Vol. 25: 95–107, 2003 | CLIMATE RESEARCH Clim Res | Published December 5

Clustering and upscaling of station precipitation records to regional patterns using self-organizing maps (SOMs)

Robert G. Crane[1,*], Bruce C. Hewitson[2]

[1]Department of Geography, The Pennsylvania State University, 103 Deike Building, University Park, Pennsylvania 16802, USA
[2]Department of Environmental and Geographical Sciences, University of Cape Town, Private Bag, Rondebosch 7701, South Africa

ABSTRACT: Self-organizing maps (SOMs), a particular application of artificial neural networks, are used to proportionately combine precipitation records of individual stations into a regional data set by extracting the common regional variability from the locally forced variability at each station. The methodology is applied to a 100 yr record of precipitation data for 104 stations in the Mid-Atlantic/Northeast United States region. The SOM combines stations with common precipitation characteristics and identifies precipitation regions that are consistent across a range of spatial scales. A variation of the SOM application identifies the temporal modes of the regional precipitation record and uses them to fill missing data in the station observations to produce a regional precipitation record. A test of the methodology with a complete data set shows that the 'missing data' routine improves the regional signal when up to 80% of the data are missing from 80% of the stations. The improvement is almost as pronounced when there is a bias in the missing data for both high-precipitation and low-precipitation events.

KEY WORDS: Upscaling · Regional precipitation · Regionalization

1. INTRODUCTION

The need to present regional expressions of larger scale change or variability in global climate data sets has led to a recent focus on downscaling methodologies using empirical or numerical techniques. These downscaling methodologies have been used to derive the station-level response to larger scale forcing, but more frequently the need is for extracting a regional response, where the region is defined geographically or simply as a finer resolution grid-scale. In either case, when it comes to deriving training data sets for empirical analyses or validation data for numerical models, we are faced with the opposite problem from downscaling—the need to upscale from individual station data to regional patterns.

Upscaling from station to regional data presents 2 distinct problems: the first is the need to identify homogeneous climate regions, and the second is to account for differing lengths and gaps in individual station records. In upscaling procedures, regions are often defined arbitrarily or according to external factors (such as political or other environmental boundaries), and any available station records within the region are often averaged for a given time interval. For some applications this may not be an issue, but there are many applications for which such an approach would be inappropriate. In a climate diagnostic mode, for example, when examining the regional response to a supposed larger scale forcing, it makes sense to define the regions in terms of stations that have similar climate characteristics; for example, Knapp & Yin (1996) examined the relationships between geopotential heights and temperature distributions in the Southeast United States by first defining climatic regions using a factor analysis of mean annual temperature data. Comrie & Glen (1999) used a principal components-based regionalization of

*Email: crane@essc.psu.edu

precipitation in the SW USA / northern Mexico to examine variability in monsoon rainfall. Similarly, Waylen et al. (1996) examined the spatial and temporal response of annual precipitation patterns in Costa Rica to the Southern Oscillation. They first used cluster analysis to group stations with similar precipitation distributions into 4 major regions, and then they derived individually the response of each region's mean annual precipitation to the Southern Oscillation Index.

In particular, station precipitation records tend to exhibit a large degree of spatial inhomogeneity due to small-scale events and local forcing, hence the need to identify homogeneous regions and to derive a regional precipitation signal. If the temporal characteristics of the record are also of interest, e.g. when identifying temporal patterns or trends in the data, it is also important that stations do not have a disproportionate influence on the temporal pattern either because of anomalous conditions at individual stations or because of differences in their length of record. The challenge, therefore, is to proportionately combine individual station records into a regional data set, with a methodology that removes the effects of missing data and produces a measure of the underlying regional precipitation, i.e. one that extracts the common regional variability from the locally forced variability at each station. The present paper develops such a methodology based on a particular form of artificial neural network known as the Kohonen self-organizing map (SOM), using station precipitation data for the northeast United States.

2. SELF-ORGANIZING MAPS

Self-organizing maps, or SOMs (Kohonen 1989, 1990, 1991, 1995), provide a mechanism for visualizing complex relationships in multi-dimensional data sets. SOMs are developed through an iterative training procedure in which elements of the SOM are mapped to representative regions of the input data space. The process is outlined in Fig. 1 for an input data matrix of n variables and m observations. The SOM is an arbitrary XY array, where each cell or node in the array is described by a $(1,n)$ reference vector. In other words, the length of the reference vector matches the number of variables in the input data set. Several alternatives exist for assigning the initial values of the reference vectors, the most common method being to assign random values. Another common approach is to base them on the loadings of the first 2 eigenvalues of the input data matrix, but in either case the technique proceeds by competitively matching each input data record with a node in the SOM, comparing the data record with each of the SOM node reference vectors. The 'winning' node is the one with the closest match to the data vector, usually calculated as a measure of Euclidean distance.

The reference vector of the winning node is then updated by adjusting it slightly to reduce the difference between the node reference vector and the input data vector. The amount of adjustment is set by a user-defined factor referred to as the *learning rate*. An important element of the SOM learning process, however, is not just that the winning node is adjusted — but

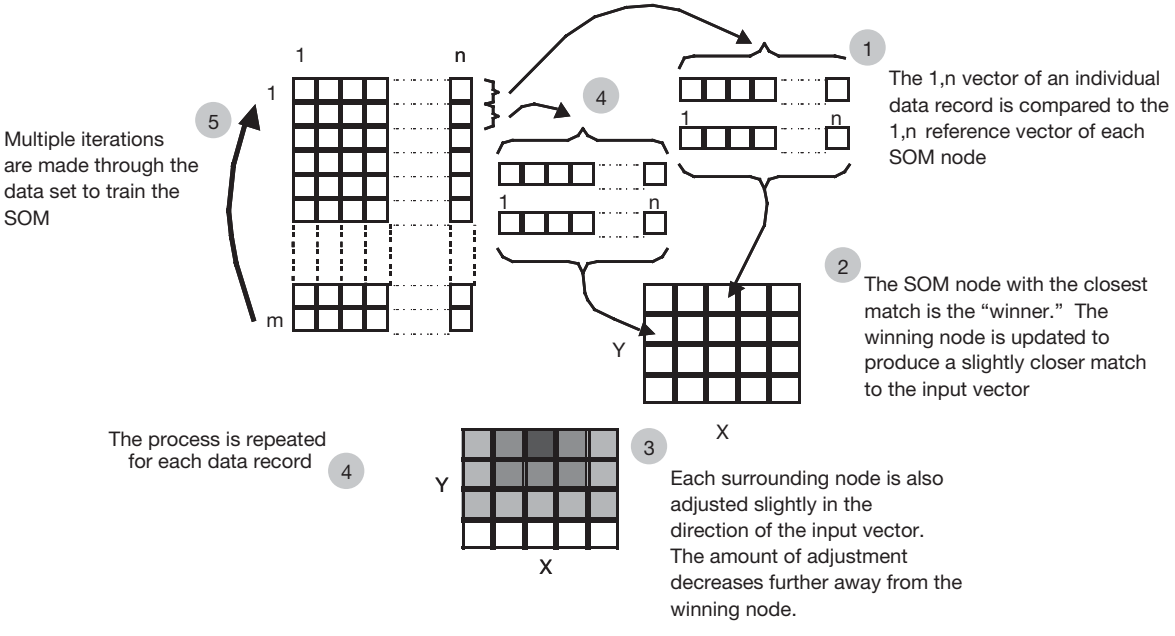

Fig. 1. The iterative self-organizing map (SOM) training procedure

that surrounding nodes are also adjusted. There are various methods for doing this, but one common approach is to adjust the surrounding nodes in proportion to their distance away from the winning node (closer SOM nodes have proportionately larger changes). The size of the update kernel is also set by the user and usually decreases with successive passes through the data set. The procedure continues with each successive data record, and it is then repeated for a large number of iterations through the data set until there is no further change in the SOM reference vectors. In the present case, we use 50 000 iterations and repeat the training a second time. In the first training phase we set a broad update kernel to produce an initial separation of SOM nodes. We then repeat the training, using the final reference vectors from the first training phase as the new initial SOM vectors, but using a smaller update kernel to refine the distribution in regions of high data densities. Once the training is complete, each of the SOM nodes has a reference vector that describes a location in the original *N*-dimensional data space. The iterative training procedure locates nodes as a combined function of the density of data points and the difference between the clusters of points. Together with the use of the update kernel to adjust nearby nodes during each training step, this ensures that observations that are very similar to each other define nodes that are close together on the SOM, while observations that are very different map to nodes that are located further apart in the SOM space.

