Tellus (2005), 57A, 35–42
Printed in UK. All rights reserved

TELLUS

A new validation scheme for the evaluation of multiparameter fields

By FRIEDRICH-WILHELM GERSTENGARBE*, MARTIN KÜCKEN and PETER C. WERNER,
Potsdam Institute for Climate Impact Research Telegrafenberg, PO Box 601203, D-14412 Potsdam, Germany

(Manuscript received 30 April 2003; in final form 19 May 2004)

ABSTRACT

On the basis of an extended cluster analysis algorithm, we present a new validation method for the evaluation of simulation experiments characterized by more than one parameter. This method allows the assessment of any parameter combination in space and time. As an example for the effectiveness of the algorithm, the results of two regional climate model runs and observational data have been tested and interpreted.

1. Introduction

Models to describe complex systems have become more and more detailed. A problem that has been recently discussed is the evaluation of such models especially when they are very complex. The very complexity of such models means that severe limits are placed on our ability to analyse and understand the model processes, interactions and uncertainties. In general, it is easy to define the error in simulations of single variables in climate models, for instance by the 'Taylor diagram' (IPCC 2001). However, the validation of coupled variables or processes is a problem that has not yet been solved satisfactorily. In IPCC (2001, p. 474) it was pointed out that "while we do not consider that the complexity of a climate model makes it impossible to ever prove such a model 'false' in any absolute sense, it does make the task of evaluation extremely difficult ...." To make progress in this field is the outline of the paper. Therefore, a tool is presented which can be used to validate the complex behaviour of the model results (Section 2). The basis of this tool is an extended non-hierarchical cluster analysis (see Appendix A) and a new error handling which permits the detection of error sources. The application of these methods shows how far this tool can deliver more information on the quality of climate model outputs (Section 3).

2. Methods

The question that should be answered by the new method is the following: how well can a model represent complex parameter combinations of a regime?

*Corresponding author.
e-mail: gerstengarbe@pik-potsdam.de

To answer this question the following working steps are necessary.

1. Definition of a complex relation

The quality of a climate model is primarily based on the correct determination of the interactions between individual meteorological parameters. Thus, temperature and precipitation, for instance, describe the thermic-hygric situation. Therefore, it is necessary to determine which physical situation should be described in the model.

2. Selection of meteorological parameters

Meteorological parameters required to investigate the complex relation are selected. This is done for a reference data set (for instance, observed data of meteorological stations) as well as for a model data set (grid data), the quality of which should be verified by the reference data set. It is required that the position of the grid and the time period correspond with each other.

3. Derivation of characteristic parameters

Each meteorological parameter can be described more precisely by a set of its characteristics $\boldsymbol{p}$, for instance, maximum, minimum, sum, mean value and mean variance, etc., the selection of which is determined by the specific purpose of the investigation. Should the mean conditions be tested, mean value and sum are to be chosen, whereas maximum and minimum are chosen for extreme value observations. This means that these characteristic parameters are to be calculated for the reference data set as well as for the climate model.

4. Extended cluster analysis

The patterns that are typical for the reference data set will be determined with the aid of the extended cluster analysis (see Appendix A). These patterns are the basis for the comparison of the model results with the reference data set.

5. Validation I

In a first validation step, we verify for each grid point of the model data set whether the respective parameter combination corresponds with the one of the respective station in the reference data set, i.e. whether there is a cluster correspondence. If this is not the case, we search for the cluster to which the grid point can best be suited. The classification is done by using a distance measure, here the Euclidean distance. The Euclidean distance is used to relate the parameter combination of the model data set to the cluster of the reference data set that has the smallest distance to the group centroid (see Appendix A). No error exists if the cluster number of a simulated field grid point agrees with the cluster number of the respective station within the reference data set. If they do not correspond, an error handling is necessary as follows. Let us define the different clusters of the station of the reference data set and of the grid point of the simulated field with a and b. As a simple rule, it can be said that the more the pair of clusters a and b are separated from each other, the greater the error is. So the number of overlaps $O_{a,b}$ can be used to estimate this error. Thus, we proceed as follows. We calculate the ratio $R_{a,b}$ from the current number of overlaps and the maximum number of overlaps between two clusters a and b of the reference data set:

$$R_{a,b} = O_{a,b}/O_{a,b}^{\max}. \tag{1}$$

We calculate the maximum ratio $R^{\max}$ of all possible combinations of $R_{a,b}$:

$$R^{\max} = \max(R_{a,b}) \text{ with } a, b = 1, \cdots, K \text{ with } a \neq b. \tag{2}$$

