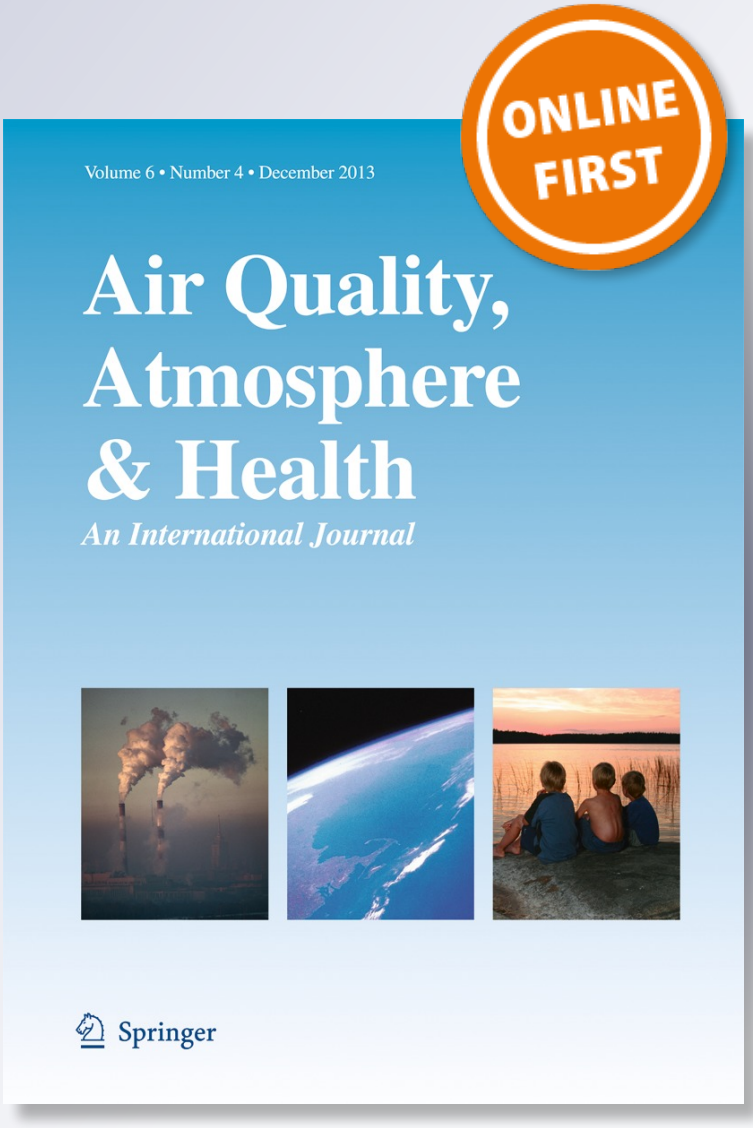
Application of a chemical transport model and optimized data assimilation methods to improve air quality assessment

Camillo Silibello, Andrea Bolignano, Roberto Sozzi & Claudio Gariazzo

Air Quality, Atmosphere & Health
An International Journal

ISSN 1873-9318

Air Qual Atmos Health
DOI 10.1007/s11869-014-0235-1



Springer

Air Qual Atmos Health
DOI 10.1007/s11869-014-0235-1

Application of a chemical transport model and optimized data assimilation methods to improve air quality assessment

Camillo Silibello · Andrea Bolignano · Roberto Sozzi · Claudio Gariazzo

Received: 3 October 2013 / Accepted: 1 January 2014

Abstract The combined use of air quality monitoring data and state-of-the art dispersion models provides a more realistic representation of the spatial distribution of pollutants and allows a reduction in the uncertainties involved in the assessment of the exposure in epidemiological studies. Data assimilation is a method which combines such information to produce an optimal representation of the state of the atmosphere. In this work, we tested two approaches to merge these information sets: the successive corrections method (SCM) and the statistical optimal interpolation (OI). These methods have been extended in order to take into account the spatial representativeness of measurements. PM_{10}, NO_2, and O_3 concentration fields produced by an air quality modeling system, run with two nested domains covering much of Central Italy and the Rome urban area, have been used to identify the optimal values for the horizontal and vertical scaling distances that are key parameters for the SCM and OI methods. A statistical analysis of the results obtained from the application of these methods demonstrated that lower RMSE values resulted from the use of the OI method. Further, $PM_{2.5}$ modeling results over the Rome urban area and additional measurements collected during experimental campaigns, performed within the population exposure to polycyclic aromatic hydrocarbons (EXPAH) LIFE+ Project, allowed the evaluation of this approach in reconstructing $PM_{2.5}$ levels at EXPAH monitoring sites, which were not used in the data assimilation process. The results confirmed the potential of these methods to improve the estimation of modeled concentrations, by taking into account local phenomena not resolved by the model, but clear from the observations, and also in providing more reliable data to be used in exposure studies.

