Nonlinear climatology and paleoclimatology: Capturing patterns of variability and change with Self-Organizing Maps

David B. Reusch *

EMS Earth and Environmental Systems Institute, The Pennsylvania State University, University Park, PA 16802, USA

A R T I C L E   I N F O

Article history:
Received 26 February 2009
Received in revised form 30 July 2009
Accepted 4 September 2009
Available online 18 September 2009

Keywords:
Antarctica
Atmospheric circulation
Self-Organizing Maps
Nonlinear
Polar climate
Meteorology

A B S T R A C T

Self-Organizing Maps (SOMs) provide a powerful, nonlinear technique to optimally summarize complex geophysical datasets using a user-selected number of “icons” or SOM states, allowing rapid identification of preferred patterns, predictability of transitions, rates of transitions, and hysteresis/asymmetry in cycles. For example, SOM-based patterns concisely capture the spatial and temporal variability in atmospheric circulation datasets. SOMs and the SOM-based methodology are reviewed here at a practical level so as to encourage more-widespread usage of this powerful technique in the atmospheric sciences. Usage is introduced with a simple Antarctic ice core-based dataset to show analysis of multiple variables at a single point in space. Subsequent examples utilize a 24-year dataset (1979–2002) of daily Antarctic meteorological variables and demonstrate usage with 2-D, gridded data, for single and multiple variable scenarios. These examples readily show how SOMs capture nonlinearity in time series data, concisely summarize voluminous spatial data, and help us understand spatial and temporal change through, for example, pattern frequency analysis and identification of preferred pattern transitions.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Many techniques are useful for extracting patterns from a large geophysical dataset, such as the examples of daily-mean anomalies of the atmospheric circulation data considered here or records of monthly Antarctic sea-ice extent. It may prove useful and informative, for example, to note the strength of the loading of the data from a particular month on the first two principal components of the dataset. It may also be instructive to note that the data from a given month are very similar to those of some earlier, well-known month (e.g., “this looks like the ice that trapped Endurance in the Weddell Sea ice pack in January, 1915”). In this hypothetical example, “January, 1915” serves as an icon, to which other data fields can be compared. Such icon-based classification schemes can work in parallel with pattern-extraction tools such as Principal Components Analysis (PCA). However, a particular data field (such as “January, 1915”) may contain unique features that reduce its utility as a more general icon. To overcome this difficulty (i.e., over specificity of the results), the technique of Self-Organizing Map (SOM) analysis (Kohonen, 2001) provides an objective way to optimally extract a user-specified number of icons, or SOM states, from an input dataset. Here “optimal” indicates that the patterns capture dataset variability in a generalized way that accurately summarizes key features without being overly drawn to specific details.

Self-Organizing Maps (Kohonen, 2001) are an analysis tool from the field of artificial neural networks that uses so-called unsupervised training to find salient features (icons) of an input dataset without prior specification (or knowledge) of the “correct” output. The features (or patterns) found by the network are often of direct interest but they are also frequently used to divide the input data into distinct classes (i.e., classification) that further help us to understand the data. In short, SOMs support unsupervised classification of large, multivariate geophysical datasets through creation of a spatially organized set of generalized patterns of variability which, collectively, represents the probability density function (PDF) of the input data (e.g., Hewitson and Crane, 2002; Reusch et al., 2005a,b). That is, a SOM analysis produces a succinct, discrete, nonlinear (Kohonen, 2001) classification/summary of complex, continuum input data.

The extracted patterns of a SOM analysis are returned in a grid or “map”, with similar patterns placed near each other and the most-extreme patterns at the corners, i.e., there is a spatial organization and order in the pattern set. Often, the patterns at the ends of one diagonal are similar to the positive and negative phases of the first principal component of the input data, with the second principal component correspondingly at the ends of the other diagonal; however, this is not required. For example, in the ice-core data analysis example below (Fig. 2), high recent values and low early-record values are placed at opposite ends of one diagonal.
of the map, while important asymmetries in the behavior of specific variables are found on the other diagonal. In neither case, however, are the end-diagonal patterns mirror images, as would be found in a strictly linear analysis.