We calculate the standardized relative ratio $Q^{\text{norm}}{}_{a,b}$ for the detection of errors with changing clusters:

$$Q_{a,b}^{\text{norm}} = 100 \times (1 - R_{a,b}/R^{\max}). \tag{3}$$

Error classes can be determined using $Q^{\text{norm}}{}_{a,b}$ as is specified in Table 1.

6. Validation II

If a grid point of the model data set has a different cluster classification than the respective station of the reference data set, it has to be clarified to what extent the individual meteorological parameters differ from each other. Let us consider the situation that the station is assigned to cluster a according to the reference data and to a different cluster b according to the model data. In this case we can determine what parameters are mainly responsible for this error. They will be the ones with the smallest Euclidean distance to the cluster centroid of b, i.e. with the largest distance to the cluster centroid of a. The standardized ratios of the distances are a measure of how much the individual parameters contribute to the error. A complete overview of all working steps can be found in the flow diagram of Fig. 1.

Table 1. Definition of error classes

Error class	$Q_{\text{norm}}^{a,b}$	
0.00	≥0%	<5%
0.05	≥5%	<10%
0.10	≥10%	<25%
0.25	≥25%	<50%
0.50	≥50%	<75%
0.75	≥75%	<90%
0.90	≥90%	<95%
1.00	≥95%	≤100%

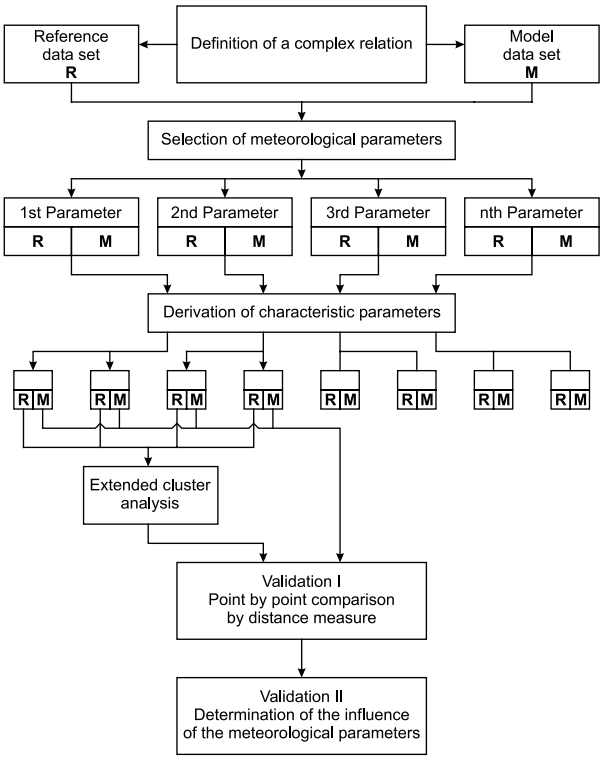

Fig 1. Validation scheme.

