Available online at www.sciencedirect.com
SCIENCE DIRECT®
Progress in Oceanography 59 (2003) 223–239
Progress in Oceanography
www.elsevier.com/locate/pocean

Using self-organizing maps to identify patterns in satellite imagery

A.J. Richardson *,[1], C. Risien, F.A. Shillington

Remote Sensing Unit, Oceanography Department, University of Cape Town, Rondebosch 7701, South Africa

Accepted 28 July 2003

Abstract

Satellite remote sensing has revolutionized modern oceanography, providing frequent synoptic-scale information that can be used to deduce ocean processes. However, it is often difficult to extract interpretable patterns from satellite images, as data sets are large and often non-linear. In this methodological paper, we describe the self-organizing map (SOM), a type of artificial neural network adept at pattern identification. The ability of the SOM to extract patterns from a variety of satellite data, including scatterometer and thermal imagery, is illustrated by example. We characterize inter-annual, seasonal and event-scale variability by using the SOM and relate the output to auxillary variables by using a number of techniques that enhance interpretation. Practical recommendations for the fruitful application of SOMs are given. Although the SOM has only rarely been used in oceanography previously, it is a promising applied mathematical tool for pattern extraction from many types of data, especially large and complex satellite data sets.

Keywords: Self-organizing map; Satellite imagery; Pattern recognition

Contents

* Corresponding author. Tel.: +44-1752-633290; fax: +44-1752-600015.
E-mail address: anr@mail.pml.ac.uk (A.J. Richardson).
[1] Present address: Sir Alister Hardy Foundation for Ocean Science (SAHFOS), The Laboratory, Citadel Hill, PL1 2PB, UK.

0079-6611/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pocean.2003.07.006

1. Introduction

To improve our understanding of the functioning of marine systems, oceanographers rely heavily on remotely-sensed data. Extracting interpretable patterns is often made difficult by the large volumes of data collected. Moreover, there is a growing appreciation of the complexity and non-linearity of living and physical systems, leading to a shift in emphasis from obtaining quantitative descriptions of systems, to identifying qualitative features and patterns (Capra, 1996).

A technique that has been applied fruitfully to extracting interpretable patterns from large and complex data sets is the artificial neural network (ANN). ANNs are computer algorithms that simulate the information processing abilities of the brain by mimicking its basic structure (Dayhoff, 1990). They consist of a network of interconnected simple processing units or nodes that process information in parallel. This feature enables neural networks to learn patterns within data, much like the human brain, rather than being pre-programmed.

As applied mathematical tools for extracting patterns from data, ANNs have a number of advantages over traditional statistical methods (Wassermann, 1989). First, they can solve non-linear problems of almost infinite complexity (Dayhoff, 1990). Second, they are more robust in handling noisy and missing data than traditional methods. This is especially desirable for satellite data from visible and infra-red sensors that often have a considerable portion of the image not visible because of cloud. Last, they do not require prior knowledge and assumptions about the data, such as normality or equality of variances (Chen & Ware, 1999).

This study focuses on the application of one common type of unsupervised ANN particularly adept at pattern recognition and classification, the self-organizing map (SOM, Kohonen, 1984, 1997). In unsupervised or competitive nets such as the SOM, the nodes compete to best represent the data. There have been over 5300 published papers based on the SOM.[2] These have been in a wide variety of areas including assessing beer quality, identification of breast cancer, analysing insect courtship songs, predicting bankruptcies, speech and fingerprint recognition, performing searches on the WEB, and in controlling autonomous robots (see Kaski, Kangas & Kohonen, 1998 for a database of these and other SOM papers arranged thematically). SOMs have also been used extensively in the geographical sciences for synoptic climatology (Hewitson & Crane, 1994; Hewitson & Crane, 2002).

Until recently, SOMs have had little exposure in oceanography: Ainsworth (1999) and Ainsworth and Jones (1999) have used SOMs to improve chlorophyll estimates from satellite by better pixel classification; Richardson, Pfaff, Field, Silulwane and Shillington (2002) and Silulwane, Richardson, Shillington and Mitchell-Innes (2001) have used SOMs to identify characteristic chlorophyll profiles in the ocean; and Hardman-Mountford, Richardson, Boyer, Kreiner, Boyer and Bartholomae (this issue) have applied the SOM to altimeter data and related it to the recruitment of Namibian sardine.

[2] see http://www.cis.hut.fi/research/refs/

The aim of this study is to present the SOM technique to researchers in the oceanographic community that work on satellite data. It is not the intention of this study to give a detailed theoretical description of the SOM algorithm, which can be found in Kohonen (1984, 1997). Rather, it is meant to give a brief overview of the technique, and then demonstrate its utility by applying it to two very different satellite data sets. The first example uses the SOM to identify characteristic synoptic-scale wind fields that have been obtained from scatterometer data. By following the temporal evolution of these wind fields using the SOM, two extreme wind events are contrast. The second example uses the SOM to identify characteristic synoptic-scale patterns of sea surface temperature (SST) that have been obtained from 18 years of weekly thermal satellite imagery. The seasonal cycle of SST is identified, highlighting the considerable within month variability. A SOM on SST anomalies (SST after the climatology is removed) is then used to identify years with stronger than usual upwelling, or that are generally warmer than usual. A method for relating SOM patterns to ancillary variables is then illustrated by relating the thermal patterns to anchovy recruitment in the southern Benguela system. Finally, the SOM technique will be placed in perspective by discussing its strengths and weaknesses, and by comparing it with other approaches for the analysis of multidimensional data.

1.1. The philosophy behind the self-organizing map

The general principle of how the SOM identifies patterns in data is illustrated using a simple, fictitious two-dimensional data set (Fig. 1, modified from Hewitson & Crane, 2002). These data contain both a non-linearity and a breakpoint. The SOM places nodes representing common states or patterns within the input data. Several important features are evident. First, the strong non-linearity for $x < 3$ in the data is well represented by the nodes of the SOM. Second, the SOM has fitted the data well on both sides of the breakpoint. Third, nodes are relatively closer together where there are more data and relatively further apart where there are fewer data. This allows greater resolution of features where there is sufficient data to do so. It is noteworthy that the outlier at (1,4) is not represented by the SOM (no node was placed near it) because it was considerably different from other data. Last, the SOM assumes the data are continuous and spans the breakpoint by placing a node in the centre. This may be beneficial if there are missing data, as the SOM interpolates to "fill in" the gaps. If there are real discontinues in the data, a subsequent frequency analysis, which will be discussed below, would reveal that no input data were similar to this pattern.

The key to the ability of the SOM to extract patterns is the way that it learns; this is embodied in its training algorithm. Artificial neural networks learn by an iterative process, whereby input data are presented

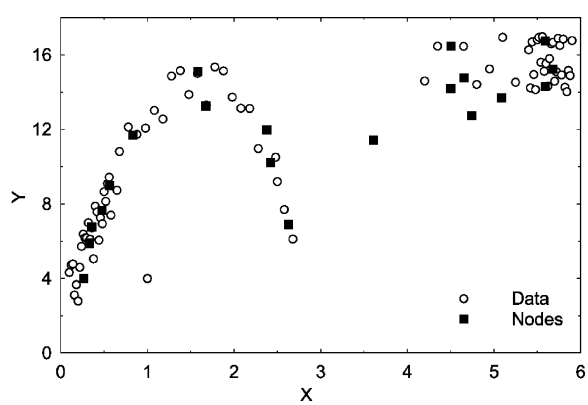

Fig. 1. A fictitious two-dimensional example illustrating the SOM (modified from Hewitson & Crane, 2002).

successively to the network. The initial step in this iterative procedure is to randomly distribute nodes in the data space. Thus, in our two-dimensional example, nodes were initially a random cloud of points in two dimensions. After successive presentations of the input data to the network, the nodes approach positions that best represent the input data. More generally, input data have more dimensions than only two. To visualize their corresponding patterns, the SOM maps multidimensional input data onto a lower (usually two) dimensional continuous output map or array of patterns.