Keywords Air quality models · Spatial analysis · Data assimilation · Emission inventories · Exposure modeling

C. Silibello (✉)
ARIANET Srl, via Gilino, 9, 20128 Milan, Italy
e-mail: c.silibello@aria-net.it

A. Bolignano · R. Sozzi
ARPA Lazio, Via Garibaldi, 114, 02100 Rieti, Italy

C. Gariazzo
INAIL Research Center, Via Fontana Candida,
1, 00040 Monteporzio Catone, Rome, Italy

Published online: 28 January 2014

Introduction

A more realistic representation of the spatial distribution of pollutants is the final goal of "data fusion" (DF) methods that combine information from air quality monitoring networks with other sources such as reanalysis data, satellite data, and data sets obtained from statistical models or chemical transport models (CTMs) (Denby et al. 2009). This approach permits the improvement of air quality assessment and consequently, reduces uncertainty in exposure estimates (Physick et al. 2007). According to Zhang et al. (2012), "data assimilation" (DA) can be defined as a subset of DF in which CTM results are one of the sources of data used to produce an optimal representation of the state of the atmosphere. DA methods are used to overcome the limitations of CTMs and simple interpolation of observations by combining available information in a coherent way. A growing interest on DA methods in air quality modeling is shown by several works dealing with methods in which:

- CTMs directly incorporate monitoring data during the modeling process: Elbern and Schmidt 2000; Denby et al. 2006; Adhikary et al. 2008; Tombette et al. 2009;
- Different data sources (satellite data, air quality fields from models, land-use data, etc.) are used in either a statistical or geometric way: Denby et al. 2005; Horálek et al. 2005;

Borrego et al. 2011; Delle Monache et al. 2008; Petritoli et al. 2011.

The role of combined monitoring and modeling is also being investigated by a sub-group (http://fairmode.ew.eea.europa.eu/monitoring-modeling-sg1) of the Forum for Air quality Modeling in Europe whose main aims are as follows:

- To promote the best practices on the combined use of models and monitoring for EU Directive 2008/50/EC (EC 2008) related applications;
- To develop and to apply quality assurance practices in combining models with monitoring;
- To provide guidance on station representativeness and station selection for the combined use of monitoring with modeling.

A survey conducted in this framework (Denby and Spangl 2010) has shown that nudging methods, residual kriging (Janssen et al. 2008), statistical interpolation, sequential methods (optimal interpolation, Ensemble Kalman filter), and variational methods (3D, 4D-var) are widely used in Europe. A review of major techniques for chemical DA, used in regional real-time air quality forecasting models, is presented in Zhang et al. (2012). Zhang et al. stress the need for continuous development and improvement of DA techniques to reduce inaccuracies and errors in forecast products.

In this work, we have investigated two sequential methods to combine model results and observations: the successive corrections method (SCM) and the statistical optimal interpolation (OI). These methods are generally able to give the best estimate of the chemical state of the atmosphere providing that suitable values are used for the scale parameters defining the influence of the observations along the horizontal and vertical directions. OI and SCM methods and the description of the spatial representativeness of measurements are illustrated in Section "Assimilating observations into CTMs." These methods have been applied to a dataset from an Air Quality Modeling System (AQMS) which simulated PM_{10}, NO_2, and O_3 levels over two nested domains covering Central Italy and Rome urban area, respectively. The description and the evaluation of the modeling system are given in Section "AQMS application." The procedure adopted to identify the optimal values for the horizontal and vertical scale parameters, for the above mentioned species and domains, and the analysis of SCM and OI method results are described in Section "Application of DA methods." Section "Application to the EXPAH dataset" illustrates the application of the OI method to the population exposure to polycyclic aromatic hydrocarbons (EXPAH) LIFE+ Project dataset that demonstrates the capacity of this approach to improve the representation of the spatial distribution of $PM_{2.5}$ levels over the urban area of Rome.

Assimilating observations into CTMs

Concentration fields provided by CTMs at each grid point are considered as "the first guess (also known as background field or prior information) of our best estimate of the state of the atmosphere prior to the use of the observations (Kalnay 2003)." A more realistic representation of the spatial distribution of pollutants is the final goal of DA methods that integrate such information. In this work, we have used two methods: an empirical approach based on the SCM and the well-known statistical OI. These methods allow the interpolation of arbitrarily located observations onto a regular grid, using a background field as first guess, and taking into account the error variances of both the model results and the observations. The merged field is optimal in the sense that it has the lowest error variance.

Optimal interpolation

Within the OI framework, the analyzed (optimal) state vector x^A is given by:

$$x^A = x^G + \underbrace{BH^T\left(HBH^T + R\right)^{-1}}_{K}\left(Y - Hx^G\right)$$

where x^G is the background state vector, Y is the observation vector, H is the observation operator that extracts from a state vector the corresponding values at the location of the observations, B is the background error covariance matrix, R is the observation error covariance matrix, and K is the so-called gain matrix. Here, we have used the module developed by Alexander Barth (available at: http://ocgmod1.marine.usf.edu/OI/optiminterp-0.2.5.tar.gz) to implement the OI method. The following assumptions are made in this module: (1) the observation errors are uncorrelated (e.g., off-diagonal elements of R are zero and R is consequently assumed to be diagonal: $R=\sigma_0^2 I$ where I is the identity matrix and σ_0^2 is the observation error variance defined as follows: $\sigma_0^2=\langle\varepsilon_0^2\rangle$, where $\varepsilon_0=Y-Hx^T$ and x^T is the "true state"); (2) B is assumed to decrease exponentially with the square of the distance along each dimension:

$$B(i,j) = \sigma_B^2 \exp\left(-\frac{d_h^2(i,j)}{L_h^2}\right)\exp\left(-\frac{\Delta z_{i,j}^2}{L_z^2}\right)$$

Here, $d_h(i,j)$ is the horizontal distance between the i^{th} and the j^{th} grid points, Δz_{ij} is their vertical distance, σ_B^2 is the background error variance (defined as follows: $\sigma_B^2=\langle\varepsilon_B^2\rangle$, where $\varepsilon_B=x^G-x^T$), and L_h and L_z are the horizontal and vertical scaling distances. The implication of assuming B as exponentially decreasing is to spread out spatially the information from local observations that contribute to the corrections of the state variables in neighboring locations. Defining ε^2 as the ratio of the observation error variance to the background error variance ($\varepsilon^2=\sigma_0^2/\sigma_B^2$) and dividing the two error covariance matrixes R and B by σ_B^2, the diagonal elements of R become equal to $\varepsilon^2 I$ and the parameter ε^2 becomes a single-tuning parameter (Messina et al. 2011). The effect of the inclusion of error statistics is to provide a blend between the background field and the local data, depending on the value of ε^2. For relatively small observation errors and large background errors, the analysis will converge toward the observations; while for large observation errors, the results will be more dependent on the background fields.

