INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **21**: 1401–1414 (2001)
DOI: 10.1002/joc.702

CLASSIFICATION OF RAINFALL VARIABILITY BY USING ARTIFICIAL NEURAL NETWORKS

SILAS CHR. MICHAELIDES[a,*], CONSTANTINOS S. PATTICHIS[b] and GEORGIA KLEOVOULOU[b]
[a] *Meteorological Service, Nicosia, Cyprus*
[b] *Department of Computer Science, University of Cyprus, Cyprus*

Received 16 February 2001
Revised 6 June 2001
Accepted 7 June 2001

ABSTRACT

In this paper, the usefulness of artificial neural networks (ANNs) as a suitable tool for the study of the medium and long-term climatic variability is examined. A method for classifying the inherent variability of climatic data, as represented by the rainfall regime, is investigated. The rainfall recorded at a climatological station in Cyprus over a long time period has been used in this paper as the input for various ANN and cluster analysis models. The analysed rainfall data cover the time span 1917–1995. Using these values, two different procedures were followed for structuring the input vectors for training the ANN models: (a) each 1-year subset consisting of the 12 monthly elements, and (b) each 2-year subset consisting of the 24 monthly elements. Several ANN models with a varying number of output nodes have been trained, using an unsupervised learning paradigm, namely, the Kohonen's self-organizing feature maps algorithm. For both the 1- and 2-year subsets, 16 classes were empirically considered as the optimum for computing the prototype classes of weather variability for this meteorological parameter. The classification established by using the ANN methodology is subsequently compared with the classification generated by using cluster analysis, based on the agglomerative hierarchical clustering algorithm. To validate the classification results, the rainfall distributions for the more recent years 1996, 1997 and 1998 were utilized. The respective 1- and 2-year distributions for these years were assigned to particular classes for both the ANN and cluster analysis procedures. Compared with cluster analysis, the ANN models were more capable of detecting even minor characteristics in the rainfall waveshapes investigated, and they also performed a more realistic categorization of the available data. It is suggested that the proposed ANN methodology can be applied to more climatological parameters, and with longer cycles. Copyright © 2001 Royal Meteorological Society.

KEY WORDS: artificial neural networks; cluster analysis; Mediterranean; rainfall variability; self-organizing feature map; weather variability

1. INTRODUCTION

The variability of the weather has always been a major concern to mankind, because it critically affects almost all human activity, and especially the planning ahead of weather sensitive operations. For a better understanding of this variability in weather, scientists have sought to understand the factors that play an important role, and several theories have evolved in this respect. The use of sophisticated climatic models is becoming an increasingly established tool in the effort to understand climatic cause and effect relationships. Through simulated cases, these models offer the opportunity to estimate the impact that various physical processes (natural or man-made) have on possible climatic changes and trends. However, a first step towards a more definite approach for understanding and explaining the variability in the various components of the weather is to quantify and document this variability, and the present paper is an endeavour in this respect. Statistical techniques have traditionally been used in the study of climatic variables. In the present paper, an attempt is being made to identify, and thereby classify, climatic

* Correspondence to: Meteorological Service, Nicosia 1418, Cyprus; e-mail: silas@ucy.ac.cy

variability in Cyprus by using an artificial neural network (ANN) modelling approach. More specifically, the objective of this study is to identify and classify similar patterns in the 1- and 2-year rainfall distributions by using an unsupervised ANN learning paradigm based on the self-organizing feature maps algorithm of Kohonen (1990, 1995). The results of the ANN models are compared with cluster analysis findings derived using the agglomerative hierarchical clustering algorithm.

The motivation behind the use of ANNs lies in their capacity for making no assumptions about the underlying probability density functions, finding near-optimum solutions from incomplete data sets, and the fact that learning is accomplished through training (Haykin, 1994). As a result of these characteristics, ANNs have widely been used in modelling a large variety of dynamic systems that are characterized by non-linearity (Sanchez-Sinencio and Lau, 1992; Arbib, 1995). Modelling of non-linear systems by using ANN ranges from medical applications (Micheli-Tzanakou, 1995), signal and image processing (Luo and Unbehauen, 1997), engineering applications (Bulsari *et al*., 1996) and many more. Such a non-linearity is a prime characteristic of issues related to the atmospheric and hydrologic sciences. In these sciences, ANNs have recently been utilized in classifying cloud observed in satellite imagery (Peak and Tag, 1992; Bankert, 1994), tornado prediction (Marzban and Stumpf, 1996), rainfall run-off estimation (Furundzic, 1998), agrometeorology (Francl and Panigrahi, 1997), sea-surface temperature forecasting (Tangang *et al*., 1998), spatial meteorological analysis (Kalogirou *et al*., 1998), completing time series of meteorological elements (Kalogirou *et al*., 1997), and in forecasting minimum temperature (Schizas *et al*., 1994). It appears that ANN may tackle a large number of problems, and to the best of our knowledge, this is the first time that ANN modelling is investigated for the classification of weather variability.

The analogue method for meteorological analysis is the non-deterministic search of historical weather states that closely resemble a given state, and has been suggested by Lorenz (1969). This method did not find immediate application, and could not be introduced into operational practice, partly because of the lack of appropriate data sets, and partly owing to the lack of appropriate non-linear mathematical tools. However, with more data becoming available, the analogue method has been re-introduced, and appears to be promising (Van den Dool, 1989). ANN modelling, which falls under the category of the generic non-linear analogue techniques has, in actual fact, revived the idea of analogue weather analysis (Elsner and Tsonis, 1992; Nicolis, 1998).

The island of Cyprus is situated in the northeastern most corner of the Mediterranean basin and, therefore, has a typical eastern Mediterranean climate. In the eastern Mediterranean, the climatology of the combined temperature and rainfall regimes is characterized by cool-to-mild and wet winters and by warm-to-hot and dry summers (Taha *et al*., 1981; Furlan, 1997). Another important characteristic of the rainfall regime is that it exhibits a large temporal and spatial variability.

An understanding of the temporal variability of rainfall can be useful in the more efficient planning of several human activities. For example, the distribution of rainfall during the year determines, to a large extent, the growth of plants and, hence, food production; therefore, the understanding of the temporal variability of rainfall can aid in the choice of the most suitable crop variety. It is this temporal variability that the present research attempts to identify, and subsequently classify by using ANN.