3. Application of the method to the validation of climate simulations

3.1. Model and data base

To test the method, two different simulation runs of a regional climate model were used. The first run was carried out with the

basic version of the non-hydrostatic model (M1) of the German Weather Service (DWD; Doms and Schättler, 1999) and the second one with the same model but with a parametrization of the cloud ice (M2) (Schättler and Doms, 2000). The cloud ice scheme is designed for applications using a horizontal grid spacing of about 3–50 km to take microphysical processes in stratiform mixed-phase and ice clouds into account. The experiments are carried out in a non-hydrostatic mode with a resolution of 7 km and with 40 layers in vertical. The model domain (1/16° horizontal resolution) covers the region from 3°W to 53°E and from 43°S to 68°N. DWD analysis data from August to October 1995 were used as initial and boundary conditions. The observation data of 372 weather stations in an area from 9°W to 16°E and from 50°S to 55°N were used as a reference data set. This area was selected for the investigation of water budget components and covers the catchment of the River Elbe.

As meteorological variables, the near-surface temperature and precipitation in combination were selected to describe the thermo-hygric complex. These near-surface parameters were chosen because both are of fundamental importance for the evaluation of the climate regime. The derived parameters are the monthly mean 2-m temperature (t2m) and the monthly sum of daily precipitation (prec). The model variables are staggered on an Arakawa-C/Lorenz grid (Arakawa and Lamb, 1981) with scalars defined at the centre of a grid box (prec and t2m) and the normal velocity components defined on the corresponding box faces. For our validation, we assume that the spatial differences between the centre of a model cell and the actual position of the validation station can be neglected.

3.2. Tasks and execution

During the validation, two questions have to be answered.

(1) How large is the difference between simulated results and observed data (reference field) averaged over the investigation period of three months?

(2) How strong is the influence of the two meteorological parameters on the difference between simulation and observation?

A simulation run with a duration of three months was carried out for each model. The statistical parameters were estimated for each month and each grid point. For the cluster calculation (validation I), we used two parameters per month (monthly mean of the air temperature and the monthly sum of precipitation), a total of six parameters for the whole period of three months. To answer the first question, the simulation runs of M1 and M2 were compared grid point by grid point with the cluster results of the reference case (observational data) for the whole period. This comparison between simulation and observations gives an idea of the existing cluster shifts. A comparison of the differences between the model versions provides an answer to question 1. To solve question 2, the influence of each meteorological parameter has to be calculated as described in Section 2 (validation II).

3.3. Results

The results discussed below are not only intended as an assessment of the quality of any specific climate model but as an example of how well the newly developed method may be applied.

Figure 2 shows the cluster distribution for the observed data. 12 clusters were separated, i.e. all stations with the same cluster number showing the same statistical characteristics.

Figure 3a shows the deviations between the simulation carried out with M1 and the observations according to the error classes defined in Table 1. In the ideal case (when the grid points of simulation and observations are in the same cluster) the simulation exactly represents the mean conditions of the meteorological variables and only error class 0 (white) occurs. One can immediately see that errors of different degrees occur (shading from light to dark green) if a shifting of cluster affiliation exists. The spatial error structure provides information about what may possibly have caused them.

The calculations of the M2 version were carried out in the same way. The results are shown in Fig. 3b. It is obvious that the results differ in comparison to Fig. 3a.

Table 2 provides an overview of the mean values of the error deviations for the validation area before and after activating the cloud ice parametrization. The table shows that, in M1, precipitation and near-surface temperature could only identically be reproduced (error class 0) for approximately 20% of all stations whereas the number of stations for which this has not at all been possible (error class 1) is twice as high. The model's insufficient calculation of the monthly precipitation amount is the reason why the majority of the grid cells do not or only partially correspond with the observations. These statements are also true for M2. The cloud ice parametrization used here does not lead to an improvement of the model quality for all grid cells.

Table 3, however, shows a more differentiated picture of the efficiency of the cloud ice parametrization. At 113 out of 372 stations, a shift in the error class can be seen (improvement, 49; deterioration, 64). Furthermore, at 40 of these stations, there is not only a shift in the error class but a complete reversal of the error class from 0 to 1, or vice versa. In most of these cases, precipitation is responsible for this complete reversal (improvement, 12; deterioration, 18). This fact is not amazing because cloud ice parametrization should be able to improve the precipitation flow in general. However, the used parametrization not only selectively affects model grid points with precipitation deficiencies but in summary also model grid points with an already stable precipitation balance in M1.