Successive correction method

The second approach is based on the Bratseth technique (Bratseth 1986), a successive correction method that converges to OI due to the inclusion of background and observation error statistics (Kalnay 2003). The analysis is initialized with a background field (first guess modeled field) which is then modified by taking into account the local data. In this work, we have followed the approach described in the ARPS Data Assimilation System to implement this method (see http://www.caps.ou.edu/ARPS/arpsdoc.html).

The analysis of a model variable, s, at grid point x is performed using an iterative approach as follows:

$$s_x(n) = s_x(n-1) + \sum_{j=1}^{nobs} \alpha_{xj}\left[s_j^0 - s_j(n-1)\right]$$

where n is the iteration, s_j^0 is the observations, $s_j(n-1)$ is the analysis values at the observation locations at the previous iteration (on the initial pass over the grid the background field is used) and α_{xj} are the weights at the grid points The analysis value at the observation location i, $s_i(n)$, is given by the following expression:

$$s_i(n) = s_i(n-1) + \sum_{j=1}^{nobs} \alpha_{ij}\left[s_j^0 - s_j(n-1)\right]$$

where $s_i(n-1)$ is the analysis value at the previous iteration (on the initial iteration it corresponds to the background value interpolated to the observation location). The weights α_{xj} and α_{ij} at the grid points and at the observation locations, respectively, are given by:

$$\alpha_{xj} = \frac{\rho_{xj}}{m_j}; \alpha_{ij} = \frac{\left(\rho_{ij} + \varepsilon^2 \delta_{ij}\right)}{m_j}$$

where ρ_{xj} and ρ_{ij} are the correlation coefficients of the grid points and observations, respectively, δ_{ij} is the Kronecker delta which is one when $i=j$ and zero otherwise and ε^2 is, as stated before, the ratio of the observation error variance to the background error variance. The correlation coefficients ρ are assumed to be Gaussian functions, allowing the weights to decrease asymptotically to zero with increasing distance between the observation and analysis points:

$$\rho(i,k) = \exp\left(-\frac{d_h^2(i,k)}{L_h^2}\right)\exp\left(-\frac{\Delta z_{i,k}^2}{L_z^2}\right)$$

Here, $d_h(i,k)$ is the horizontal distance between the i^{th} (grid point or observation) and the k^{th} observation and Δz_{ik} the vertical distance between them. In the formula for the weights, the parameters m_j represent the local data density around the analysis point and are given by following expression:

$$m_j = \varepsilon^2 + \sum_{i=1}^{nobs} \rho_{ij}$$

Spatial representativeness of measurements

When combining the measurements and model results, it is important to take into account the so-called "lack of representativeness" errors, which can be defined as "the typical deviances or differences that occur between model calculated and observed concentrations, if their spatial and/or temporal positions, or averaging characteristics, do not match" (Walker et al. 2006). Observational error variances, therefore, derive from two different sources: instrumental and associated with local phenomena (e.g., emissions, local flows, and turbulence) at spatial scales not resolved by the underlying model. The second error is denoted "error of representativeness." Assuming that these errors are not correlated (Kalnay 2003), the observational error variance σ_0^2 is the sum of the instrument error variance and the representativeness error variance. According to Elbern et al. (2007), the representativeness error can be expressed by the following formula:

$$\varepsilon_{repr} = \varepsilon_{abs}\sqrt{\frac{\Delta x}{L_{repr}}}$$

where Δx is the grid resolution of background field, L_{repr} is the characteristic length of the observations (e.g., the radius of

influence associated with different types of ground based stations), and ε_{abs} is a tuning parameter called "characteristic absolute error." Pagowski et al. (2010) found experimentally that $\varepsilon_{abs} = \frac{1}{2}\varepsilon_{instr}$ and suggest the following values for L_{repr}: 10, 4, and 2 km respectively for rural, suburban, and urban stations. Consequently, the following expression for the ε^2 parameter is obtained:

$$\varepsilon^2 = \frac{\sigma_0^2}{\sigma_B^2} = \frac{\sigma_{instr}^2 + \sigma_{repr}^2}{\sigma_B^2} = \frac{\langle \varepsilon_{instr}^2 \rangle + \langle \varepsilon_{repr}^2 \rangle}{\sigma_B^2}$$
$$= \frac{\langle \varepsilon_{instr}^2 \rangle + \frac{\Delta x}{4L_{repr}} \langle \varepsilon_{instr}^2 \rangle}{\sigma_B^2} = \frac{\sigma_{instr}^2}{\sigma_B^2}\left(1 + \frac{\Delta x}{4L_{repr}}\right)$$

or alternatively:

$$\varepsilon^2 = \hat{\varepsilon}^2 \cdot \frac{\left(1 + \frac{\Delta x}{4L_{repr}}\right)}{\left(1 + \frac{\Delta x}{4\max(L_{repr})}\right)}$$

where:

$$\hat{\varepsilon}^2 = \frac{\sigma_{instr}^2}{\sigma_b^2}\left(1 + \frac{\Delta x}{4\max(L_{repr})}\right)$$

$\hat{\varepsilon}^2$ represents the actual tuning parameter: ε^2 will coincide with $\hat{\varepsilon}^2$ for stations having the maximum influence radius L_{repr} (e.g., remote stations) and will have higher values for stations with lower spatial representativeness (e.g., results will be more dependent on the background field). Consequently, the spatial features of the assimilated fields will also depend on the values assumed by the characteristic lengths associated with each monitoring station. These DA methods and the procedure adopted to identify the optimal values for L_h and L_z (which are pollutant and domain scale dependent) have been tested on a dataset, described in the following section. The dataset was produced using an AQMS developed to analyze the effects of different emission/dispersion characteristics on air quality over the Lazio region (Central Italy) and the Rome conurbation.

