On the Use of Air Quality Monitoring Networks for the Evaluation of Nitrogen Oxide Emission Inventories

I.B. Konovalov[1,2] and M. Beekmann[*,2]

[1]*Institute of Applied Physics, Russian Academy of Sciences, Nizhniy Novgorod, Russia*

[2]*Laboratoire Interuniversitaire des Systèmes Atmosphériques (LISA) CNRS, Université ParisEst et Paris 7, Créteil, France*

Abstract: The usefulness of ground based air quality monitoring data for diagnostics of uncertainties in gridded emission inventories is examined. A general probabilistic procedure for comparison of levels of uncertainties in different emission datasets is developed. It implies the evaluation of the agreement between modeling results obtained with these emission datasets and corresponding measurements. This procedure is applied to the evaluation of different datasets for European gridded nitrogen oxide (NO_x) emissions by using the AirBase monitoring data and the CHIMERE chemistry-transport model. Numerical experiments are performed for two different types of spatial distributions of emission uncertainties and five different types of monitors. The results are also generalized for various levels of uncertainties in simulated and measured data. It is found, in particular, that most informative, from the point of view of diagnostics of NO_x emission uncertainties, are the measurements of NO_2 at rural background sites and measurements of ozone at suburban sites situated in the vicinity of intensive sources of emissions. A more precise conclusion regarding the relative accuracy of two emission datasets can be drawn with a larger number of monitors in a network and a higher accuracy of the model and measurements. For example, with a network of 50 rural background NO_2 monitors, the probability of choosing the more certain emission data set is more than 90 percent, if differences in uncertainty of two sets are more than 50 percent. Practical recommendations for designing or evolving surface measurement networks, in light of the study results, are given.

Keywords: Nitrogen oxide emissions, tropospheric ozone, monitoring networks, inverse modeling.

1. INTRODUCTION

Emissions of gases and particulate matter into the atmosphere are one of the major factors controlling the atmospheric composition and its changes. Emission inventories provide inputs into the atmospheric models, which are widely used in order to understand complex physical and chemical processes in the atmosphere, to develop air quality management strategies and to predict variations in climate. Numerous studies pointed out the uncertainties in available emission data as an important source of inaccuracies in model results [1-8]. Thus it is not surprising that significant efforts are devoted to the development of emission inventories on urban regional, continental and global scales [9-17]. A common way of elaborating emission estimates involves the processing of available statistical information regarding different activities, such as industry, energy production, transport and others. An alternative approach is based on the inverse methods, which are intended to optimize emission parameters in atmospheric models by reducing the difference between simulated and measured concentration fields. Because neither of these approaches is free from potential uncertainties in resulting emission estimates, and because different assumptions and input data used in emission inventories may lead to substantially different emission estimates, it is important to have an independent way for validating and/or comparing different emission data.

It is not uncommon that available emission estimates are validated by comparing simulations performed with a model (in which the respective emission data are used as inputs) with independent measurements [18-21]. A general goal of such validation is to demonstrate that one emission dataset is more accurate than another. It was also suggested [22-24] that independent measurements can be used for optimization of emission estimates obtained by means of inverse modeling. Such optimization can be helpful because weights assigned to a priori information and measurements in Bayesian inversion schemes are usually quantified by parameters that are poorly known. The subjective estimation of the weights may lead to uncontrollable uncertainties in the a posteriori emissions.

However, it is easy to realize that improving agreement of model results with measurements after emission optimization does not necessarily mean that emissions used in the model are also improved. Indeed, it is probable that even perfect emission data can be "optimized" further so as to compensate model and measurement errors. Therefore, there is a general problem concerning the significance of an improvement in agreement between simulated and measured concentration fields as a measure of improvement of an emission inventory. To the best of our knowledge, this problem has not been yet addressed in a peer-reviewed literature.

*Address correspondence to this author at the LISA, UMR 7583 CNRS, Univ. Paris12 & 7, 61 Av. Général de Gaulle, 94010 Creteil Cedex, France; E-mail: beekmann@lisa.univ-paris12.fr

1874-2823/08 2008 Bentham Open

In this paper, we propose a general statistical framework that permits quantification of statistical significance of a conclusion regarding the relative accuracy of two (or more) emission datasets, which can be made after comparing the measured and simulated data. More specifically, a main goal of this study performed in the framework of the GeoMon EC/FP6 project is to quantify the usefulness of ground based monitoring of nitrogen dioxide (NO_2) and ozone (O_3) near-surface concentrations for selection of the best dataset of nitrogen oxide (NO_x) emissions among from those available. A complementary goal is to optimize parameters of a surface network (e.g. density and location of monitors), having in mind its possible application for diagnostics of NO_x emission inventories. In particular, the dependence of results on the number and type (rural, urban, etc.) of sites will be discussed. Nitrogen oxides are known to be among the major pollutants responsible for photochemical smog events. They also have a strong impact on the oxidizing capacity of the troposphere, and in this way they can affect concentrations of some green-house gases, such as methane and ozone. Studying NO_x emissions is made easier due to availability of sufficiently accurate satellite measurements of tropospheric NO_2 column amounts (see, e.g., [25, 26]) which allow elaboration of measurement based NO_x emission estimates [20, 23, 24, 27-29]. The statistical framework discussed here can be used in studies of emissions of other species, such as carbon dioxide, methane, carbon monoxide and aerosols, which are intensively measured both by the ground based networks and from space.

2. METHOD

2.1. Problem Definition and Summary of the Method

Let us assume that we have two datasets for seasonally averaged gridded NO_x emissions. These emissions are used as inputs for a chemistry transport model which can provide simulated concentrations of certain species. Let us assume also that we have a dataset of measured concentrations of the same species. We can further define and evaluate some measure of agreement between the simulated and measured concentrations. Ideally, it could be expected that more accurate emissions would provide better agreement between the simulations and measurements. However, because of model and measurement errors, the result may be opposite. Such a situation is illustrated in Fig. (**1**), where it is assumed that a model is perfect but a measurement is not: the emissions $\boldsymbol{E_2}$ are obviously less accurate but produce better agreement with measurements than emissions $\boldsymbol{E_1}$.

In the following, we do not separate the model and measurement errors, but consider them in combination, as it is common in inverse modeling studies (see, e.g., [30] (It is indeed hardly possible to quantify them separately, particularly because observed differences between simulations and measurements may include a representativeness error, which can be attributed both to a model and measurements). The problem is therefore to quantify the probability of the erroneous conclusion regarding the relative accuracy of a pair of emissions datasets, given a combined model and measurement error and relative difference in the accuracies of the emission datasets. It is reasonable to expect that the decision - error probability will be smaller in situations where differences in the accuracies of the emission datasets are larger. The solution presented here is based on the Monte Carlo method. We use the data of a typical "bottom-up" emission inventory (here, this is the EMEP emission inventory [16]) as a substitute for unknown true emissions. Such emissions are referred below to as the standard emissions and the model results obtained with such emissions are referred to as the reference case. We consider all other possible emission data sets as random realizations from statistical ensembles characterized by the assumed level of uncertainties in a given emission inventory. In order to simulate such a statistical ensemble, we add random perturbations (which are independent of time) to the standard emissions. It seems clear that by adding random perturbations to gridded emissions, we probably get less accurate emissions. All emission datasets obtained in such a way are used to simulate near surface concentrations of NO_2 and O_3, which are then compared with corresponding measurements. Finally, we assess the required probability of the decision error as a fraction of cases, for which the perturbed (i.e. worse) emissions produced better results in terms of pre-defined statistics, such as the mean square error (MSE) or coefficient of determination (R^2). This is done for different set-ups of surface networks (species, number and type of sites). The numerical experiments were performed with the CHIMERE chemistry transport model (http://euler.lmd.polytechnique.fr/chimere/). We

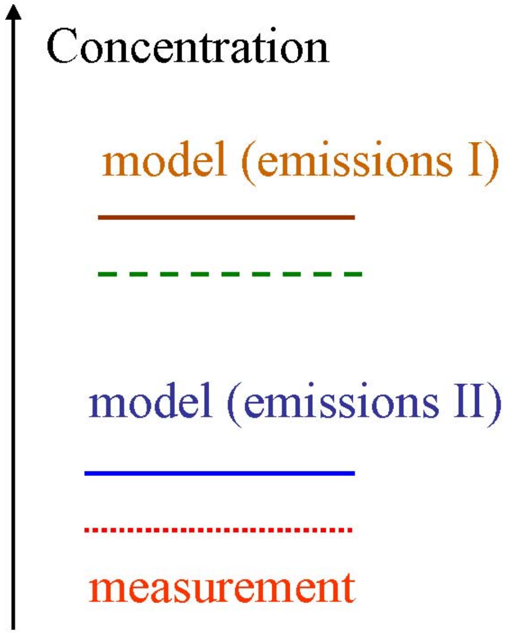

Fig. (1). An illustration demonstrating that a model (which is assumed here to be perfect) with better emissions (E_1) does not necessarily yield better agreement with an imperfect measurement than the model with biased emissions (E_2). It is easy to imagine a similar situation when there are both model and measurement errors.

used one of the standard CHIMERE domains, CONT3, which covers the Western and Central Europe with a resolution of 0.5^0x0.5^0 and includes 3082 grid cells. The meteorological conditions and the anthropogenic emissions were specified for the summer season (June-August) of 2005.

2.2. Basic Definitions and Formulations

Let E_0 and E_t be the vectors of the standard gridded NO_x emission rates assigned in the model and true (unknown) emissions, respectively. Here and below, all emission vectors represent average NO_x emission rates over three summer months (June-August). Taking into account that emissions are strictly positive values, it is more convenient to deal with their natural logarithms: $e_0=e_t+\varepsilon_0$, where ε_0 is the emission estimate error. Then the uncertainty of the standard emissions can be quantified by means of the following variance:

$$U_0^2 = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_{0i}^2 \tag{1}$$

where i is the index of a grid cell and N is the total number of grid cells. Following the probabilistic approach, we consider uncertainties of a given dataset of emissions as a sample from some probability distribution having a certain mean, $<\boldsymbol{\varepsilon_0}>$ with <> denoting the average over the ensemble), and variance, $\boldsymbol{\sigma_0}^2$. Accordingly, U_o^2 should also be considered as a random variable. If the probability distribution of $\boldsymbol{\varepsilon_0}$ (vector of ε_{0i} values for grid points i=1,…N) is close to the normal distribution (a reasonable assumption), then the variance of U_o^2 is, asymptotically, inversely proportional to N. This observation means that for typical grids used in emission inventories with $N>10^3$, the variance of U_0^2 is likely quite negligible in comparison with the value of U_0^2 itself, and any sample value of U_0^2 characterizes the average variance of $\boldsymbol{\varepsilon_0}$:

$$U_0^2 \cong \frac{1}{N}\sum_{i=1}^{N}\sigma_{0i}^2 \tag{2}$$

We generate the perturbed emissions, $\boldsymbol{e}_p$, by adding random perturbations, $\boldsymbol{\varepsilon_p}$, to $\boldsymbol{e_0}$. After repeating this operation many times, we can get a sufficiently large statistical ensemble of the perturbed emission datasets, $\{\boldsymbol{e}_p\}$. The uncertainty of a given set of the perturbed emissions can be quantified similar to Eq. (1):

$$U_p^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\varepsilon_{0i}+\varepsilon_{pi}\right)^2 \tag{3}$$

Let $\boldsymbol{\sigma_p}^2$ be the variance of $\boldsymbol{\varepsilon_p}$ and θ_p^2 be the average of $\boldsymbol{\sigma_p}^2$ over the grid:

$$\theta_p^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_{pi}^2 \tag{4}$$

If $\boldsymbol{\varepsilon_0}$ and $\boldsymbol{\varepsilon_p}$ are statistically independent variables (we can require this by definition), then, using Eqs. (2), (3) and (4), we obtain:

$$U_p^2 \cong U_0^2 + \theta_p^2 \tag{5}$$

This relation means, in particular, that the perturbed emissions are more uncertain than the standard emissions. It should be kept, in mind however, that this conclusion is valid only if the covariance of $\boldsymbol{\varepsilon_0}$ and $\boldsymbol{\varepsilon_p}$ is, on average, much smaller than both U_0^2 and θ_p^2. This would not be the case if, for example, both the standard emissions were uniformly biased and $<\boldsymbol{\varepsilon_p}>$ were not zero.

