International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

A combined Neural Network and Optimal Interpolation approach for PM10 forecast over Po Valley

C. Carnevale [a], **G. Finzi**[a], **E. Pisoni**[a], **M. Volta**[a]

[a] *University of Brescia, Department of Information Engineering, Brescia, Italy*
(carneval@ing.unibs.it, finzi@ing.unibs.it, enrico.pisoni@ing.unibs.it, lvolta@ing.unibs.it)

Abstract: Over recent years, the high levels of air pollution have become a quite important problem due to their direct impact on human health. Due to these health risks, EU directive (2008/50/EC) recommends member states to ensure that timely information about actual and forecasted levels of pollutants are provided to the public. In order to follow these guidelines, prevent critical episodes and inform the public, environmental authorities need fast and reliable forecasting systems. In literature, air pollution forecasting systems can be broadly split in two main category: (1) deterministic systems, solving mass-balance differential equations for a large number of pollutant in atmosphere and (2) non deterministic data driven systems, based on the identification of stochastic models using measurement data collected by monitoring networks. In this work, an optimal interpolation technique has been used to integrate daily PM10 concentrations forecasted by artificial neural networks in 120 monitoring stations with the results of a deterministic chemical transport model. The methodology has been applied to a Northern Italy domain, characterized as one of the most polluted and industrialized area in Europe. Among measurement data from 2003-2007, year 2008 has been selected to test the performance of the integrated modelling system. The evaluation shows very good performances for all the stations, with high agreement in both mean value and 95th percentile computed over the entire year and correlation coefficient usually higher than 0.7 and small values of normalized mean error.

Keywords: PM10 forecast; ANN; Optimal Interpolation

1 INTRODUCTION

Over recent years, the high levels of air pollution have become an important problem due to their direct impact on the environment and on human health. In particular, PM10 long term and peaks exposure can cause a variety of effects, ranging from minor irritation of the respiratory system to premature death. Because of health risks due to pollution exposure, EU directive (2008/50/EC) recommends member states to ensure that timely information about actual and forecasted exceedances of pollutants threshold are provided to the public. In order to follow these guidelines environmental authorities need fast and reliable forecasting systems.

The forecast of air pollution concentrations is a very challenging task as it involves non-linear relationships and uncertain variables. In literature this problem has been solved using two different approach: (a) deterministic models (3-D chemical transport models) [Rouil et al., 2009], [Manders et al., 2009] and (b) statistical models [Pisoni et al., 2009], [Carnevale et al., 2011].

Deterministic chemical transport models (CTM) are extremely complex systems, which formalize the physical and chemical processes involved in the formation of the pollutants by means of the mass-balance equation. These models give a detailed description of the atmospheric variables over urban to global scale, but are computationally time-consuming and require very detailed and often uncertain information concerning meteorology and emissions to initialize and run the model. For these reasons, the performances of deterministic models, when run in forecasting configuration, are usually low and they are used in simulation configuration for different tasks as posteriori and scenario analysis [Brandt et al., 2001], [Manders et al., 2009], [Rouil et al., 2009].

In contrast, point-wise data driven statistical models describe a particular phenomenon based on a mathematical relationship without providing a description of its physics and chemistry. The main advantages of these kind of models is that usually they require less computing time to be simulated, are easier to be implemented than deterministic ones and can reach higher performances. The drawbacks are ineherent to their data driven nature, allowing them to reproduce system behavior only if the system is not affected by any structural change during data collection period, and to perform computation only where the data are collected, i.e. in the monitoring station locations [Schlink and Volta, 2000].

In order to avoid this drawback, the use of spatialization procedure can be applied. A large variety of interpolation methods has been developed for different applications [Cressie, 1993], starting from point-wise measurement data and creating spatially distributed fields [Hooyberghs et al., 2006; Janssen et al., 2008] often performing integration with different spatialized variable (cokriging, [Singh et al., 2011]). Nevertheless, these methods usually did not take into account the uncertainty in the involved data sources.

