Using Self-Organizing Maps to Identify Coherent CONUS Precipitation Regions

LEIF M. SWENSON AND RICHARD GROTJAHN

Department of Land, Air and Water Resources, University of California, Davis, Davis, California

(Manuscript received 15 May 2019, in final form 12 August 2019)

ABSTRACT

Extreme precipitation events have major societal impacts. These events are rare and can have small spatial scale, making statistical analysis difficult; both factors are mitigated by combining events over a region. A methodology is presented to objectively define ''coherent'' regions wherein data points have matching annual cycles. Regions are found by training self-organizing maps (SOMs) on the annual cycle of precipitation for each grid point across the contiguous United States (CONUS). Using the annual cycle for our intended application minimizes problems caused by consecutive dry periods and localized extreme events. Multiple criteria are applied to identify useful numbers of regions for our future application. Criteria assess these properties for each region: having many more events than experienced by a single grid point, good connectedness and compactness, and robustness to changing the number of regions. Our methodology is applicable across datasets and is tested here on both reanalysis and gridded observational data. Precipitation regions obtained align with large-scale geographical features and are readily interpretable. Useful numbers of regions balance two conflicting preferences: larger regions contain more events and thereby have more robust statistics, but more compact regions allow weather patterns associated with extreme events to be aggregated with confidence. For 6-h precipitation, 12–15 regions over the CONUS optimize our metrics. The regions obtained are compared against two existing region archetypes. For example, a popular set of regions, based on nine groups of states, has less coherent regions than defining the same number of regions with our SOM methodology.

1. Introduction

One of the fundamental problems in researching extreme events is finding a large enough sample size to make the statistics robust. A common way to build sample size is to aggregate events within some geographic area (e.g., Kunkel et al. 1993; Karl and Knight 1998; Grotjahn and Faure 2008; Kunkel et al. 2012). Aggregation can be effective for precipitation extremes because precipitation varies across smaller scales than other atmospheric variables (e.g., temperature and pressure) (Hewitson and Crane 2005).

Other climate regionalizations have been made before, most notably in Karl and Knight (1998), Kottek et al. (2006), and Bukovsky (2011). Previous regionalizations are not suitable to aggregating extreme precipitation events for one or more reasons: the regions are too large, the regions are too discontinuous over mountainous regions, the regions are partly defined from elevation, the regions are defined from combinations of meteorological variables, or the regions are partially based on local vegetation. In addition, these regionalizations are created or modified by subjective factors like consensus among researchers or intuition. Our method uses precipitation solely and the shape and number of the regions result from predefined criteria.

This paper presents an objective way to select geographic areas for grouping extreme precipitation events by training self-organizing maps (SOMs) on the normalized annual cycle of precipitation. Therefore, each region contains points having similar seasonal cycle. The seasonal cycle is normalized so that the regions are not influenced by the size of the total annual precipitation.

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-19-0352.s1.

Corresponding author: Leif M. Swenson, lmswenson@ucdavis.edu

DOI: 10.1175/JCLI-D-19-0352.1

Removing the total allows us to find larger-scale patterns and not have SOM-based regions that merely show topographic elevation or proximity to the ocean. Seasonality is emphasized because extreme precipitation events and the meteorological drivers behind them can be mainly seasonal (Kunkel et al. 2012). In some areas wintertime precipitation is almost exclusively caused by frontal systems whereas other areas receive most of their summertime precipitation from convective systems. Training a SOM on the normalized annual cycle of precipitation will therefore more likely contain similar extreme precipitation events occurring at different places in the region. The paper is organized as follows. Section 2 details the different datasets used. Section 3 describes the methodology. Section 4 tests the approaches. Sections 5 and 6 apply the criteria and compare reanalyses. Section 7 compares the maps obtained by our method to other regionalizations in the literature, and concluding remarks are provided in section 8.

2. Data

Climate Forecast System Reanalysis (CFSR) precipitation data (Saha et al. 2010a) are emphasized for training the self-organizing map. CFSR uses the Climate Forecast System model (CFS) data to generate a continuous best estimate of the state of the ocean–atmosphere system (Saha et al. 2010b). CFSR is chosen since it has all the fields we need to create algorithms (in later work) that diagnose the meteorological drivers of precipitation. The CFSR temporal resolution of four times a day will allow our diagnostics to capture the individual process(es) driving each extreme event. CFSR incorporates hourly input data. The CFS model has T382, or approximately 35-km horizontal resolution, using a sigma-pressure hybrid vertical coordinate with 64 levels and a top pressure of ~0.266 hPa. The CFSR has 0.5° resolution in both latitude and longitude and temporal resolution of 6 h. Our time period is 1 January 1979–31 December 2010.

As a cross check we also apply our methodology to the Climate Prediction Center's (CPC) unified precipitation data (Chen et al. 2008a). The CPC data are based on a rain gauge network spanning the conterminous United States (CONUS) and have been interpolated to a latitude–longitude grid using an optimal interpolation objective analysis technique as in Xie et al. (2007) and Chen et al. (2008b). These gridded precipitation data have a resolution of 0.25° in both latitude and longitude and are recorded daily as an accumulation from 1200 UTC of the day before to 1200 UTC of the current day. We use data from 1 January 1950 through 31 December 2018.

3. Methodology

The goal is to train a SOM to divide the CONUS into regions with similar precipitation characteristics. These regions are called ''coherent regions'' in this report. Because extreme precipitation is likely seasonal, we use the annual cycle of precipitation to create our regions. SOMs are a type of artificial neural network first introduced by Kohonen (1982). SOMs utilize a competitive and unsupervised learning algorithm to produce a lower-dimensional representation of the input data; in this case almost 6000 average annual cycles, one for each grid point, are grouped into a much more manageable and representative ~15 average annual cycles, one for each region. Each region's annual cycle is the average of the average annual cycle of each grid point within that region. Another feature is that the ~15 representative annual cycles are ordered by similarity, that is, the annual cycles of region $K = 1$ and of the highest K value region (e.g., $K \sim 15$) are the most dissimilar. This feature makes it simple to see the full range of patterns extracted from the input data. It should be noted that the two most dissimilar regions are often adjacent, as seen in Figs. 1a–f. This indicates that two decidedly different regimes abut one another. Where this occurs with a jagged or messy border suggests that, perhaps due to interannual variations, the boundary is uncertain. Precipitation has a highly skewed distribution, which makes any measure of the annual cycle very noisy. It is therefore useful to take the cube root of precipitation before creating the measure of the annual cycle as that operation is observed to transform precipitation data to an approximately normal distribution (Stidd 1953). Processes are described in the next section that were discarded (sections 4a–c) in favor of using the cube root, which will be discussed further in section 4d. All leap days are then removed, and a long-term daily mean (LTDM) is created by averaging all the 1 January data from every year, the 2 January data from every year, and so on. This is done for each grid point individually so the end result is a time series of the cube root of precipitation with 365 values for each grid point. We care most about the timing of precipitation and want the SOM to be able to easily compare climatologically wet and dry areas. Because of these concerns, each time series is adjusted so that the range of the data is from 0 to 1. The procedure is to subtract the minimum value of each time series from every value in the time series and divide the result by the new maximum of that time series. This normalized annual cycle of precipitation will henceforth be referred to as the LTDM-n. The LTDM-n allows the methodology to compare the occurrence of the wettest days at different