Similar to (5), we can derive:

$$\theta_p^2 \cong \frac{1}{N}\sum_{i=1}^{N}\left(e_{pi}-e_{0i}\right)^2 \tag{6}$$

The ensemble of emission datasets $\{\boldsymbol{E_p}\}$ can be used in the model to generate a corresponding ensemble of simulated concentrations, $\{\mathbf{C_m}\}$, which are extracted from model results to be linked with available observations, $\mathbf{C_o}$. The components of the matrixes $\mathbf{C_m}$ and $\mathbf{C_o}$ represent concentrations of a given species from a set of certain locations and days.

The agreement between $\mathbf{C_m}$ and $\mathbf{C_o}$ can be measured by means of some standard statistics, Z. Here we employ two statistics, defined as follows:

$$Z_1 = \frac{1}{LM}\sum_{k=1}^{L}\sum_{j=1}^{M}\frac{\left(C_m^{jk}-C_o^{jk}-\overline{C}_m^k+\overline{C}_o^k\right)^2}{\overline{C}_o^{k2}} \tag{7}$$

$$Z_2 = \frac{1}{L}\sum_{k=1}^{L}(1-R_k^2) \tag{8}$$

where k and j are indexes of the monitoring station and of the day of observation, respectively, L and M are the total numbers of the monitoring stations and the days of observations considered, R^2 is the coefficient of determination defined for time series of the measured and simulated concentrations at a given location, and a horizontal bar above the symbol of concentration denotes the average over the entire period (here, three summer months) considered. We consider daily average and daily maximum concentrations for nitrogen dioxide and ozone, respectively. The first statistics is the centered and normalized mean square error (NMSE), which is an analog of the root mean square error (RMSE), while the second statistics is based on the standard coefficient of determination. While choosing these statistics we took into account that RMSE is one of the most commonly used statistics and that the normalization enables, at least to some degree, equalizing contributions from locations with different levels of air pollution. Accounting for the bias in the simulated concentrations is intended to accentuate the differences in temporal variations of concentrations, rather than discrepancies of their means. The coefficient of determination is also a quite commonly used statistics, and it is also expected to be rather insensitive to differences in the mean concentrations. Meanwhile, it is known that RMSE and R^2 may lead, sometimes, to rather different conclusions [31-33]. Accordingly, our experiments allow us to test to what extent the considered probability of the decision error is sensitive to the choice of the comparison statistics.

Since the perturbed emissions are more uncertain (see Eq. (5)), we could expect that both Z_1 and Z_2 calculated with the perturbed emissions would be larger than the same statistics quantified with the standard emissions. That is, ideally, we should always get $Z(\boldsymbol{E_p})>Z(\boldsymbol{E_0})$. However, because of model and observation errors, it may not always be the case. Accordingly, our task is to quantify the probability, ρ_{err}, of obtaining $Z(\boldsymbol{E_p})<Z(\boldsymbol{E_0})$. If the comparison of simulated and measured concentrations is used to identify the best emission inventory, then ρ_{err} can be considered as the probability of a wrong conclusion regarding the comparative accuracy of two databases, one of which has larger uncertainty than the other. More formally,

$$\rho_{err}=p(Z(\boldsymbol{E_2})<Z(\boldsymbol{E_1})|\ U_2^2-U_1^2=\theta_p^2) \qquad (9)$$

where $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$ is a pair of emission datasets with unknown uncertainties U_1 and U_2 defined similarly to (1) and (2). In other words, given an estimate θ_p of the difference in uncertainties in the pair of available emission datasets $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$ and the fact that $Z(\boldsymbol{E_1})>Z(\boldsymbol{E_2})$, the conclusion that $\boldsymbol{E_1}$ is more uncertain than $\boldsymbol{E_2}$ may be wrong with the probability equal to ρ_{err}.

In the framework of the method considered here, the probability ρ_{err} is simply estimated as the fraction of cases, for which $Z(\boldsymbol{E_0})$ is larger than $Z(\boldsymbol{E_p})$. In other words, given the ensemble $\{\boldsymbol{\varepsilon}_p^1, \boldsymbol{\varepsilon}_p^2,\ldots, \boldsymbol{\varepsilon}_p^K\}$ of K perturbations of the standard emissions $\boldsymbol{E_0}$ and the number J of the cases for which $Z(E_0)>Z(E_p)$, the decision-error probability ρ_{err} is estimated as follows:

$$\rho_{err} \cong \frac{J}{K} \qquad (10)$$

It may be useful to note that θ_p can be estimated from the above as:

$$\theta_{p\max}^2 \cong \frac{1}{N}\sum_{i=1}^{N}\left(e_{2i}-e_{1i}\right)^2 \qquad (11)$$

This estimation (see also Eq. (6)) stems from the assumption that the emission datasets contain both some common and independent errors. It is easy to see that the root mean square difference of independent errors equals θ_{pmax}. Along with the statistics defined in accordance with Eq. (7) and (8), we have considered also a somewhat different statistics. The aim is to evaluate the statistics Z_1 and Z_2 for each site separately (L=1) and then to count the number of sites for which $Z_{1,2}(\boldsymbol{E_p})-Z_{1,2}(\boldsymbol{E_0})>0$. The ratio of this number to the total number of sites gives us a kind of a nonparametric statistics. On average, this ratio should be larger than 0.5. The decision-error probability can then be estimated as a fraction of cases for which this ratio is smaller than 0.5.

2.3. Numerical Experiment Settings

Obviously, the response of the concentration fields to emission perturbations may depend not only on the magnitude of these perturbations but also on their spatial structure. Since we cannot consider here all imaginable distributions of emission uncertainties, in order to test the role of this factor we considered two different cases denoted below as RAN and COV.

The RAN case represents a situation when the emission uncertainties in different grid cells are statistically independent. The standard deviation of emission perturbations in each grid cell is defined to be the same, and equals θ_p.

The COV case is intended to test the role of spatial covariances in emission uncertainties. Such covariances may arise due to systematic biases in assumed emission factors or activity data. The nature and magnitude of such covariances may be different for different emission inventories. Here we assumed that the emission uncertainties are the same for the same country and the same SNAP (Selected Nomenclature for Air Pollution) sectors. Accordingly, the same random perturbation is applied, in the COV case, to all gridded annual data of the EMEP NO_x emission inventory that is attributed to the same SNAP sector and the same country. Such perturbed annual emissions are then processed in a standard way (see Section 3.2) by the CHIMERE emission preprocessor to yield values of total anthropogenic NO_x emissions in a given grid cell. As a consequence, emission perturbations in different grid cells (inside the same country) where most of emitted NO_x is associated with the same type of activity are quite similar.

Note that we also considered a situation when one of the analyzed emission inventories is derived from satellite data by means of inverse modeling. Specifically, we evaluated uncertainties in different grid cells by means of a Monte-Carlo experiment described in [23]. We found that the estimates of ρ_{err} obtained are rather similar to those obtained for the RAN case, and for that reason this situation is not considered below separately.

We performed 30 model runs for each case and 3 different levels of the standard deviation θ_p of the perturbations: 0.1, 0.4 and 0.6. Note that according to our estimations [23], the logarithmic standard deviation of emission uncertainties in the NO_x emissions specified in the CHIMERE CTM (based on the EMEP inventory) is about 0.6. Therefore, simulations with $\theta_p>0.6$ could hardly present any practical interest. It was argued also, in the same paper, that the use of satellite data enables considerable reduction of uncertainties: the uncertainty in the a posteriori emissions was estimated to be about 0.4. This means that a typical value of θ_p corresponding to the situation where the surface measurements are employed in order to prove that the emissions derived from satellite data are indeed more accurate than the a priori emissions is, probably, in the range from 0.4 to 0.5.

Lognormal perturbations to emission rates lead to a shift of the total emissions; that is, the average of perturbations in emissions, $<exp(\varepsilon_{pi})>$, is larger than unity. In order to test the effect associated with this shift, special cases are additionally considered, in which $Z(\boldsymbol{E_p})$ are compared with $Z(\boldsymbol{E_0}<\exp(\boldsymbol{\varepsilon_p})>)$. That is, in these cases, we scaled the standard emissions, such that the average of random perturbation, considered relative to the scaled emissions, would be zero.

2.4. Measurements

We use the data of measurements of near surface concentrations of nitrogen dioxide and ozone obtained from the AirBase air quality database system (http://air-climate.eionet.europa.eu/databases/airbase/). AirBase contains air quality monitoring information submitted by the participating European countries. NO_2 and O_3 measurements are reported on daily and hourly basis, respectively. We considered only those monitors that provided data for at least 90 percent of the days in the considered period (summer of 2005). In the case of ozone measurements, a day was accounted for only if at least 22 hour data were provided. Based on these criteria, we selected 1893 NO_2 monitors and 1395 O_3 monitors.

Originally, the monitors were classified according to the zone of location and the type of environment. The zones defined in the AirBase stations are rural, suburban and urban, while the types are background, industrial and traffic. Taking into account the goals of our study, we considered the following five categories: (1) rural background, (2) suburban background, (3) urban background, (4) suburban-sources, and (5) urban-sources. That is, we excluded the rural stations situated near industrial units and big roads, and do not make the distinction between "industrial" and "traffic" monitors. The location of selected monitors is shown in Fig. (**2**).

