MODIFIED METHOD OF CLUSTERING OF ARGUMENTS AS AN EFFICIENT TECHNIQUE FOR STATISTICAL ESTIMATE OF THE CHARACTERISTICS OF THE FREE ATMOSPHERE UNDER CONDITIONS OF INFORMATION DEFICIT

V.S. Komarov, V.I. Akselevich, and A.V. Kreminskii

Institute of Atmospheric Optics,
Siberian Branch of the Russian Academy of Sciences, Tomsk
Received April 9, 1993

Feasibility of and prospects for the modified method of clustering of arguments (MMCA) are discussed as applied to the problems of forecasting (retrieving) the vertical profiles of temperature and the zonal and meridional components of wind velocity in the free atmosphere from the data obtained at lower levels, with no experimental data on the forecasted parameters. Examples show the advantages and efficiency of such an approach, which significantly extends the range of its application to the problems of atmospheric sounding, including lidar sensing techniques.

Now, because of greater requirements on quality of the data of atmospheric monitoring imposed by different branches of the national economy, new means and techniques for routine monitoring of the state of atmospheric environment are getting a tremendous boost. Among them the techniques of active remote sensing, which use laser sources of radiation, feature definite advantages,[1,2] since they provide estimate of the atmospheric fields with exceptionally high temporal and spatial resolution and retrieval of the necessary geophysical data in real time.

However, despite the apparent advantages of the techniques of laser sounding over *in situ* (e.g., radiosonde) measurements, they have some grave disadvantages as well. The main disadvantages are insufficient accuracy of the data obtained in the free atmosphere (that is, above 2–3 km) with ground–based lidar systems and rapid increase of the measurement error as the length of the atmospheric beam path increases. For example, air temperatures are retrieved by lidars at distances longer than 2 km with an error worse than 2°C (see Ref. 2), which exceeds the rejection criterion for air temperature data (1°C for the troposphere) established by the World Meteorological Organization.

Therefore, it is currently recommended to apply an integrated approach, which would compensate for those drawbacks. It is based on a joint use of accurate lidar data obtained at low altitudes (up to 1–3 km), complimented above by the meteorological parameters retrieved using various mathematical models (e.g., hydrodynamic or physico–statistical). This approach is particularly efficient for estimate of the vertical profiles of temperature and wind velocity, which are the basic atmospheric parameters used to forecast the weather and to model the climate.

However, it appears impossible to solve the problem of vertical profiling of meteorological elements for hydrodynamic forecast models, since such models are cumbersome. They feature limiting resolution and prediction range and call for the data of observations that would cover almost the whole hemisphere. This entails lengthy computations and supercomputers. In addition, the contribution from errors in the initial data to the errors in the output model results reaches 18–25% for hydrodynamic forecast models.[3]

All things considered, it appears more feasible to use the physico–statistical techniques, which have found wide application to the problem of retrieving and forecasting the vertical profiles of meteorological parameters (see, e.g., Refs. 4–6). Of such techniques, the newly developed modified method of clustering of arguments stands out.[7] It is quite simple, obviates the necessity of large amount of initial data and lengthy computations, does not call for any preliminary averaging of long–term series of the empirical data, and finally, offers a possibility to synthesize a prognostic model under conditions of partly or fully uncertain knowledge of the structure of the modeled process and properties of noise in the data used. We use this approach to solve the problem of retrieving such characteristics of the free atmosphere as its vertical temperature and wind profiles from the data of spatiotemporal observations.

Before proceeding to the results of statistical estimate (retrieval) of the characteristics of the free atmosphere by the MMCA algorithm, we consider its certain theoretical grounds.

The basic idea of the MMCA algorithm is as follows.[7] Based on a sample of experimental data, a certain set is automatically generated of prognostic models of different structure within a prescribed class of functions, from which one or several best models are selected (against a certain performance criterion), and then, using the model chosen, the spatiotemporal forecast (or retrieval) itself is made.

Apparently, to solve the formulated problem, one needs to know the type and length of the sample of experimental data; to prescribe the class of basis functions (operators), from which the set of the prognostic models is formed; to prescribe the way by which the structure of different models will be generated; and, to choose a technique for estimate of the parameters of generated models and a technique for minimization of the performance criterion.