In this paper, an integrated modelling system to forecast daily mean PM10 concentrations up to three days in advance is presented. The system integrates artificial neural networks and deterministic model simulation results applying an optimal interpolation techniques allowing to take into account uncertainty in both information sources. The methodology has been applied to a Northern Italy domain, characterized as one of the most polluted and industrialized area in Europe, for the year 2008.

2 THE INTEGRATED SYSTEM FORMALIZATION

Let us define a domain containing S unequally spaced monitoring stations at locations $X = \{x_i\}_{i=1}^{S}$ and Ω regular grid cells at locations $W = \{w_j\}_{j=1}^{\Omega}$. Let $\bar{y}(X, \cdot)$ be the measurements of the variable to forecast; $y(X, \cdot)$ be the point-wise forecast at the location of monitoring stations and $\hat{y}(W, \cdot)$ be the interpolated forecast over the regular grids of whole domain. The problem is defined as to obtain the τ (τ=1, 2, 3) days ahead forecast values $\hat{y}(W, t + \tau)$ all over the domain from the available data at the location of the monitoring stations. To achieve the goal, statistical forecasting models and geospatial interpolation techniques have been integrated. The integrated modelling system consists of two main parts:

1. Statistical forecasting models to provide point-wise forecast: $y(X, \cdot)$;

2. A geospatial interpolation model to perform interpolation of point-wise forecast over whole domain: $\hat{y}(W, \cdot)$.

2.1 Point-wise statistical forecasting models

In the integrated modelling system, statistical forecasting models are responsible for providing point-wise forecast up to three days in advance ($\tau = 1, 2, 3$) at the locations X of the monitoring stations. Three different models have been identified for each day of forecast, as follows:

$$y(X, t+\tau) \quad = \quad f_\tau(X, \bar{y}(X, t)) \tag{1}$$

where:

$y(X, t+\tau) \in R$ is the variable to be forecasted;

$\bar{y}(X, t)$ is the measured variable;

The function f_τ has been modelled by feedforward artificial neural networks [Carnevale et al., 2009],[Demuth et al., 2009]. Different neural network structures have been tested and evaluated, ranging from simple feedforward Neural Network to recurrent/dynamic neural network. The evaluation phase shows that feedforward ANN ensure better performances, neverthless the nature of the phenomena. Feedforward neural network are expressed by means of a vector function $f_{FN} : \Re^Q \rightarrow \Re^L$, where Q and L are the dimensions of the input and output vectors of the net respectively; the l-th element of the vector function f_{FN} for the n-th pattern ($v^n \in R^Q$) is defined as (M is the number of the neurons in the hidden layer):

$$f_{FN}(v_l^n) = af_2(\sum_{m=1}^{M}(OW_{l,m} \cdot a_m^n) + g_l) \tag{2}$$

and

$$a_m^n = af_1(\sum_{q=1}^{Q}(IW_{m,q} \cdot v_q^n) + b_m) \tag{3}$$

where af_1 and af_2 are two real nonlinear continues functions, called activation function of the hidden layer (af_1) and of the output layer (af_2). The matrices IW ($M \times Q$) and OW ($L \times M$) are the input and output weight matrix respectively, and b ($M \times 1$) and g ($L \times 1$) vectors are the bias terms estimated by means of the well-known Levenberg-Marquardt backpropagation (BP) algorithm [Hagan et al., 1996].

