Principal Component Analysis for Extremes and Application to U.S. Precipitation

YUJING JIANG AND DANIEL COOLEY

Colorado State University, Fort Collins, Colorado

MICHAEL F. WEHNER

Lawrence Berkeley National Laboratory, Berkeley, California

(Manuscript received 3 June 2019, in final form 1 May 2020)

ABSTRACT

We propose a method for analyzing extremal behavior through the lens of a most efficient basis of vectors. The method is analogous to principal component analysis, but is based on methods from extreme value analysis. Specifically, rather than decomposing a covariance or correlation matrix, we obtain our basis vectors by performing an eigendecomposition of a matrix that describes pairwise extremal dependence. We apply the method to precipitation observations over the contiguous United States. We find that the time series of large coefficients associated with the leading eigenvector shows very strong evidence of a positive trend, and there is evidence that large coefficients of other eigenvectors have relationships with El Niño–Southern Oscillation.

1. Introduction

There is great current interest in understanding patterns and trends of extreme weather events. Of particular recent interest has been the quantification of the influence of anthropogenic climate change on specific individual extreme weather events (National Academies of Sciences, Engineering, and Medicine 2016). Climate change affects extreme weather locally through thermodynamically driven processes as well as nonlocally through changes in the statistics of the large-scale meteorological patterns conducive to extreme weather.

This work proposes a new tool for exploring patterns and trends of extreme weather; specifically, we propose an extremes analog to principal component analysis (PCA). To illustrate the method, we apply it to 3-day precipitation data from continental U.S. (CONUS) weather stations during hurricane season and investigate overall trends of extreme precipitation as well as relationships to El Niño–Southern Oscillation (ENSO). However, the method is not specific to precipitation studies and could be applied to explore any climate variable of interest.

PCA (also, empirical orthogonal function analysis) is a popular tool in climate science that reduces a large set of variables into a smaller, more interpretable, set (Wilks 2011, chapter 12). High-dimensional dependence is viewed through the lens of the ordered basis of eigenvectors of the covariance matrix. PCA is best suited to variables that are approximately Gaussian and whose dependence follows the elliptical contours of a Gaussian density. Because short-term precipitation is positively skewed and almost always contains a substantial fraction of values that are exactly zero, PCA is more often applied to monthly or season precipitation (e.g., Uvo 2003; He et al. 2017), although some PCA studies of shorter duration precipitation have been performed (e.g., Widmann and Schär 1997).

More importantly, because it arises from the covariance matrix, standard PCA is poorly suited to study *extreme* behavior of any variable of interest. The covariance matrix describes dependence at the center of the distribution and may not accurately capture dependence in the joint tail. Also, when studying extreme behavior, there is almost always a direction of interest; for example, in our study we seek to learn about large precipitation events, and are uninterested in this study in small precipitation events. Covariance measures linear dependence in both directions from center, and dependence among stations for low precipitation likely differs from that for high precipitation, as high precipitation is

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-19-0413.s1.

Corresponding author: Daniel Cooley, cooleyd@stat.colostate.edu

DOI: 10.1175/JCLI-D-19-0413.1

often very localized, whereas lack of precipitation is generally more widespread.

Extreme value analysis is the branch of statistics specifically aimed at describing the tail of the distribution. Extreme value theory classifies the upper tail of a distribution as "bounded" (having a finite upper bound), "light" (infinite upper bound and essentially decreasing like an exponential function), or "heavy" (infinite upper bound and decreasing like a power function). To focus on the tail, extremes methods use only a subset of the extreme data and the rest are discarded. When studying extremes for a single climate variable, fitting the generalized extreme value (GEV) distribution to the block maxima or fitting the generalized Pareto distribution (GPD) to large values over some threshold have become a common practice in climate science (Wilks 2011, section 4.4.5).

Our method begins with the notion of tail dependence, and a few studies in atmospheric science have investigated tail dependence in the bivariate case. Timmermans et al. (2019) use an extremal dependence measure to compare extreme precipitation in gridded data products for the continental United States, Weller et al. (2013) use the same measure to compare the precipitation extremes in the observational record to those produced by reanalysis, and Kuhn et al. (2007) use a related measure to describe the extremal dependence in precipitation extremes at different locations in South America. Ben Alaya et al. (2018) use bivariate extreme value theory to model the relationship between two components underlying a calculation of probable maximum precipitation. There has also been an abundance of work developing spatial models for extremes and applying them in various settings; Davison et al. (2019) give a recent summary. Spatial extremes models are parameterized by making simplifying assumptions about the spatial behavior of dependence and are well suited for modeling aggregated effects across multiple locations of extreme events. Spatial extremes models are typically used to model local or regional extremes, but have not been applied on the continental scale. Furthermore, these models become difficult to fit as the number of locations increases.

The method we propose here differs in that it is primarily for *exploration* of extremal behavior when the dimension is very large: our application looks at over 1000 station locations spread over the continental United States. Our method is rooted in multivariate extreme value theory; specifically, our approach relies on the framework of multivariate regular variation. There are other representations for multivariate extremes such as the multivariate extreme value distributions (de Haan and Ferreira 2006, chapter 6) and the multivariate GPD (Rootzén and Tajvidi 2006), but all have very closely related dependence structures. Similar to traditional PCA, we will summarize the tail dependence information in a matrix of pairwise extreme dependence metrics. We then perform an eigendecomposition of this tail pairwise dependence matrix (TPDM; Cooley and Thibaud 2019), and view extremal dependence through the lens of a resulting eigenbasis.

2. Extremal dependence and eigendecomposition

a. A framework for multivariate extremes

The foundation of our method is the framework of multivariate regular variation. Essentially, a random vector that is multivariate regularly varying is one that is heavy-tailed in all its dimensions. Importantly, the definition of multivariate regular variation only describes the upper tail; thus, like the GEV and GPD univariate extremes models, the framework focuses on extreme behavior and does not characterize the full distribution. The probabilistic behavior of a multivariate regularly varying random vector is most easily described after polar transformation, as the magnitude and direction of the vector are approximately independent for large observations.

Let $\mathbf{X}$ be a regularly varying random vector taking values on $[0,\infty)^p$. We work on the p-dimensional positive orthant, as this allows us to focus on the large values and ignore the small values, which are of no interest. A formal definition of regular variation requires ideas of convergence; more details can be found in Cooley and Thibaud (2019), and Resnick (2007) gives comprehensive treatment of regular variation. For our purposes here, it suffices to say that we assume if A is a set consisting of large values (sufficiently far away from the origin), then

$$P(\mathbf{X}\in A)\underset{\sim}{\propto}\int_{(r,\mathbf{w})\in A}\alpha r^{-\alpha-1}\,dr\,dH(\mathbf{w}). \tag{1}$$

Here, the symbol $\underset{\sim}{\propto}$ denotes "approximately proportional to," $\alpha>0$, r refers to the magnitude or radial component of the location, $\mathbf{w}$ is a location on the unit sphere $\mathbb{S}=\{\mathbf{w}\in\mathbb{R}_+^p:\|\mathbf{w}\|=1\}$, and H is a measure on the unit sphere $\mathbb{S}$. The heavy-tailed nature of the distribution is shown in that r in the integrand has power-law behavior given by α. As α decreases, the tail becomes heavier, and α is the reciprocal of ξ, the shape parameter of GEV distribution in Coles (2001) and elsewhere. Figure 1 illustrates regular variation's polar representation in two dimensions.