The general SOM training algorithm is outlined heuristically below; mathematical details can be found in Kohonen (1997). A schematic of the training of the SOM is shown in Fig. 2. The number of nodes (output patterns) is defined by the user and is dependent upon the level of detail desired in the analysis. (This issue will be addressed more fully in the Conclusions.) Each node in the output layer is connected via a weighted link to each of the inputs, which in our examples are satellite images. Prior to training, the weights are initialized with starting (usually random) values. This is equivalent to starting with random images on each of the nodes. The first input sample (or row vector) is then compared with each node in the network and the node with the least (Euclidean) difference between itself and the input vector is termed the *winner* and becomes the centre of an *update neighbourhood*. The update neighbourhood is the area within which nodes and their associated weights will be updated, such that each weight vector converges to the input pattern (i.e. positive feedback). Within the update neighbourhood, the degree of adjustment of the weights decreases away from the winning node, according to the spatio-temporal decay function or kernel. Thus, the winning node becomes similar to the input sample and surrounding nodes develop representations of similar, but not the same, patterns. In this way, the nodes in a self-organizing map compete to best represent the particular input sample. This process is repeated for each of the input samples as they are passed sequentially to the SOM. For the SOM to converge on patterns within the data, input samples are then cycled through the SOM a large number of times. During this iterative process, the rate at which the winning nodes converge to the input samples is termed the *learning rate*. Throughout training, the learning rate and size of the update neighbourhood (the *update radius*) decrease, so that the initial generalized patterns are progressively refined. After the training phase, the SOM consists of a number of patterns characteristic of the data, with similar patterns nearby and dissimilar patterns further apart.

Once underlying patterns or classes have been identified, there are several analyses that enhance interpretation. For instance, the SOM can be used to classify the input data into these classes (i.e. the pattern which best represents each of the input samples can be found). In addition, there is also a measure of how well each sample is represented by the SOM: this is the Euclidean distance between the input vector and the pattern it is most similar to. Thus, a large error for a particular input sample means that it is not well represented by the pattern it was mapped to. This can be used to identify outliers. Also there is often additional information (e.g. time, space, auxillary environmental variables) about each input sample that can be related to the patterns identified. The application of the SOM to satellite imagery and analyses to enhance its interpretation will become clearer from the examples.

2. Application of the SOM: some examples

The SOM was performed using the SOM_Pak software Version 3.1 (Kohonen, Hynninen, Kangas, & Laaksonen, 1995), which is produced by, and freely available from the Neural Network Research Centre at the Helsinki University of Technology.[3]

[3] http://www.cis.hut.fi/research/som_lvq_pak.shtml.

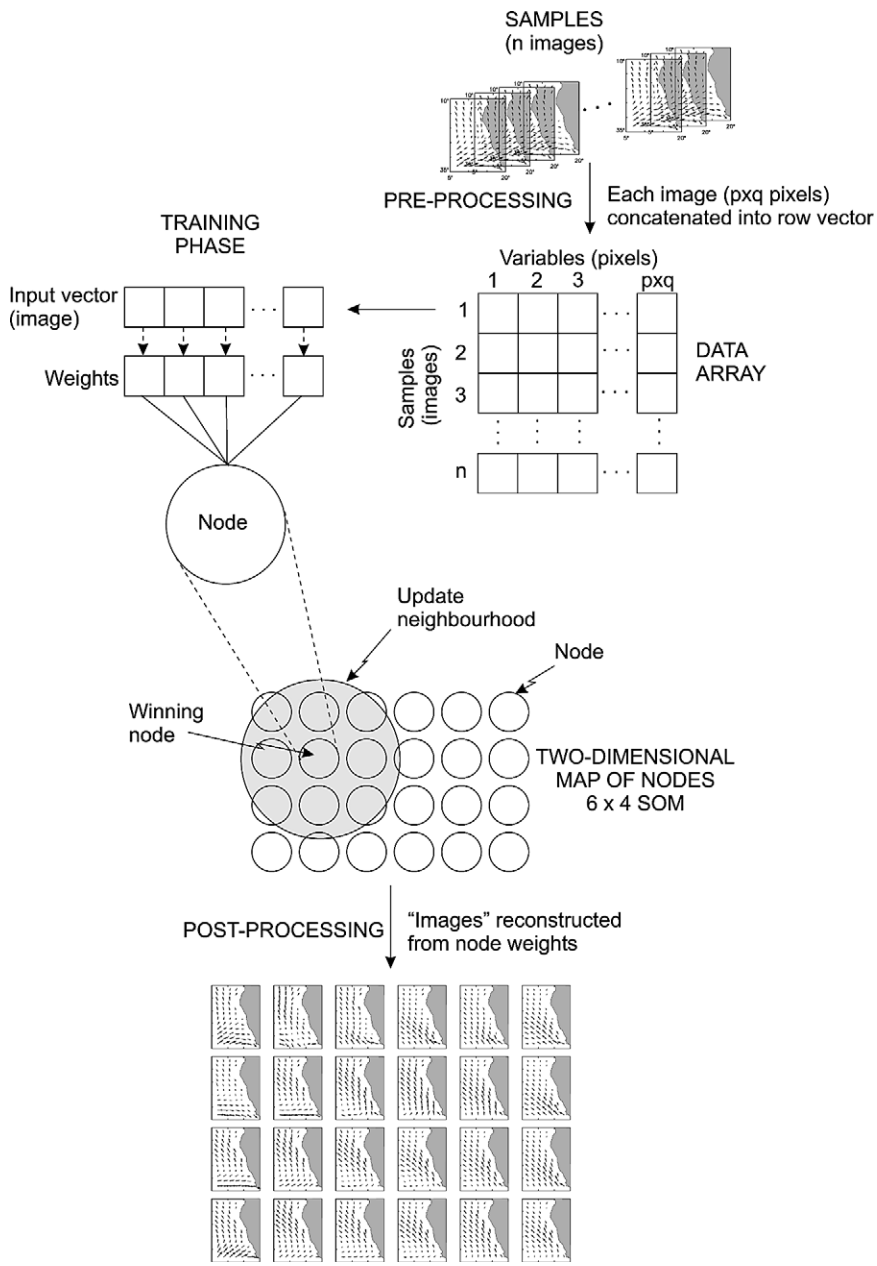

Fig. 2. A schematic depicting the implementation of the self-organizing map, including the pre-processing, training, and post-processing phases.

2.1. Example 1: Scatterometer data

This example shows how the SOM can be used to extract characteristic wind patterns from scatterometer data. The 16-month data record analysed extends from 1 August 1999 to 29 November 2000. The *u* and *v* wind stress components have been sub-sampled from the original QuikSCAT Level 3 Daily Gridded

Ocean Wind Vectors data set. The region of interest was off the west coast of southern Africa, between 10–35°S and 5–20°E. To aid description of this and the following example, a map of the region is presented in Fig. 3. In this region there are 1664 data points. Data were smoothed temporally, to 2-day composites, and spatially, to a 0.5° resolution.

2.1.1. Pre- and post-processing

To apply the SOM to satellite images, some pre- and post-processing of the data is necessary. Each wind vector could be described as either *u* and *v* components or wind speed and direction. Either is suitable and will give very similar results; *u* and *v* components were chosen for simplicity. Each scatterometer image was transformed into a single row vector by concatenating each row in the image (Fig. 2). All *u* components were placed in the first half of the rows followed by all *v* components. The land mask was removed, so that input data consisted only of sea pixels. The final input matrix consisted of 1970 columns (985 grid points × 2 components) × 244 rows (number of 2 day periods). After the SOM was performed, scatterometer images were recreated from the node weights by recombining the *u* and *v* component of each grid point and reconstructing the two-dimensional shape of the image. The land mask was then reinserted.