To set the scene for the analysis that follows, it is useful at this point to exhibit briefly the large-scale causes of the temporal variability in the meteorological parameter studied here over the area of the eastern Mediterranean. During the winter period, most of the precipitation in the area originates from frontal depressions that move into the area, mostly from the west or northwest. In summer, however, the region is dominated by an interaction of the high pressure which is related to the subtropical system west of Africa, and the low pressure which forms an extension of the Asian monsoon circulation, thus leading to a generally prolonged dry season. Variations in the location, magnitude and timing of these pressure patterns have a significant impact on the precipitation and temperature regimes (Kutiel, 1987, 1988). Although the above large-scale features dominate the variability in temperature and rainfall regimes, local-scale phenomena can add to the temporal

variability; for example, short lasting orographically or thermally induced precipitation formation in unstable airmasses can be an important source of such variability.

The paper is organized as follows. In Section 2, the rainfall data preprocessing steps are described. In Section 3, the self-organizing feature maps algorithm is given, as well as the steps carried out for developing the ANN models. In Section 4, the agglomerative hierarchical clustering algorithm is described. Section 5 covers the results of the ANN classification, and Section 6 contrasts the ANN and cluster analysis classification. In Section 7, a validation of the classification is performed, and in Section 8, the concluding remarks are briefly presented.

2. DATA DESCRIPTION

In the present paper, a single meteorological variable has been used, namely the rainfall. The raw data consist of the daily measurements of this element over a long period of time for the climatological station of Nicosia in Cyprus. The average monthly distribution of rainfall for this station is shown in Figure 1. The raw data used in this study are the daily accumulated rainfall (in mm), recorded at 08:00. Local Standard Time (LST = Universel Temps Coordonné (UTC) + 2 h). The rainfall records cover the period from 1917 to 1995. The rainfall time series used consists of each month's average daily rainfall. To obtain comparable values, this time series was normalized on the basis of the actual number of days in each month.

Using the above time series, two different procedures were followed for structuring the input vectors for training the ANN and cluster analysis models:

(a) each 1-year subset consisting of 12 elements (the monthly average values for the 12 months in each year) formed a case, and
(b) each 2-year subset consisting of 24 elements (the monthly average values for 24 months in 2 consecutive years) formed a case. In this subset, an overlap of 1-year (50%) was applied.

In both (a) and (b) each case represents the temporal distribution of the appropriate meteorological parameter for the respective 1-year or 2-year period. The unsupervised learning self-organizing feature maps ANN algorithm was used to group these cases based on their temporal waveshapes, as described in Section 3. The number of output classes tested in a series of experimental runs carried out in order to classify these cases ranged between 8 and 20. Four different ANN models were trained for classifying the 1-year (12 input) and 2-year (24 input) rainfall distributions; the output nodes (i.e. classes) were set to 8, 12, 16 and 20. The 12 and 24 input rainfall distributions were also used for cluster analysis, as described in Section 4. A dendogram was constructed that was pruned at different levels, which lead to different partitions of the input data.

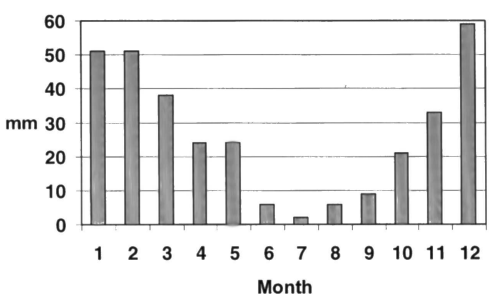

Figure 1. The average monthly distribution of rainfall for Nicosia

3. SELF-ORGANIZING FEATURE MAPS ALGORITHM

Learning in ANN is achieved through systematic training. Training of ANN is a matter of adjusting weights (the strength of a connection) in a systematic manner. Irrespective of weight adjustment, neural network training can take place in one of two ways: supervised or unsupervised. In supervised training, input and output data are supplied to the network in an effort to teach it to produce the desired output vectors. In unsupervised training, data are simply entered into the network without any human intervention or side information. Training is achieved through the formation of internal constructions that capture regularities in their input vectors. In each training method, learning is completed when the network reaches a certain stability criterion. Initially, connecting weights are set to small random values. Input data are then supplied to the network, causing it to pass through state changes that subsequently introduce changes to the values of the weights. Stability is reached when no further weight changes are caused. In effect, a neural network learns through adaptation that is determined by the learning rule of the methodology used.

The self-organizing feature maps algorithm of Kohonen (1990, 1995) was used in this study. The proposed ANN model derived from this algorithm is an one or two dimensional array of neuron-like logic units that are weight-connected to an input pathway where the 12 or 24 element feature vectors, derived in Section 2, describing each pattern are supplied (the ANN architecture used for classifying the 1-year rainfall distributions is given in Figure 2). Input vectors were presented sequentially in time, without specifying the desired output. The presentation of a number of input vectors and the adjustment of weights accordingly will lead to a two-dimensional map, with weights specifying clusters that resemble the input space, such that the point density function of the vector centres tends to approximate the probability density function of the input vectors. The algorithm was implemented using the MATLAB neural networks toolbox, as outlined below (Demuth and Beale, 1994):

- **Step 0.** Initialize weights from N input to M output nodes with small random values. Set the initial radius of the neighbourhood to cover the whole array.
- **Step 1.** Present new input.
- **Step 2.** Compute the Euclidean distance, or an arbitrary norm d_j between the input and each output node j using:

$$d_j = \sum (x_i(t) - w_{ij}(t))^2$$

  where $x_i(t)$ is the input to node i at time t and $w_{ij}(t)$ is the weight from input node i to output node j at time t.
- **Step 3.** Select the output node c with the minimum distance d_j.

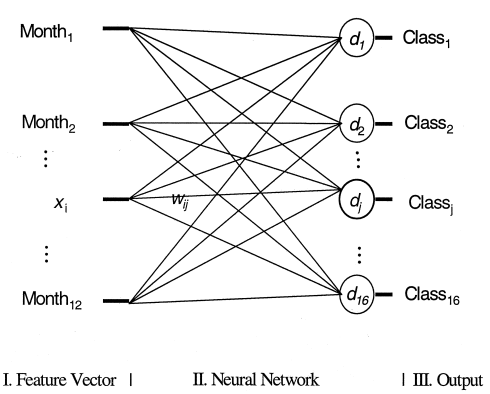

Figure 2. ANN architecture used for classifying the 1-year rainfall distributions

- **Step 4.** Update weight to node c and neighbourhood $N_c(t)$ using:

$$w_{ij}(t+1) = A_1 x_i(t) + A_2 w_{ij}(t)$$

$$j \in N_c(t) \;\; 0 \le i \le N-1 \qquad A_1 = A\left(1 - \frac{t}{T_1}\right) \qquad A_2 = 1 - A_1$$

  with A being a monotonically decreasing gain factor $0 < A < 1$, and T_1 being the period for updating the radius of the neighbourhood.
- **Step 5.** Go to step 1 for a specified number of epochs.

