BOREAL ENVIRONMENT RESEARCH 7: 301–304 ISSN 1239-6095
Helsinki 2 October 2002

Selection of representative stations by means of a cluster analysis for the BAMAR region in the PIDCAP period

Hermann Oesterle

Potsdam Institute for Climate Impact Research (PIK), Telegrafenberg, P.O. Box 601203, D-14412 Potsdam, Germany

Oesterle, H. 2002: Selection of representative stations by means of a cluster analysis for the BAMAR region in the PIDCAP period. — *Boreal Env. Res.* 7: 301–304. ISSN 1239-6095

A cluster analysis procedure is presented that allows to select stations that are typical for different meteorological conditions. The estimations of mean daily minimum and maximum air temperatures, and also mean daily precipitation were used as parameters of the classification procedure. The means are calculated using verified synoptic data which are collected in the course of interval from August to October 1995 at 950 synoptic sites allocated on the Baltex Model Area (BAMAR) region. As a result 83 representative stations are selected, and used for validation of the LM-model at PIK.

Introduction

It is more advisable to use real data for comparative validation of models with different resolution than grid data from objective analyses. For the purpose of such investigations, it is desirable to have a set of stations which represent all climatic zones of each sub-region. The cluster analysis procedure allows to select representative stations for these zones. Synoptic data were used to identify such stations for the BAMAR region in the PIDCAP period (August–October 1995). Mean values of daily minimum and maximum air temperature and daily precipitation were computed for the PIDCAP period using verified data. Computations were made for stations with data availability of more than 67%.

Cluster analysis procedure

A cluster analysis procedure was used as described in Lund (1963) and others (Oesterle and Shapovalova 1974, Oesterle and Enke 1994, Werner *et al.* 2001). The essence of this analysis is as follows:

For each pair of k available objects, in our situation for each pair of stations, X_m and X_l, the difference between the objects is computed according to the formula

$$d(X_m - X_l) = \sum_{i=1}^{n} \mathrm{abs}(X_{mi} - X_{li}) \qquad (1)$$

where n is the objects' dimensionality, $\mathrm{abs}(X_{mi} - X_{li}) = 0$ if the value is not larger than the threshold value db_i, and $\mathrm{abs}(X_{mi} - X_{li}) = 1$ if the value is

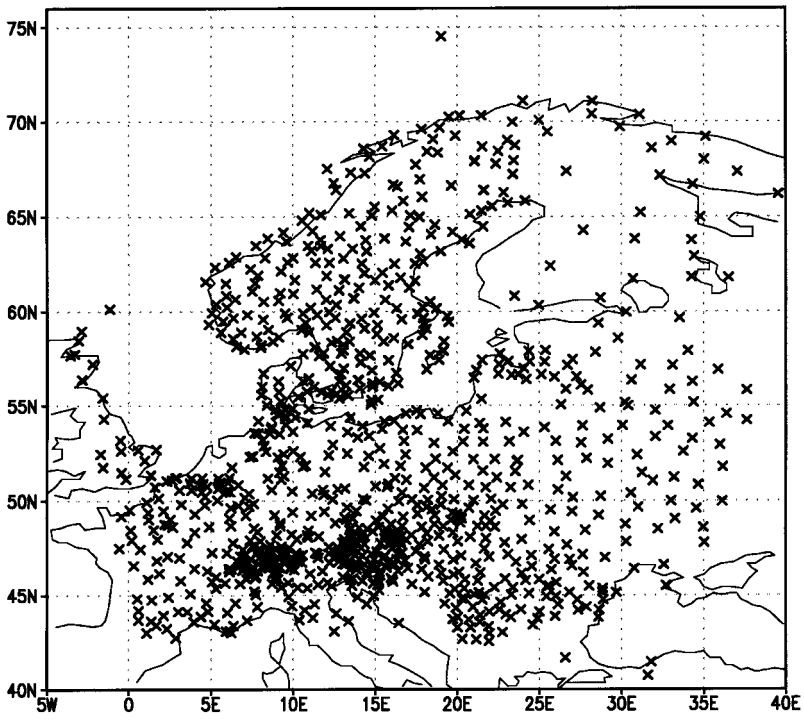

Fig. 1. 950 stations (x) with temperature and precipitation data of sufficient quality used in the cluster analysis for the PIDCAP period.

larger than db_i which is given for each component describing the vector. Thus we have normalized values for different components of the distance vector for calculation of the distances matrix d_k. A threshold value of likeness defines a maximal hyper-sphere around each object. A rank is defined as the number of objects that belong to the hyper-sphere (the number of "neighbours"). Any tested object, which has the highest rank is declared to be the standard object for the first cluster.

In the next steps, the standard object and its "neighbours" will be deleted from the matrix, and the procedure will be repeated for the remaining objects. The procedure will be stopped if it is not possible to find more than two "neighbours".

After the selection of standard objects, all other objects are classified according to the minimum-distance to them. This second step is necessary because in the first step any object was included in the previous cluster although they were nearer to other standard objects.

Then the classification will be improved. Therefore, mean values were computed for each component in each cluster, and all objects are exposed to the classification on the principle of the minimum-distance to these "mean standard objects" or "cluster centroids". Stations with a minimum distance to the cluster centroids were selected as representative stations (standard stations) for each cluster.