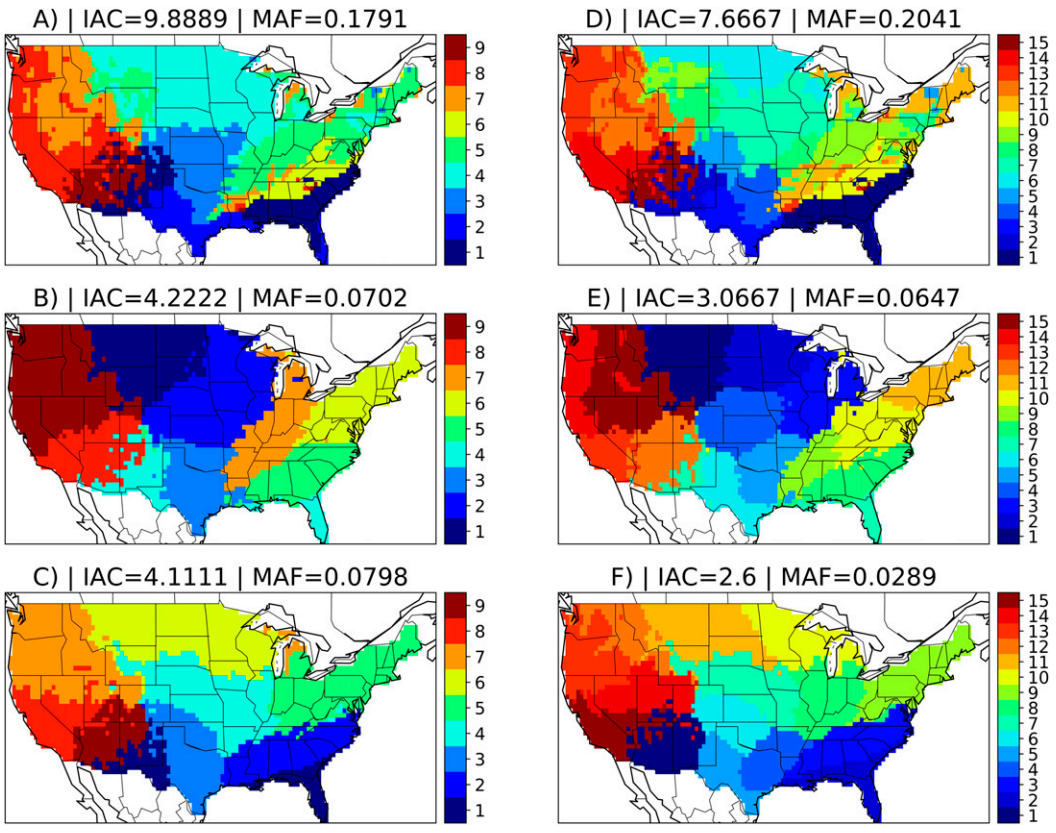

FIG. 1. SOM regions created from (a) the first six harmonics of the long-term daily mean (LTDM) of precipitation at each grid point, $K = 9$; (b) the LTDM of precipitation, adjusted to vary from 0 to 1, at each grid point where $K = 9$; and (c) the LTDM of the cube root of precipitation, adjusted to vary from 0 to 1, at each grid point where $K = 9$. (d) As in (a), but $K = 15$. (e) As in (b), but $K = 15$. (f) As in (c), but $K = 15$. IAC is the mean value of the isolated area count, and MAF is the mean value of the minor areas fraction. Smaller values are preferable for both IAC and MAF.

locations instead of simply creating regions based on annual rainfall amount.

Without some form of LTDM it is very common to get long strings of zeros in a daily precipitation accumulation time series. These strings create a false similarity between distant, climatologically dry, areas. Taking an LTDM reduces the chances of zero values occurring and separates climatologically dry areas better. An added benefit is that a representative seasonal cycle will be created for each region. Future applications may desire to focus on a specific season (e.g., winter in California) and having seasonality built into the SOMs is advantageous.

To test that the regions created by the SOM are statistically distinguishable, a test based upon the ''false discovery rate'' (FDR) is used (Johnson 2013). This test checks if the LTDM-n at each grid point in a region is significantly different than the LTDM-n in every other region at the 5% level. This method provides a useful upper bound on the number of regions we can find across the CONUS. This upper bound is 32 regions for the CFSR data and 63 regions for the CPC data. To find an ''optimal'' number of regions (K_B) rather than merely rely on the upper bound presented in (Johnson 2013), five metrics of four criteria are considered: connectedness, robustness, compactness, and the number of extreme events in each region during the record.

We prefer that each SOM-based region be contiguous, a property we label *connectedness*. Connectedness has two elements. First, we want to minimize the number of separate areas that comprise each SOM region. Second, we prefer each SOM region to be mainly a single larger area and that any other areas be individually and collectively small. To measure this connectedness attribute, two metrics were designed. The first counts the number of isolated areas that belong to each region, where an isolated area is a continuous group of grid points from one single region entirely surrounded by grid points from other regions. This metric for the mean number of isolated areas composing the regions

will be called the isolated area count (IAC). This average number of separate areas per region varies from a minimum of 1 to a maximum of about 10 in CFSR data and about 30 in CPC data. The second measure of connectedness recognizes that a region broken into several small disconnected areas and one large connected area is preferable to a region broken into approximately equally sized disconnected areas. To this end the second metric computes, for each region, the ratio of grid points not belonging to the largest isolated area to the region's total number of grid points. This metric of the fraction of areas in the region that are not part of the largest area will be called the minor areas fraction (MAF). The CONUS average of the MAF is used. We specify that no region may have connectedness metrics above 6 and 0.25, respectively, for the two metrics discussed above.

Another criterion is *robustness*, which refers to how much each SOM region boundary changes when the number of SOM regions allowed is incremented; the less change to existing region boundaries by the addition of another region the better. To measure this quantity, for each grid point we count the other grid points that are in the same region as the particular grid point for a given value of K. If, for $K + 1$ regions, any of the counted grid points are no longer in the same region as the particular grid point, we discard them from our count and divide the reduced count by the original count (for K regions). Even if the region were to grow in size, we are only considering grid points that were part of the original count. Hence, 1 is the largest this metric can be. It is expected that regions will shrink when K is increased but it is possible for a region to grow in size. If a particular region grows in size from a map with K regions to a map with $K + 1$ regions then robustness can also be one for grid points in that region. However, one or more adjacent regions will have shrunk, thereby lowering the robustness scores of their grid points. The map average of these ratios (i.e., the average of the ratios at all the grid points) is between 0 and 1. A map average near 1 is ideal, with lower values indicating that a map with K regions is not as robust. We specify that a map average ratio be ≥ 0.65 for the map to be considered adequately robust.

The last criterion, *compactness*, is intended to foster our eventual goal of compositing events within each region. Long and thin regions are a hindrance to compositing because when a region extends too far in one direction, lining up the origins of particular events becomes problematic. In each region we calculate the nondimensional ratio (or compactness ratio) of the square root of the total area encompassing the largest connected group of grid points divided by the perimeter of that area to evaluate how compact each region is. We specify that no region may have a compactness ratio below 0.075. For reference a perfectly circular region would have a compactness ratio of ~0.28. The set of $K = 15$ regions has a median compactness ratio of 0.14.

Thresholds introduced are not general. After looking at maps for K values of 2–63 we subjectively decided on these thresholds of 6 and 0.25 for connectedness, 0.65 for robustness, and 0.075 for compactness. The thresholds are a way of establishing minimal qualifications for each criterion.

4. Approaches considered to constrain the annual cycle

This study uses the normalized LTDM of the cube root of precipitation, but it is useful to see other approaches considered and rejected. Different methods can find an annual cycle. These include different types of harmonic analysis where more or fewer harmonics are retained depending on the time scale of interest. One could also use a measure of central tendency for each day of the year across all years in the data. Four measures of the annual cycle of precipitation are discussed.

One notes that very robust, connected, and compact regions can be made simply by training a SOM on the full, unprocessed, daily precipitation record at each grid point. For our purposes there are two issues with this method of regionalization. First, using the full record does not emphasize seasonality, which is undesirable for the reasons discussed above. Second, this method yields a precipitation time series for each region that is 32 years long and much less readily interpretable than a measure of the annual cycle. A figure of the regions produced by training a SOM on the full, unprocessed, daily precipitation records is available in the online supplemental material.

a. Harmonic LTDM to capture the annual cycle

Because we are most interested in what time of year precipitation tends to occur in a specific area, the logical first step in creating SOM-based precipitation regions is to train the SOM on the annual cycle of precipitation. Prior studies of the annual cycle of temperature (Grotjahn 2011; Grotjahn and Zhang 2017) found that a limited set of harmonics captures a smoothly varying LTDM. Harmonics work because the day-to-day variability remains quite large even when averaging 60 years of data. Given their success in creating a smoothly varying annual cycle by retaining only the first six harmonics of the LTDM of temperature, wind, and geopotential height, we tested a similar filtering on precipitation. This filtering creates a smoothly varying

annual cycle of precipitation, here called the harmonic long-term daily mean (HLTDM). Choosing the number of harmonics to retain in order to adequately represent the annual cycle of precipitation is unclear. In Wang and LinHo (2002) 12 harmonics are used to capture the onset of the Asian monsoon, but in an earlier paper (Wang 1994) only 4 harmonics are used. The older paper was not concerned with the precise timing of the onset, but rather with how the intensity changes from year to year. The newer paper sought to identify the time of monsoon onset, which required a more detailed representation of the annual cycle. To choose the number of harmonics here, we observed the average difference between the LTDM and the HLTDM at each grid point versus how many harmonics we kept. This analysis led us to use six harmonics because after the sixth harmonic the additional reduction in difference between the LTDM and the HLTDM, in our map average, becomes small relative to the reduction in difference gained when adding each of the first six harmonics.