After all the vectors in the training set are presented once at the input, the procedure is repeated many times, with the vectors presented in a different random order each time. This part of the algorithm eventually organizes the weights of the one or two-dimensional map, such that topologically close nodes become sensitive to inputs that are physically similar. Nodes will be ordered in a natural manner, reflecting the different classes of the training set. In the evaluation phase, an input vector was assigned to the output node (winning node) with the closest weight vector. The number of epochs for all models investigated was 1000.

4. AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Clustering is the grouping of similar objects, and the word clustering, according to Hartignan (1975), is almost synonymous with classification. The clustering methods address the problem of unsupervised classification where unlabelled data are grouped into a set of classes according to their similarity relations. There are numerous clustering methods, including hierarchical clustering, splitting methods, merging methods, and others (Duda and Hart, 1973; Hartignan, 1975; Everitt, 1977). Among the best known methods of clustering are the hierarchical clustering procedures, mainly because of their conceptual simplicity (Norusis, 1993). These procedures can be further divided into two distinct classes, agglomerative and divisive. Agglomerative (bottom-up) procedures start with n clusters, and form the sequence by successively merging clusters, whereas divisive (top-down) procedures start with all the samples in one cluster, and form the sequence by successively splitting clusters. The computation needed to go from one level to another is usually simpler for the agglomerative procedures.

In this paper, the agglomerative hierarchical clustering algorithm was used for clustering the 12 or 24 element temperature and rainfall feature vectors. The Statistical Package for Social Sciences—SPSS (Norusis, 1993) was used and the major steps in agglomerative clustering are outlined below, as documented in Duda and Hart (1973):

- **Step 0.** Let $\hat{c} = n$ and $S_{i=0}\{x_i\}, \; i = 1, \ldots, n$.
- **Step 1.** If $\hat{c} \le c$, stop.
- **Step 2.** Find the nearest pair of distinct cluster, say S_i and S_j.
- **Step 3.** Merge S_i and S_j, delete S_i, and decrement $\hat{c}$ by one.
- **Step 4.** Go to step 2.

This procedure was terminated when the number of clusters was one, i.e. $c = 1$. A dendogram was constructed that was cut at various levels which lead to different partitions (i.e. clusters or classes) of the feature vectors investigated. At any level, the distance between nearest clusters can provide the dissimilarity value for that level. In this study, the Euclidean distance was used and the average linkage between groups method was adopted for combining clusters based on the following criterion:

$$d_{\text{avg}}(S_i, S_j) = \frac{1}{n_i n_j} \sum_{x \in s_i} \sum_{x' \in s_j} \|x - x'\|$$

The average linkage between groups method, which is also called unweighted pair-group method using arithmetic averages (Norusis, 1993), defines the distance between two clusters as the average of the

distances between all pairs of cases in which one member of the pair is from each of the clusters. This differs from the other linkage methods, in that it uses information about all pairs of distances, not just the nearest or the furthest. For this reason, it is usually the preferred method of cluster analysis compared with the single and complete linkage methods (Norusis, 1993).

5. ANN MODELLING CLASSIFICATION

In the following, the results for the climatic classification by using the ANN methodology, as described above, are discussed. For the 1-year rainfall data, various ANN models were tested. The distributions were classified by using different ANN models, in which the number of output nodes increase from 8 to 20. It was observed that by increasing the number of output nodes, increasingly less rainfall distributions belong to the various classes. This is because, through increasing the number of output nodes, the respective ANN model is directed towards resolving more detailed characteristics of the rainfall distributions and is, therefore, able to differentiate on the basis of minor characteristics of these distributions. In fact, increasing the number of output nodes beyond a value creates some classes with no distribution membership at all. Also, with a very small number of output nodes, there appears to be a difficulty in the identification of minor characteristics of the various rainfall distributions. In this case, the model recognizes only the crude characteristics of these distributions, and consequently, the respective classification is not offered to differentiate substantially between various classes.

In the present paper, the results of the rainfall distribution classification for 16 output nodes are presented. The number of output nodes was selected as being the largest number of output nodes for which each class contains at least three rainfall distributions. In addition, the number of 16 output nodes was selected by the expert meteorologist, following a qualitative evaluation of the ANN results (bearing in mind the characteristics of the climatic variability in Cyprus). When the number of output nodes is set to the next higher number, i.e. 20, the respective ANN model yields classes that contain two or less rainfall distributions.

The rainfall classification results for the 1-year distributions are shown in Table I. Classes 2 and 13 are the classes with the highest membership, each representing seven 1-year rainfall distributions. In the 79-year period studied here, the most rare classes appear to be those numbered 3 and 5. It is interesting to note that very few classes contain similar distributions for 2 (or more) consecutive years, implying that the chance that 2 consecutive years have similar yearly rainfall cycles is quite small.

Figure 3 displays the prototype 1-year rainfall distributions which were established on the basis of this grouping. Apparently, it is not possible to discuss within the context of the present paper each and every characteristic of all the prototype distributions. However, it is obvious by looking at these prototypes that almost all of them have distinct features that make them different from the others; it is these features that the ANN model was able to discern. For example, class 1 represents distributions with moderate rainfall during wintertime but with relatively higher rainfall in May. In both classes 2 and 3, December is particularly wet; in class 2, however, wet months are also February and March, whereas in class 3, January is also a wet month. Class 3 appears to have common characteristics with class 7, but in the former, the overall rainfall intensity is higher. Class 4 refers to distributions with very wet January and December. Classes 5 and 6 share common characteristics; the major difference is that the amount of November rainfall is greater in class 5. January and December are the wettest months in class 8. Class 9 has a characteristic relative maximum of rainfall in June. In class 10, the 3 first months are wet, whereas the remaining months have very low or no precipitation. Class 11 presents an almost smooth decrease in rainfall between January and June, and an almost equally smooth increase from July till December. In class 12, the wettest months are January and October. Classes 13 and 14 comprise particularly dry years, with maximum rainfall recorded in November. Lastly, classes 15 and 16 display a similar distribution of rainfall, with small differences though. In both of these classes, March appears to be the wettest month.

