METHODS OF SELF-ORGANIZATION OF MODELS IN PROBLEMS OF PHYSICAL-STATISTICAL PREDICTION OF ALTITUDE PROFILES OF THE TEMPERATURE

A.N. Kalinenko and T.G. Teushchekova

Institute of Atmospheric Optics,
Siberian Branch of the Academy of Sciences of the USSR, Tomsk
Received January 9, 1989

Possible applications of algorithms of the method of grouping of arguments (MCA) in problems of medium- and long-term prediction of the altitude profiles of the temperature are discussed. These algorithms make it possible to construct simple predictive models with optimal complexity based on a special table of the initial experimental data (predictors), which hold a long history of the process, and with the help of criteria set by experts. The effectiveness of this approach in medium-term prediction of the altitude profiles of the temperature is demonstrated for a concrete example.

The rapid development of passive and active optical systems for remote study of the natural resources of the earth and the need for a more complete data base for a wide range of applied problems in the optics of the atmosphere, for the solution of which it is necessary to take into account the characteristics of the altitude distribution of the thermodynamic parameters and other optically active components of the atmosphere, have made it necessary to develop new reliable methods for real-time and predictive (with different terms of the prediction) estimates of the state of the atmospheric-optical channels of Information transmission.

In this paper, following the methodology of Ref. 1, we make an attempt to use the algorithms of the method of self-organization of predictive models, developed at the Institute of Cybernetics of the Academy of Sciences of the Ukrainian SSR (see, for example, Refs. 2 and 3), for making a probabilistic estimate of the changes in the altitude profiles of the temperature as a function of time; such profiles play an important role in calculations of the optical characteristics of a channel (for example, the transmission functions). We emphasize that this paper concerns precisely the probabilistic estimation of the time-dependent changes in the altitude profiles of one of the optically active components of the atmosphere — the temperature — and not weather forecasting (temperature forecasting) in the usual sense of this word, though for brevity we shall use the term "forecasting". Following Refs. 2 and 3 we shall describe the basic idea of the method.

Self-organization of models, similar to the widely employed regression analysis, is an experimental method of modeling, since it is based on the analysis of tabulated observational data obtained in a passive or active experiment. In regression analysis, however, the structure of the model must be determined arbitrarily. This makes it possible to construct the model only in the region where the number of coefficients of the model is equal to or less than the number of tabulated points. We also point out that in regression analysis the model is evaluated according to the criterion of the rms error, calculated based on all tabulated points. We shall term such criteria internal. However any internal criterion for comparing models leads to a false rule: the more complicated the model the more accurate it is.

In contradistinction to regression analysis. In self-organization of models the so-called inductive approach is employed. In this approach the structure of the forecasting model is determined by testing many possible models according to external selection criteria that are set by experts, i.e., calculated from data that are not employed in the synthesis of the models. According to the principle of self-organization, as the structural complexity of the model is gradually increased the values of the external criteria at first decrease, after which they start to increase, i.e., there exists a minimum that determines the model with optimal complexity. Thus a small table of starting data is stored in a computer and a class of functions and criteria for selecting the model are indicated. Using algorithms of the method of grouping of arguments (MCA) and by sorting through a large number of possible models the computer finds based on a given external criterion a unique model with optimal complexity giving an objective forecast.

We shall point out two other features of the MCA that make it possible to improve the forecast of complicated processes for which it is difficult (and often impossible) to construct an arbitrary complicated physical model isomorphic to the mechanism of the object being modelled and adequately describing it.

0235-6880/90/01 050–7 $02.00

1. Forecasting can be performed without a complete data base (i.e., without measurements of all existing arguments). Modeling by the MCA is the reverse of the idea of increasing the data base of the model by taking into account the maximum number of all actions. This makes it possible to reduce to a minimum the required a priori information that must be stored in the computer.

2. Self-organization of the physical and forecasting models with the initial data strongly distorted by noise is possible. Algorithms of the MCA now make it possible to reconstruct a physical model of the object in the case when the noise in the initial data is several times stronger than the regular signal.[4] In the works mentioned above,[2, 3] in our opinion, the advantages of using this method in problems of forecasting complex processes have been quite convincingly illustrated for different concrete examples.

