Vol. 22: 13–26, 2002 | CLIMATE RESEARCH Clim Res | Published August 8

Self-organizing maps: applications to synoptic climatology

B. C. Hewitson[1,*], R. G. Crane[2]

[1]Department of Environmental and Geographical Sciences, University of Cape Town, Private Bag, Rondebosch 7701, South Africa
[2]Department of Geography, Pennsylvania State University, University Park, Pennsylvania 16827, USA

ABSTRACT: Self organizing maps (SOMs) are used to locate archetypal points that describe the multi-dimensional distribution function of a gridded sea level pressure data set for the northeast United States. These points—nodes on the SOM—identify the primary features of the synoptic-scale circulation over the region. In effect, the nodes represent a non-linear distribution of overlapping, non-discreet, circulation types. The circulation patterns are readily visualized in a 2-dimensional array (the SOM) that places similar types adjacent to one another and very different types far apart in the SOM space. The SOM is used to describe synoptic circulation changes over time, and to relate the circulation to January station precipitation data (for State College, Pennsylvania) in the center of the domain. The paper focuses on the methodology; however, the analysis suggests that circulation systems that promote precipitation have decreased over the last 40 yr—although January precipitation at State College has actually increased. Further analysis with the SOM indicates that this is due to a change in precipitation characteristics of the synoptic-scale circulation features, rather than to their frequency of occurrence.

KEY WORDS: Self-organizing maps · Synoptic climatology · Downscaling · Climate change · Synoptic classification

1. INTRODUCTION

Synoptic climatology has a rich methodological heritage of techniques designed to relate synoptic-scale atmospheric circulation to a local climate or environmental response. While its historic development can be traced back to the late 19th century (e.g. Köppen 1874, Abercromby 1883, 1887), synoptic climatology was established as a distinctive climatological sub-field with the publication of 'Synoptic climatology: methods and applications' by Barry & Perry (1973). Here it was defined as 'obtaining insight into local or regional climates by examining the relationship of weather elements, individually or collectively, to atmospheric circulation processes.'

The most common approach to synoptic climatology is to partition the atmospheric state into broad categories (either in terms of spatial pattern or the multivariate characteristics of an airmass), and to relate these synoptic categories or 'types' to some dependent variable such as local temperature, acid deposition, etc. Synoptic classification has often been used as a data reduction technique in process studies that examine interactions or relationships between the circulation and local environmental parameters (e.g. Crane 1978, Yarnal 1984, Wigley & Jones 1987). Synoptic classification has also been used to extend data records: where the circulation record is much longer than the record for the environmental parameter of interest; a transfer function is developed between the environmental parameter and the synoptic types, and the circulation record is used to extend the local environmental data back in time (e.g. Barry et al. 1981).

Early approaches used manual classification to define synoptic types (e.g. Lamb 1950); although effective, these manual techniques are extremely labor intensive. Subsequent automated approaches based on quantitative algorithms gave rise to a plethora of techniques based on a few core procedures involving

*E-mail: hewitson@egs.uct.ac.za

some form of correlation, cluster, and/or eigenfunction analysis. In all cases, the approach of generalizing the circulation into characteristic modes or synoptic types required a fine balance between producing a small enough number of types to easily visualize and conceptualize the circulation, while avoiding so much generalization that the strength of any relationship to a local climate variable was lost. The major problems with this approach are due to the degree of within group variability produced. It is also common that days in the same synoptic type can often be associated with a very different local response, or that the same response can be obtained from different synoptic types. The fundamental characteristics of synoptic classification techniques are effectively summarized in Yarnal (1993).

Underlying this traditional approach to synoptic classification is the premise that the continuum of weather states may be effectively divided into a small number of categories with clear discernable boundaries. This premise clearly has limitations: while typing weather systems gives a good first-order insight into the basic characteristics of the climate system, much of the information is inherently subsumed by the degree of generalization imposed.

In response to this, Hewitson & Crane (1992a,b) proposed that the system can be treated as a continuum with a continuous function, and that quantitative relationships between the atmosphere and local surface variables can be developed in the form of a downscaling transfer function. This procedure is diametrically opposed to the classic synoptic climatology approach. In this case there are no synoptic types, but rather a transfer function approximating the continuum of the cross-scale relationship. However, while this approach effectively re-captures much of the relationship information lost in the generalized typing schemes, it has its own shortcomings in that it is often difficult to interpret physical processes from the relationship represented by the cross-scale function. For example, local precipitation may be derived as some weighted function of atmospheric circulation and humidity measures on a larger spatial grid.

Nonetheless, such empirical downscaling, as it has become termed, is now in widespread use and, as with synoptic typing schemes, has promoted numerous methodologies ranging from simple linear regression (e.g. Sailor & Li 1999), to non-linear artificial neural nets (ANNs) (e.g. Hewitson & Crane 1996) and stochastic weather generators (e.g. Bellone et al. 2000).

Both of these techniques—synoptic typing and empirical downscaling—effectively accomplish the principle objectives of synoptic climatology, yet they represent widely divergent approaches; thus, a middle ground that treats the atmosphere as a continuum yet retains interpretability, would represent a significant advantage.

2. SELF-ORGANIZING MAPS

Self-organizing maps (SOMs) (Kohonen 1989, 1990, 1991, 1995) offer an alternative approach to synoptic climatology that provides a mechanism for visualizing the complex distribution of synoptic states, yet treats the data as a continuum. SOMs are in widespread use across a number of disciplines (e.g. Joutsiniemi et al. 1995, Palakal et al. 1995, Chen & Gasteiger 1997), but have little exposure to date in the climatological literature. SOMs were introduced to the physical geography community as part of a broader discussion on neural nets (Hewitson & Crane 1994), but only the ANN component has become widely adopted. Hewitson (1999, 2001) and Crane & Hewitson (1998) include SOMs as part of an ANN-based downscaling procedure; Main (1997) used SOMs to investigate seasonal cycles in general circulation models (GCMs); Hudson (1998) uses SOMs to evaluate frequency changes of synoptic events in a GCM perturbation experiment. SOMs have also been used as a mechanism for climate classification by Malmgren (1999) and Cavazos (1999, 2000), and for cloud classification by Ambroise et al. (2000).

In many respects SOMs are analogous to more traditional forms of cluster analysis. Given an N-dimensional cloud of data points, the SOM will seek to place an arbitrary number of nodes within the data space such that the distribution of nodes is representative of the multi-dimensional distribution function, with the nodes being more closely spaced in regions of high data densities. Most cluster algorithms are designed to identify groups that minimize the within-group differences while maximizing the between-group differences. There are numerous ways in which these differences can be defined, including, for example, cluster algorithms that define group centroids and measure the distance between the centroid and each group member and the distance from the group centroid to all other centroids. Alternatively, the algorithm may measure the distance between all points in the group and between each point in the group and each point in a neighboring group. Some algorithms will allow observations to belong to more than 1 group, although in many cases this results, in effect, from a post-processing step in which, once the groups are defined, a probability of group membership is calculated for each point in all of the groups. A comparison of several clustering algorithms (Ward's minimum variance, average linkage and centroid) used in a synoptic classification procedure is described by Kalkstein et al. (1987).

SOMs differ from traditional cluster algorithms in 2 significant characteristics, the first being the way in which groups are defined. While the end result of the SOM analysis is some form of data clustering, unlike a

clustering algorithm the basic SOM methodology is not primarily concerned with grouping data or identifying clusters. As noted above, SOMs attempt to find nodes or points in the measurement space that are representative of the nearby cloud of observations and, when taken together, describe the multi-dimensional distribution function of the data set.

The initial step in the SOM routine is to define a random distribution of nodes within the data space. The nodes are defined by a reference vector of weighting coefficients, where each coefficient is associated with a particular input variable. If, for example, the initial data set comprises a time series of sea level pressure observations on a 10 × 20 spatial grid, each node in the SOM will have a reference vector of 200 coefficients. For every node, the *n*th coefficient in the reference vector will be associated with the *n*th input variable. Thus, each node has an associated reference vector equal in dimension to the input data. As each data record is presented to the SOM, the similarity between the data record and each of the node reference vectors is calculated, usually as a measure of Euclidean distance. The reference vector of the 'best match' node is then modified such as to reduce the difference with the input vector by some user-defined factor, or *learning rate*. The data record does not become part of a group at this time; it is simply used to adjust the location of the SOM node in the data space.