AQMS application

The city of Rome and the Lazio Region are characterized by high ambient concentrations of particulate matter, NO_2, and O_3. The annual average concentration of NO_2 is above the EU standard value recommended for human health protection (40 μg m^{-3}), while the hourly regulatory limit of 180 μg m^{-3} for ozone is exceeded several times a year at some monitoring stations. As for PM_{10}, the mean annual value is above the regulatory limit of 40 μg m^{-3} at urban traffic stations, and the maximum number (35) of exceedances of the daily human health protection limit, fixed at 50 μg m^{-3}, is unattained in busy streets. According to the EU Directive (EC 2008), air quality assessment and management has to be performed by means of a combined use of monitoring data, emission inventories, and modeling techniques. In this perspective, an AQMS has been developed with a background domain, covering a large portion of Central Italy (66×58 cells at 4 by 4 km) and a target domain including Rome urban area (61×61 cells with 1 by 1-km resolution).

AQMS description

The AQMS is based on the Flexible Air quality Regional Model (FARM) (Gariazzo et al. 2007; Silibello et al. 2008) and includes subsystems used to reconstruct flows and related turbulence parameters, to apportion data from the emission inventories to grid cells and to calculate the air quality indicators required by the EC directives. FARM employed the SAPRC-90 (Carter 1990) chemical mechanism and the *aero3* modal aerosol scheme from the CMAQ framework (Binkowsk 1999; Binkowski and Roselle 2003). The meteorological fields were produced by the prognostic and non-hydrostatic model RAMS (Cotton et al. 2003) using a two way nested grid system. Figure 1 shows the domains used by the RAMS and FARM models.

Deposition velocities of gas-phase species, horizontal/vertical diffusivities and natural emissions fields (sea salt, wind-blown dust, and biogenic VOCs) required by FARM were calculated using an interface module (Finardi et al. 2005) depending on meteorological parameters (e.g., wind speed, solar radiation, temperature) and land-use characteristics (e.g., soil type). This AQMS is routinely used by the Lazio Region Environmental Protection Agency (ARPA Lazio) to produce air quality forecasts, to assess air quality, and to evaluate the impact of different emission control strategies over the region and Rome.

Traffic emissions

A detailed methodology, based on the Traffic Emission Factors Improved Calculation (TREFIC) model (Nanni et al. 2004), was adopted to estimate road traffic emissions (Gariazzo et al. 2007). TREFIC follows the COPERT III approach and includes, for particulate matter, the emission factors developed by the International Institute for Applied Systems Analysis (IIASA) (IIASA 2001) that consider both exhaust and non exhaust (tires, brakes, road coating) sources. Vehicle flow data for the highway network, provided by the

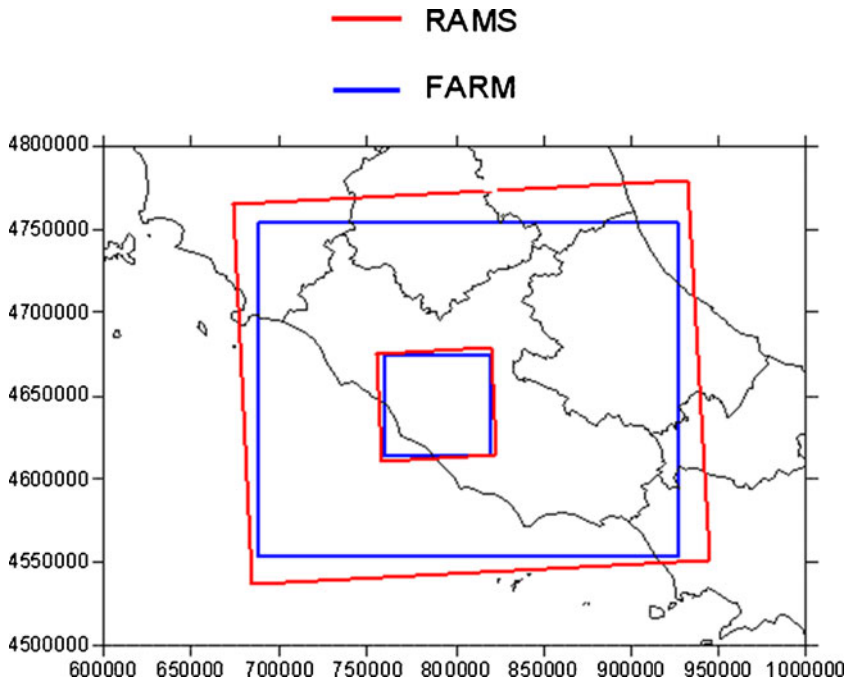


Fig. 1 Nested domains used by RAMS and FARM models over Lazio (Central Italy) and Rome conurbation (4 and 1-km horizontal resolutions)

Italian Association of Motorway and Tunnel Concessionaire Companies, together with information on the characteristics of the vehicle fleet, are utilized to estimate the emissions. Traffic data on the main roads and limitations on the circulation of some categories of vehicles (e.g., non catalytic, EURO 1, etc.) in specific zones, are also taken into account to estimate the related emissions for Rome urban road network.