The SOM-based approach brings a number of methodological benefits as well. Foremost is that SOMs offer an alternative to more traditional linear techniques, such as principal component analysis (PCA), that is more robust (e.g., able to interpolate into areas of the input space not present in the available training input), less complex, and less subjective while also accommodating nonlinear relationships in the data (Kohonen, 2001; Reusch et al., 2005a,b). SOMs also provide a completely independent uniformitarian or multivariate analysis pathway and thus provide independent results for comparison with more traditional techniques. SOM-based analysis thus complements linear techniques without replacing them. SOMs also provide a powerful visualization approach for studying structure in large, complex datasets. In the case of atmospheric circulation data, for example, the patterns capture the full range of synoptic conditions while also treating the data as a continuum. Because each input data record maps uniquely to one SOM pattern, SOM analysis easily allows characterization of, for example, time-trends in frequency of occurrence, preferred transitions that may point toward predictability, and cyclicity of preferred patterns. These aspects and others will be detailed below.

Readers interested in going beyond the examples presented here have a growing number of examples to choose from of SOM usage in meteorology, climatology and oceanography. Recent work with a regional/synoptic focus covers the polar regions as well as the temperate latitudes of both hemispheres. In the polar regions, examples include characterizing extreme events at Barrow, Alaska in a synoptic context (Cassano et al., 2006a,b), the identification and characterization of a low-level jet on the Ross Ice Shelf, Antarctica (Seefeldt and Cassano, 2008), and the atmospheric circulation of West Antarctica (Reusch et al., 2005a,b). In the North Atlantic, synoptic forcing of precipitation in Greenland (Schaunemann et al., 2009) and the spatial/temporal characteristics of the North Atlantic Oscillation (Reusch et al., 2007; Johnson et al., 2008) have also been examined with SOMs. Outside the polar/subpolar regions, SOMs have helped studies of precipitation in the Iberian Peninsula (Morata et al., 2006), Australian drought and its relationship to a variety of tropical forcings (Verdon-Kidd and Kiem, 2009), and the assessment of seasonal forecast model skill in Peru (Gutierrez et al., 2005). Lastly, SOM usage in oceanography is also expanding (e.g., Liu and Weisberg, 2005; Liu et al., 2009).

With respect to model forecasts, SOMs have been particularly useful in evaluating the growing number of GCM results. A key to trusting GCM projections of the future is whether the models can realistically reproduce the observed past. For example, Leloup et al. (2008) evaluated GCM skill for the 20th century by using SOM-based analyses of sea surface temperatures in the tropical Pacific to study ENSO characteristics in a suite of GCMs. Similar studies that also include evaluation of future projections (focusing on sea-level pressure and precipitation) have been done for the Arctic (Cassano et al., 2006a,b), the Antarctic (Lynch et al., 2006; Uotila et al., 2007) and southern Africa (Hewitson and Crane, 2006).

Section “Methodology” provides details on the SOM methodology and a very simple example. The balance of the paper discusses a series of examples with increasing data complexity to highlight the key features of the SOM-based approach. Section “A realistic example: multiple variables at a single point” introduces a real example using multiple variables at a single location. Remaining sections cover spatial examples with one (Sections “A two-dimensional example: atmospheric model fields” and “Further interpretation, and Temporal change and transitions and trajectories”) and multiple (Section “Multivariable, multidimensional SOMs”) variables in the analysis.

2. Methodology

The successful application of SOMs revolves around two key concepts: the architecture and topology of the SOM grid, and the parameter selection and training process that yields the SOM patterns. All aspects of SOM usage, such as the frequency maps and other analyses described in later sections, build on these ideas, thus, it is important to understand these foundations. Note that the freely available SOM-PAK software (Kohonen, 2001) has been used here as in our previous work (e.g., Reusch et al., 2005a,b).

2.1. Architecture and topology

Architecturally, a SOM is composed of a finite set of nodes with a grid topology such that each node has either four (rectangular) or six (hexagonal) neighbors (fewer for those nodes located on a grid edge or corner). A particular SOM instance is usually referred to by its grid dimensions. For example, a 4 × 3 SOM has 12 nodes. The size of the grid (number of patterns) directly influences the amount of generalization: smaller (larger) node arrays have fewer (more) available patterns to characterize the n-dimensional data space, so the final patterns developed during training will tend to do more (less) generalization of the input. Grid size is thus a first-order experimental parameter. Because the pattern set is relatively small (and finite), complexity is reduced to working with the set of patterns instead of the (usually much larger) original dataset. Individual SOM patterns are typically identified either by an (x,y) coordinate pair or a sequence number within the two-dimensional array (counting left-to-right, top-to-bottom). Coordinate pairs identify patterns consistently across different grid sizes while sequence numbers have notational simplicity (and will be used here).