A major difference with most cluster algorithms is that it is not only the closest node that is updated during this process, but all surrounding nodes are also incrementally adjusted toward the input vector in inverse proportion to their distance from the 'winning' node. The user determines the size and shape of this update kernel. In this application we actually train the SOM twice. The first set of training iterations use random starting points for the node vectors and a relatively large update kernel that is close to the size of the smaller SOM dimension. This produces a first broad distribution of nodes. The second set uses the final node vectors of the first run as the starting points and a smaller update kernel to refine the mapping.

This iterative process continues during several cycles through the data set until there are no more changes in the node locations. The net result is that the SOM will cluster nodes in regions of the data space that have high data densities (where there is more information). The SOM reference vectors are iteratively adjusted such that they span the data space, and each node represents a position approximating the mean of the nearby samples in data space. This procedure effectively identifies 'archetypal' points that span the continuum of the data. The node to which each sample maps with the lowest error at the end of the training is recorded, and new data can be assigned to one of the nodes, assuming they are from the same population as the training data. If we think of the SOM nodes as defining eventual group or sub-group centroids, this technique explicitly recognizes that groups are not discrete, non-overlapping entities and allows individual observations to contribute to the definition of more than 1 node or group.

There are many different forms of cluster analysis and, at one level, we could regard SOMs as simply another clustering option. However, most clustering algorithms make some assumption about the data structure or are based on underlying statistical model that describes the data distribution. Where a large group lies very close to a small group (in the data space), for example, some algorithms would group both together while others might split off part of the larger group and join it to the smaller group, if that maximizes the similarity measure being used. Similarly, some algorithms are appropriate for distinguishing between spheroidal or hyperspheroidal clusters, but not linear clusters etc. One of the advantages of the SOM approach is that it is much more versatile. Because of its iterative nature, and because it locates nodes that span the data space, the results of the SOM are less dependent on the data conforming to a specific distribution or underlying model.

Fig. 1 demonstrates the principles of a SOM using a simple artificial 2-dimensional data set. The data set is constructed using a random number generator to create a skewed distribution and to incorporate nonlinearity and a break in the data, with 2 values per observation. The red points in Fig. 1 denote the data samples. To develop the SOM mapping, the node weight vectors are initialized with random numbers prior to a training phase. The SOM is trained using these data and the final SOM node locations are shown as the blue points in Fig. 1. The example serves to illustrate 3 key aspects of the SOM behavior:

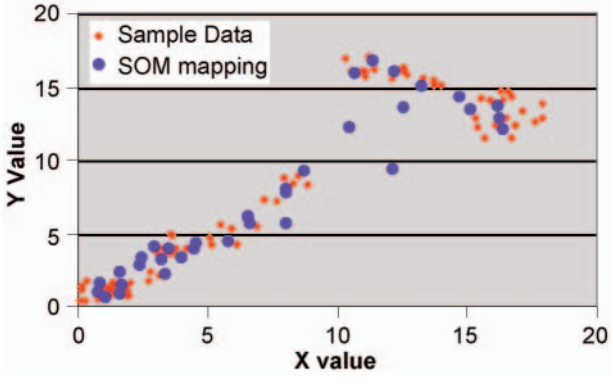

Fig. 1. Distribution of data points and self-organizing map (SOM) nodes in simple 2-dimensional space

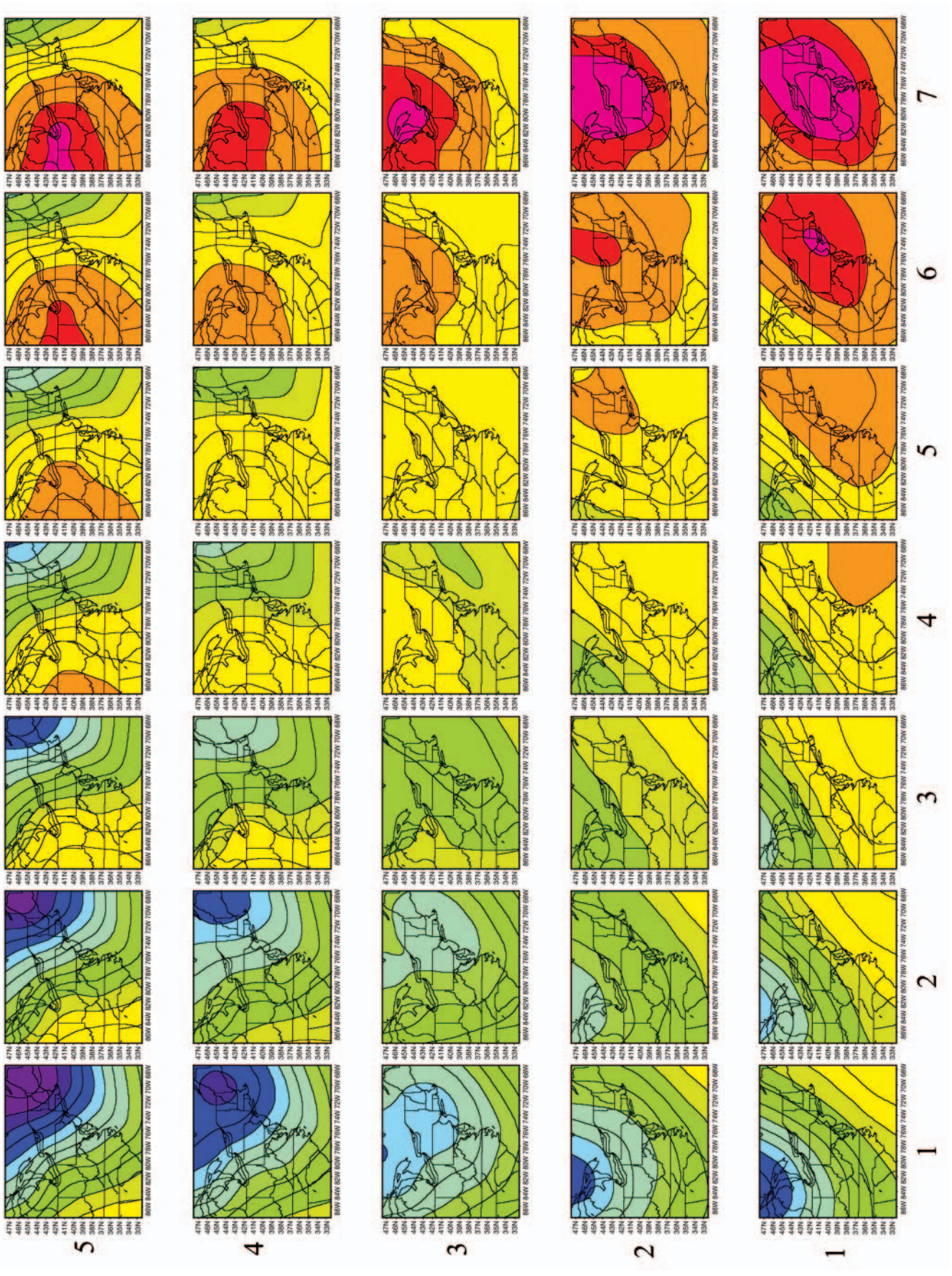

Fig. 2. A 5 × 7 SOM of January sea-level pressure (SLP) for the north-east United States. Blues represent relatively low pressure, while reds indicate high pressure

- The SOM assumes the data are continuous. In Fig. 1, the SOM locates a series of points or nodes that are spread through the data cloud. The result is that the SOM attempts to span the break in the data, which may or may not be an advantage: if there are missing data then the SOM provides a means to interpolate, while if there are genuine discontinuities in the data, the interpolation will place nodes within the discontinuous region. However, when calculating frequencies of occurrence on each node (the number of observations mapped to a particular node), as in Fig. 6, there would be zero observations mapped to that node.
- Fewer SOM nodes are allocated where there is sparse data, while more SOM nodes are allocated to regions of the data space where there is greater information in the data set. This behavior allows discrimination between more subtle variations — where the information to do so exists. By placing more nodes where there are more data points, the SOM attempts to represent the details of the data distribution at whatever level of generalization (SOM dimension) is used.
- The SOM captures the non-linear characteristics of the data. While the measure of similarity between the data and the reference vector is linear, the iterative training procedure allows the SOM to account for non-linear data distributions such as the one shown in Fig. 1.

