FORECASTING OF OZONE DAILY MAXIMA WITH A STATISTICAL PROGNOSTIC MODEL: METHODOLOGY AND RESULTS OF ITS VALIDATION

Eugene Genikhovich, Lev Sonkin and Victoria Kirillova
Voeikov Main Geophysical Observatory, St. Petersburg, Russia

INTRODUCTION

High levels of variability, both in space and in time, of concentrations of atmospheric pollutants could result in a comparatively poor performance of forecasts of the air pollution. The axial ground level concentrations from a single point source, for example, cannot be predicted with the mean square error less than about 100% (*Genikhovich*, 2003). As a remedy, one could switch to predicting more stable statistical characteristics of the air pollution when the magnitude of the stochastic component of this variability is reduced. In particular, the regulatory dispersion model OND-86, widely used in Russia and other NIS countries, which is aimed for simulating the field of annual upper limits (98th percentiles) of short-term concentrations, performs with the error of about 25% (*Genikhovich*, 1995).

For the purpose of short-term stochastic forecasting the air pollution, the noise could be filtered out by the use of the so called parameter P that characterizes a synchronous component of variations of concentrations of pollutants at all urban monitoring stations (see *Berlyand*, 1991; *Sonkin*, 2002). For each of pollutants and given observational time, this parameter is determined as the ratio of the number of monitoring stations "m", where measured short-term concentrations are exceeding the corresponding long-term mean values more than in 50%, to the total number of monitoring stations "n" in the city under consideration. It was found that P was closely correlated with the first coefficient of expansion of the urban concentration field into series of the natural orthogonal functions. Being a dimensionless indicator, P characterizes the unfavorable meteorological conditions that could lead to high levels of air pollution over the whole urban area. In 2002, daily routine forecasts of P were produced in 235 Russian cities and actively used there for making decisions aimed to reducing the anthropogenic emissions. In this paper, we present a different approach to filtering out the noise in stochastic forecasting schemes based on the use of daily maxima of concentrations of air pollutants at each of urban monitoring stations.

DESCRIPTIOIN OF DATA SETS AND METHODS IN USE

The routine operational monitoring of air pollution in Russian cities is based on manual sampling of the air either three or four times a day followed by the laboratory chemical analysis. The list of measured species includes "major" pollutants (NO, NO_2, SO_2, CO and TSP), which are registered in each city, as well as so called "specific" pollutants that differ for different cities and depend on the chemical composition of the emissions in the city considered. On the whole, the number of pollutants monitored in all Russian cities exceeds 50. The number of monitoring stations is varying from one up to twenty and more depending on population of the city. Automatic air quality monitoring systems are operating in major cities like Moscow, St. Petersburg and others. In the present work, the methodology of forecasting the air pollution is constructed in such a way that it could be applied to both, manual and automatic monitoring data. The development of the statistical prognostic model for daily maxima of concentration is carried out in two steps (i) formulating and teaching the model, and (ii) forecasting with the use of the model. In its turn, the methodology of formulating and teaching the model includes the following steps:

(1) selection of predictors ;
(2) censoring of the sample (if needed);

(3) transformation of the predictant to the normal distribution;
(4) transformation of nonlinear dependencies in linear ones.

The resulting stochastic model is formulated then using the stepwise regression. Significant predictors are selected here with the use of the Fisher- and Student criteria.

RESULTS OF THE DATA ANALYSIS

Statistical prognostic models were developed for the cities of Krasnoyarsk, Siberia (for carbon disulfide and hydrogen fluoride) and Ufa, the European part of Russia (for ethylbenzene and benzene). All pollutants but benzene were emitted from a comparatively small number of sources, so that one can expect a high level of inhomogeneity of corresponding concentration fields (the most difficult situation to predict the air pollution). Daily sets of routine monitoring data were used to determine their maximum values. Prognostic schemes were constructed separately for the warm and cold seasons of the year. First of all, normalized concentrations of pollutants, B, were calculated with the use of the mean value of concentration of the pollutant in question during corresponding seasons. The list of the "potential" predictors included the wind speed and direction measured at the meteorological station, the depth of the surface inversion of the air temperature and the height of the lower boundary of the elevated one, the synoptic predictor (see definition below), and values of B measured "on the previous day". The censoring of the samples was carried out using the criteria P > 0.2. The predictors were transformed to normal ones using the standard procedure described in numerous textbooks on mathematical statistics (see for example *Pugachev*, 1979). The transformation of nonlinear dependencies in linear ones was done using the following expression:

$$[X]=P\{B|X\}, \qquad (1)$$

where *P* is the symbol of the mathematical expectation and B|X means the value of B corresponding to given X. In particular, the synoptic predictor, S, is defined as [SS] where SS is one of the number of "typical" synoptic situations selected "manually" from the analysis of the surface pressure maps.