The statistical characteristics of the prototypes worked out on the basis of the classification are given in Table I. Class 15 has the lowest mean daily rainfall, amounting to 0.597 mm/day, whereas class 4 has

Table I. Classification of the 1-year rainfall distributions for the 16 class ANN and cluster analysis models

Class	Years	*n*	Mean	S.D.	Min	Max	Median
ANN							
1	1920, 1923, 1924, 1963, 1983	5	1.045	1.007	0.000	3.616	0.828
2	1918, 1929, 1936, 1946, 1948, 1964, 1975	7	1.150	1.330	0.000	5.052	0.731
3	1919, 1938, 1949	3	1.141	1.732	0.000	8.997	0.608
4	1926, 1930, 1934, 1944	4	1.339	1.830	0.000	6.039	0.529
5	1921, 1954, 1976	3	1.253	1.125	0.000	3.727	1.044
6	1952, 1961, 1962, 1971, 1979, 1988	6	1.062	1.075	0.000	3.628	0.863
7	1941, 1955, 1956, 1978, 1991	5	0.774	1.092	0.000	4.281	0.275
8	1917, 1947, 1951, 1968	4	1.020	1.095	0.000	4.613	0.596
9	1922, 1928, 1939, 1994	4	1.044	1.272	0.000	4.520	0.341
10	1927, 1931, 1967, 1980, 1981, 1990	6	0.835	1.004	0.000	3.910	0.405
11	1933, 1950, 1953, 1959, 1974	5	0.782	0.814	0.000	2.713	0.554
12	1937, 1940, 1942, 1958, 1965, 1989	6	0.899	1.145	0.000	4.981	0.456
13	1932, 1935, 1945, 1973, 1984, 1986, 1992	7	0.787	0.956	0.000	4.247	0.438
14	1925, 1977, 1982, 1985	4	0.647	0.706	0.000	2.742	0.367
15	1966, 1970, 1972, 1993, 1995	5	0.597	0.603	0.000	2.403	0.435
16	1943, 1957, 1960, 1969, 1987	5	0.891	1.093	0.000	4.432	0.470
Cluster analysis							
1	1920, 1923, 1963, 1983	4	1.060	1.021	0.000	3.616	0.850
2	1925, 1932, 1935, 1945, 1970, 1973, 1984, 1986, 1992, 1995	10	0.697	0.859	0.000	4.247	0.370
3	1917, 1933, 1940, 1941, 1950, 1953, 1955, 1956, 1957, 1958, 1959, 1960, 1968, 1972, 1976, 1977, 1978, 1982, 1985, 1987, 1989, 1991, 1993	23	0.762	0.872	0.000	4.281	0.467
4	1922, 1927, 1928, 1931, 1967, 1974, 1980, 1981, 1990	9	0.847	1.016	0.000	3.910	0.372
5	1962, 1971	2	1.081	0.980	0.000	2.737	0.950
6	1918, 1919, 1924, 1936, 1938, 1946, 1948, 1952, 1961, 1964, 1979, 1988	12	1.046	1.140	0.000	4.706	0.764
7	1939, 1947	2	1.068	1.164	0.000	3.323	0.418
8	1943, 1969	2	1.025	1.363	0.000	4.432	0.458
9	1951, 1966	2	1.054	1.111	0.000	4.613	0.773
10	1942, 1965	2	1.150	1.332	0.000	4.981	0.787
11	1954	1	1.347	1.294	0.000	3.343	1.009
12	1929, 1975	2	1.373	1.623	0.000	5.052	0.632
13	1921	1	1.468	1.335	0.000	3.727	1.379
14	1937, 1994	2	1.100	1.483	0.000	4.742	0.516
15	1926, 1930, 1934, 1944	4	1.339	1.830	0.000	6.039	0.529
16	1949	1	1.365	2.583	0.000	8.997	0.351

See Figure 3 for the 1-year waveshapes.