Maps created using six harmonics (Figs. 1a,d) were unsatisfactory in that they consistently failed our tests of connectedness by exceeding the thresholds mentioned above. The most prominent cause of disconnectedness is a persistent link between Wyoming and Illinois, which belong to disconnected parts of the same region. This is illustrated in Fig. 1d for region 8 where southern Illinois and eastern Missouri are part of the same region as most of Wyoming. The similarity between the HLTDMs of the geographically separate areas of Wyoming and Illinois, among other disconnected regions, indicates that gross seasonality is not enough to define coherent precipitation regions. Subseasonal variations should play a stronger role in identifying our regions if we desire connectedness, especially on smaller scales. With that in mind we experimented with keeping more harmonics when constructing our HLTDM and we did find that the more harmonics that were kept, the better the map scored in each of our criteria. However, part of that better score is coming from large variations on very short time scales still present after averaging the data over the full period of record, whereas we prefer matching the broader seasonal cycle.

b. *Long-term daily median of precipitation*

To retain day-to-day variation, we considered constructing a time series from a central tendency of the precipitation distribution for each day of the year for each grid point. Because precipitation is known to have a very skewed distribution (Ison et al. 1971) we tested the median as our central tendency. Unfortunately, there are multiple locations in Arizona and New Mexico where the median precipitation is 0 for every day of the year. This does not reflect the seasonality in precipitation that exists in those areas and makes the long-term daily median a poor choice to represent the annual cycle of precipitation.

c. *LTDM of precipitation with and without a 3-day running mean*

Since a long-term daily median is a poor choice to measure the annual cycle we trained SOMs on the LTDM of precipitation at each grid point (Figs. 1b,e). This LTDM is much noisier and less intuitive than a HLTDM or long-term daily median but does avoid the issues discussed in the previous two subsections. This trade-off also creates regions that have better connectedness than regions made using a HLTDM or long-term daily median but there are still areas of significant disconnectedness. Even averaging 32 years of data, there remains significant variation from one day to the next in the LTDM. Large daily variations are problematic so we attempt to soften them with a simple nonrecursive smoother so we used a simple nonrecursive smoother in time on the raw precipitation data before creating the LTDM. The smoothed value for day X equals one quarter of day $X - 1$ plus one-half of day X plus one quarter of day $X + 1$. A LTDM is then created from the smoothed data at each grid point and used to train the SOM. The smoothing altered the shapes of the SOM regions, but the connectedness metric was not sufficiently improved. Disconnected regions moved, but their number was not appreciably reduced. The smoothing had insufficient benefit to any of the four criteria overall.

d. *Working with the cube root of daily precipitation data*

It is well known that precipitation data fit a gamma distribution (Ison et al. 1971) better than a normal distribution. Therefore, transforming the gamma-distributed precipitation data to a more normal distribution by taking the cube root of precipitation (Stidd 1953) was tested. This operation has a much larger effect on large precipitation values than on small precipitation values, effectively reducing the impact of extreme data. This reduction is valuable because we want the SOMs to be based on the seasonal cycle, which facilitates coherence in space and has links to extreme precipitation mechanisms mentioned earlier. Additionally, we found that even the LTDMs of neighboring grid points had large differences in their peak values. Taking the cube root of the raw precipitation data de-emphasizes the spatial variation of these peak values. This method still retains the seasonal cycle of precipitation and incorporates a reduced form of subseasonal variation leading to maps that consistently score better in all four criteria than the maps based on any

of the methods discussed above. For low K values the map using the cube root of precipitation performs marginally better than the map without taking a cube root. The superiority of this method over using the LTDM without taking the cube root grows as K increases and is easily seen for maps with higher values of K (Fig. 1).

5. Results

a. Criteria values as a function of K

When making composites we increase the number of events sampled by expanding the area we aggregate over. Because we are analyzing extreme events here we use the raw (i.e., no normalization or taking of the cube root) precipitation data. We define an extreme event to be a 6-h period that exceeds the 95th percentile for precipitation accumulation at a particular grid point, after discarding time steps with zero precipitation. For each region we count the number of time steps when any grid point reports an extreme as a ''regional event.'' Creating a threshold for the number of these regional events that each region must have would necessarily base the threshold on the driest (fewest precipitation periods) region found by the SOM. Instead, we want each region to aggregate more extreme events than could be found from any single grid point within it. To this end we calculate a regional extremes ratio (RER). The RER is the number of time periods when a point somewhere in the region exceeds the 95th percentile divided by the number of time periods when there is rain somewhere in the region. Since RER is a ratio based on exceeding the 95th percentile, then the RER for an individual grid point would be essentially 0.05 or 5%. We would like our SOM regions to capture at least 4 times as many extreme events as an individual grid point would, so we apply a RER threshold of 0.2 or 20% criterion to our analysis. The RER allows comparison of extremely dry and extremely wet areas of CONUS more than would a fixed number of events. This ratio also allows intercomparison of different datasets having differing periods of data, grid intervals, and/or time intervals. Figure 2 shows the relationship in CFSR data between K and RER in the region with the lowest RER along with the 20% threshold. From Fig. 2 it is clear that using more regions means that each of those regions aggregate fewer events and asymptotically approach the value (RER = 5%) for a single grid point. This simple RER threshold indicates that values of $K > 15$ are eliminated from further consideration in the CFSR data.

However, Figs. 3a–c show that as the number of regions is increased the compactness ratio and IAC improve dramatically and MAF sees modest improvement

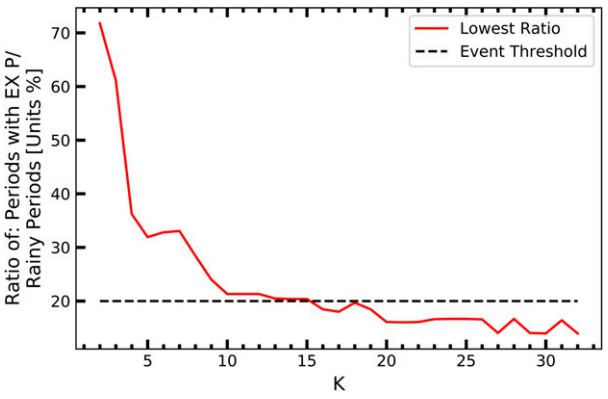

FIG. 2. The ratio of periods with extreme precipitation (>95% value) to periods with nonzero precipitation (RER) in the CFSR data contained within the region with the lowest RER for each value of K is shown in red. Our threshold of 20% (meaning 4 times as many periods with an extreme somewhere in the region as occur at a single grid point) is shown by the dashed line.

in the worst region's ratio. But increasing the number of regions decreases the number of events in each region by decreasing the average areal extent of each region over which we can aggregate events. Therefore, a balance is sought between two competing goals. The first goal is to create regions that are large enough to have meaningfully more events to aggregate compared to considering a single grid point. The second goal is to create regions that attain high scores in compactness and low (better) scores in IAC and MAF. Creating a map with fewer regions helps the first goal while creating a map with more regions helps the second goal.

Of the K values that pass the event threshold described (2–15) the values that do not meet our robustness criteria (values < 0.65 in Fig. 3d) are discarded. The remaining values (K = 3, 5–13, and 15) are ranked by their median score in compactness ratio, and their worst region's scores in MAF and IAC. The worst region's score was judged to be more important than the median for the purpose of selecting the optimal value of K due to the lack of variation in the median scores of MAF and IAC. In the case of IAC this was particularly true for $K > 10$. The ''optimal'' K value with the best average rank is 15 in this paper. One notes that $K = 12$ also does very well in this comparison and is the most compact of all the maps considered.