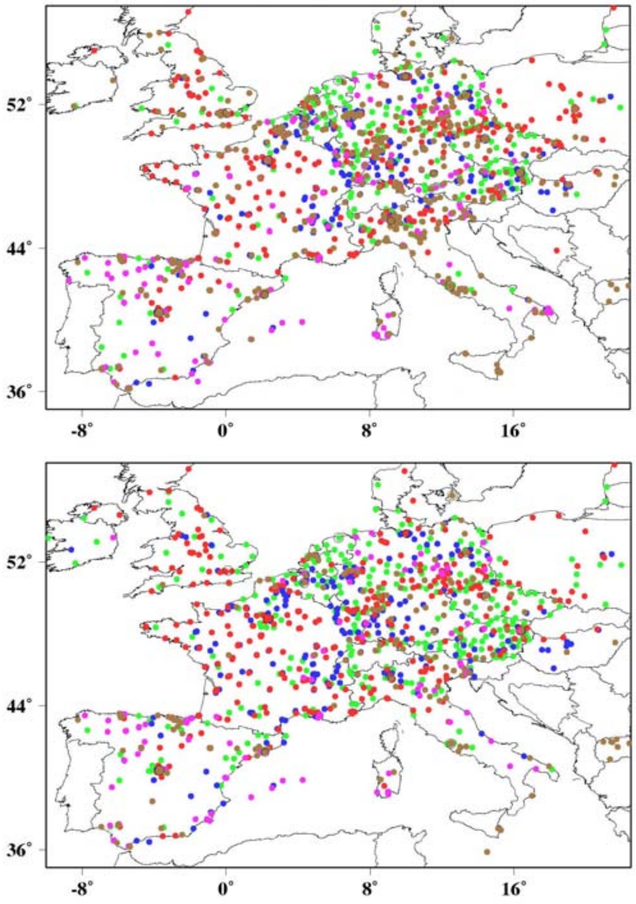

Fig. (2). The locations of the selected NO_2 (upper panel) and O_3 (lower panel) monitors presented in the Airbase database. Rural background, suburban background, urban background, "suburban-sources" and "urban-sources" monitors are marked in green, blue, red, purple, and brown, respectively.

2.5. Simulations

We used the CHIMERE CTM [34], which is a Eulerian 3D model designed to simulate and predict air pollution on the urban and continental scales. This model has already been successfully used in numerous studies, and so we mention below only its basic features essential for this study. A detailed description of CHIMERE and related references are available on the web at http://euler.lmd.polytechnique.fr/chimere/. We used the latest version (V200709) of CHIMERE available at the time when this study was started, which features the support of parallel computing, NetCDF input/output interface and accounting for deep convection. In this study, the simulations are performed on $0.5^0x0.5^0$ horizontal resolution, with 12 layers defined as hybrid coordinates; the top of the upper layer is fixed at 200 hPa pressure level. Meteorological input data were calculated off-line with horizontal resolution of 50×50 km^2 using the MM5 non-hydrostatic meso-scale model (http://www.mmm.ucar.edu/mm5/). MM5 was initialised with GFS (Global Forecast Model) rotating forecast data (http://www.cpc.ncep.noaa.gov/products/wesley/ncep_data/). Lateral boundary conditions were prescribed using monthly average values of climatological simulations by the LMDz-INCA model (see http://www-lsceinca.cea.fr/welcome_real_ time.html). The MELCHIOR2 simplified chemical mechanism was used, which includes 44 species and about 120 reactions.

The anthropogenic emission data used in this study are based on the "expert" annual data of the EMEP emission inventory [16] for the year 2005. The annual EMEP emissions are processed by the standard CHIMERE interface to yield hourly emissions for the summer season. Daily, weekly and seasonal factors applied to the annual data are provided for different SNAP sectors by IER, University of Stuttgart [35]. Biogenic emissions of isoprene, pinene and NO are parameterised as proposed in [36], using distributions of tree species on a country basis provided in their work and the inventory of NO soil emissions in [37].

3. RESULTS

The results of our study are presented below in the following order. Firstly, we analyze to what degree different monitors located within a given distance from each other can be considered as independent with respect to the information on NO_x emission uncertainties provided by their measurements. Such analysis allows us to define certain criteria in the monitor selection in order to insure that the results of this study are not sensitive to a specific configuration of a monitoring network. The results of this analysis are also of interest in the context of monitoring network design. Next, we consider some basic characteristics of the probability distribution of the error statistics Z (see Eqs. 7 & 8). In particular, we consider their scaling properties with respect to changes of the number of monitors in a network. After that, we present the main results of this study, which concern the evaluation of the probability of wrong conclusions regarding the relative accuracy of the two emission datasets. Finally, we show how the decision-error probability could change if the level of uncertainties in the simulated and measured concentrations was different.

3.1. Geostatistical Analysis of Spatial Distributions of the Error Statistics

It seems obvious that when two monitors of the same type are situated close to each other, they can only provide essentially the same information as a single monitor. Therefore, the usefulness of the monitoring network for diagnostics of emission uncertainties may depend on the distance between monitors. Rather than considering all possible configurations of the network, it seems more reasonable to focus on the analysis of a network of essentially independent monitors. Accordingly, as a first step, we have to define conditions under which the monitors can be considered as independent. A common way to study the degree of correlation within spatial random fields proceeds by the calculation of the variogram [38]. Spatial scales of variability of air pollution have already been analyzed earlier, but in a different context [39, 40]. For convenience, we consider here a normalised variogram, γ_n, defined as follows:

$$\gamma_n(d) = \frac{\left\langle \left[Z_1(x) - Z_1(y) \right]^2 \right\rangle}{\left\langle Z_1(x)^2 \right\rangle + \left\langle Z_1(y)^2 \right\rangle} \qquad (12)$$

where Z_1 is one of the statistics defined by Eq. (7) and (8), $\boldsymbol{x}$ and $\boldsymbol{y}$ are coordinates of the monitors and $d=\|\boldsymbol{x}-\boldsymbol{y}\|$ is the distance between the monitors, and averaging is performed over the ensemble of runs with randomly perturbed emissions. If the emission perturbations in different grid cells and corresponding perturbations of Z_1 become statistically independent as d increases, the variogram should approach unity. We evaluated γ_n for each possible pair of monitors of the same category. We then calculated a running average over different pairs as a function of d. The running window included 50 data points. The results of the analysis are presented in Fig. (**3**). It can be seen that in the case with spatially uncorrelated emission perturbations (RAN), the monitors can be regarded as being independent from a distance of about 50-70 km in the case of NO_2 and 100-200 km in the case of ozone. Note that our analysis is limited by the relatively coarse resolution (0.5^0x0.5^0) of the CHIMERE simulations. Specifically, it can hardly provide the reliable information on the degree of independence of monitors separated by a distance smaller than the model grid size (~ 50 km).

In the COV case, the critical distance is not so well defined and anyway, significantly larger, as would have been expected, because the distance of covariation of the emission perturbations is also larger. The differences between results for different categories of monitors are small in the RAN case, but rather considerable in the COV case. Evidently, urban monitors are more independent than rural and suburban. This is in agreement with the intuitive expectation that concentrations of nitrogen dioxide and (especially) ozone are more sensitive to changes of local (large) emission rates in urban locations than in rural regions. In contrast to urban areas, concentration fields in rural background locations tend to be determined predominantly by transport rather than by local emissions. We also performed a similar analysis with the Z_2 statistics, but the results were quite similar to those with Z_1 statistics, and so they are not presented here.

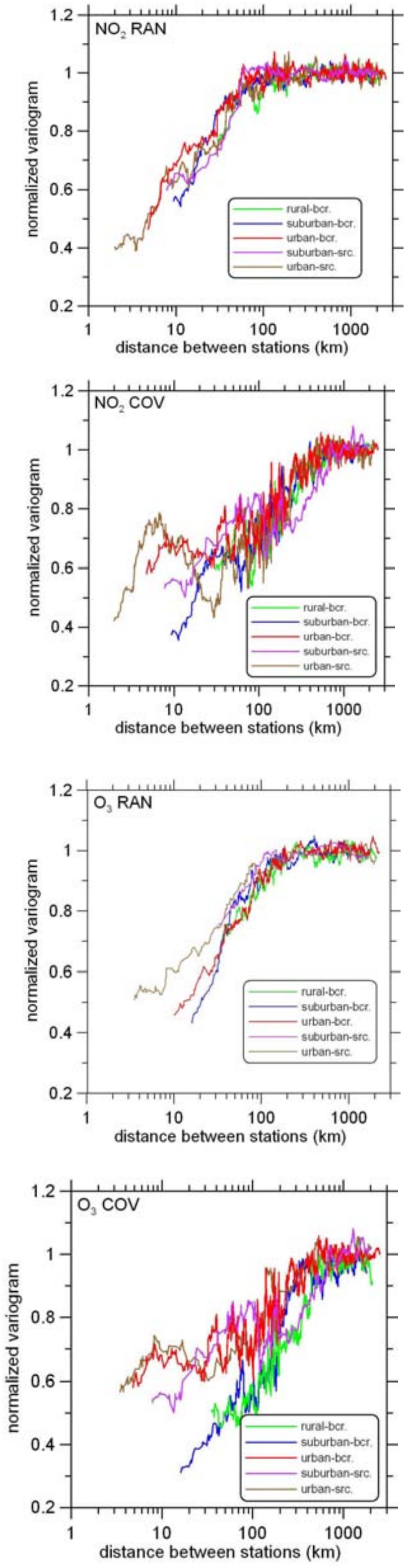

Fig. (3). Normalized variograms (see Eq. (12)) evaluated for perturbations of the error statistics Z_1 caused by random perturbations in the standard NO_x emissions. The results for all different pairs of monitors (not shown) are averaged using the running average method (with 50 data points in the window). Variograms are presented for different pollutants, different types of spatial distributions of emission perturbations, and for different categories of monitors (see legends in the figures).

As a consequence of this analysis, in all experiments discussed below, we consider randomly selected subsets of the monitors, such that the distance between each pair of monitors is greater than 100 km. Such a defined “critical” distance is in agreement with the results discussed above for the RAN case. Unfortunately, the existing network does not allow us to select a sufficiently large number of independent monitors in the COV case. Thus it is necessary to keep in mind that the monitors considered in the COV case are not quite independent.

3.2. Some Properties of Probability Distributions of the Error Statistics

Fig. (**4**) presents examples of the histograms of the difference $Z_1(\boldsymbol{E})$-$Z_1(\boldsymbol{E_0})$ calculated for an ensemble of model runs with the randomly perturbed emissions. We performed 30 “random” runs, but the histograms are also plotted using only 20 runs in order to check if the available sample of 30 values is sufficiently representative of the probability distribution of $Z_1(\boldsymbol{E})$-$Z_1(\boldsymbol{E_0})$. The histograms are presented only for the RAN case, but for all categories of the monitors. Values of Z_1 are evaluated for a set of 50 randomly selected monitors satisfying the requirement defined in the previous section. In this study, we are interested mainly in the negative branch of these distributions as we have to evaluate the probability of $Z_1(\boldsymbol{E})$-$Z_1(\boldsymbol{E_0})<0$. This probability can be roughly estimated as the total area confined by the histogram on the left of the zero line ($Z_1(\boldsymbol{E})$-$Z_1(\boldsymbol{E_0})=0$).