Each SOM node has an associated reference vector representing the node's generalized pattern. Reference vectors have the same dimensionality as the original data samples once those samples have been converted to vector form (i.e., if the samples being analyzed are not one-dimensional, they must be converted to vector form before processing). After training (see below), the generalized patterns are a projection of the multidimensional input data onto the two-dimensional array of reference vectors that preserves the topological structure of the original data.

Because relative distances between nodes (reference vectors) in data space are not fixed (as the term “grid” may imply) but vary as a function of the information content and distribution of the raw data, Sammon maps (Sammon, 1969) are often used to visualize these relative distances. Sammon mapping, as used here, projects the multidimensional reference vectors for each generalized SOM pattern into a 2-D space. This aids visualization of inter-pattern relationships by being able to plot the SOM nodes based on relative neighbor-to-neighbor similarity, rather than using the simpler, but less informative, regularly spaced grid format normally used. However, the regularly spaced grid format is quite adequate for the purposes of displaying relative frequencies and related attributes of each node.

2.2. Parameter selection and training

As with many higher order data analysis tools, SOM usage involves an iterative process of selecting model parameters (e.g., grid size), then using the model with these parameters (i.e., training to create the generalized patterns, a step that is itself iterative). The main parameters typically available (e.g., with the software used here, SOM-PAK (Kohonen, 2001)) are grid size (recall the link to level of generalization), number of training iterations, starting neighborhood size and learning rate (how much each reference vector is changed after matching with data records). Grid size and iterations…
are fixed for a given SOM while the neighborhood size and learning rate are gradually reduced from the starting value as the iterations progress. It is often useful to explore this parameter space at least along the grid-size dimension, to get a feel for how the generalization level changes. Also important to test is the number of iterations. Adjusting this number from low to high values will normally lead to a situation in which the patterns no longer change significantly as the reference vectors converge on a stable solution. Iterations beyond this threshold will just be wasting time and resources. Note that, as with other analysis techniques, there are scenarios where a SOM can produce invalid or unusable results, e.g., converging to patterns that exactly match the data. However, this is rare and typically requires either very simple input datasets (where too many iterations will cause problems) or datasets with too few training examples for the size of the data samples. For example, based on tests with real-world data, a sample with 65,000 grid points will need more than 100 training records to produce a successful SOM. Apart from these extremes, SOMs are quite robust.

Multiple runs with varied parameters will naturally result in multiple sets of SOM patterns for any given dataset and the consequent need to evaluate a mix of objective and subjective metrics before deciding which set is best (or most useful) for your analysis before proceeding. Examples of objective metrics may be found in Kohonen (2001), Kiviluoto (1996), Uriarte and Martin (2006), and references therein. An example of a subjective metric is to look at the Sammon maps to assess the shape of the grid topology. A SOM that needs more training may have a twisted or folded grid as a result of nodes not yet being placed correctly in the data space.

Once a grid size and other parameters have been chosen, the next major step in SOM usage is the creation of the generalized patterns from the input data, an iterative process known as training. A key step in SOM training, and in later analyses using the generalized patterns, is the mapping of input data samples to the closest matching reference vector (usually based on the Euclidean distance between the input and the reference vectors). During training, this step identifies which reference vectors (and neighbors) are to be updated. After training, mapping is used to classify the input data (since input records with common patterns will map to the same or nearby SOM nodes) and is widely used to study, for example, frequency-of-occurrence characteristics of different subsets of the data (based on time or other criteria, such as state of the Southern Oscillation Index (SOI)), also known as frequency mapping.

Details on SOM training are readily available in the literature (e.g., Hewitson and Crane, 2002; Reusch et al., 2005a,b) although it can be difficult to locate “best practices”. As such, it remains useful to review multiple sources for their perspectives on SOM usage. Briefly, training is an iterative and unsupervised process in which the reference vectors representing each SOM node (or pattern) are first initialized, then adjusted repeatedly based on distances (i.e., differences) between best-matching reference vectors and each input record. Training ends (in SOM-PAK) once the user-specified number of iterations through the dataset have been completed.

Reference vectors are typically initialized in one of two ways: randomly or by using the first two eigenvectors. Random initialization distributes the reference vectors evenly across the input data space, but without any spatial organization. Eigenvectors provide an initial spatial organization in a strictly linear sense (diagonal corner patterns will be mirror opposites and the Sammon map will be rectangular) which speeds up subsequent training since the SOM is starting with some organization already defined. However, eigenvector calculation can be memory intensive and thus this approach is limited to data samples smaller than a system-dependent size. Random initialization has the advantages of being fast and usable regardless of the data sample size, but subsequent training takes more processing time.