Assuming (1), the probabilistic behavior of extreme events is characterized by the tail index α and the

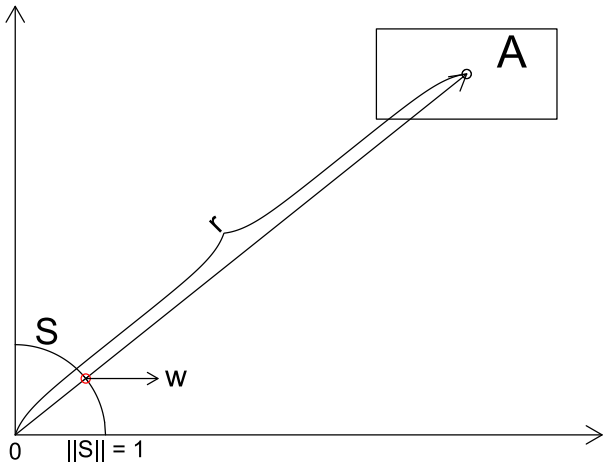

FIG. 1. Illustration of the polar form of regular variation in two dimensions. As points become large, the magnitude given by the radial component r becomes independent of the direction or angular component $\mathbf{w} \in \mathbb{S}$. Dependence information is contained in the measure H that lives on $\mathbb{S}$ and can be thought of as a (perhaps unnormalized) distribution of the angular components.

angular measure H, which describes tail dependence. In two dimensions as in Fig. 1, dependence increases as the mass of H concentrates in the center of $\mathbb{S}$, as this implies that when the magnitude of $\mathbf{X}$ is large, both components tend to be large since values of $\mathbf{w}$ near the center of $\mathbb{S}$ have roughly equal values.

Asymptotic independence is a fundamental notion of tail dependence. Let $x_1(p)$ and $x_2(p)$ denote the pth quantile of random variables X_1 and X_2, respectively. Note that X_1 and X_2 are asymptotically independent if $\lim_{p\to 1} P[X_2 > x_2(p) | X_1 > x_1(p)] = 0$, and asymptotically dependent if this limit is greater than zero. If X_1 and X_2 are jointly regularly varying and asymptotically independent, then the mass of H exists only on the axes. The standard regular variation framework we use here is most often used for describing dependence in the asymptotically dependent setting and can be extended to calculate probabilities associated with jointly extreme events for asymptotically independent regularly varying random variables (Resnick 2002). We will comment about asymptotic dependence for the precipitation data we analyze in the discussion in section 4.

In relatively small dimensions, the angular measure H can be modeled either parametrically or nonparametrically. However, in large dimensions there are neither applicable models nor sufficient information in the subset of extreme events to fit H. Rather than completely model the high-dimensional angular measure, in the next section we summarize the dependence contained in H via the TPDM, a matrix of bivariate tail dependencies.

In the remainder of the paper, we will refer to the "scale" of the components $X_i, i = 1, \ldots, p$. Formally, we say X_i has scale b if $\lim_{x\to\infty} P(X_i > x)/x^{-\alpha} = b$. If X_i is regularly varying with unit scale, then bX_i will have scale b. In standard PCA, scale is described by variance, but variance speaks about the scale of the random variable from its center (mean), whereas scale here describes behavior in the random variable's tail.

The regular variation framework described above requires that each of the variables is heavy-tailed with a common tail index α. Often in extremes studies, the data do not exhibit this property. Transforming the marginal distributions is common to extremes studies and can be defended by theoretical results [Resnick 1987, proposition (5.10)]. Furthermore, transforming the data is not uncommon outside of extremes: in standard multivariate analyses, data may be transformed to be approximately Gaussian, and modeling dependence via copulas requires transformation to uniform marginals. Figure 2 shows data from two stations on the original scale, and after a transformation has been applied so that the marginals are regularly varying with $\alpha = 2$. Also shown is a simple estimate of H, that is, a histogram of the

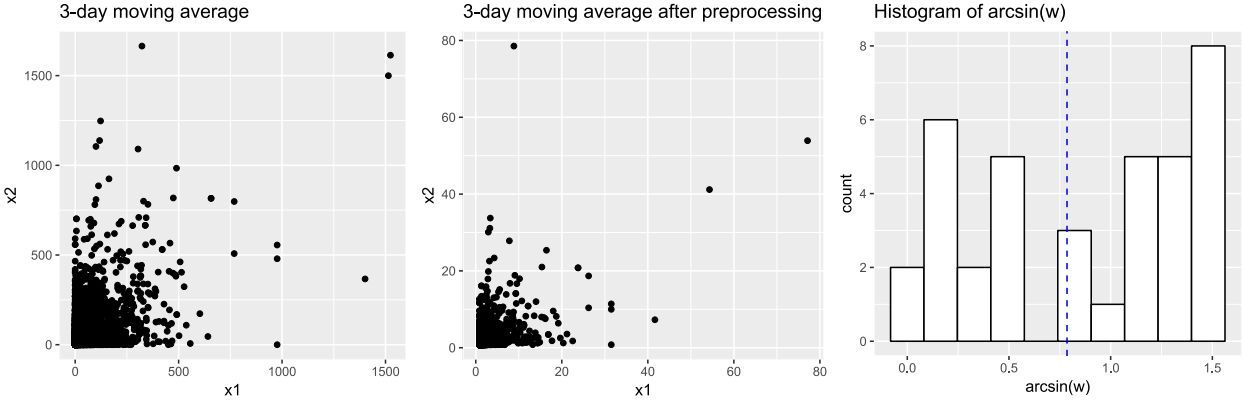

FIG. 2. Illustration of marginal transformation: (left) original precipitation data from two stations, (center) the data after transformation so that each marginal is regularly varying with $\alpha = 2$, and (right) a histogram of the angular components of the largest 2% of the data.

angular components corresponding to exceedances of the empirical 0.98 quantile.

b. Tail pairwise dependence matrix

We assume $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is a p-dimensional regularly varying random vector with index $\alpha = 2$ and angular measure H. The TPDM, $\mathbf{\Sigma}_X$, is the matrix whose (i, j)th element is

$$\sigma_{i,j} = \int_{\mathbb{S}} w_i w_j \, dH(w), \quad i,j = 1, \ldots, p. \tag{2}$$

In the bivariate case $\sigma_{i,j}$ was defined by Larsson and Resnick (2012) and referred to as the extremal dependence measure.