To construct a forecasting model of the altitude profiles of meteorological parameters by the methods of self-organization the following elements are required together with a powerful multifunctional computing complex with well-developed software and an information base:

1) well-prepared initial data (sample);

2) selection of a method for generating forecasting models that is adequate for the problem and a rule for gradually increasing the structural complexity of the models and instructions for constructing possible models of different complexity on a given class of reference functions;

3) threshold selection; setting of heuristic criteria for selecting models: and,

4) analysis of the results obtained and estimation of the accuracy.

We shall study in greater detail the application of the MCA for forecasting the altitude profiles of the temperature based on aerological data. For the predictors and predictants we employed the time series of the leading coefficients $a_1(t)$ in the expansion of the altitude profiles of the deviations of the temperature $\Delta T_z(t)$ in the natural orthogonal functions $\varphi_1(z)$ (z is the altitude) of the corresponding covariation matrices.[6] The elements of these matrices describe the interlevel correlation couplings of the variations of the temperatures at the level of the aerological station and on the standard isobaric surfaces (850, 700, 500, 400, 300, 250, 200, 150, and 100 mb) for a quite long series of observations, preceding the period of forecasting (as a rule, this is a ten year series), and a chosen season (winter, spring, summer, or fall) or month of the year. This technique and its advantages are explained in detail in Ref. 7. Here we only note that choosing for the predictors the leading coefficients in the expansion of $\Delta T_z(t)$ in the natural orthogonal functions makes it possible to reduce sharply the volume of calculations and to preserve the main characteristic features of the behavior of the altitude profiles of the temperature in a given region that are determined by the main circulation processes in the atmosphere.

To prepare the table of initial data for forecasting the altitude profiles of the temperature we employed a specialized program-information complex that includes the following modules:

1) a data base of aerological data on magnetic tape, reading, writing, and conversion of data;

2) checking and analysis of data, rejection of unreliable data, and spline interpolation of missing values;

3) formation of ordered arrays of data on isobaric surfaces or at the nodes of a kilometer grid;

4) calculation of the statistical characteristics of the meteorological quantities (averages, standard deviations, variances, etc.);

5) construction of the correlation and covariation matrices;

6) calculation of the eigenvectors and eigenvalues of these matrices;

7) expansion of the initial data in the obtained eigenvectors;

8) reconstruction of the initial profiles using the expansion coefficients;

9) filtering of the expansion coefficients and estimation of their statistical characteristics; and,

10) display or printing of the results, obtained in each block, in a tabulated or graphical form.

With the help of this complex we analyzed data from different aerological stations in order to determine the accuracy and rate of convergence of the expansions of the Initial data In the natural orthogonal functions (NOFs). We constructed the initial temperature profile

$$\Delta T_z(t) = \sum_{i=1}^{v} a_i(t)\varphi_i(z), \qquad (1)$$

where $\Delta T_z(t) = T_z^1(t) - \bar{T}_z$ is the deviation of the temperature, measured at the time t at the level z, from the multiyear average value $\bar{T}_z$, $v \le P$, and P is the rank of the covariation matrix, was reconstructed from the expansion coefficients $a_1(t)$ and the corresponding eigenvectors of the covariation matrices $\varphi_1(z)$ (the i-th natural orthogonal function) for different aerological stations In the northern hemisphere and for different seasons.