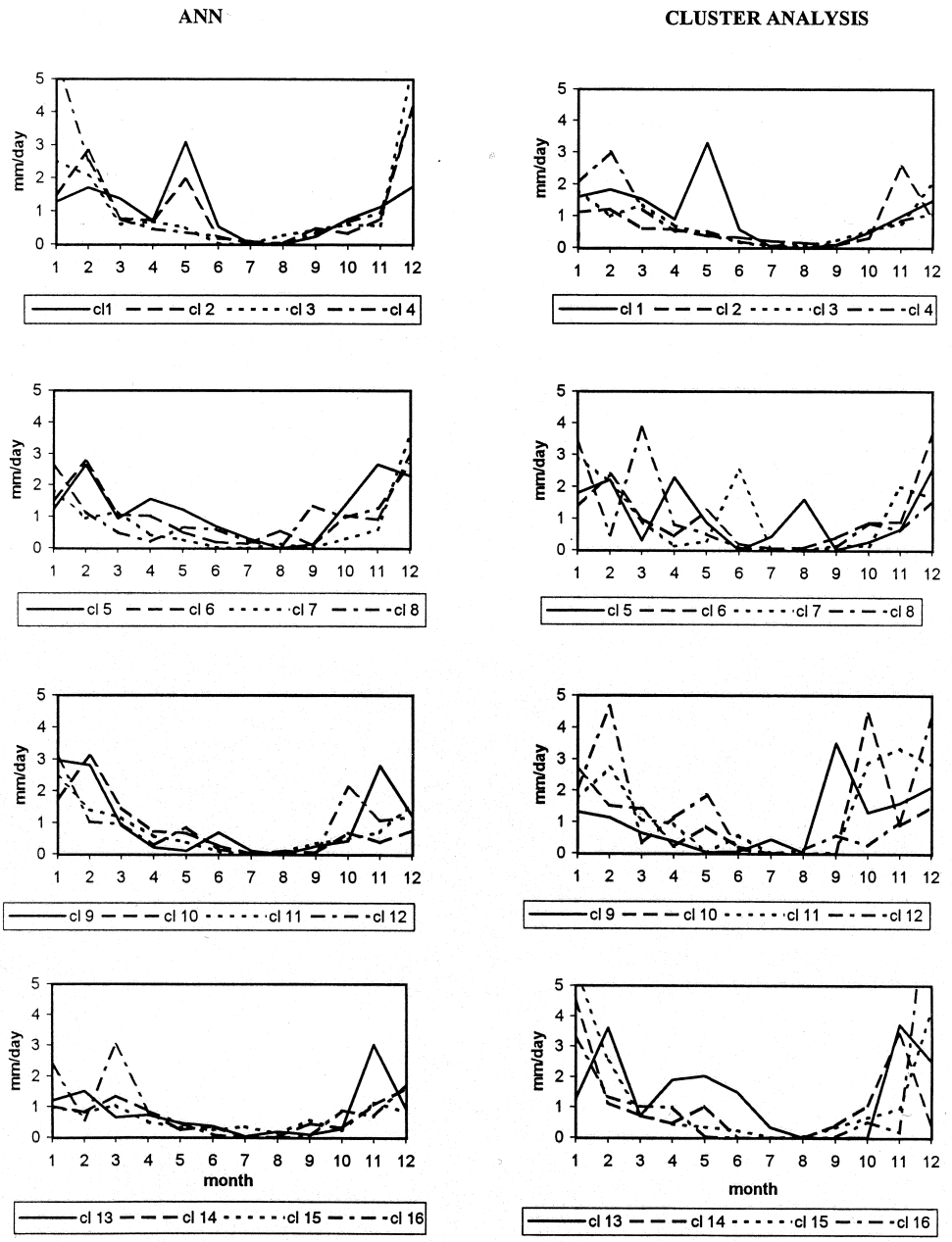

Figure 3. The prototype 1-year rainfall distributions assigned to each class for the ANN and cluster analysis models

the highest mean daily rainfall, amounting to 1.339 mm/day. The highest maximum is associated with class 3. All of the classes have at least 1 completely dry month.

The classification performed for the 2-year rainfall distributions is tabulated in Table II. Class 8 has the highest frequency, containing eight 2-year distributions, whereas classes 6, 8 and 9 have the lowest frequency, with three members each. The possibility that two or more consecutive or overlapping 2-year rainfall distributions share the same pattern is quite small. A similar finding was stated above for the 1-year rainfall distributions.

Table II. Classification of the 2-year rainfall distributions for the 16 class ANN and cluster analysis models

Class	Years	*n*	Mean	S.D.	Min	Max	Median
ANN							
1	1925–1926, 1935–1936, 1937–1938, 1945–1946, 1970–1971	5	0.946	1.114	0.000	6.003	0.617
2	1932–1933, 1966–1967, 1973–1974, 1984–1985, 1986–1987	5	0.798	0.907	0.000	4.247	0.505
3	1921–1922, 1922–1923, 1927–1928, 1939–1940, 1980–1981, 1992–1993	6	0.904	1.044	0.000	3.910	0.370
4	1920–1921, 1931–1932, 1953–1954, 1972–1973, 1982–1983, 1983–1984, 1985–1986, 1993–1994	8	0.944	1.038	0.000	4.520	0.662
5	1940–1941, 1943–1944, 1960–1961, 1963–1964, 1967–1968, 1990–1991	6	0.906	1.161	0.000	6.039	0.479
6	1957–1958, 1976–1977, 1987–1988	3	0.839	0.824	0.000	3.345	0.511
7	1924–1925, 1954–1955, 1961-1962, 1979–1980	4	0.916	1.035	0.000	3.910	0.550
8	1923–1924, 1950–1951, 1965-1966	3	0.999	1.011	0.000	4.613	0.767
9	1933–1934, 1947–1948, 1948–1949	3	1.044	1.568	0.000	8.997	0.373
10	1951–1952, 1955–1956, 1968–1969, 1977–1978	4	0.921	1.111	0.000	4.613	0.441
11	1942–1943, 1959–1960, 1962–1963, 1971–1972, 1988–1989	5	0.877	1.028	0.000	4.981	0.548
12	1958–1959, 1969–1970, 1981–1982, 1989–1990, 1994–1995	5	0.742	1.014	0.000	4.520	0.363
13	1917–1918, 1918–1919, 1928–1929, 1929–1930, 1974–1975	5	1.161	1.385	0.000	5.577	0.585
14	1938–1939, 1946–1947, 1952–1953, 1978–1979, 1991–1992	5	1.007	1.093	0.000	4.281	0.603
15	1919–1920, 1936–1937, 1941–1942, 1949–1950, 1956–1957, 1964–1965, 1975–1976	7	1.017	1.304	0.000	8.997	0.576
16	1926–1927, 1930–1931, 1934–1935, 1944–1945	4	1.083	1.455	0.000	6.039	0.529
Cluster analysis							
1	1918–1919, 1921–1922, 1922–1923, 1924–1925, 1927–1928, 1932–1933, 1935–1936, 1938–1939, 1939–1940, 1940–1941, 1945–1946, 1946–1947, 1947–1948, 1952–1953, 1953–1954, 1954–1955, 1955–1956, 1956–1957, 1957–1958, 1958–1959, 1959–1960, 1960–1961, 1961–1962, 1967–1968, 1968–1969, 1969–1970, 1970–1971, 1973–1974, 1976–1977, 1977–1978, 1978–1979, 1979–1980, 1980–1981, 1981–1982, 1984–1985, 1985–1986, 1986–1987, 1987–1988, 1988–1989, 1989–1990, 1990–1991, 1991–1992, 1992–1993	43	0.872	1.013	0.000	4.432	0.467
2	1919–1920, 1962–1963, 1971–1972, 1975–1976	4	1.022	1.033	0.000	4.286	0.781
3	1917–1918, 1923–1924, 1928–1929, 1937–1938, 1963–1964, 1974–1975	6	1.064	1.202	0.000	5.052	0.760
4	1951–1952, 1966–1967	2	1.045	1.062	0.000	4.613	0.781
5	1920–1921, 1983–1984	2	1.180	1.136	0.000	3.727	0.836
6	1931–1932, 1972–1973, 1993–1994	3	0.788	0.976	0.000	4.520	0.482
7	1982–1983	1	0.831	0.930	0.000	3.616	0.554
8	1950–1951, 1965–1966	2	1.010	1.042	0.000	4.613	0.773
9	1936–1937, 1941–1942, 1964–1965	3	0.969	1.290	0.000	4.981	0.435
10	1994–1995	1	0.857	1.222	0.000	4.520	0.436
11	1925–1926, 1933–1934, 1943–1944	3	1.011	1.496	0.000	6.039	0.382
12	1926–1927, 1930–1931, 1934–1935, 1944–1945	4	1.083	1.455	0.000	6.039	0.529
13	1949–1950	1	1.085	1.881	0.000	8.997	0.644
14	1942–1943	1	1.089	1.382	0.000	4.981	0.548
15	1929–1930	1	1.404	1.770	0.000	5.577	0.635
16	1948–1949	1	1.162	1.983	0.000	8.997	0.453

See Figure 4 for the 1-year waveshapes.