The second major difference between SOMs and traditional clustering algorithms is that the SOM presents an effective means of visualizing the relationships between the nodes. If a time series of synoptic charts is envisioned, the charts can be placed on a flat surface such that similar synoptic states are piled together or placed in adjacent piles. Synoptic states that are very dissimilar are widely separated, and transitional states are placed between the groups. In the same manner, a SOM will arrange the distribution of nodes into a 2-dimensional array (the self-organizing 'map'), where similar nodes are located close together in the array and dissimilar nodes are further apart, causing the SOM to produce a mapping or projection of the multi-dimensional data distribution function onto a 2-dimensional plane. This property of the SOM stems from the iterative nature of the process and the use of the spatial update kernel described above.

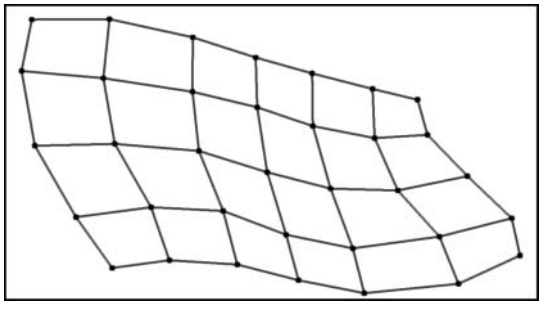

Fig. 3. SOM node distortion surface

The update kernel assures that similar nodes will be located near each other in the SOM array. Even where the initial random distribution of weight vectors causes very different nodes to be located next to each other, ultimately one node will begin to dominate that region of the SOM and subsequent iterations will force the 2 very different nodes further apart in the SOM space. The typical result is what we see in Fig. 2, where very different synoptic states map to the corners and edges of the SOM. For the present data set, a similar result is always obtained regardless of the initial distribution of weights. Any run of the SOM with different starting points will always place the high-pressure mode currently in the bottom right of the SOM (node 1,7 in Fig. 2) into one of the corners, and it will always map the patterns with the low pressure in the north-east (node 5,1) to the opposite corner.

The SOM mapping will always locate similar nodes close to each other in the SOM space. However, the regular array presented in Fig. 2 (and subsequent figures) can be a little misleading. This arrangement does not present a quantitative measure of similarity. The 4 nodes in the lower right (1,6; 1,7; 2,6; 2,7) may be much more similar to each other than are the nodes in the top left (5,1; 5,2; 4,1; 4,2). Similarly the difference between nodes 1,7 and 2,7 may be much less than the difference between nodes 2,7 and 3,7 or 3,7 and 4,7, etc. Knowing how similar or dissimilar nodes are to each other could make a difference in subsequent analyses and applications using the SOM. A simple measure of similarity is obtained by computing the Euclidean distance between nodes in the original measurement space. As the nodes represent *L*-dimensional vectors in the original data space and the object is to show how these are related to each other in a lower-dimensional space (2 dimensions in this case), one approach could be to use a Sammon mapping scheme (Sammon 1969). In the present case, we compute the distance between each node and its adjacent nodes and show this distance as a distortion surface (Fig. 3). The distortion mapping shows that nodes 1,6, 1,7, 2,6 and 2,7 are indeed closer to each other than are nodes 5,1, 5,2, 4,1 and 4,2.

A practical software package for implementing SOMs is freely available (http://www.cis.hut.fi/research/som-research/) along with extensive references and guidelines for practical implementation. This paper does not seek to duplicate the extensive theoretical discussion on SOMs, as these are well described in the literature, beginning with Kohonen (1989, 1990, 1991, 1995). Rather, the remainder of the paper focuses on

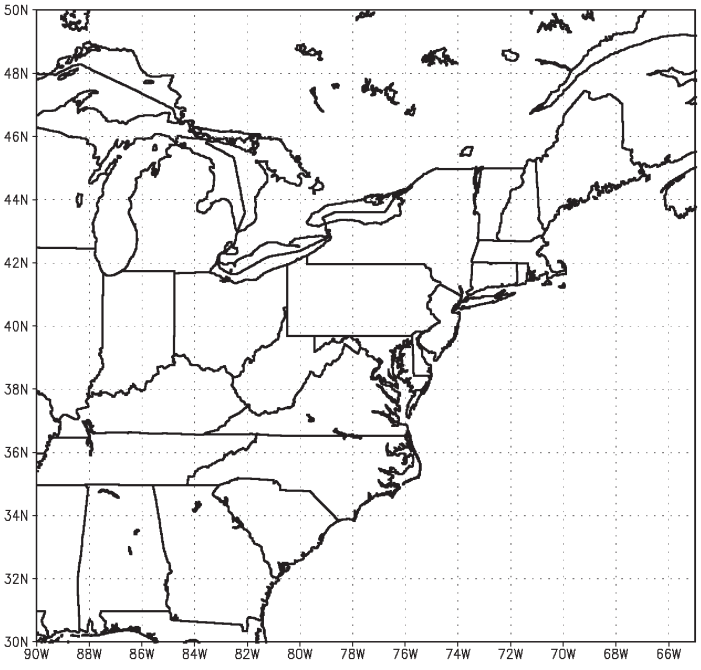

Fig. 4. Spatial domain used for the SLP SOM analysis

the application of a SOM to a synoptic climatology application relating sea-level pressure (SLP) fields over the north-eastern USA to station precipitation in the center of the domain.

3. SOM MAPPING OF SEA-LEVEL PRESSURE

The atmospheric data used in this example are NCEP reanalysis fields of SLP (Kalnay et al. 1996). The data are 12 hourly 2.5° × 2.5° gridded data spanning a domain centered on Pennsylvania (90–67.5° W, 30–60° N; Fig. 4). Data are for Januarys only, from 1958–1997, and provide a 40 yr climatology of 14 880 samples for the northeastern USA. No pre-processing is undertaken on the data.

The size of the SOM array is defined by the user and determines the degree of generalization that will be produced by the SOM—the more nodes, the finer the representation of detail, while the fewer nodes, the broader the level of generalization. However, the same broad patterns are revealed at each level of generalization. A 3 × 4 SOM or a 7 × 9 SOM would both show the same broad pattern with nodes that represent dominant high-pressure systems grouped in one corner of the SOM and nodes representing low-pressure systems in the opposite corner, with mixed patterns in between. The 5 × 7 array is analogous to using 35 clusters in more traditional methodologies, although in the case of SOMs the 35 'classes' will represent synoptic states spanning the continuum as represented by the data samples. Other synoptic studies over this domain have used a relatively small number of synoptic types (e.g. Comrie 1992, Yarnal & Frakes 1997). While 35 SOM nodes is significantly more, with the ease of visualization shown below it affords greater resolution of the synoptic scale variability of the circulation over the region.