Although it focuses on extremal dependence due to its reliance on H, the construction of TPDM is similar to that of standard covariance matrix and consequently it has similar properties. Thus, if X_i has scale b then the ith diagonal element $\sigma_{i,i}$ is b^2, and $\sigma_{i,j} = 0$ if and only if X_i and X_j are asymptotically independent. Important for PCA, $\mathbf{\Sigma}_X$ is symmetric and positive definite, thus its eigenvectors are real and eigenvalues are positive. If the marginal distributions are transformed to have a common scale of one, then the TPDM is like a correlation matrix with diagonal entries of one. An additional property not shared generally by covariance/correlation matrices is that the TPDM is also completely positive: there exists a $p \times q$, with $q \geq p$, nonnegative matrix $\mathbf{B}$ such that $\mathbf{\Sigma}_X = \mathbf{B}\mathbf{B}^{\mathrm{T}}$. Although we do not use this property in this work, complete positivity yields a construction method for generating a random vector with a given TPDM (Cooley and Thibaud 2019).

To estimate the TPDM, let $\mathbf{x}_t, t = 1, \ldots, n_{\text{samp}}$ be the transformed observations for all stations on day t. Elements of the TPDM are estimated using pairs of $\mathbf{x}_t$'s elements. Define the radial component $r_{t,ij} = \sqrt{x_{t,i}^2 + x_{t,j}^2}$, and let $(w_{t,i}, w_{t,j}) = (x_{t,i}, x_{t,j})/r_{t,ij}$. We estimate $\sigma_{i,j}$ as the following:

$$\hat{\sigma}_{i,j} = 2n_{ij,\text{exc}}^{-1} \sum_{t=1}^{n_{\text{samp}}} w_{t,i} w_{t,j} \mathbb{I}(r_{t,ij} > r_{0,ij}), \tag{3}$$

where $r_{0,ij}$ is some high threshold for the radial components, and $n_{ij,\text{exc}}$ is the number of observations whose r_{ij} is greater than the corresponding high threshold. The indicator function $\mathbb{I}$ forces estimation to be based on the pairs with the largest radial component $r_{0,ij}$. Choosing it involves the usual difficulties often found in choosing a threshold in an extreme value analysis. The estimate $\hat{\sigma}_{i,j}$ should be relatively constant for sufficiently high thresholds; however, assessing when such a level has been achieved is often done via diagnostic plots. When p is large, viewing p-choose-2 diagnostic plots is not feasible. We suggest viewing diagnostic plots for a number of the possible $\hat{\sigma}_{i,j}$, and then choosing $r_{0,ij}$ to correspond to a common high quantile above which the examined $\hat{\sigma}_{i,j}$ appeared to be relatively constant.

One issue with this pairwise estimate is that the estimated $\hat{\Sigma}_X$ is not guaranteed to be positive definite. Once an initial matrix is estimated via (3), we use the nearPD function in the R package Matrix to implement the Higham (2002) method to find the positive definite matrix nearest to the estimated $\hat{\Sigma}_X$ in terms of the Frobenius norm. In this study, we were motivated to perform the pairwise estimation because of the spatial extent of the CONUS study region and the localized behavior of extreme precipitation; when studying precipitation for a small region of Switzerland, Cooley and Thibaud (2019) thresholded in terms of the entire vector of observations.

c. PCA decomposition for extremes

Critical to ordinary PCA is the fact that the eigenvectors of the covariance matrix form an orthonormal basis for the p-dimensional reals, and this basis is ordered in importance by the eigenvalues that yield the amount of variance explained by each eigenvector. Critical to our method will be obtaining an ordered orthonormal basis for the p-dimensional *positive orthant*. To have a basis, one must first have a vector space. Cooley and Thibaud (2019) create a vector space for the p-dimensional positive orthant by applying the transformation $\mathbf{x} = \tau(\mathbf{y}) = \log\{1 + \exp(\mathbf{y})\}$ componentwise to the vector $\mathbf{y} \in \mathbb{R}^p$. The important characteristic of this transformation is that $\tau(\mathbf{y}) \approx \mathbf{y}$ for large $\mathbf{y}$, and therefore the transformation has negligible effect on large values. Vector addition and scalar multiplication of a vector are defined via this transformation, and regular variation is preserved by this particular transformation.

Further, Cooley and Thibaud (2019) show that applying this transformation to the eigenvectors of the TPDM yields an orthonormal basis for the positive orthant. This basis is ordered by eigenvalues that yield the scale explained by each eigenvector. Let $\mathbf{\Sigma}_X = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues with $\lambda_1 \geq \lambda_p \geq 0$, and $\mathbf{U}$ is a matrix with columns $\mathbf{u}_i, i = 1, \ldots, p$ being the corresponding eigenvectors. The eigenvectors for the positive orthant are $\mathbf{e}_i = \tau(\mathbf{u}_i)$.

Let $\mathbf{x}_t$ be the realization of the regularly varying random vector $\mathbf{X}$ with TPDM $\mathbf{\Sigma}_X$ at time t. Let

$$\boldsymbol{v}_t = \mathbf{U}^{\mathrm{T}} \tau^{-1}(\mathbf{x}_t). \tag{4}$$

Then $\boldsymbol{v}_t$, a vector in the p-dimensional reals, is the vector of principal components for $\mathbf{x}_t$; that is, it is the vector of coefficients of the eigenbasis:

$$\mathbf{x}_t = \upsilon_{t,1} \circ \mathbf{e}_1 \oplus \cdots \oplus \upsilon_{t,p} \circ \mathbf{e}_p, \tag{5}$$

where $\circ$ and $\oplus$ are the transformed multiplication and addition of Cooley and Thibaud (2019).

The PCA decomposition becomes useful from the knowledge that most of the information in $\mathbf{x}_t$ is contained in the leading terms of (5). In a standard PCA study, the leading eigenvectors are often visualized and interpreted. Orthogonality implies that the eigenvectors contain no redundant information, and interpretation is done sequentially. Here, each eigenvector is the direction of greatest scale remaining after the scale accounted for by the previous eigenvectors has been removed. Time series of the leading principal components $\upsilon_{t,i}$ can be investigated to find behavior in the often large-scale effects described by the corresponding eigenvectors.

3. Analysis of U.S. extreme precipitation

a. Data description

We obtain daily precipitation data over the U.S. continent between 1950 and 2016 from the Global Historical Climatology Network (GHCN)-Daily dataset (Menne et al. 2012). We limit our investigation to the months of August, September, and October, which roughly corresponds to the height of hurricane activity in the Atlantic, although it is important to note that we analyze all extreme precipitation regardless of whether it was associated with a hurricane event. We select stations which have fewer than 5% missing values during this period. There are 1140 stations and 6164 days in the analyzed dataset.

b. Data preprocessing

We choose to analyze data that correspond to a 3-day moving average of the daily precipitation amounts. That is, let $z_{t,i}$ denote the observed precipitation on day t at station i, and let $x_{t,i}^{(\text{orig})} = z_{t,i} + z_{t+1,i} + z_{t+2,i}$. The superscript simply denotes that $x_{t,i}^{(\text{orig})}$ is on the original scale before further transformation as explained below. Selecting a 3-day moving average to analyze alleviates some of the problem of a single extreme precipitation event being partially recorded over two separate days; that is $z_{t,i}$ and $z_{t+1,i}$ are actually due to the same event. It also may help alignment problems between stations, for instance where $z_{t,i}$ and $z_{t+1,i}$ are due to the same event. However, taking a 3-day moving average does induce dependence in the $x_{t,i}^{(\text{orig})}$ terms, which must be accounted for in the subsequent analysis. Our 3-day average was motivated by the duration of the events we wish to explore, but the extremal PCA analysis could be applied to data of any duration of interest.