The prototypes for the 2-year rainfall cycles, based on the classification described above, are shown in Figure 4. As before, these prototypes comprise the average distribution, which is formed by taking the mean daily rainfall of all the members in each class. Although the ANN model was able to identify even minor characteristics between the 2-year rainfall distributions, it is not possible to discuss each and every one of these characteristics that differentiate the prototypes. Here, only some distinguishable features are discussed. In both classes 1 and 2, Decembers in the first year appear to be the driest of the winter season; in the second year, December rainfall reaches a maximum in class 1, whereas in class 2, it remains at the same level as in the first year. In class 3, the wettest month is February in the 2 consecutive years. Class 4 is characterized by a maximum in rainfall during November of the second year. Classes 5 and 6 share many common characteristics, as far as their broad features are concerned. In classes 7, 8 and 11, rainfall is quite erratic throughout the 2-year period. Class 9 is marked with an exceptionally wet December of the second year. In class 10, the rainfall maxima are observed in December of each of the 2 years comprising the time period studied. In class 12, the wettest month is January in the first year, with the rest of the period being characterized by moderate amounts of rainfall. Class 13 exhibits two peaks in May for both years. In classes 14 and 15, the maximum rainfall is noted in December of the first year. Class 16 displays an exceptionally wet January of the first year, whereas its counterpart of the second year is markedly less wet.

Table II displays the statistical characteristics of the 16 prototypes for the 2-year rainfall cycle. Class 13 represents the wettest 2-year cycles, and class 12 the driest. The lowest and highest standard deviations are associated with class 6 and class 9, respectively. Also, classes 9 and 15 are associated with the highest maximum mean daily rainfall in any month, and classes 3 and 7 are associated with the lowest maximum.

6. CLUSTER ANALYSIS CLASSIFICATION

In order to compare the above-presented ANN classification to a traditional statistical classification technique, it was decided to perform such a classification adopting the cluster analysis method. The number of classes at which cluster analysis was carried out was maintained as with the ANN classification. Maheras and Arseni-Papadimitriou (1984) have performed a classification of rainfall records in Athens by using the cluster analysis method. They have analysed a 120-year period by adopting 14 classes. They have also performed an interpretation of these classes regarding the temporal characteristics of the respective rainfall distributions.

The results of the cluster analysis are placed in the same context as the ANN results (Tables I and II and Figures 3 and 4). It must be stressed at this point that the characteristics of the classes established by using cluster analysis should not necessarily be similar to the respective classes established by ANN. Apparently, the characteristics of each class are determined by the distributions that belong to it, with each of the two methods grouping the distributions following a different procedure, as described in Section 3, for ANN, and in Section 4, for cluster analysis.

The cluster analysis classification of the 1- and 2-year rainfall distributions are presented in Tables I and II, respectively, together with the statistical characteristics of each class. Unlike the ANN classification, the cluster analysis classification revealed that the large majority of the 1-year distributions fall into just six classes, with three classes having four members, whereas the remaining classes have at most two members. The same feature of the cluster analysis classification, placing most of the distributions into only a few classes, is also recognized for the 2-year rainfall distributions. Indeed, from Table II, almost half of the distributions are identified with class 1, with most of the other classes containing a single distribution. The prototype 1- and 2-year distributions revealed by cluster analysis are plotted in Figures 3 and 4, respectively.

Cluster analysis was carried out adopting the average linkage between groups method for combining clusters, as described in Section 4. The following criteria for combining clusters were also investigated