Each node is a vector that describes the mean value of the nearby observations in the multi-dimensional data space. As the number of nodes increases (i.e. with larger SOMs), they are located in smaller and closer clusters; hence, more detail is available in the parts of the data space with the highest number of data points. If sufficient nodes are available, the SOM will also locate nodes in data-sparse regions or between other nodes where there are no existing data points. Thus the locations of the nodes span the data space, but are concentrated in regions of higher data densities. As such, the distribution of nodes reflects the multi-dimensional data density distribution function of the original data set. The number of nodes available to describe the distribution function depends on the size of the SOM array. More detail is revealed as the SOM size increases, but at any particular size the SOM nodes provide the best description of the distribution function at that level of generalization. SOMs thus provide a non-linear (and non-orthogonal) alternative to using principal components analysis (PCA) or eigenvector analysis to describe data structure. At the same time, because the nodes represent the nearby data points, SOMs can also be regarded as a form of cluster analysis, where the groups are defined by individual nodes or by several adjacent (and related) nodes. The use of SOMs to group days with similar sea-level pressure (SLP) patterns is given in Hewitson & Crane (2002), together with a more detailed description of the SOM procedure and the differences between SOMs and more traditional forms of cluster analysis.

A useful feature of the SOM is that the SOM array provides a graphic visualization of the distribution function. Taking SLP as an example, if the input data set is a time series of gridded SLP fields over a region, the reference vectors on the SOM nodes approximate the mean SLP fields for the surrounding cloud of data points. Each of these node vectors can thus be displayed as an SLP map (Fig. 2). The SOM array, in effect, becomes a projection of the multi-dimensional data set onto a 2-dimensional plane. Note in Fig. 2 how pressure fields that are very different are located far apart in the SOM space, while those that are similar to each other are close together. Once the SOM has been trained, the data can be presented to the SOM again, and we can note to which node each observation is mapped. We can then look at the frequency of occurrence on each node and conduct a range of analyses, such as computing the variance of each node, transitions between nodes, changes in frequency through time etc., as in Hewitson & Crane (2002). In the present paper, we use the same procedure to group stations with similar precipitation climatologies and, in a second step, use the same techniques to fill in missing data records and to extract a regional precipitation signal. While the processing time obviously varies with the size of the data set and the number of iterations used, the types of SOM analysis described here take only a few minutes on a high-end PC or workstation.

3. REGIONALIZATION

Precipitation data for 1900 to 1999 were extracted from the Daily Historical Climatology Network, National Climatic Data Center, Asheville, NC (http://lwf.ncdc.noaa.gov/oa/climate/research/ushcn/daily.html) for the NE USA from 70–80° W and 35–43° N. The data set for this region contains 104 stations with varying lengths of record. In this application the SOM is used in an analogous fashion to traditional forms of cluster analysis. We begin by deriving a similarity matrix of the precipitation data. Any measure of similarity could be chosen, but in this case we use the variance-covariance matrix. This allows the magnitude of the common variance between stations to also be a factor in determining similarity, and not just the temporal phase matching that would result from using the correlation matrix. The input data for the SOM is thus the matrix of ($n - 1 \times n$) covariances, where n is the

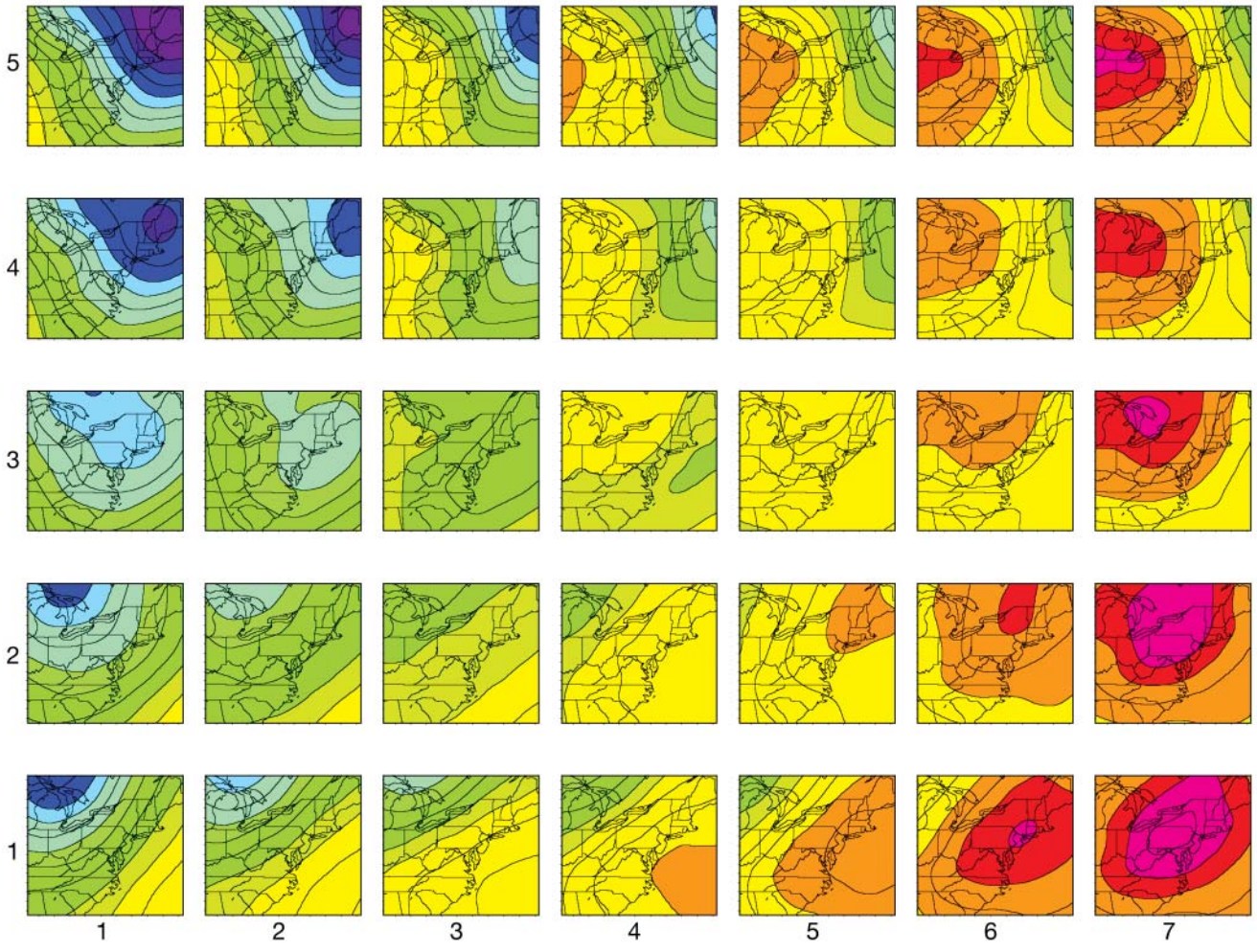

Fig. 2. A 5 × 7 SOM of January sea-level pressure (SLP) for the NE USA. Blues represent relatively low pressure, while reds indicate high pressure. (After Hewitson & Crane 2002)

number of stations and each station in the data set is represented by a row of covariances from the similarity matrix.

The longitude and latitude of the station are included as additional variables in the input data. In this case, the longitude and latitude are scaled to match the variances, and they are simply 2 out of the 105 variables for each station. It would be possible to increase the weighting on the location information if it were important to ensure that distant locations with similar rainfall characteristics did not map to the same SOM node. Conversely, in areas of complex terrain, the location information could be given a zero weighting to avoid nearby stations at very different elevations being forced onto the same node. Looking at larger scale patterns, this weighting function could also be adjusted to explicitly seek regions that are not contiguous (e.g. when looking for teleconnection patterns). Although we only use the longitude and latitude of the stations here, other parameters such as elevation and aspect could be included in the training data set. In this particular application all covariances are used for each station. As the SOM does not require a complete set of data for each record, comparisons with the SOM reference vectors are based on whatever variables are available for any data record. Consequently, it is also possible to include a correlation or covariance threshold value and base the SOM training process only on data that exceed the threshold.