As explained in section 2a, the regular variation framework leading to the TPDM assumes each univariate marginal distribution is regularly varying with $\alpha = 2$. Since this is not true of our data, further transformation is required. We transform to obtain $x_{t,i} = G^{-1}\{\hat{F}_i[x_{t,i}^{(\text{orig})}]\}$, where $G(x) = \exp(-x^{-2})$ is the cumulative distribution function (cdf) of a Frechét random variable with scale 1 and $\alpha = 2$ and $\hat{F}_i$ is an estimated marginal cdf from the data at location i. Choosing to additionally have a common scale is analogous to performing standard PCA analysis on the correlation matrix instead of the covariance matrix. Whether it makes more sense to work with data with a common scale depends on one's aim (Wilks 2011, section 12.1.4), but a consequence of our doing so is that "extreme" precipitation is defined relative to the climate of the location.

The simplest method for obtaining $\hat{F}_i$ is to use a rank transform; however, extremes studies that aim to estimate probabilities of multivariate extreme events beyond the range of the data require a parametric model (usually GPD) to be fit to the upper tail. Here, a parametric tail model is not required. Due to the dependence induced in $x_{t,i}^{(\text{orig})}$ by the 3-day moving average, a simple rank transform would ignore this dependence. In the online supplemental information, we provide the details of a method where we take the average of three linearly interpolated cdf estimates obtained from three lag-3 subsequences of $x_{t,i}^{(\text{orig})}$. We show that applying this estimate $\hat{F}_i$ better retains clustering in the generated $x_{t,i}$.

As in a traditional PCA analysis, the transformed data $\mathbf{x}_t$, $t = 1, \ldots, 6164$ are treated as independent and identically distributed, and the estimated TPDM $\hat{\Sigma}_X$ is obtained as described in section 2b. After viewing several diagnostic plots, we choose $r_{0,ij}$ to correspond to the 0.98 quantile.

c. Interpretation of eigenvectors

The eigenvectors $\mathbf{u}_i$, $i = 1, \ldots, p$ obtained through standard eigendecomposition of $\hat{\Sigma}_X$, are transformed to $\mathbf{e}_i = t(\mathbf{u}_i)$, which form an ordered orthonormal basis for $\mathbb{R}_+^p$. We will concentrate our attention on the first six basis vectors, and $\mathbf{e}_i$, $i = 1, \ldots, 6$, are shown in Fig. 3. Just as eigenvalues in standard PCA correspond to the amount of variance explained by each principal component, Cooley and Thibaud (2019) show that the scales of the regularly varying principal components are given by the eigenvalues, and the first six explain 41% of the total scale. As in standard PCA, the orthogonality of the basis vectors makes interpretation of $\mathbf{e}_i$ more difficult as i increases, and $\mathbf{e}_i$ can be thought of the direction of maximum scale after accounting for the information contained in the previous basis vectors. Note in Fig. 3 that $\mathbf{e}_i > 0$, and the origin in this vector space is log(2). Therefore, we will

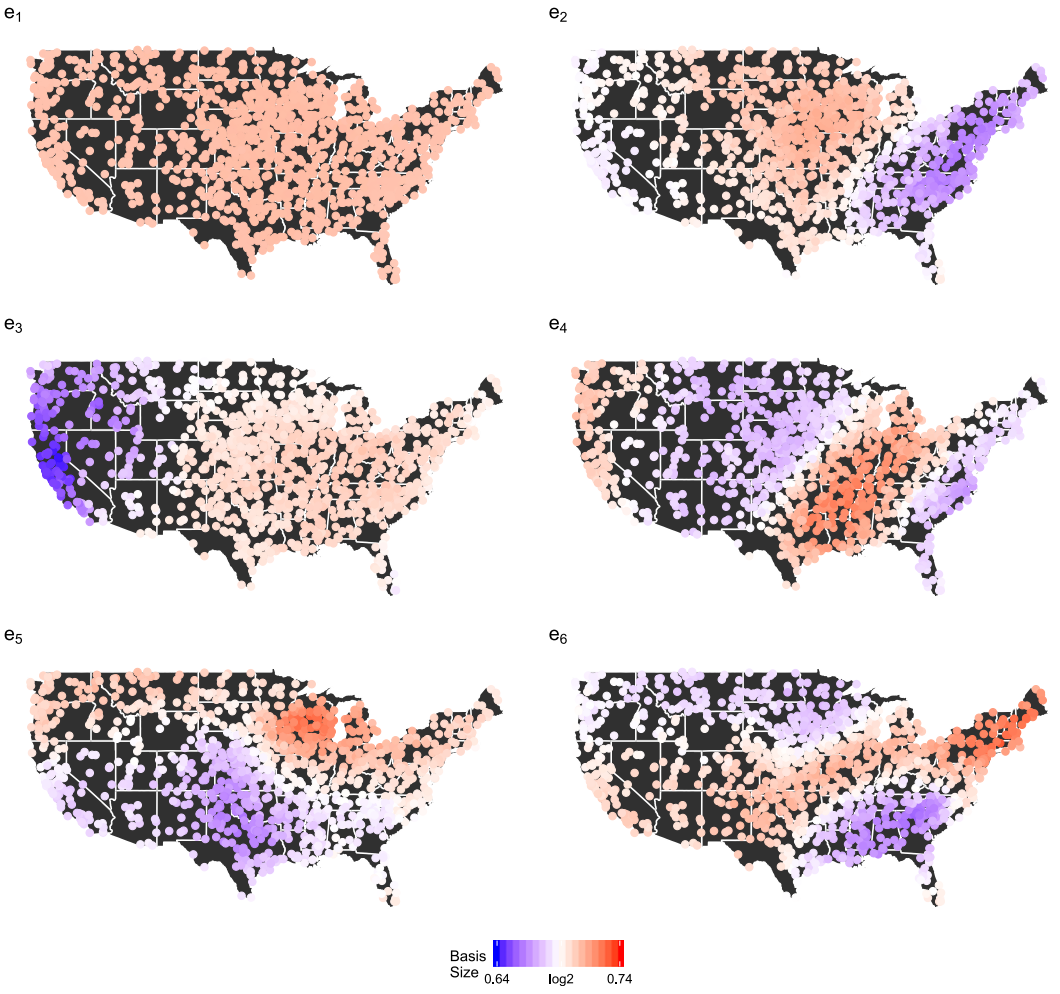

FIG. 3. Plots of the eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_6$. Each colored dot corresponds to a station location and areas lacking stations are shown in black. Of particular interest are $\mathbf{e}_1$, which gives a continental signal and whose values are all positive [>log(2)], $\mathbf{e}_2$ and $\mathbf{e}_4$, which give signals on the East Coast, with $\mathbf{e}_4$ more narrowly defined, and $\mathbf{e}_6$, which contrasts the Northeast with the Southeast.

refer to a value as "positive" if it is greater than log(2), and "negative" if it is less than log(2). Positive values are colored red and negative values are colored blue in Fig. 3.

