 1979

FIG. 1. (top) Pseudo-precipitation (PP), as well as (bottom) precipitation Pr on 1 Jun 1979, derived from the NARR reanalysis (m). White areas in the bottom plot are either outside of the NARR domain or, otherwise, have zero Pr.

The matrices of the model coefficients $\mathbf{A}$ and the level residuals are found by a regularized multiple linear regression (MLR) and depend on the seasonal cycle at the monthly resolution. While the residuals of the first and second level may involve serial correlations, the last level's residual $\mathbf{r}^{(3)}$ is typically white in time (otherwise, additional levels can be added). Note that while the model construction procedure is sequential from the first level down to the last level, the Eqs. (4)—when rewritten as one equation containing the time-lagged variables—are formally equivalent to the autoregressive moving average model (ARMA: Box et al. 1994).

The model (4) can provide independent realizations of observations that are statistically very similar to the input data. At this stage of model simulation, the residual forcing at the third model level $\mathbf{r}^{(3)}$ is replaced by a random forcing, which can involve simultaneous or lagged spatial correlations between different PC "channels" and depend on the simulated state $\mathbf{x}$ (effectively *making the model nonlinear*). One can also use the EMR model for statistical forecasting of the out-of-sample data. Trivial linear transformation of the simulated PCs provides the data simulation or forecasts in the original physical space.

While the original LIM models, as well as the EMR methodology above, have been typically applied to fairly low-dimensional subsets of meteorological data, Kravtsov et al. (2016, 2017) demonstrated its applicability to larger or higher-resolution datasets such as regional surface temperature (Kravtsov et al. 2017) and precipitation. In the latter case, most relevant to the present project, the EMR modeling of combined T_a and PP fields resulting from an hourly, 16-km-resolution Japan regional reanalysis was successfully used by AIR Worldwide (Boston, Massachusetts) for flood-risk assessment over Japan (B. Dodov 2016, Director of Flood Modeling, personal communication). We here build an analogous combined T_a/PP daily EMR model over CONUS and use it to identify potentially predictable large and extreme precipitation events to be included in the final thinned subsample.

(ii) EMR application to NARR T_a/PP data

All model construction steps, including the identification of seasonal cycle and initial data compression, were done using the NARR's 1979–99 (training period) data. We built our EMR model (4) in the phase space of 3000 common EOFs of the daily 2-m air temperature and pseudo-precipitation [section 3b(1)] anomalies with respect to the mean seasonal cycle computed by the linear regression of raw daily data onto the first five harmonics of the annual cycle. The maps of climatological standard deviation of these anomalies (over the 1979–99 period) are shown in supplemental Fig. S1. The EOF identification only used land grid points (hence, the assessment of model performance should in principle also focus on the land region). We first computed 1000 leading EOFs of T_a and 3000 leading EOFs of PP field, normalized the corresponding individual PCs by the standard deviation of the leading PC of each field and applied an additional EOF rotation to the dataset of concatenated T_a/PP individual normalized PCs, finally retaining the leading 3000 common PCs so obtained. These PCs were again normalized by the standard deviation of their own PC-1, while the corresponding dimensional EOF patterns were found by regressing the individual fields onto these common PCs. Note that these patterns only represent the actual common EOFs over the land region and should be interpreted as a teleconnection pattern over ocean.

To initialize model forecasts performed over the validation period (2000–20), we projected the anomaly data there (again, with respect to the 1979–99 mean seasonal cycle) onto common T_a/PP EOFs computed above. For the back transformation, to produce the patterns in physical space from a map of individual-day PC loadings, as obtained, for example, from our EMR model simulations, one is to simply add all of the 3000 individual EOF patterns multiplied by the corresponding loadings, on top of the mean seasonal cycle. The EOF truncation errors associated with the procedure above are shown in the supplemental Figs. S2 (training period) and S3 (validation interval) and demonstrate a fairly high accuracy (small errors) over CONUS for both T_a and PP data, sufficient for a faithful representation of extreme hydroclimatic events in the region.

The EMR model construction and simulation technical steps follow Kravtsov et al. (2017), except here we are only modeling the evolution of daily fields and thus disregard the subdaily and monthly model tiers employed there. Note that all of the model operators in (4) are season-dependent at monthly resolution. For example, to estimate the model parameters for January, we consider the December–January–February (DJF) subset of daily data and use a regularized PLS version of multiple linear regression for each of the three model levels sequentially. At the simulation stage, the third-level residual $\mathbf{r}^{(3)}$ is simulated by pulling its randomized 5-day snippets from the library of actual residuals obtained during the model construction stage. This random forcing selection is also season-dependent, so that, for example, if the current time step is in January, the DJF subset of $\mathbf{r}^{(3)}$ library is used for that purpose. To avoid unnecessary discontinuities, the consecutive random forcing snippets were overlapped by two days and added with the weights $(\sqrt{3}/2, 1/2)$ and $(1/2, \sqrt{3}/2)$ before phasing out the previous snippet of $\mathbf{r}^{(3)}$ completely.

We used the EMR model above in two ways: first to produce, from random initial conditions, 100 synthetic realizations of the 2-m air temperature and precipitation 1979–99 evolution and assess how well the model captures the observed statistical characteristics of these fields (section 4a). Second, we ran 0–10-day 100-member ensemble forecast of temperature and (pseudo)precipitation for each of the 2000–20 initial conditions to assess the model's predictive skill (section 4b) and eventually utilized these forecasts to develop and test an innovative methodology for reforecast thinning (section 4c). Since our interest here is in extreme precipitation events, we will focus below on the simulation of precipitation; the present EMR performance in modeling temperature will be considered elsewhere.

3) Case selection using EMR ensemble forecasts

In principle, the EMR ensemble-mean hindcasts of the 0–10-day total precipitation P_{10} can be processed in exactly the same way as the GEFSv12 reforecasts to produce an alternative representative subset of events of impact, as described in section 3a; the outcome of such a procedure, which results in the thinned sample we will refer to as sample$_{B,}$ are briefly discussed at the very end of section 4c. However, a large size of the EMR hindcast ensemble, achievable due to this model's numerical efficiency, makes it possible to develop an alternative methodology that involves relative entropy of the EMR hindcasts; this methodology is described in detail in section 4c. We will call the thinned sample produced by this EMR based method simply a "sample" or an "EMR-RE sample." Note, once again, that the RE methodology requires a large number of ensemble members to compute the distributions of the forecasted fields such as precipitation with sufficient accuracy. It is, therefore, cannot be applied to 5-member GEFSv12 ensemble reforecasts.

4. Results

a. *Using the EMR model as an emulator of daily precipitation evolution*

Preliminary inspection of the EMR-model simulations of daily precipitation Pr—which equals PP when PP > 0 and is zero otherwise—reveals model biases in the distribution of precipitation events (not shown). To eliminate these biases, we apply quantile mapping (for each of the DJF, MAM, JJA, and SON seasonal subsets) to each 1979–99 model simulation of *pseudo-precipitation* to make the simulated local distributions of this quantity identical to those based on the original 1979–99 NARR data. Specifically, the observed 1979–99 and simulated 2000–20 PP time series at a given grid point and for a given season (DJF, MAM, JJA, and SON) were sorted in the ascending order, upon which the sorted 2000–20 simulated values were replaced by the sorted 1979–99 observed values, then put back in the original order (cf. Hamill 2018). This procedure automatically ensures the identical local (i.e., a given grid point's) *precipitation* distributions between the model and NARR reanalysis as well. However, the spatiotemporal characteristics of sequences of daily precipitation maps are entirely due to dynamics embedded in the EMR model's propagator. Examples of such sequences for the warm and cold season are shown in Fig. 2 and supplemental Fig. S4, respectively and give one a visual impression of how well the model matches the space–time structure of the observed stationary and propagating precipitation patterns; the external link to longer sequences is also available in the supplemental information.

We also compute, for future use, daily time series of day 0–10 cumulative precipitation (P_{10}) and display its seasonal means and 99th percentiles in Figs. 3 and 4, respectively. Note that while the simulated local *daily* precipitation distributions are fixed due to quantile mapping, the simulated and observed distributions of P_{10} can be different if the spatial scales or persistence/intermittency of the simulated precipitation differ from the observed characteristics. However, this does not seem to be the case here: the simulated P_{10} seasonal cycle is entirely consistent with observations (Fig. 3); for example, the southeastern United States experiences highest rainfall totals in JJA, while the northwestern United States—in DJF. The EMR simulated P_{10} biases are small, under 2 mm over the vast majority of grid points, and, on average, are at 2%–3% of climatology (depending on the season) over land (and rarely exceeding 10% of climatology at any given location). We note here again that the EMR model is only trained to accurately model precipitation over land, where the T_a/PP EOFs were computed and is thus not expected to perform equally well over ocean. Statistically significant P_{10} climatological differences indeed only occur at a small number of grid points either over the ocean or, otherwise, in locations with small precipitation totals.

The simulated P_{10}'s 99th percentile (Fig. 4) is an overestimate compared to observations, including large areas over land, so the EMR model does tend to overpredict the magnitude of extreme events. This reflects, perhaps, overly persistent local precipitation anomalies, but the overall match between the simulated and observed P_{10} distributions is still fairly reasonable: the P_{10} 99th percentile patterns are captured extremely well, with spatial correlation coefficient between the EM based and reanalysis patterns exceeding 0.95 for each season, while the seasonal biases in the P_{10} 99th percentile magnitude over land remain under 25% of its climatology on average (and rarely exceed 50% of the climatology at any given grid point).

b. *EMR model predictive skill*

To initialize the EMR model forecasts starting from a given day n within the 2000–20 validation interval, we assume that the observable state vectors $\mathbf{x}$ at days n, $n - 1$, and $n - 2$ are all known. This, however, still requires us to solve for the values of the hidden-level variables $\mathbf{r}^{(1)}$ and $\mathbf{r}^{(2)}$ at the initial day n, which involves two pre-steps of the model (4) driven by a random $\mathbf{r}^{(3)}$ forcing that ensure dynamical consistency [within the model (4)] of the hidden-state variables with the observables $\mathbf{x}_n$, $\mathbf{x}_{n-1}$, $\mathbf{x}_{n-2}$. After these pre-steps, the model is integrated forward in a normal way until the time $n + 10$. This procedure is repeated for all of the available initial conditions. Upon transformation back to physical space, the collection of PP forecasts for a given lead time is, again, *quantile mapped* to the 1979–99 local daily PP distributions; finally, zeroing out the negative values of this quantile mapped PP forecast gives the final forecast of the daily precipitation at this lead time, for each initial condition. Summing up the precipitation forecasts for the days n to $n + 10$ makes up the final P_{10} forecast for each initial condition; we produced an ensemble of 100 such forecasts under different realizations of the random forcing. Below we will focus on these P_{10} forecasts when estimating the EMR model's forecast skill.