It can be seen that the shape of the distributions for different categories of monitors is rather different, and it hardly resembles any theoretical probability distribution. As evident from the comparison of the histogram calculated with the samples of 20 and 30 values, we can hardly claim that these histograms represent the real probability distributions sufficiently well. Nevertheless, the cumulative probability of $Z_1(\boldsymbol{E})$-$Z_1(\boldsymbol{E_0})$ being negative is rather insensitive to the size of the sample. Furthermore, we made sure that the major conclusions of this study did not change if an ensemble of 20 model runs were considered instead of 30. Based on the results of the test shown in Fig. (**4**) and other similar tests we concluded that the uncertainty in estimates of ρ_{err} obtained with 30 model runs is about 0.1. Accordingly, it is unlikely that the results presented below would significantly change if we performed much more model runs. It is obvious that the decision error probability is related to the width of the distribution of Z: if the width increased from zero to infinity, ρ_{err} would increase from zero to 0.5. In turn, the width of the distribution can be quantified by means of the variance.

It is well known that the variance of the mean of N uncorrelated random variables which have the same standard deviation is inversely proportional to N. Similarly, we can expect that the sample variance of values Z obtained for a set of N independent monitors will decrease proportionally to N^{-1}. Such a decrease indeed shows up in Fig. (**5**), which presents values of the normalized (sample) standard deviation of Z_1, σ_{z1}, calculated with θ_p=0.4:

$$\sigma_{z1} = \left\{ \frac{\left\langle \left[Z_1 - \langle Z_1 \rangle \right]^2 \right\rangle}{Z_1(E_0)^2} \right\}^{1/2} \qquad (13)$$

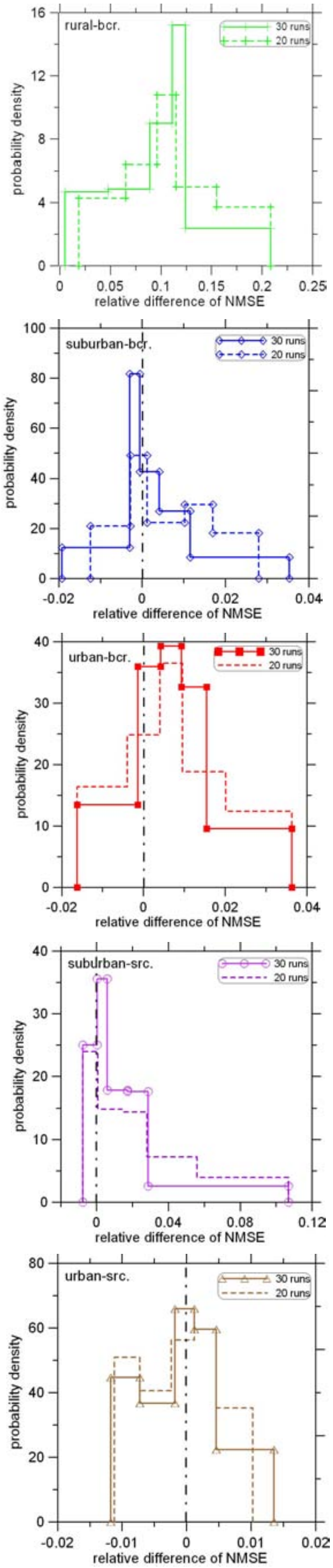

Fig. (4). The histograms of the difference $Z_1(\boldsymbol{E})$-$Z_1(\boldsymbol{E_0})$ calculated for ensemble of 30 and 20 model runs with the randomly perturbed emissions in the RAN case.

Values of σ_{Z1} shown in Fig. (**5**) are calculated as the average over sets of *N* monitors selected randomly from the AirBase network. The process of selection is repeated many times until stable results are reached.

However, when considering differences between the curves for different categories of monitors, it is necessary to be aware that these differences are determined, to a large extent, by the differences in NMSE, which is used as the normalizing factor, rather than differences in the sensitivity of concentration fields to changes in emissions. Note that without the normalization, the magnitude of standard deviation of Z_1 would depend strongly on the level of the measured concentrations, which is very different for different types of monitors.

Ideally, the decision probability error, ρ_{err}, should also decrease as the number of monitors increases. However, this is not necessarily so if $<Z(\boldsymbol{E_p})>$ is smaller than $<Z(\boldsymbol{E_0})>$. Examples of the dependencies of ρ_{err} on *N* (θ_p=0.4) are shown in Fig. (**6**). It is seen that in most cases, the decision probability error does indeed decrease as *N* increases, but in some cases does not. The factors that determine ρ_{err} and the differences between results for different categories of monitors are discussed in the next sections. Here it is important to note that, in majority of cases, the dependence of ρ_{err} on *N*, at least between *N*=10 and *N*=50, is quasi-logarithmic.

This is an important observation because it allows us to simplify any further presentation. Below, we are going to present values of ρ_{err} obtained only with *N*=10 and *N*=50. If necessary, the reader can estimate the decision probability error for intermediate values of *N* using the following approximate relationship:

$$\rho_{err}(N) \cong \frac{(p(10)-p(50))}{\log(5)}(1-\log N)+p(10) \qquad (14)$$

This empirical dependence can also be used in order to roughly estimate ρ_{err} when *N*>50.

3.3. The Decision-Error Probability Estimated Using the Error Statistics Z_1 and Z_2

Figs. (**7**) and (**8**) present estimates of ρ_{err} as a function of the standard deviation of emission perturbations, θ_p, in the case of the NO_2 monitoring network. Not surprisingly, ρ_{err} decreases, in most cases, with the increase of θ_p, although this decrease is, sometimes, rather irregular. That is, the more uncertain the perturbed emissions are, the "easier" they can be distinguished. Unfortunately, due to large computational costs, it was not feasible to estimate ρ_{err} for a larger number of values of θ_p. In Figs. (**7**) and (**8**), results are compared for randomly perturbed emissions and for cases where the reference emissions were scaled to match the average of the perturbed emissions. The difference between straight and dashed curves in these figures thus corresponds to effect of the uniform emission changes.

It is rather unexpected that ρ_{err} can be so sensitive to uniform changes in emissions, in particular because these changes are, on average, much smaller than respective random perturbations. Furthermore, it is seen that the impact of uniform changes in the standard emissions on ρ_{err} can be very different for different categories of monitors and for different error statistics. For example, such a change causes an increase of ρ_{err} for rural monitors but a strong decrease for urban monitors (in case of the Z_1 statistics). The interpretation of these results is hampered by the fact that both

Table 1. Basic Statistical Characteristics of the Measured and Simulated Daily Mean Concentrations of Nitrogen Dioxide and Daily Maximum Concentrations of Ozone Averaged Over All Monitors of a Given Category

Statistics/Monitor Type	NO_2				O_3			
	C_{ob}	C_m	NRMSE	r	C_{ob}	C_m	NRMSE	r
rural-background	8.88	6.89 (8.16)	0.421 (0.459)	0.356 (0.348)	103.6	104.1 (104.8)	0.161 (0.162)	0.779 (0.775)
suburban-background	17.4	8.51 (10.0)	0.370 (0.379)	0.397 (0.385)	102.4	106.7 (107.3)	0.164 (0.165)	0.809 (0.806)
urban background	19.6	7.50 (8.86)	0.348 (0.354)	0.381 (0.367)	97.6	105.0 (105.4)	0.169 (0.170)	0.784 (0.779)
suburban-sources	23.0	6.39 (7.41)	0.379 (0.386)	0.397 (0.380)	100.4	111.1 (111.8)	0.179 (0.182)	0.699 (0.694)
urban-sources	36.7	8.42 (9.91)	0.318 (0.320)	0.385 (0.371)	95.0	112.2 (112.7)	0.192 (0.196)	0.713 (0.707)

The seasonal (June-August) averages of the measured and simulated concentrations, C_{ob} and C_m, are given in units of µg/m^3. The normalized centered RMSE (NRMSE) is calculated in accordance to Eq. (7); r is Pearson's correlation calculated for daily time series. Values of the statistics are provided for the standard emissions and (in brackets) for the COV case (in which the response of concentrations to changes of NO_x emissions is larger than in the RAN case) with θ_p=0.6.

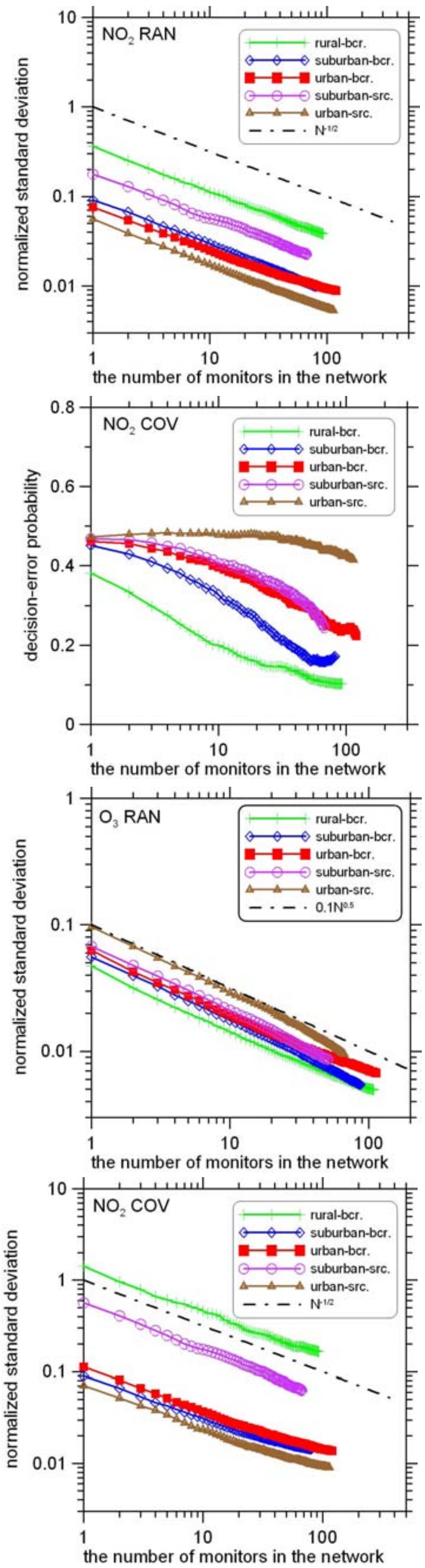

Fig. (5). The normalized (sample) standard deviation of Z_1 (see Eq. (13)), σ_{z1}, calculated with θ_p=0.4. The dashed lines shows the theoretical dependence (σ_{z1}~$N^{-1/2}$); these lines begin here from arbitrary values.