An odd characteristic is a persistent link between the Florida (FL) peninsula and the New Mexico/Texas (NM/TX) border area, especially when K is small ($K < 10$). At first glance this seems very strange because peninsular Florida has a warm, humid, tropical climate whereas the New Mexico and Texas border area is arid (Kottek et al. 2006). This linkage also shows up in the

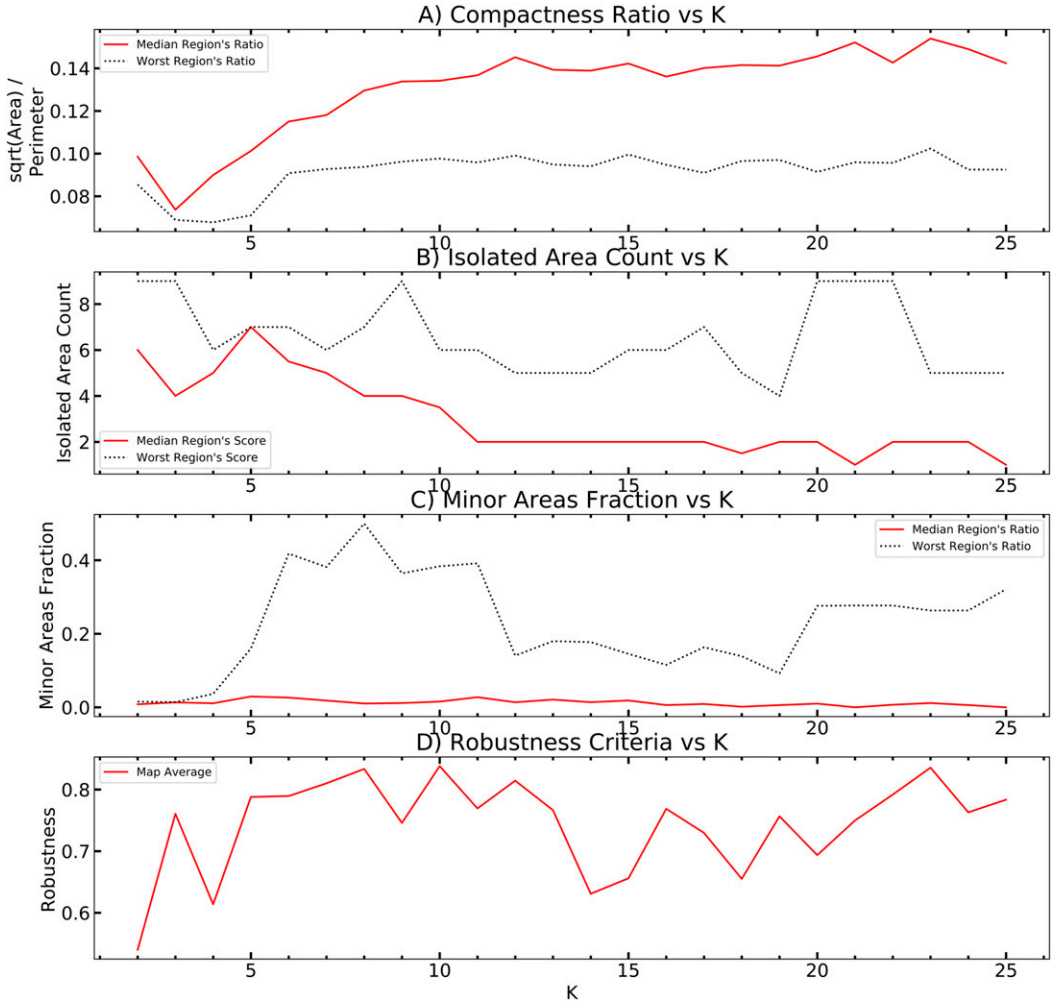

FIG. 3. Performance criteria for CFSR-based SOM regions created from the LTDM of the cube root of precipitation, adjusted to vary from 0 to 1, at each grid point. (a) Number of regions vs compactness ratio. The dashed curve is the least compact region, and the red curve is the median region's compactness ratio. (b) Number of regions vs isolated area count. The dashed curve is the worst performing region, and the red curve is the median isolated area count of the regions. (c) Number of regions vs the minor areas fraction. The dashed curve is the worst preforming region, and the red curve is the median minor areas fraction of the regions. (d) Number of regions vs robustness. The red curve shows the map average for robustness for the shown values of K. Preferred metric values are higher in (a) and (d), and lower in (b) and (c).

CPC dataset, which is based on observations, ruling out model error as causing this unexpected similarity. Our procedure to normalize the data at each grid point causes the grouping of New Mexico with Florida. Both areas share a late summer relative peak in precipitation and, while Florida is much wetter, the normalization of precipitation magnitudes causes Florida's humid maritime seasonal cycle to closely match New Mexico's arid continental cycle. For compositing events, having similarly large spatially separated areas grouped together is very undesirable. We design a process to identify and label as separate regions large, spatially separate areas that the SOM analysis assigns to one region. This process is discussed further in section 6.

b. *Variations within the CONUS of the annual cycle*

Figure 4 illustrates how the normalized annual precipitation cycle varies across the CONUS; the spatial variation is similar for smaller and larger numbers of regions. The $K = 12$ regions version is emphasized because it has the most compact regions of all maps that were considered ($K = 2$–15). All along the west coast the wet season starts in November and ends around May or June. Regions 10, 11, and 12 are then primarily

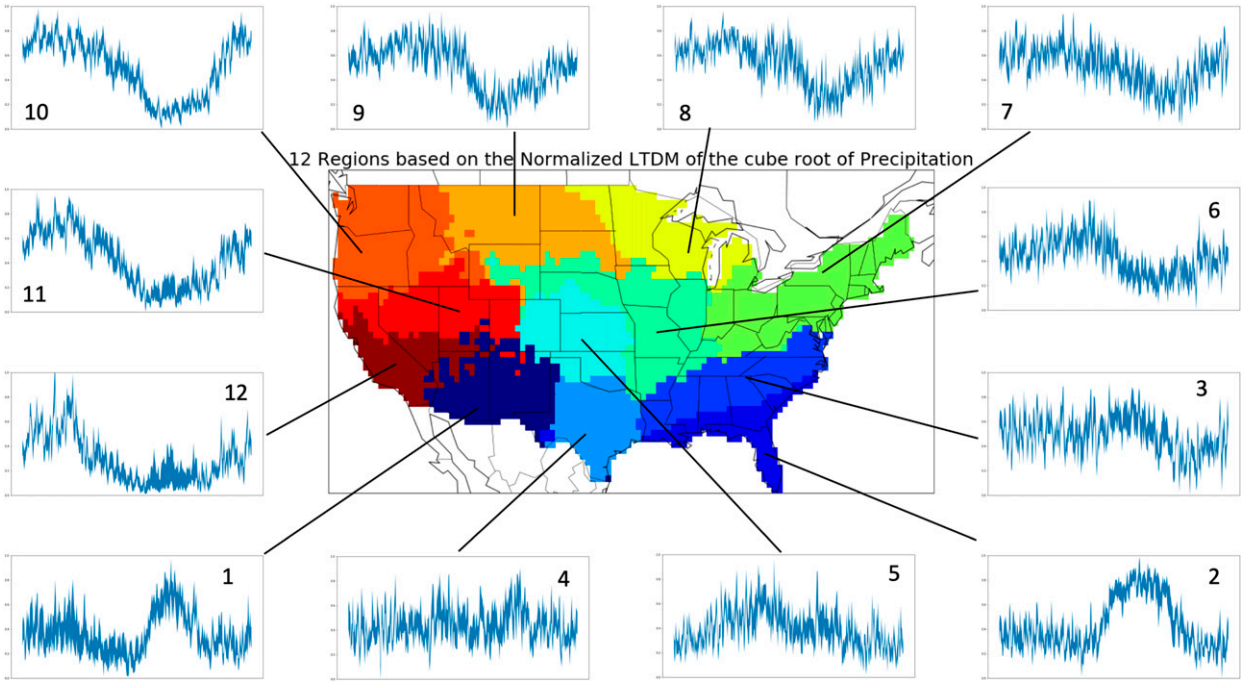

FIG. 4. SOM regions with $K = 12$ created from the LTDM of the cube root of precipitation, adjusted to vary from 0 to 1, at each grid point. The surrounding plots show the representative LTDM of the cube root of precipitation, adjusted to vary from 0 to 1, at each grid point. The middle 50% of grid points in each region are contained in the shaded area of each subplot. The subplots each begin at 1 January on the left and end on 31 December on the right. The value $K = 12$ is chosen because it separates the NM/TX border region from Florida in these CFSR data and has the highest compactness ratio of all maps considered ($K = 2$–15).

differentiated by the relative strength of their dry seasons compared to the median value of their annual cycle. This strength decreases as we move from north to south. In the southwest, region 1 encompasses most of Arizona and New Mexico; it has both a weak wintertime wet season (especially January–March) and much stronger late summer to early fall wet season (July–September). Hence, this area gets some wintertime precipitation, probably from Pacific storm tracks, but is more influenced by the North American monsoon (NAM; Adams and Comrie 1997) in the late summer. The large late summer peak in precipitation makes the annual cycle here similar to the annual cycle in the Florida peninsula as discussed above. The northern Great Plains are dominated by wintertime precipitation with a fairly strong dry season from July into November (region 9). Moving south through the Great Plains we see a gradual flattening of the annual cycle down into south Texas (regions 6, 5, and 4). In the southeast and inland from Florida, region 3 has a weak dry period (October–December) leading into a winter and spring that are near the annual mean precipitation rate with a slightly wetter summer. The northeast (region 7) is fairly wet until August and then becomes wet again in November. Eight of the twelve regions in this map are to some degree wintertime dominated when it comes to precipitation.