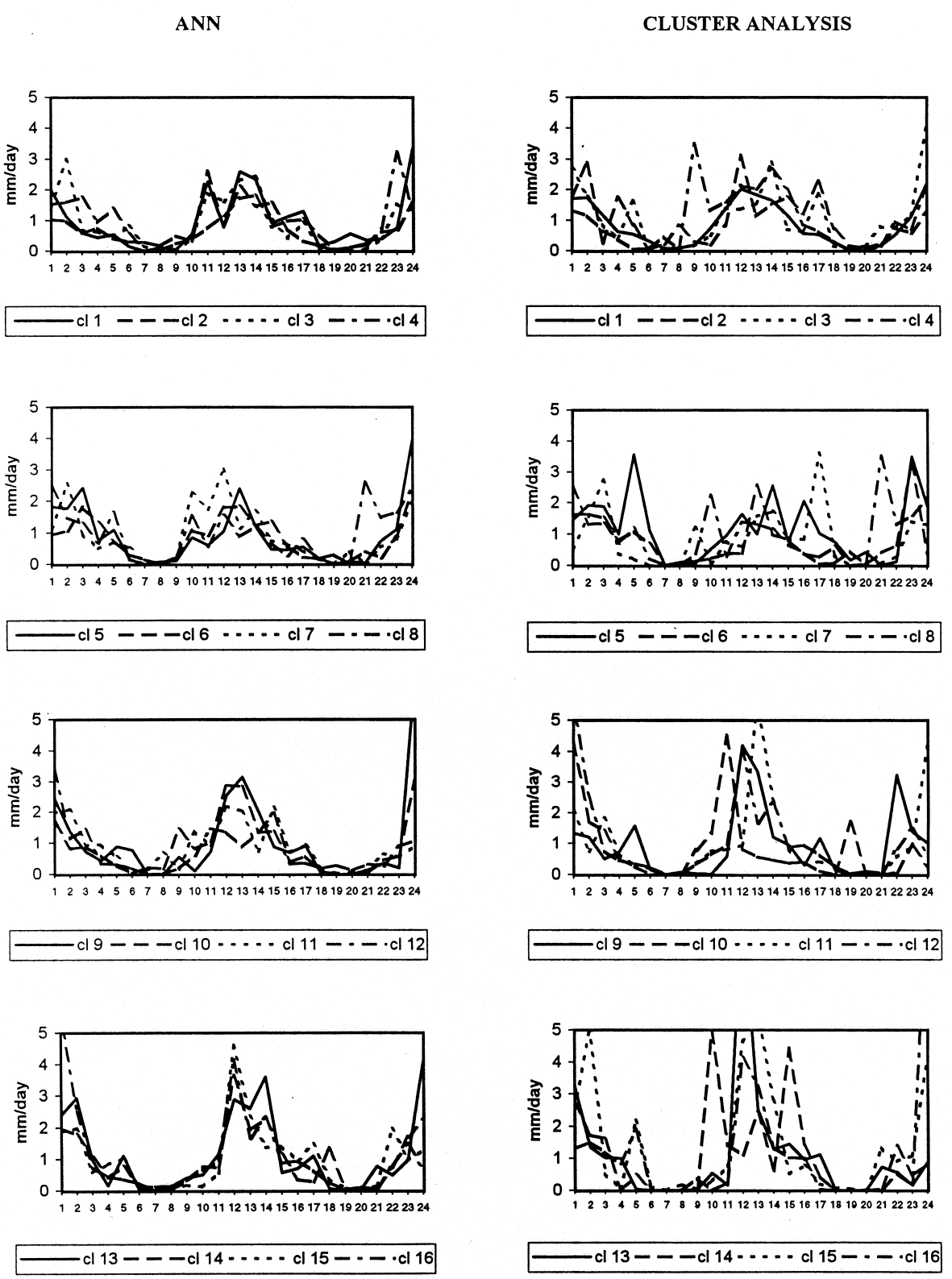

Figure 4. The prototype 2-year rainfall distributions assigned to each class for the ANN and cluster analysis models

as these are implemented in the SPSS package (Norusis, 1993): Ward's, median, nearest neighbour and furthest neighbour. These methods gave similar findings to the ones tabulated in Tables I and II, based on the average linkage between groups method.

From the discussion above, it is obvious that the ANN and the cluster analysis classifications differ primarily in their ability to identify minor characteristics of the distributions. The ANN classification categorizes the distributions more uniformly, whereas the cluster analysis groups most of the distributions into a few distinctive classes with the remaining of the distributions behaving as outliers. For example, cluster analysis findings, model 12, has 43 members, whereas models 7, 10, 13, 14, 15 and 16 have one member each (see Table II). Furthermore, bearing in mind the form of the climatic variability of Cyprus, the cluster analysis classification cannot be considered as adequate.

7. VALIDATION OF THE ANN AND CLUSTER ANALYSIS CLASSIFICATION

In order to validate the classification procedure described above, the rainfall distributions for the more recent years of 1996, 1997 and 1998 were utilized. For this task, the respective distribution was undergone through a series of comparative tests against the established prototypes assigned to each class. The categorization of each distribution was subsequently determined on the basis of the squared error. In particular, a distribution was assigned to a specific class, against which the sum of square error (SSE) is found to be minimal. For each of the ANN and cluster analysis procedures, the 1- and 2-year distributions for the validation years were assigned to particular classes, and the results are shown in Table III.

The difficulty of the cluster analysis classification to categorize discretely the rainfall distributions for the 3 validation years is also seen in Table III. Indeed, the ANN classification was able to differentiate adequately between the three 1-year and the three 2-year distributions. On the contrary, when applying the cluster analysis classification, the categorization for the 1-year distributions assigns all the validation years to a single class, namely, class 3. Similarly, all the 2-year validation distributions are indiscriminately assigned to a single class, namely, class 1.

The results of the above validation support the conclusion reached in Section 6, which is that the cluster analysis classification tends to accumulate most of the distributions in just a few classes, whereas the ANN classification tends to categorize the distributions more uniformly between the various classes. We may generalize by saying that the cluster analysis validation has categorized the validation years as belonging to those respective classes that have the highest distribution membership. The bias of the cluster analysis classification towards particular cloudily described distributions that was revealed in Section 6 is strongly supported by the above validation results.

Table III. Validation of the ANN and cluster analysis models for rainfall

	Year(s)	ANN		Cluster analysis	
		Class	SSE	Class	SSE
1	1996	11	0.873	3	1.029
2	1997	14	2.383	3	4.247
3	1998	14	1.979	3	2.481
4	1995–1996	2	10.063	1	12.205
5	1996–1997	12	6.588	1	5.825
6	1997–1998	6	6.857	1	6.246

SSE gives the sum of squared difference between the prototype distribution and the feature vector of the year(s) under evaluation.

8. CONCLUDING REMARKS

Using ANN, the study's objective of identifying and classifying similar rainfall distributions has been met. A simple procedure was introduced in order to group similar patterns belonging to the same class in relation to the temporal distribution of rainfall. With this technique, years belonging to the same class are considered to be 'similar' (with regard to the temporal distribution of the particular climatological element) making no reference to the statistical characteristics of its distribution; rather, the temporal distribution is treated by the ANN as a single 'element'. The same classification procedure was adopted for the 2-year subsets in the time series, with 1 year overlapping in two consecutive 'elements'.

From the examination of the results, of a series of different ANN models with varying class numbers, a number of 16 classes for the rainfall regime appeared to be the optimum for further detailed analysis and discussion. The criteria used for the adoption of these classifications of the rainfall distributions were empirically established, but there appears to be a relationship between the number of input nodes used and the largest number of output nodes, for which a reasonable classification can be achieved.

The analysis presented above has shown that ANN can successfully be used to identify and classify similar patterns of rainfall. The ANN models trained were capable of detecting even minor characteristics and differentiating between various classes. Comparison of the ANN and cluster analysis methods (Section 6), together with the validation evidence (Section 7), indicate that the ANN classification procedure is superior to the cluster analysis classification because the former has put more refinement into the respective classification, and performed a more realistic categorization of the available data.

Ben-Gai *et al.* (1994) discuss long-term rainfall variability over Israel, and ascribe these changes to both local land use management and global-scale forcing associated with an increase of sea surface temperature. Although it is beyond the scope of this paper to interpret the observed variability, it should be mentioned here that the variability detected in this research could also have its causes at both the local-scale and global-scale.

The authors believe that the classification reached by using ANN can become a valuable tool in the study of climatic variability. It provides a unique measure of the dimension of the inherent variability of the weather that is attributed to the temporal evolution of the parameter in study. This dimension is often overlooked by standard statistical approaches.