For this application, the training is carried out using 4 different-sized SOM arrays: 2×3, 3×4, 4×6, and 6×8. The training data set is the same for all SOM dimensions (i.e. the covariance of the station data using all available data for the 104 stations from 1900 to 1999). The results are shown in Figs. 3 to 5. Six groups, A–F, are identified that correspond to the 6 nodes of the 2×3 SOM. The nodes are numbered as in Fig. 4, and the numbers in Fig. 3 match the node numbers for the

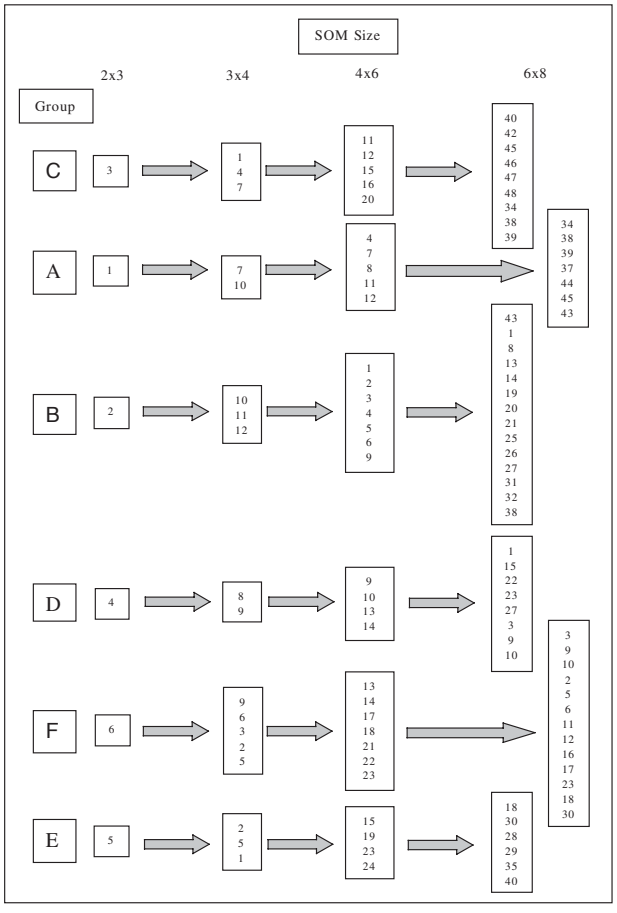

Fig. 3. SOM groupings for each level of SOM generalization

4 SOMs. Fig. 3, therefore shows that the stations that fell on node 3 of the 2×3 SOM all mapped to nodes 1, 4 and 7 in the 3×4 SOM; nodes 11, 12, 15, 16 and 20 of the 4×6 SOM; and nodes 34, 38, 39, 40, 42, 45, 46, 47 and 48 of the 6×8 SOM. The groupings of the nodes in SOM space are shown in Fig. 4. For example, Group F consists of stations that are mapped to Node 6 of the 2×3 SOM (Fig. 3; Fig. 4a). These same stations map to nodes 2, 3, 5, 6 and 9 in the 3×4 SOM (Figs. 3 & 4b), and to nodes 13, 14, 17, 18, 21, 22 and 23 of the 4×6 SOM (Figs. 3 & 4c). The locations of the stations are shown in Fig. 5, identified by the SOM node to which they map.

Comparing Figs. 3–5 we see that the station groupings are consistent at all levels of generalization. As each SOM starts from a different random initialization of the node vectors, Group F maps to the top-right corner of the 2×3 SOM, the top-left corner of the 3×4 and 6×8 SOMs, and the bottom-right corner of the 4×6 SOM. However, the same relative distribution of groups is always present. The stations that make up Groups A and F consistently map to opposite corners of the SOM space (Fig. 4), indicating that the greatest differences exist between these 2 groups. Groups B and C are different; there is no overlap between the groups (Figs. 3 & 4), but both have some overlap with Group A. Similarly, Groups D and E are different, but both overlap with Group F. There is virtually no overlap at all between Groups A, B and C and Groups D, E and F. At the greatest level of generalization it would appear that there are 2 broad regions divided by the solid boundary in Fig. 5.

Looking at one area in greater detail (Region F), there are 29 stations mapping to Node 6 in the 2×3 SOM (Fig. 6a). This group is located in the Piedmont and Coastal Plain region of North Carolina and Virginia, and includes both shores of the Chesapeake Bay. Most of these stations are divided into 2 main subgroups in the 3×4 SOM (Groups 3 and 6; Fig. 6b). As before, Group 3 extends from the coast to the mountains in North Carolina and the southern and western parts of Virginia. The coastal and central parts of Virginia make up Group 6. Group 6 is also found on both shores of the Chesapeake Bay. However, there is no clear geographic distinction between groups at the head of the bay, where there is a mix of stations in Groups 5, 6 and 9. Note also that some of stations along the SE edge of the mountains that were included in Group 6 in the 2×3 SOM become part of the mountain group, Group 2, in the 3×4 SOM.

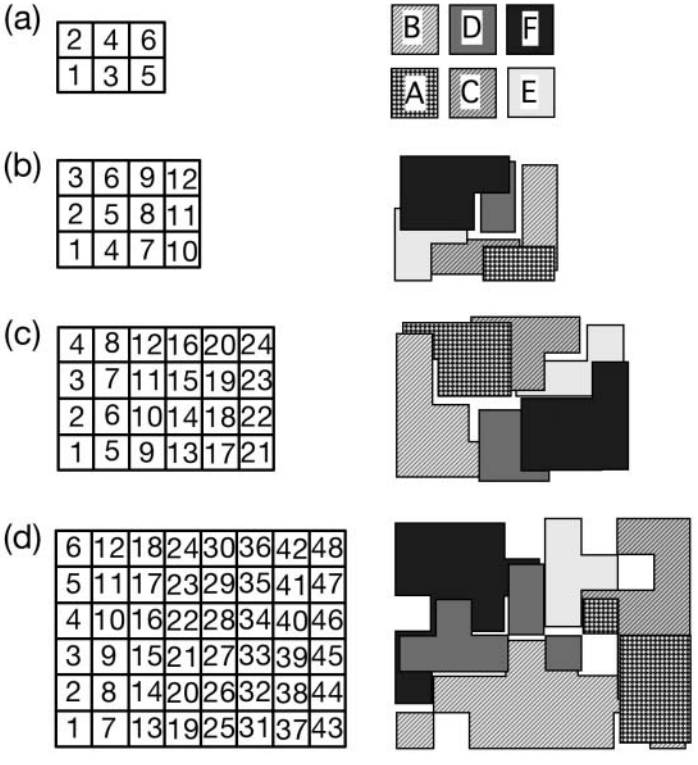

Fig. 4. SOM groupings, showing how the nodes listed in Fig. 3 are arranged in SOM space. Refer to Fig. 5 to see the stations mapped to each node