We will also compare the EMR model forecasts with the benchmark damped persistence forecasts of daily precipitation:

$$p_{n+m} = r_m p_n + (1 - r_m)\overline{p}, \tag{5}$$

where r_m is the precipitation's lag-m autocorrelation and $\overline{p}$ is the climatology, both computed for each season's subset of the 1979–99 NARR's daily precipitation data. The damped persistence P_{10} forecasts are obtained from (5) as the sum of p_{n+m} for $m = \overline{0, 10}$.

1) DETERMINISTIC SKILL

We first discuss some traditional deterministic measures of skill by comparing the observed P_{10} values with their ensemble-mean EMR based prediction. Figure 5 provides cool-season examples of such a comparison for select cases of substantial observed P_{10} episodes over CONUS (see Fig. S5 for analogous warm-season comparisons). These cases were arbitrarily selected from the subset of "cases-of-interest" (sample$_A$) identified by the GEFSv12-based methodology described in section 3a. Visual inspection confirms reasonable EMR

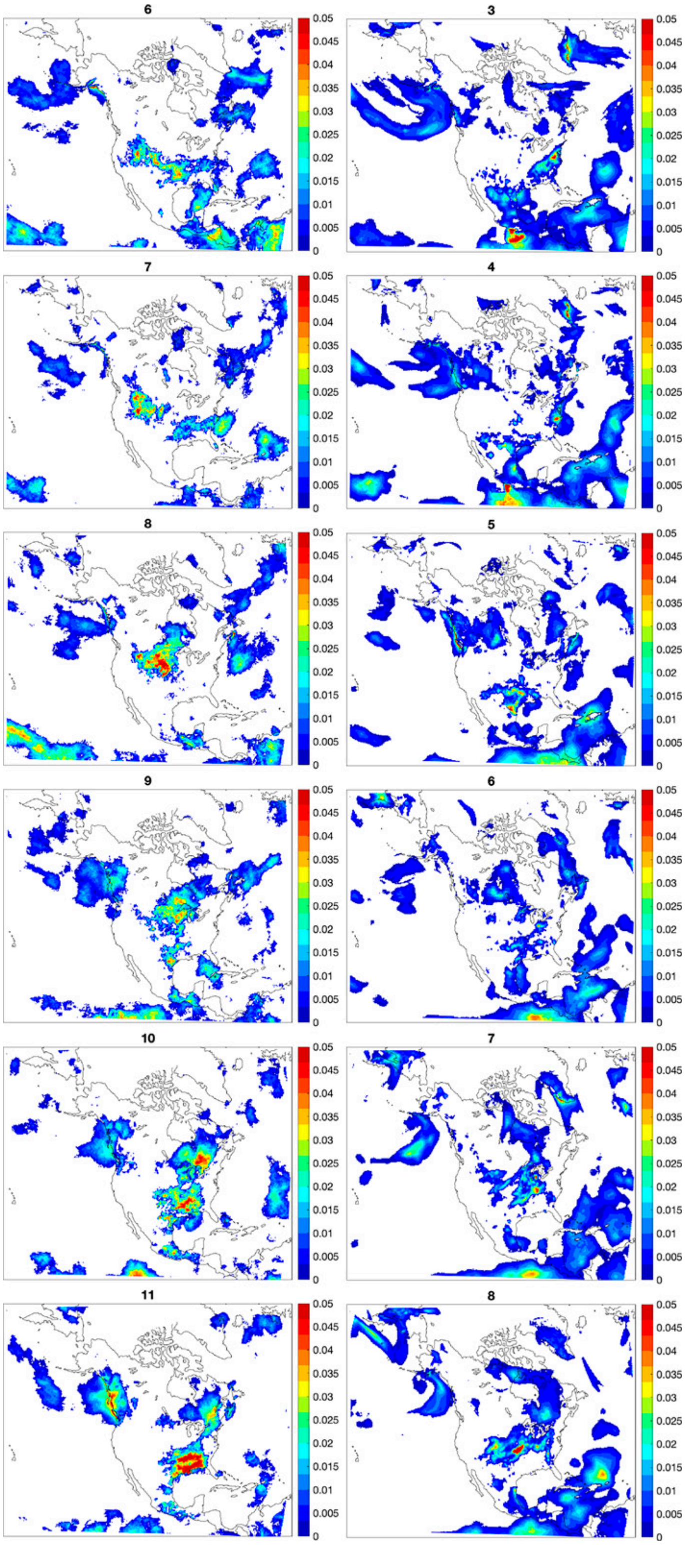


FIG. 2. A JJA-season sequence of daily surface precipitation maps (m) from (left) arbitrary (random) realization of EMR model and (right) NARR reanalysis. Day “1” in a panel caption would correspond to 1 Jun 1979. White areas in the bottom plot are either outside of the NARR domain or, otherwise, have zero Pr.

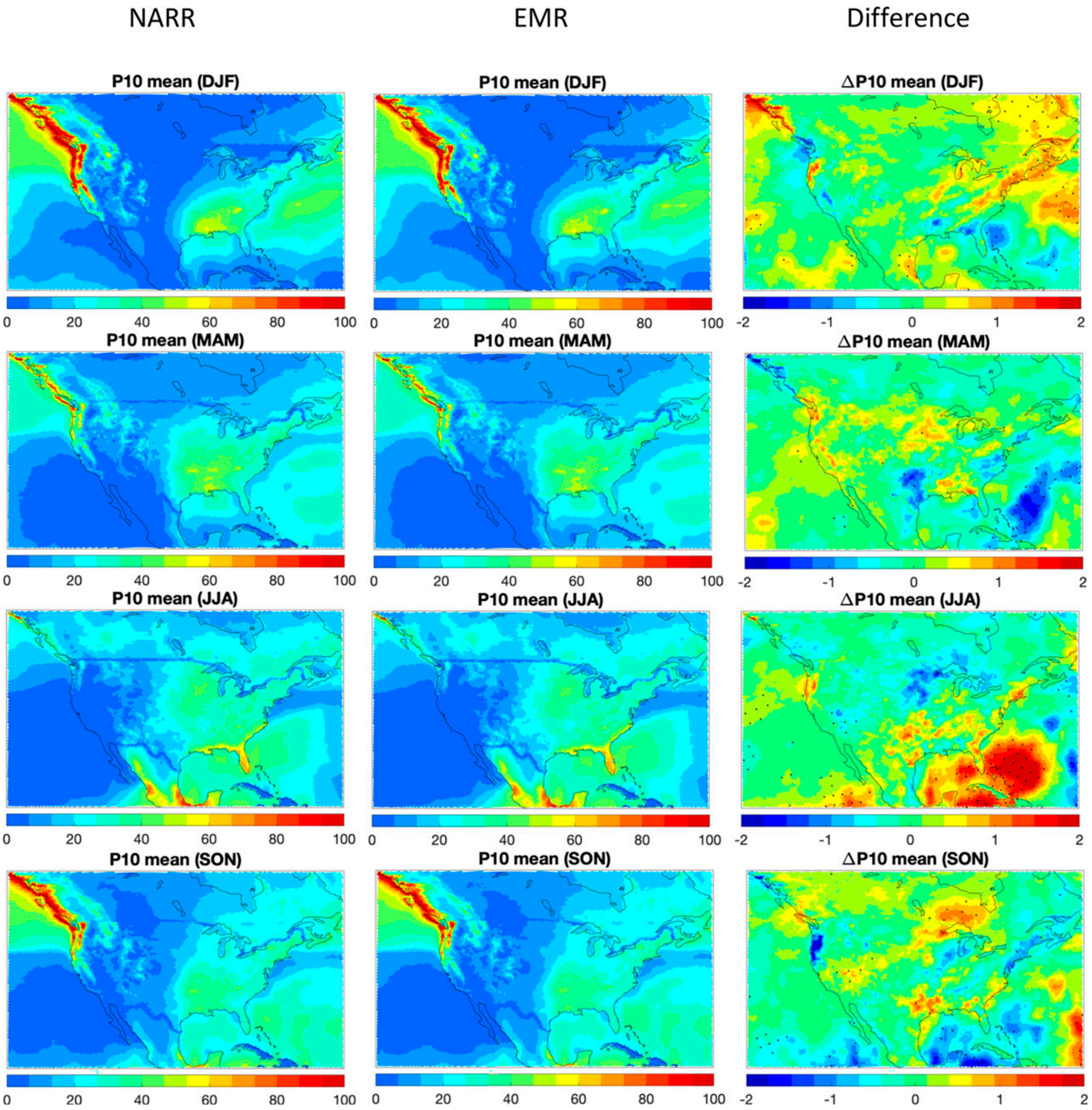


FIG. 3. The 1979–99 seasonal climatology of the 0–10-day total precipitation at the surface—P_{10} (mm). (left) Climatology based on NARR reanalysis; (center) climatology based on an ensemble of 100 EMR model simulations; and (right) the difference between the simulated and NARR based P_{10} climatology, with stippling indicating the regions over which this difference is of the same sign for more than 97 realizations (so, effectively, it is statistically significant at the 5% level).

forecasts (left column) of the spatial scale, shape, location, and magnitude of the observed large P_{10} events (center column), qualitatively and quantitatively similar to analogous GEFSv12 forecasts (right column). For example, the EMR model outperforms the GEFSv12 forecast of P_{10} along the Pacific coast on 11 January 2000 (second row, black box) in better capturing both the pattern and magnitude of precipitation there, whereas GEFSv12 forecast overpredicts/underpredicts precipitation in the southern/northern part of this region, respectively. Both prediction systems over the same region on 26 November 2001 (third row, black box) show a mixed performance: the EMR model captures the pattern, but underpredicts the magnitude of precipitation, while GEFSv12 forecast is much better in terms of the magnitude in the southern part of the region but is essentially the same with the EMR forecast in the northern part of the region, also under-predicting the magnitude of the observed P_{10}. At the same time, the GEFSv12 forecast over