We found that depending on the geographic characteristics of a station four to six of the first coefficients are, as a rule, sufficient to reconstruct the initial temperature profiles with adequate accuracy (not worse than the instrumental error In the measurement using a radiosonde); this accounts for 85 to 95% of the total variance of all variations of the temperature. These coefficients, calculated for the London station, which has quite reliable and long continuous series of observations, were used for developing the method of forecasting of altitude profiles of the temperature with the help of multlseries[8] and single-series[9] algorithms for separating the harmonic trend with nonmultiple frequencies according to the principle of self-organization of models. These

algorithms were adapted to an ES-1055M computer. They were supplemented with selection criteria, programs for reconstructing, based on the predicted values of the expansion coefficients, the altitude profiles of the temperature and comparing them with the profiles reconstructed from the actual values of the coefficients and real realizations for the same periods of the radiosonde data, obtained from an aerological data base, as well as estimates of the accuracy of the forecast and programs for tabulated and graphical output of the results. The modeling showed that both algorithms give fairly good results, but each algorithm has its own advantages and disadvantages:the single-series algorithm requires less computer memory and less computing time (it is approximately an order of magnitude faster), but it is too sensitive to noise. A closed numerical experiment showed that for a noise level of ~ 1% of the useful signal the parameters of the model are no longer reconstructed. Therefore this algorithm cam be used only for prefiltered data. The multiseries algorithm, which systematically removes the harmonic trend from the initial time series and fragments, requires more computer memory and more computer time, but it is much less sensitive to noise and makes it possible to separate from the process being approximated a large number of harmonics.

The process of approximating the initial time series, given by the discrete function f_n on the interval [1, N_A] (called the teaching sequence), with a finite of a trigonometric series with m nonmultiple frequencies $\omega_k \hbar \neq k\omega_1$ (i.e., frequencies that are not multiples of the period of the observations)

$$y_n = A_0 + \sum_{k=1}^{m-1} [A_k \sin \omega_k n + B_k \cos \omega_k n], \qquad (2)$$

where $\omega_i \neq \omega_j$ for $i \neq j$, $0 < \omega_1 < \pi$, and $k = 1, 2, \ldots, m$, where m is the number of harmonics), is divided into three stages.

At the first stage m coefficients α_p, $p = 0, 1, \ldots, m-1$ giving the best satisfaction of the relations

$$\sum_{p=1}^{m-1} \alpha_p (f_{l+p} + f_{l-p}) = f_{l+m} + f_{l-m}. \qquad (3)$$

are determined.

To reduce the sensitivity of the algorithm to noise we introduce a moving average of the time series (smoothing)

$$\bar{y}_k = \frac{1}{l} \sum_{l=k-l_0}^{k+l_0} f_l; \quad l_0 = \frac{1}{2}(l-1),$$

i is an odd number.

At the second stage the frequencies ω_k, $k = 1, 2, \ldots, m$ are found from the equation

$$\alpha_0 + \sum_{p=1}^{m-1} \alpha_0 \cos p\omega = \cos m\omega,$$

which with the help of the recurrence relation

$$\cos k\omega = 2\cos[(k-1)\omega]\cos\omega - \cos[(k-2)\omega]$$

is reduced to an algebraic equation of degree m for $\cos \omega_k$.

At the third stage the coefficients of the model A_k and B_k, which appear linearly In the model (2), are determined from the condition

$$\sum_{l=1}^{N_A} (f_l - y_l)^2 \to \min$$

Since the frequencies have already been calculated the estimates A_k and B_k can be obtained by Gauss method. The implementation of the principles of self-organization consists of the following. The initial time series f_n, given on the interval [1, $N = N_A+N_B+N_C$], is divided into three sequences: the teaching sequence, whose length is for estimating the parameters of the model (2); a checking sequence, whose length is N_B, for determining the value of the external selection criterion; and, an examination sequence, whose length is N_C, for estimating the quality of the forecast. The sequence lengths N_A, N_B, and N_C are set by the user.

Trends with the first, second, and m_{max}-th harmonic components are separated using the points of the teaching sequence; F (freedom of choice) of the best trends are selected from them according to the given selection criterion. Trends of different complexity (from 1 to m_{max} frequencies) are formed for each of the F remainders (the difference between the initial time series and the model) and the F best ones are selected out of the overall number of trends of the second series. This procedure is repeated at the next series of selection. The maximum number of selection series is an input parameter.