The efficiency of the steps from (1) to (4) is illustrated on Fig. 1. The versions of the prognostic models and corresponding indicators of their performance presented on different panels of this figure are described in Table 1. Here, $\langle B \rangle$ and σ are mean values and standard deviations of maximum dimensionless concentrations, "meas" and "pred" correspond to measured and predicted B, Corr means the coefficient of the linear correlation between measured and predicted values of B; mean values and standard deviations of the rations of measured to predicted B are given in two last columns. The results of validation of these models upon independent data sets showed a reasonably good performance with correlation coefficients around 0.7 and predictability around 80%. An example of validation of the prognostic model on independent data is given in Table 2.

Table 1. Comparison of the efficiency of different versions of the prognostic model

Panel	Transformation (step)	Sample volume	$\langle B \rangle_{meas}$	$\langle B \rangle_{pred}$	σ_{Bmeas}	σ_{Bpred}	Corr	$P_{(meas/pred)}$	$\sigma_{(meas/pred)}$
a	-	160	1.82	1.81	2.26	1.22	0.54	1.10	1.61
b	(2)	70	2.82	2.82	2.62	1.51	0.58	1.10	0.91
c	(2)+(4)	70	2.82	2.77	2.62	1.64	0.78	1.09	0.65
d	(2)+(3) + (4)	70	2.82	2.82	2.62	1.96	0.84	1.05	0.57