The first basis vector $\mathbf{e}_1$ has all positive values, which is due to TPDM's property of complete positivity. Another noticeable feature is that there is little variation among the values over the contiguous United States. In section 3d, we will see that during an extreme event, this "continental" signal has a large positive coefficient, resulting in elevated values for all stations, and that subsequent eigenvectors further allocate the extreme behavior to more local regions.

The second basis vector $\mathbf{e}_2$ shows large negative values on the eastern third of the country and moderately positive values in the midcontinent. If paired with a negative coefficient, this basis would allocate extreme behavior to the east. Vector $\mathbf{e}_3$ shows a strong negative signal on the West Coast. This may seem counterintuitive at a first, since August–October is a season where the precipitation is typically not extreme on the West Coast. However, recall that the TPDM is estimated after each marginal is transformed to have a common scale, and thus extreme is defined relative to the climate for this region during this period.

The fourth basis vector $\mathbf{e}_4$ shows a narrower East Coast signal compared to the pattern of $\mathbf{e}_2$, with the transition between negative and positive roughly coinciding with the location of the Appalachian mountains. Vector $\mathbf{e}_5$ shows a contrast between the upper Midwest and south-central United States, while $\mathbf{e}_6$ shows a contrast between the Southeast and the Northeast. Since we were motivated by hurricane season, our subsequent analysis will focus on the East Coast, thus we will be particularly interested in $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_4$, and $\mathbf{e}_6$.

d. Time series of coefficients, partial basis reconstruction

Figure 4 gives the time series of the principal components $v_{t,i}$ for $i = 1, \ldots, 6$. Note that the scales of these plots decrease according to their eigenvalues, which are respectively 297.1, 47.9, 44.1, 29.5, 25.4, and 22.3.

For illustration, we examine the eigenbasis reconstruction as in (5) for the 3 days beginning 16 September 1999, when Hurricane Floyd made its landfall in Cape Fear, North Carolina, and then moved northward roughly following the coast. The first six coefficients $v_{t,1}, \ldots, v_{t,6}$ for this time are marked with a red $\times$ in Fig. 4, and the observed precipitation (after transformation) is shown in the top-left panel of Fig. 5. The coefficient $v_{t,1}$ has a large positive value of 99.9, which taken by itself would give large values across the continent. Coefficients $v_{t,2}$ and $v_{t,4}$ have values of -145.4 and -86.9, which have the largest magnitudes in these respective time series. These negative values, when combined with $v_{t,1}$, allocate the precipitation to the East Coast and diminish the signal for the rest of the country. The coefficients $v_{t,3}$ and $v_{t,5}$ have moderate values as the observed precipitation for this day generated little signal either on the West Coast or in the upper Midwest/south-central regions. The coefficient $v_{t,6}$ has a large positive value of 100.1 since the signal due to Hurricane Floyd was seen in the Northeast rather than the Southeast.

Figure 5 shows panels of the reconstruction of the Hurricane Floyd event via (5). It shows that the complete reconstruction matches the observations. The truncated reconstruction with 2, 6, 10, and 20 eigenvectors shows the increased resolution of the event as the number of eigenvectors increases. It is noteworthy that even with 20 eigenvectors included, one still does not see the fine detail of the very high levels of rain in North Carolina. Because very extreme precipitation tends to have a limited spatial effect, it is not surprising that it would require a large number of eigenvectors to see detailed effects such as this event's precipitation levels in North Carolina.

e. Further analysis of basis coefficients

One of the advantages of PCA is that the decomposition allows one to examine and test the time series of the principal components for temporal trends and also for relationships with large-scale oscillations such as ENSO. We begin with a 0.95 quantile regression of the first principal component, which tests to see if there is a linear trend in time in the continental signal. We chose 0.95 as it is high enough to be commonly considered "extreme" but low enough that an adequate amount of data remains to estimate parameters with acceptable levels of uncertainty. The estimated slope of the 0.95 quantile is 0.0019 units per year with a 95% confidence interval of (0.0011, 0.0027). Both a standard test (which does not account for temporal dependence) and a block-resample permutation test (which does) give p values of less than 1/1000, and thus there is very strong evidence for an upward trend in the large values of this continental signal. The fitted 0.95 quantile regression line is shown in the $v_{t,1}$ panel of Fig. 4.

To assess relationships between the principal components and ENSO, we obtain a yearly index by averaging NOAA's Oceanic Niño Index (ONI)[1] data for August, September, and October. We then shade the principal component time series plots (Fig. 4) according to whether ENSO is in its low (blue; <0.5°C), high (red; >0.5°C), or neutral (gray) phase. Examining the plot of the second principal component, we were struck by the appearance that many of the large negative values (corresponding to large precipitation events on the East Coast) appeared to occur in the La Niña (low ENSO) phase. Setting a threshold at the overall negated 0.95 quantile, we found that the proportion of days that exceeded this threshold during the La Niña phase was 0.067, which was greater than the 0.043 found when in its neutral or high phase. A test of whether these proportions are equal returns a p value of 0.0003, giving strong evidence that ENSO affects this East Coast signal. We further tested if the distribution of (negative) exceedances of this threshold differed with ENSO phase. A likelihood ratio test of H_0, the negatively large values follow a common GPD over all phases versus H_1: the distribution of these values is different in that the low phase rejects H_0 with a p value of 0.0475, providing some suggestion that not only do the exceedance rates differ, but also the tails themselves might differ. Although ours is a study of extreme precipitation and not exclusively hurricanes, these results about principal component 2 (PC2) are in accordance with other studies (e.g., Gray 1984; Patricola et al. 2014) showing that El Niño inhibits the Atlantic tropical cyclone activity.

Figure 6 shows a bivariate scatterplot of PC2 (East Coast signal) and PC6 (Northeast–Southeast contrast). As a large negative values of PC2 indicate a large event on the East Coast, we focus on the left side of the plot. First, it can be seen that these two principal components are not asymptotically independent as there are many points with large negative values for PC2 and large values (both positive and negative) for PC6. The points are colored to indicate ENSO phase, with blue indicating low, red indicating high, and green indicating neutral. Upon visual examination, we noticed that many of the points with large negative value for PC2 and large positive value for PC6 (indicating a large value in the Northeast) appeared to occur in the low ENSO phase. To perform a statistical test, we considered values in the two

[1] See https://www.esrl.noaa.gov/psd/data/correlation/oni.data.