Figures 4a and 4b show the spatial structures of the influence of each meteorological parameter on the error for model versus

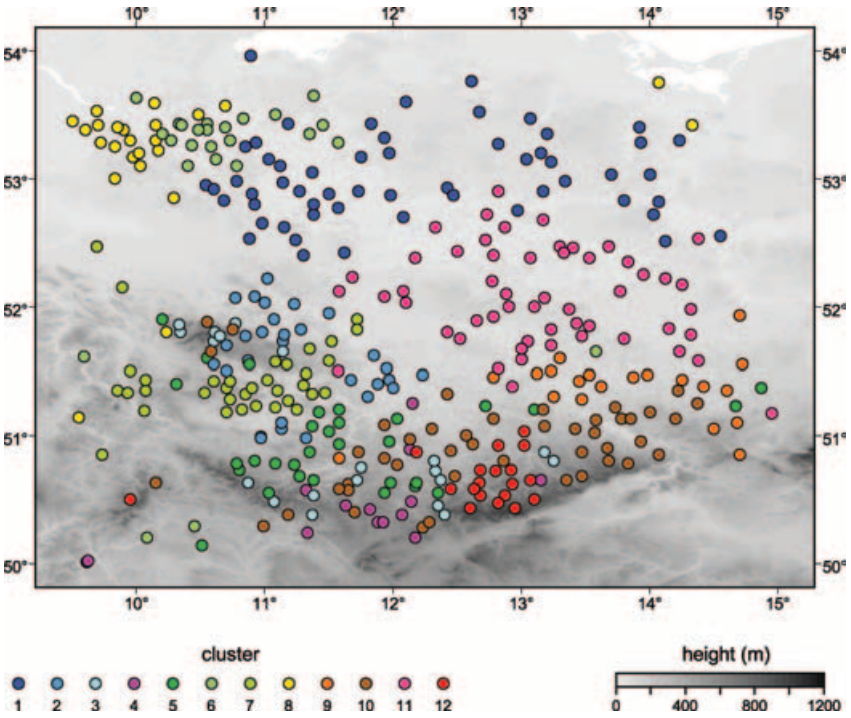

Fig 2. Cluster distribution of the observed data.

observations for M1 and M2. To get a clear overview three types were defined:

(1) grid points without cluster deviation (white);
(2) grid points with cluster deviation mainly caused by the influence of air temperature (red range);
(3) grid points with cluster deviation mainly caused by the influence of precipitation (blue range).

It is obvious that in both cases the error structure is primarily controlled by the precipitation (58.3%). Also, the model is not able to reproduce the temperature conditions correctly (41.7%).

4. Summary

It has been shown that the method presented here is suitable for describing complex relations (patterns) based on different parameter combinations. Using pattern comparison the differences between reference and simulation data sets can be made visible in space. Investigations on the temporal behaviour are published in Kücken et al. (2002). Additionally, one can estimate the influence of each single parameter on the error. Thus, a tool is made available for the modeller to analyse simulations quickly and conveniently. The influence of cloud ice parametrization on the prediction quality of the model that has been investigated with this method could spatially be exactly quantified in the validation area. The introduction of cloud ice parametrization did not lead to a general improvement of the model results. This is due to the fact that the amount of precipitation was overestimated in the model with the new cloud ice scheme. At the same time, the complex validation showed that the calculated near-surface temperature values are, as expected, influenced only in a few cases by the introduction of this scheme. In principle, it is not enough to introduce a cloud ice parametrization to better describe the physical processes that are the basis for the precipitation flow. The model dynamics, however, requires better fine-tuning with this parametrization.