As has already been mentioned, in our case we used the spatiotemporal observations of the form

0235-6880/94/02 121-04 $02.00

$$\{ Y_{h,\,t},\ h = 0,\ 1\ ,\ \dots\ ,\ h^*;\ t = 1,\ 2,\ \dots\ ,\ N \}$$
$$\{ Y_{h,\,t}\ ,\ h = 0{,}1\ ,\ \dots\ ,\ \overline{h}\ \mathrm{J}\ h^*;\ t = N + 1\} \qquad (1)$$

for our initial experimental data. Here h is altitude and t is time of observations. For our basis functions we selected mixed difference dynamic–stochastic models of the form

$$Y_{h,\,N+1} = \sum_{s=1}^{N^*} A_{h,\,\tau}\ Y_{h,\,N+1-\tau} + \sum_{j=0}^{h-1} B_{h,\,j}\ Y_{j,\,N+1} + \varepsilon_{h,\,N+1}\ ,$$
$$(h = \tilde{h} + 1,\ \dots\ ,\ h^*)\ , \qquad (2)$$

where N^* is the serial number of the time lag ($N^* < [N - h - 1]/2$); $A_{h,\,1}, \dots, A_{h,\,N}$ and $B_{h,\,0}, \dots, B_{h,\,h-1}$ are the unknown parameters of the model; and, $\varepsilon_{h,\,N+1}$ is the model discrepancy.

We followed Ref. 7 to determine the best model (2) and to make successful forecasts on its basis. Taking all the initial data (1), we divide them preliminarily into the sample A (it contained observations up to the instant $t = N - 1$ inclusive) and the sample B, which included only the observations at the instant $t = N$. In addition, two special methods were used, namely:

1) The method of directed cluster sampling to optimize the structure of the model, including two–stage model selection based on:

– *forecast resultant error* (after Akaike) of the form

$$\mathrm{FRE} = \frac{(N - N^* - 1) + s}{(N - N^* - 1) - s}\ \mathrm{RSS}(s)\ , \qquad (3)$$

where

$$\mathrm{RSS}(s) = \sum_{j\,=\,1}^{N-N^*-1} [\hat{Y}_{h,\,N-j} - \hat{Y}_{h,\,N-j}(s)]^2$$

is the residual sum of squares for the current model $\hat{Y}_{h,\,N-j}(s)$, containing s nonzero estimates of its parameters. Here the value of $\hat{Y}_{h,\,N-j}$ is estimated using the expression

$$\hat{Y}_{h,\,N-j} = X\ \hat{\mathrm{Q}},\quad X \in M_{(N-N^*-1)\times(N^*+h)}\ ,\quad \hat{\mathrm{Q}} \in R^{N^*+h}\ , \qquad (4)$$

where $\hat{\mathrm{Q}} = [\hat{A}_{h,\,1} \dots \hat{A}_{h,\,N^*}\ \hat{B}_{h,\,0} \dots \hat{B}_{h,\,h-1}]^{\mathrm{T}}$ is the minimum estimate of the parameters over the sample A, calculated from special formulae (here T denotes the operation of transposition), R^k is Euclidian space of k–dimensional vectors, and $M_{m\times p}$ is space of $m \times p$ matrices;

– *root–mean–square forecast error* from the reference sample (sample B):

$$|\,Y_{h,\,N} - \hat{Y}_{h,\,N}(\mathrm{s})\,|^2 \to \min\ , \qquad (5)$$

where the minimum is sought over all the $N^* + h$ structures, each being determined by the individual model $\hat{Y}_{h,\,N-j}(s)$.

2) The method of minimax estimate used to obtain the estimates of the model parameters, that guarantees the quality of the respective forecast estimated using the inequality

$$E\,|\,E(Y_{h,\,N+1}) - \hat{Y}_{h,\,N+1}\,|^2 \le \delta_{h,\,N+1}\ ,\ (h = \overline{h} + 1,\ \dots\ ,\ h^*)\ , \qquad (6)$$

where $E(\cdot)$ is the operator of mathematical expectation, performing averaging over all the realizations of observational errors, $Y_{h,\,N+1}$ and $\delta_{h,\,N+1}$ are the minimax estimates, depending on the variance of observational errors and *a priori* information on the maximum permissible errors in the forecast.

The technique described above was applied to the problem of retrieving the vertical profiles of temperature (T) and the zonal (V_x) and meridional (V_y) components of wind velocity in the free atmosphere. Performance criterion and efficiency of the MMCA were then estimated from the long–term (1961–1975) observations at four aerological stations: Keflavik (63°57′ N, 22°37′ W), Stavanger (58°33′ N, 05°38′ E), Rome(41°48′ N, 12°38′ E), and Miami (25°49′ N, 80°17′ W), which are in different physical and geographical regions of the northern hemisphere. Retrieval accuracy for the above physical parameters was estimated using the standard (root–mean–square) errors δ and the relative errors δ/σ in per cent (here σ is the root–mean–square deviation characterizing the natural variability of each studied parameter).