6. Comparison between CFSR and CPC

When our criteria with respect to K are examined to find the optimal value of K for SOMs trained on the CPC data, broadly similar results are found, with a few key differences (Figs. 5 and 6). Comparing the regions with the fewest unique events, for a given value of K, in the

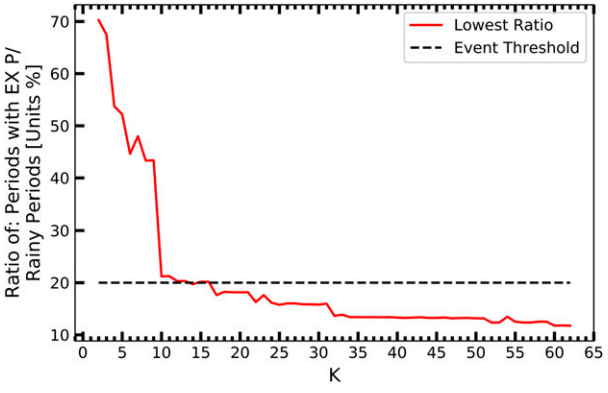

FIG. 5. As in Fig. 2, but for CPC data.

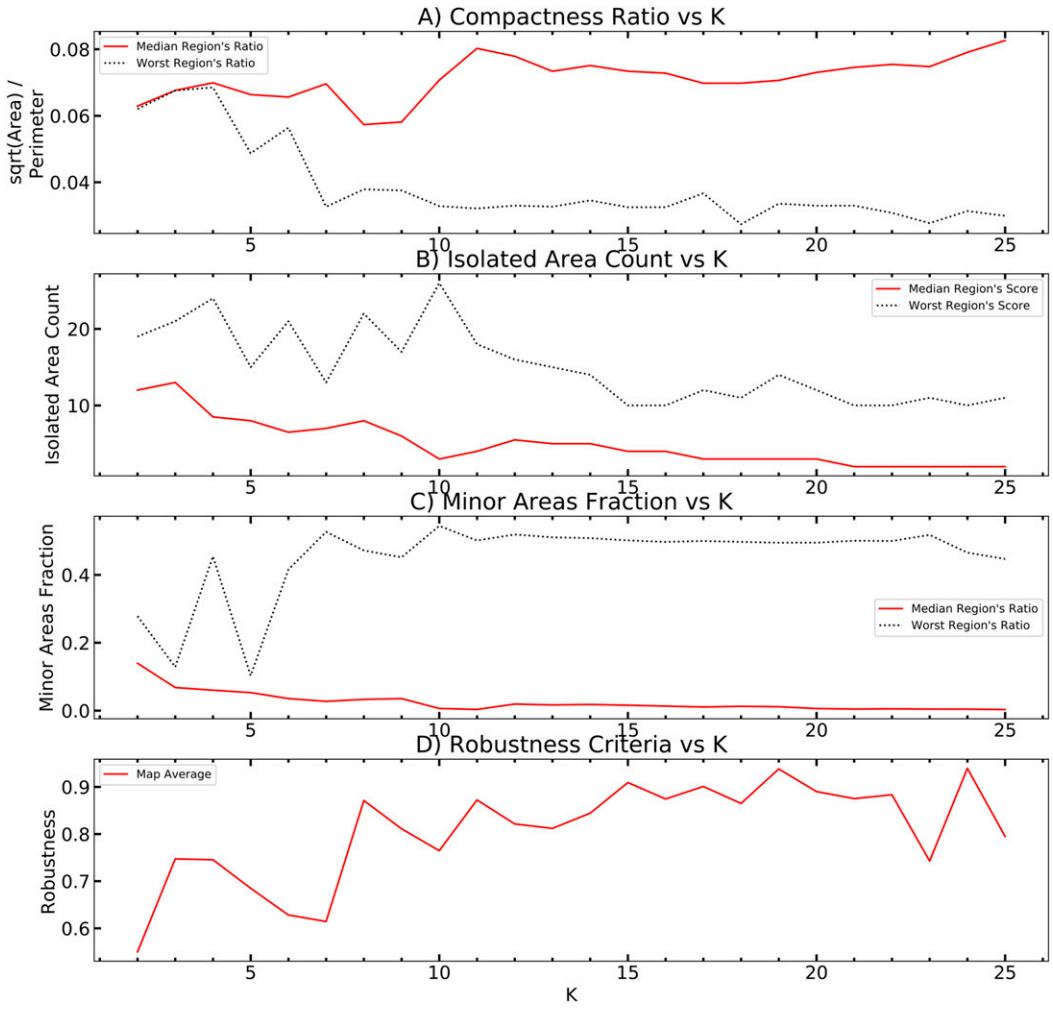

FIG. 6. As in Fig. 3, but for CPC-based SOM regions.

CPC- and CFSR-based maps (not shown) shows that the CPC-based maps have far fewer unique events. If the region with the lowest RER, for a given value of K, in the CPC-based maps is compared to the region with the lowest RER, for the same value of K, in the CFSR-based maps, they will have a very similar value of RER (comparing Figs. 5 and 2). So, even though the different regions have different numbers of unique events due to the temporal and spatial sampling and length of the data, the RER is comparable, as intended. The median compactness ratio of the CPC-based maps generally increases with K apart from a dip into lower values for $K = 8$ and 9 (Fig. 6a). Unlike CFSR-based maps, the worst region's compactness ratio generally decreases with increasing K for $K > 5$. As the number of regions increases both the median and worst region's IAC decreases steadily (Fig. 6b). The worst region in the CPC-based maps exhibits this trend much more strongly than does the worst region in the CFSR-based maps. The median region's MAF is very comparable to that of the CFSR-based maps for $K > 5$, both having small values that vary little. The worst region's MAF in the CPC-based maps is fairly constant between 0.4 and 0.5 except for $K = 2, 3$, and 5 where it falls below 0.3 (Fig. 6c). This is very different from the CFSR-based maps, which have low values (~0.1) from $K = 12$ through $K = 19$ and generally higher values (0.3–0.4) elsewhere (Fig. 3c). The relationship between robustness and K of the CPC-based maps is similar to the CFSR-based maps but with overall higher robustness and different local peaks (Fig. 6d).

To find the optimal value of K for the CPC-based maps we used the same 20% threshold of regional events divided by periods with nonzero precipitation. Applying this threshold requires $K < 17$, excluding $K = 14$. This threshold is similar to the CFSR value despite the large

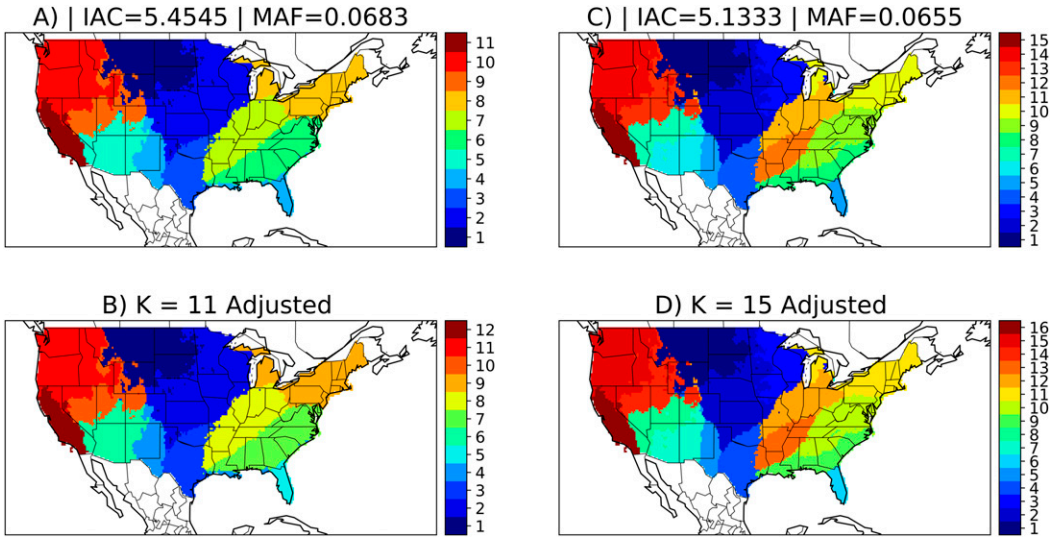

FIG. 7. SOM regions from the CPC data: (a) $K = 11$, (b) $K = 11$ again but with region 4 separated into two distinct regions to create 12 regions, (c) $K = 15$, and (d) $K = 15$ again but with region 5 separated into two distinct regions to create 16 regions. IAC is the mean value of the isolated area count, and MAF is the mean value of the minor areas fraction. These CPC data are created from the LTDM of the cube root of precipitation, adjusted to vary from 0 to 1, at each grid point, same as was done for CFSR data.

difference in the upper bounds provided by the false discovery rate for the CFSR-based maps compared to the CPC-based maps. Again, this is because RER is a far more limiting factor than FDR.