5. Appendix A: Extended Non-Hierarchical Cluster Analysis

The aim of cluster analysis is the separation of several elements into homogeneous groups. In a first step, an equal number of L elements e_i (with $e_i = f(p_{i1}, \ldots p_{iN}$, where N is the number of parameters) from a total of M elements (grid points) has to be distributed to a defined number of K_0 clusters $c_1, \ldots c_k$ (initial partition) so that each cluster receives $L = M/K_0$ elements as follows:

$$\begin{aligned} &e_1, \ldots, && e_L \in c_1 \\ &e_{L+1}, \ldots, && e_{2L} \in c_2 \\ &. && . \\ &. && . \\ &. && . \\ &e_{(k-1)L+1}, \ldots, && e_{kL} \in c_k \end{aligned} \qquad \text{(A1)}$$

A so-called group centroid $\bar{e}$ is then calculated for each cluster c_k. It is the cluster mean value using normalized parameters:

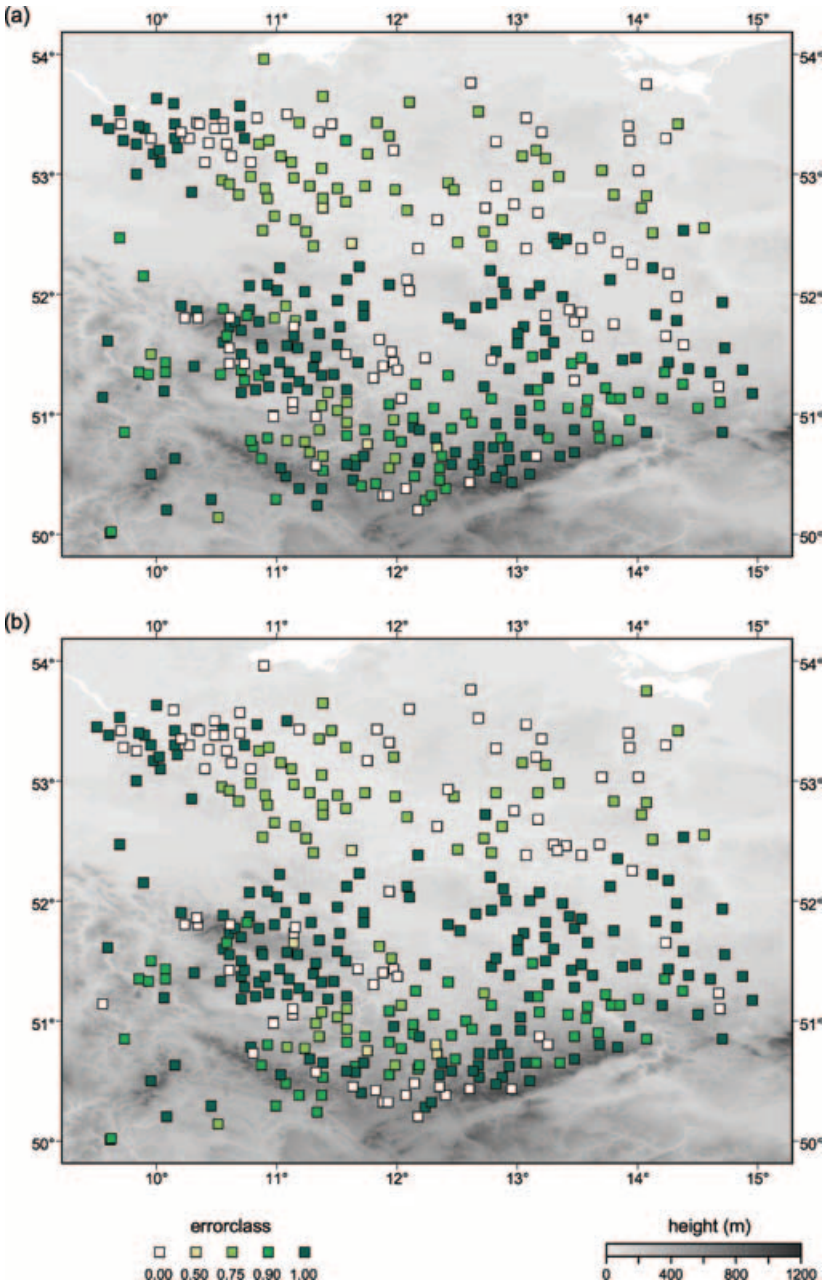

Fig 3. Spatial error distribution of the comparison (a) M1 versus observations and (b) M2 versus observations for the period August 1, 1995 to October 31, 1995.