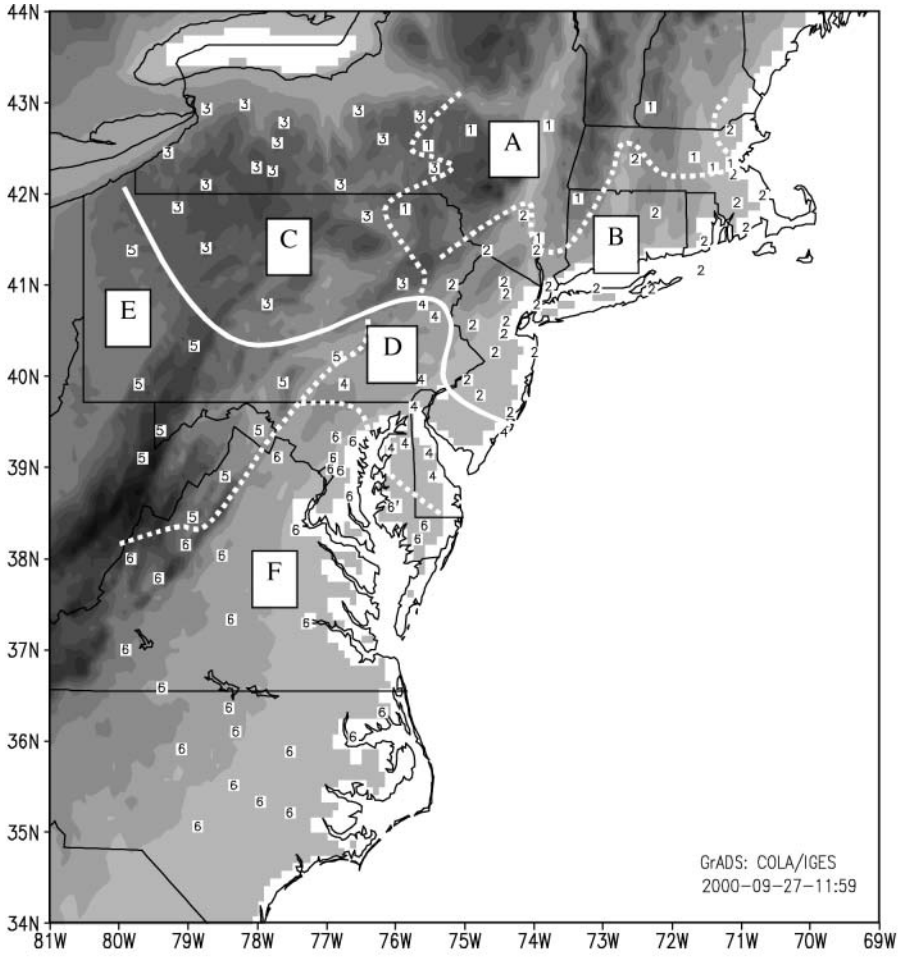

Fig. 5. Results of the 2×3 SOM regionalization. The letters are the 6 regions that correspond to nodes 1–6 of the 2×3 SOM. The numbers show the station locations and the SOM node to which each location maps. The white lines show the approximate boundaries between regions. Note, however, that in some areas the station density is not sufficient to draw these boundaries with any degree of accuracy

Further differentiation occurs in the 4×6 SOM (Fig. 6c). Group 3 in the 3×4 SOM is divided into 2 groups—a Coastal Plain group and a Piedmont group (21 and 22). The mountains stay as a distinct group (23), but there is further differentiation around the Virginia coast and the Chesapeake Bay. Fig. 6 also shows where the nodes (groups) map in the SOM space. Remember that the nodes with the greatest differences map to opposite corners of the SOM (Figs. 2 & 4). Consequently, we see in the 2×3 SOM that Node 6 maps to one corner. In the 3×4 SOM, Node 3 becomes the core of the region, with Groups 2 and 6 being most similar. Groups 5 and 9 are at the edge of the region both geographically and in SOM space. Similarly, with the 4×6 SOM, Group 21 lies at the core of the region adjacent to Groups 17 and 22, which are also next to 21 in SOM space. Groups 13, 14, 18 and 23 all border the region—again geographically and in SOM space. This example shows quite clearly that the regionalization is consistent in all of the SOMs, with the level of generalization increasing as the SOM size is reduced.

At the broadest level and for the region as a whole there appear to be 2 precipitation regimes—a northern and a southern regime, with the border running through Pennsylvania. Each of these regimes can be further divided into 3 distinct geographic subregions, and further subdivisions can be made as the SOM size increases. However, there is obviously no single correct grouping. As SOM size increases, the SOM nodes provide an increasingly closer fit to the original data distribution function. As would be expected, stations that map to nodes in the core of the region consistently map to the same or neighboring nodes as the SOM size increases. This feature of the SOM classification can be a way of identifying 'robust' core climate regions that could be used for a more rigorous (less noisy) assessment of climate variation or change. Stations on the periphery may change groupings as the larger SOMs provide greater discrimination between stations with smaller differences. It should also be remembered that, as with any type of climate regionalization, the boundaries between regions are not discrete. While the stations mapping to Node 22 (Fig. 6c), for example, have similar precipitation characteristics, as a group they will also show some similarity with the precipitation characteristics of the surrounding stations. The choice of the number of groups (and SOM size) depends on the level of detail or generalization required for the particular application. We use the 2×3 SOM groups in the analysis that follows, as this level of generalization provides distinct geographic regions, with each region having a sufficient number of stations to determine a regional precipitation signal.

Table 1 shows the correlations between the 6 groups. The correlations are based on the length of the record available for each pair of stations. Data are available from 1900 for 4 of the groups. Group 3 has data from 1916 and Group 5 from 1926. There are reasonable correlations between the groups, although there are some low correlations between some pairs (notably Groups 3 and 6, and Groups 3 and 4). Fig. 7 shows the annual means for all 6 groups. All groups show the

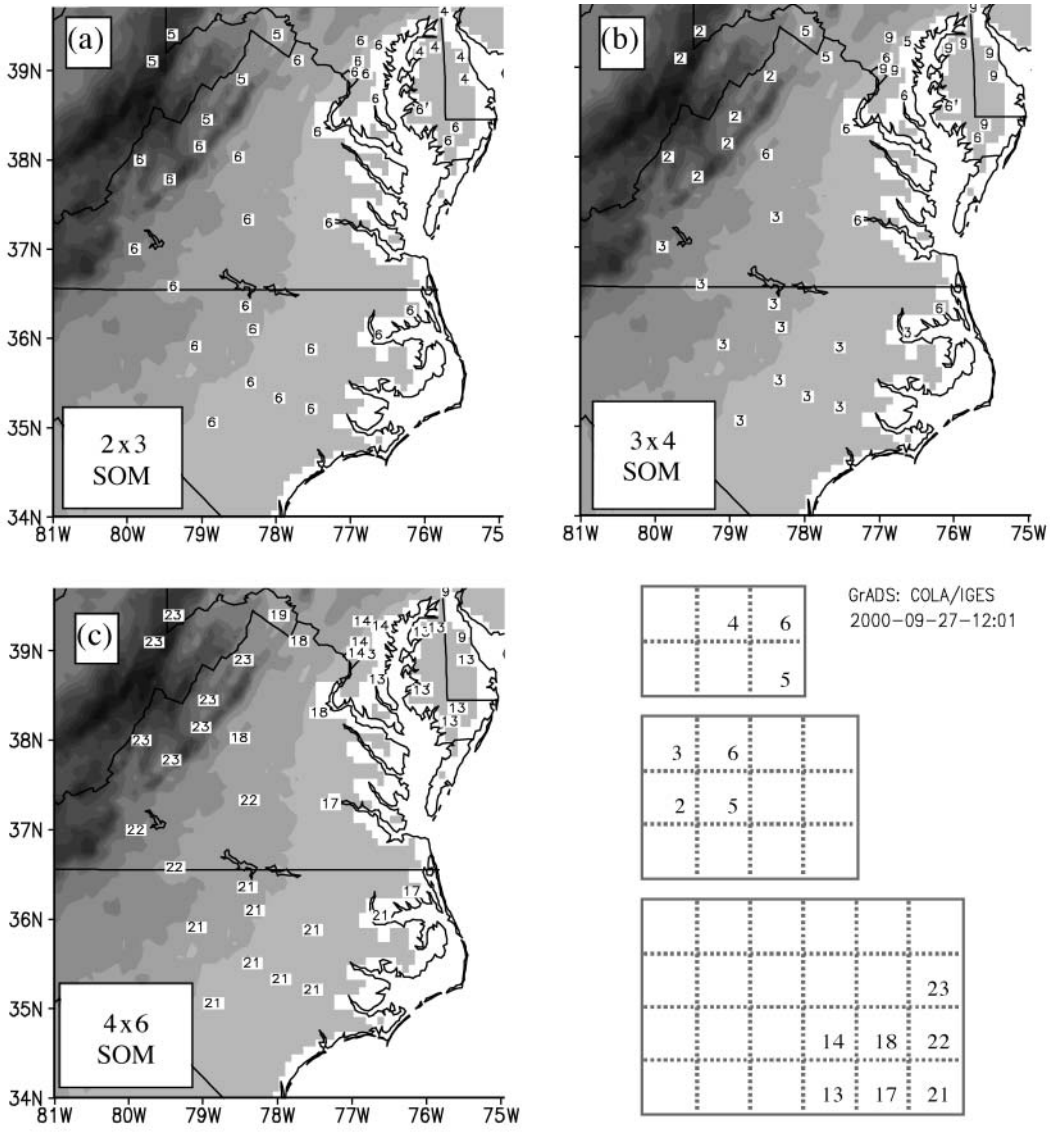

Fig. 6. Consistency in regionalization (grouping) as SOM size changes. Note that the level of generalization increases as the SOM size decreases