SOM generalization arises from the use of only a fraction of the difference in the iterative reference vector adjustments, as determined by the learning rate. Because of this, reference vectors never exactly match any of the data samples. As training progresses, the reference vector updates converge to values such that each node represents a distinct portion, or sub-space, of the multidimensional input space. (Unlike summation of PCA components, the sum of these reference vectors does not reconstruct the original input data,) Because neighbors of the best match are also updated during training (but to a lesser degree), adjacent nodes have the strongest similarity, with similarity decreasing with increasing distance away from any given node. For mathematical stability during the learning process, the SOM grid is normally asymmetric rather than being square (e.g., $5 \times 4$ or $5 \times 3$, not $5 \times 5$ or $4 \times 4$) (Kohonen, 2001). Such asymmetry in the pattern space helps the SOM capture the primary axes of variance in the data by separating variance axes in SOM space.

2.3. Data considerations

Theory and experience indicate that SOMs are applicable to a wide variety of datasets, but there are still potential pitfalls with respect to real-world usage that need to be noted. Of particular importance when working with datasets containing multiple variables is that all variables need to share a similar scaling. This is typically done through data standardization using each variable’s mean and standard deviation. For example, if surface temperature and geopotential heights are being jointly analyzed (see Section “Multivariable, multidimensional SOMs”), unstandardized temperatures ($^\circ C$) might be in the range $–50$ to $–10$ and mid-tropospheric geopotential heights ($m$) in the range of 5000–6000. The SOM will attempt to span this range of variability and, more than likely, will produce results that do not represent either variable very well. With standardized variables, the SOM will be working with a single, overlapping range of variability and provide much better results.

As noted above, too few training records may lead to invalid SOM results, although determining this threshold in advance is highly qualitative. Intuitively, the number of training records likely scales with the number of grid points per sample for spatial data. While this aspect of SOM usage is difficult to quantify, Sammon maps, quantization error, RMS differences and other techniques may be used to assess SOM skill at least semi-quantitatively.

One area in which SOM usage typically differs from a number of other analysis techniques is that the full input dataset is usually used to “train” the SOM, i.e., to create the generalized patterns. In other techniques, particularly other methods from the artificial neural network field (e.g., Reusch and Alley, 2002), reserving input data for testing (and validations) is essential to getting valid analysis results. Other methods, such as jackknifing and bootstrapping, reserve a portion of the input data to develop confidence metrics (e.g., Efron and Tibshirani, 1993). In contrast, a primary goal of SOM analyses is to use the generalized patterns developed from the full dataset to classify/summarize all of the data. Reserving data is thus potentially counterproductive. Whether techniques such as bootstrapping would be applicable and useful in SOM analyses, e.g., for the development of confidence metrics, is beyond the scope of this paper.

2.4. A very simple analysis

To provide a concrete example of data classification, a very simple (and unrealistic, but useful for certain concepts) dataset was created and analyzed with a $3 \times 2$ SOM (Fig. 1). The four patterns
in Fig. 1a were each replicated 100 times to create an input dataset with 400 samples. Each sample then had a small amount of uniform noise added to make the dataset slightly more challenging for the SOM algorithm. The beginning of this input dataset (the first 10 samples) is shown in Fig. 1b. (Note that this particular dataset is not well-suited to using the SOM for pattern extraction due to this extreme simplicity.) After random initialization (Fig. 1c), the SOM patterns (reference vectors) are not quite identical and show features of all the available input patterns (i.e., the input space). In Fig. 1d, the SOM has been run for 500 iterations, i.e., long enough to show well-organized structure. These patterns can reliably classify all of the input data because data classification relies on the best match (rather than exact matches). For example, a group characterized by the first input pattern will be identified by matches with SOM pattern 3; input samples 4 and 9 will be in this group (the number of the best matching SOM pattern in Fig. 1d is shown below each sample in Fig. 1b). The Sammon map (Fig. 1e) readily shows how different features of the patterns are related, e.g., those patterns with distinct smiles (three, five and six), are closer to each other than the rest of the patterns. Note that the $3 \times 2$ SOM does not fully discriminate between the four idealized patterns: input samples 1 and 5 both match best with SOM pattern 1. This is pri-
marily a reflection of the simplicity of the data rather than a flaw in the technique.