Training of the SOM was accomplished by randomly initializing the node vectors, and then training in 2 successive passes of 50 000 iterations each. During the first pass the learning rate (the measure of how much a node vector is adjusted to a data sample) was kept at the default for the software, with an initial radius of update around a best-matched node set at the smaller

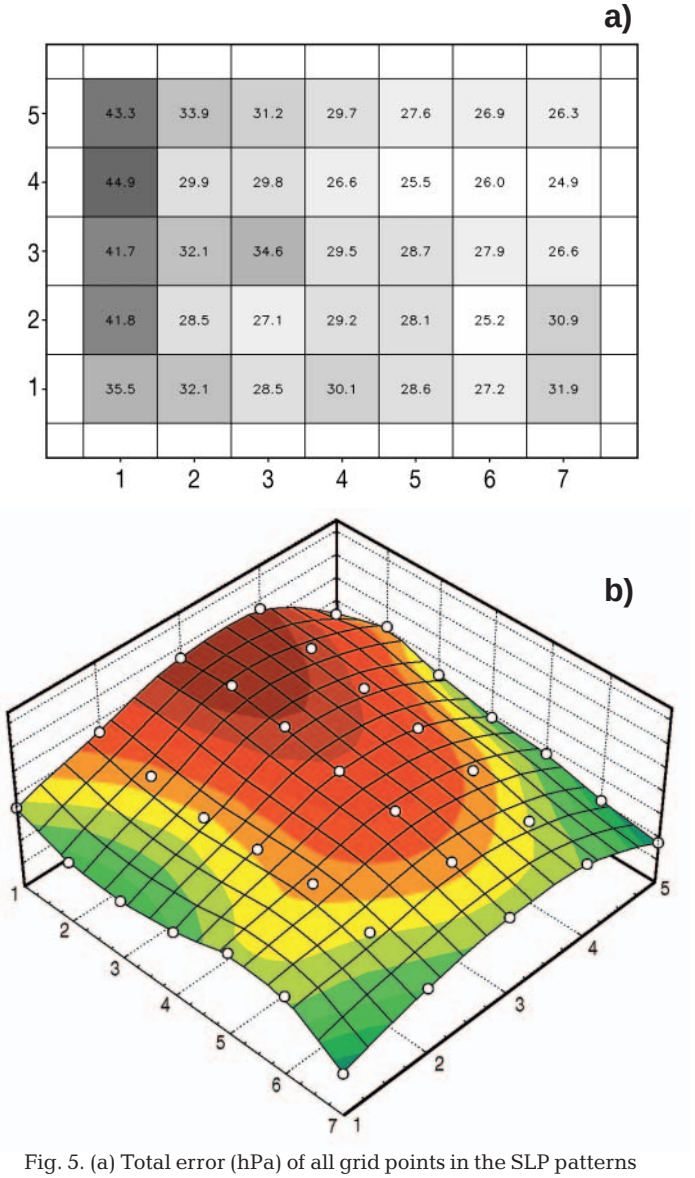

Fig. 5. (a) Total error (hPa) of all grid points in the SLP patterns mapped to each SOM node. (b) Data as in (a), but from a 2nd training of the SOM, displayed as a 3-dimensional error surface

of the array dimensions (5). This relatively large update radius is used in order to develop the broad mapping across the array of nodes. During this training pass the update radius is progressively reduced. In the second training pass the initial update radius is set to 3 (again progressively reduced during training), with a training rate half that of the first pass. This second set of iterations develops the finer details of the mapping.

Each node's vector now forms a reference vector for a particular synoptic state. As such, the vector may be represented as a spatial SLP map, generating a matrix of SLP maps arrayed in relative position to the nodes in the SOM array. These represent the continuum of synoptic states that would be obtained if the synoptic charts were spread across a table as described earlier. The final 5 × 7 array of reference vectors are shown in Fig. 2. Similar synoptic states are located adjacent to one another in the SOM mapping, while dissimilar states are at opposite extremes of the SOM space. The continuum of states is easily visualized in the variation from the dominant high-pressure systems through transition states to synoptic fields dominated by deep low-pressure systems.

The size of the data space represented by each node, or the variance of synoptic states related to each node, is evaluated by determining the error with which the sample values map to a given node. The sum-of-squared differences between each sample and the node reference vector provide a measure of whether the node represents a broad region of the data space or a clearly defined synoptic state. In traditional cluster analysis this would be analogous to the measure of within group variance. The magnitude of the error at each node is shown in Fig. 5a. Training a new SOM produces nearly identical results, as shown in Fig. 5b which displays the error as a 3-dimensional error surface in Fig. 5b. Each shaded box in Fig. 5a represents one of the nodes in the 5 × 7 array. With reference to Fig. 2 (the node reference vectors) it can be seen that the highest variability of synoptic states on a node (shown by higher errors) is associated with the transient low-pressure systems, while the lowest errors are, not surprisingly, associated with the relatively stationary dominant high-pressure systems. If the errors are summed across the SOM array, the mean absolute error can be used as a measure for determining optimum SOM size. The circulation can be mapped to SOMs with several different dimensions and the change in error used to select the SOM to be retained for further climatological analysis. However, as noted above, the same broad groupings appear at all dimensions and the dimensions simply determine the degree of generalization that will be obtained. A subjective decision on SOM size, based on the particular application of interest, can be just as valid a criterion for selecting SOM dimensions. In this case we chose a 5 × 7 array in order to illustrate the detail that can be obtained, while keeping the SOM small enough to display the results in a single diagram (i.e. Fig. 2). The SOM procedure, at this point, has achieved what would be equivalent to the typing phase in a traditional synoptic climatological analysis. Each 'type' (node) represents a range of states within the continuum described by the original data space. In this case, however, the relative relationship between the 'types' is clearly visualized, and these 'types' do not represent discrete classes. At this stage, a number of simple yet powerful analyses are possible that provide insight into the nature of the SLP fields and their relationship to, in this example, station precipitation data.

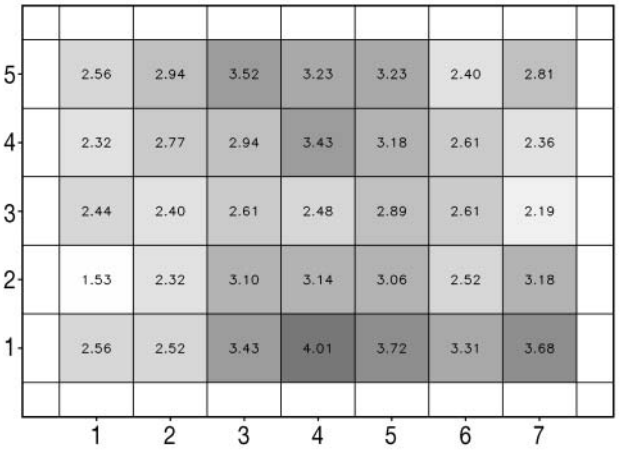

Fig. 6. Climatological 40 yr mean frequency (%) of days in a month mapping to each SOM node

4. FREQUENCY ANALYSES OF SYNOPTIC SYSTEMS

One of the simplest investigations that can be undertaken with the trained SOM is to look at the frequency of occurrence of synoptic systems. After training, the data were presented to the SOM to determine which node exemplified that particular synoptic state. A grey-scale map showing the frequency of occurrence of synoptic states across the SOM space is constructed by accumulating the number of days mapped to each node.

Fig. 6 shows the frequency of days mapped to each node over the 40 yr of January SLP. Note that Fig. 6 and subsequent figures plot data on the same SOM array to facilitate comparison with Fig. 2. However, they should be interpreted with consideration given to the similarity and error mappings shown in Figs. 3 & 5. Each shaded square in Fig. 6 represents 1 node in the SOM array, while the numbers are the percentage frequency of occurrence. The figure indicates that fre-

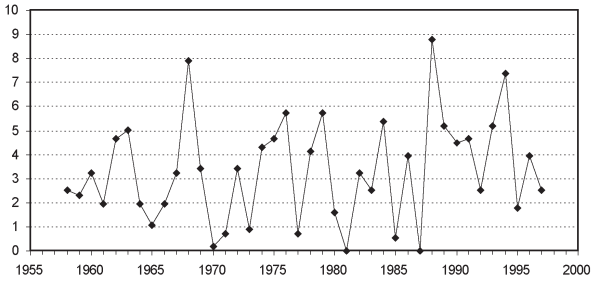

Fig. 7. January annual frequency (average of nodes 5,1, 6,1 and 7,1)

quencies are distributed fairly evenly over the nodes, with relative minima in the left and central portions of the array. Relative maxima are found in the bottom-right and the top- and bottom-central portions of the SOM array. The minima are associated with indeterminate patterns in the center of the array, or nodes in the left part of the array that represent strong low-pressure systems dominating much of the region (Fig. 2). The higher frequencies in the bottom-right quadrant represent strong central high-pressure systems, while the relative maxima in the top- and bottom-central parts of the array are associated with transitional patterns. While a traditional synoptic typing approach would identify the central high and low-pressure systems as independent synoptic types, it would also attempt to assign the remaining days to classes with high- or low-pressure systems located in different parts of the region. Conversely, the SOM clearly indicates that the region is dominated by transitional states rather than discrete synoptic types.