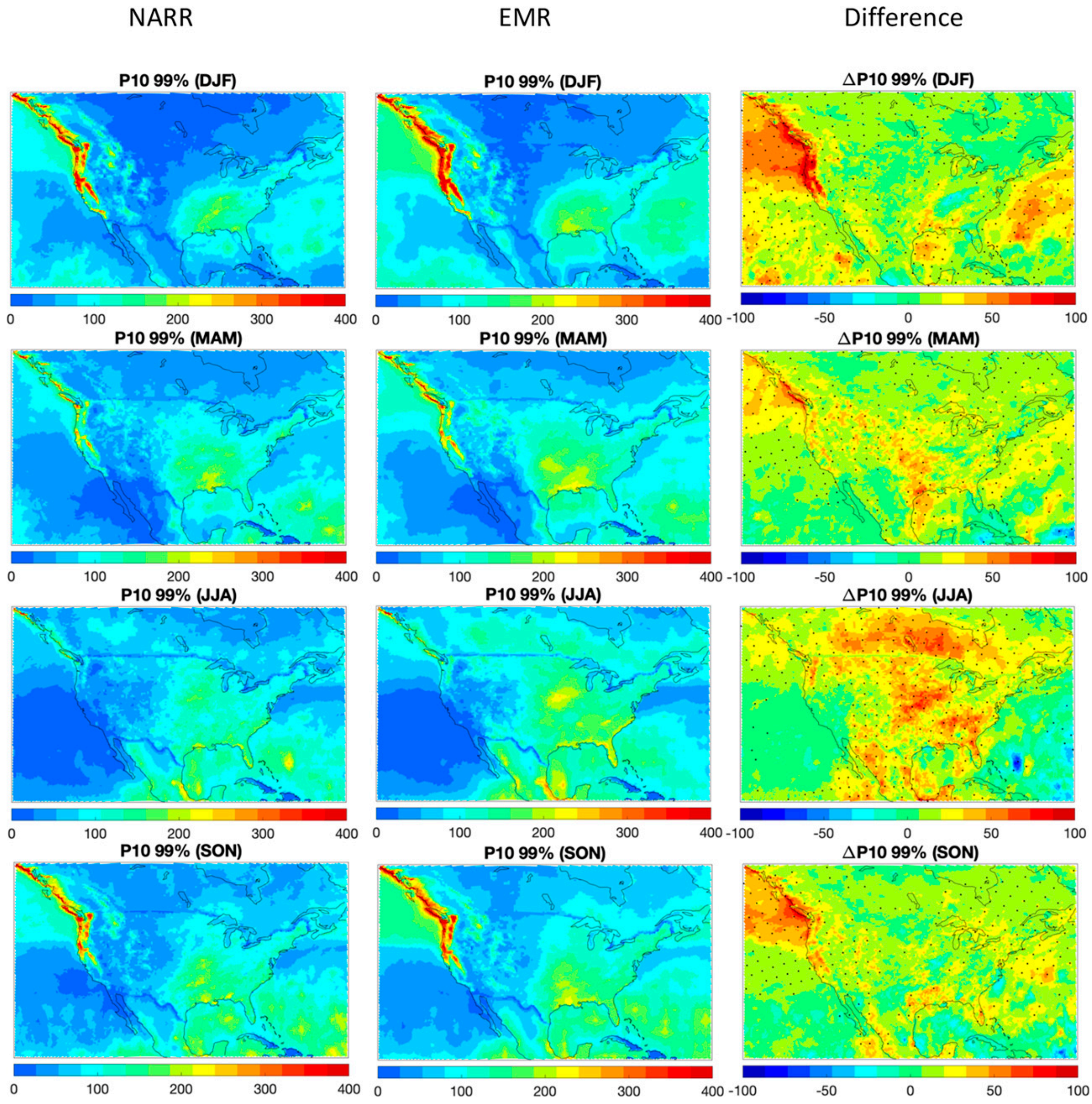


FIG. 4. As in Fig. 3, but for the 99th percentile of P_{10}.

the southeastern United States (third row, red box) successfully reproduces a band of heavy precipitation there, which the EMR model completely misses. A few days later though (fourth row, red box), the EMR model matches the magnitude of precipitation better than the GEfSv12 forecast, which overpredicts the maximum precipitation by as much as 50 mm.

Despite a reasonable case-by-case performance, the overall correlations between the observed and forecasted P_{10} time series (for each season) (Fig. 6, left), while positive, are fairly low, at the 0.2–0.3 level in most areas, with the exception of a few season-dependent regions reaching potentially useful levels of 0.5–0.6, including the northwestern United States in MAM/SON, and the monsoon region covering Arizona–New Mexico in JJA. However, these correlations, even when small, are consistently higher than those for the damped-persistence forecasts (Fig. 6, right).

The root-mean-square (rms) distance between the observed and forecasted P_{10} time series (Fig. 7) is generally close to the P_{10}'s climatological standard deviation (so that the ratio of the two shown in Fig. 7 is close to 1), with EMR model forecasts outperforming damped persistence forecasts in some of the southern areas—in particular, over regions across Mexico except for the DJF season—as well as over the northwestern

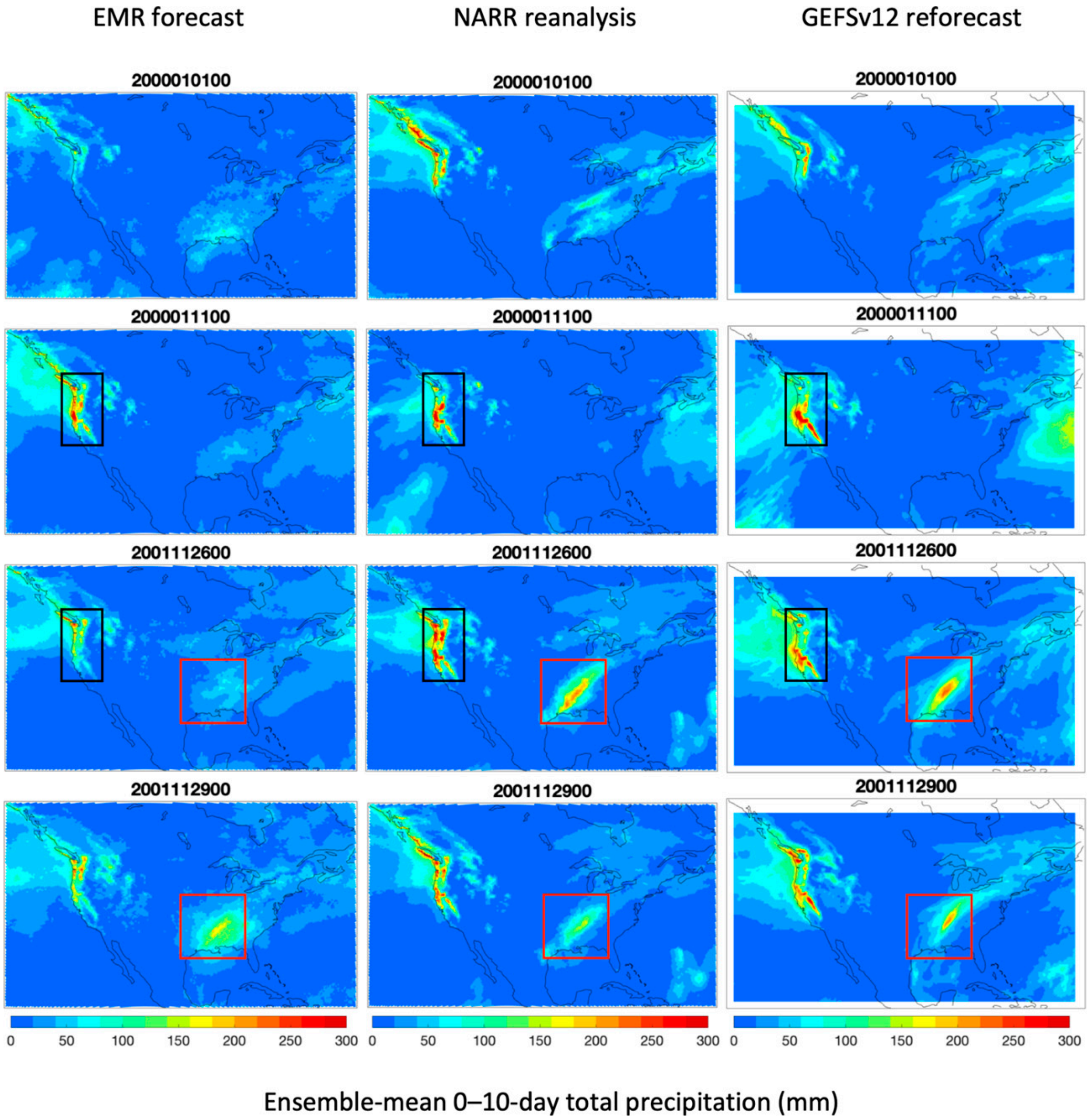


FIG. 5. Examples of (cool season) P_{10} forecasts (mm) using (left) EMR model and (right) GEFSv12 system, along with the (center) actual P_{10} maps based on NARR reanalysis. The forecast initialization time (the same across each row) is shown in panel titles in the YYYYMMDDHH format. The black and red rectangles indicate regions further discussed in text.

United States in MAM and SON, consistent with Fig. 6, but performing similar to damped persistence forecasts elsewhere.

Overall, the deterministic measures of skill suggest, at best, a modest performance of the EMR model in forecasting P_{10}. This, however, may be in part due to unsuitability of these measures to describe the forecast quality of a discontinuous and highly intermittent—in space and time—state variable such as precipitation. In particular, considering the ensemble-mean forecast only completely disregards much of the useful information associated with the entire ensemble of forecasts.