For this extremely simple dataset, too much training or too many nodes may lead to an “over trained” SOM where the SOM
patterns converge to exact patterns from the data and/or SOM nodes become duplicated. An example of this is shown in Fig. 1f. With more nodes available than are needed, the SOM has combined nodes 1 and 2, 3 and 4, and 9 and 10. The Sammon map for this SOM (Fig. 1g) shows this even more clearly. The utility of such a degenerate pattern set depends on the goals of the analysis. In this case, information in the Sammon map allows us to combine patterns and produce a useful classification of the input data. Input samples 1 and 5 now map to unique SOM patterns (1 and 9, respectively) and the input patterns with smiles map to the other grid corners. SOM patterns 5–8 and 11 represent intermediate patterns not present in the data.

3. A realistic example: multiple variables at a single point

As a first example of a SOM analysis of real-world data, consider a dataset of chemistry from an Antarctic ice core (Kurbatov et al., 2006), i.e., multiple variables changing over time at a single point in space. Seven chemical species (e.g., Na, K) have been measured in the ice at high resolution then averaged for 15-year intervals over an approximately 11,000 year period. Algorithmically, SOM analysis treats all data samples as a vector, regardless of the original dimensionality of the data or the number of distinct variables involved. In effect, each data sample becomes one row in the input matrix for the SOM code. In this example, each row of the SOM input matrix is a sample for a specific 15-year interval while the columns represent the seven chemical species in each sample. As this is just an example, details of what the data represent will be left to the literature (e.g., Legrand and Mayewski, 1997; Reusch et al., 1999; Kurbatov et al., 2006). Instead, focus on the details and organization of the generalized patterns in Fig. 2a that were extracted from the data by a 6 × 5 SOM analysis (the data are displayed as anomalies from the long-term mean to better highlight pattern differences). There are three key aspects to note about the patterns and their organization. First, any given pattern more closely resembles its immediate neighbors than it does patterns that are farther away. Second, within the limits of the discrete nature of the SOM, there are smooth transitions between patterns, i.e., no abrupt changes are seen moving along any given row or column. These first two points highlight how the SOM algorithm has created a “self-organized” set of spatially organized, generalized patterns that smoothly represent the range of the input data. Third, patterns in each corner show extreme values of some (upper left, lower right) or all (upper right, lower left) of the data. Here the SOM has identified two modes of variability: all the chemistry acting together (all high or low) or just two species acting together. As will be explained further below, these corner patterns are not just the positive and negative phases of the first and second principal components of the input data. Instead, they express aspects of the data that, while not necessarily “nonlinear”, cannot readily be captured by strictly linear analysis methods (e.g., traditional PCA).

As noted earlier, Sammon mapping (Sammon, 1969) provides a method for using spatial proximity (distance) to show dissimilarity between patterns, i.e., the farther apart two patterns are in a Sammon map, the less similar they are. Fig. 2b is an example of this for the SOM patterns of Fig. 2a. Ignoring the upper right patterns for the moment, the original regular grid shape of the SOM, i.e., as shown in Fig. 2a, is readily seen in slightly distorted form. That is, the bulk of the patterns remain relatively evenly spaced with respect to each other, and thus it would probably be wise to limit interpretation of the minor differences in spacing for most of this grid. The main exception to this is found in the upper right set of patterns. This group is clearly offset from the rest of the patterns indicating greater dissimilarity between this trio and their nearest neighbors. It turns out that these patterns represent the bulk of the data for roughly the last 800 years. In this period, the data are distinctly higher and well above the long-term linear increasing trend, i.e., a nonlinear component has developed late in the record. A less dramatic example of clustering may be noted at lower left where another set of patterns stands slightly apart from the rest (representing the low values found in the earliest part of the record). In short, the dominant asymmetry of the Sammon map shape captures the late-record nonlinear trend in chemistry values by clearly separating the highest value patterns from those representing the linear trend.

4. A two-dimensional example: atmospheric model fields

SOMs are equally applicable to 2-D data, e.g., meteorological fields defined on latitude/longitude grids. This 2-D, single variable approach has featured in a number of studies of sea-level pressure fields. Examples include the Arctic work of Cassano et al. (2006a,b) and Schuenemann et al. (2009), the Antarctic work of Lynch et al. (2006) and Uotila et al. (2007), and the Australian drought study of Verdon-Kidd and Kiem (2009).