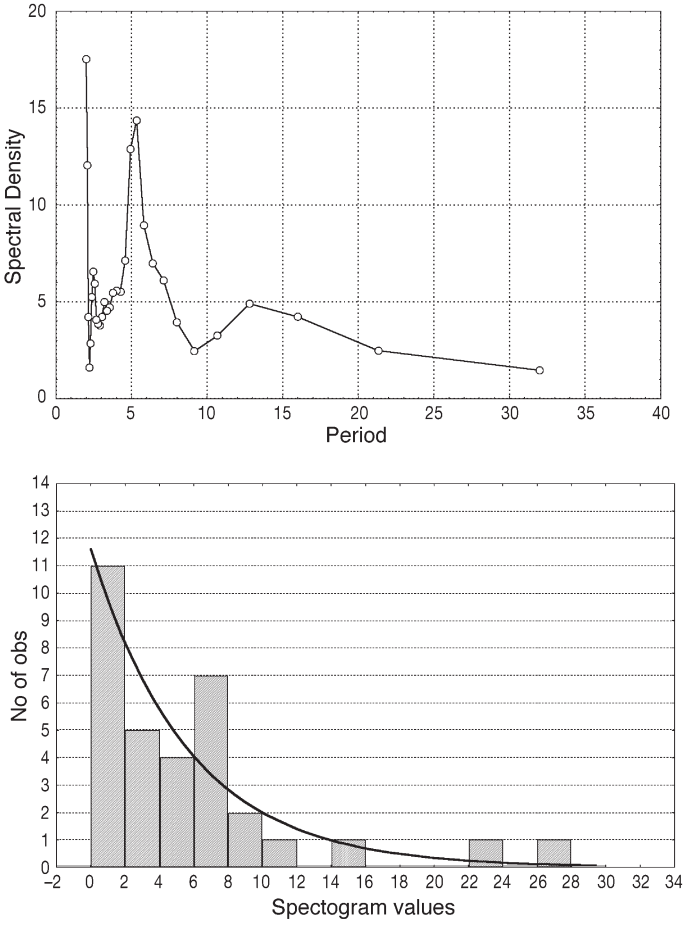

Fig. 8. Fourier analysis of January annual frequencies (dominant high-pressure system)

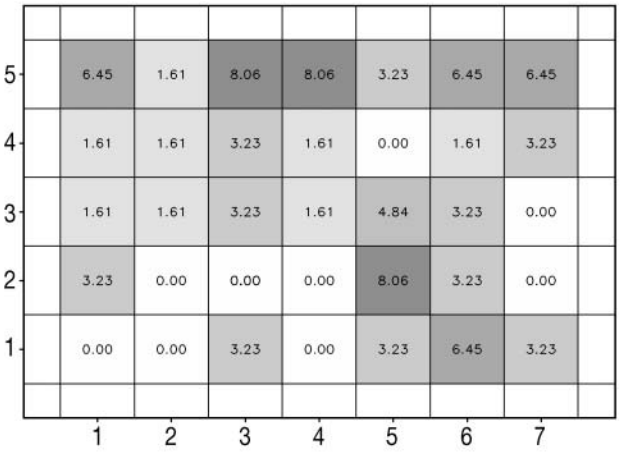

Fig. 9. Number of days mapping to each SOM node for January 1978 (low-precipitation month)

Plotting the frequency of occurrence (number of days mapping to each node) through time presents information on the temporal behavior of the synoptic states. For example, the time series of the average monthly frequency of occurrence for the 3 nodes in the bottom-right corner of the SOM map (nodes 5,1, 6,1 and 7,1; the dominant high-pressure patterns), are shown in Fig. 7. For this general synoptic state there appears to be multi-year periods of preferential modes, with particularly low frequencies in the early 1970s and mid 1980s, and peak occurrences in 1968, 1988, and 1994. A Fourier analysis of the same frequency time series reveals a strong spectral peak in the 6 to 8 d period. Fig. 8 shows the spectral density plot of the Fourier analysis, along with the histogram of the periodiogram with the white noise curve overlain. The peak in this case is likely representative of the normal mid-latitude cyclic nature of the synoptic systems.

The frequency of occurrence of synoptic events can also be analyzed with respect to other environmental parameters. Figs. 9 & 10 show the frequency distribution for a dry and wet January as measured by the station precipitation at State College, Pennsylvania, in the center of the domain. Not surprisingly the wet year dis-

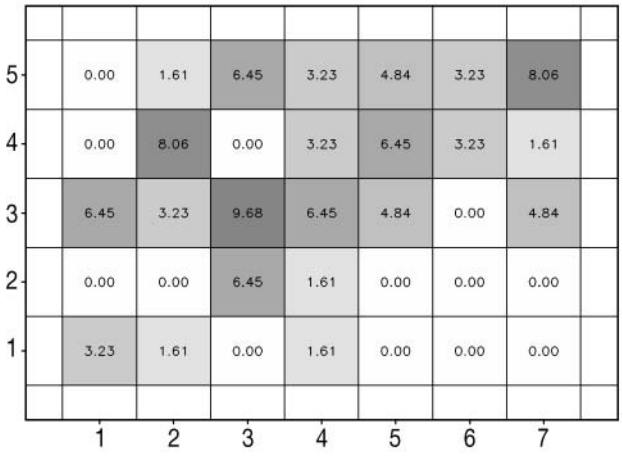

Fig. 10. Number of days mapping to each SOM node for January 1981 (high-precipitation month)

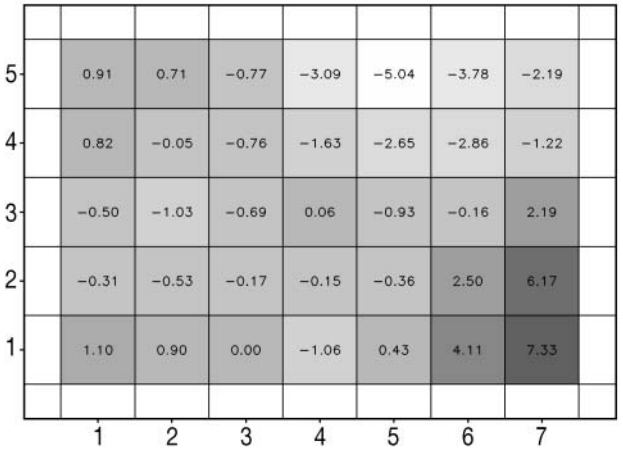

Fig. 11. Trend in the percentage change of number of days of occurrence in a month over 40 yr

plays a significantly higher frequency of low-pressure systems than does the dry year, and actually has zero strong high-pressure events of the type analyzed earlier.

Having considered the inter-annual variability and frequency of occurrence in individual years, a logical extension would be to examine the 40 yr long-term trend in the frequency of synoptic systems. For demonstration purposes, this is calculated by determining the trend in frequency at each node over the 40 yr. However, as adjacent nodes may be very similar to each other, it is possible that very similar synoptic states could map to either node. With only 40 samples in the data set (1958–1997), marginal differences in intensity of related synoptic patterns (adjacent nodes) from year to year could potentially have a strong impact on the fitted trend line. Determining the significance of the trends would require a more sophisticated analysis combining frequencies from related nodes to produce the trend line. Several approaches to grouping nodes are possible. For example, one could re-run the SOM with fewer nodes to increase the level of generalization—producing greater differences between individual nodes. Within the larger SOM, nodes could be grouped according to a subjective grouping of adjacent and similar patterns, or by using some measure of inter-node distance (as in Fig. 3). Alternatively, a PCA of the node vectors could produce a linear combination of related nodes—similar to the approach adopted by Jones & Kelly (1982) in an analysis of the Lamb catalogue of daily weather maps of the British Isles. Here we present the trends at each node and simply note the cohesiveness of the trend pattern across the nodes as an indication of the importance of the trend.

Fig. 11 displays the trends at each node as the percentage change in frequency over 40 yr. The figure shows an apparent coherent pattern of trend focused on the synoptic states mapping to the right-hand side of the SOM array (see Fig. 2 for reference). In particular, there is a positive trend in the frequency of occurrence for days dominated by strong high-pressure systems, compensated by a decrease in days with a moderate continental high and low pressure to the north-east.