2.1.2. Synoptic wind states

A 6 × 4 SOM array for the 244 scatterometer images is shown in Fig. 4. Each map represents a typical synoptic state within the data, constructed from the weights on that particular node. Patterns are spread in a continuum across the two-dimensional node space. On the left side of the SOM array, winds are predominantly westerly to southwesterly (south of 25°S), whereas on the right side winds are southerly to southeasterly throughout the region. Once the characteristic patterns were identified, input images were then re-

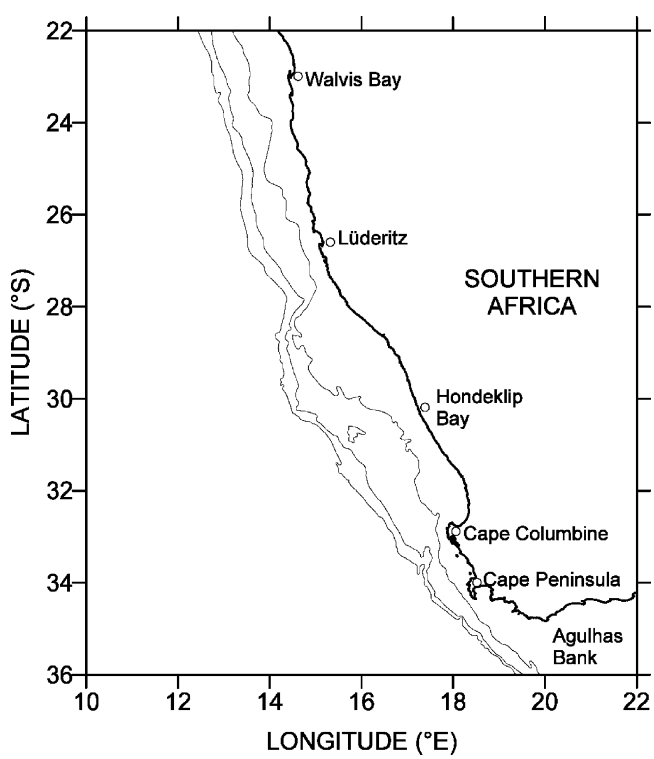

Fig. 3. A map of the southeast Atlantic showing points of interest referred to in the text.

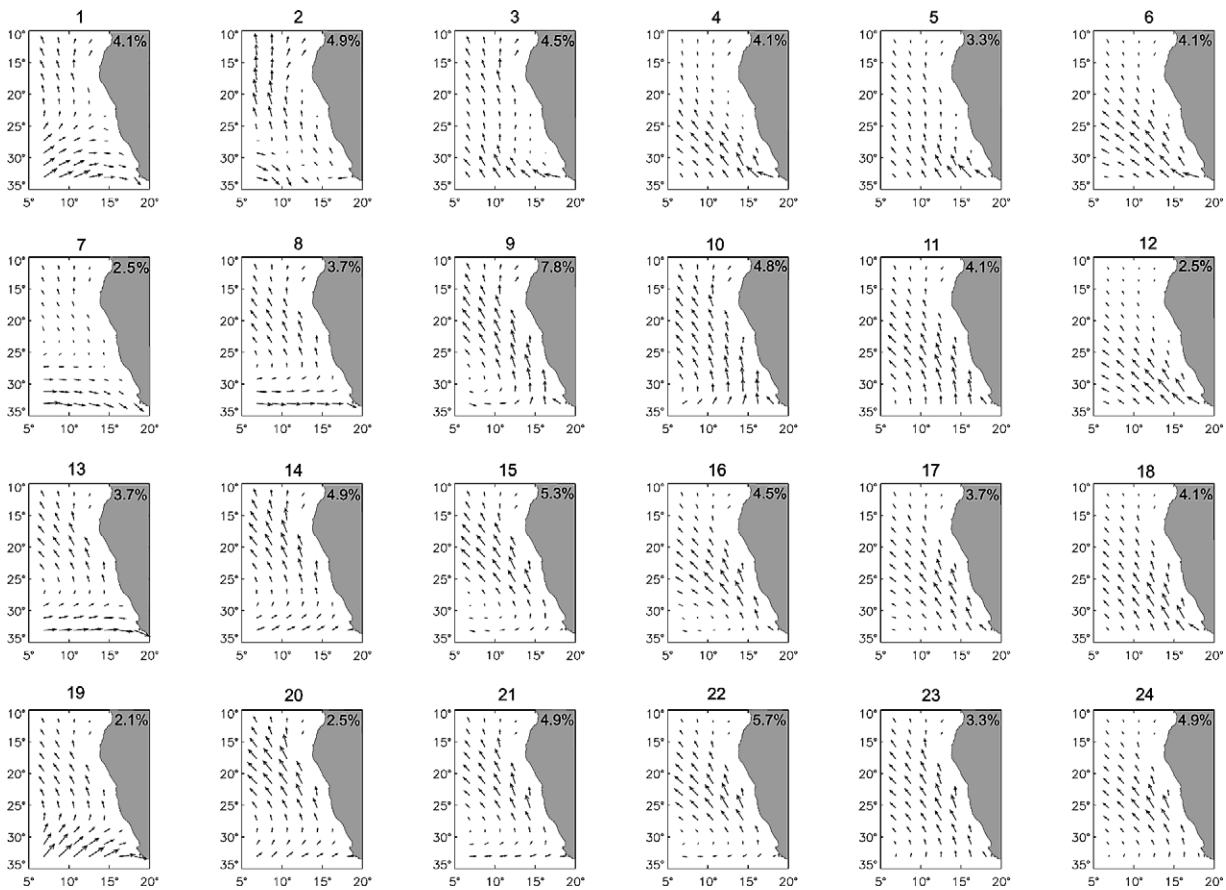

Fig. 4. A 6 × 4 SOM of 244 scatterometer images (two-day composites) depicting wind vectors over the southeast Atlantic. A two-step training schedule was used: the initial training consisted of 1000 iterations, a learning rate of 0.2, and an update radius of 5; the second training consisted of 50,000 iterations, a learning rate of 0.02, and an update radius of 0. The relative frequency of each pattern is shown in the top right corner of each pattern.

presented to the SOM to find out which pattern they are most similar to. This was then used to determine the frequency of occurrence of each pattern (presented as the relative frequency in the top right of each pattern in Fig. 4). All SOM output patterns exist in the data (no zero frequencies). The most common synoptic wind pattern, occurring in 7.8% of the images, was node 9.

2.1.3. Frequency analysis of synoptic states

The relative frequency of each synoptic state can be viewed seasonally (Fig. 5). This map can be visualized as being superimposed on the SOM array, so that the co-ordinates of the frequency map correspond to those of the SOM array. For example, the top left corner of the frequency map corresponds to the top left corner of the SOM array.