2) PROBABILISTIC CHARACTERISTICS OF SKILL

A perhaps more suitable measure of skill for precipitation should involve probabilistic characteristics associated with ensemble forecasts of this quantity. An example of such a measure is shown in Fig. 8, which plots the climatological ratio of the interquartile range (IQR) of the EMR model forecasts to the climatological IQR of P_{10}. This quantity is related to

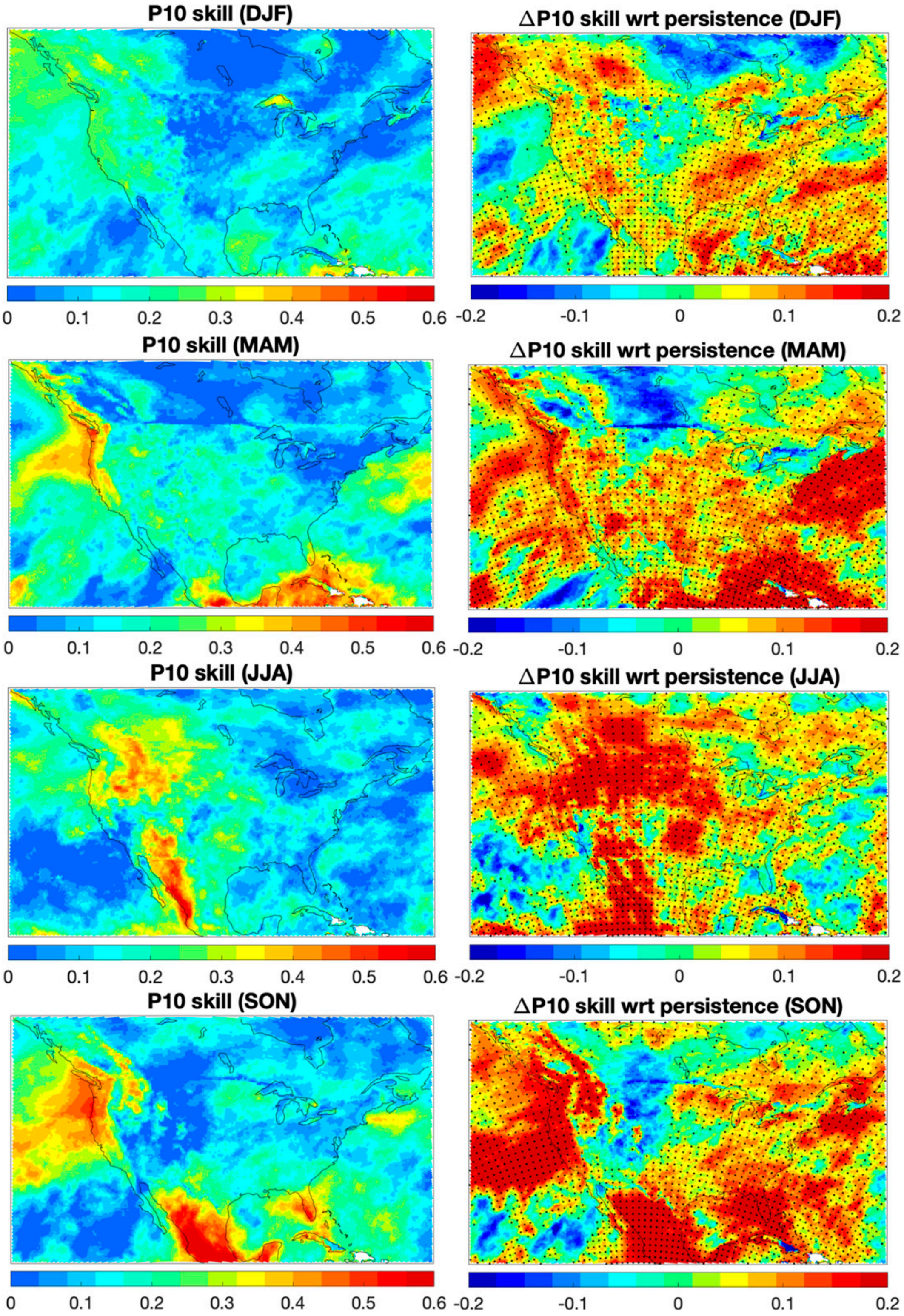

P10 skill (DJF)
ΔP10 skill wrt persistence (DJF)
P10 skill (MAM)
ΔP10 skill wrt persistence (MAM)
P10 skill (JJA)
ΔP10 skill wrt persistence (JJA)
P10 skill (SON)
ΔP10 skill wrt persistence (SON)
0 0.1 0.2 0.3 0.4 0.5 0.6
-0.2 -0.1 0 0.1 0.2

FIG. 6. The EMR model precipitation forecast skill. (left) Correlation between (1-day lead time) EMR forecast (ensemble mean of 100 members) and daily P_{10} time series from NARR reanalysis, for each season. (right) The difference between forecast skill of the EMR model and (daily) damped persistence forecast of P_{10} (see text for details). Stippling indicates the areas of positive differences, where the EMR forecast outperforms the damped persistence forecast.

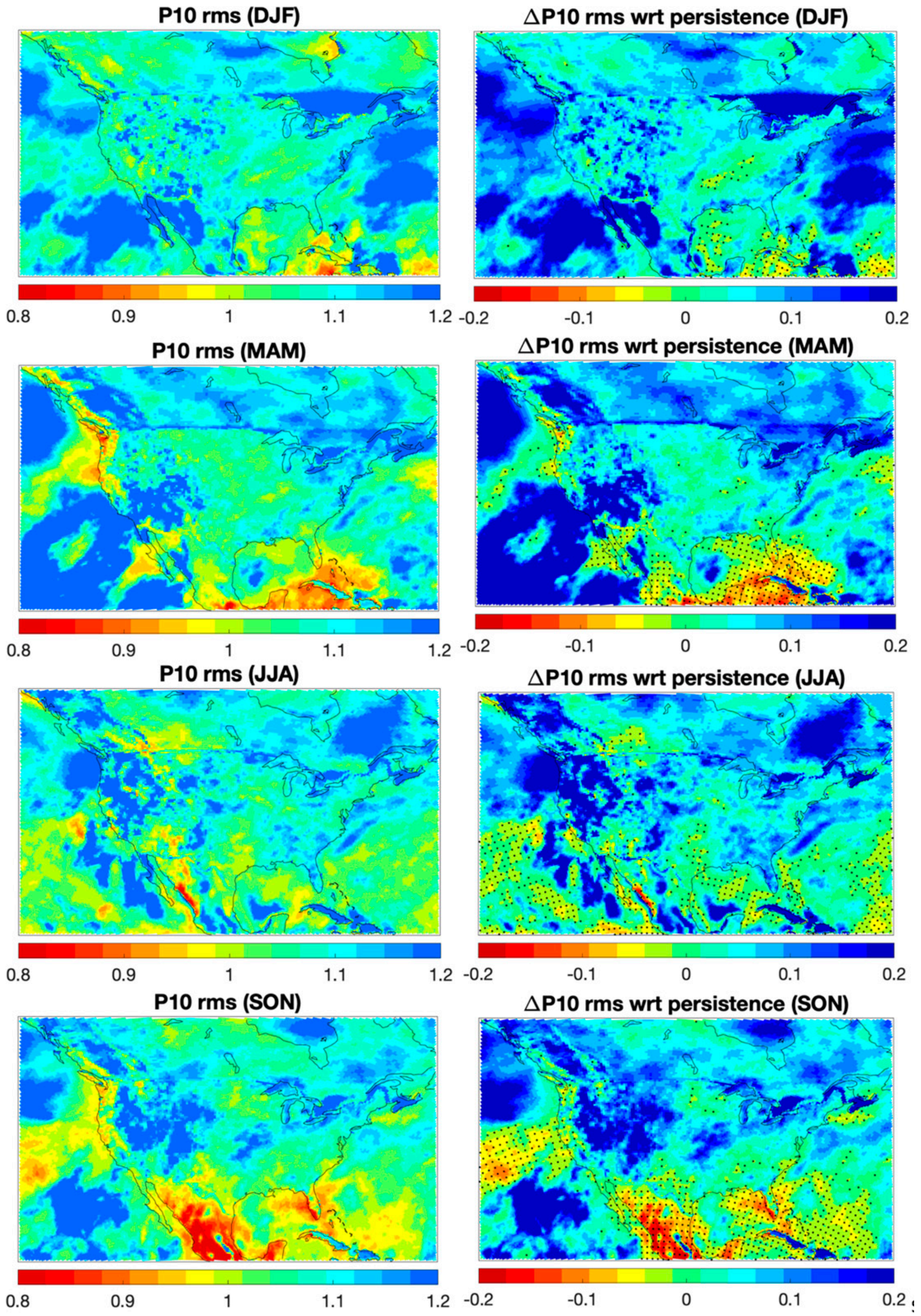

P10 rms (DJF)
ΔP10 rms wrt persistence (DJF)
P10 rms (MAM)
ΔP10 rms wrt persistence (MAM)
P10 rms (JJA)
ΔP10 rms wrt persistence (JJA)
P10 rms (SON)
ΔP10 rms wrt persistence (SON)
0.8
0.9
1
1.1
1.2
-0.2
-0.1
0
0.1
0.2

FIG. 7. EMR model's P_{10} forecast (2000–20) root-mean-square (rms) error as a fraction of (1979–99) climatological standard deviations, for each season. Note the inverted color scale; otherwise, the same layout and conventions as in Fig. 6.

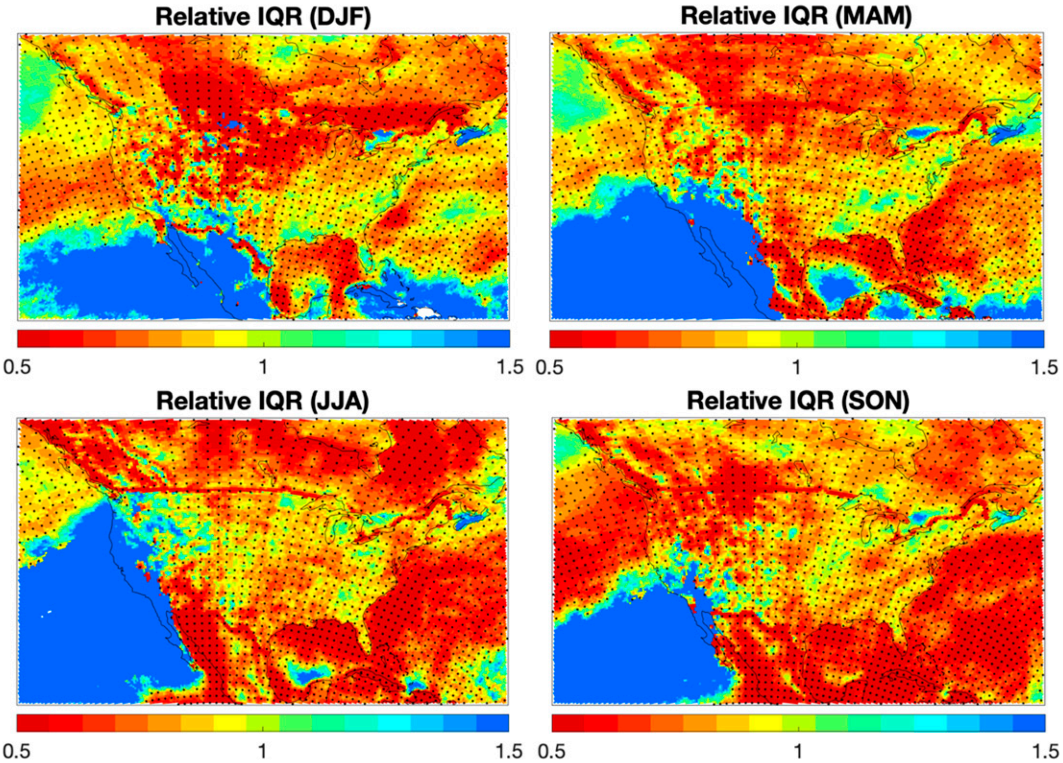


FIG. 8. The (average 2000–20) EMR model's P_{10} forecast interquartile range (IQR)—based on an ensemble of 100 forecasts—relative to the (1979–99) climatological IQR of P_{10}, for each season. The ratios below unity (stippling) indicate an enhanced forecast utility relative to that of climatology forecast. Note the inverted color scale.

the so-called potential predictability (see Kleeman 2002 and references therein), with the values less than 1 (this value corresponds to climatology forecast) and increasingly closer to zero indicating a progressively more reliable forecast. Based on this measure, the EMR model provides potentially useful forecasts throughout the region of interest, including CONUS, except for the western (warm season) or southwestern (cool season) United States and Mexico, where the EMR predictions are over-dispersed when forecasting the behavior of the North Pacific/Hawaiian high (also see below).