Finally, in terms of frequency analyses, the synoptic event frequencies can be related to other non-local atmospheric processes, in particular indices of identified teleconnection features. By way of example, the North Atlantic Oscillation is correlated to the frequency of occurrence on each node and shown in Fig. 12. The correlations on individual nodes are not strong and, for the same reason as above, the significance levels are not calculated. However, again the patterns across the nodes display a strong cohesiveness that lends some strength to consideration that the correlations are real.

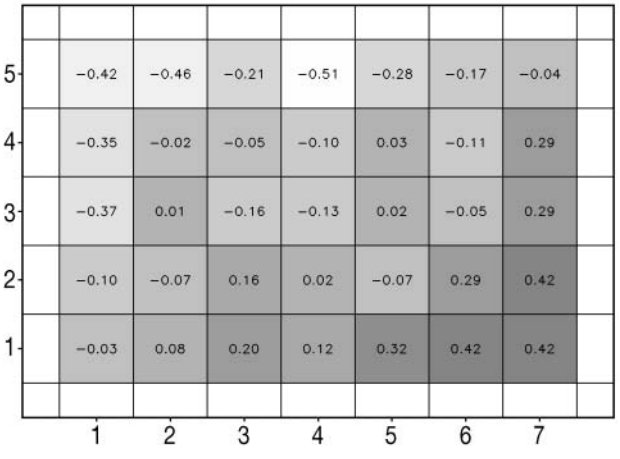

Fig. 12. Correlation of monthly frequencies at each node with the North Atlantic Oscillation from 1958–1997

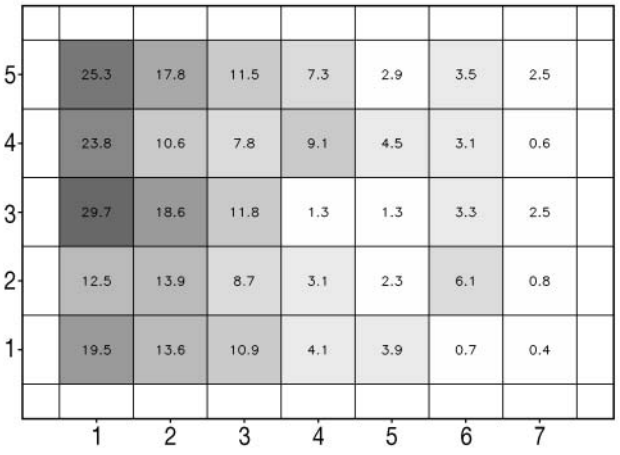

Fig. 13. Average precipitation for days mapping to each SOM node

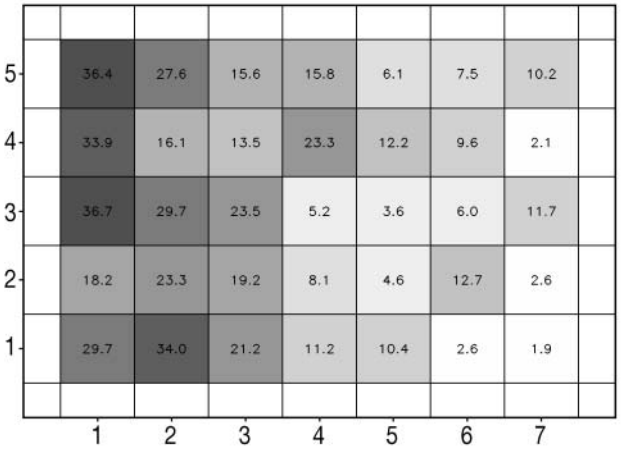

Fig. 14. Standard deviation of precipitation (mm) of days mapped to each SOM node

5. RELATION TO SURFACE CLIMATE: STATION PRECIPITATION

The next phase in any synoptic climatology is to relate the atmospheric circulation to some dependent variable, usually a local-scale climate or environmental parameter. This example uses daily station precipitation from State College. For the first step the mean and variance of the rainfall related to each synoptic state is determined. For a given SOM node this is accomplished by determining the rainfall for all days where the circulation maps to the given node.

Fig. 13 shows the average precipitation per node, while Fig. 14 shows the standard deviation of the precipitation on each node. Not surprisingly, the distribution closely follows the synoptic states representing large low-pressure systems. The precipitation is strongest under synoptic states mapping to the upper-left quadrant of the SOM space, extending to more moderate precipitation values in the lower-left quadrant. The right-hand sector of the SOM space demonstrates minimal precipitation.

The standard deviation associated with each node closely matches the distribution of precipitation, with a secondary peak in the more transitional synoptic states in the upper-central region of the SOM space. In winter the recording station region is subject primarily to frontal precipitation, with secondary contributions arising from its location at the margin of the lake effect precipitation region under northwest flow regimes. As such the higher standard deviation for the nodes in the top central portion of the SOM space is physically consistent with the northwest flow regimes represented by these synoptic states.

Taking these relationships into account some interesting physical relationships may be inferred in the context of the frequency trends shown earlier (Fig. 11). The trends indicate a rise in the frequency of occurrence for nodes in the lower-right quadrant of the SOM space- nodes associated with strong high-pressure systems dominating the domain. The rise in frequencies in this region of the SOM space are compensated largely by a decrease in frequency in the nodes associated with north-westerly flow over the region—those synoptic states associated with lake effect precipitation. The inference here is that the synoptic states that reduce precipitation are increasing, while synoptic states that enhance precipitation have been decreasing over time. Because of this, a decreasing trend in station precipitation over the 40 yr could be expected; however, the station record (Fig. 15) shows a positive trend in monthly mean precipitation; hence the changes implied above must be compensated for elsewhere.

The precipitation trends inferred from changes in circulation must also be interpreted in the light of trends in average precipitation for a given synoptic state. To investigate this, the trend in precipitation matching synoptic events mapped to each node is calculated and shown in Fig. 16 as a percentage of the mean monthly average. Fig. 16 shows that days that

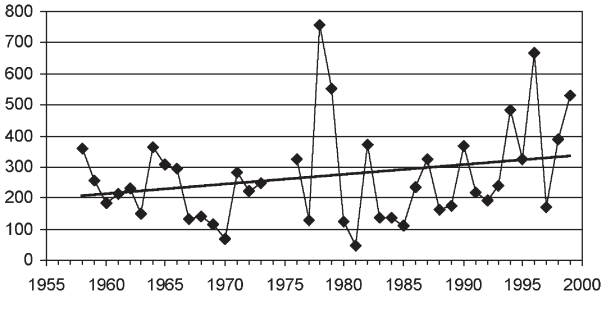

Fig. 15. Monthly mean precipitation at State College, Pennsylvania, with a linear trend line

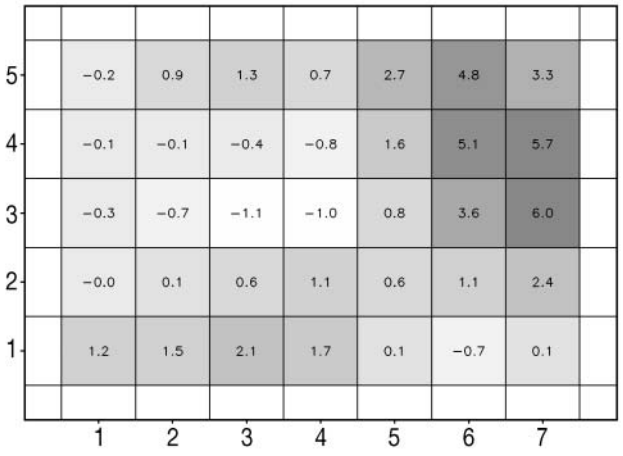

Fig. 16. Percentage change per year in precipitation from synoptic systems on each SOM node