Frequency maps highlight the seasonal variability in synoptic states. Viewing Figs. 4 and 5 together, the frequency of node mappings for spring and autumn are spread relatively evenly across the entire SOM array, with maxima (~12%) at nodes 9 and 14. Node 9 depicts relatively strong southeasterly winds close to the coast. Moving north from the Cape Peninsula (~34.5°S) towards Lüderitz (~26°S) and Walvis Bay (~23°S), these southeasterlies increase in both stress and offshore extent, with the area of maximum stress

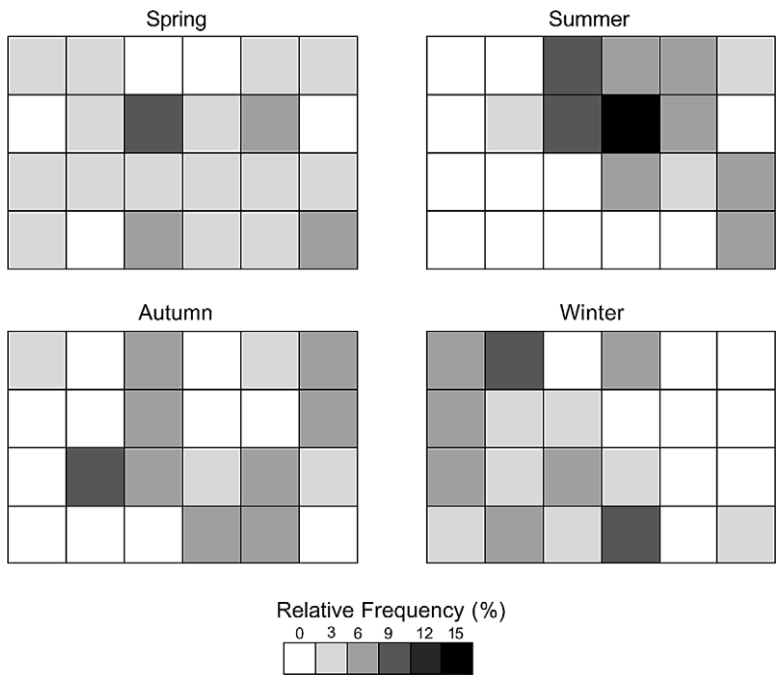

Fig. 5. Seasonal frequency maps of the 6 × 4 SOM of scatterometer images. Percentages for each season sum to 100.

at ~25°S, 12°E. By contrast, node 14 depicts a band of relatively strong southeasterly winds between 15 and 28°S, with a maximum at ~18°S, 11°E. Light southwesterly winds are found off the Cape Peninsula. These winds weaken and become more southerly in direction towards 28°S.

The most frequent pattern in summer (~18%) is node 10, which has a more uniform south to southeasterly wind field compared to those of spring and autumn. Winds tend to be marginally stronger in the southern Benguela region, south of 28°S, with those off the Cape Peninsula having a strong southeasterly component. Moving up the coast, winds tend to become more southerly.

The frequency map for winter differs markedly from those of other seasons, with most states on the left side of the SOM array. Patterns generally have westerly, northwesterly, and southwesterly winds south of 28°S. By comparison, southeasterly winds occur further north between 10 and 28°S. The winter frequency plot is bimodal, with nodes 2 and 22 having a frequency of ~12% each. Node 2 has strong northwesterly winds south of 28°S. Winds closer to the coast tend to be weaker than those offshore; additionally they tend to blow in a more easterly direction. Further up the coast there is once again an increase in wind stress, with winds north of 26°S containing a strong southerly component. Node 22 contains a substantially different pattern compared to node 2. There is a strong anticyclonic component to the circulation, with high wind stress confined to a band between 18 and 28°S. Within this band, the strongest wind stress is off the coast at ~26°S.

The seasonal evolution of patterns illustrated by the SOM can be easily interpreted in terms of the movement of the South Atlantic Anticyclone (SAA). The SAA is maintained throughout the year, providing the pronounced anticyclonic circulation that is apparent in the SOM array. In summer, the SAA is centred at approximately 30°S, 5°E (Shannon, 1985), leading to strong southeast winds over the majority of the region. In winter, the SAA is centred at 26°S, 10°E, leading to strong to gale force westerly and southwesterly winds between 26 and 35°S.

2.1.4. Identifying trajectories

Since each synoptic wind pattern maps to a node, the temporal evolution of synoptic events can be traced by following the trajectory through time of SOM patterns. To illustrate this technique, two intense southerly wind events were selected, one in October 1999 and the other in February 2000. Their trajectories show rapid changes in synoptic states during these events (Fig. 6).

2.1.4.1. 14–22 October 1999 On the 14 October, the region of maximum southerly wind was between 20–26°S. By the 16 October, these winds had become more easterly and moved further south, a condition that persisted on the 18 October. These winds then moved even further south and became more easterly on the 20 October. On the 22 October, there was a rapid movement back to the initial synoptic state of the 14 October. From the corresponding synoptic weather charts (from the South African Weather Services), the southward movement of maximum southeasterlies is a consequence of the ridging of the SAA.

2.1.4.2. 11–19 February 2000 On the 11 February, there was a reasonably uniform southerly wind field over most of the area, extending as far east as 10°E in the south. On both the 13 and 15 February, the synoptic state had changed dramatically, with a belt of intense southeasterly winds between 23 and 28°S. By the 17 and 19 February, this band of strong southeasterly winds had shifted south. The synoptic wind system had returned to a state similar to that of the start of the event by the 21 February.

Examining the temporal evolution of synoptic events using the SOM provides a useful method to compare events. For instance, both events came back to very close to the same initial state, and both included a ridging anticyclone, although the February event was much weaker. Moreover, in each trajectory, the same pattern was sometimes repeated in consecutive images, indicating that the system was stationary for 3–4

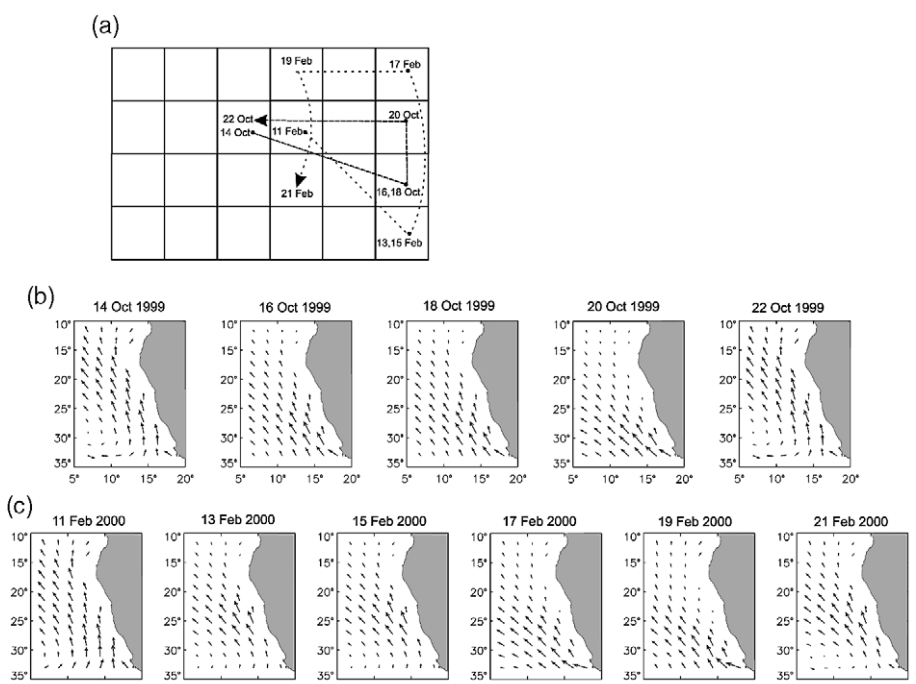

Fig. 6. Trajectories through SOM space of two intense wind events. The first wind event led to a wind maximum off the coast at 34°S, and the second event led to a wind maximum off the coast at 26°S. (a) Trajectories overlaying SOM array, (b) the nodes visited during the October 1999 wind event and (c) the nodes visited during the February 2000 wind event.

days. An interesting contrast between the two events identified by the SOM was the strong southeasterly winds between 23 and 28°S in the middle of the February 2000 event, but not the October 1999 event. These strong southeasterly winds were caused by an interaction between a mid-latitude cyclone and the SAA, which led to compression of the isobars between 23 and 28°S.