Other emissions

Diffuse emissions were taken from the national emission inventory and were projected to the simulated year using national trends differentiated for each pollutant and activity (see http://www.sinanet.isprambiente.it/it/sia-ispra/serie-storiche-emissioni/serie-storiche-delle-emissioni-nazionali-di-inquinanti-atmosferici-1980-2010/view that reports the total emissions of airborne pollutants from 1980 to 2010). A detailed analysis of the emissions from the heating sector and the on-field burning of stubble has revealed a significant underestimation of PM_{10} and NMVOC emissions, probably due to the evaluation of the firewood used and the stubble burnt (Caserini et al. 2007). As reported by Caserini et al. (2007), the large variability of PM emission factors for wood combustion (depending on type of wood, combustion devices, etc.) is an important source of uncertainty for PM emissions. For this reason, the values proposed in the RAINS-Italy model (Vialetto et al. 2005; Zanini et al. 2005) were used for these two sectors and for other sources not included in the national inventory such as construction and other residential combustion activities (fireworks, cigarette smoking, meat frying, food preparation, and barbeques). The largest industrial facilities were considered as point sources with emission rates derived from stack measured data and owner declarations to local control authorities.

Boundary conditions

Boundary conditions for the 4-km FARM grid were provided by QualeAria modeling system (http://www.aria-net.eu/QualeAria/index_en.html) which simulates regional scale air quality over the Italian peninsula starting from national and European emission inventories, synoptic scale weather analysis, and global scale air quality indicators. It was developed within the research project FUMAPEX, funded by the European Commission within the Fifth Framework Program, and the COST Action ES0602 collaboration framework.

AQMS evaluation

In this work, we have used hourly PM_{10}, NO_2, and O_3 concentrations from the AQMS over the two domains for the year 2012 (base case simulation). The performance of the AQMS in the base case simulation was initially evaluated by comparing predicted concentrations with observed values measured at air quality monitoring network stations of the Lazio Region (Table 1). Table 1 shows the station coordinates, type, representativeness, and measured pollutants. The station representativeness values reported in the table were proposed by the personnel managing the network on the basis of their location (site, buildings, and emission sources nearby). It has to be emphasized that the assignment of proper L_{repr} values, in large metropolitan areas, is particularly challenging because of the

Table 1 Lazio Region monitoring network

Station	UTM *x* (km)	UTM *y* (km)	*z* a.s.l. (m)	L_{repr} (km)	Type	PM_{10}	NO_2	O_3
Francia	**787.6**	**4,649.8**	**43**	**1**	**Urban traffic**	**V**	**V**	
Grecia	**791.2**	**4,642.7**	**49**	**1**	**Urban traffic**	**V**	**V**	
Fermi	**788.0**	**4,640.5**	**26**	**1**	**Urban traffic**	**V**	**V**	
Tiburtina	**794.3**	**4,645.9**	**32**	**1**	**Urban traffic**	**V**	**V**	
Arenula	**788.3**	**4,643.8**	**31**	**2**	**Residential**	**V**	**V**	**V**
Guidonia	**808.7**	**4,656.0**	**89**	**2**	**Residential**	**V**	**V**	
Ciampino	**799.7**	**4,633.6**	**134**	**2**	**Residential**	**V**	**V**	
Preneste	**793.8**	**4,643.9**	**37**	**3**	**Urban/industrial**	**V**	**V**	**V**
Cinecittà	**796.2**	**4,640.1**	**53**	**3**	**Urban/industrial**	**V**	**V**	**V**
Bufalotta	**792.9**	**4,650.0**	**41**	**3**	**Urban/industrial**	**V**	**V**	**V**
Cipro	**786.0**	**4,645.1**	**31**	**3**	**Urban/industrial**	**V**	**V**	**V**
Malagrotta	**777.6**	**4,641.3**	**50**	**3**	**Urban/industrial**	**V**	**V**	**V**
Villa Ada	**790.8**	**4,648.3**	**50**	**5**	**Urban background**	**V**	**V**	**V**
Guido	**771.0**	**4,642.7**	**61**	**10**	**Rural background**	**V**	**V**	**V**
Cavaliere	**803.4**	**4,648.6**	**48**	**10**	**Rural background**	**V**	**V**	**V**
Civitavecchia Porto	730.6	4,664.3	6	1	Urban traffic	V		
Civitavecchia Albani	731.4	4,664.6	34	1	Urban traffic	V		
Cassino	903.3	4,604.3	41	1	Urban traffic	V	V	
Frosinone Scalo	860.8	4,617.1	161	1	Urban traffic	V	V	
Latina Romagnoli	825.1	4,598.4	23	1	Urban traffic	V	V	
Civita Castellana	781.6	4,687.5	139	1	Urban traffic	V	V	
Colleferro Oberdan	833.1	4,627.6	219	2	Residential	V	V	V
Colleferro Europa	833.5	4,627.0	223	2	Residential	V	V	
Civitavecchia	731.8	4,663.8	26	2	Residential	V	V	V
Ceccano	861.7	4,611.0	130	2	Residential	V	V	
Ferentino	853.7	4,624.4	316	2	Residential	V	V	
Alatri	860.9	4,628.6	445	2	Residential	V	V	
Latina Tasso	826.8	4,597.7	21	2	Residential	V	V	V
Latina Scalo	829.3	4,605.3	18	2	Residential	V	V	
Rieti	817.5	4,701.9	397	2	Residential	V	V	V
Viterbo	755.8	4,701.3	338	2	Residential	V	V	V
Anagni	845.1	4,630.0	401	5	Urban background	V	V	
Frosinone Mazzini	862.2	4,618.9	153	5	Urban background	V	V	V
Allumiere	740.3	4,671.4	542	10	Rural background	V	V	V
Fontechiari	889.2	4,623.6	388	10	Rural background	V	V	V
Aprilia	804.5	4,611.3	83	10	Rural background	V	V	
Leonessa	825.2	4,721.0	948	10	Rural background	V	V	V
Acquapendente	735.5	4,735.6	377	10	Rural background	V	V	V