While providing a measure of forecast utility, the potential predictability does not directly compare the forecast with the actual observed precipitation value for the time of forecast. To do so, we here introduce an additional forecast skill measure—the forecast success rate—by counting the frequency of forecasts for which the observed P_{10} value is within the IQR range of the EMR forecast ensemble. The EMR model forecast success rate has large areas with values exceeding 0.5 (the observed value of P_{10} is within the IQR of EMR forecasts 50% of the time or more) and sometimes nearing the value of 1 (Fig. 9, left). By this metric, the EMR forecasts seem most reliable off of the Pacific coast, but this is simply an artifact of an over-dispersion documented over the same region in Fig. 8. Combining the results of Figs. 8 and 9, one can argue that the EMR model forecasts have low dispersion and >50% success rates over much of the United States from the Great Plains to California consistently throughout the seasons (or for about 50% of the grid points over the United States). We also combined the damped persistence forecasts of P_{10} with the mean and IQR range of the corresponding EMR forecast to compute the success rate associated with the damped persistence forecast: in particular, the "range" associated with a damped persistence forecast f_p was set to be $f_p - \Delta_m$, $f_p + \Delta_p$, where Δ_m and Δ_p are the offsets between the EMR model's ensemble mean and its 25th and 75th percentiles, respectively. We verified that the damped persistence forecast success rate defined in this way is substantially lower than the EMR model's success rate (Fig. 9, right).

Hence, the EMR model produces reliable, low-dispersion forecasts that tend to track the observed precipitation (signal), much more so than the damped persistence forecasts. Kleeman (2002) argued that a forecast's relative entropy

$$R = \sum_i p_i \ln\frac{p_i}{q_i}, \tag{6}$$

where p_i is climatological distribution and q_i is that for the prediction, can be very useful in characterizing prediction utility as it naturally captures both the signal and dispersion components of skill. We computed R by estimating the seasonal climatological distributions p_i over the 1979–99 training period for each grid point in terms of the normalized histograms using 20 equally spaced bins between the minimum and maximum P_{10} value at this grid point, and then computing the

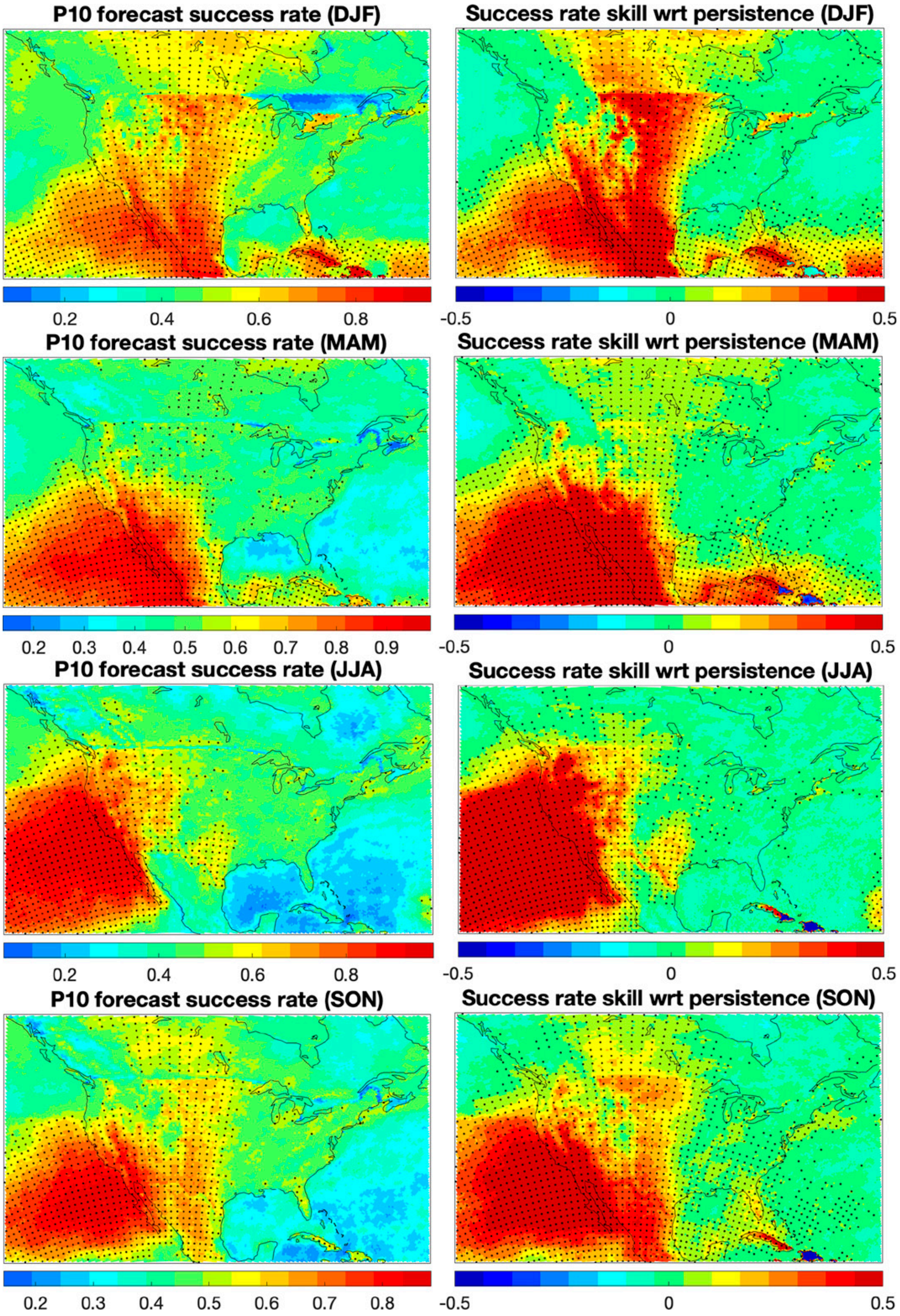

P10 forecast success rate (DJF)
0.2 0.4 0.6 0.8
Success rate skill wrt persistence (DJF)
-0.5 0 0.5
P10 forecast success rate (MAM)
0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Success rate skill wrt persistence (MAM)
-0.5 0 0.5
P10 forecast success rate (JJA)
0.2 0.4 0.6 0.8
Success rate skill wrt persistence (JJA)
-0.5 0 0.5
P10 forecast success rate (SON)
0.2 0.3 0.4 0.5 0.6 0.7 0.8
Success rate skill wrt persistence (SON)
-0.5 0 0.5

FIG. 9. (left) The EMR forecast success rate defined as the fraction of P_{10} forecasts (over all initial conditions, in each season separately) for which the actual P_{10} value from the NARR reanalysis is within the IQR of (100-member) ensemble forecasts; stippling shows the areas with success rate exceeding the value of 0.5 (associated with the climatology forecast). (right) The difference between the EMR success rate and the success rate associated with the damped persistence forecast combined with the IQR of the EMR model (see text for details); stippling denotes the areas of positive differences, where the EMR model outperforms damped persistence forecast.

distribution q_i of the 100 EMR forecasts using the same bins, for each initial condition in the 2000–20 period. With the natural logarithms in (6), the units of information in which the relative entropy is measured are referred to as nats (as opposed to bits for base-2 logarithms). The relative entropy measures how different the forecast distribution is from a climatological distribution. However, the *expectation* of R (characterizing climatological difference between forecasts and observations) would tend to be lower for the forecast schemes that are more skillful than others. For example, the climatological relative entropy associated with the damped persistence forecasts is expected to be higher than that for the EMR forecasts. This is indeed the case in Fig. 10 (left), where the relative entropy here was only computed and shown over the grid points at which the 99th percentile of P_{10} exceeded 50 mm (cf. Fig. 7, left). The basin-mean of the ratios (at each grid point) of the relative entropy associated with the damped persistence forecasts to the climatological R of the EMR forecasts over land exceeds 2.6 for all seasons. Yet, over time, the relative entropy associated with individual P_{10} forecasts can greatly exceed its climatological value (Fig. 10, right): the analogous basin mean of ratios of the 99th percentile of the EMR forecasts' relative entropy to the climatological value of R is 5.1, 4.0, 2.9, and 3.4 for the DJF, MAM, JJA, and SON seasons, respectively. In section 4c below, we will develop a subsampling strategy in which the forecasts with large values of the quantity R are tagged to define and sample potential large and extreme precipitation events.

c. *EMR-based probabilistic algorithm for thinning reforecast sample size*

Note that the cases displayed in Figs. 5 and S5 were selected using the ad hoc algorithm based on heavy precipitation in GEFSv12 reforecasts (section 3a; sample$_A$). The multipage image files with analogous maps for other selected cases are available through a web page referenced in the supplemental material. As mentioned before, the same algorithm was applied to the EMR model's ensemble-mean P_{10} forecasts, which are also available through the supplementary website; see a brief discussion at the end of this section.

We here also developed and applied an alternative strategy, which selects the dates based on the large value of the EMR forecasts' relative entropy. In particular, we computed, for each day, the average among the top 10% relative-entropy gridpoint values over CONUS. The grid points over which this was done were also preselected to have the seasonal 1979–99 P_{10} 99th percentile exceeding 5 cm, thus excluding the white areas in Fig. 10. Each day in the record was then ranked based on its relative entropy score. To eliminate possible effects of any long-term relative entropy trends, we selected 40% of the highest-score dates from the first and 40% of the highest-score dates from the second half of the original 2000–20 sample. We next edited out the member with a higher R from all the pairs of consecutive high-relative-entropy days identified above, and then from the pairs separated by two days. This procedure results in the identification of 1095 cases separated by at least two days out of the total 7671 days comprising the 2000–20 period. We argue that this subset is an optimal subset, which includes the majority of the high-impact events, while being also representative of the climatological P_{10} distribution. If more frequent sampling is required, the additional dates for reforecasts can be added at random from the remainder of the record.