In this example (and the multivariate example that follows), the fields are from a 24-year dataset based on Polar MM5 model runs (Monaghan et al., 2006) and are at daily resolution (averaged from 6-hourly) on a 60-km grid. As mentioned in the earlier discussion of reference vectors, 2-D grid data are reshaped into a 1-d vector with the same total number of data values. This vector becomes one row (1 day) of the SOM input matrix, where each column represents a grid point, i.e., a measurement of the same variable but at a different point in space. The most common structure for the SOM input data is thus analogous to what is done for S-mode PCA (Cattell, 1952; Richman, 1986).

As mentioned earlier, the size of the SOM grid is a subjective choice and depends, in part, on what level of generalization is desired, which, in turn, depends on the goals of the analysis. Here the main goal was to get a concise summary, so a 4 × 3 grid was selected (as seen in the earlier artificial dataset example, a 3 × 2 grid is mathematically possible but is usually overly concise). Fig. 3 shows the patterns and associated Sammon map from an analysis of 2-m daily temperatures for all Julys in this 24-year dataset (744 days). Temperatures are displayed as anomalies to better show deviations from long-term means. Patterns dominated by warm anomalies tend to appear at the left side while those with widespread cold anomalies are at the right, showing that the SOM has clearly captured the highs and lows of the data (but not the extremes since these are generalized patterns). As before (Fig. 2), adjacent patterns are more similar than those farther away, and patterns change smoothly across rows and columns. The reduction in the number of patterns versus Fig. 2 affects the smoothness of changes, but this is to be expected. Patterns in opposite corners are clearly not mirror images (i.e., not positive/negative PC phases) but remain reasonably complementary (especially 1 and 12). The Sammon map is more regular than that seen in Fig. 2 but still suggests a measure of nonlinearity in the data: the diagonals (dashed lines) are clearly not the equal length that they would be in a linear analysis, e.g., PCA.

5. Further interpretation and temporal change

With further examination of the patterns in Fig. 3, three somewhat subjective groups can be created based on spatial distribution of the temperature anomalies (Fig. 4). The first group includes patterns with warm anomalies predominantly in East Antarctica (patterns 1–3, 5 and 9). A complementary group made up of East Antarctic cold anomalies consists of patterns 8 and 12. Lastly, two patterns (7 and 11) are (loosely) dominated by warm anoma-
lies in the Ross Sea region of West Antarctica (although pattern 11 might also fit in the cold East Antarctic group).

A very common analysis step when working with SOMs is to compare pattern frequency (how often a pattern appears) for different time periods or other subsets of the input data. The two main approaches for subsetting the data are to just select specific time slices (e.g., first half, second half), look for trends, or to use an external index (e.g., an index for the North Atlantic Oscillation, or NAO) to identify time steps when the index is in an extreme state (usually high and low values). The latter may help identify changes in the patterns associated with changes in external factors (e.g., what does West Antarctica look like during El Niño events?).

Fig. 3. Generalized patterns (upper) and associated Sammon map (lower) for an analysis of July daily 2-m temperatures, land-only, 1979–2002, on a rectangular 4 × 3 SOM grid. Data from Monaghan et al. (2006).
Niño?) or to better characterize conditions during extremes of the index (e.g., what do North Atlantic temperatures look like when the NAO is in its high state?). In this example, a simple comparison is done between the first (1979–1983) and last (1998–2002) 5 years of the dataset. These 5-year periods are arbitrary, so care must be taken in interpreting the changes seen (as with any time series analysis involving just two points). With that in mind, the frequency changes (see caption for details) seen in this comparison remain intriguing since they suggest a distinct shift in Antarctic temperature variability: East Antarctic patterns (cold and warm) dropped by 27 percentage points (occurring 74% of the time in the first 5 years but only 47% in the last 5 years) while Ross Sea pattern occurrences shifted from roughly 1 in 12 (8%) to over 1 in 4 (27%). Most of the East Antarctic change comes from a 20-point drop in the frequency of cold patterns. In short, if one wanted to explain Antarctic temperature variability in 1979–1983, chances

---

**Fig. 4.** Pattern groups found in $4 \times 3$ 2-m temperature analysis: (a) warm East Antarctica, (b) cold East Antarctica, and (c) warm Ross Sea sector. Frequency analysis shows the following changes between 1979–1983 and 1998–2002: (a) 28–16%, (b) 36–16%, and (c) 8–27%. Patterns not shown changed from 28% to 41%.
are there was either a warm or cold anomaly of reasonable spatial extent in East Antarctica. For 1998–2002, the probabilities still favor (warm) East Antarctica but the odds of a moderately warm West Antarctic anomaly pattern have increased substantially.