2.2 Optimal Interpolation

The spatialization of point-wise forecasts has been performed using the results of a 3D modelling system over a grid domain by means of Optimal Interpolation [Cressie, 1993]. This technique is an evolution of simpler geostastical techniques (kriging, cokriging, inverse distance weighting) allowing to take into account the uncertainty in the data sources. For this reason, Optimal Interpolation is a scientfically sounding option when, as in this case, the integration has to be performed between two models. If the uncertainty

of the two data sources can be expressed by zero mean uncorrelated errors with known covariances, the analysis field obtained by optimal interpolation can be expressed in the form of the *Best Linear Unbiased Estimator (BLUE)*, which is defined by the equations:

$$\begin{aligned} \hat{y}(w, t+\tau) &= x_b(W, t_b) + K(y(X, t+\tau) - Hx_b(W, t_b)) \quad (4) \\ K &= PH^T(HPH^T + R)^{-1}. \quad (5) \end{aligned}$$

where:

$x_b(t_b)$, $t_b < t$ is the background (model) field, that can have a temporal resolution different than point-wise forecast;

H is a linear operator mapping the background field grid W in the monitoring location space X;

K is the gain vector;

P is the model error covariance matrix;

R is the point-wise covariance matrix.

The definition of error covariance matrices P and R is a crucial decision for the quality of the analysis. They determine the weights of gain matrix K, hence how the correction is applied to the background in order to obtain the analysis. The main parameters are variances (diagonal terms of the matrices) and covariances, which define how the innovation is spread and smoothed on the whole domain. The description of background error covariance is quite complex. An hypotesis needed to estimate the statistics is to assume the ergodicity of the system: thus, statistics are considered stationary and uniform over the domain [Bouttier and Courtier, 2001]. In this work, the Gaussian model as been used:

$$P = [p_{i,j}] = \exp(-\frac{d_{i,j}^2}{2L^2}) \quad (6)$$

where $d_{i,j}$ is the distance between two points of the domain (i and j) and L is an influence length which defines the decay of covariance with respect to the distance. The equation is usually split into its horizontal and vertical components. Moreover, assuming the background error variance v as a feature of the model, hence a constant on the whole domain, Equation (6) can be rewritten as (indexes i and j are omitted for simplicity):

$$\begin{aligned} P &= [p] = \exp(-\frac{d_h^2}{2L_h^2}) \cdot \exp(-\frac{d_v^2}{2L_v^2}) \cdot v \\ &= \tilde{P}v. \quad (7) \end{aligned}$$

About the description of the matrix R, for monitoring stations, the performances of the forecasting for the different monitoring location can be considered independent; this implies that covariances between the observation errors can be set to zero, leading to a diagonal R matrix. Furthermore, it can also be assumed the same error variance r for all

the forecasting, considering the overall performances of the point-wise forecast all over the domain. Therefore, under these assumptions the matrix R can be written as:

$$R = rI \tag{8}$$

With the above hypothesis, the matrix K can be rewritten as:

$$\begin{aligned} K &= \tilde{P}vH^T[v(H\tilde{P}H^T + \frac{r}{v}I)]^{-1} \\ &= \tilde{P}H^T(H\tilde{P}H^T + \sigma I)^{-1}. \end{aligned} \tag{9}$$

where σ, is the ratio between the error variances of observations and background, respectively.

3 THE CASE STUDY

The methodology has been applied to compute daily forecasted maps up to 3 days over a domain including the whole Northern Italy ($640x410km^2$, Figure 1), with a resolution of $10x10km^2$, for the entire 2008 year. In this area, frequent stagnanting conditions and heavy urban and industrial emissions caused very high PM10 concentrations both in winter and summer seasons.