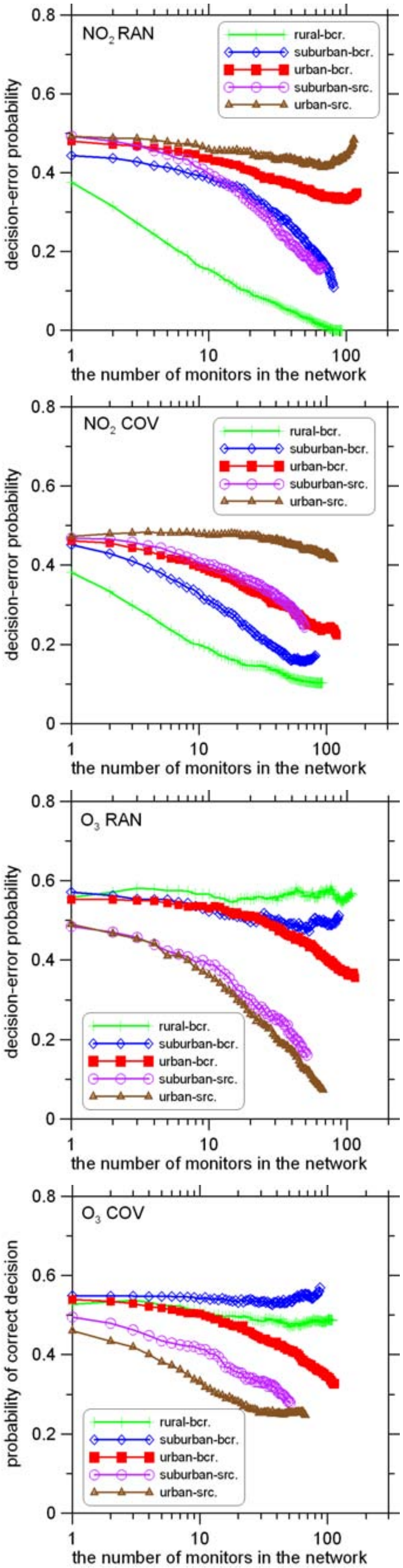

Fig. (6). The dependencies of the decision-error probability on the number of monitors in the network. The dependencies are calculated for the statistics Z_1 and θ_p=0.4.

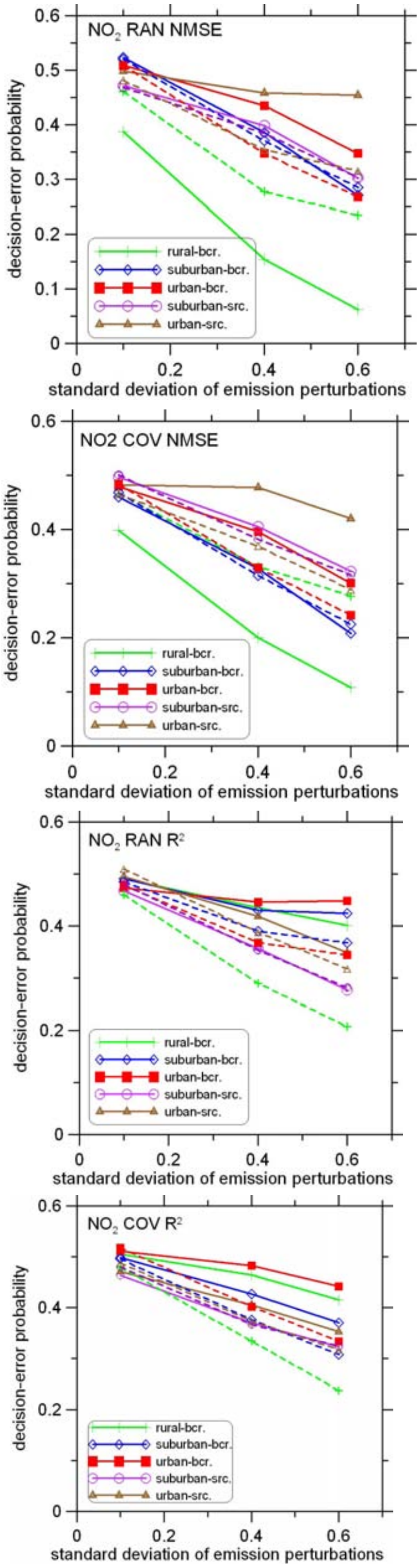

Fig. (7). Estimates of the decision-error probability as a function of the standard deviation of emission perturbations, θ_p, in case of a network of 10 NO_2 monitors. The results are provided for two error statistics Z_1 and Z_2 which are based on NMSE and R^2, respectively (see figure legends). The dashed curves present the results for the cases where the standard emissions were scaled so that to account the change of the total emissions as a result of logarithmic perturbations.

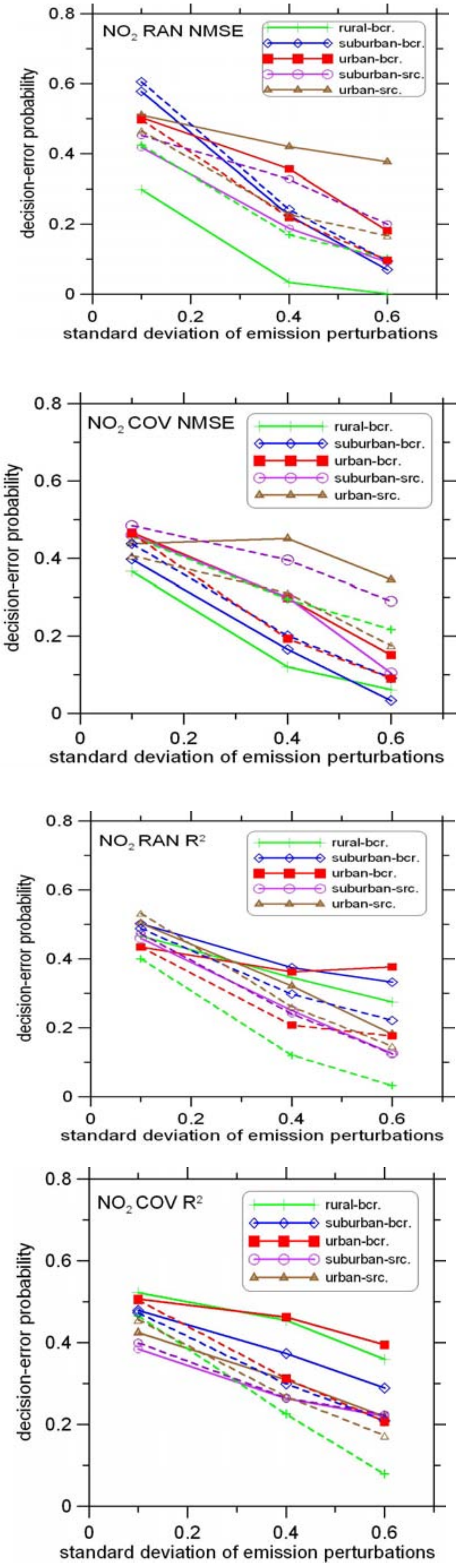

Fig. (8). The same as in Fig. (**7**) but for a network of 50 monitors.

statistics, Z_1 and Z_2, by definition should be relatively insensitive to changes of the mean level of the simulated concentrations which could take place as a result of uniform changes in emissions. Probably, the lower decision-error probability obtained for urban monitors in the cases with the uniformly increased emissions is due to the fact that the model strongly underestimates the observed concentrations at urban areas (see Table **1**). In this situation, the uniform increase of NO_x emissions improves not only the average level of the simulated concentrations but their temporal variability as well (very small average concentrations of NO_2 would be, obviously, associated with small absolute variability). The underestimation of the measured concentrations is much smaller in rural areas, and the uniform scaling of NO_x emissions deteriorates the temporal variability of the simulated concentrations. A part of the sensitivity of ρ_{err} to a uniform emission change may be an artifact of the limited number of simulations; although our preliminary analysis (see Sect. 3.2) indicates that the respective "random" effects should not be large.

The estimates of ρ_{err} obtained using the second error statistics, Z_2 (which is based on the coefficient of determination), are rather different from those obtained with the first statistics. These differences are again not easy to explain, but it is clear that some of them are related to the fact that NMSE is quite sensitive to the differences in absolute values of the simulated and measured concentrations, while R^2 can be large even if one of the concentrations is strongly biased (multiplicatively) but temporal variations of both concentrations are still synchronous. In general, the Z_1 statistics provides better results than the Z_2 statistics with a few exceptions (such as in the case of the "urban-source" category).

The differences between the results for different kinds of spatial distribution of emission uncertainties (RAN and COV) are also noticeable, although, on the whole, not very large. This shows that a spatial covariance in emission uncertainties has little impact on the decision-error probability if it is evaluated using NO_2 measurements.

Figs. (**9**) and (**10**) present the same kind of estimates as shown in Figs. (**7**) and (**8**) but obtained using ozone measurements. It can be seen that the decision error probabilities are in general much larger than for NO_2 monitors. Results for different types of monitors are even more diverse than in the case of nitrogen dioxide measurements. The sensitivity to the changes of the average level of emissions is also much larger. This large sensitivity is, however, not surprising since it is well known that ozone has a relatively long lifetime, and so its concentration in a given grid cell is a cumulative result of emissions over many grid cells. In particular, rural monitors provide ρ_{err} which are about or even larger than 0.5 (when no uniform scaling is introduced). This result (see green solid lines on the figures) can be caused by the fact that CHIMERE tends to underestimate the ozone level when (or where) the observed ozone concentration is high [24]. The increase of the total emission rate over the domain as a result of logarithmic perturbations of emissions increases the ozone formation rate and, consequently, improves the

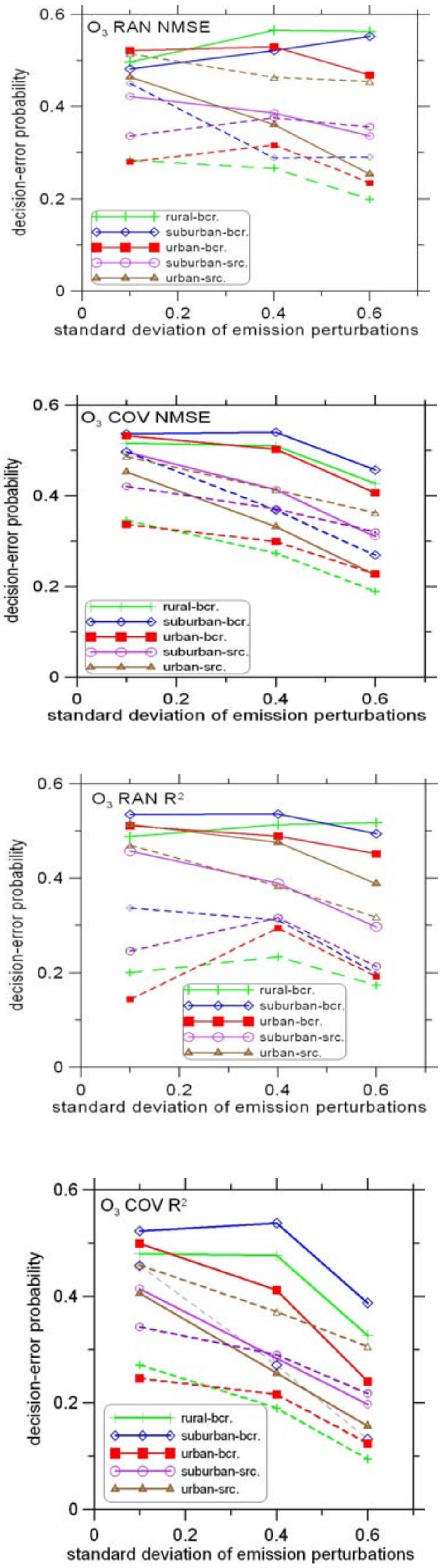

Fig. (9). The same as in Fig. (**7**) but for a network of 10 O_3 monitors.