The size of the latter sample is also consistent with that of the sample$_A$, which has 520 cases per each of the semiannual cool and warm seasons (over 2000–19 period, with the following breakdowns: DJF—188, MAM—276, JJA—247, and SON—329 cases). The corresponding breakdowns for the present sample are DJF—282, MAM—290, JJA—240, and SON—283 cases, featuring a more uniform seasonal distribution of cases, with more DJF cases and fewer SON cases compared to the GEFSv12 based subsample. The two samples turn out to be largely independent, with only 198 (~20%) matching dates over the 2000–19 period. A few examples of the P_{10} observed and predicted maps based on the present sample are shown in Figs. 11 and S6, and others are available through the supplementary material. The third column of these figures shows the distribution of the EMR forecasts' relative entropy on a given day, which tends to track the areas of large and extreme precipitation (recall that the relative-entropy-based selection criterion was only applied over CONUS, rather than over a larger region of the NARR reanalysis).

To assess relative performance of the two methods, we computed distributions of P_{10} associated with each sample and compared them with the climatological distribution of P_{10}. An example of these distributions in Fig. 12 demonstrates that the EMR based sample provides a better match to the NARR based P_{10} climatological distribution than the GEFSv12 based subsample, which tends to be excessively heavy tailed: DJF panel gives a particularly clear example of this for the location chosen. The positive bias of the GEFSv12 based sample—perhaps natural, given the selection criterion built on the direct occurrence of the large or extreme precipitation—is also evident in the maps of the climatological mean (Fig. 13) and, to a somewhat lesser extent, in the maps of the 99th percentile (Fig. S7) of the distributions based on the full and subsampled data. Overall, the present sample has a distribution of 0–10-day total precipitation that is closer to the distribution based on the full data compared to that of GEFSv12 based sample$_A$, while capturing the majority of high-impact precipitation events. It should be noted, however, that the ultimate test of the success of the subsampling will be the accuracy of postprocessed precipitation guidance based on the sample at hand, and not the fidelity against the NARR data. For example, heavier precipitation periods preferentially sampled by the GEFSv12 algorithm by design may be particularly important for establishing the statistical relationships in situations with heavy precipitation that are of greatest interest.

Finally, we note here that the thinned sample$_B$ obtained using the same algorithm as for the GEFSv12 data, but applied to the EMR precipitation forecasts, produced results inferior of those associated with either the EMR-RE sample or the GEFSv12-based sample$_A$ in terms of the similarity of climatological precipitation distributions based on the thinned and full available data samples (Figs. S8 and S9). This may

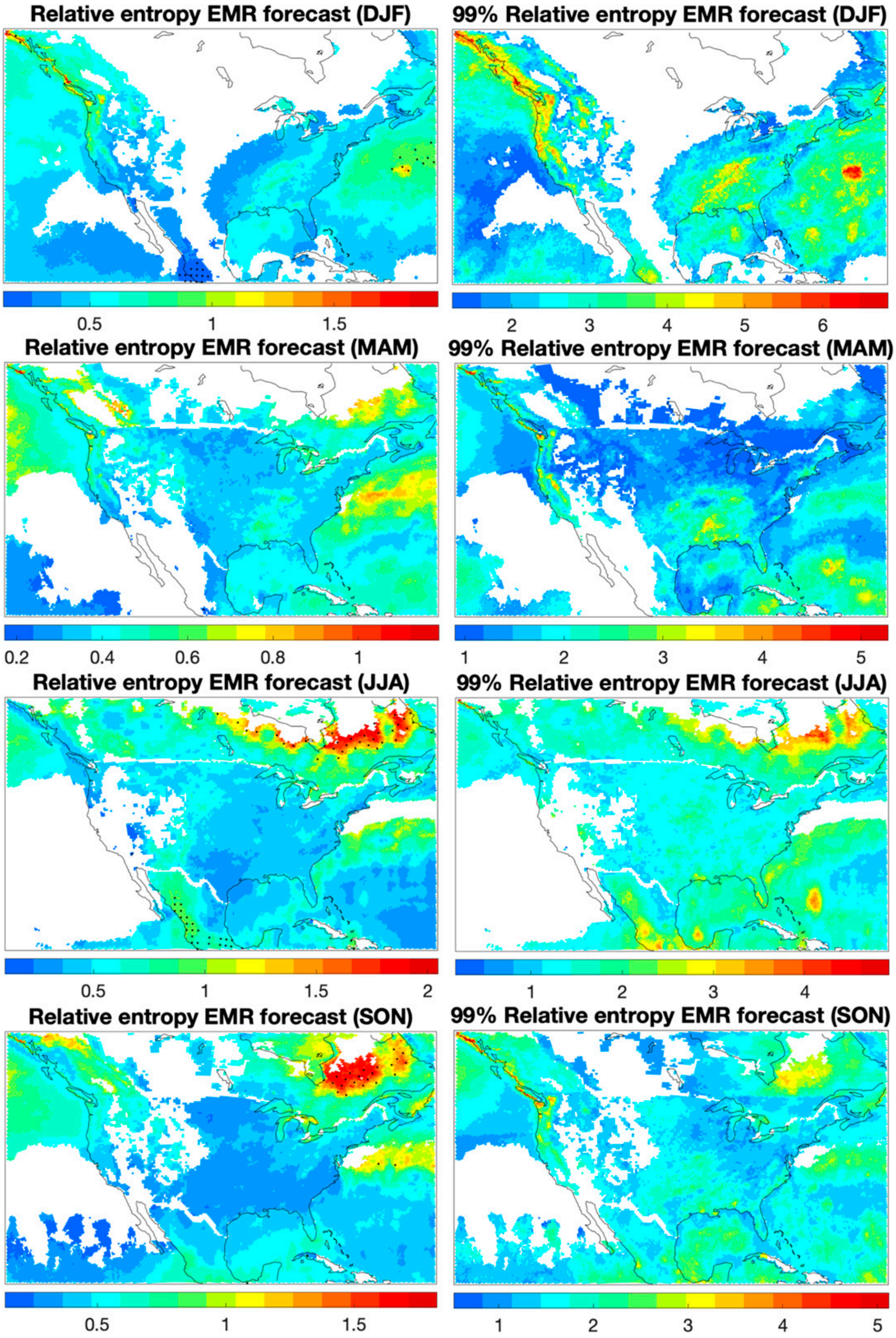

Relative entropy EMR forecast (DJF)
0.5
1
1.5
99% Relative entropy EMR forecast (DJF)
2
3
4
5
6
Relative entropy EMR forecast (MAM)
0.2
0.4
0.6
0.8
1
99% Relative entropy EMR forecast (MAM)
1
2
3
4
5
Relative entropy EMR forecast (JJA)
0.5
1
1.5
2
99% Relative entropy EMR forecast (JJA)
1
2
3
4
Relative entropy EMR forecast (SON)
0.5
1
1.5
99% Relative entropy EMR forecast (SON)
1
2
3
4
5

FIG. 10. Relative entropy of EMR forecasts (nats). (left) The expectation (climatology), with stippling showing the areas where this expectation exceeds that associated with the damped persistence forecast (see text for details); (right) the 99th percentile. Note that the relative entropy here was only computed and shown over the grid points at which the 99th percentile of P_{10} exceeded 50 mm (cf. Fig. 7, left); the areas in which this is not the case are colored white.

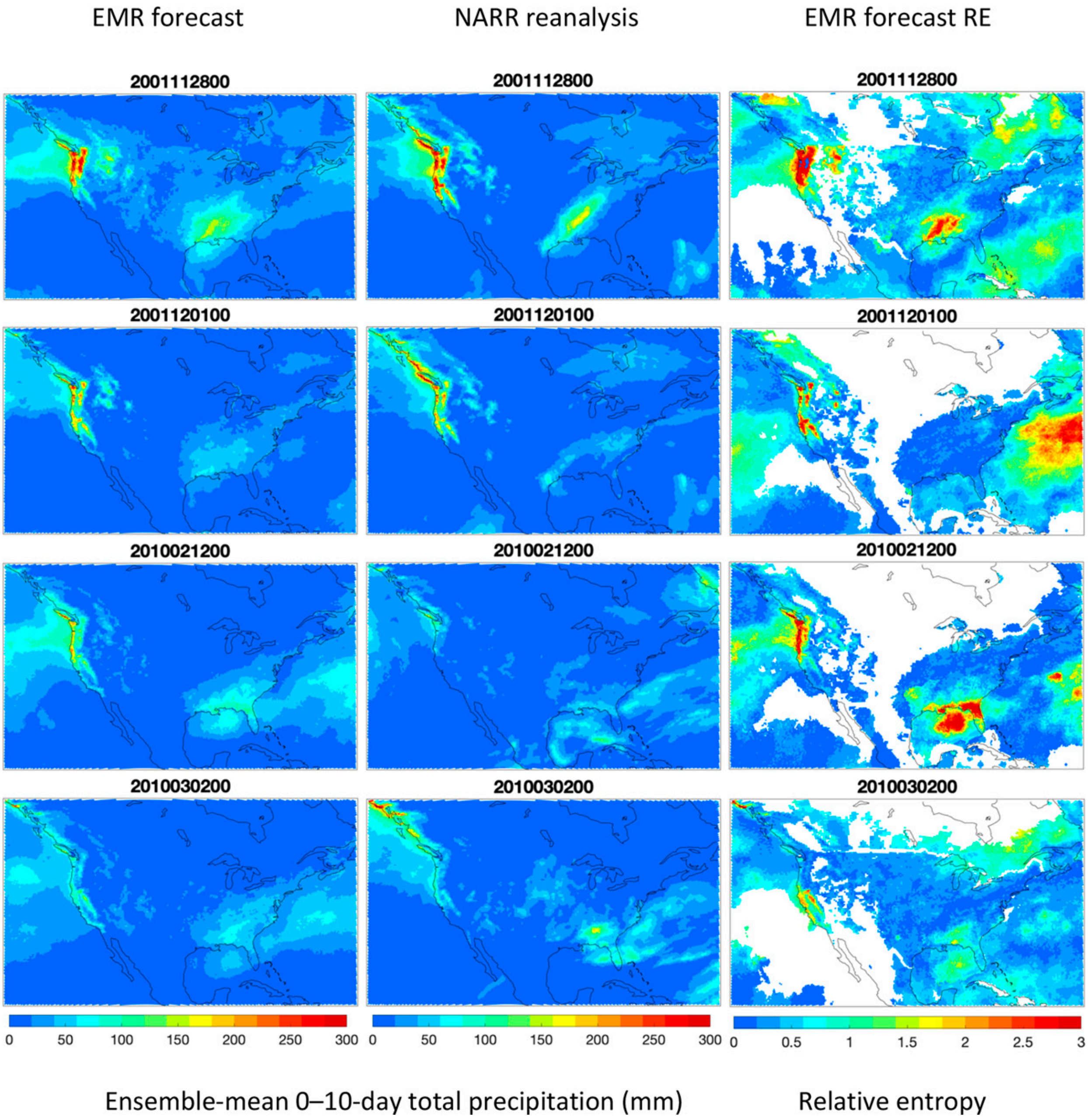


FIG. 11. Examples of (left) (cool season) P_{10} forecasts (mm) using EMR model and (center) the actual P_{10} maps based on NARR reanalysis. (right) The corresponding map of the relative entropy. The forecast initialization time (the same across each row) is shown in panel titles in the YYYYMMDDHH format. Note that, similar to Fig. 10, the relative entropy in the right column plots was only computed and shown over the grid points at which the 99th percentile of P_{10} exceeded 50 mm (cf. Fig. 7, left); the areas in which this is not the case are colored white.

be due to the fact that the EMR forecasts of P_{10} have a smaller deterministic skill than analogous high-end GEFSv12 reforecasts.