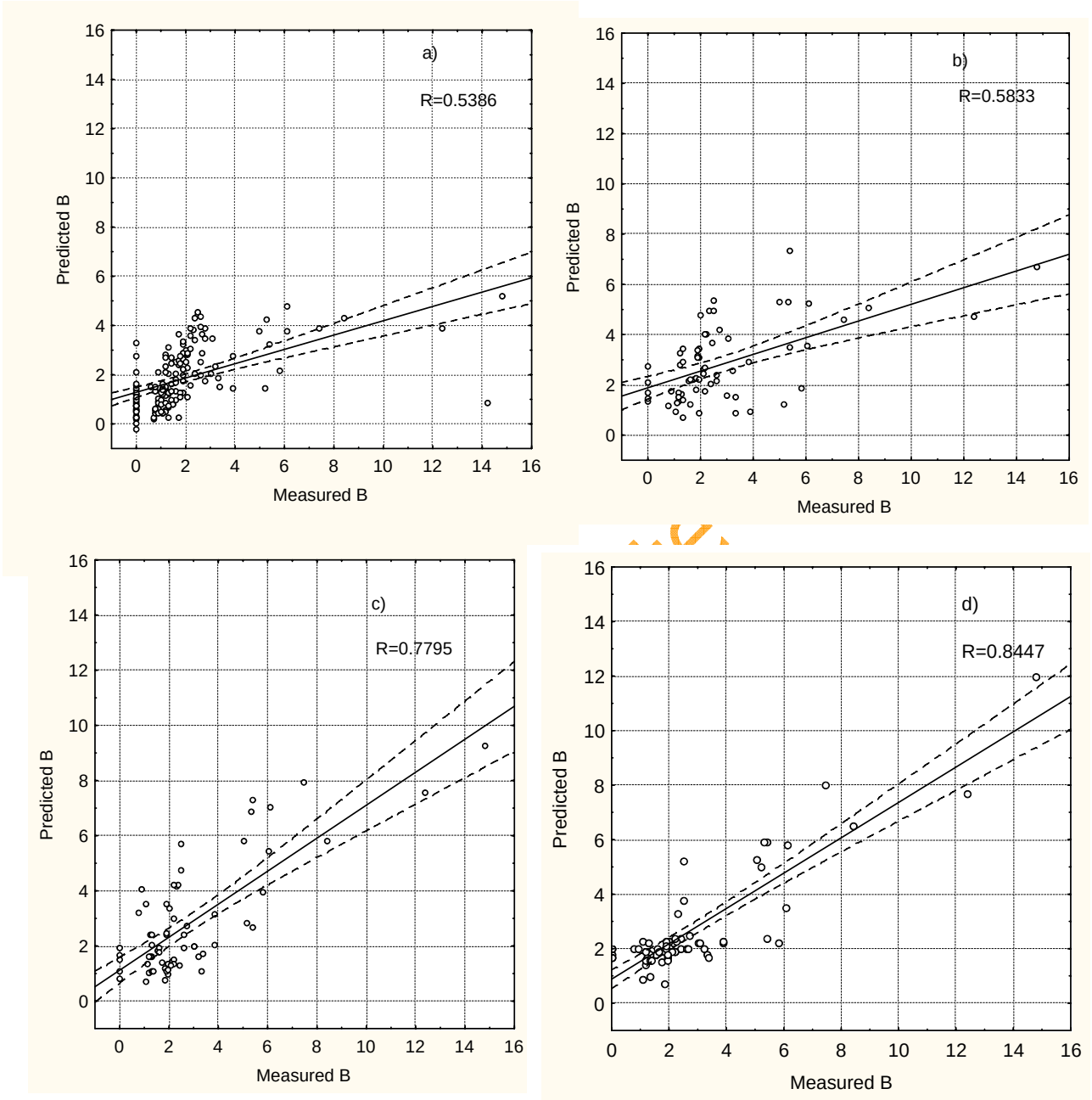

Fig. 1; Predicted vs. measured dimensionless maximum concentrations. Variants of the transformations in use are listed in Table 1. Solid lines represent linear regressions ("best fit") and dashed lines characterize the 95% confidence intervals for these regressions

A prognostic model of daily maximum concentrations of ozone was developed using data of the automatic monitoring system which consists of two open-path gas analyzers that monitor concentrations of SO_2, NO, NO_2, benzene, toluene and ozone. The instruments are located at the heights of 3 m ("street level") and 15 m ("roof level") in the street canyon in downtown of St. Petersburg. The system has been operating since 1998. Its description and results of data analysis are presented by *Genikhovich et al.* (2005).

Table 2. Dimensionless concentrations of carbon disulfide B_{meas} (numerator) and B_{pred} (denominator) for days when high concentrations were observed in Krasnoyarsk at four or more monitoring stations

Date	Monitoring station							
	1	3	5	7	8	9	20	21
7.07.84	6.32 / 2.46	4.11 / 5.11	4.32 / 3.42	5.03 / 5.81	------	5.24 / 5.20	4.30 / 4.81	8.22 / 4.60
20.07.84	0.00 / 1.34	0.00 / 1.48	0.000 / 1.02	5.82 / 3.82	5.67 / 3.44	6.41 / 3.62	9.19 / 3.18	0.00 / 2.00
21.07.84	2.35 / 3.48	1.57 / 2.60	6.68 / 4.59	6.10 / 7.05	4.34 / 4.89	6.34 / 6.44	9.96 / 6.17	3.17 / 4.16
03.08.84	9.10 / 4.88	4.68 / 6.39	0.00 / 7.26	7.45 / 7.91	9.00 / 6.02	5.58 / 8.75	6.92 / 11.62	7.56 / 5.09
04.08.84	4.61 / 6.10	6.82 / 7.03	6.28 / 5.69	5.39 / 7.28	7.95 / 6.64	4.93 / 6.65	4.92 / 8.32	3.87 / 6.31
02.07.85	6.09 / 4.12	16.11 / 9.03	21.83 / 10.60	8.41 / 6.54	13.61 / 9.33	25.46 / 12.10	24.60 / 10.96	9.71 / 5.68
03.07.85	6.18 / 5.37	15.87 / 9.32	13.83 / 10.91	14.86 / 12.0	4.00 / 4.51	12.30 / 10.13	16.11 / 11.35	-----

The following parameters were used as an initial set of predictors in the prognostic model:

O_3 – the previous daily maximum of ozone concentrations;
T – the air temperature;
U – the wind speed;
D – the wind direction;
NO_2 – the nitrogen dioxide concentration.

The notation "7" is also used further to mark the predictors corresponding to measurements at 7 a.m. on the prognostic day. The synoptic predictor, S, is not used in the preliminary prognostic scheme presented in this paper. It has been found, however, that there is an evident correlation between synoptic situations and maximum ozone concentrations. In particular, the highest concentrations were observed in the summer in the cases with weak pressure gradients accompanied by high air temperatures and thunderstorms in the region of St. Petersburg; low ozone concentrations corresponded to the cold and rainy weather when the city was situated at the NW-, N-, and NE periphery of the cyclone. That is why one can expect a certain improvement of the model performance when including S in the list of predictors.