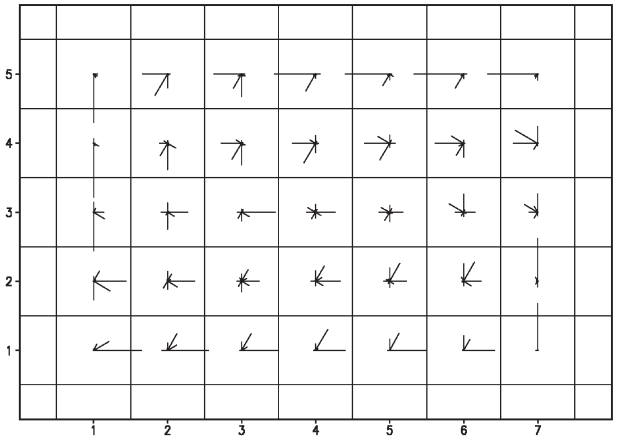

Fig. 17. Proportional frequencies of back trajectory directions in SOM space (each node standardized by total frequency of days on the node)

map to nodes 4,6, 4,7 and 3,7, with a high-pressure system to the north and west of Pennsylvania, have increasing precipitation over the 40 yr record. The same synoptic circulation results in more precipitation at present than it did 40 yr ago. This indicates that the changes in precipitation at the station are not simply related to changes in the frequency of occurrence of particular synoptic patterns, but also to changes in the precipitation conditions and processes occurring within a circulation type. While it is tempting to infer a climate change signal (for example, increased boundary layer moisture from global warming), this would require further analysis beyond the intent of methodological examples in this paper.

6. EXTENSION TO TEMPORAL TRAJECTORIES

As adjacent nodes in the SOM are related to each other (because similar days map to adjacent or nearby nodes), the SOM array can also be used to examine the temporal evolution of synoptic events. This is accomplished by tracking the trajectory in time across the array of nodes. The assumption underlying this approach is that the incremental change in synoptic state from one time step to another is small enough that the sequential movement across the SOM space may be tracked. Thus, for each node, the frequency of transitions from one node to nearby nodes (either forward or backward in time) is calculated.

In this analysis, the forward and backward trajectories across SOM space are determined in terms of the number of times a change occurs from one node to another in each of 12 directions, or sectors, toward or away from the original node. Figs. 17 & 18 are equivalent to transition matrices and show the preferential trajectories forward and backward in SOM space. For each node the length of the line in each sector displays the relative frequency of transition in that direction away from or toward the node. The frequencies are normalized by the total frequency of occurrences on the node in order to facilitate comparisons between nodes. Reference to Fig. 6 shows the absolute total frequency of occurrence from one node to another.

Fig. 19 displays the percentage of time the circulation is stationary at each node—when the synoptic system maps to the same node on consecutive observations—and should be used in conjunction with Figs. 17 & 18 for interpreting synoptic sequences. The most obvious aspect of the trajectories is the cyclic nature of the weather systems, notably that there is a preferential clockwise evolution with time (in SOM space). Comparing these figures with Fig. 2 shows that this behavior represents a sequence of transitory high- and low-pressure systems across the region. Beginning at

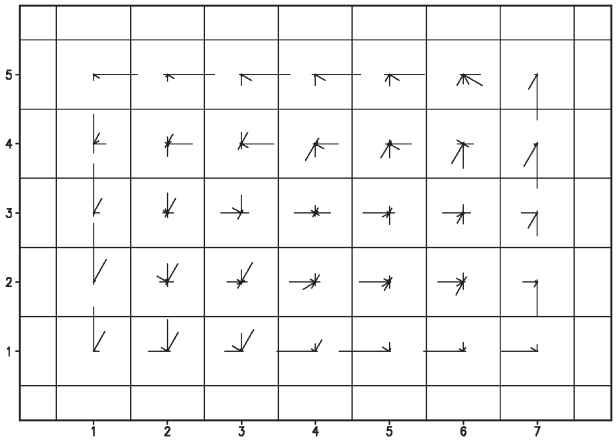

Fig. 18. Proportional frequencies of forward trajectory directions in SOM space (each node standardized by total frequency of days on the node)

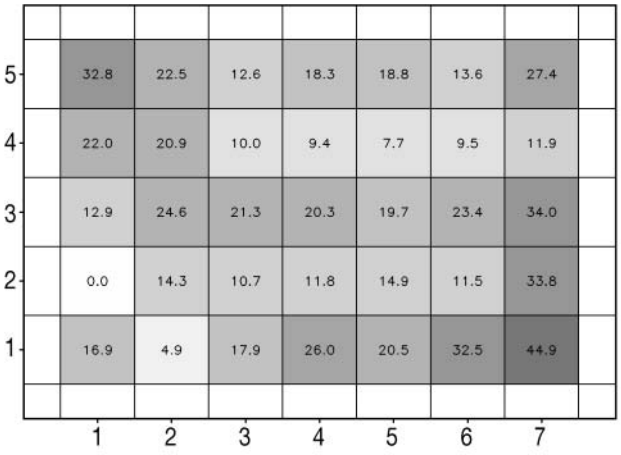

Fig. 19. Frequencies (%) of days on each node where no transition is made to another node in the next time step

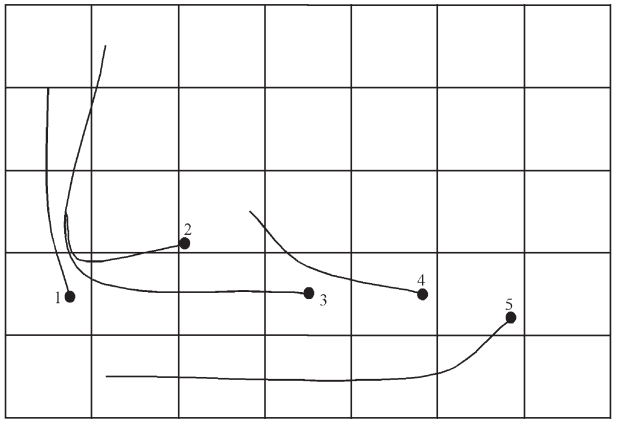

Fig. 20. Back trajectories of the mean of each cluster in SOM space over 36 h for rain events in the top 0.15%. Solid circle indicates starting position. Clusters determined with Ward's clustering

the bottom-right corner of the array, we see a central high-pressure system gradually being replaced by a low-pressure system tracking in from the north-west, moving across the northern part of the region, and being replaced by high pressure moving in behind it.

Secondly, it can be seen that the transitions to or from synoptic states in the center of the SOM space are more variable and have less clearly defined trajectory pathways than the peripheral nodes. Furthermore, while some nodes display a single mode of preference others, such as 1,4, 2,4, 2,5 and 2,6 show systems arriving from several directions, although all broadly from the same quadrant. The trajectory analysis of the mean evolution pathways offers a good overview of the preferential development of weather systems over the region; however, it is also possible to focus on particular events in order to understand special cases, such as extreme precipitation events. For this, the top 1.5% of rainfall days have been extracted, and the back trajectory coordinates over 36 h determined. The trajectories as defined by their node *x,y* coordinates are then grouped with simple cluster analysis (using Ward's algorithm) to define 5 groups with an equal number of trajectories in each. The mean trajectory for each of these groupings is then calculated and plotted as a representation of the typical evolution pathways that may lead to heavy precipitation events. This approach was adopted simply to illustrate some characteristic trajectories. An alternative would be to examine trajectories individually, or to examine exceptional trajectories that are significantly different from those in these groups.

Fig. 20 shows these pathways. Two of the 5 mean trajectories (1 and 2) are very similar when interpreted in the light of the synoptic states (see Fig. 2). These 2 terminate in a synoptic state with a large low-pressure system positioned to the northeast of the station, and both are relatively slow moving, as evidenced by the short trajectory paths. Trajectory 3 is a fast moving synoptic system, terminating in a similar state to 1 and 2, but is a more intense version consistent with the speed of transition. Trajectory 4 is again relatively slow moving and associated with the central nodes on the SOM space—largely weaker pressure gradients and terminating in a state that favors flow into the region from the south. Trajectory number 5 is quite dissimilar to the others. It shows the fastest development of all (traversing ~6 nodes in 36 h), starting from a strong dominant high-pressure system that transitions to a strong low-pressure system arriving from the northwestern quadrant over Canada. Each trajectory evolves into a precipitation event in the top 1.5% of events, and highlights the distinctly different pathways that may give rise to similar magnitude precipitation events.