6. Transitions and trajectories

Another common analysis, particularly with time series data, looks at which pattern comes next for each input record, e.g., if this pattern represents today, what pattern follows tomorrow? Known as transition analysis, results can identify preferred sequences of change in the data and may provide a measure of predictability. Results can be summarized as a table, but visualization using arrows scaled to percent frequency on the SOM grid is usually more informative. For example, by following the longest arrow out of each pattern, the presence/absence of cycles and their statistical likelihood can be readily discovered.

Fig. 5 is based on such an analysis using the 2-m temperature SOM patterns of Fig. 3. Here details of the transition analysis have been summarized in a cartoon of a hypothetical cycle in the data. This cycle is "hypothetical" because the actual data rarely progress all the way through over the 10 days represented by the 10 patterns. However, the result remains useful in helping us to understand how the atmosphere behaves by showing the changes most commonly seen in the data.

Two frequency values (as percentages) are associated with each pattern in Fig. 5. The first is the likelihood of persistence, i.e., that the next day will look much like the current day and will thus also match best with the same pattern (shown as a percentage within each pattern). For example, if pattern 5 is the best match for today, there is a 57% chance that it will also be the best match for tomorrow. In nearly all cases this value exceeds 50% indicating that persistence is generally relatively high (or at least that the chance of a significantly different pattern is on the order of 50/50) on a daily timescale. It is possible that the likelihood of persistence is high simply because the patterns are so generalized and thus are missing many details at a synoptic level that would be significant to change at that scale. Alternatively, the winter atmosphere over much of the Antarctic ice sheet is often quite stable due to formation of a strong surface temperature inversion in the absence of solar radiation (and the presence of the distinctly cold surface of the ice sheet) (e.g., Parish, 1982; Warren, 1996). A more detailed comparison of the data and the patterns would help resolve this question.

The second value associated with Fig. 5 patterns, adjacent to the arrows, shows the most likely change to occur when persistence does not apply. Here the next day is sufficiently different from today that it matches better with a different pattern. These percentages are based on the number of events left after persistence has been removed, i.e., if 4 of 10 events involved persistence, then the transition percentage shown is based on the remaining six events. For example, if pattern 1 is the best match for today, there is a 63% chance of staying there tomorrow, but if there is change, there is a 53% chance of moving to pattern 10 (and a 47% chance of moving to one of the remaining patterns). In two cases, multiple arrows are shown leaving a pattern due to transition percentages being equal (pattern 3) or very close (pattern 8).

The set of patterns and transitions in Fig. 5 suggests a cycle in the data where temperature anomalies develop through a trajectory of spatial changes that move roughly counterclockwise around the continent. Starting in pattern 4, a warm anomaly in East Antarctica (possibly part of a remnant dipole with the Weddell Sea and West Antarctica) strengthens (patterns 3, 2, 1) and moves around into West Antarctica (patterns 5, 9, 10) before dissipating and switching to a cool East Antarctic anomaly (patterns 11, 12, 8). As mentioned above, though, this is a highly generalized version of events and the likelihood of the circulation actually behaving in exactly this way over consecutive days is exceedingly small. Furthermore, the actual data may transition into the cycle at any point and may leave the cycle at any point. This cartoon is meant to show only the highest probability transitions, not the transitions that

---

![Fig. 5](image_url) Fig. 5. A cycle in 2-m temperature patterns from the 4 x 3 analysis. Persistence is indicated by percentages shown within each pattern. Most likely next pattern is shown with arrow out of each pattern. Percentages near arrows indicate likelihood of transition after persistence has been removed.
always happen. However, this summary does still give useful insight on the behavior of surface warm and cold anomalies during winter on the ice sheet.

7. Multivariable, multidimensional SOMs

SOM analysis is not limited to one variable at multiple points or multiple variables at one point but can also operate on multivariable/multidimensional (MV/MD) data (although the spatial dimension is usually just limited to 2-D). SOMs have been used in this way to study the North Atlantic Oscillation (Reusch et al., 2007) and West Antarctic atmospheric circulation (Reusch et al., 2005a,b), assess the skill of seasonal forecast models (Gutierrez et al., 2005), and evaluate climate change projections of southern African precipitation (Hewitson and Crane, 2006).