2.2. Example 2: Thermal imagery

This example shows how the SOM can be used to extract characteristic synoptic-scale sea surface temperature patterns, using data from the Advanced Very High Resolution Radiometer aboard the NOAA-series of polar-orbiting environmental satellites (POES). The data set comprises 18 years (1982–1999) of weekly composites (864 images). The spatial resolution is 4.5 km at the equator. The region of interest was the southern Benguela region off the west coast of South Africa, between 30–35°S and 16–20°E.

2.2.1. Pre- and post-processing

Each SST image was transformed into a single row vector by concatenating image rows as for the scatterometer images. Land was removed from the analysis, so that the input data consisted of 7020 sea pixels. As thermal infra-red sensors cannot penetrate cloud, cloud was coded as a missing value and then had no influence on the analysis. The ability to cope with missing data without a priori interpolation is a useful aspect of the SOM that is not the case for other multivariate approaches. The input matrix consisted of 7020 columns (pixels) $\times$ 864 rows (weeks). After the SOM was performed, the final weights for each node were reformed into two-dimensional images and the land was reinserted.

2.2.2. Synoptic SST patterns

A 5×3 SOM of 864 weekly sea surface temperature images is shown in Fig. 7. There is a continuum of change across the SOM array, with cool patterns on the left and warm patterns on the right. Superimposed on this general pattern is cooler water off capes (indicating upwelling) in the patterns toward the bottom. These upwelling cells are less developed toward the top of the SOM array. The relative frequency of occurrence of each pattern is also shown in Fig. 7, indicating that all SOM output patterns exist in the data. The most common synoptic SST pattern, occurring in 11.6% of the images, was node 6.

2.2.3. Identifying the seasonal cycle

The movement of patterns around the SOM array is summarized by plotting the centroid (mean position) of each month (Fig. 8(a)). This representation not only highlights the seasonal cycle, with winter on the left side of the SOM array and summer on the right side, but reveals its asymmetry; temperatures gradually cool until the austral winter, remain cold for a few months, and then rapidly warm to summer conditions. A more detailed examination of the seasonal cycle of SST is shown in Fig. 8(b). This frequency analysis highlights the variability in patterns within each month; a mean pattern does not capture this variability. For instance, April is the most variable month with offshore water temperatures as low as 16°C or as high as 22°C. This is reflected in April having 10 synoptic states, whereas May to October have only 6 or 7 states each. There is also considerable variability in December, with upwelling sometimes off the Cape Peninsula (pattern 9), sometimes off both the Cape Peninsula and Cape Columbine with weak upwelling further north (patterns 14 and 15), and other times off the Cape Peninsula, Cape Columbine and Hondeklip Bay (pattern 13).

2.2.4. Characterizing inter-annual variability

Owing to the strong seasonal cycle, a SOM was performed on SST data after the seasonal climatology had been removed by subtracting the weekly climatology for the 18-year period from each weekly image (Fig. 9). There is a continuum of SST anomaly patterns across the SOM array, from cooler patterns on

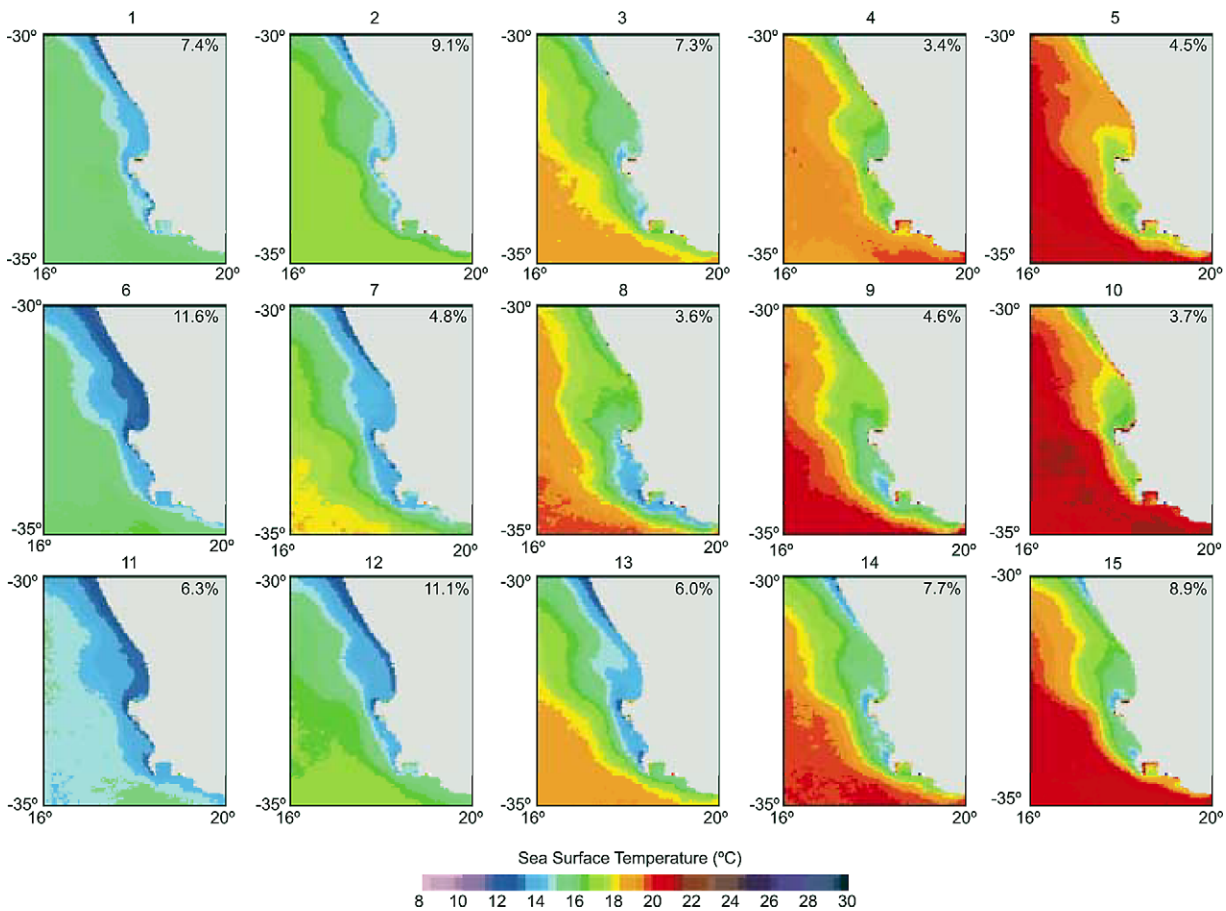

Fig. 7. A 5 × 3 SOM of 864 weekly sea surface temperature images in the southern Benguela region. A two step training schedule was used: the initial training consisted of 1000 iterations, a learning rate of 0.2, and an update radius of 4; the second training consisted of 10,000 iterations, a learning rate of 0.1, and an update radius of 0. The relative frequency of each pattern is shown in the top right corner of each pattern.

the left to warmer patterns on the right. Some individual patterns are noteworthy: pattern 4 shows warming in the north; pattern 10 shows warming everywhere; and pattern 11 shows almost overall cooling. The frequency of occurrence of each pattern is also shown in Fig. 9, indicating that the most common synoptic SST anomaly pattern, occurring in 12.7% of the images, was node 7.