5. Summary and discussion

In this study, we developed a novel methodology for multiscale statistical modeling of precipitation by utilizing the empirical model reduction (EMR) technique (Kravtsov et al. 2005, 2017) applied to the NARR reanalysis. The key element of the new algorithm is the usage of the pseudo-precipitation PP—whose positive values are associated with the actual precipitation, while negative values represent the column integrated water vapor saturation deficit—as a part of the climate state vector to be simulated by the EMR model. The PP field thus carries information about both the mesoscale precipitation features and synoptic-scale environmental background comprised of large-scale meteorological patterns (LSMP)

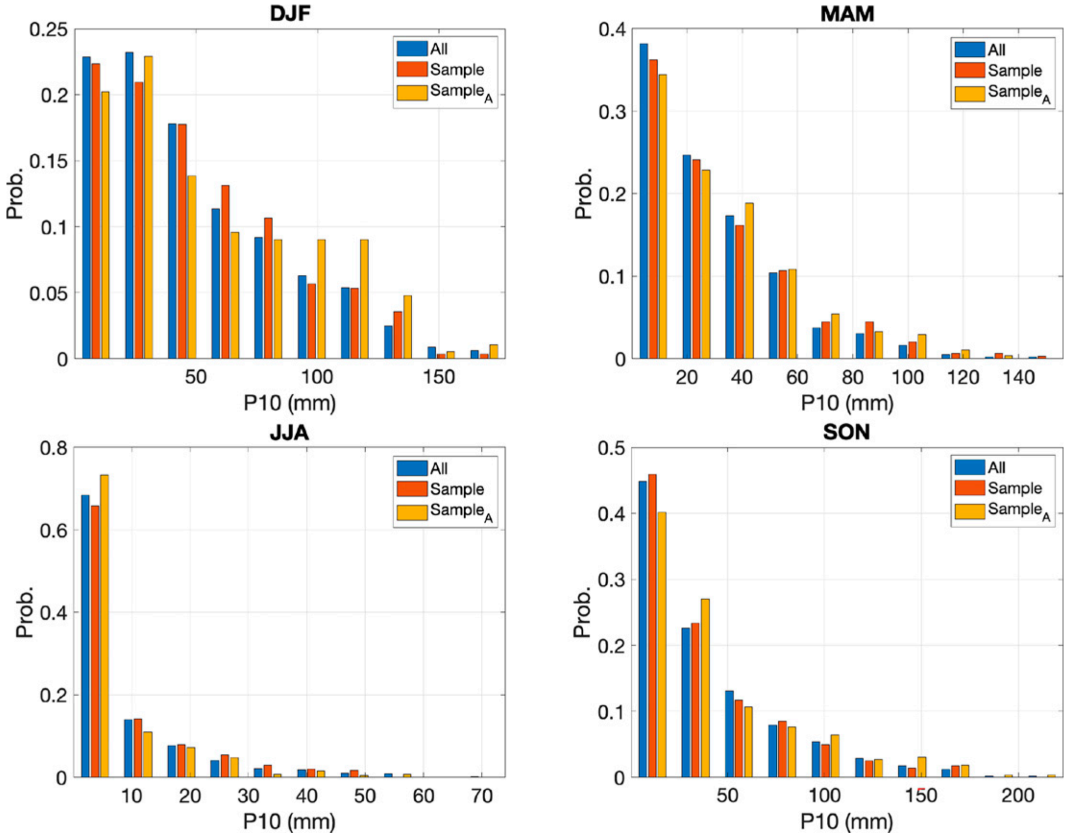


FIG. 12. The P_{10}'s probability density function (PDF) estimates at 47.4°N, 122.4°W (Seattle, WA) based, for each season, on the entire daily P_{10} data (blue), and two thinned subsamples ~1/7 the size of the whole available data: a subsample based on relative entropy of EMR forecasts (EMR-RE) (sample, red) and the one (sample$_A$, yellow) based on ensemble-mean GEFSv12 P_{10} forecasts associated with significant precipitation events over CONUS (see text for details).

potentially conducive to high-impact precipitation events. This EMR model was found to provide a seamless spatiotemporal statistical description of the precipitation-producing weather systems across a wide range of spatial scales over the entirety of CONUS and to possess a significant predictive skill, especially in a probabilistic sense.

We defined the events-of-impact in terms of the relative entropy (Kleeman 2002) of the EMR based ensemble hindcasts of the 0–10-day total surface precipitation P_{10} over the 2000–20 period and identified an optimal—arguably minimal—subset of dates proved to provide local precipitation distributions consistent with those based on the full dataset. By contrast, an alternative statistical methodology for selecting such dates based directly on the magnitude of P_{10} in high-end ensemble-mean reforecasts of precipitation produced subsamples with a more substantial heavy-precipitation bias. Thinning the frequency of reforecasts—the task that motivated this research in the first place—is extremely important in a variety of hydrological modeling applications to be described in a future paper.

Note that our selecting reforecast cases for their presumed importance in one metric (here, 0–10-day precipitation) may bias the sampling properties for different kinds of important extreme events, which might include hurricanes, mixed precipitation events, severe weather, extreme surface temperatures or winds, among others. For example, heavy precipitation events are forecast better using the quantile approach with respect to precipitable water than the absolute magnitude of the precipitable water (Kunkel et al. 2020). Such biases, however, would be a limitation of any method that seeks to limit the reforecast sample size.

Another possible limitation of the EMR methodology developed here is that the EMR model is trained on the earlier data, while the ongoing climate change may skew the more recent historical record in various ways, introducing a bias into EMR forecasts associated with the latter record. For our present application, we believe that such biases associated with the P_{10} statistics are relatively small, as evidenced by a fairly uniform in time distribution of dates in our thinned samples. For example, the number of important cases identified in the first and second halves of the 2000–20 record is similar.

Our new EMR methodology for statistical modeling of precipitation is fundamentally different from more traditional techniques, which typically work with individual precipitation records at a local level and/or postulate ad hoc connections

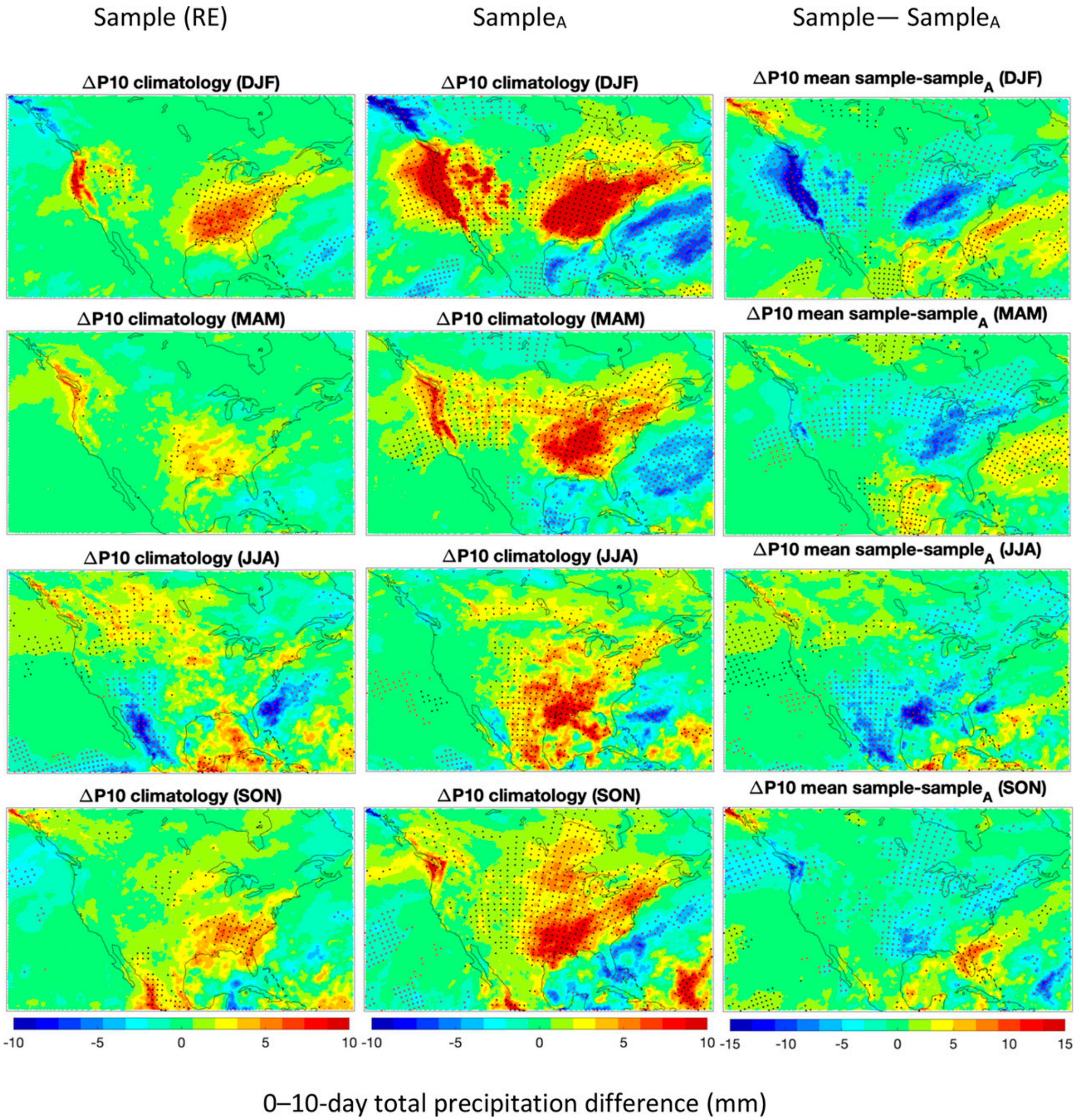


FIG. 13. The differences between the estimates of P_{10}'s climatological mean based on (left) EMR-based thinned sample and the entire seasonal P_{10} data; (center) GEFSv12-based thinned sample$_A$ and the entire seasonal P_{10} data; and (right) EMR-based thinned sample and GEFSv12-based thinned sample$_A$. Stippling shows areas where the differences are statistically significant at the 5% level according to the two-sided bootstrap test involving surrogate random subsamples of the same size as either the EMR-based thinned sample or GEFSv12-based thinned sample$_A$.

with a limited number of large-scale predictors (see section 2a), in that it automatically accounts for spatiotemporal multiscale structure of precipitation dynamics, thereby providing a unified framework to model diverse precipitation environments. At the same time, it is still extremely numerically efficient and thus easily permits large-ensemble simulations/forecasts which are essential for monitoring and fully utilizing probabilistic characteristics of precipitation, in contrast to full-blown dynamical models and systems (such as GEFSv12) necessarily limited in the number of ensemble members due to prohibitive computational expenses.