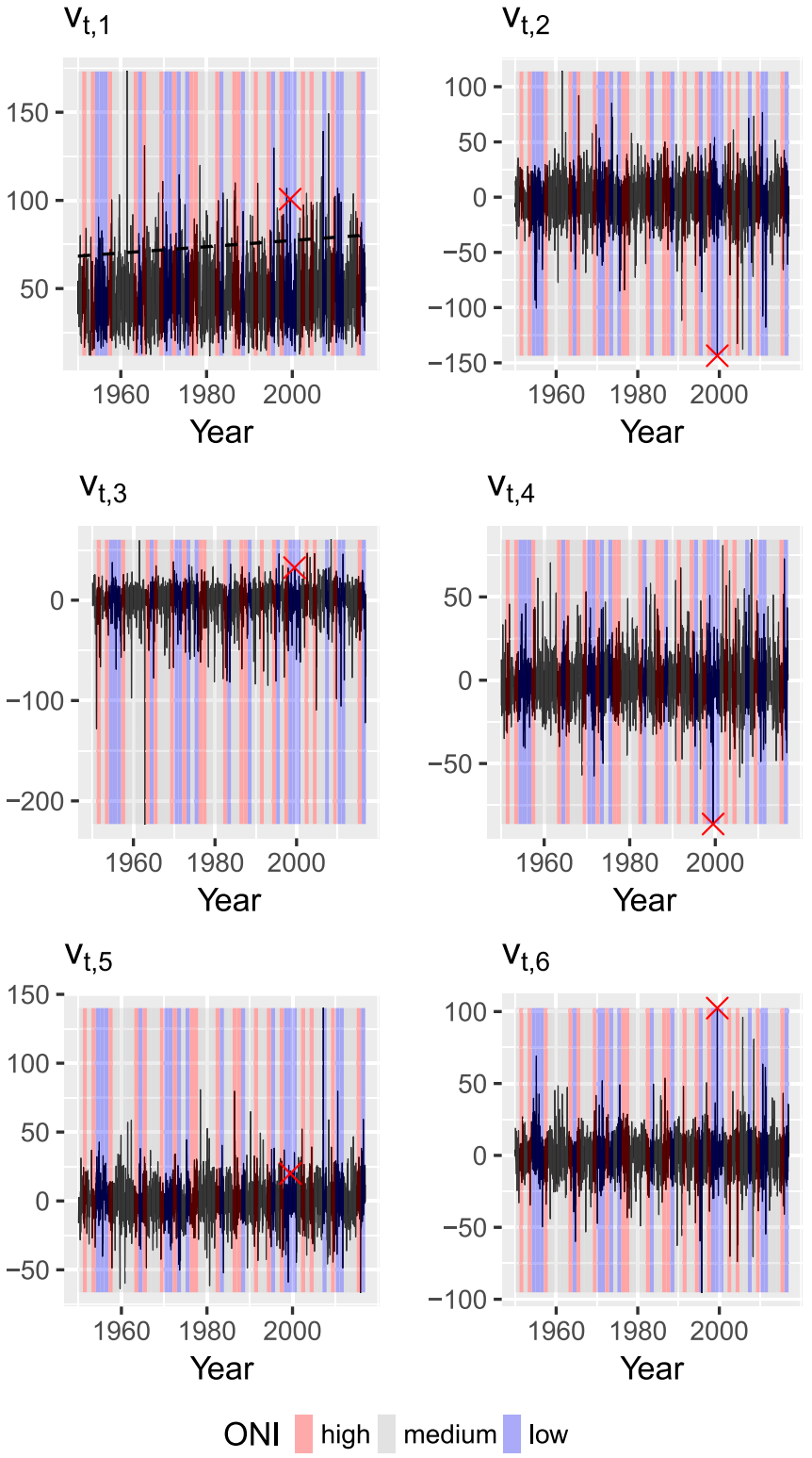

FIG. 4. Time series of the principal component scores $v_{t,i}$ for $i = 1, \ldots, 6$; that is, the coefficients corresponding to the first six eigenvectors. The date corresponding to Hurricane Floyd, 16 Sep 1999, is marked with a red $\times$. The shading corresponds to ENSO phase with blue, red, and gray indicating low, high, and neutral values, respectively. The dashed line for $v_{t,1}$ corresponds to the estimated 0.95-quantile regression line.