The optimal trend is determined based on the minimum of one or several external selection criteria. In this work we tested (see Ref. 8) the regularity criterion, which determines the standard deviation of the model on the checking sequence N_B, the correlation coefficient of the values of the initial series and the model on N_B, the forecasting accuracy criterion, which determines the standard deviation of the model on the examination sequence N_C, and the criterion of nonstationariness of the remainder, which can be used for long observational intervals, since the correlation function is used to calculate it.

After the optimal harmonic trend is found the forecasting accuracy is determined: the values of the standard deviation and the correlation coefficient on the teaching, checking, and examination sequences are calculated.

The elementary statistical characteristics are calculated for the remainder (the difference between the initial time series and the optimal trend): min

and max of the values, I–IV initial moments, I–IV central moments, the confidence intervals of the average, the variance, and the standard deviation for levels of significance 0.80 and 0.96; this also makes it possible to judge the accuracy of the optimal trend.

The optimal model is calculated independently for each leading expansion coefficient $a_1(t)$, after which $T_z(t_n)$ are reconstructed using the formula (1) on a fixed number of steps n ahead and are compared with T_z^1 at the corresponding data from the aerological data bank.

Tens of test calculations for different initial data were performed with the help of these algorithms using the data from a ten-year (1961–1970) series of observations at the London station. They permit drawing the conclusion that these algorithms of MCA make it possible to solve successfully the problem of forecasting the values of the altitude profiles of the temperature 10 to 15 days in the future. Some typical examples of the comparison of the forecasted and actual temperature profiles for the summer season, which is characterized by high variability, are presented in Tables I and II. The actual values of the temperature from the aerological data base are presented in column 1 (and all subsequent odd columns), and the deviations of the values of the temperature reconstructed from the first six predicted expansion coefficients from the actual value are presented in column 2 (and all subsequent even columns). These data were obtained for comparatively short teaching sequences from 15 to 20 steps long, and different modifications of the regularity criterion were employed for selecting the models. An example of the computational results for a long teaching sequence N_A, equal to 60 points, and a 20-day forecast starting at August 10, 1967 is shown in Fig. 1 (the criterion of nonstationarity of the remainder was employed). Here the altitude profiles of the temperature up to 100 mb on August 15, 1967, August 19, 1967, and August 29, 1967 (the observations were conducted for 12 hours) are shown. The solid line shows the values of the temperature reconstructed using six expansion coefficients based on the initial data from the data bank. The circles indicate the initial temperatures from the data bank (with the corresponding standard deviations). The broken lines show the temperature profiles predicted for the corresponding dates (also using six coefficients). It is obvious that even the least successful test calculation gives a satisfactory forecast from August 10 to August 19, 1967 of the temperature profiles up to altitudes corresponding to 300 mb (~ 9 km).

TABLE I.

Comparison of the actual and predicted altitude profiles of the temperature. London station. Summer t970. Time 00 h.