To highlight inter-annual differences, a frequency map of SST anomaly states was constructed for three contrasting years, viz. 1982, 1993 and 1998 (Fig. 10). For instance, 1982 in the southern Benguela was generally mild to cool, with most patterns on the left side of the SOM. By contrast, 1998 was an extremely warm year, as all patterns correspond to overall warm conditions. The year 1993 was very different to the other years, as it was extremely variable, with almost all synoptic SST anomaly states occurring.

The temporal evolution of individual patterns of interest can also be followed. For instance, pattern 10 (Fig. 9) represents warm conditions; its time series shows that 1986, 1991–1992 and 1999 were warm years (Fig. 11). By contrast, pattern 13 (Fig. 9) shows negative SST anomalies extending northwest from the Cape Peninsula, representing increased upwelling. An annual time series of the frequency of this pattern shows that there is considerable inter-annual variability, with intense upwelling off the Cape Peninsula in

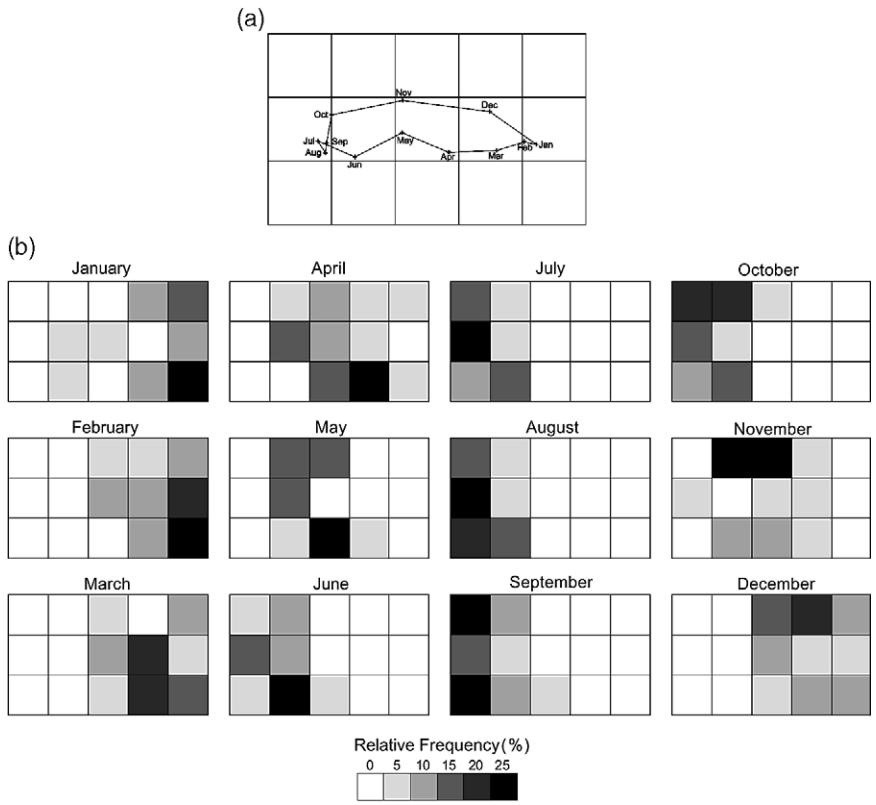

Fig. 8. (a) A map of centroids for each month illustrating the seasonal cycle in SOM space. This centroid map overlays the 5×3 SOM array. (b) Monthly frequency maps of the 5×3 SOM of sea surface temperature. Percentages for each month sum to 100.

1982–1984, 1989, 1994–1995 and 1999 (Fig. 11). Interestingly, 1999 had both increased upwelling off the Cape Peninsula and very warm overall conditions.

2.2.5. Relating inter-annual variability to auxillary variables

The SOM can also be related to auxillary variables by correlation of individual patterns with an external time series. As an illustration, the time series of SST anomaly patterns is related to the recruitment of anchovy in the southern Benguela system. Anchovy in this region spawn between September to February each year (Huggett et al., 1988), and it is thought that recruitment (the number of young fish that become available to the fishery each year) is dependent upon physical oceanographic conditions during spawning (Cury & Roy, 1989; Cury et al., 1995; Shin, Roy, & Cury, 1998). In particular, it has been suggested recently that intense upwelling off the Cape Peninsula may be detrimental for recruitment because of advective loss of spawning products, whereas moderate upwelling north of Cape Columbine is good because it provides food for developing larvae (Hutchings, Bloomer, Boyd, Crawford, Huggett, Kerstan et al. 1998; Roy, Weeks, Rouault, Nelson, Barlow & van der Lingen, 2001). Node 8 represents the condition where it appears that warm water is advected from the spawning grounds on the Agulhas Bank to the area north of the Cape Peninsula and moderate upwelling is found north of Cape Columbine. This hypothesis was examined by relating the frequency of node 8 between September to February each year to the subsequent recruitment. Time series of anchovy recruitment and node 8 each year is shown in Fig. 12 (note that

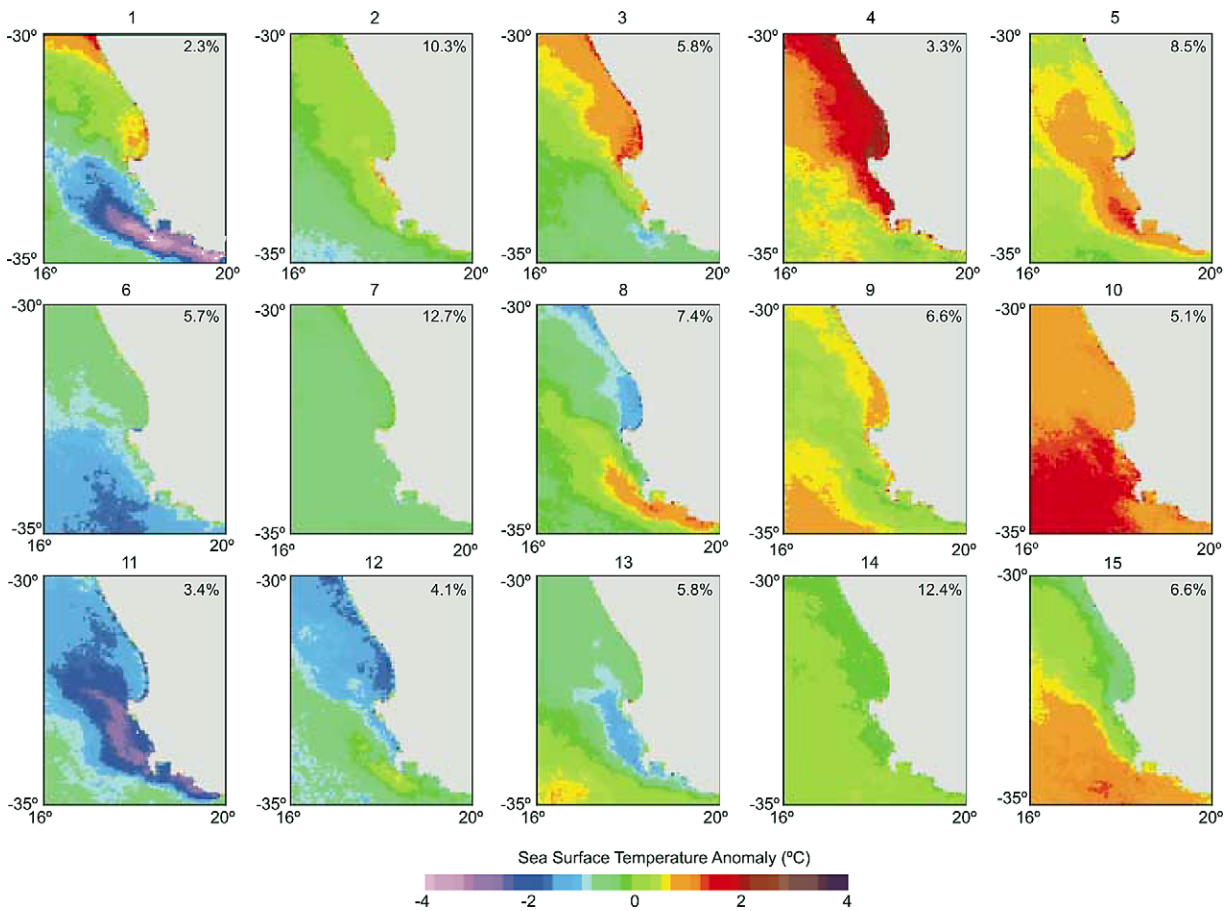

Fig. 9. A 5 × 3 SOM of 864 weekly SST anomaly images (SST images after the weekly climatology has been removed). The training schedule used was the same as the one for the SOM on SST (see Fig. 7). The relative frequency of each pattern is shown in the top right corner of each pattern.

recruitment estimates are only available from 1985 onwards). There is a significant correlation between the two series ($r = 0.57$, $n = 15$, $p < 0.05$), but not for any other nodes.