TABLE I. Standard retrieval errors δ *and root–mean–square deviations* σ *of temperature* (T,°C) *and the zonal* (V_x, m/s) *and meridional* (V_y, m/s) *components of wind velocity from the data obtained at the ground and barometre altitude of* 850 hPa.

Barometre altitude, hPa	T			V_x			V_y		
	δ	σ	δ/σ, %	δ	σ	δ/σ, %	δ	σ	δ/σ, %
1	2	3	4	5	6	7	8	9	10
	Station Keflavik								
	Winter								
700	2.5	5.4	46	1.9	8.3	23	5.1	9.6	53
500	3.5	5.6	62	3.0	11.3	26	6.9	13.0	53
400	3.3	4.6	72	4.4	13.6	32	8.7	15.6	56
300	3.2	3.5	91	4.4	14.7	30	8.9	16.9	53
	Summer								
700	1.9	3.4	56	2.5	6.4	39	3.6	6.6	55
500	2.5	3.8	66	3.1	9.1	34	3.7	9.1	40
400	2.9	4.1	70	3.8	11.5	33	7.2	11.3	63
300	3.1	3.8	82	4.0	14.6	27	6.8	14.0	48
	Station Rome								
	Winter								
700	2.6	4.4	59	5.0	9.1	55	6.7	9.2	73
500	3.0	4.5	66	6.0	11.5	52	8.8	13.4	66
400	2.9	4.4	66	8.4	14.4	58	9.8	16.5	59
300	3.0	3.3	91	11.1	17.9	62	13.8	19.4	71
	Summer								
700	2.5	3.5	71	4.3	7.5	57	4.3	6.6	65
500	2.1	3.3	64	5.7	8.4	67	6.0	8.7	69
400	2.3	2.8	82	7.2	10.6	68	8.2	10.5	78
300	2.5	2.6	96	9.0	13.2	69	10.6	13.6	78
	Station Miami								
	Winter								
700	1.9	2.5	76	1.8	5.9	30	3.8	6.4	60
500	1.5	2.4	62	2.2	7.8	28	4.9	7.2	68
400	1.4	2.5	56	2.7	9.5	28	5.8	9.0	64
300	1.4	2.2	64	4.0	11.7	34	6.5	11.0	59
	Summer								
700	0.9	1.3	69	0.9	4.3	21	2.7	3.7	73
500	1.0	1.3	77	1.2	4.9	24	2.2	3.7	59
400	0.9	1.3	69	1.3	5.8	22	3.1	4.9	63
300	1.0	1.6	62	1.3	7.4	18	4.3	5.9	73

At the first stage of solving the formulated problem, numerical experiments were conducted to estimate the dependence of efficiency of the MMCA on the choice of the meteorological parameter, on the number of the prescribed structures of the prognostic models, and on the order of matrices of input variables, which depends on the number of profiles used.

Our numerical experiments demonstrated that:

1. The modified method of clustering of arguments is an efficient technique for numerical estimate of the characteristics of the free atmosphere (in our case these are the air temperature and the zonal and meridional components of wind velocity) from the data obtained at lower levels including the ground and barometre altitude of 850 hPa, where reliable lidar data are available. This is clear from Table I, which lists, by way of example, the values of absolute (σ) and relative (δ/σ, %) retrieval errors for T, V_x, and V_y at different levels in the troposphere. These were retrieved from the data obtained at two levels: the ground and barometre altitude of 850 hPa (~ 1.5 km) at the stations Keflavik, Rome, and Miami only.

2. With rare exception, this method yields the best results of retrieval of wind velocity components (particularly its zonal component V_x). In almost every case, independent of station site, season, and level, the relative retrieval error for V_x remains within 18–60%, while the error of numerical estimate of V_y does not exceed 60–70% (see Table I).

3. The most successful 12–hour forecast (retrieval) of the vertical profiles of tropospheric temperature and wind is given when ten structures are specified, which define the best structure of the prognostic model (in the sense of the retrieval quality), when a statistical sample numbering from 7 to 16 profiles is used.