number of days	earth		Isobaric surfaces, mBar						Isobaric surfaces, mBar							
			850		700		500		400		300		250		200	
	APT*	DF**	APT	DF	APT	DF	APT	DF	APT	DF	APT	DF	APT	DF	APT	DF
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	12.0	-2.8	3.3	-2.2	-2.6	-1.0	-16.7	0.8	-27.9	0.6	-42.9	-1.1	-50.4	-3.3	-52.2	-1.1
2	10.2	-1.4	5.2	-0.9	0.5	1.8	-13.0	3.4	-23.8	3.9	-39.7	2.1	-51.1	-3.0	-59.1	-8.0
3	16.0	6.4	7.6	-3.1	0.8	-1.4	-14.4	-1.0	-25,9	-1.1	-42.3	-2.6	-52.5	-4.9	-58.6	-6.7
4	14.2	1.3	8.2	-3.7	1.2	-1.5	-13.7	-0.4	-25.1	-0.3	-41.6	-2.4	-51.3	-4.4	-52.6	0.5
5	13.6	-2.6	6.2	-1.5	-1.3	-0.8	-16.3	0.2	-27.7	0.5	-42.1	0.1	-48.0	0.5	-47.5	6.6
6	11.6	-5.6	2.6	-2.0	-4.6	-1.9	-18.5	-0.5	-28.6	0.9	-36.7	6.2	-36.3	11.7	-41.9	9.7
7	8.8	-6.0	2.0	-2.8	-2.4	-0.4	-15.0	2.2	-25.5	3.3	-40.4	3.0	-43.7	1.6	-49.3	4.6
8	11.2	-0.2	6.8	-1.4	-0.4	-1.2	-15.7	-0.9	-27.8	-1.2	-44.6	-2.1	-52.0	-0.3	-45.1	11.7
9	15.2	6.4	8.9	-2.1	1.2	-1.3	-13.6	-0.4	-24.6	0.2	-40.3	0.4	-50.2	-0.2	-57.7	-3.2
10	13.8	3.8	5.8	-3.0	-3.5	-3.7	-20.0	-4.7	-31.7	-4.8	-42.7	-0.6	-43.4	6.4	-45.1	7.7
11	11.4	-4.8	5.1	-1.2	-1.0	0.8	-15.4	2.1	-26.9	2.0	-44.2	-2.6	-54.4	-8.5	-55.1	-4.7
12	14.6	-5.0	8.7	1.8	2.0	2.6	-12.4	4.3	-23.1	5.0	-38.2	2.1	-48.8	-4.2	-59.9	-8.4
13	16.2	-0.5	11.3	3.3	3.7	2.9	-10.9	3.9	-21.8	4.6	-37.2	3.5	-48.0	0.2	-57.2	-2.7
14	13.8	1.6	7.2	-0.9	-1.1	-1.9	-16.6	-2.2	-28.0	-2.3	-41.5	-0.8	-46.4	2.3	-47.6	4.7
15	12.7	5.0	8.4	0.4	0.5	0.8	-14.9	0.7	-26.3	0.9	-40.9	2.3	-48.3	3.1	-52.8	-0.4

APT — actual profile of temperature
DF — deviation from forecasting

TABLE II.

Comparison of the actual and predicted altitude profiles of the temperature. London station. Summer 1970. Time 12 h.

number of days	earth		isobaric surfaces, mBar						isobaric surfaces, mBar							
			850		700		500		400		300		250		200	
	APT*	DF**	APT	DF	APT	DF	APT	DF	APT	DF	APT	DF	APT	DF	APT	DF
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	13.2	-2.1	2.3	-2.4	-0.3	1.7	-12.2	4.4	-22.1	5.5	-36.7	5.0	-47.9	-0.1	-57.8	-8.3
2	18.2	1.4	8.0	0.1	1.5	0.1	-13.3	-0.4	-24.4	-0.8	-39.6	-2.4	-49.4	-4.5	-57.3	-6.9
3	20.4	1.7	6.6	-1.7	-0.5	-1.3	-14.2	-0.6	-24.8	-0.4	-40.8	-2.8	-51.7	-6.2	-61.8	-8.6
4	16.2	-2.5	6.6	0.6	0.0	1.8	-14.6	2.5	-26.2	2.3	-42.5	-0.1	-50.6	-2.5	-46.6	5.0
5	12.4	-5.7	1.9	-4.5	-6.4	-4.6	-21.6	-4.5	-32.8	-4.4	-43.4	-0.4	-43.2	5.3	-42.8	5.3
6	15.6	-2.5	-1.3	-6.8	-4.0	-4.8	-16.8	-2.5	-26.9	-1.1	-38.7	2.7	-43.5	6.4	-48.1	4.1
7	16.2	-2.7	0.9	-7.5	-2.8	-4.2	-15.4	-1.9	-26.2	-1.7	-42.1	-3.2	-50.2	-2.9	-45.6	9.1
8	18.0	-0.4	7.7	1.1	-0.4	0.3	-16.0	-0.6	-27.9	-1.8	-44.4	-6.5	-51.6	-9.1	-47.1	1.7
9	15.4	-0.2	7.0	2.5	-0.5	1.9	-15.4	1.5	-26.8	1.1	-42.8	-1.9	-50.2	-4.7	-47.2	0.0
10	16.0	0.3	2.3	-4.0	-1.4	-0.9	-14.1	1.0	-24.7	1.7	-39.3	2.2	-46.9	2.3	-45.6	6.0
11	12.8	-6.6	9.8	0.4	2.5	0.7	-12.0	1.3	-22.7	1.8	-36.9	-2.0	-46.6	0.5	-58.3	-4.4
12	18.0	-2.4	13.4	5.4	4.3	4.3	-11.2	4.4	-22.0	4.8	-37.4	3.4	-48.3	-1.0	-60.8	-8.0
13	16.4	-1.8	10.1	4.3	2.4	4.5	-12.2	5.1	-23.0	5.6	-38.6	4.3	-48.6	-0.2	-55.3	-6.6
14	18.8	1.7	5.8	-1.0	-0.9	-0.5	-15.7	-0.5	-27.3	-1.0	-44.6	-4.2	-54.0	-7.2	-52.3	-4.2
15	20.8	3.5	9.0	1.3	1.3	0.3	-14.2	-0.6	-25.7	-1.2	-41.9	-3.1	-51.8	-4.5	-58.0	-3.6