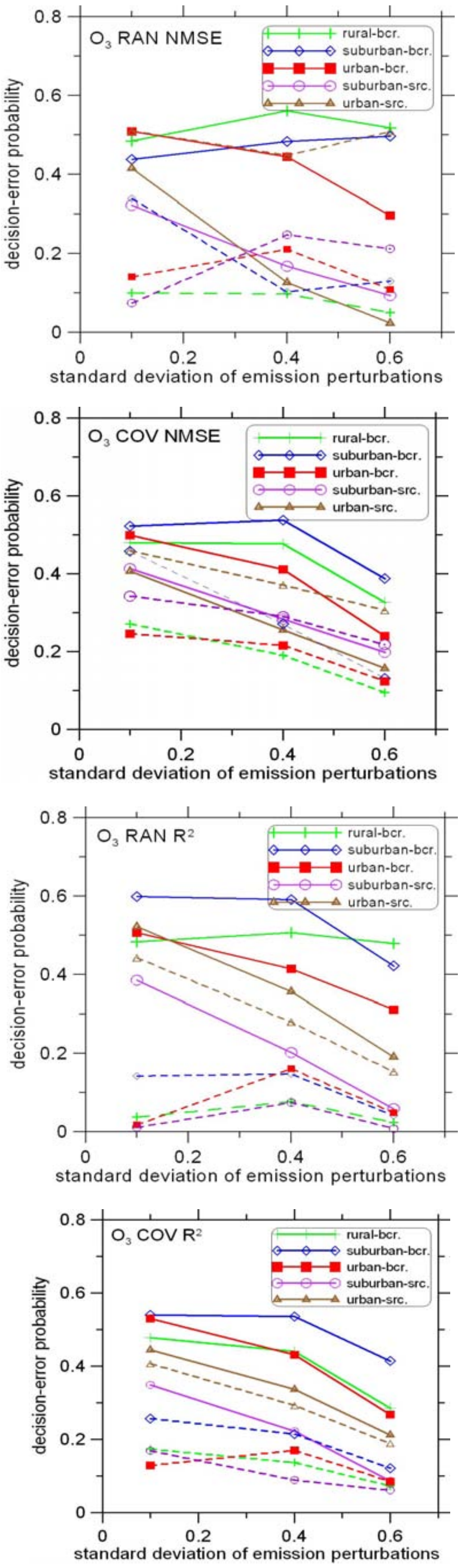

Fig. (10). The same as in Fig. (**9**) but for a network of 50 O_3 monitors.

agreement between the simulated and modeled ozone concentrations. When this change in the total emission rate is compensated (see green dashed lines), ρ_{err} evaluated with the rural monitors becomes lower. In case of "urban-source" monitors, the situation is reverse: the model tends to overestimate the observed ozone level (see Table **1**), since it insufficiently resolves large point sources that control ozone concentration *via* titration. Accordingly, emission perturbations worsen the performance of the model, as it was intended. The estimates of ρ_{err} obtained with "urban-source" monitors after compensation for the change in the total emission significantly differ for the two different statistics; such "irregular" differences are, probably, a result of a complex interaction of several factors. The differences between the results obtained for different kinds of spatial distributions of emission uncertainties and for different levels of magnitudes of uncertainties are of a complex nature and will not be analyzed here. This large effect of uniform emission changes makes ozone measurements less suitable than NO_2 measurements to discern differences in the random uncertainty of emissions.

Note that, the sensitivity of estimates of ρ_{err} to changes in the total emission rate is smallest for "suburban-source" monitors, although values of ρ_{err} in this case are still larger than for rural NO_2 measurements. Therefore, on the one hand, our results suggest that the "suburban-source" monitors can still be useful for diagnostics of uncertainties in spatial distributions of NO_x emissions. However, on the other hand, the large differences between the results for the "urban-source" and "suburban-source" monitors, which can only be explained by differences in systematic errors in the model (see Table **1**), indicate that the relatively good results for the "suburban-source" monitors may be partly due to specific features of the model used, and that they may not be easily reproducible with another model.

Accordingly, more research is needed to assess the usefulness of the ozone monitors; their use in the networks aimed at diagnostics of NO_x emission uncertainties cannot be unambiguously recommended at the moment.

The results obtained with the nonparametric statistics which count the number of sites demonstrating improvement or deterioration in the agreement between the simulations and measurements (in terms of RMSE or R^2) were found to be essentially the same as those obtained with the respective statistics Z_1 or Z_2 and, for that reason, are not presented here.

The only noticeable difference is that the decision-error probability are, in general, slightly larger with the nonparametric statistics. This can be caused by some asymmetry of this kind of statistics due to the special situation when the number of sites for which the agreement is improved is equal to the number of sites for which it is reduced. Such a situation would be counted as associated with the wrong decision thus increasing the decision-error probability. On the whole, it may be concluded that the nonparametric statistics do not demonstrate any obvious advantages. They can, nevertheless, be used effectively instead of or in parallel with the basic statistics considered here.

3.4. The Dependence of the Decision-Error Probability on the Level of Uncertainties in the Model Results and Measurements

The results presented above are obtained with a certain model (CHIMERE) and certain monitors (those presented in AirBase). It is obvious, that any other models and measurements which feature different levels of uncertainties can yield different values of the decision-error probability. Accordingly, in order to make the results of this study more general, it is useful to consider how the decision probability error depends on the uncertainties in the simulations and measurements. Following an approach common for inverse modeling studies, we do not consider the model and measurements errors separately, but, rather, characterize their total error by means of one parameter. Although such an approach does not allow us to take into account possible differences in statistical properties of measurement and model errors (such as the differences in their spatial and temporal covariances), these properties are generally not sufficiently well known anyway. Moreover, the results of our experiments cannot depend on the error covariances when the error statistics considered here are determined by local concentration changes caused by spatially uncorrelated emission perturbations (as for the RAN experiments).

The main idea of the subsequent analysis is to replace the real measurements with synthetic data obtained as a combination of the modeled and observed concentrations:

$$\mathbf{C}_{os}=\beta(\mathbf{C}_o-\mathbf{C}_m)+\mathbf{C}_m, \qquad (15)$$

where β is a constant scaling coefficient. It is easy to check that

$$\mathrm{RMSE}_s=\beta\times\mathrm{RMSE}_0, \qquad (16)$$

where RMSE_s and RMSE_0 are the standard RMSE calculated with the synthetic and real measurements, respectively. This relation holds also if RMSE is defined as square root from the centered mean square error defined by Eq.(3). Figs. (**11**) and (**12**) present estimates of the decision-error probability as a function RMSE defined as follows:

$$RMSE=\frac{1}{LM}\left[\sum_{k=1}^{L}\sum_{j=1}^{M}\left(C_m^{jk}-C_o^{jk}-\overline{C}_m^{k}+\overline{C}_o^{k}\right)^2\right]^{1/2} \qquad (17)$$

The decision-error probability has been calculated with β=0.25;0.5;0.75;1;1.5. Results corresponding to these values are marked in figures by symbols.

It can be seen that, in most cases, ρ_{err} logarithmically increases with the increase of RMSE. Significant deviations from the logarithmic law are observed mainly when the ρ_{err} is rather small, below 0.1.

Note that, when results are compared for different categories of monitors, ρ_{err} tends to be larger for those categories for which RMSE is larger (for a given value of β). For example, urban monitors are associated with both larger RMSE

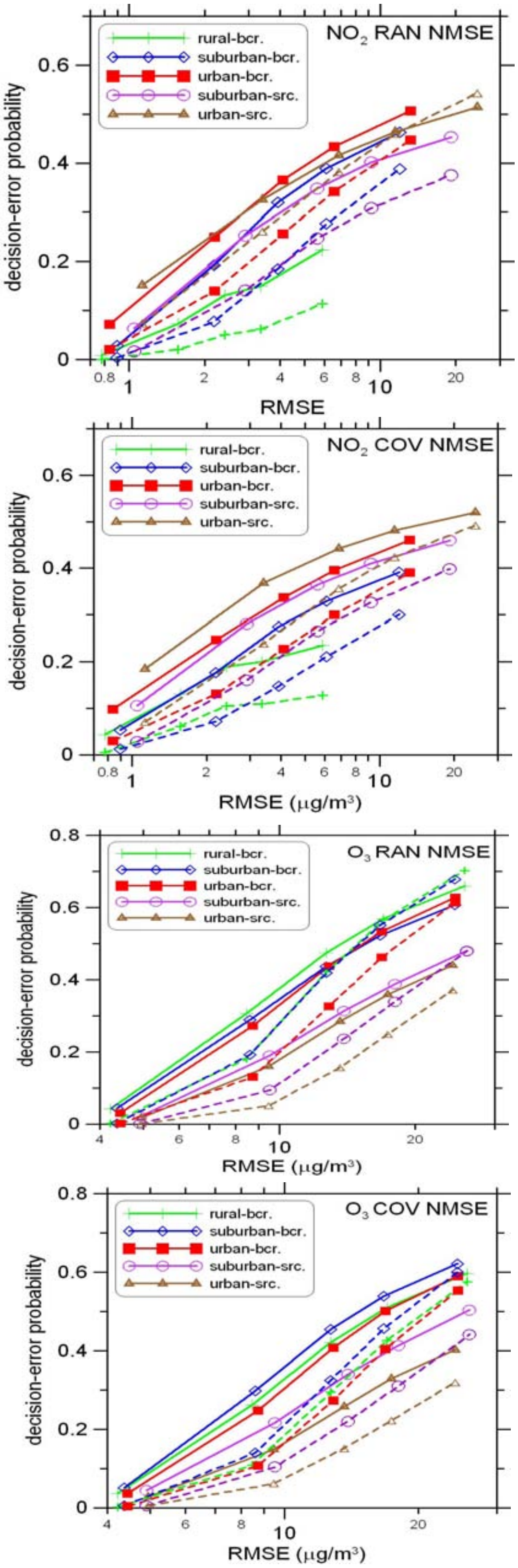

Fig. (11). Estimates of the decision-error probability as a function of the RMS difference between the measured and simulated concentrations for networks of 10 NO_2 and O_3 monitors. The estimates are based on the error statistics Z_1. The estimates are shown for the cases with θ_p=0.4 (solid lines) and θ_p=0.6 (dashed lines).