$$\bar{e}_k = \frac{1}{L} \sum_{i=(k-1)L+1}^{kL} e_i . \tag{A2}$$

The Euclidean distance (Steinhausen and Langer, 1977) between the elements and the centroid defines the following target function $a(g)$ at each grouping step g:

$$a(g) = \sum_{k=1}^{K} \sum_{i \in k} |e_i - \bar{e}_k|^2 . \tag{A3}$$

In this sense, each grouping step can be seen as a displacement of the element e_i into the cluster whose centroid is closest to e_i.

Table 2. Distribution of error deviations for M1 and M2

	M1	M2
Mean value deviations	0.71	0.72
Error class 0	22.04%	22.04%
Error class 0.5	1.08%	1.34%
Error class 0.75	18.28%	16.40%
Error class 0.9	18.55%	14.52%
Error class 1.0	40.05%	45.70%
T2m influence >50%	41.7%	41.7%
Prec influence >50%	58.3%	58.3%

Table 3. Shifts in the error classes in M2 compared to M1

Type of shift	Cluster improvement	Cluster deterioration
All error classes	49 stations	64 stations
Complete change of the error class 0.1	16 stations	24 stations
T2m influence >50% with a complete change of the error class	4 stations	6 stations
Prec influence >50% with a complete change of the error class	12 stations	18 stations