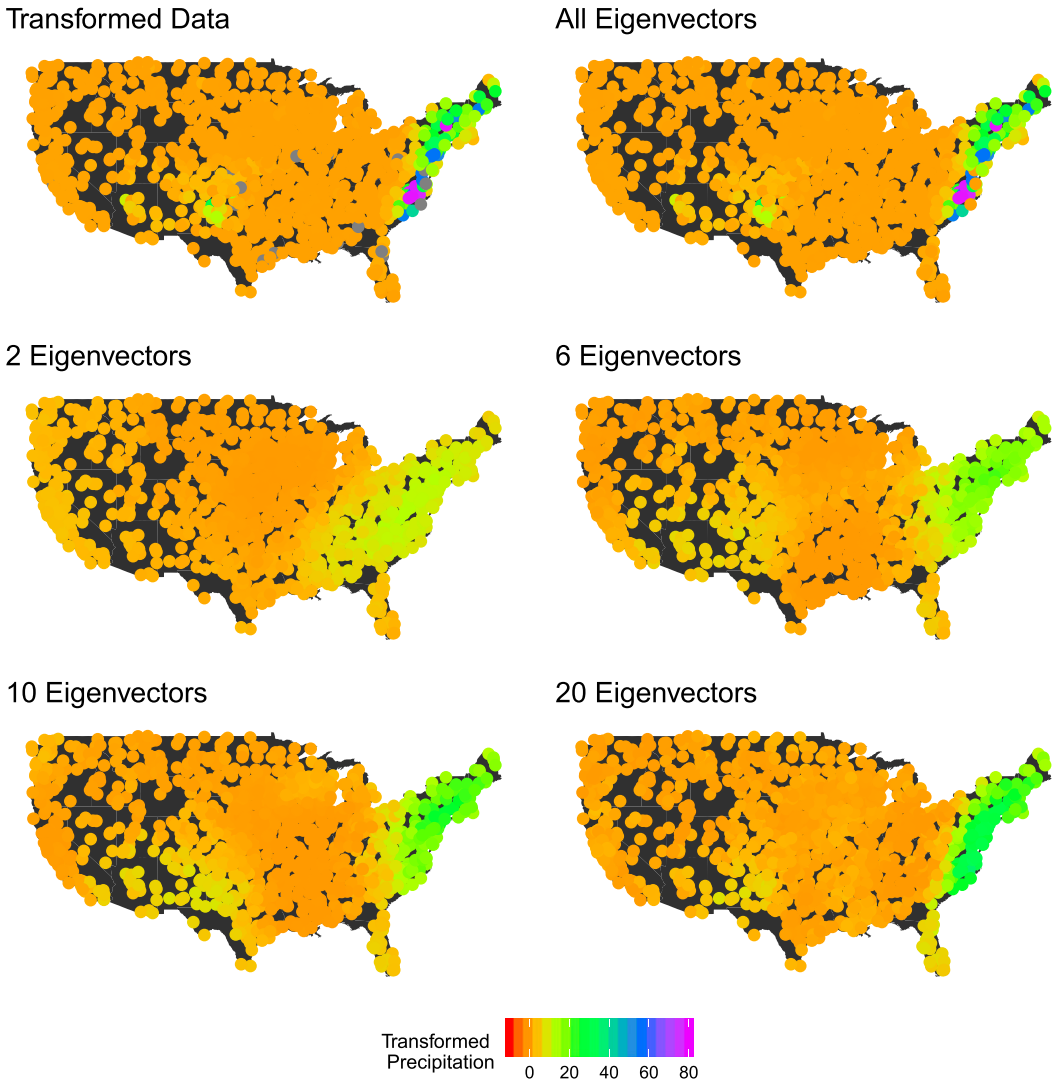

FIG. 5. Representations of the data for 16 Sep 1999, the date corresponding to Hurricane Floyd. Shown are a plot of the transformed data, a complete basis reconstruction, and truncated reconstructions with 2, 6, 10, and 20 eigenvectors.

boxes shown that capture areas where both principal components are large in magnitude. A chi-square test of equivalence of proportions of ENSO phases in the two boxes yields a p value of 0.0215. Thus, there is evidence that the proportion of events in the upper box (indicating a Northeast event) occurring during the low ENSO phase is greater than the proportion of events in the lower box during a low ENSO phase. We are unaware of any previous investigation that suggests that hurricane-season extreme precipitation in the Northeast United States is linked to La Niña conditions.

4. Discussion

We have presented a method for exploring extreme behavior of high-dimensional data by decomposing the data via a basis arising from a matrix summarizing pairwise extremal dependence. The method is analogous to PCA, but tailored for extremes. The exploratory nature of the method differs from previous atmospheric science extremes work, which has primarily aimed to quantify risk (e.g., provide an estimate of a 100-yr event) or to model phenomena (e.g., fit a max-stable process to weather stations' annual maxima). We apply the method to U.S. precipitation data and find strong evidence for a positive trend in the coefficients of the first principal component and find evidence for relationships between other principal components and ENSO. The method is general and can be used for any variable of interest.

Our exploration of the behavior of the principal components led us to perform several hypothesis tests. It could be argued that the hypothesis tests we conducted

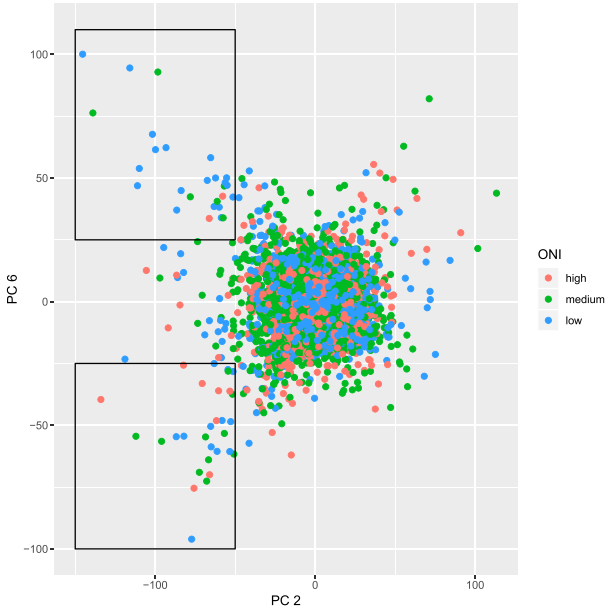

FIG. 6. Bivariate scatterplot of PC2 and PC6. Since negative values of PC2 correspond to large events on the East Coast, we focus on this portion of the plot. Recall that positive values of PC6 correspond to large events in the Northeast and negative values correspond to large events in the Southeast. The ONI indicates ENSO levels, with high values and low values corresponding to El Niño and La Niña, respectively.

suffer from test selection bias since we chose to perform them *after* examining the data (which the PCA method allows us to do in a novel way). For example, the test shown in Fig. 6 was one of a possible 15 (6-choose-2) tests pairing the first six principal components; if one were to perform all 15 tests, some multiple testing correction should be applied to the p value. Furthermore, the boxes indicated in Fig. 6 were chosen based on viewing this data and are unique to this pair of principal components, thus adding to the possibility of test selection bias. It is clearly important to keep in mind the possibility of test selection bias when interpreting the p values of these tests. However, in an exploratory analysis such as this, it is natural to pose questions based on what one discovers from the data exploration, so test selection bias is perhaps unavoidable. The real value of a test like the one illustrated in Fig. 6 is to suggest further avenues for exploration that could perhaps lead to a confirmatory analysis.

An interesting aspect of applying this method to CONUS precipitation data is that extreme precipitation is localized in its spatial extent. As seen in Fig. 5, the area where Hurricane Floyd's most extreme precipitation occurred is quite small. Thus, it is not surprising that the first six eigenvectors only explain 41% of the total scale for precipitation, and that it requires a large number of eigenvectors to recreate the small-scale features of an extreme event. Similar behavior is found when traditional PCA is applied to data with localized dependence, and PCA is often most useful in such cases because the large-scale behavior can be more difficult to visualize directly from data that appear to be dominated by local behavior. Still, we recognize that interpretation is challenging. By aggregating the localized signals from storms across more than 60 years of data, we are able to find evidence for large-scale trends and relationships with ENSO. But the leading eigenvectors we analyze do not yield information about the spatial extent of individual storms. If one needed to do a risk assessment associated with individual storms, a different type of extreme value analysis would be required.

Clearly, precipitation is not asymptotically dependent at continental scales. As asymptotic independence is a degenerate case in the regular variation framework that underlies the TPDM and our PCA decomposition, one might conclude that our method is ill-suited for this study. If one were interested in *modeling* extreme precipitation events across CONUS (a dubious proposition), one would absolutely need a model that could capture the nuanced tail dependence in order to accurately estimate probabilities in the joint tail. Here, our aim is to *explore* patterns in extreme precipitation, and the lens of the most efficient basis provided by our PCA decomposition provides a new avenue for exploration.