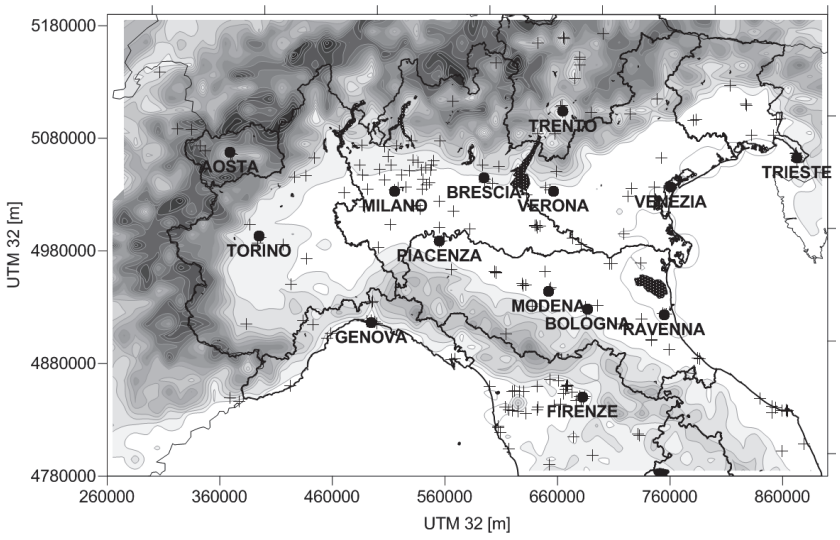

Figure 1: Test case domain, with monitoring station locations (crosses).

The measurement data of years 2000-2008 have been obtained from AIRBASE database. From the initial dataset of 220 stations, only the ones having at least two years of valid data has been selected (Figure 1). For each of the 120 selected stations, two dataset have been created: (1) the data from 2000 to 2007 have been used to train the neural network, (2) the 2008 has been chosen as validation (test) year.

A single feedforward neural network has been identified for each monitoring stations, using as input the PM10 levels observed for t, $t-1$ and $t-2$. The background fields

x_b has been obtained processing the simulations of Transport and Chemical Aerosol Model (TCAM) [Carnevale et al., 2008] over the study domain. In particular, monthly mean PM10 maps computed for the 2004 year in the frame of QUITSAT project with a resolution of $10x10km^2$ over the study domain in simulation configuration have been used.

3.1 Integrated forecasting model validation

The integrated forecasting model system validation has been performed for the 2008 year for $\tau = 1, 2, 3$, using a cross-validation procedure. During this procedure, the system has been applied for 100 yearly iterations using the 80% of the stations (86 stations) for the optimal interpolation and the remaining 20% for the validation. At each iteration, the database split has been randomly performed and the results collected.

Figure 2 presents the boxplot for the metastation performances for different τ. The methodology ensure very good performances, in particular for $\tau = 1$, with very high value of correlation for all the cross-validation tests with median, first and third quartile of the boxplot close to 0.85 and low value of Normalised Mean Absolute Error. The threshold indexes, computed for the law threshold of $50\mu g/m^3$ provides high performances both in terms of Hit and False Alarm ratio. Moreover, it can be noticed that the performances of the metastation slightly change with respect to the selection of validation stations, stating the good spatial coverage of the point-wise forecasting over the domain. Performances for $\tau = 2$ and $\tau = 3$ are quite similar and consistently lower than the $\tau = 1$ ones, stating a fast decreasing in the cause-effect relationship strength with the increasing of forecast period.

Figure 3 shows the mean value of correlation and normalised mean absolute error as a function of the minimum distance between each validation station and the closest station used in the Optimal Interpolation re-analysis computation. In both cases, the degradation of the performances is lower for 1 step forecast, showing a correlation of at least 0.75 for distance greater than 25 km and a quite constant value of normalised mean absolute error. For 2 and 3 step forecast the performance degradation is higher, in particucular for minimum distance greater than 15 km.

4 CONCLUSION

The paper presents the formalization and application of a methodology to perform the forecast of PM10 levels up to 3 days in advance over a mesoscale domain. The methodology has been applied for the year 2008 to Northern Italy, a domain often affected by high PM10 levels exceeding european legislation. The results shows that the system ensures excellent performances for 1 day forecast both in terms of statistical (correlation coefficient and normalised mean absolute error) and threshold (hit and false alarm ratio) indicators. Neverthless a consistent decreasing is noticed, the performances present relatively good value also for larger forecasting period. The presented methodology represents a fast and reliable solution to the Regional Authority that has to provide to population information on pollution levels. Neverthless, a lot of work has still to be done, in particular to improve the estimation of error covariance matrixes and to overcome the relatively strong assumptions needed by the optimal interpolation technique.