same general pattern, with a tendency toward higher values and greater variability in the 1970s through to the mid-1980s, compared to the 1960s and the more recent record. Comparing records for earlier periods may not be valid because of the small number of stations reporting in each group (this is discussed further below). For the second half of the record, the 2 coastal groups (2 and 6) tend to show higher precipitation amounts and greater variability than the inland groups (3 and 5). There are, however, some differences between the groups. In the 1990s, for example, Groups 1 and 2 have similar patterns, dropping from a peak in 1990–91 and with little change in 1992, 1993 and 1994. Group 3, on the other hand, has a sharp peak in 1993. Other differences in detail are also apparent earlier in the record, where precipitation in Group 3 increases steadily from 1974 to 1978, while Groups 1 and 2 show much more variability. Similar differences can also be seen between Groups 4, 5 and 6. Note that there is considerable difference between Groups 2, 3 and 4 in

Table 1. Correlation coefficients for the 6 groups identified by the 2×3 SOM. The correlations are based on annual means for the available length of record

Group	Group				
	2	3	4	5	6
1	0.66	0.39	0.70	0.60	0.48
2		0.44	0.81	0.71	0.64
3			0.32	0.47	0.09
4				0.76	0.75
5					0.65

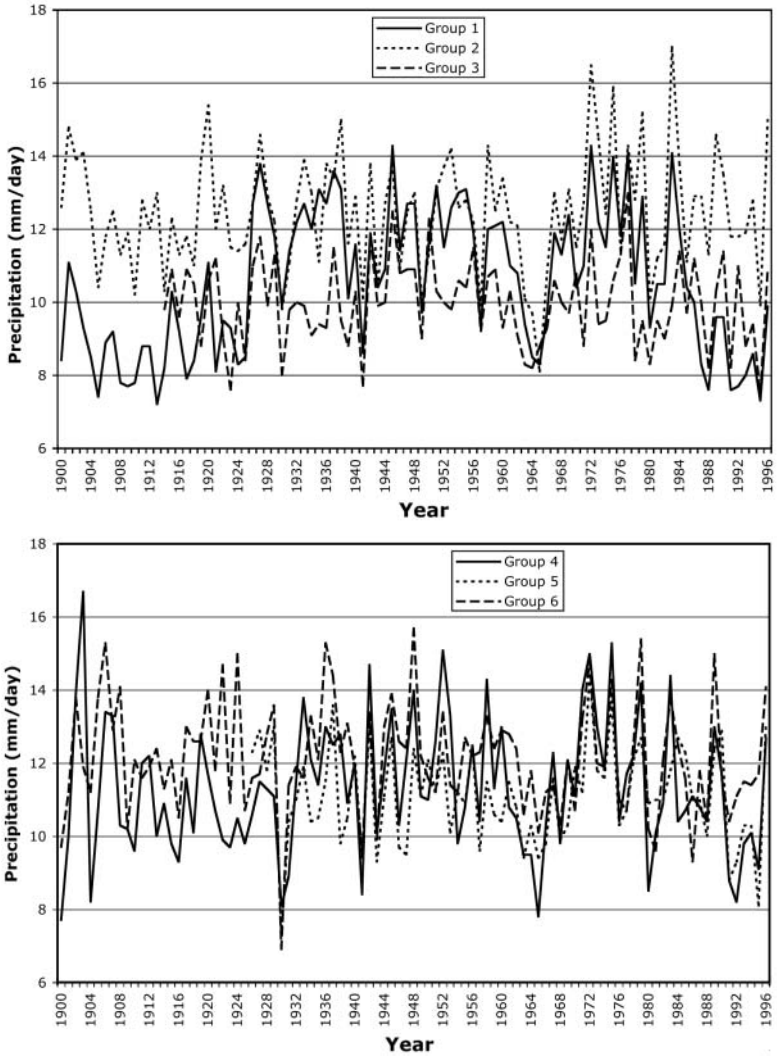

Fig. 7. Annual precipitation for (a) Groups 1 to 3 and (b) Groups 4 to 6 for 1957–1996

the early 1990s. These 3 regions come together around the head of the Chesapeake Bay. A study of river flow and water or chemical inputs to the Chesapeake Bay using recent precipitation data could thus be very dependent on exactly which stations are selected from New Jersey and eastern Pennsylvania. By the same token, this also suggests that the chemical inputs may vary depending on which regions are experiencing greater or lesser precipitation. Not all of the stations in the basin experience the same precipitation anomaly.

Fig. 8 shows the mean monthly precipitation for all 6 regions. Groups 4 to 6 show similar distributions, with a clear summer maximum peaking in July. Groups 1 to 3, however, are somewhat different. Group 3 has a summer maximum, but peaking earlier in the year, while Groups 1 and 2 have much less of a distinct seasonal distribution. Again, the indication is that regional differentiation may be significant for diagnostic analyses of Mid-Atlantic/NE USA rainfall and that simple area averaging of station data may not be an appropriate approach.

4. TIME-SERIES RECONSTRUCTION

To reconstruct a regional precipitation time series we again use a SOM—but in a mode that is more analogous to a nonlinear, oblique, principal component analysis. In this case we begin with the daily precipitation records for the stations in the regions defined in the previous step. The station data for any given region are presented to a SOM that has sufficient nodes to match the length of the original data record (e.g. approx. 200 × 180). If every data record were unique and all records were equally different from one another, each day would map to an individual node in the SOM space, and the node reference vectors would be identical to the station precipitation values. In practice, because many of the days in the record are similar to one another, the observations actually map to a much smaller number of nodes. Each node thus represents a characteristic regional precipitation distribution. Each day that maps to the node has a similar distribution of station precipitation values, and the SOM describes the temporal modes of the precipitation record. Using the large SOM provides the highest possible resolution of the distribution function.

We assume that, for a given set of forcing functions (synoptic circulation, atmospheric humidity, elevation, etc.) there will be a regional precipitation response. The actual precipitation at any station will comprise the station's contribution to that regional signal (a deterministic response), plus a stochastic element that reflects random variability at the station. As we wish to relate the regional precipitation to the larger scale forcing (for climate diagnostic analysis, climate-model validation, developing climate-change scenarios, etc.), it is the deterministic component—the regional response—that is of interest. For a data set with N stations, each node in the SOM represents a small region within the N-dimensional distribution function. Remember that each coefficient in the node reference vector represents a particular station's precipitation; so through the iterative training process, the node coefficients tend toward the mean precipitation at each sta-