In this study we chose not to detrend the data. Consequently, this allowed us to quantify the significance of the trend found in the coefficients for the "continental" signal $v_{t,1}$. Our estimation of the TPDM, like the estimation of the covariance matrix in traditional PCA, assumed that the data were independent and identically distributed. An alternative modeling strategy would be to first detrend the data prior to estimating the TPDM. This of course would have involved selecting a detrending method (parametric or nonparametric), but had the trend been estimated on the "continental" scale, the extremal PCA analysis on the detrended data would likely have not seen a trend in the leading coefficients $v_{t,1}$. As with traditional PCA, but also time series analysis or geostatistical modeling, a researcher must often choose what to include as a nonstochastic factor (i.e., a trend), and what to leave in to as a stochastic component.

Formal detection and attribution analyses of long-term trends in nonextreme climate variables have used PCA methods to identify large-scale patterns of change (sometimes called fingerprints) and to test whether these trends are attributable to anthropogenic or natural forcings (e.g., Santer et al. 2004). Extensions of these methods to temperature and precipitation extremes have transformed extreme variables so that standard

PCA could be performed (Min et al. 2011, 2013; Zhang et al. 2013). The PCA method for extremes presented here may offer an alternative method for the formal detection and attribution of observed trends in extreme temperature and precipitation without making such transformations. Similar to such analyses of changes in nonextreme climate variables, comparisons of the observed patterns of extreme PCA components to those of climate model simulations with and without various natural and anthropogenic forcings would be straightforward, and well-established methods of assessing significance could then be applied. We plan on investigating such approaches in our future research.

Links to R code and data for replicating the results in this paper are available at https://www.stat.colostate.edu/~cooleyd/, as are functions to apply the methods to new datasets.

Acknowledgments. Yujing Jiang has been supported by NSF DMS-1243102. Daniel Cooley has been partially supported by the aforementioned grant as well as DMS-1811657. Michael Wehner was supported by the Regional and Global Climate Modeling Program of the Office of Biological and Environmental Research in the Department of Energy Office of Science under Contract DE-AC02-05CH11231.

REFERENCES

Ben Alaya, M., F. Zwiers, and X. Zhang, 2018: Probable maximum precipitation: Its estimation and uncertainty quantification using bivariate extreme value analysis. *J. Hydrometeor.*, **19**, 679–694, https://doi.org/10.1175/JHM-D-17-0110.1.

Coles, S. G., 2001: *An Introduction to Statistical Modeling of Extreme Values*. Springer, 208 pp.

Cooley, D., and E. Thibaud, 2019: Decompositions of dependence for high-dimensional extremes. *Biometrika*, **106**, 587–604, https://doi.org/10.1093/biomet/asz028.

Davison, A. C., R. Huser, and E. Thibaud, 2019: Spatial extremes. *Handbook of Environmental and Ecological Statistics*, A. Gelfand et al., Eds., CRC Press, 711–744.

de Haan, L., and A. Ferreira, 2006: *Extreme Value Theory: An Introduction*. Springer, 418 pp.

Gray, W. M., 1984: Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb quasi-biennial oscillation influences. *Mon. Wea. Rev.*, **112**, 1649–1668, https://doi.org/10.1175/1520-0493(1984)112<1649:ASHFPI>2.0.CO;2.

He, J., C. Deser, and B. J. Soden, 2017: Atmospheric and oceanic origins of tropical precipitation variability. *J. Climate*, **30**, 3197–3217, https://doi.org/10.1175/JCLI-D-16-0714.1.

Higham, N. J., 2002: Computing the nearest correlation matrix—A problem from finance. *IMA J. Numer. Anal.*, **22**, 329–343, https://doi.org/10.1093/imanum/22.3.329.

Kuhn, G., S. Khan, A. R. Ganguly, and M. L. Branstetter, 2007: Geospatial–temporal dependence among weekly precipitation extremes with applications to observations and climate model simulations in South America. *Adv. Water Resour.*, **30**, 2401–2423, https://doi.org/10.1016/j.advwatres.2007.05.006.

Larsson, M., and S. I. Resnick, 2012: Extremal dependence measure and extremogram: The regularly varying case. *Extremes*, **15**, 231–256, https://doi.org/10.1007/s10687-011-0135-9.

Menne, M. J., I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston, 2012: An overview of the Global Historical Climatology Network-Daily database. *J. Atmos. Oceanic Technol.*, **29**, 897–910, https://doi.org/10.1175/JTECH-D-11-00103.1.

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

——, ——, ——, H. Shiogama, Y.-S. Tung, and M. Wehner, 2013: Multimodel detection and attribution of extreme temperature changes. *J. Climate*, **26**, 7430–7451, https://doi.org/10.1175/JCLI-D-12-00551.1.

National Academies of Sciences, Engineering, and Medicine, 2016: *Attribution of Extreme Weather Events in the Context of Climate Change*. National Academies Press, 186 pp.

Patricola, C. M., R. Saravanan, and P. Chang, 2014: The impact of the El Niño–Southern Oscillation and Atlantic meridional mode on seasonal Atlantic tropical cyclone activity. *J. Climate*, **27**, 5311–5328, https://doi.org/10.1175/JCLI-D-13-00687.1.

Resnick, S., 1987: *Extreme Values, Regular Variation, and Point Processes*. Springer-Verlag, 320 pp.

——, 2002: Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, **5**, 303–336, https://doi.org/10.1023/A:1025148622954.

——, 2007: *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 404 pp.

Rootzén, H., and N. Tajvidi, 2006: Multivariate generalized Pareto distributions. *Bernoulli*, **12**, 917–930, https://doi.org/10.3150/bj/1161614952.

Santer, B. D., and Coauthors, 2004: Identification of anthropogenic climate change using a second-generation reanalysis. *J. Geophys. Res.*, **109**, D21104, https://doi.org/10.1029/2004JD005075.

Timmermans, B., M. Wehner, D. Cooley, T. O'Brien, and H. Krishnan, 2019: An evaluation of the consistency of extremes in gridded precipitation data sets. *Climate Dyn.*, **52**, 6651–6670, https://doi.org/10.1007/s00382-018-4537-0.

Uvo, C. B., 2003: Analysis and regionalization of northern European winter precipitation based on its relationship with the North Atlantic oscillation. *Int. J. Climatol.*, **23**, 1185–1194, https://doi.org/10.1002/joc.930.

Weller, G. B., D. Cooley, S. R. Sain, M. S. Bukovsky, and L. O. Mearns, 2013: Two case studies on NARCCAP precipitation extremes. *J. Geophys. Res. Atmos.*, **118**, 10 475–10 489, https://doi.org/10.1002/jgrd.50824.

Widmann, M., and C. Schär, 1997: A principal component and long-term trend analysis of daily precipitation in Switzerland. *Int. J. Climatol.*, **17**, 1333–1356, https://doi.org/10.1002/(SICI)1097-0088(199710)17:12<1333::AID-JOC108>3.0.CO;2-Q.

Wilks, D., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Academic Press, 704 pp.

Zhang, X., H. Wan, F. W. Zwiers, G. C. Hegerl, and S.-K. Min, 2013: Attributing intensification of precipitation extremes to human influence. *Geophys. Res. Lett.*, **40**, 5252–5257, https://doi.org/10.1002/grl.51010.