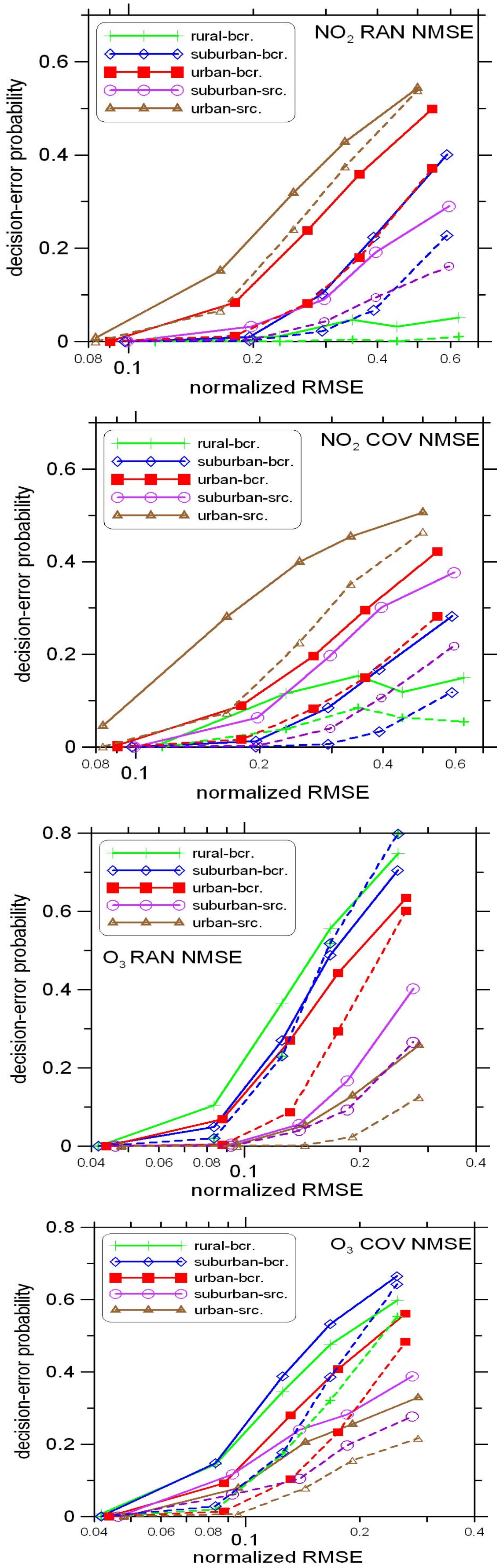

Fig. (12). The same as in Fig. 11 but for networks of 50 monitors.

and ρ_{err} than rural monitors. This relation could indeed be expected since the model and measurement errors are the only reason for possible decision errors. However, it is obvious also that differences in RMSE cannot explain all the differences between values of ρ_{err} for different categories of monitors. For example, “suburban-sources” NO_2 monitors are associated with larger RMSE than urban background monitors; nevertheless, they tend to yield smaller ρ_{err}.

3.5. Practical Implications

The results presented above can be used in practice in several ways. First, the estimates of ρ_{err} can be used for testing different hypotheses regarding the relative accuracy of two emission datasets. Second, they can be used to optimize the choice of values of a priori uncertainty in emissions and observations, needed for optimal estimation of updated emissions. Third, the results can be used to guide the design of networks with respect to their ability to minimize errors in emissions.

For example, if we assumed (based, e.g., on results of the inverse modeling) that the difference in the accuracy of two emission datasets is θ_p, then it would be straightforward to test whether or not this assumption is consistent with independent measurements: the error statistics for a presumably more accurate emission dataset should be smaller, otherwise, we would have to conclude (with the probability ρ_{err} depending on θ_p, the category and the number of monitors) that our assumption is wrong. In a more general case, we could hypothesize that θ_p is distributed, with some probability density $\eta(\theta_p)$, inside a certain interval $[\theta_p^{\min}, \theta_p^{\max}]$. In that case, we could evaluate the cumulative decision-error probability, P_{err}, as follows:

$$P_{err} = \int_{\theta_p^{\min}}^{\theta_p^{\max}} \rho_{err}(\theta_p)\eta(\theta_p)d\theta_p \tag{18}$$

Here ρ_{err} can be approximated as a function of θ_p using, e.g., a linear interpolation between its estimates for fixed values of θ_p. A value of θ_{pmin} can be put to zero, while θ_{pmax} can be estimated as the difference of the emission logarithms in accordance to Eq. (11). Of course, a simple comparison of model results obtained with two different emission fields will unlikely lead, in a general case, to a definite conclusion about the relative quality of the emission datasets. Indeed, there is always a possibility that two datasets can be very different but have the same level of uncertainties. The decision might be more definite if not only the mere fact that $Z(\boldsymbol{E_1})$ is different from $Z(\boldsymbol{E_2})$ is taken into account but also the magnitude of the difference between them. In principle, the estimation of P_{err} in such a situation could be made using essentially the same method (and the results of the same model runs) as discussed here. However, in practice, this would be a rather daunting task, since it would require tabulating the probability distributions of $|Z(\boldsymbol{E_1})\text{- } Z(\boldsymbol{E_2})|$ for all probable values of θ_p.

Some conclusion about the relative accuracy of two emission datasets can be made also in a much simpler way, namely by considering the distribution of signs of differences $Z(\boldsymbol{E_2})$-$Z(\boldsymbol{E_1})$ calculated for individual measurement sites. Let us denote the number of sites for which $Z(\boldsymbol{E_2})$-$Z(\boldsymbol{E_1})>0$ as J and the total number of sites as L. If the level of uncertainties in $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$ is the same and if the considered sites can be treated as being independent (this condition means that they should satisfy the criteria discussed in Section 4.1), then we can expect that the probability of the ratio of J to L being larger than 0.5 can be described, approximately, by the binomial distribution. Accordingly, by performing a standard binomial test, we can estimate the probability, P_r, that the actual differences between the results obtained with two emission datasets are completely random and, therefore, that the accuracy of $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$ is essentially the same. If the emissions $\boldsymbol{E_1}$ are significantly better than $\boldsymbol{E_2}$, it is probable that the ratio of J to L is much larger than 0.5 and P_r is small. It should be noted, however, that the meaning of P_r is quite different from that of P_{err}. Indeed, P_{err} is defined so that to take into account, at least in ideal, all possible realizations of errors in emission data and in the simulated and measured concentrations, while P_r is itself a random number which may be different for different random realization of uncertainties in $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$. Even if $\boldsymbol{E_2}$ is better than $\boldsymbol{E_1}$, there is still some probability that the ratio of J and L would be equal or even smaller than 0.5. Nevertheless, the simple binomial test can be really helpful; in fact, it has already been used in several inverse modeling studies [21, 23, 24].

The results of this study can also be helpful in the context of the network design, assuming that the purpose of a monitoring network is not only to register the level of air pollution at a limited number of sites, but also to provide information elucidating the sources of the observed pollution. Our results indicate that rural NO_2 background monitors are most "powerful", among the considered categories of NO_2 and O_3 monitors, with respect to the detection of uncertainties in NO_x emission datasets. The existing network already allows detecting the emission uncertainties corresponding to θ_p=0.4 (~50% emission uncertainty) with the probability of the error smaller than 0.1. However, more monitors are needed in order to further reduce this detection threshold, in particular if smaller differences in emission uncertainties need to be detected. In Western Europe, the increase of the number of rural NO_2 monitors should be particularly encouraged in France, Great Britain, Italy, Spain and Portugal, where the existing networks are sparse, especially when compared with the rural NO_2 monitoring network in Germany. Of course, developing of air quality monitoring in Eastern Europe, where it is now almost absent, is an especially pressing issue.

New rural sites will be most efficient if they satisfy the requirements which were discussed in Section 3.1. Specifically, the distance between the neighboring monitors should not be smaller than 100 km in order to assure independent measurements; and moreover, taking into account that the uncertainties in the spatial distribution of emissions may covariate (as in the COV case), it can be recommended to keep an even larger minimum distance. On the other hand, there is no risk that the density of a monitoring network can ever become really excessive, even when it will be no longer possible to locate new monitors sufficiently far from existing ones. The "power" of individual monitors will be smaller than that in the case of sufficiently sparse network, but still the efficiency of the whole network will continue to increase as new monitors are introduced. Meanwhile, in case of a very dense network, it is especially important that the monitors should be distributed homogeneously.

Although, in accordance to our results, ozone measurements are less efficient indicators of the NO_x emission uncertainties, they still can be useful, at least, as a source of supplementary information which could enable more reliable conclusions regarding NO_x emission uncertainties. In particular, we found that the most useful information in the given context is provided by suburban ozone monitors situated in the vicinity of strong sources. Probably, the major effect detected by these monitors is the titration of ozone by freshly emitted nitrogen monoxide; this explains why they are so sensitive to local changes of NO_x emissions. Note that monitors of such a type are rather rare in many European countries (e.g., in France, Great Britain, and Italy). However, as it was emphasized in Section 3.3, more research is yet needed in order to investigate to what extent the estimates of ρ_{err} are independent of specific features of the model (e.g., of its spatial resolution). In general, ozone measurements can be considered as less useful, because the differences in results with and without total emission adjustment show large spread.

Our results show that the efficiency of the monitoring network depends not only on the number of monitoring sites but also on the accuracy of the measurements and the model. The disagreement between the model and measurements is especially large in the case of NO_2 concentrations (see Table **1**). On the one hand, a significant part of this error is likely associated with the representativeness of the measurements and can be diminished through the choice of more representative locations of monitoring sites and the increase of resolution of the model. The results of this study could be used to find a compromise (through optimizing ρ_{err}) between a larger network with larger observational errors and a smaller network with a smaller error. On the other hand, rather low correlation between the modeled and measured time series indicates that there may also be other serious reasons for the strong disagreement between the model and the measurements. It is obvious that further significant efforts should also be devoted to both the model development (including chemistry, physics and the spatial-temporal allocation of pollution sources) and advancements in the measurement techniques.

4. CONCLUSIONS

It is not uncommon that measurements of near surface concentrations of different species are used in combination with respective modeled concentrations for validation or comparison of different emission datasets. Ideally, it is expected that better emission estimates should yield better agreement between the simulated and measured concentrations. However, due to the presence of model and measurement errors, an opposite result is also probable. In this report, we propose a general probabilistic procedure aimed at quantification of the statistical significance of the conclusion regarding the relative quality of two (or more) emission datasets, which can be drawn from comparison of measurements with model results obtained with corresponding emission datasets. We applied this procedure to the case of European gridded NO_x emissions by using the AirBase monitoring data for nitrogen dioxide and ozone concentrations and the CHIMERE chemistry-transport model. We quantified the probability that better agreement between the simulated and measured concentrations is obtained using more uncertain emission dataset for five categories of monitors, such as “rural-background”, “suburban-background”, “urban-background”, “suburban-sources” and “urban-sources”. We used different error statistics which are based on the use of the mean square error and the coefficient of determination. The numerical experiments were performed for three different types of spatial distributions of emission uncertainties, including the distribution of uncertainties of NO_x emission estimates derived from satellite measurements.