7. CONCLUSIONS

Synoptic climatology is a sub-discipline with a rich heritage in the methodology of typing weather systems. Such an approach, while valuable in that it allows broad generalizations of cross-scale relationships, inherently obscures much of the detail through the degree of generalization. The opposite methodological extreme of transfer function downscaling circumvents this, but may reduce interpretability. Where the precipitation is described as some linear or non-linear function of local circulation and humidity parameters, it may be harder to attribute the precipitation characteristics to particular atmospheric processes.

An SOM-based approach, however, allows rapid and powerful generalization of weather systems into an easily visualized array of synoptic states spanning the continuum of events. The method allows the user to define the degree of generalization required (by defining the dimension of the SOM), without losing the ability to visualize the results. In addition, the SOM facilitates the investigation of the temporal aspects of the synoptic systems, from long-term frequencies of events through to the temporal evolution of individual weather systems.

This paper presents a range of analytical approaches that can be accomplished simply by applying an SOM to SLP data. Such an analysis is easily extended to multivariate circulation data (for example, coupling SLP with 500 hPa geopotential heights). Once the SOM has been developed, the relationship of the SOM modes to other environmental parameters (such as precipitation) is easily determined. It would also be possible to include other climate or environmental data in the input data set that derives the SOM mapping, producing a SOM that is analogous to the airmass approach to synoptic climatology (e.g. Kalkstein & Corrigan 1986).

The simplicity of evaluating the frequency characteristics of daily resolution synoptic events offers one particularly powerful application for the evaluation of climate model performance and the investigation of the underlying circulation changes projected under global warming. At present much of the validation of GCMs is undertaken with monthly or seasonal mean fields. An SOM, however, provides a means for evaluating the daily fields that make up the more commonly used monthly means, and for investigating the potential changes of regional circulation. Given the present difficulty of developing regional climate change scenarios, and the urgent need for such scenarios (see, for example, the recommendation from the Intergovernmental Panel on Climate Change to 'improve the integrated hierarchy of global and regional climate models with a focus on the simulation of climate variability, regional climate changes, and extreme events.' Houghton et al. 2001), an SOM analysis presents a valuable tool for the climate downscaling community.

Acknowledgements. We wish to thank the 3 anonymous reviewers for their helpful comments on an earlier draft of the manuscript. This work was supported in part by a grant to the Pennsylvania State University from the US National Science Foundation Program on Human Dimensions of Global Change (SBR-9521952), Center for Integrated Regional Assessment; the Penn State Office of International Programs; the US Environmental Protection Agency Cooperative Agreement No. CR 826554; and the South Africa Water Research Commission Project K5/1012.

LITERATURE CITED

Abercromby R (1883) On certain types of British weather. Q J R Meteorol Soc 9:1–25

Abercromby R (1887) Weather: a popular exposition of the nature of weather changes from day to day. Kegan Paul, London

Ambroise C, Seze G, Badran F, Thiria S (2000) Hierarchical clustering of self-organizing maps for cloud classification. Neurocomputing 40

Barry RG, Perry AH (1973) Synoptic climatology: methods and applications. Methuen & Co Ltd, London

Barry RG, Elliott DL, Crane RG (1981) The palaeoclimatic interpretation of exotic pollen peaks in Holocene records from the eastern Canadian Arctic: a discussion. Rev Palaeobot Palynol 33:153–167

Bellone E, Hughes JP, Guttrop P (2000) A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. Clim Res 15:1–12

Cavazos T (1999) Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas. J Clim 12:1506–1523

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

Chen L, Gasteiger J (1997) Knowledge discovery in reaction databases: landscaping organic reactions by a self-organizing map. J Am Chem Soc 119:4033–4042

Comrie AC (1992) A procedure for removing the synoptic climate signal from environmental data. Int J Climatol 12: 177–183

Crane RG (1978) Seasonal variations of sea ice extent in the Davis Strait-Labrador Sea area and its relationships with synoptic-scale atmospheric circulation. Arctic 31:437–447

Crane RG, Hewitson BC (1998) Doubled CO_2 precipitation changes for the Susquehanna Basin: Downscaling from the GENESIS general circulation model. Int J Climatol 18:65–76

Hewitson B (1999) Deriving regional precipitation scenarios from general circulation models. Water Research Commission 751/1/99, Pretoria

Hewitson B (2001) Global and regional climate modelling: application to Southern Africa. Water Research Commission 806/1/01, Pretoria

Hewitson BC, Crane RG (1992a) Regional-scale climate prediction from the GISS GCM. Global Planetary Change 97: 249–267

Hewitson BC, Crane RG (1992b) Large-scale atmospheric controls on local precipitation in tropical Mexico. Geophys Res Lett 19(18):1835–1838

Hewitson BC, Crane RG (1994) Neural computing: applications in geography. Kluwer Academic Publishers, Dordrecht

Hewitson BC, Crane RG (1996) Climate downscaling techniques and applications. Clim Res 7:85–95

Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linded PJ, Dai X, Maskell K, Johnson CA (eds) (2001) Climate change 2001: the scientific basis. Cambridge University Press, Cambridge

Hudson DA (1998) Antarctic Sea ice extent, southern hemisphere circulation and South African rainfall. PhD thesis, University of Cape Town

Jones PD, Kelly PM (1982) Principal component analysis of the Lamb catalogue of daily weather types: Part 1, Annual frequencies. J Climatol 2:147–157

Joutsiniemi SL, Kaski S, Larsen TA (1995) Self-organizing

map in recognition of topographic patterns of EEG spectra. IEEE Trans Biomed Eng 42:1062–1068
Kalkstein LS, Corrigan P (1986) A synoptic climatological approach for geographical analysis: assessment of sulfur dioxide concentrations. Ann Assoc Am Geogr 76:381–395
Kalkstein LS, Tan G, Skindlov JA (1987) An evaluation of three clustering procedures for use in synoptic climatological classification. J Clim Appl Meteorol 26:717–730
Kalnay E and 21 others (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77(3):437–471
Kohonen T (1989) Self-organization and associative memory, 3rd edn. Springer-Verlag, Berlin
Kohonen T (1990) The self-organizing map. Proc IEEE 78(9): 1464–1480
Kohonen T (1991) Self-organizing maps: optimization approaches. In: Proceedings of the International Conference on Artificial Neural Networks, Espoo, Finland, June 1991, p 981–990
Kohonen T (1995) Self-organizing maps. Springer-Verlag, Heidelberg
Köppen W (1874) Über die Abhängigkeit des klimatischen Charakters der Winde von ihrem Ursprunge. Rep. Met. (St. Petersburg), 4(4); cited in Barry RG, Perry AH (1973) Synoptic climatology: methods and applications. Methuen & Co Ltd, London
Lamb HH (1950) Types and spells of weather around the year in the British Isles: annual trends, seasonal structure of the year, singularities. Q J R Meteorol Soc 76: 393–429
Main JPL (1997) Seasonality of circulation in Southern Africa using the Kohonen self organising map. MSc thesis, University of Cape Town
Malmgren BA, Winter A (1999) Climate zonation in Puerto Rico based on principal components analysis and an artificial neural network. J Clim 12:977–985
Palakal MJ, Murthy U, Chittajallu SK, Wong D (1995) Tonotopic representation of auditory responses using self-organizing maps. Math Comput Model 22:7–21
Sailor DJ, Li X (1999) A semiempirical downscaling approach for predicting regional temperature impacts associated with climate change. J Clim 12:103–114
Sammon JW Jr (1969) A nonlinear mapping for data structure analysis. IEEE Trans Comput C-18(5):401–409
Wigley TML, Jones PD (1987) England and Wales precipitation: a discussion of recent changes in variability and an update to 1985. Int J Climatol 7:231–246
Yarnal B (1984) Relationships between synoptic-scale atmospheric circulation and glacier mass balance in southwestern Canada during the International Hydrological Decade, 1965–74. J Glaciol 30:188–198

Editorial responsibility: Brent Yarnal, University Park, Pennsylvania, USA

Submitted: June 14, 2000; Accepted: December 17, 2001
Proofs received from author(s): July 19, 2002