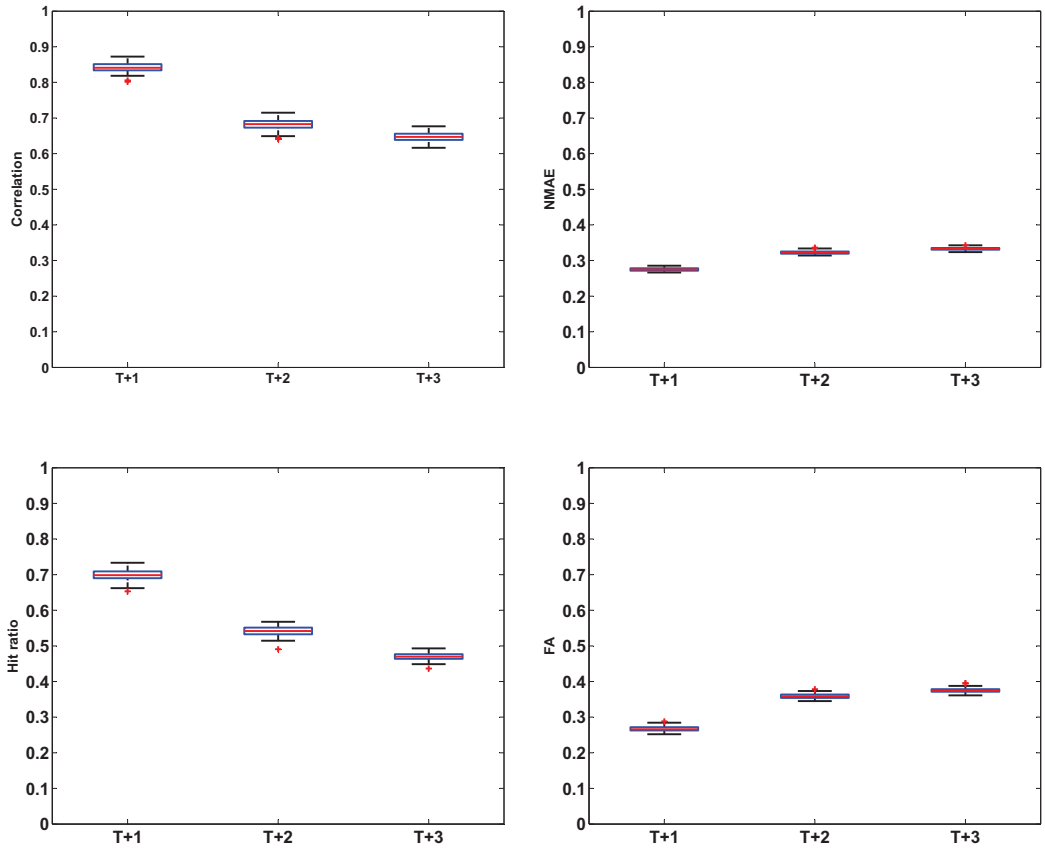

Figure 2: Metastation cross-validation performances: correlation (a), normalised mean absolute error (b), hit ratio (c) and false alarm ratio (d)

5 Acknowledgments

This work has been supported by Regione Lombardia and CILEA Consortium through a LISA Initiative (Laboratory for Interdisciplinary Advanced Simulation) 2010 grant [link:http://lisa.cilea.it].

References

Bouttier, F. and P. Courtier. Data assimilation concepts and methods. Technical report, ECMWF, 2001.

Brandt, J., J. Christensen, L. Frohn, F. Palmgren, R. Berkowicz, and Z. Zlatev. Operational air pollution forecasts from European to local scale. *Atmospheric environment*, 35:91–98, 2001.

Carnevale, C., E. Decanini, and M. Volta. Design and validation of a multiphase 3D model to simulate tropospheric pollution. *Science of the Total Environment, The*, 390 (1):166–176, 2008.