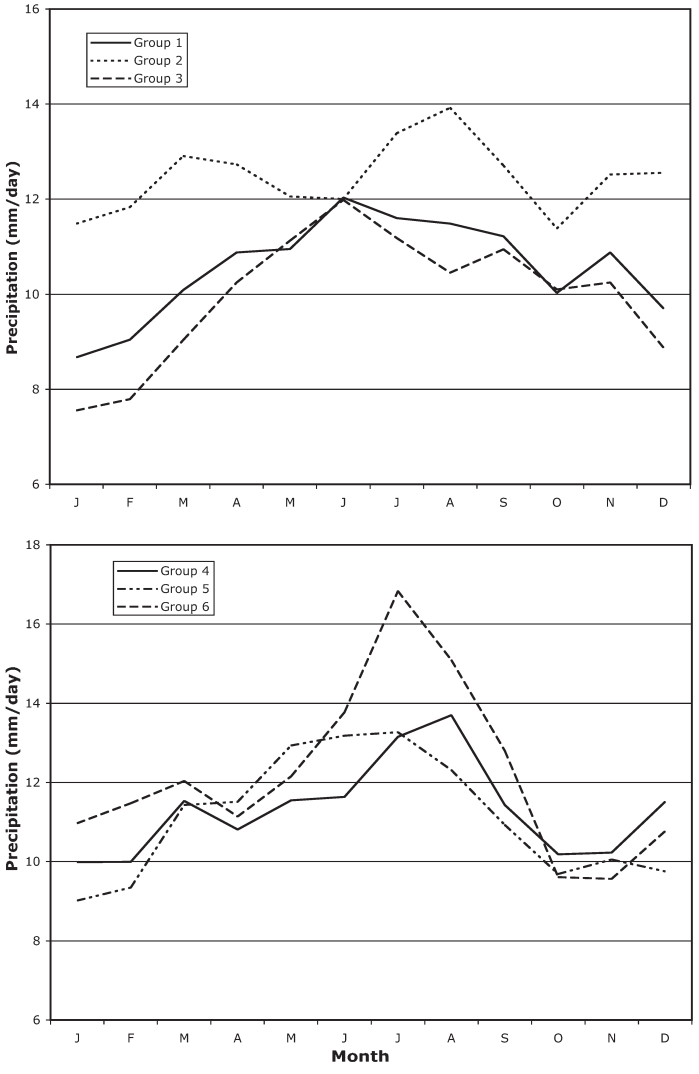

Fig. 8. Mean monthly precipitation for (a) Groups 1–3 and (b) Groups 4–6 in the 2 × 3 SOM (1957–1996)

tion for all days that mapped to that node (i.e. all days with similar regional precipitation characteristics). We take that mean station value to represent the station's contribution to the regional precipitation.

For example, if we consider a single SOM node, and 20 days map to that node, this indicates that the pattern of precipitation across the *N* stations is very similar for those 20 days. For a given station, the corresponding coefficient in the SOM node reference vector represents the mean precipitation for days with that particular regional precipitation distribution. In effect, what we are doing through the SOM is taking the multi-station time series of precipitation, rearranging the data into a multi-dimensional frequency distribution, and at any point in that distribution (defined by a SOM node) we are suggesting that, for any station, the mean value of the nearby data points represents that station's contribution to the regional precipitation. The variability around that localized mean reflects subregional-scale forcing and random variability at each station. The contribution of each station to the regional precipitation record is the station's coefficient in the reference vector for the SOM node to which the daily precipitation maps. The regional precipitation is calculated by mapping every day to a SOM node and replacing the daily station values with the appropriate coefficient from the SOM node. These values are averaged across the stations and through time to obtain the regional precipitation. Where the stations are unevenly spaced, this could be a weighted mean that reflects the station distribution. In this case, the regions are small and the stations fairly evenly distributed, and we simply average the stations to obtain the re-created regional precipitation.

Not only does this approach extract the regional signal from the station data, it also fills in isolated missing values. Some form of regression model is often used to fill in missing data in a multi-dimensional data set where a relationship is derived between Stn A and a nearby Stn B (with which it has a high correlation), or between Stn A and a group of stations, with Stn A as the dependent variable. The regression model is used to fill in missing values for Stn A. A problem with this approach is that the replacement value is simply a linear function of the other variables and thus contains the same variance structure as the independent variables.

With the SOM, if 20 days map to a given node, the SOM coefficients for that node approximate the mean precipitation for each station based on the available data for each station from those 20 days. By replacing the individual daily data with the SOM coefficients to construct the regional data set, we also fill in missing data. Even if a station has missing data for every day that maps to a particular node, there will still be a value for the corresponding coefficient because this will have been updated each time the surrounding nodes were updated during the training process. There will only be missing days in the reconstructed data set if a station had missing data on every day that mapped to a given node and on every day that mapped to all surrounding nodes. Where a station has a missing data record, we are using the SOM to identify all days on which the regional precipitation characteristics are similar to the missing day, and then using the station's mean precipitation for those days to replace the missing value. Hence, we use as much information as is available to determine the precipitation regime for a given day (based on all the stations present), but then use only the station's

precipitation characteristics under that regime to replace a missing value. Thus, the variance structure of the reconstructed time series for the station is internally consistent and independent of the other stations in the group.

4.1. Testing the methodology

The utility of the SOM approach for developing a regional precipitation signal and re-creating missing data is tested using a daily precipitation data set from 52 rainfall stations over the East Coast of the USA, spanning 1950–1999. As these are to be used as the 'truth' to test the methodology, the stations and time period were selected to provide the largest and most complete dataset possible—There are only 45 missing observations out of a total 949 624 records for the entire dataset. The SOM methodology is usually very effective at handling missing data. For example, if there are 10 stations in a region and, on any given day, 1 or 2 stations are missing observations, the SOM will have sufficient information from the remaining stations to fill in the missing data. This should work even if a single station has a large number of missing days; the SOM simply uses the other stations and adds data from the intermittent station whenever it is available. As the number of missing stations increases, however, the SOM has less information to define the nodes and the results should be less effective.

To test the SOM approach, we randomly flag 20% of the stations out of the dataset and then randomly drop 20% of the data for these stations. We run the SOM re-creation and use the SOM values to replace the missing observations. We compute the mean daily precipitation across all stations using the full observed dataset (truth), the sparse dataset (the observed data with missing observations), and the SOM re-created dataset. We then compute the difference between the 'sparse' and 'true' data and the 're-created' and 'true' data, and calculate the percentage improvement of the SOM re-created data over the 'sparse' data (Table 2). The results show the mean absolute error in mm per day. Negative values in the table show where the re-created values were closer to the truth than the sparse data.

Having run the analysis for 20% missing days, it was repeated for 40, 60 and 80% missing days, and the complete sequence repeated for 40, 60 and 80% of the stations. The stations and observations were selected at random, and the analysis was repeated for 10 sets of randomly generated series in all of the categories. Table 2 gives the mean ± SD from the 10 experiments. Note also that, as the missing stations and observations were selected at random, there should be no bias in the missing data. As we are not selecting data from any particular part of the distribution, we would expect little change in the mean. However, with up to 80% missing data from 80% of the stations, the SOM re-creation still produces a more accurate representation of the dataset than does a straight average of the available data. The results support the hypothesis presented earlier – that the SOM is effective at filling in missing data from individual stations and is less effective only when the number of missing stations and the amount of missing data are both large. The improvement in the SOM reconstruction compared to the straight averaging procedure only drops below 15% when 80% of the stations have 60% missing data, or when 60% of the stations have 80% missing data. It is interesting to note that the improvement is smallest (in terms of percentage difference) when the number of missing stations and days is small (when there is enough data for the straight averaging to give a reasonable estimate of the daily means), and when the number of missing stations and days is large (when the SOM procedure has less data to work with). The improvement produced by the SOM is greatest at intermediate levels of missing stations and days. However, the SOM re-creation always gives better results than simply averaging the data in the incomplete dataset.