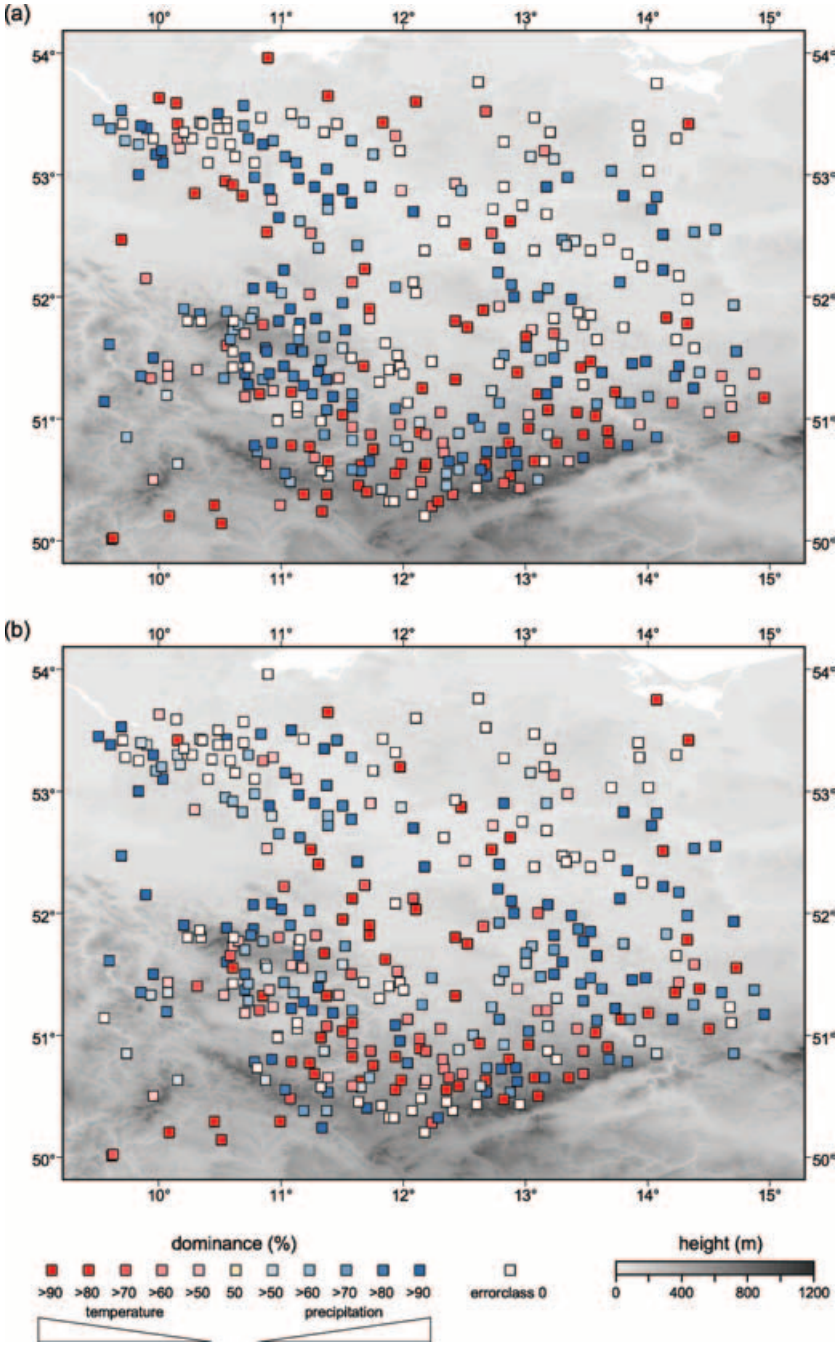

Fig 4. Spatial influence of the meteorological parameters surface air temperature (red) and precipitation (blue) on the error of the comparison (a) M1 versus observations and (b) M2 versus observations for the period August 1, 1995 to October 31, 1995.

Thus, the target function can be made smaller:

$$a(g)\forall g \rightarrow \min . \tag{A4}$$

This procedure is repeated until a local minimum of the target function is reached.

Gerstengarbe and Werner (1997) have developed a procedure to test the quality of cluster separation as follows. After having reached the local minimum, each cluster contains a certain number of elements. Each element is defined by n parameters, that is, it is located in an n-dimensional parameter space. Hence, each cluster is represented by a scatter plot of elements in the parameter space. Overlaps may occur between the scatter plots of individual clusters. This means that the parameter space of a cluster a may pass into that of cluster b. The number of parameters in the common space of two clusters can be defined as overlaps of cluster a with respect to cluster b

$$O_{a,b} = \sum_{i_a=1}^{L_a} \sum_{i_b=1}^{L_b} \sum_{j=1}^{N} o_{i_a,i_b,j} a, b = 1, \cdots, k \; a \neq b \tag{A5}$$

with

$$o_{i_a,i_b,j} = \begin{cases} 1 p_{i_b,j} \geq p_{i_a,j} \\ 0 p_{i_b,j} < p_{i_a,j} \end{cases}.$$

The maximum possible number of overlaps between clusters a and b is denoted as follows: $O_{a,b}^{\max} = NL_aL_b$ (where L_a is the number of elements in cluster a and L_b is the number of elements in cluster b). This number is reached if both clusters cover the same region in the n-dimensional parameter space. A statistically significant cluster separation depends on the number of overlaps.

A student t-test can be used to see whether $\bar{O}$ and $O^{\max}$ originate from the same basic population, where $\bar{O}$ is the mean and $O^{\max}$ is the maximal possible number of overlaps of all combinations of cluster pairs.

The clusters can be separated only when the null hypothesis is rejected. In this case the following procedure must be performed additionally, otherwise the clustering has to start with another initial partition. The ratio $v_{a,b}$ of the actual to the maximum possible number of overlaps is determined for each cluster pair $v_{a,b} = O_{a,b}/O_{a,b}^{\max}$.$\bar{v}$ is the mean of all $v_{a,b}$. A statistically significant separation between a and b exists if $v_{a,b} > \bar{v}$. Where $v_{a,b} > \bar{v}$, the quality of the separation still needs to be determined. The null hypothesis for this case is formulated as follows. The overlaps between two clusters a and b are not significantly different from the mean number of overlaps $\bar{O}$. For the confirmation or rejection of the null hypothesis, the following χ^2 test can be applied using the maximum possible number of overlaps $O_{a,b}^{\max}$, the actual number of overlaps $O_{a,b}$ and the mean of all actual numbers of overlaps $\bar{O}$ of all combinations of cluster pairs