A possible extension of the ANN classification discussed in this study could be in the domain of forecasting medium and long-term trends in the rainfall regime. In this respect, current measurements completed for only part of the year (e.g. for the first 3 months) can be contrasted to the respective part of the available prototypes. Further investigation involves whether the task of matching these current measurements to the prototype patterns could give a (probabilistic) forecast of the pattern of the distribution for the remaining time period.

Work in progress, also, includes experimentation with more climatological parameters, and with longer cycles. Experimentation with longer cycles than the ones used in the present study could reveal the existence of a long suspected definite or preferred cycle in rainfall (Xanthakis, 1973). It must be emphasized, however, that in such an endeavour, it is essential that the input data comprise a sufficiently long time-series of the respective meteorological parameter.

ACKNOWLEDGEMENTS

The database used in the present analysis was retrieved from the archives of the Meteorological Service of Cyprus.

REFERENCES

Arbib MA. 1995. *The Handbook of Brain Theory and Neural Networks*. The MIT Press: Cambridge, MA.

Bankert RL. 1994. Cloud classification of a AVHRR imagery in maritime regions using probabilistic neural network. *Journal of Applied Meteorology* **33**: 909–918.

Ben-Gai T, Bitan A, Manes A, Alpert P. 1994. Long-term changes in annual rainfall patterns in southern Israel. *Theoretical and Applied Climatology* **49**: 59–67.
Bulsari AB, Kallio S, Tsaptsinos D. 1996. *Solving Engineering Problems with Neural Networks*. Abo Akademis Tryckeri: Turku, Finland.
Demuth H, Beale M. 1994. *MATLAB Neural Network Toolbox User's Guide*. The Maths Works, Inc.: Natick, MA.
Duda RO, Hart PE. 1973. *Pattern Classification and Scene Analysis*. Wiley: New York.
Elsner JB, Tsonis AA. 1992. Nonlinear prediction chaos and noise. *Bulletin of the American Meteorological Society* **73**: 49–60.
Everitt B. 1977. *Cluster Analysis*. Heineman Educational Books: London, UK.
Francl LJ, Panigrahi S. 1997. Artificial neural network models of wheat leaf wetness. *Agricultural and Forest Meteorology* **88**: 57–65.
Furlan D. 1997. The climate of southeast Europe. In *Climates of Central and Southern Europe*, vol. 6, Wallen CC (ed.). Elsevier Scientific Publishing Co.: Amsterdam; 185–235.
Furundzic D. 1998. Application example of neural networks for time series analysis: rainfall-runoff modelling. *Signal Processing* **64**: 383–396.
Hartignan JA. 1975. *Clustering Algorithms*. Wiley: New York.
Haykin S. 1994. *Neural Networks—A Comprehensive Foundation*. Macmillan College Publishing: New York.
Kalogirou SA, Neocleous CC, Michaelides SC, Schizas CN. 1997. A time series reconstruction of precipitation records using artificial neural networks. In *Proceedings of the Fifth European Congress on Intelligent Technologies and Soft Computing—EUFIT '97*, vol. 3, Zimmermann HJ (ed.). Aachen: Germany; 2409–2413.
Kalogirou SA, Neocleous CC, Michaelides SC, Schizas CN. 1998. Regeneration of isohyets by considering landscape configuration using artificial neural networks. In *Proceedings Fourth International Conference on Engineering Applications of Neural Networks*, Gibraltar, Bulsari AB, Fernadez de Canete J, Kallio S (eds); 383–389.
Kohonen T. 1990. The self-organizing map. *IEEE Proceedings* **78**: 1464–1480.
Kohonen T. 1995. *Self-Organizing Maps*. Springer Series in Information Sciences: Berlin.
Kutiel H. 1987. Rainfall variations in the Galilee (Israel). I—Variations in the spatial distribution in the periods 1931–1960 and 1951–1980. *Journal of Hydrology* **94**: 331–344.
Kutiel H. 1988. Rainfall variations in the Galilee (Israel). II—Variations in the temporal distribution between 1931–1960 and 1951–1980. *Journal of Hydrology* **99**: 179–185.
Lorenz EN. 1969. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences* **26**: 636–646.
Luo FL, Unbehauen R. 1997. *Applied Neural Networks for Signal Processing*. Cambridge University Press: New York.
Maheras P, Arseni-Papadimitriou A. 1984. Application of the methods of factorial analysis for the study of the organization of the rhythm of the monthly precipitation values in Athens. *Zeitschrift fur Meteorologie* **34**: 100–105.
Marzban C, Stumpf GJ. 1996. A neural network for tornado prediction based on doppler radar-derived attributes. *Journal of Applied Meteorology* **35**: 617–626.
Micheli-Tzanakou E. 1995. Neural networks in biomedical signal processing. In *The Biomedical Engineering Handbook*, Bronzino JD (ed.). CRC Press and IEEE Press: New York; 917–932.
Nicolis C. 1998. Atmospheric analogues and recurrence time statistics: toward a dynamical formulation. *Journal of Atmospheric Sciences* **55**: 465–475.
Norusis MJ. 1993. *SPSS for Windows Professional Statistics Release 6.0*. SPSS Inc.: Chicago, IL.
Peak JE, Tag PM. 1992. Toward automated interpretation of satellite imagery for navy shipboard applications. *Bulletin of the American Meteorological Society* **73**: 995–1008.
Sanchez-Sinencio E, Lau C. 1992. *Artificial Neural Gibraltar Networks*, *Paradigms*, *Applications and Hardware Implementations*. IEE Press: New York.
Schizas CN, Pattichis CS, Michaelides SC. 1994. Forecasting minimum temperature with short time-length data using artificial neural networks. *Neural Network World* **2**: 219–230.
Taha MF, Harb SA, Nagib MK, Tantawy AH. 1981. The climates of the Near East. In *Climates of Southern and Western Asia*, *World Survey of Climatology*, vol. 9, Takahashi K, Arakawa H (eds). Elsevier Scientific Publishing Co.: Amsterdam; 183–255.
Tangang FT, Hsieh WW, Tang B. 1998. Forecasting regional sea surface temperatures in the tropical Pacific by neural network models, with wind stress and sea level pressure as predictors. *Journal Geophysical Research* **103**: 7511–7522.
Van den Dool HM. 1989. A new look at weather forecasting through analogues. *Monthly Weather Review* **117**: 2230–2247.
Xanthakis J. 1973. *Solar Activity and Precipitation*. Springer: New York.