Carnevale, C., G. Finzi, E. Pisoni, V. Singh, and M. Volta. An integrated air quality

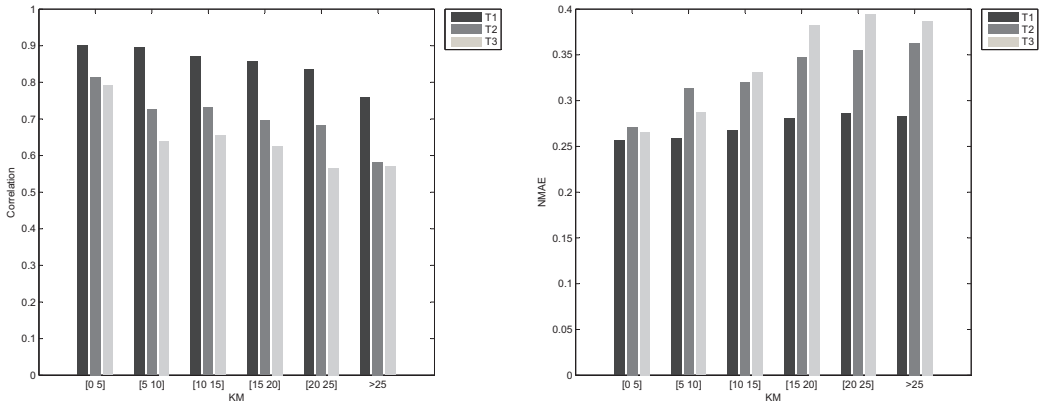

Figure 3: Variability of Correlation and NMAE as a fucntion of the minimum distance with spatialization stations.

forecast system for a metropolitan area. *Journal of Environmental Monitoring*, 13(12): 3437–3447, 2011. cited By (since 1996) 0.

Carnevale, C., G. Finzi, E. Pisoni, and M. Volta. Neuro-fuzzy and neural network systems for air quality control. *Atmospheric Environment*, 43:4811–4821, 2009.

Cressie, N. *Statistics for Spatial Data (revised ed)*. Wiley, New York, 1993.

Demuth, H., M. Beale, and M. Hagan. Neural network toolbox 6 users guide. *The MathWorks Inc.: Natick*, 2009.

Hagan, M., H. Demuth, M. Beale, et al. *Neural network design*. PWS Boston, MA, 1996.

Hooyberghs, J., C. Mensink, G. Dumont, and F. Fierens. Spatial interpolation of ambient ozone concentrations from sparse monitoring points in belgium. *Journal of Environmental Monitoring*, 8:1129–1135, 2006.

Janssen, S., G. Dumont, F. Fierens, and C. Mensink. Spatial interpolation of air pollution measurements using corine land cover data. *Atmospheric Environment*, 42:4884–4903, 2008.

Manders, A., M. Schaap, and R. Hoogerbrugge. Testing the capability of the chemistry transport model LOTOS-EUROS to forecast PM10 levels in the Netherlands. *Atmospheric Environment*, 43(26):4050–4059, 2009.

Pisoni, E., M. Farina, C. Carnevale, and L. Piroddi. Forecasting peak air pollution levels using NARX models. *Engineering Applications of Artificial Intelligence*, 2009.

Rouil, L., C. Honoré, R. Vautard, M. Beekmann, B. Bessagnet, L. Malherbe, F. Meleux, A. Dufour, C. Elichegaray, J. Flaud, et al. Prev'air: An Operational Forecasting and Mapping System for Air Quality in Europe. *Bulletin of the American Meteorological Society*, 90(1):73–83, 2009.

Schlink, U. and M. Volta. Grey box and component models to forecast ozone episodes: A comparison study. *Environmental Monitoring and Assessment*, 65:313–321, 2000.

Singh, V., C. Carnevale, G. Finzi, E. Pisoni, and M. Volta. A cokriging based approach to reconstruct air pollution maps, processing measurement station concentrations and deterministic model simulations. *Environmental Modelling & Software*, 26:778–786, 2011.