Table 2. Comparison of the error due to missing data in the observed and SOM reconstructed datasets for precipitation (mm d^{-1}). Missing days and stations with missing data were selected at random

Missing stations (%)	Missing days (%) 20	40	60	80
Error for the 'sparse' dataset (mean ± SD)				
20	0.40 ± 0.18	0.51 ± 0.17	0.60 ± 0.16	0.64 ± 0.16
40	0.55 ± 0.16	0.69 ± 0.14	0.80 ± 0.13	0.88 ± 0.13
60	0.64 ± 0.15	0.83 ± 0.13	0.97 ± 0.12	1.07 ± 0.11
80	0.72 ± 0.14	0.97 ± 0.11	1.16 ± 0.10	1.33 ± 0.09
Error for the SOM corrected dataset (mean ± SD)				
20	0.34 ± 0.18	0.41 ± 0.18	0.48 ± 0.16	0.53 ± 0.17
40	0.43 ± 0.18	0.55 ± 0.16	0.66 ± 0.15	0.73 ± 0.15
60	0.50 ± 0.17	0.67 ± 0.15	0.82 ± 0.13	0.96 ± 0.11
80	0.57 ± 0.16	0.80 ± 0.13	1.02 ± 0.11	1.24 ± 0.09
Difference and % difference between SOM corrected and 'sparse' datasets				
20	−0.07 −16.2	−0.10 −19.9	−0.12 −19.5	−0.11 −17.8
40	−0.11 −20.6	−0.14 −20.8	−0.14 −17.7	−0.15 −17.1
60	−0.14 −22.2	−0.16 −19.2	−0.15 −15.2	−0.11 −10.5
80	−0.15 −20.7	−0.17 −17.5	−0.14 −12.4	−0.09 −6.6

Table 3. Percentage improvement in the data record using SOM reconstructed data compared to a dataset with missing values, where the missing values are biased to the high and low ends of the distribution (see text for details). The shaded areas show the conditions for which the SOM reconstruction improved the dataset

		Missing days (%)			
		20	40	60	80
High	Run 1	−26.4	−20.0	−16.7	−1.4
	Run 2	−24.5	−18.0	−16.7	−9.2
	Mean	−25.5	−19.0	−16.7	−5.3
Low	Run 1	−21.3	−20.0	−13.3	53.8
	Run 2	−23.2	−20.1	1.6	22.8
	Mean	−22.3	−20.1	−5.8	38.3

Table 3 shows the results when a bias is added to the missing stations. The stations are ordered by their mean precipitation from highest to lowest. Table 3 shows the effects of removing the highest 20%, 40%, etc., of observations from the top 20% of high-rainfall stations, and the lowest 20%, 40%, etc., of observations from the bottom 20% of low-rainfall stations. In this case we see that the SOM re-creation is superior in all cases, except when 80% of the observations are deleted in the low-rainfall case.

4.2. Application to the regional datasets

Examples of the SOM re-created regional precipitation are given in Fig. 9. Fig. 9a shows the precipitation for the 6 regions defined by the 2×3 SOM. The regional values are obtained simply by averaging all available station data for each year. These represent the 'observed' data. Fig. 9b shows the regional data recreated through the SOM procedure. Note that the match between the two is very close back to the middle of the twentieth century. During this period there is clear decadal-scale variability, with all of the regions showing relatively high precipitation amounts in the 1950s, lower precipitation in the 1960s, a marked increase in the 1970s and 1980s, and declining values in the late 1980s and 1990s. Prior to the 1950s, however, there is a substantial difference between the observed and re-created data. The observed data show higher values and little change through time for most of the regions. The SOM re-created data, on the other hand, show lower values and an increasing trend in the first half of the century. Part of the reason for this is given in Fig. 10, which shows the number of stations reporting in each region through the period of record. There are stepwise increases in the number of stations in the mid-1920s and late 1940s. During the early part of the record, the stations in Regions 2, 4 and 6, in particular, have relatively high rainfall amounts compared to the stations that were added later in the record. Consequently, when the available data are simply averaged in the 'observed' dataset, the regional precipitation is relatively high. When the regional precipitation is re-created through the SOM, we see that precipitation was actually lower, and there has been an increasing trend through time.

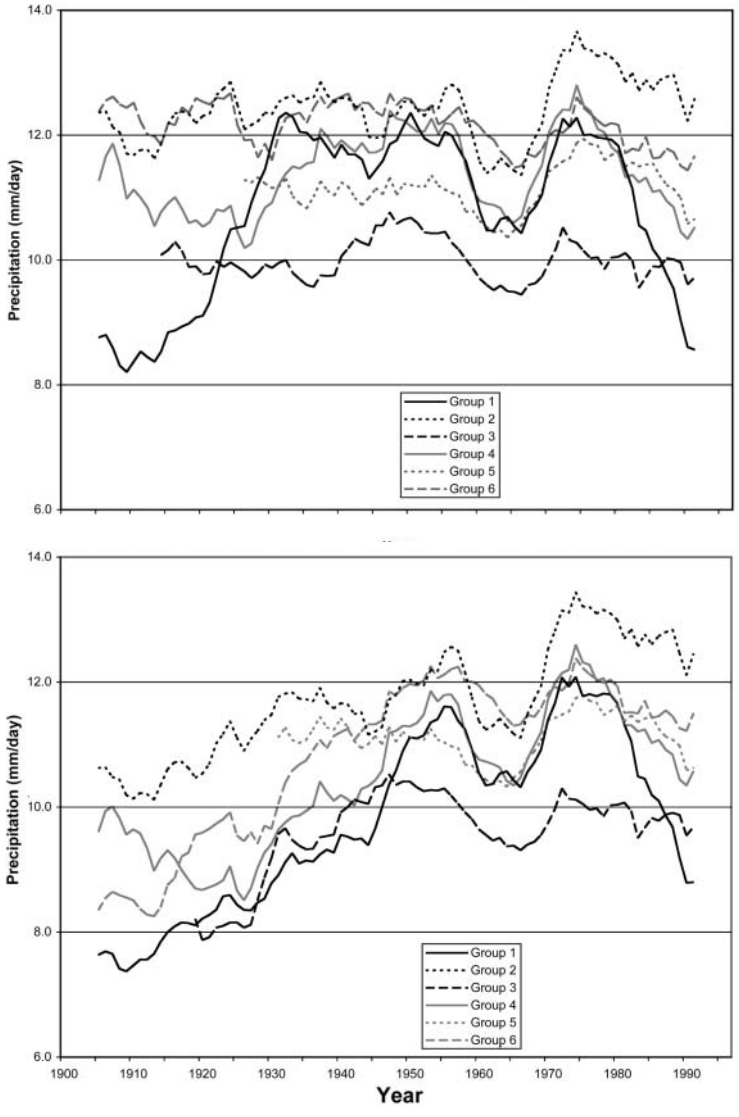

Fig. 9. Mean annual daily precipitation for the 6 regions identified by the 2×3 SOM (11 yr running mean). (a) Raw data (averages of the stations in the regions); (b) SOM re-created time series

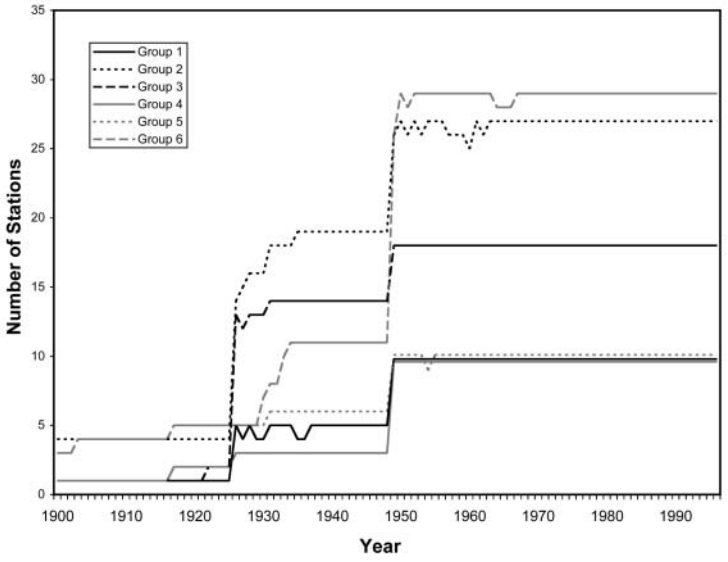

Fig. 10. Number of stations reporting in each of the 6 regional groups

Even more dramatic differences are seen in the number of days per year with extreme precipitation events. Fig. 11 shows the number of days each year with measurable precipitation, averaged across the stations in the group, of <8 mm d^{-1} (i.e. low-precipitation events), and Fig. 12 shows the number of days per year when precipitation amounts for the group exceeded 80 mm d^{-1} (approximately the top 15% of days in the record). In both cases the influence of the number of reporting stations is obvious in the observed data. The SOM re-created data improves the dataset overall, and particularly in the middle part of the record, but the number of stations in the early part of the record is too few even for the SOM to fully correct.