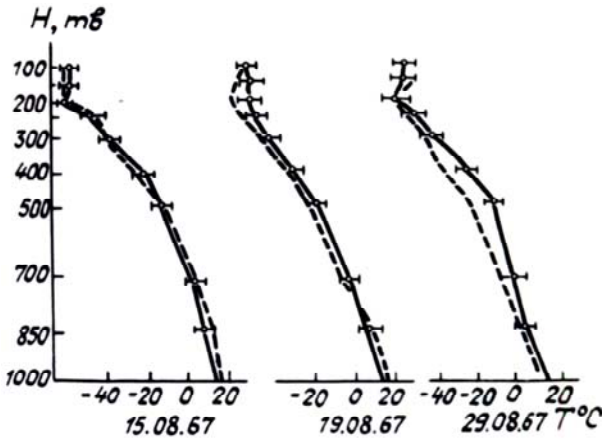

FIG. 1. Altitude profiles of the temperature. Comparison of the predicted and actual values: the solid curves show the actual values, the broken curves show the predicted values, and the circles show the initial data together with the standard deviation.

Thus the estimates presented, made without any optimization of the Input parameters and selection criteria in the algorithms, demonstrate, from our viewpoint, that the method of grouping of arguments is useful for forecasting the altitude profiles of the meteorological parameters of the atmosphere based on aerological data. Since these algorithms are very sensitive to variations in the input parameters and the selection criteria, it will be necessary to perform a special series of mathematical experiments in order to optimize these parameters and to increase the forecasting accuracy and term, including also for other meteorological parameters and other stations.

REFERENCES

1. V.S. Komarov, A.N. Kalinenko, and S.A. Mikhailov, Atmos. Opt. **2**, No. 5, 513 (1989).
2. A.G. Ivakhnenko and I.A. Müller, *Self-Organization of Forecasting Models*, [in Russian], (Tekhnika, Kiev, 1985); VEB Technlk Verlag, Berlin (1984).
3. A.G. Ivakhnenko and V.S. Stepashko, *Noise Insensitivity of Modeling*, (Naukova Dumka, Kiev, 1985).
4. A.G. Ivakhnenko and V.S. Stepashko, Avtomatika, No. 4. 26 (1982).
5. Yu.L. Babikov and V.T. Kalaida, *Applied Support for Collective- Use Systems* (Nauka, Novosibirsk, 1985).
6. A.V. Meshcherskaya et al., *Natural Components of Meteorological Fields*, (Gidrometeoizdat, Leningrad, 1970).
7. V.E. Zuev and V.S. Komarov, *Statistical Models of the Temperature and Caseous Components of the Atmosphere*, (Gidrometeoizdat, Leningrad, 1986).
8. A.G. Ivakhnenko [ed.], *Hand on Typical Modeling Programs*, (Tekhnika, Kiev, 1980).
9. Yu.P. Yurchkovskiĭ and N.V. Popkov, Avtomatika, No. 6, 9 (1986).