Atmospheric Environment 43 (2009) 6425–6434

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

# Combining deterministic and statistical approaches for $PM_{10}$ forecasting in Europe

I.B. Konovalov [a,b,*], M. Beekmann [b], F. Meleux [c], A. Dutot [b], G. Foret [b]

[a] Institute of Applied Physics, Russian Academy of Sciences, Nizhniy Novgorod, Russia
[b] Laboratoire Inter-Universitaire de Systèmes Atmosphériques, CNRS, Université Paris-Est and Université Paris 7, Créteil, France
[c] Institut National de l'Environnement Industriel et des Risques, INERIS, Verneuil en Halatte, France

ARTICLE INFO

*Article history:*
Received 31 January 2009
Received in revised form
22 June 2009
Accepted 24 June 2009

*Keywords:*
Particulate matter
Air quality
Statistical model
Neural network
Data assimilation
CTM

ABSTRACT

Well documented adverse health effects of airborne particulate matter (PM) stimulate intensive research aimed at understanding and forecasting its behaviour. Forecasting of PM levels is commonly performed with either statistical or deterministic chemistry-transport models (CTM). In this study, we investigate advantages of combining deterministic and statistical approaches for $PM_{10}$ forecasting over Europe one day ahead. The proposed procedure involves statistical postprocessing of deterministic forecasts by using $PM_{10}$ monitoring data. A series of experiments is performed using a state-of-the-art CTM (CHIMERE) and statistical models based on linear regressions. It is found that performance of both CTM simulations and "pure" statistical models is inferior to that of the combined models. In particular, the root mean squared error of the deterministic forecasts can be reduced, on the average, by up to 45 percent (specifically, from 12.8 to 6.9 $\mu g/m^3$ at urban sites in summer) and the coefficient of determination can be almost doubled. Importantly, it is found that the combined models for rural sites in summer and for urban and suburban sites in both summer and winter are representative, on the average, not only for a given monitoring site used for their training, but also of territories of similar type of environment (rural, suburban or urban) within several hundreds of kilometers away.



## 1. Introduction

Respirable particulate matter (PM) adversely affecting human health (see, e.g., Pope, 2000; Moshammer and Neuberger, 2003; Hauck et al., 2004) is commonly considered as one of the major factors contributing to air pollution problems in urbanized regions. In order to control PM levels, certain standards of PM concentrations are defined in the legislation of many countries. In particular, the Council Directive 2008/50/EC established target values for annual and daily mean $PM_{10}$ (particles below 10 $\mu g/m^3$ in diameter) concentrations in the EU. Measurements of $PM_{10}$ are routinely performed in Europe at several hundred measurement sites. $PM_{10}$ levels are operatively forecasted either explicitly or as a part of national air quality indices (EURAD Project, 2009; IRCEL-CELINE, 2009; PREV'AIR, 2009; UK Air Pollution Forecast, 2009) in order to warn the population against potentially dangerous atmospheric conditions.

$PM_{10}$ forecasts (as well as forecasts of other major pollutants) can be performed by means of either deterministic or statistical methods. Deterministic forecasts (see, e.g., Honoré et al., 2008) are elaborated by chemistry transport models (CTM), which attempt to explicitly describe all major physico-chemical and meteorological processes responsible for the evolution of atmospheric aerosol, while statistical methods (Chaloulakou et al., 2003; Hooyberghs et al., 2005; Perez and Reyes, 2006; Zolghadri and Cazaurang, 2006; Slini et al., 2006; Stadlober et al., 2008; Hoi et al., 2009) use time series of past measurements in order to define associations between meteorological conditions and $PM_{10}$ concentration. A major advantage of CTM based forecasts is an uniform spatial coverage, while statistical models are best representative of only a given measurements site. On the other hand, insufficient knowledge of pollutant sources and inaccuracies in description of physico-chemical processes can lead to rather strong biases in CTM results (Vautard et al., 2007; Stern et al., 2008).

The goal of this study is to investigate the prospects of the combined use of deterministic and statistical methods for $PM_{10}$ forecasting. An idea of such a combination is to use a statistical model in order to correct predictions made by a deterministic model. Basically, this is a classical procedure known in meteorology (Glahn and Lowry, 1972) as Model Output Statistics (MOS). A similar

* Corresponding author at: Igor B. Konovalov, Institute of Applied Physics of Russian Academy of Sciences, 46 Ulyanov Str., 603950 Nizhny Novgorod, Russia. Tel.: +78314164902; fax: +78314160616.
*E-mail address:* konov@appl.sci-nnov.ru (I.B. Konovalov).

1352-2310/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2009.06.039

procedure is a part of downscaling methods (Wilby and Wigley, 1997) aimed at improving weather forecasts at individual sites. A common approach involves building a linear regression model between a set of variables provided by a numerical model and an observed variable; this regression is further used to correct the numerical weather forecasts for a given site (e.g., Wilson and Valée, 2003; Taylor and Leslie, 2005). Here we analyze the potential of this procedure in improving $PM_{10}$ forecasts not only at individual measurement sites, but also in locations where $PM_{10}$ monitors are absent.