5. CONCLUSIONS

The recent interest in global change and global warming has led to an increased focus on the interactions between human and physical systems. The attention has broadened to encompass a range of biophysical processes with a growing recognition that, while many of these are global in extent, their interaction with human systems is at the regional scales that are important to planners and policy makers. Regional boundaries are frequently defined topographically, hydrologically, or in terms of economic and political factors. Where precipitation is an important variable in the analysis, however, it also makes sense to define regions according to their precipitation characteristics.

This paper demonstrates a mechanism for defining regions that are internally consistent across a range of spatial scales. The technique uses the iterative discriminatory power of self-organizing maps (SOMs) to identify stations with similar precipitation characteristics. The analysis indicates that the Mid-Atlantic/NE USA region can be broadly divided into a northern and a southern precipitation regime, with the border running approximately across Pennsylvania. Both regimes can be further subdivided into smaller regions, depending on the level of generalization required. The same procedure can be used for defining other climatic regions, either in terms of individual

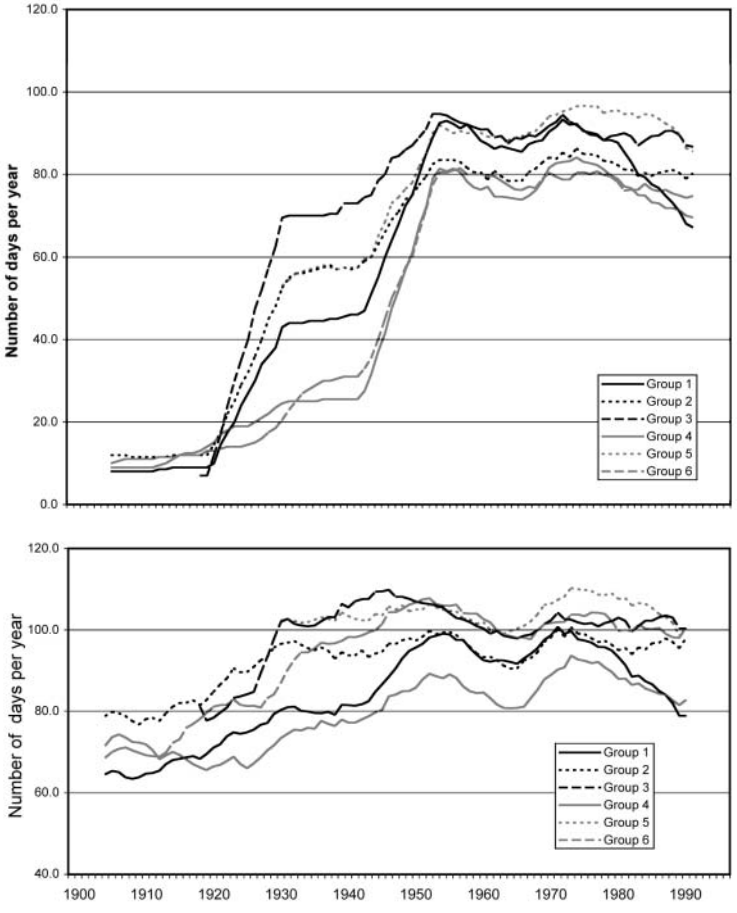

Fig. 11. Number of days per year with low-rainfall events (precipitation greater than zero and less than 8 mm d^{-1}). (a) Raw data for the 6 groups identified by the 2×3 SOM; (b) SOM re-creation for the same groups

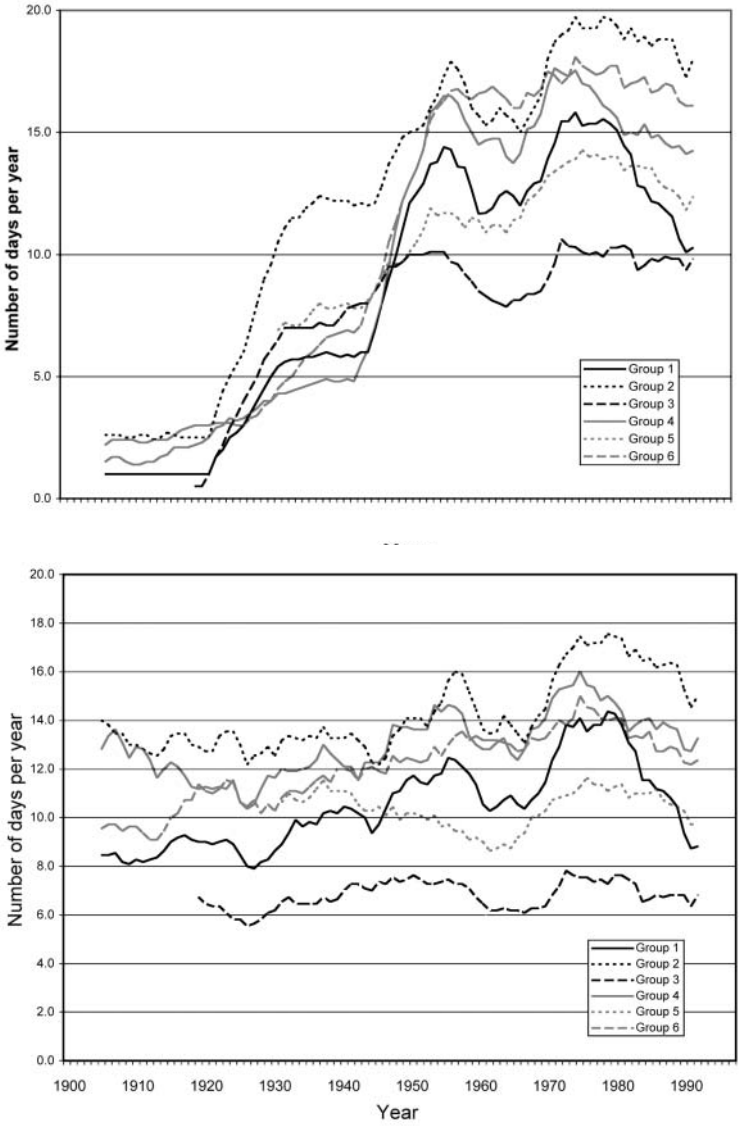

Fig. 12. Number of days per year with high-rainfall events (precipitation greater than 80 mm d^{-1}). (a) Raw data for the 6 groups identified by the 2×3 SOM; (b) SOM re-creation for the same groups

characteristics (e.g. temperature or precipitation) or using multivariate criteria. A different application of the SOM is used to recreate regional precipitation signals and to fill missing data in the temporal record. The analysis suggests that the SOM re-creation improves the time series back through the 1930s, but there are too few stations for earlier periods to establish an accurate regional signal.

Acknowledgements. This work was supported, in part, by EPA Cooperative Agreement Number R-830533-01-0 to A. Fisher, The Pennsylvania State University.

LITERATURE CITED

Comrie AC, Glen EC (1999) Principal components-based regionalization of precipitation regimes across the southwest United States and northern Mexico, with an application to monsoon precipitation variability. Clim Res 10:201–215

Hewitson BC, Crane RG (2002) Self-organizing maps: applications to synoptic climatology. Clim Res 22:13–26

Kohonen T (1989) Self-organization and associative memory, 3rd edn. Springer-Verlag, Heidelberg

Kohonen T (1990) The self-organizing map. Proc IEEE 78:1464–1480

Kohonen T (1991) Self-organizing maps: optimization approaches. In: Kohonen T, Mkisara K, Simula O, Kangar J (eds) Proceedings of the International Conference on Artificial Neural Networks, Espoo, Finland, June 1991. Elsevier, Amsterdam, p 981–990

Kohonen T (1995) Self-organizing maps. Springer-Verlag, Heidelberg

Knapp PA, Yin ZY (1996) Relationships between geopotential heights and temperature in the South East United States during wintertime warming and cooling periods. Int J Climatol 16:195–211

Waylen PR, Quesada ME, Caviedes CN (1996) Temporal and spatial variability of annual precipitation in Costa Rica and the Southern Oscillation. Int J Climatol 16:173–193

Editorial responsibility: Andrew Comrie, Tucson, Arizona, USA

Submitted: January 30, 2003; Accepted: August 12, 2003
Proofs received from author(s): November 11, 2003