One does not expect the metrics presented here to find the same optimal value of K for both datasets, but one expects the values of K to be comparable. The median compactness is lowest for $K = 8$ and 9. The worst compactness ratio is highest for $K = 2$–4 and 6. The worst compactness ratios show a decreasing trend with increasing K, which is opposite from the CFSR data. Additionally, the compactness ratios found for CPC data are about half as large as those found for CFSR data. The median region's IAC is very noisy for $K = 2$–10 with $K = 10$ being the largest value. For $K > 10$ the median region's IAC is much smaller. The CPC IAC is generally larger than for the CFSR data. As with IAC, the median region's MAF is small for $K > 5$. The worst region's MAF is lowest (i.e., best) for $K = 3$ and 5 while the other values are higher, between 0.4 and 0.5. The CFSR MAF values are similar except for the worst regions from $K = 12$–19, where they are around half as much. The robustness of CPC-based maps is highest (i.e., best) for $K = 8$, 11, and 15. Compared to CFSR-based maps the CPC-based maps have smaller differences between the most and least robust maps. By these considerations CPC-based maps with $K = 15$ regions perform best.

In CPC data, $K = 15$ is not large enough to have Florida (FL) separate from the New Mexico–Texas (NM/TX) border region through the SOM analysis; instead, $K \geq 41$ is required for this to happen. However, maps with ≥41 regions have far fewer events per region than ideal. Again, the link between FL and the NM/TX border region is caused by those separate areas experiencing their respective wet seasons at nearly the same time of year. Because of the distance between FL and NM, the relatively similar sizes of these two subregions, the region they both belong to should be separated into two distinct regions for analysis and making of composites. This approach is recommended for all regions with a minor area that is greater than a quarter the size of the major area; for reference, the FL part of region 4 is 0.41 the size of the NM/TX part in Fig. 7a. The FL part of region 5 is 0.53 the size of the NM/TX part in Fig. 7c. In the analysis shown in this paper, the only regions that would meet this requirement encompass FL and NM/TX and the separation results in a map with 12 and 16 regions and is shown in Figs. 7b and 7d. One notes that this threshold does identify other regions for separation at high values of K (≥20). For some of these maps Michigan (MI) is its own region and the two parts of MI are disconnected from one another.

7. Comparison between SOM and other regionalizations

Karl and Knight (1998) used nine regions to analyze precipitation trends over the CONUS based on combining entire states. State boundaries do not necessarily correspond to meteorological "boundaries" or

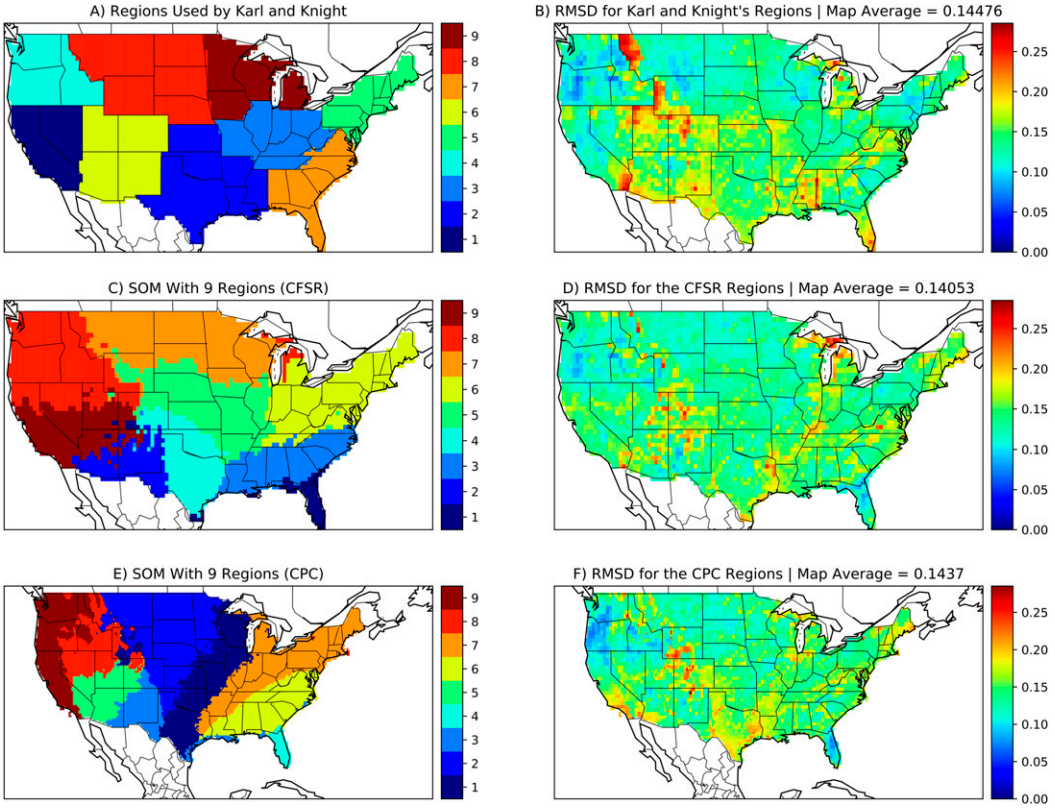

FIG. 8. (a) Our representation of the nine regions used by Karl and Knight (1998). (b) The root-mean-square difference (RMSD), described in section 5b, of the regions in (a). (c) SOM regions created from the LTDM of the cube root of precipitation (CFSR data), adjusted to vary from 0 to 1, $K = 9$ to match the number of regions in (a). (d) As in (b), but for the regions shown in (c). (e) As in (c), but for CPC data. (f) As in (d), but for the regions shown in (e). Note that regions in (e) are calculated from CPC data but the RMSD is calculated with CFSR data to be comparable to (b) and (d). For both (c) and (e) Florida was manually separated from the NM/TX border region to form the ninth region. This does not substantially affect the results shown in (d) and (f). Smaller RMSD is desirable.

climatological zones. Nonetheless, their choice of regions has been popular so it is important to see how their regions compare with our SOM-based regions. Their map (Fig. 8a) is constrained by state boundaries, and therefore scores very well in our connectedness metrics. The Karl and Knight map also does a fairly good job of grouping areas by their annual cycle of precipitation in many parts of the CONUS, although this was not their stated goal. To compare the Karl and Knight map quantitatively to the SOM method, we created a SOM with nine regions (see Fig. 8c). Both Figs. 8c and 8e were created with eight regions, and the ninth region comes from separating Florida from the New Mexico–Texas border region as described in the previous section. We want a quantitative measure of how much each grid point is like the rest of the grid points in its region in terms of its annual cycle. For each of Figs. 8a and 8c we calculate the root-mean-square difference (RMSD) between the LTDM-n at each grid point and the mean of each other grid point's LTDM-n within the region the original grid point belonged to. The RMSD is found by taking the squared difference in LTDM-n at each day between one grid point and the average value of the LTDM-n within the region on that day, then taking the average of all days. Small values of this RMSD indicate that a specific grid point is very much like the rest of its region, and large values indicate a grid point that is quite different. The result for each map is plotted in Figs. 8b and 8d. We do not believe $K = 9$ creates regions that are small enough to confidently aggregate events within. This is because the $K = 9$ map has a smaller compactness ratio compared to a map with $K = 15$ (Fig. 3a). Nonetheless, Karl and Knight use nine regions so we make our comparison using nine SOM-based regions. Even using a small ($K = 9$) number of regions, the SOM regions have grid points with a more consistent annual cycle than do the Karl and Knight regions. While we see a number of areas with particularly poor consistency