As a result, it is found, in particular, that, among the considered categories of monitors, the measurements of NO_2 at rural background sites provide the most efficient and reliable information from the point of view of diagnostics of NO_x emission uncertainties. It is shown also that relatively small changes in the total emissions can have a larger impact on the accuracy of the simulated concentrations than uncertainties in the spatial distribution of emissions. This effect is especially important in case of ozone measurements. This is why these measurements can be considered as less suitable than NO_2 measurements for the purpose of our study.

The results of this study can also be considered in the context of the monitoring network design and planning. In particular, our results indicate that more rural NO_2 monitors are needed. While the rural NO_2 network is very sparse in some Western European countries (France, Great Britain, Italy, Spain and Portugal), the corresponding measurements are almost totally absent in Eastern Europe. In order to insure the efficiency of the monitoring network with the respect of detection of uncertainties in NO_x emission data, it could be recommended to distribute monitors rather homogeneously in space, such that the distance between neighboring monitors would be about 100 km or more. As it could be expected, our analysis has shown that the information provided by a pair of ozone monitors is less independent than the information provided by a pair of NO_2 monitors located in a similar environment within the same distance from each other, although the difference between corresponding characteristics of O_3 and NO_2 monitors turned out, rather unexpectedly, to be quite small.

It is found that the probability of wrong decision regarding the relative accuracy of two emission datasets can be reduced at the expense of larger number of monitors in a network and higher accuracy of a model and measurements. With the given model and measurements, the decision error probability is found to be rather significant in some practically interesting cases, and therefore, further development of models and measurement techniques should be encouraged. The dependence of the decision-error probability on the number of monitors and RMSE (combining the model and measurement errors) is, commonly, rather weak (logarithmic).

The next steps in this direction may include the analysis of the utility of surface NO_2 and O_3 measurements for validation of estimates of multi-annual changes in NO_x emissions. The similar probabilistic procedure can also be used in case of other species, such as carbon monoxide, methane, and carbon dioxide which are measured both from the ground and from the space.

ACKNOWLEDGEMENTS

The authors acknowledge the support provided the European Commission through the GEOMON FP6 project. I.B. Konovalov also acknowledges the support of the Russian Foundation for Basic Research (grant No. 08-05-00969-a) and Russian Academy of Sciences (in the framework of the Programme for Basic Research “Physics of Atmosphere; Electrical Processes, Radiophysical Methods of Research”). The authors acknowledge the use of helpful routines created by G. Curci (Centre of Excellence for the forecast of Severe Weather (CETEMPS), Dipartimento di Fisica, Università degli Studi dell'Aquila, Coppito - L'Aquila, Italy) which facilitated the work with the AirBase data. The help of G. Foret (LISA/CNRS) with providing meteorological data for the CHIMERE simulations is also much appreciated. The authors are grateful to the anonymous reviewers for their useful remarks and suggestions.

REFERENCES

[1] Hanna SR, Chang JC, Fernau ME. Monte Carlo estimates of uncertainties in predictions by photochemical grid model (UAM-IV) due to uncertainties in input variables. Atmos Environ 1998; 32: 3619-28.

[2] Bergin MS, Noblet GS, Petrini K, Dhieux JR, Milford JB, Harley RA. Formal uncertainty analysis of a Lagrangian photochemical air pollution model. Environ Sci Technol 1999; 33: 1116-26.

[3] Placet M, Mann CO, Gilbert RO, Niefer MJ. Emissions of ozone precursors from stationary sources: a critical review. Atmos Environ 2000; 34: 2183-204.

[4] Beekmann M, Derognat C. Monte-Carlo uncertainty analysis of a regional-scale transport chemistry model constrained by measurements from the Atmospheric Pollution Over the Paris Area (ESQUIF) campaign. J Geophys Res 2003; 108(D17): 8559.

[5] Taghavi M, Cautenet S, Arteta J. Impact of a highly detailed emission inventory on modeling accuracy. Atmos Res 2005; 74: 65-88.

[6] Deguillaume L, Beekmann M, Menut L. Bayesian Monte Carlo analysis applied to regional-scale inverse emission modeling for reactive trace gases. J Geophys Res 2007; 112: D02307.

[7] Eyring V, Stevenson DS, Lauer A, *et al.* Multi-model simulations of the impact of international shipping on Atmospheric Chemistry and Climate in 2000 and 2030. Atmos Chem Phys 2007; 7: 757-80.

[8] Pison I, Menut L, Bergametti G. Inverse modeling of surface NOx anthropogenic emission fluxes in the Paris area during the Air Pollution Over Paris Region (ESQUIF) campaign. J Geophys Res 2007; 112: D24302.

[9] Butler TM, Lawrence MG, Gurjar BR, van Aardenne J, Schultz M, Lelieveld J. The representation of emissions from megacities in global emission inventories. Atmos Environ 2008; 42: 703-19.

[10] Dentener F, Stevenson D, Cofala J, *et al*. The impact of air pollutant and methane emission controls on tropopspheric ozone and radiative forcing: CTM calculations for the period 1990-2030. Atmos Chem Phys 2005; 5: 1731-55.

[11] Klimont K, Cofala J, Amann M, Streets DG, Ichikawa Y, Fujita S. Projections of SO2, NOx, NH3 and VOC emissions in East Asia up to 2030. Water Air Soil Pollut 2001; 130: 193-8.

[12] Olivier JGJ, Berdowski JJM. Global emissions sources and sinks. Guicherit R, Heij BJ, Eds. The Climate System; Balkema Publishers/Swets, Zeitlinger Publishers: Lisse, The Netherlands 2001; pp. 33-78.

[13] Olivier JGJ, van Aardenne JA, Dentener F, Ganzeveld L, Peters JAHW. Recent trends in greenhouse gas emissions: regional trends and spatial distribution of key sources. Environ Sci 2000; 2: 81-99.

[14] Pulles T, van het Bolscher M, Brand R, Visschedijk A. Assessment of global emissions from fuel combustion in the final decades of the 20th century. Application of the emission inventory model TEAM. Technical Report A-R0132B, Netherlands Organisation for Applied Research (TNO): Apeldoorn, The Netherlands 2007.

[15] Samaali M, Francois S, Vinuesa JF, Ponche JL. A new tool for processing atmospheric emission inventories: Technical aspects and application to the ESCOMPTE study area. Environ Model Softw 2007; 22: 1765-74.

[16] Vestreng V, Breivik K, Adams M, *et al.* Inventory Review 2005, Emission Data reported to LRTAP Convention and NEC Directive, Initial review of HMs and POPs, Technical report MSC-W 1/2005, 2005.

[17] Zhang Q, Wei Y, Tian W, Yang K. GIS based emission inventories of urban-scale: A case study of Hangzhou, China. Atmos Environ 2008; 42(20): 5150-65.

[18] Pétron G, Granier C, Khattatov B, *et al*. Monthly CO surface sources inventory based on the 2000-2001 MOPITT satellite data. Geophys Res Lett 2004; 31: L21107.

[19] Müller JF, Stavrakou T. Inversion of CO and NO_x emissions using the adjoint of the IMAGES model. Atmos Chem Phys 2005; 5: 1157-86.

[20] Martin RV, Sioris CE, Chance K, *et al*. Evaluation of space-based constraints on global nitrogen oxide emissions with regional aircraft measurements over and downwind of eastern North America. J Geophys Res 2006; 111: D15308.

[21] Konovalov IB, Beekmann M, Burrows JP, Richter A. Satellite measurement based estimates of decadal changes in European nitrogen oxides emissions. Atmos Chem Phys 2008; 8: 2623-41.

[22] Krakauer NY, Schneider T, Randerson JT, Olsen SC. Using generalized cross-validation to select parameters in inversions for regional carbon fluxes. Geophys Res Lett 2004; 31: L19108.

[23] Konovalov IB, Beekmann M, Richter A, Burrows JP. Inverse modelling of the spatial distribution of NO_x emissions on a continental scale using satellite data. Atmos Chem Phys 2006; 6: 1747-70.

[24] Konovalov IB, Beekmann M, Richter A, Burrows JP. The use of satellite and ground based measurements for estimating and reducing uncertainties in the spatial distribution of emissions of nitrogen oxides. arXiv: physics/0612144 (www.arxiv.org). 2006.

[25] Richter A, Burrows JP. Tropospheric NO_2 from GOME measurements. Adv Space Res 2002; 29: 1673-83.

[26] Martin RV, Chance K, Jacob DJ, *et al.* Koelemeijer RBA: An improved retrieval of tropospheric nitrogen dioxide from GOME. J Geophys Res 2002; 107(D20): 4437.

[27] Leue C, Wenig M, Wagner T, Klimm O, Platt U, Jahne B. Quantitative analysis of NOx emissions from GOME satellite image sequences. J Geophys Res 2001; 106: 5493-505.

[28] Martin RV, Jacob DJ, Chance K, Kurosu T, Palmer PI, Evans MJ. Global inventory of nitrogen oxide emissions constrained by space-based observations of NO_2 columns. J Geophys Res 2003; 108: 4537.

[29] Jaeglé L, Martin RV, Chance K, *et al.* Satellite mapping of rain-induced nitric oxide emissions from soils. J Geophys Res 2004; 109: D21310.

[30] Tarantola A. Inverse problem theory; methods for data fitting and model parameter estimation; Elsevier 1987.

[31] Taylor KE. Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 2001; 106: 7183-92.

[32] Konovalov IB, Beekmann M, Vautard R, *et al.* Comparison and evaluation of modelled and GOME measurement derived tropospheric NO_2 columns over Western and Eastern Europe. Atmos Chem Phys 2005; 5: 169-90.

[33] Tilmes S, Brandt J, Flatoy F, *et al*. Comparison of five Eulerian air pollution forecasting systems for the summer of 1999 using the German ozone monitoring data. J Atmos Chem 2002; 42: 91-102.

[34] Schmidt HC, Derognat C, Vautard R, Beekmann M. A comparison of simulated and observed ozone mixing ratios for the summer of 1998 in western Europe. Atmos Environ 2001; 35: 6277-97.

[35] GENEMIS (Generation of European Emission Data for Episodes) project. EUROTRAC Annual Report 1993, Part 5, EUROTRAC International Scientific Secretariat. Garmisch-Partenkirchen: Germany 1994.

[36] Simpson D, Winiwarter W, Borjesson G, *et al.* Inventorying emissions from nature in Europe. J Geophys Res 1999; 104: 8113-52.

[37] Stohl A, Williams E, Wotawa G, Kromp-Kolb H. A European inventory of soil nitric oxide emissions on the photochemical formation of ozone in Europe. Atmos Environ 1996; 30: 3741-55.

[38] Wackernagel H. Multivariate Geostatistics; Springer 2003.

[39] Tilmes S, Zimmermann J. Investigation on the spatial scales of the variability in measured near-ground ozone mixing ratios. Geophys Res Lett 1998; 25(20): 3827-30.

[40] Ito K, Thurston GD, Nadas A, Lippmann M. Monitor-to-monitor temporal correlation of air pollution and weather variables in the North-Central U.S. J Expos Anal Environ Epidem 2001; 11: 21-32.

Received: June 30, 2008 Revised: September 11, 2008 Accepted: October 7, 2008