$$\chi^2 = \frac{(O_{a,b} - \bar{O})^2 \cdot \left(2O_{a,b}^{\max} - 1\right)}{(O_{a,b} + \bar{O})^2 \cdot \left(2O_{a,b}^{\max} - O_{a,b} - \bar{O}\right)} \tag{A6}$$

with one degree of freedom. The result of this test can be interpreted in the following way. If the calculated χ^2 value is greater than a given threshold of significance, the frequency of overlaps exceeding the mean value $\bar{O}$ differs significantly from the χ^2 value. The separation between the clusters a and b is therefore statistically not significant.

As mentioned above, the cluster calculation must be started with a certain number of clusters (initial number). This number of clusters can influence the cluster result. Therefore, it is necessary to estimate the optimum initial number of clusters.

The starting point for the calculation of the initial cluster number is the target function $a(g)$. We know that the target function is constructed such that the partition for which the function reaches a minimum defines the best grouping of the clusters. Now we calculate the target function for an increasing number of initial clusters (for $q = 2, 3, 4, \ldots, K_0$). We obtain a sequence of K_0 independent target function values. This sequence can be incorporated in the following estimation of the optimum initial number of clusters. Realizing that each value of the target function corresponds to a specific initial number of clusters, we define the optimal initial number as the inflection point within the sequence of target function values where the trend of the target function values disappears and no further significant changes occur. This idea can be solved practically with the following steps:

(1) calculate the differences between consecutive values of the target function sequence and creation of a difference series $d_i (i = 1, \ldots, m)$ with $m = K_0 - 1$ values;

(2) apply the Pettitt (1979) test to estimate the beginning of a trend (or inflection point) within the difference series.

Continuously increasing the initial number of clusters, the Pettitt test finally defines the position within the difference series d_i (of the target function values) which divides this series into one part with significant changes of the values and the other part without significant changes. This cluster number defines the optimal initial number of clusters which we finally use for the cluster separation.

6. Acknowledgments

We are indebted to the German Weother Service for providing the data used in this investigation.

References

Arakawa, A. and Lamb, V. R. 1981. A potential enstrophy and energy conserving scheme for the shallow water equations. *Mon. Wea. Rev.* **109**, 18–36.

Doms, G. and Schättler, U. 1999. The non-hydrostatic limited-area model LM (LokalModell) of DWD – Part I: scientific documentation. German Weather Service, Offenbach/M.

Forgy, E. W. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* **21**, 768.

Gerstengarbe, F.-W. and Werner, P. C. 1997. A method to estimate the statistical confidence of cluster separation. *Theor. Appl. Climatology* **57**, 103–110.

Gerstengarbe, F.-W., Werner P. C. and Fraedrich, K. 1999. Applying non-hierarchical cluster analysis algorithms to climate classification: Some problems and their solution. *Theor. Appl. Climatology* **64**, 143–150.

IPCC 2001. *Climate Change 2001—The Scientific Basis*. Cambridge University Press, Cambridge, 881 pp.

Kücken, M., Gerstengarbe, F.-W. and Werner, P.C. 2002. Cluster analysis results of regional climate model simulations in the PIDCAP period. *Boreal Env. Research* **7**, 219–223.

Pettitt, A. N. 1979. A non-parametric approach to the change-point problem. *Appl. Statistics* **28**, 126–135.

Schättler, U. and Doms, G. 2000. The non-hydrostatic limited-area model LM (LokalModell) of DWD – Part III: user guide. German Weather Service, Offenbach/M.

Steinhausen, D. and Langer, K. 1977. Clusteranalyse—Einführung in Methoden und Verfahren der Automatischen Klassifikation. Walter de Gruyter, Berlin, 411 pp.