Classification results

From 1560 stations available (obtained from the Meteorological Data Centre for BALTEX at the Deutscher Wetterdienst), 950 stations were used for which it was possible to compute mean values for the entire time period (*see* Fig. 1). Other stations had less than 67% of complete data to compute mean values for temperature and precipitation.

These stations were put in the classification procedure. The mean values of minimum and maximum air temperature, mean daily precipitation for the time period and latitudes and longitudes of the stations are the parameters that were used to describe the stations. Co-ordinates of the stations were used to select representative stations for each sub-region.

In the experiments, the threshold value db_i was 2 °K for temperature, 1 mm for precipitation

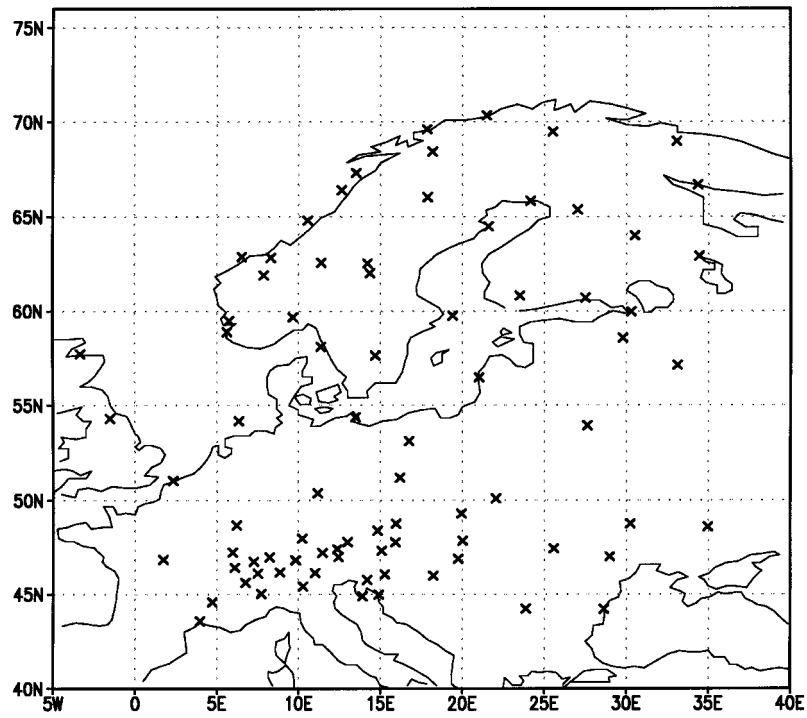

Fig. 2. 83 standard stations (x) were selected by the classification method which used temperature, precipitation and co-ordinates of stations as parameters.

and 5° for co-ordinates. First, a classification was done on the basis of temperature only. In this case, 21 clusters were selected for all regions. On the basis of temperature and co-ordinates, 57 clusters were selected, and on the basis of temperature, precipitation and co-ordinates, the most complete classification of 83 clusters was obtained (*see* Fig. 2).

The highest number of clusters was selected for the mountains (Alps) and coastal-mountains (Norway) regions. In the window 45–50°N and 5–20°E (~Alps) 28 clusters were selected as an example, where as in its neighbouring, rather flat sub-region between 50–55°N and 5–20°E, only 5 clusters were selected.

In the analysis, the variation of the values at the standard stations in the Alpine sub-region was:

— for max. temperature: from 5.6 °C (WMO-number 6780, elev. 1828 m) to 23.4 °C (WMO-number 16088, elev. 97 m);
— for min. temperature: from 0.8 °C (WMO-number 6780, elev. 1828 m) to 13.7 °C (WMO-number 16088, elev. 97 m);
— for precipitation: from 42 mm/month (WMO-number 16061, elev. 710 m) to 165 mm/month (WMO-number 6702, elev. 1675 m).

In the second sub-region with a less structured orography, these differences were less substantial:

— for max. temperature: from 16.2 °C (WMO-number 10558, elev. 630 m) to 19.9 °C (WMO-number 12415, elev. 124 m);
— for min. temperature: from 9.0 °C (WMO-number 12230, elev. 73 m) to 13.7 °C (WMO-number 10113, elev. 13 m);
— for precipitation: from 18 mm/month (WMO-number 12415, elev. 124 m) to 57 mm/month (WMO-number 10558, elev. 630 m).

Conclusions

The approach allows to select stations from a large number of stations which are typical for regions with different meteorological regimes. These stations have basic information available for describing temperature and precipitation fields and can be considered as representative stations

used for validation of different models. All these stations were used for validation of the LM-model at PIK.

References

Lund I.A. 1963. Map-pattern classification by statistical methods. *J. Appl. Meteorol.* 2: 56–65.

Oesterle H. & Shapovalova V. [Оестерле Х. & Шапо-валова В.] 1974. [Study of some climatological characteristics of precipitation in Central Asia for short-term weather forecasting]. *SANIGMI-Report* No. 11(92): 54–66. [In Russian].

Oesterle H. & Enke W. 1994. Zeitreihenanalyse von Temperatur und Niederschlag der letzten 100 Jahre für ausgewählte Stationen in Osteuropa. *PIK-Report* No. 1: 202–203.

Werner P.-C., Gerstengarbe F.-W., Fraedrich K. & Oesterle H. 2000. Recent climate change in the north Atlantic/European sector. *Int. J. Climatol.* 20: 463–471.

Received 23 January 2002, accepted 10 April 2002