At the second stage of solving the formulated problem, we conducted numerical experiments aimed at additional assessment of the efficiency of the minimax approach to the statistical estimate of the characteristics of the free atmosphere. To this end, we used a procedure by which the temperatures and the zonal and meridional components of wind, retrieved by the algorithm of the MMCA, were compared to those retrieved by the method of multidimensional extrapolation (MMDE). The latter has found wide application to estimating the parameters of the free atmosphere from the data obtained at the levels below and above those being retrieved (see, e.g., Refs. 4, 9, and 10). By the MMDE, the vertical profiles of meteorological parameters were retrieved from the data obtained at lower levels using the expression of the form

$$\hat{a}_0 = \overline{a} + \boldsymbol{\omega}' (\mathbf{a}_i - \overline{\mathbf{a}}_i) , \qquad \boldsymbol{\omega}' = S^{-1} \mathbf{s} , \qquad (7)$$

where $\hat{a}_0$ and $\overline{a}_0$ are the retrieved and the average values of a given meteorological parameter at some retrieval level; $\mathbf{a}_i$ and $\overline{\mathbf{a}}_i$ are the vectors of the instantaneous and average values of the same parameter at lower levels; $\boldsymbol{\omega}'$ is the parameter of multiple regression, estimated with the help of the selected covariance matrix $\|S\|$ and the selected k–dimensional vector of relation between the predictor and the predicant $\mathbf{s}$.

By way of example, Table II lists retrieval errors for the vertical profiles of temperature and the two components of wind velocity obtained by the MMCA and the MMDE algorithms from the data of station Rome in winters of 1970–1975.

TABLE II. Root–mean–square deviations σ *and standard retrieval errors* δ *of the temperature* (T,°C) *and the zonal* (V_x, m/s) *and meridional* (V_y, m/s) *components of wind velocity from the data obtained at the ground and barometre altitude of* 850 hPa.

Barometre altitude, hPa	T			V_x			V_y		
	δ_1	δ_2	σ	δ_1	δ_2	σ	δ_1	δ_2	σ
700	2.6	3.0	4.4	5.0	8.0	9.1	6.7	7.3	9.2
500	3.0	5.7	4.5	6.0	8.9	11.5	8.8	9.8	13.4
400	2.9	6.3	4.4	8.4	9.2	14.4	9.8	9.9	16.5
300	3.0	3.6	3.3	11.1	12.4	17.9	15.3	15.8	19.4

Analyzing Table II, one sees that the values of standard errors δ calculated on the basis of these two methodological approaches, are quite comparable to each other. Moreover, usually independent of altitude (level) and the chosen meteorological parameter, the retrieval accuracy for the profiles of T, V_x, and V_y is much higher if the method of modified clustering of arguments is employed. This fact argues for the use of this method for statistical estimate of characteristics of the free atmosphere.

Thus, based on the statistical estimates, we may conclude that the MMCA algorithm is acceptable for retrieving the characteristics of the free atmosphere from the data obtained at lower levels. This is also true if we consider the lidar data obtained at the ground and barometre altitude of 850 hPa. As compared to the method of multidimensional extrapolation, this method obviates the necessity of preliminary statistical generalization of a large array of the data of long–standing aerological observations, from which the parameters of multiple regression are otherwise to be calculated. It should be emphasized that the algorithm of the MMCA may yield even greater efficiency in case the input data are obtained with high spatial and temporal resolution, and an account is made of the synoptic situation during the experiment (in other words, when a teaching sample is formed). However, all such problems are outside the scope of the present paper and will be the subject of our further studies.

REFERENCES

1. V.E. Zuev and V.V. Zuev, *Remote Optical Sensing of the Atmosphere* (Gidrometeoizdat, St. Petersburg, 1992), 232 pp.
2. R. Mezheris, *Remote Laser Sensing* (Mir, Moscow, 1987), 550 pp.
3. K. Miyakoda, in: *Theoretical Foundations of Medium–Range Weather Forecast* (Gidrometeoizdat, Leningrad, 1979), pp. 5–78.
4. V.S. Komarov, Meteorol. Gidrol., No. 4, 16–19 (1970).
5. A.I. Bagrov and E.A. Loktionova, Tr. Gidromet. Tsentr, No. 212, 42–46 (1978).
6. A.N. Kalinenko and T.D. Teushchekova, Atmos. Opt. **3**, No. 1, 50–56 (1990).
7. Yu.L. Kocherga, Avtomatika, No. 5, 80–86 (1991).
8. V.E. Zuev and V.S. Komarov, *Statistical Models of Temperature and Gaseous Components of the Atmosphere* (D. Reidel Publishing Company, Dordrtecht–Boston–Lancaster–Tokyo, 1987), 306 pp.
9. V.S. Komarov, Proc. VNIIGMI–MCD, No. 42, 22–26 (1977).
10. A.S. Marchenko, L.A. Minakova, and A.G. Semochkin, in: *Application of Statistical Methods to Meteorology* (Siberian Branch of the Academy of Sciences of the USSR, Novosibirsk, 1971), pp. 82–121.