Stations coordinate (UTM: Zone 32), Station representativeness (from personnel network judgment); measured pollutants (PM_{10}, NO_2 and O_3). Bold data are associated to stations belonging to the Rome domain

high spatial unevenness (e.g., open areas, street canyons, high and low populated districts) which influences the pollutant dispersion and transformation, and consequently, the spatial representativeness of observations. Due to the time resolution of the monitoring data, daily average PM_{10} concentrations have been used for the full year, while hourly average concentrations estimated for the months of January and July have been considered for NO_2 and O_3, respectively. We have limited the analysis to these months for NO_2 and O_3 because (1) during these periods, the investigated pollutants reach their highest levels (January for NO_2 and July for O_3); (2) the calculation of the optimal values for horizontal and vertical scaling distances is believed to be independent from the considered period and more related to the model's domain

characteristics (dimension and spatial resolution) and the spatial distribution of monitoring stations (different for the different pollutants). Moreover, the 2-months analysis permits to reduce the time required to estimate these parameters.

Figure 2 reports the evaluation of the base case simulations considering yearly averaged values for PM_{10} and the monthly mean of the highest daily values for NO_2 (January) and O_3 (July). These results are useful to check the model capability in predicting the extreme concentrations which are important from a health perspective. From this figure, it can be seen that for the Rome domain, the AQMS predictions are within a factor of two of the observations (with a better agreement for PM_{10} and O_3). It is also clear that the observed PM_{10} and NO_2 levels are underestimated when the regional scale is analyzed. A more comprehensive evaluation of model results confirmed that AQMS simulations are more accurate over the higher resolution domain. As an example, the ranges of the Pearson correlation coefficient (r) for the considered pollutants are the following:

- PM_{10}=0.52≤r≤0.71 for Rome domain and 0.18≤r≤0.71 for Lazio domain;
- NO_2=0.35≤r≤0.59 for Rome domain and 0.04≤r≤0.56 for Lazio domain;
- O_3=0.54≤r≤0.77 for Rome domain and 0.18≤r≤0.67 for Lazio domain.

Application of DA methods

The results presented in the previous section can be improved using DA techniques, provided that the optimal values of horizontal and vertical scaling distances L_h and L_z associated with each of the three pollutants within the two domains, are found.

To choose the optimal values for these parameters, we have performed several numerical experiments using the SCM method which require relatively little computational effort; as it converges to the OI method (see above), the results can be extended to the OI method. For $\hat{\varepsilon}^2$, we assumed a value of 0.1. This means that observations associated with stations having the maximum influence radius ($max(L_{repr})$, rural background stations) are assumed to be more accurate than the background estimates (Daley 1991).

Calculation of optimal values for L_h and L_z

To identify the optimal values of L_h and L_z, the following combinations for these parameters were considered:

- L_h Lazio domain: 5, 10, 20, 30, 40, and 50 km; Rome domain: 1, 5, 10,15, 20, and 25 km;
- L_z 100, 200, 400, 800, and 1,000 m for both domains.

For each combination of parameters values (e.g., pollutant: PM_{10}; domain: Lazio; L_h=5 km; L_z=100 m), the statistical performance indexes have been calculated for each monitoring station excluding it from the SCM analysis ("leave-one-out cross validation"). We applied this procedure iteratively for all stations, for the three pollutants (PM_{10}: daily averaged concentrations for the full year; NO_2 and O_3: hourly averaged concentrations for January and July, respectively) and for each combination of L_h and L_z. Tables 2 and 3 show the combinations of these parameters which minimize the root mean square error (RMSE) at each monitoring site. These tables report also the median, the average, and the proposed values. A value of 500 m for the vertical scaling distance may be considered optimal for both domains. For L_h, the following values were identified as follows: 20 km for both domains for NO_2; 20 and 10 km for regional and local domains for PM_{10}, and 40 and 20 km for O_3. It has to be emphasized that these values are related to pollutant characteristics, model

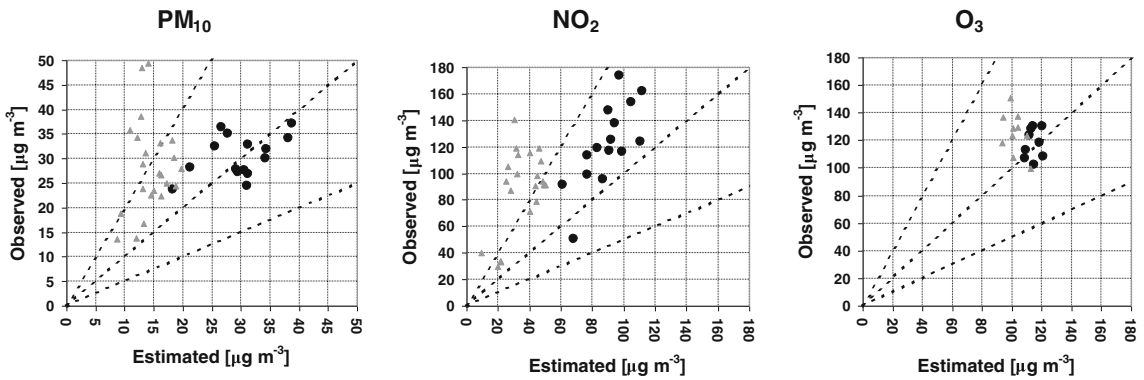


Fig. 2 Evaluation of the base case simulations. *PM_{10}* yearly average values; *NO_2* January monthly mean of highest daily values; *O_3* July monthly mean of highest daily values. *Black circles* and *gray triangles* correspond to the Rome and Lazio domains