3. Conclusions

The SOM is a powerful technique for pattern identification in large and complex satellite data sets. It is intuitively appealing, because it allows easy visualization of underlying patterns and it handles a variety of data types, as illustrated by the scatterometer and SST examples. The SOM technique may be especially useful for deducing dynamic oceanic processes from satellite imagery, as it assumes there is an underlying continuum of patterns.

This study highlights a number of techniques that can be used to enhance interpretation of the SOM array, in an attempt to better interpret underlying patterns and processes on the inter-annual, seasonal and event scales. Seasonal cycles can be described by using centroid maps to describe annual trajectories. The relative importance of each synoptic state can determined by frequency maps. Particular events can be

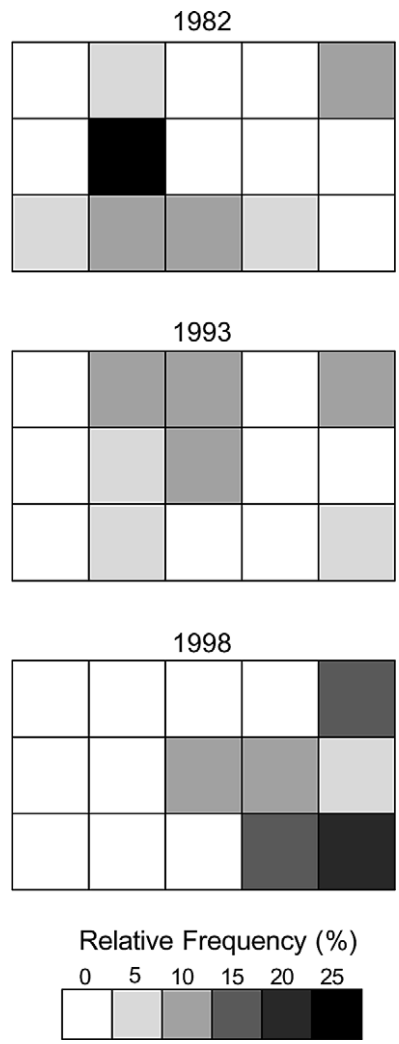

Fig. 10. Annual frequency maps for 1982, 1993 and 1999 of the 5 × 3 SOM of SST anomalies. Percentages for each year sum to 100.

traced using the trajectory through time over the SOM array. The time series of a particular node of interest can be related to auxillary variables, as was done with the anchovy recruitment and SST example. It is also possible to analyse the frequency through time of a particular node (or group of nodes) using conventional time series techniques such as spectral analysis (Hewitson & Crane, 2002) or wavelet analysis.

Although it is not the intention or within the scope of this study to compare various multivariate techniques directly, a general comparison of SOM with other more-traditional multivariate techniques may help to place SOM in a broader context. The SOM array itself is similar to a combination of cluster analysis and multidimensional scaling (MDS). Patterns in the SOM array represent characteristic states, and are similar to a cluster as there are generally a number of similar input patterns that have mapped to each state. The SOM array, however, captures relationships among all clusters by placing them in a two-dimensional array. Dendrograms do not represent the relationship among groups well because the relative pos-

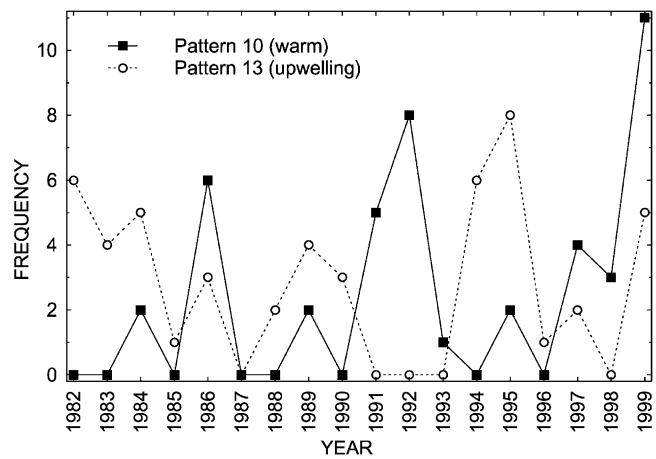

Fig. 11. Time series of the annual frequency of overall warm conditions (pattern 10, see Fig. 9) and upwelling (pattern 13, see Fig. 9) from 1982 to 1999.

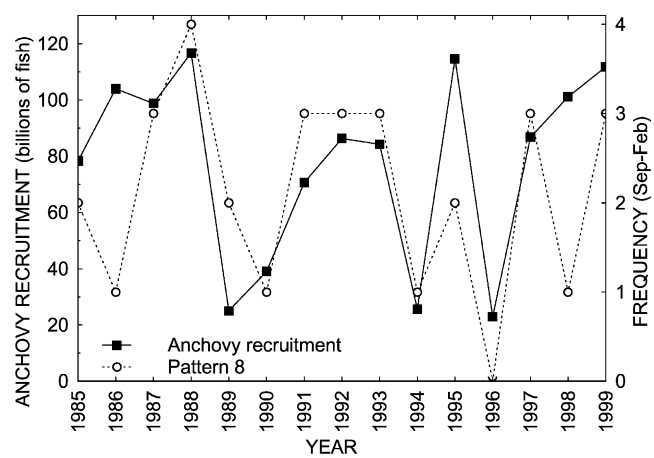

Fig. 12. Time series of anchovy recruitment (1985 to 1999) in the southern Benguela and frequency of pattern 8 (see Fig. 9) during the anchovy spawning season (September to February).

itions of the groups are arbitrary. Both MDS and empirical orthogonal functions (principal components analysis) capture relationships among samples well, but do not explicitly define groups.

A weakness of the SOM approach (that is to a lesser extent also applicable to MDS) compared with other more-conventional multivariate techniques is that there is not a robust statistical theory underpinning it, as there is for parametric statistics. Thus, there is no significance testing of the patterns identified. Empirical orthogonal functions analysis provides statistical testing of the underlying principal components. This statistical rigour is at the cost of stricter assumptions that are not always met.

A perceived difficulty in using the SOM is that the number of patterns chosen is arbitrary, as the dimensions for the SOM array are chosen by the researcher. If the data has a continuum of underlying patterns, the number of patterns chosen is dependent upon the level of detail desired in the analysis. For instance, a very large SOM array identifies a large number of patterns revealing detailed structure, whereas a small SOM array identifies a few generalized patterns. However, if the data have a finite number of patterns,

the SOM would still identify these: those patterns would have high frequency and other patterns that were produced in the SOM array by interpolation would have zero frequency. The decision on the number of groups to specify in a SOM analysis is analogous to the researcher making an arbitrary decision on the level of similarity required to identify groups in cluster analysis.