The scatter plot of predicted vs. observed daily maximum ozone concentrations is shown on Fig. 2 (left-hand panel). The regression line drawn there corresponds to the coefficient of correlation of 0.76. This result was obtained with the use of O3, NO2, V as well as T7 and D7 as predictors. When tested upon independent data, however, the performance of this prognostic scheme was damaged due to the absence of the efficient procedures for censuring the low ozone concentrations. Still, when measured ozone concentrations were higher than 70 $\mu g/m^3$, the intervals of predicted and measured concentrations were closely overlapping one with another (with the correlation coefficient at the level of 0.45). The right-hand panel on Fig. 2 demonstrates that the prognostic scheme can be significantly improved when predicting the ratio $(O_3 - O_37)/O_37$ rather than ozone itself.

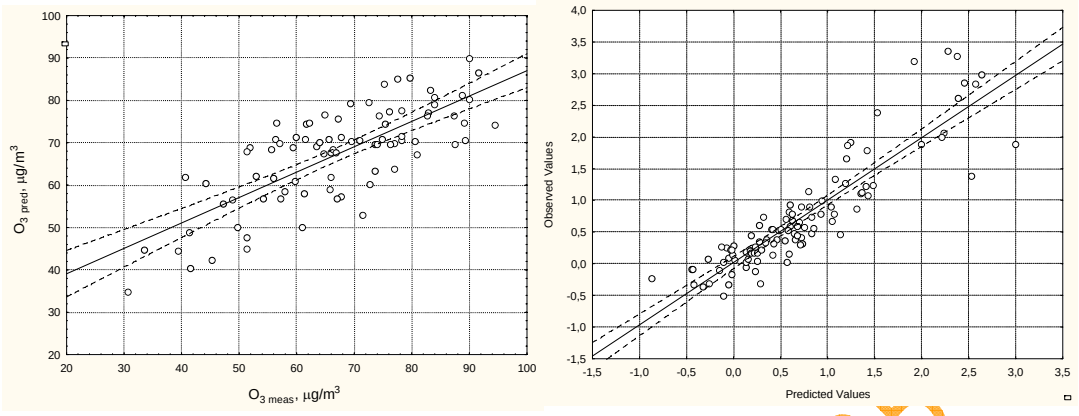

Fig. 2; Predicted vs. observed values of the daily maximum ozone concentrations (left-hand panel) and their relative deviations from concentrations measured at 7 a.m. (right-hand panel).

CONCLUSION

The stochastic forecasting models of air pollution have obvious deficiencies in comparison with deterministic ones. In particular, it is difficult to extrapolate the prognostic results to the whole urban area outside the locations of the monitoring stations. However, stochastic models can be easier initialized because they do not use any information about the actual emissions of primary and secondary pollutants. They seem to be simpler in practical applications and provide much wider opportunities to using the data of monitoring of the urban air pollution. When validated upon experimental data, forecasts with stochastic models usually outperform those with deterministic ones, and results presented in this paper could be considered as a proof of such a statement. Finally, it should be stressed out that in the near future the approaches based on deterministic and stochastic dispersion modeling could be "unified" in the systems of the hybrid monitoring of urban air pollution.

REFERENCES

Berlyand, M. (1991) Prediction and Regulation of Air Pollution, Kluwer/A.P.

Genikhovich, E.L. (1995) Practical applications of regulatory diffusion models in Russia. *Intern. Journal of Environment and Pollution*, vol. 4-5, No 4-6, 530 - 537

Genikhovich, E. (2003) Indicators of performance of dispersion models and their reference values. *Intern. Journal of Environment and Pollution*, Vol. 20, Nos. 1-6, 321-329

Genikhovich, E.L, A.D. Ziv, E.A. Iakovleva, F. Palmgren, R. Berkowicz (2005) Joint analysis of air pollution in street canyons in St. Petersburg and Copenhagen. *Atmospheric Environment*, **39**, No 15, 2747-2757

Pugachev, V.S. (1979) Theory of Probabilities and Mathematical Statistics, Nauka Publishers, M. (in Russian)

Sonkin, L.R. (1991) Synoptical and Statistical Analysis and Short-Term Prediction of Air Pollution, Hydrometeorological Publishers, L. (in Russian)

Sonkin, L.R., Nikolaev, V.D., Ivleva, T.P., Kirillova, V.I. (2002) Forecasts of extremely high levels of air pollution in cities and regions. In: Problems of Atmospheric Boundary-Layer Physics and Air Pollution (Ed. S.S. Chicherin), Hydrometeorological Publishers, SPb, 310 – 322 (in Russian).