Table 2 Values of L_h [km] and L_z [m] that minimize RMSE performance index over the Lazio Regional basin for the following pollutants: PM_{10}, NO_2, and O_3

Station	O_3		PM_{10}		NO_2	
	L_h	L_z	L_h	L_z	L_h	L_z
Preneste	30	200	10	100	5	800
Cinecittà	30	400	20	200	20	1,000
Colleferro Oberdan	40	1,000	20	200	40	400
Allumier	30	100	40	100	20	100
Civitavecchia	40	200	40	800	40	400
Villa Ada	40	1,000	5	800	5	100
Guido	30	200	30	1,000	5	100
Cavaliere	40	100	40	100	40	1,000
Bufalotta	40	100	10	100	10	100
Cipro	30	200	5	800	20	800
Arenula	30	100	5	800	5	1,000
Malagrotta	40	800	10	100	10	100
Fontechiari	40	1,000	40	200	20	200
Frosinone Mazzini	30	200	10	100	5	800
Latina Tasso	40	400	5	100	5	800
Rieti	40	800	40	1,000	40	1,000
Leonessa	40	1,000	30	1,000	20	1,000
Viterbo	30	1,000	40	1,000	40	1,000
Acquapendente	40	200	40	200	30	100
Median	40	200	20	200	20	800
Average	40	500	20	500	20	600
Proposed	40	500	20	500	20	500

Table 3 Values of L_h (km) and L_z (m) that minimize RMSE performance index over Rome conurbation for the following pollutants: PM_{10}, NO_2, and O_3

Station	O_3		PM_{10}		NO_2	
	L_h	L_z	L_h	L_z	L_h	L_z
Preneste	5	800	5	800	20	1,000
Cinecittà	20	1,000	10	1,000	20	100
Villa Ada	20	100	10	100	1	800
Guido	20	100	25	100	5	800
Cavaliere	20	100	15	100	20	1,000
Bufalotta	20	100	15	800	15	800
Cipro	15	800	10	800	20	100
Arenula	10	200	5	1,000	20	100
Malagrotta	20	100	10	100	20	100
Median	20	100	10	800	20	800
Average	20	400	10	500	20	500
Proposed	20	500	10	500	20	500

domain characteristics (dimension and spatial resolution), and the spatial distribution of the monitoring stations (which is different for the different pollutants).

Comparison between base case simulation and DA results

Based on the optimal combinations for L_h and L_z, the OI and SCM methods were applied and the results compared with those obtained in the base case simulation using the RMSE performance index calculated as follows (the ideal value is 0 μg m^{-3}):

$$RMSE(t) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[y_m^0(t) - y_m^f(t)\right]^2}$$

where $m=1, 2, \ldots, M$ are the monitoring sites indexes, $y_m^o(t)$ and $y_m^f(t)$ are the observed and modeled concentrations at time t. The RMSE analysis in Fig. 3 clearly indicates that the use of OI method gives better results with generally lower RMSE values than the SCM (in this figure, a secondary ordinate axis has been used because of the different scales for the base case and the assimilated RMSEs). Moreover, this figure shows the following:

- PM_{10}: lower RMSEs during warmer periods and in the higher resolution domain. These results may be ascribed to uncertainties in emissions (particularly from domestic wood stoves) and dispersion conditions during colder periods not properly modeled. The emission uncertainties seem to be greater for the regional domain, as it includes rural areas whose use of biomass for domestic heating is expected to be higher than within the Rome conurbation;
- NO_2: the analysis of the base case and the assimilated results does not show significant differences between the two domains;
- O_3: SCM and OI provide RMSEs values generally lower for the regional domain whose spatial scale (and horizontal resolution) is more appropriate to represent this secondary and ubiquitous pollutant.

Figure 4 shows an example of the spatial distribution of the yearly averaged PM_{10} concentrations over the regional domain, simulated by FARM (base case) and assimilated using the OI method. This figure clearly shows a significant underestimation of AQMS (base case) for PM_{10} levels over the Southeast part of the region, whereas the OI method shows large peaks in PM_{10} concentrations in the same area. The assimilation of observations that are influenced by local emissions not resolved by the model at the adopted spatial scale significantly modifies the resulting concentration fields, determining local maxima not apparent from the AQMS results.