A major strength of the SOM approach is that the underlying patterns in a data set can be visualized in the same form as the original data. Thus, if input data are SST images, then the outputs are SST patterns. This is an advantage over other multivariate techniques such as cluster analysis, multidimensional scaling and empirical orthogonal function analysis, which output various abstractions of the original data. As the output patterns resemble the input format, they are more easily interpreted than output from conventional multivariate techniques.

Both SOM and empirical orthogonal function analysis are particularly useful for large datasets, as they allow easy visual interpretation of a large number of images (samples). As with other ANN techniques, the performance of SOMs improves with large data sets, as they are better able to learn underlying patterns. By contrast, cluster analysis and MDS, however, are difficult to visualize for a large number of input samples.

An advantage of SOMs over other multivariate techniques is that the algorithm is robust in handling missing data, without a priori estimation. SOM patterns do not have missing data, as they are "filled in" by other input data that mapped to the same node, as well as from input data that mapped to neighbouring nodes. In this sense, SOMs can be used as data interpolation techniques, estimating missing data from input data that are similar (Hewitson & Crane, 2002).

Although the SOM technique has been applied to satellite data in this study, it is equally applicable to many other data types. For instance, numerical oceanographic models output massive volumes of data, sometimes making it difficult to deduce underlying processes. SOMs provide an easy method for obtaining qualitative patterns from such large and complex datasets. Characteristic patterns identified using a SOM could also provide baseline cases for simulation studies. SOMs have been used to compare empirical and model outputs of global circulation models by identifying similarities and differences in synoptic states (Main, 1997). The SOM is becoming an invaluable tool for data mining in many scientific disciplines and it is likely that it will be used increasingly in satellite oceanography in the future.

Acknowledgements

This research formed part of the ENVIFISH program, contract number IC18-CT98-0329. The ENVIFISH program provided funding for AJR. Dudley Chelton is thanked for processing and providing the scatterometer data. Bruce Hewitson, Jeremy Main and Geoff Brundrit are thanked for their useful discussions on the use of self-organizing maps.

References

Ainsworth, E. J. (1999). Visualization of ocean colour and temperature from multispectral imagery captured by the Japanese ADEOS satellite. *Journal of Visualization, 2*, 195–204.

Ainsworth, E. J., & Jones, S. F. (1999). Radiance spectra classification from the ocean color and temperature scanner on ADEOS. *IEEE Transactions on Geoscience and Remote Sensing, 37*, 1645–1656.

Capra, F. (1996). *The web of life: a new synthesis of mind and matter*. London: HarperCollins.

Chen, D. G., & Ware, D. M. (1999). A neural network model for forecasting fish stock recruitment. *Canadian Journal of Fisheries and Aquatic Sciences, 56*, 2385–2396.

Cury, P., & Roy, C. (1989). Optimal environmental window and pelagic fish recruitment success in upwelling areas. *Canadian Journal of Fisheries and Aquatic Sciences, 46*, 670–680.

Cury, P., Roy, C., Mendelssohn, R., Bakun, A., Husby, D. M., & Parrish, R. H. (1995). Moderate is better: exploring nonlinear climatic effects on the Californian northern anchovy (*Engraulis mordax*). In R.J. Beamish (Ed.), *Climate change and northern fish populations*. *Can. Spec. Publ. Fish. Aquat. Sci.*, *127*, 417–424.

Dayhoff, J. E. (1990). *Neural network architectures—An introduction*. New York: Van Nostrand Reinhold.

Hardman-Mountford, N. J., Richardson, A. J., Boyer, D., Kreiner, A., Boyer, H., & Bartholomae, C. (2004). Relating sardine recruitment in the Northern Benguela to processes inferred from satellite-derived sea surface height using a novel pattern recognition approach. *Progress in Oceanography*, this issue, doi:10.1016/j.pocean.2003.07.005.

Hewitson, B. C., & Crane, R. G. (1994). *Neural nets: Applications in geography*. London: Kluwer Academic Publishers.

Hewitson, B. C., & Crane, R. G. (2002). Self-organizing maps: Applications to synoptic climatology. *Climate Research*, *22*, 13–26.

Huggett, J. A., Boyd, A. J., Hutchings, L., & Kemp, A. (1998). Weekly variability of clupeoid eggs and larvae in the Benguela jet current: implications for recruitment. In S. C. Pillar, C. L. Moloney, A. I. L. Payne, & F. A. Shillington (eds.), *Benguela Dynamics*. *Impacts of Variability on Shelf-Sea Environments and their Living Resources*. *South African Journal of Marine Science*, *19*, 197–210.

Hutchings, L., Bloomer, S., Boyd, A. J., Crawford, R. J. M., Huggett, J. A., Kerstan, M., Korrûbel, J. L., de Oliviera, J., Painting, S. J., Richardson, A. J., Shannon, L., Schülein, F., van der Lingen, C. D., & Verheye, H. M. (1998). Recruitment variations of the South African anchovy: a multiplicity of factors in the spawning, transport and nursery area are all important. In S. C. Pillar, C. L. Moloney, A. I. L. Payne, & F. A. Shillington (eds.), *Benguela Dynamics*. *Impacts of Variability on Shelf-Sea Environments and their Living Resources*. *South African Journal of Marine Science*, *19*, 211–225.

Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of Self-Organizing Map (SOM) Papers: 1981–1997, *Neural Computing Surveys*, *1*, 102-350. WWW Page, http://www.icsi.berkeley.edu/~jagota/NCS/.

Kohonen, T. (1984). *Self-organization and associative memory*. Berlin: Springer Verlag.

Kohonen, T. (1997). *Self-organizing maps*. Berlin: Springer.

Kohonen, T., Hynninen, J., Kangas, J., & Laaksonen, L. (1995). *SOM Pak Version 3.1*.

Main, J. P. L. (1997). Seasonality of circulation in southern Africa using the Kohonen self-organising map. Ph.D. Thesis, University of Cape Town, Cape Town, South Africa, unpublished.

Richardson, A. J., Pfaff, M. C., Field, J. G., Silulwane, N. F., & Shillington, F. A. (2002). Identifying characteristic chlorophyll *a* profiles in the coastal domain using an artificial neural network. *Journal of Plankton Research*, *24*, 1289–1303.

Roy, C., Weeks, S., Rouault, M., Nelson, G., Barlow, R., & van der Lingen, C. (2001). Extreme oceanographic events recorded in the Southern Benguela during the 1999–2000 summer season. *South African Journal of Science*, *97*, 465–471.

Shannon, L. V. (1985). The Benguela ecosystem. Part 1. Evolution of the Benguela, physical features and processes. *Oceanography and Marine Biology Annual Review*, *23*, 105–182.

Shin, Y. -J., Roy, C. & Cury P. (1998). Clupeoids reproductive strategies in upwelling areas: a tentative generalization. In M. -H. Durand, P. Cury, R. Mendelssohn, C. Roy, A. Bakun, & D. Pauly (eds.), *Global versus local changes in upwelling systems*. Orstom éditions.

Silulwane, N. F., Richardson, A. J., Shillington, F. A., & Mitchell-Innes, B. A. (2001). Identification and classification of vertical chlorophyll patterns in the Benguela upwelling system and Angola-Benguela Front using an artificial neural network, In A. I. L. Payne, S. C. Pillar, & R. J. M. Crawford (eds.), *A Decade of Namibian Fisheries Science*. *South African Journal of Marine Science*, *23*, 37–51.

Wassermann, P. D. (1989). *Neural computing theory and practice*. New York: Von Nostrand Reinhold.