Advantages of the combined use of deterministic and statistical methods for air pollution forecasting have not yet been sufficiently investigated. The closest analog of the procedure discussed in this paper is the MOS method used by Honoré et al. (2008) for forecasting of daily maximums of ozone concentration. They used deterministic ozone forecasts and predicted air temperature as input variables to linear regression models predicting errors of deterministic ozone forecasts at individual sites. The predicted errors were then spatially interpolated. This procedure has been shown to significantly improve forecasts for episodes with high ozone concentration ($>180\ \mu g/m^3$) and is now operational within the PREV'AIR air pollution forecasting system (Rouil et al., 2009). A similar idea has recently been employed by Denby et al. (2008) in order to improve a simulated spatial distribution of exceedances of threshold values ($50\ \mu g/m^3$) defined by European legislation for daily $PM_{10}$ concentrations. Specifically, they built a linear regression of daily mean simulated concentrations with measured values. A regression model was created for each day and for the whole domain (rather than for individual sites) and did not involve any meteorological parameters. Note that the problem addressed by Denby et al. (2008), contrary to the problem of forecasting, permits the use of measurements for the same day for which the concentration field is created.

Although in the framework of this study we performed experiments with both linear and nonlinear (neural network based) statistical models, only results obtained with classical linear regressions are reported in this paper. This restriction is mainly due to the space limits but also partly due to the fact that, similar to some earlier studies (e.g., Perez and Reyes, 2006), our experiments with neural networks did not reveal any large difference between the performances of $PM_{10}$ forecasts carried out by linear and nonlinear models.

The paper is organized as follows. Section 2 is devoted to description of the proposed methodology and input data. In Section 3, we first evaluate our method for sites which were used for creation of statistical models, and then we discuss the spatial coverage of forecasts. Finally, the conclusions and outlook are provided in Section 4.

## 2. Method and data description

### 2.1. Chemistry transport model CHIMERE and meteorological data

Modelling of gaseous and aerosol processes in the framework of the CHIMERE chemistry transport model is described in detail in several previous papers. In particular, most of the model's physics and gas-phase chemistry is described by Schmidt et al. (2001). The aerosol module is described by Bessagnet et al. (2004, 2008). A detailed model description, a technical documentation and the source codes of the version (V200709) used in this study are available on the web (CHIMERE chemistry transport model, 2009). Here we mention only the model features essential for our study.

The chemical scheme used in this study, MELCHIOR1, includes more than 300 reactions of 80 species. Additionally, 7 aqueous and 4 heterogeneous reactions are taken into account. For the aerosol density distribution function, a sectional representation (Gelbard and Seinfeld, 1980) is used. All particles in section $l$ are assumed to have the same chemical composition and are characterized by their mean diameter $d_l$. Eight bins from 10 nm to 10 μm are defined following a geometrical progression. The equilibrium concentration of semi-volatile inorganic species (sulfate, nitrate, chlorine, ammonium) and water content of particles is calculated using the ISORROPIA thermodynamic model (Nenes et al., 1998). Secondary organic aerosols (SOA) are assumed to be formed from biogenic (terpenes) and aromatic precursor VOC (aromatics, high chain alkanes) following the scheme of Pun et al. (2006), as described in the Bessagnet et al. (2008). This scheme follows the classical "two product mechanism" (Odum et al., 1996). A general evaluation of this mechanisms over a European domain is not possible due to lack of suitable data (Bessagnet et al., 2008), but campaign data often showed a strong underestimation of SOA simulated with this type of mechanism pointing to yet unknown SOA formation processes (Volkamer et al., 2006).

Anthropogenic emissions of aerosols and gaseous species are taken from the EMEP database (Vestreng et al., 2005). The model also takes into account resuspension of small particles, dry deposition and wet scavenging of both gases and aerosols and sea salt aerosol formation.

In this study we use a spatial domain which covers most of Western and Central Europe with a horizontal resolution of $0.5^\circ \times 0.5^\circ$ and includes 3082 grid cells. The model runs are performed with 8 layers defined using a hybrid ($\sigma$, $p$) scheme; the top of the upper layer being fixed at 500 hPa pressure level. The model was run continuously (that is, without any external re-initialization) for the period of four complete years from 2003 to 2006.

CHIMERE is forced by meteorological data simulated off-line with the MM5 V3 mesoscale model. MM5 is initialized with the GFS operational analysis. The same data are used for creation of statistical models and postprocessing of deterministic forecasts. As in Hooyberghs et al. (2005), only analyzed meteorological data (D − 1) are used to evaluate our statistical forecasting model. Thus, strictly speaking, the results presented in this paper should be considered as an upper limit for the expected performances of real forecasts since no evaluation has been performed using meteorological forecasts for following days (from D + 0). Nevertheless, Honoré et al. (2008) have shown that accuracy of $PM_{10}$ and ozone forecasts performed with the CHIMERE model degrades only insignificantly when meteorological forecasts (from D + 0 to D + 2) are used instead of operational analysis (D − 1).

### 2.2. Statistical models

We use statistical regression models which assume a certain functional relationship between a forecasted quantity and several predictors. In the linear case considered in this paper, our models are formulated as follows:

$$PM_{10}(D+1) = \overline{PM_{10}^{obs}} + \sum_{i=1}^{n} p_i(x_i - \overline{x_i}) + \varepsilon, \quad (1)$$

where $PM_{10}(D+1)$ is the 24 h average concentration of $PM_{10}$ for the day following the current day $D$, $\mathbf{x}$ is a vector of predictors which are specified below, $p_i$ are the regression coefficients and $n$ is the number of predictors, $PM_{10}^{obs}$ is the observed concentration (24 h average of the present day), and $\varepsilon$ is the forecast error. The over bar depicts the averages of the corresponding values over the training subset (see Section 2.4) of the data. Linear regressions were fitted to observations by means of the standard SVD (Singular Value Decomposition) method (Press et al., 1992).