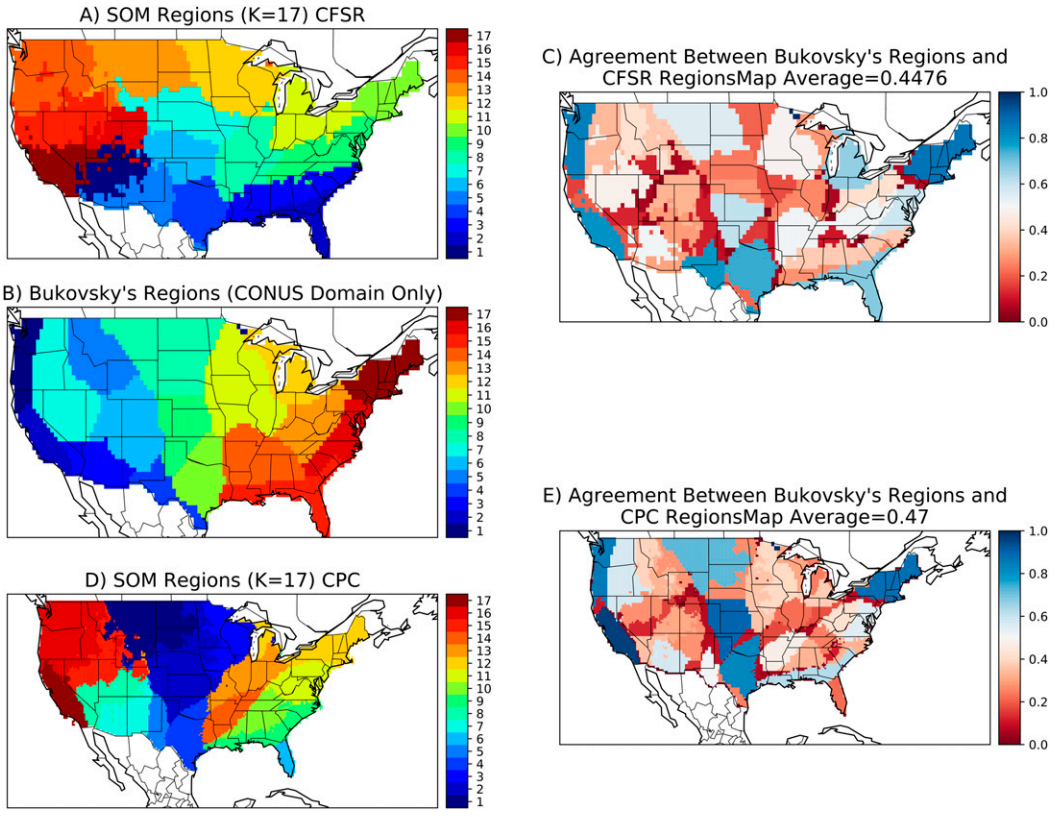

FIG. 9. (a) SOM regions created from the LTDM of the cube root of precipitation (CFSR data), adjusted to vary from 0 to 1, $K = 17$ to match the number of regions in (b). (b) Bukovsky regions plotted over the CONUS domain. (c) The agreement between the maps in (a) and (b) using the robustness method described in section 5b except both maps being compared have 17 regions. Hence, larger (darker blue) values mean greater agreement and smaller (darker red) values mean less agreement. (d) As in (a), but for CPC data. (e) As in (c), but comparing maps in (b) and (d).

in Karl and Knight's regions, namely Mississippi, western Montana, and the Four Corners states, we also see that their choice performed very well in the Pacific Northwest.

The North American Regional Climate Change Assessment Program (NARCCAP; Mearns et al. 2012) uses regions outlined in Bukovsky (2011) to capture North American regional climatology; we refer to these regions as the Bukovsky regions. For a more direct comparison to our regions, we have plotted only those 17 Bukovsky regions that exist over the CONUS (Fig. 9b). These Bukovsky regions closely follow those used by the National Ecological Observatory Network (NEON) put forward in Kampe et al. (2010). These regions are based on a statistical analysis of nine ecoclimate state variables, including temperature, precipitation, and solar insolation. One should not expect them to match exactly our regions since ours are based on normalized precipitation only. To compare our regions to the Bukovsky regions, we again match the number of regions, now $K = 17$ in Fig. 9. For each grid point we find the fraction of grid points that were in the same region as the target grid point under the Bukovsky regions compared to how many are still in the same region as the target grid point under the SOM regions, similar to our measure of robustness. This fraction varies between 0 and 1, with values of 1 indicating more agreement between the two sets of regions; the result of this calculation is shown in Fig. 9c. The map average result is less than 0.5, which means the average grid point is in a region with more than half of the grid points being different when we compare the Bukovsky regions to our SOM generated regions from CFSR data. While the maps, overall, are not very similar, one recognizes that there are areas of fairly good agreement. This is notable because of the differing ways the regions were produced. The areas of agreement are shown by bluer hues in Fig. 9c. The places where agreement is good are similar to those found in comparison with Karl and Knight, like the Pacific Northwest, New England, Southern

California, parts of the Great Plains, southern Michigan, and Florida. Overall, the method used by Bukovsky produces a map that is quite different to the one produced by the SOM method. That is expected because the two maps were designed with different purposes: Bukovsky's purpose being to create regions sensitive to changes in temperature and precipitation to aid in North American climate change assessment, while ours is to group areas purely by the annual cycle of precipitation. It is nevertheless encouraging to find some commonalities between these regionalizations. The same comparison is shown between the Bukovsky regions and SOM regions based on CPC data (Figs. 9d,e). Figure 9d was created from a SOM with 16 regions with the 17th region created by manually separating Florida from the NM/TX border region. The map average is again below 0.5 and is quite similar to Fig. 9c. The CPC and CFSR differences from the Bukovsky regions are generally similar; a couple of exceptions are for Florida and the very northernmost part of California.

8. Conclusions

This paper introduced a new and objective way to select contiguous geographic areas that experience precipitation at similar times of the year. Our purpose in devising this methodology is to identify regions for compositing information about extreme events. We have devised our method to focus on the timing of precipitation during the year because the timing is a key factor in determining which meteorological processes are primarily causing the precipitation. Our method creates ''coherent'' regions based only on the normalized annual cycle of precipitation, or LTDM-n. These regions are compact enough to aggregate extreme precipitation events within each region with confidence that the events are similar. The directness of interpretation of the LTDM-ns that represent each region makes comparisons to other regionalizations or datasets straightforward and easily quantifiable. Comparisons to other regionalizations are also aided by flexibility in the number of regions.

In an effort to optimize the choice for K (the number of SOM regions), six criteria were developed to measure four aspects of the precipitation regions. We first tested the statistical distinguishability, at the 5% level, of each pair of regions using the false discovery rate to reveal the upper bound of the number of regions that could be created. This produced upper limits of 32 regions for CFSR data and 63 regions for CPC data. We next applied a criterion called the regional extremes ratio, whose 20% threshold means all the SOM regions must have at least 4 times as many time periods with at least one grid point reporting precipitation exceeding the 95th percentile than would occur at a single grid point. In short, this RER criterion requires every region to provide notably more events for aggregation than if a single grid point was used. Also, RER facilitates comparison across datasets with different resolution and length. We created two criteria to measure connectedness because we want to track two distinct aspects of connectedness. The first aspect, measured by IAC, is how many groups of grid points are disconnected from the largest group in a region. The second aspect, measured by MAF, is how large are the disconnected groups relative to each other and to the largest group. Both criteria's map averages (i.e., the average over all K regions in a particular map) have a strong inverse trend with K. As more regions are added, the average region becomes more connected. The trend in the least connected region as K increases is much flatter. This indicates that there are only a few regions per map that have substantial values of IAC and MAF. These regions usually occur in mountainous or desert regions of the Southwest. Robustness refers to how much a map is altered by changing the value of K. The relationship between robustness and K does not exhibit a clear trend and is better described as a sawtooth function. For one to several consecutive values of K, robustness is relatively low and then for one to several K values it is relatively high. Compactness is measured by the ratio of the square root of the area of a region divided by the perimeter of that region. The map average of compactness tends to increase (improve) as the value of K increases.