In this example, four circulation variables (2-m temperature and geopotential heights at 850, 700, and 500 hPa) have been analyzed at daily resolution for January 1979–2002 (i.e., austral summer versus the austral winter of previous figures). Recalling Section "Data considerations", these variables were all standardized so as to be on similar scales before processing. As before, there is only one SOM grid for the analysis and there is also only one Sammon map version of this grid. With more variables, however, come more plots to show the patterns, usually a separate plot for each multidimensional variable for each node of the SOM grid. In this case, four 2-D plots (for temperature and the three geopotential heights) are associated with each SOM node (Fig. 6). As with one 2-D variable per node, initial visual analysis usually involves figures for each variable in grid and Sammon map formats. Following these analyses, however, it becomes very useful to show all variables together at the expense of the grid structure. For this example, such plots make it easier to develop pictures of the 3-D structure of the atmosphere for each SOM node (or just those nodes of particular interest).

Fig. 7 shows the variables associated with the SOM nodes that have increased or decreased in frequency the most when comparing the first (1979–1983) and last (1998–2002) 5 years of the record, i.e., which circulation patterns have become more or less common in the data. Because this is a multivariable analysis, each row represents just one SOM node, but each node has multiple variables associated with it. As with the single-variable analysis earlier, remember that these 5-year periods are arbitrary. With this caveat, Fig. 7 shows a notable change in dominant circulation patterns between these two time periods. For the scenario with the largest decrease in frequency (top row), strong positive geopotential anomalies are found over the Amundsen/Ross Sea region. Since this is an area with a quasi-permanent low-pressure system (the

![Fig. 6. Patterns from a 4 x 3 SOM analysis of (a) 2-m temperature and selected geopotential heights, (b) 850 hPa (Z850), (c) 700 hPa (Z700) and (d) 500 hPa (Z500). All data shown as anomalies from full record mean for each variable at all gridpoints. Blank regions in (b) and (c) indicate geopotential heights below topography, i.e., the ice sheet surface.](image-url)
Amundsen Sea Low), the anomalies indicate a distinctly weaker low in the 1979–1983 time period. Geopotential height anomalies across the rest of the domain are moderately positive. Also, coastal near-surface temperatures are warmer than normal. These circulation patterns have virtually disappeared in the late record. Conversely, for the SOM node of largest frequency increase (bottom row), which did not occur at all in the early record, the circulation has changed to a more zonal anomaly in geopotential height that is focused at lower latitudes (from approximately 65°S and lower). This gradient may reflect a stronger subtropical high in the south Pacific during this time period (1998–2002). Unlike the earlier time period (top row), the balance of geopotential height anomalies are now of opposite sign (i.e., negative) across the domain. Surface temperature anomalies are negative across the continent. These changes in dominant circulation patterns are undoubtedly related to multiple factors, including the behavior of, and connections with, ENSO during these two periods (e.g., Cullather et al., 1996) and changes in the Southern Annular Mode, or SAM (e.g., Turner et al., 2005). However, verification of these dynamical linkages is beyond the scope of this paper.

8. Concluding remarks

Much progress in the natural sciences has come from improved ability to "see" what is going on, using, for example, ice-penetrating radar or seismic techniques to collect data on the glacier bed, or satellite imagery to capture synoptic meteorological conditions. Most techniques of data analysis extract features of the dataset without allowing the user to “see” the full patterns. SOM-based analysis is an exception to this rule in that it optimally simplifies a large and complex dataset into a few patterns that a user can see. SOMs are also readily applicable to datasets with significant nonlinear characteristics, a feature that is more the norm in nature than the exception; with Sammon mapping we can also readily see these aspects. The study of frequency of occurrence, preferred transitions, rates of change, hysteresis, and other key features of our complex world is then made much easier.

SOMs also play a role by providing an independent analysis pathway that can be compared with results from other methods. Such confirmation is vital to the scientific process and the robust understanding of our data. In this way, and others, SOMs add to, and complement, more traditional analysis techniques.

Acknowledgements

This work supported by National Science Foundation (NSF) grants to the author including ANT 06-36618, ANT 05-38064 and ARC 05-31211. We also acknowledge partial support from the G. Comer Foundation and the Center for Remote Sensing of Ice Sheets (NSF 04-24589).

References