This paper showcases just one application of the new EMR precipitation model to the problem of thinning the frequency of reforecasts. Follow-up work will look into how the various sampling strategies affect precipitation forecast calibration and hydrologic forecast accuracy. We also plan to further test the EMR model's potential in a wider range of related problems around the statistical/dynamical analysis of precipitation and its predictability.

Acknowledgments. This study was funded through Cooperative Institute for Great Lakes Research (CIGLR), University of Michigan, Grant SUBK00011312 (PR, SK) and, overall, by the NOAA Grant NWS OSTI U8MWFST-PLK. We thank the two anonymous reviewers for their constructive comments, which helped improve the readability of the manuscript.

Data availability statement. The NARR reanalysis data are available at https://psl.noaa.gov/data/gridded/data.narr.html. GEFSv12 data may be accessed at https://noaa-gefs-retrospective.s3.amazonaws.com/index.html. This manuscript also has a supplementary website with data and figures generated during this study, as described in detail in the supplemental information. All MATLAB/Python scripts associated with this project are available from the authors by request.

APPENDIX

The List of Abbreviations Used

ARMA	Auto-regressive moving-average
CCA	Canonical correlation analysis
CONUS	Contiguous United States
DJF	December–January–February
EMR	Empirical model reduction
EOF	Empirical orthogonal function
GEFS	Global Ensemble Forecast System
GEFSv12	Global Ensemble Forecast System, version 12
GLM	Generalized linear model
HEFS	Hydrologic Ensemble Forecast Service
HMM	Hidden Markov model
HUC	Hydrologic Unit Codes
HUC-2	2-digit Hydrologic Unit Codes
IQR	Interquartile range
JJA	June–July–August
LIM	Linear inverse model
LSMP	Large-scale meteorological pattern
MAM	March–April–May
MLR	Multiple linear regression
NARR	North American Regional Reanalysis
NWP	Numerical weather prediction
NWS	National Weather Service
NHMM	Nonhomogeneous hidden Markov model
PLS	Partial least squares
PP	Pseudo-precipitation
Pr	Precipitation
P_{10}	10-day cumulative precipitation
PC	Principal component
rms	Root-mean-square
RE	Relative entropy
RFC	River Forecast Center
SON	September–October–November
UTC	Coordinated universal time

REFERENCES

Albers, J. R., and M. Newman, 2019: A priori identification of skillful extratropical subseasonal forecasts. *Geophys. Res. Lett.*, **46**, 12 527–12 536, https://doi.org/10.1029/2019GL085270.

Barlow, M., and Coauthors, 2019: North American extreme precipitation events and related large-scale meteorological patterns: A review of statistical methods, dynamics, modeling, and trends. *Climate Dyn.*, **53**, 6835–6875, https://doi.org/10.1007/s00382-019-04958-z.

Bolton, D., 1980: The computation of equivalent potential temperature. *Mon. Wea. Rev.*, **108**, 1046–1053, https://doi.org/10.1175/1520-0493(1980)108<1046:TCOEPT>2.0.CO;2.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel, 1994: *Time Series Analysis, Forecasting and Control.* 3rd ed. Prentice Hall, 592 pp.

Bukovsky, M. S., and D. J. Karoly, 2007: A brief evaluation of precipitation from the North American Regional Reanalysis. *J. Hydrometeor.*, **8**, 837–846, https://doi.org/10.1175/JHM595.1.

Demargne, J., and Coauthors, 2014: The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1.

Furrer, E., and R. Katz, 2007: Generalized linear modeling approach to stochastic weather generators. *Climate Res.*, **34**, 129–144, https://doi.org/10.3354/cr034129.

Grotjahn, R., and Coauthors, 2016: North American extreme temperature events and related large scale meteorological patterns: A review of statistical methods, dynamics, modeling and trends. *Climate Dyn.*, **46**, 1151–1184, https://doi.org/10.1007/s00382-015-2638-6.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hamill, T. M., 2018: Practical aspects of statistical postprocessing. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. Messner, Eds., Elsevier, 187–218.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

——, ——, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46, https://doi.org/10.1175/BAMS-87-1-33.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and Climatology-Calibrated Precipitation Analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, https://doi.org/10.1175/MWR-D-15-0004.1.

——, and Coauthors, 2022: The reanalysis for the Global Ensemble Forecast System, version 12. *Mon. Wea. Rev.*, **150**, 59–79, https://doi.org/10.1175/MWR-D-21-0023.1.

Holsclaw, T., A. M. Greene, and A. W. Robertson, 2016: A Bayesian hidden Markov model of daily precipitation over

South and East Asia. *J. Hydrometeor.*, **17**, 3–25, https://doi.org/10.1175/JHM-D-14-0142.1.

Kenabatho, P. K., N. R. McIntyre, R. E. Chandler, and H. S. Wheater, 2012: Stochastic simulation of rainfall in the semi-arid Limpopo basin, Botswana. *Int. J. Climatol.*, **32**, 1113–1127, https://doi.org/10.1002/joc.2323.

Kleeman, R., 2002: Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.*, **59**, 2057–2072, https://doi.org/10.1175/1520-0469(2002)059<2057:MDPUUR>2.0.CO;2.

Kravtsov, S., D. Kondrashov, and M. Ghil, 2005: Multi-level regression modeling of nonlinear processes: Derivation and applications to climatic variability. *J. Climate*, **18**, 4404–4424, https://doi.org/10.1175/JCLI3544.1.

——, M. Ghil, and D. Kondrashov, 2010: Empirical model reduction and the modeling hierarchy in climate dynamics and the geosciences. *Stochastic Physics and Climate Modeling*, T. Palmer and P. Williams, Eds., Cambridge University Press, 35–72.

——, N. Tilinina, Y. Zyulyaeva, and S. Gulev, 2016: Empirical modeling and stochastic simulation of sea-level pressure variability. *J. Appl. Meteor. Climatol.*, **55**, 1197–1219, https://doi.org/10.1175/JAMC-D-15-0186.1.

——, P. Roebber, and V. Brazauskas, 2017: A virtual climate library of surface temperature over North America for 1979–2015. *Sci. Data*, **4**, 170155, https://doi.org/10.1038/sdata.2017.155.

Kunkel, K. E., S. E. Stevens, L. E Stevens, and T. R. Karl, 2020: Observed climatological relationships of extreme daily precipitation events with precipitable water and vertical velocity in the contiguous United States. *Geophys. Res. Lett.*, **47**, e2019GL086721, https://doi.org/10.1029/2019GL086721.

Manzanas, R., A. Lucero, A. Weisheimer, and J. M. Gutierrez, 2018: Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? *Climate Dyn.*, **50**, 1161–1176, https://doi.org/10.1007/s00382-017-3668-z.

McCullagh, P., and J. Nelder, 1989: *Generalized Linear Models.* Chapman and Hall, 532 pp.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, https://doi.org/10.1175/BAMS-87-3-343.

Newman, M., P. D. Sardeshmukh, C. R. Winkler, and J. S. Whitaker, 2003: A study of subseasonal predictability. *Mon. Wea. Rev.*, **131**, 1715–1732, https://doi.org/10.1175//2558.1.

Penland, C., 1996: A stochastic model of Indo-Pacific sea surface temperature anomalies. *Physica D*, **98**, 534–558, https://doi.org/10.1016/0167-2789(96)00124-8.

——, and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *J. Climate*, **8**, 1999–2024, https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2.

Robertson, A. W., and M. Ghil, 1999: Large-scale weather regimes and local climate over the western United States. *J. Climate*, **12**, 1796–1813, https://doi.org/10.1175/1520-0442(1999)012<1796:LSWRAL>2.0.CO;2.

——, S. Kirshner, and P. Smyth, 2004: Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model. *J. Climate*, **17**, 4407–4424, https://doi.org/10.1175/JCLI-3216.1.

——, Y. Kushnir, U. Lall, and J. Nakamura, 2016: Weather and climatic drivers of extreme flooding events over the Midwest of the United States. *Extreme Events: Observations, Modeling, and Economics, Geophys. Monogr.*, Vol. 214, Amer. Geophys. Union, https://doi.org/10.1002/9781119157052.ch9.

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

Sinha, P., U. C. Mohanty, S. C. Kar, S. K. Dash, A. W. Robertson, and M. K. Tippett, 2013: Seasonal prediction of the Indian summer monsoon rainfall using canonical correlation analysis of the NCMRWF global model products. *Int. J. Climatol.*, **33**, 1601–1614, https://doi.org/10.1002/joc.3536.

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

Winkler, C. R., M. Newman, and P. D. Sardeshmukh, 2001: A linear model of wintertime low-frequency variability. Part I: Formulation and forecast skill. *J. Climate*, **14**, 4474–4494, https://doi.org/10.1175/1520-0442(2001)014<4474:ALMOWL>2.0.CO;2.

Wold, S., and Coauthors, 1984: *Chemometrics, Mathematics and Statistics in Chemistry.* Reidel Publishing Company, 492 pp.

Yuan, H., P. Schultz, E. I. Tollerud, D. Hou, Y. Zhu, M. Pena, M. Charles, and Z. Toth, 2019: Pseudo-precipitation: A continuous precipitation variable. NOAA Tech. Memo. OAR GSD-62, https://doi.org/10.25923/3h37-gp49.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System Version 12. *Wea. Forecasting*, https://doi.org/10.1175/WAF-D-21-0112.1, in press.

Zobel, Z., J. Wang, D. J. Wuebbles, and V. R. Kotamarthi, 2018: Evaluations of high-resolution dynamically downscaled ensembles over the contiguous United States. *Climate Dyn.*, **50**, 863–884, https://doi.org/10.1007/s00382-017-3645-6.