Our methodology was applied to CFSR and CPC precipitation data. The criteria provide a guide for choosing an optimal value for K, but in practice there were additional problems. A persistent problem is that fairly high values of K are needed in our tested data to decouple an area including the Florida peninsula from an area near the NM/TX border that are placed in the same region by the SOM algorithm. In practice, one might either opt for a larger K or if that is not feasible (i.e., if the number of events within the regions becomes too small) then manually intervene and treat the two areas as separate regions in later applications, as shown in Fig. 7.

For these CFSR data, $K = 15$ is optimal and the map has several notable features and quite compact regions (Fig. 1f). Both the East and Gulf Coasts are broken into three different regions (although they do share region 2) while the West Coast is divided into only two regions. For these CPC data, $K = 15$ is optimal and after separating the Florida and NM/TX border areas yields a map with 16 regions (Fig. 7d). In comparing the CPC map with 16 regions to the CFSR map with 15 regions the

most obvious difference is in the Pacific Northwest. Regions 12 and 13 in the CFSR are largely combined into region 15 in the CPC data. Florida and the Gulf Coast are in separate regions for the CPC but the same region for the CFSR. Another notable difference is that throughout the Great Plains the lines between regions tend to run from north to south or from east to west in the CFSR map while in the CPC map they are diagonally oriented from southwest to northeast.

Maps created from CFSR and from CPC data were compared to regions described in Karl and Knight (Fig. 8). All three maps with nine regions performed least well in our RMSD test in similar areas of the country. Substantial differences between the local annual cycle and the regional annual cycle were found along the Rocky Mountains, in the Southwest deserts and along a line drawn from east Texas to Michigan. These differences have lower magnitudes in the two SOMs than in the Karl and Knight regions. All three maps did relatively well in the Pacific Northwest and parts of the High Plains. The two SOMs did much better in Florida, while the Karl and Knight regions did better in New York State.

Maps created from CFSR and CPC data were compared to the Bukovsky regions (Fig. 9). Both SOMs agree with the Bukovsky regions in the north and south parts of the West Coast as well as east Texas and New England. The SOMs and the Bukovsky regions disagree more strongly over much of the Great Basin, western mountains, and central plains. The CFSR map agrees with Bukovsky in Florida but not in other parts of the Southeast; this relationship is reversed between the CPC and Bukovsky regions. Overall, the agreements and disagreements between the CPC and Bukovsky regions are both stronger in magnitude than those between the CFSR regions and the Bukovsky regions.

This method can be applied elsewhere in the world and the size of the regions is tunable by changing the number of regions (K) specified in the SOM. This flexibility allows the technique be used to investigate problems over different spatial scales. Here our interest is in large-scale meteorology but in principle one could find a dense observational network and examine how precipitation varies within an individual watershed.

These regions are the first step in a process to automate identification of the primary meteorological mechanism or mechanisms creating an extreme precipitation event. This SOM-based approach can be used to assess how well climate models capture the spatial changes in the annual cycle of precipitation.

Acknowledgments. U.S. Unified Precipitation data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/. CFSR data are provided by the DOC/NOAA/NWS/NCEP/EMC on their Web site at https://rda.ucar.edu/datasets/ds093.0/. We also acknowledge funding from U.S. DOE Office of Science Award DE-SC0016605, and USDA NIFA Hatch project Accession 1010971.

REFERENCES

Adams, D. K., and A. C. Comrie, 1997: The North American monsoon. *Bull. Amer. Meteor. Soc.*, **78**, 2197–2214, https://doi.org/10.1175/1520-0477(1997)078<2197:TNAM>2.0.CO;2.

Bukovsky, M. S., 2011: Masks for the Bukovsky regionalization of North America. Regional Integrated Sciences Collective, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, accessed 24 December 2018, http://www.narccap.ucar.edu/contrib/bukovsky/.

Chen, M., and Coauthors, 2008a: CPC Unified Gauge-Based Analysis of Daily Precipitation over CONUS, V1.0. Earth Science Research Lab Physical Sciences Division, accessed 1 October 2016, https://www.esrl.noaa.gov/psd/cgi-bin/db_search/DBListFiles.pl?did=125&tid=54729&vid=2415.

——, W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. W. Higgins, and J. E. Janowiak, 2008b: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.*, **113**, D04110, https://doi.org/10.1029/2007JD009132.

Grotjahn, R., 2011: Identifying extreme hottest days from large scale upper air data: A pilot scheme to find California Central Valley summertime maximum surface temperatures. *Climate Dyn.*, **37**, 587–604, https://doi.org/10.1007/s00382-011-0999-z.

——, and G. Faure, 2008: Composite predictor maps of extraordinary weather events in the Sacramento, California, region. *Wea. Forecasting*, **23**, 313–335, https://doi.org/10.1175/2007WAF2006055.1.

——, and R. Zhang, 2017: Synoptic analysis of cold air outbreaks over the California Central Valley. *J. Climate*, **30**, 9417–9433, https://doi.org/10.1175/JCLI-D-17-0167.1.

Hewitson, B. C., and R. G. Crane, 2005: Gridded area-averaged daily precipitation via conditional interpolation. *J. Climate*, **18**, 41–57, https://doi.org/10.1175/JCLI3246.1.

Ison, N. T., A. M. Feyerherm, and L. D. Bark, 1971: Wet period precipitation and the gamma distribution. *J. Appl. Meteor.*, **10**, 658–665, https://doi.org/10.1175/1520-0450(1971)010<0658:WPPATG>2.0.CO;2.

Johnson, N. C., 2013: How many ENSO flavors can we distinguish? *J. Climate*, **26**, 4816–4827, https://doi.org/10.1175/JCLI-D-12-00649.1.

Kampe, T. U., B. R. Johnson, M. Kuester, and M. Keller, 2010: NEON: The first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure. *J. Appl. Remote Sens.*, **4**, 043510, https://doi.org/10.1117/1.3361375.

Karl, T. R., and R. W. Knight, 1998: Secular trends of precipitation amount, frequency, and intensity in the United States. *Bull. Amer. Meteor. Soc.*, **79**, 231–242, https://doi.org/10.1175/1520-0477(1998)079<0231:STOPAF>2.0.CO;2.

Kohonen, T., 1982: Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, **43**, 59–69, https://doi.org/10.1007/BF00337288.

Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel, 2006: World map of the Koppen–Geiger climate classification updated. *Meteor. Z.*, **15**, 259–263, https://doi.org/10.1127/0941-2948/2006/0130.

Kunkel, K. E., S. A. Changnon, and R. T. Shealy, 1993: Temporal and spatial characteristics of heavy-precipitation events in the Midwest. *Mon. Wea. Rev.*, **121**, 858–866, https://doi.org/10.1175/1520-0493(1993)121<0858:TASCOH>2.0.CO;2.

——, D. R. Easterling, D. A. R. Kristovich, B. Gleason, L. Stoecker, and R. Smith, 2012: Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous United States. *J. Hydrometeor.*, **13**, 1131–1141, https://doi.org/10.1175/JHM-D-11-0108.1.

Mearns, L. O., and Coauthors, 2012: The North American Regional Climate Change Assessment Program: Overview of phase I results. *Bull. Amer. Meteor. Soc.*, **93**, 1337–1362, https://doi.org/10.1175/BAMS-D-11-00223.1.

Saha, S., and Coauthors, 2010a: NCEP Climate Forecast System Reanalysis (CFSR) Selected Hourly Time-Series Products, January 1979 to December 2010. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 13 November 2016, https://doi.org/10.5065/D6513W89.

——, and Coauthors, 2010b: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1058, https://doi.org/10.1175/2010BAMS3001.1.

Stidd, C. K., 1953: Cube-root-normal precipitation distributions. *Eos, Trans. Amer. Geophys. Union*, **34**, 31–35, https://doi.org/10.1029/TR034i001p00031.

Wang, B., 1994: Climatic regimes of tropical convection and rainfall. *J. Climate*, **7**, 1109–1118, https://doi.org/10.1175/1520-0442(1994)007<1109:CROTCA>2.0.CO;2.

—— and LinHo, 2002: Rainy season of the Asian–Pacific summer monsoon. *J. Climate*, **15**, 386–398, https://doi.org/10.1175/1520-0442(2002)015<0386:RSOTAP>2.0.CO;2.

Xie, P., A. Yatagai, M. Chen, T. Hayasaka, Y. Fukushima, C. Liu, and S. Yang, 2007: A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeor.*, **8**, 607–626, https://doi.org/10.1175/JHM583.1.