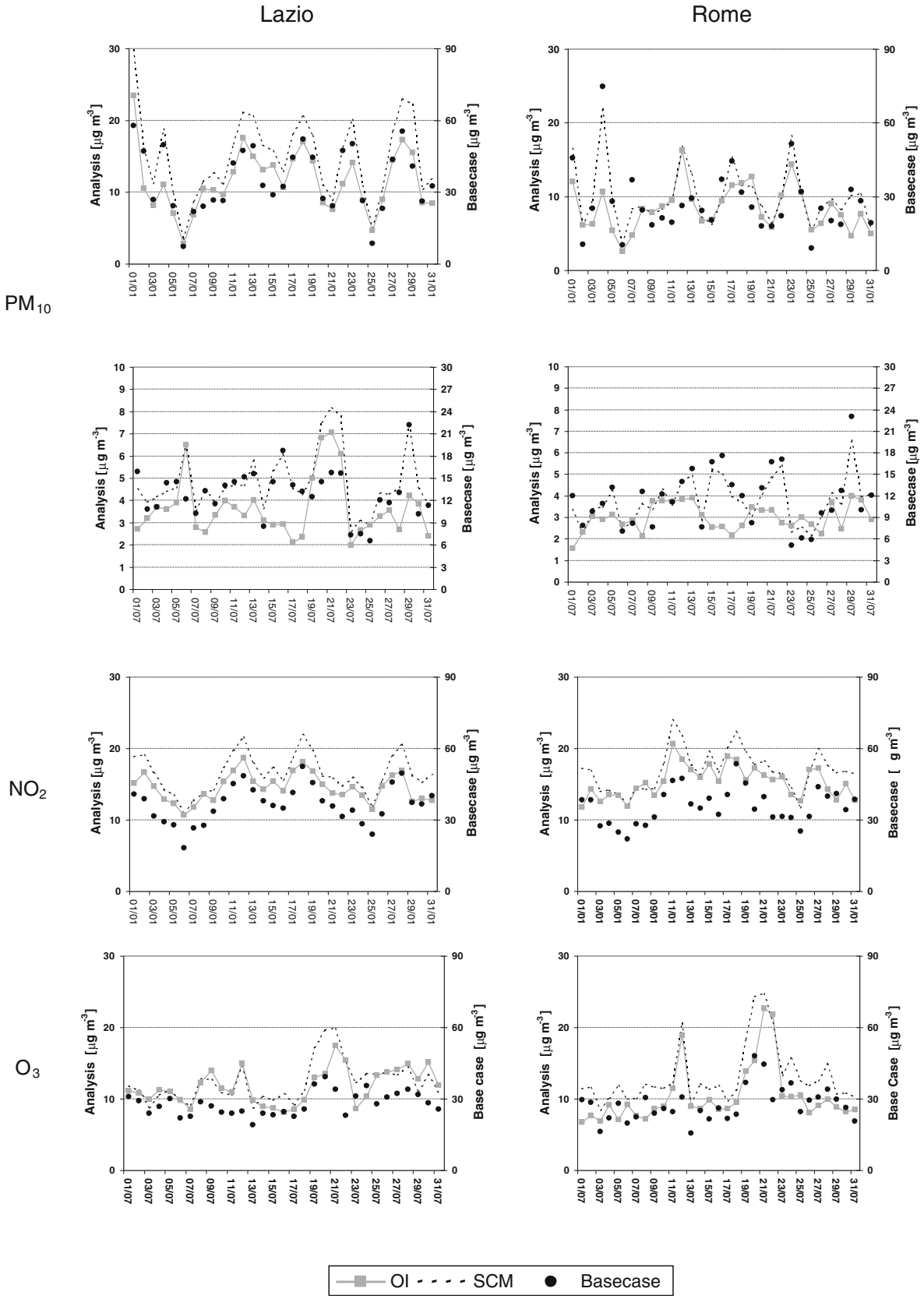


Fig. 3 Daily averaged RMSE values considering all the monitoring sites for PM_{10} (January and July 2012, *first two rows*), NO_2 (January 2012, *third row*), and O_3 (July 2012, *bottom row*) obtained from the base case, SCM, and OI data assimilation methods simulations. Note that a secondary ordinate axis has been used for the base case

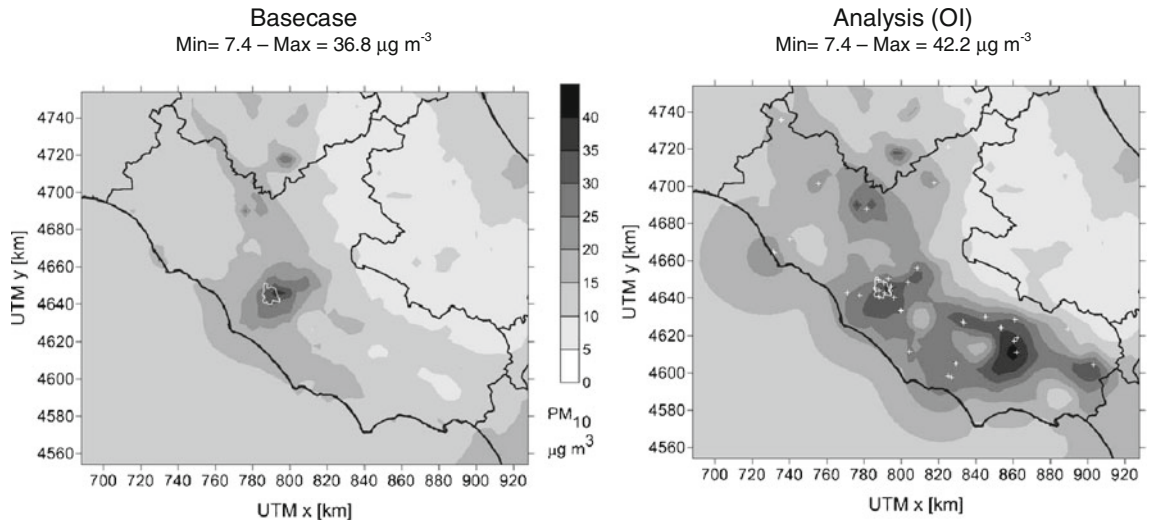


Fig. 4 Yearly averaged PM_{10} concentrations [μg m^{-3}] simulated by the AQMS (Basecase) and estimated by OI assimilation method over the regional domain. Rome urban area and monitoring sites of the Lazio region network are shown

Application to the EXPAH dataset

The identification and quantification of the polycyclic aromatic hydrocarbons (PAHs) exposure to children and elderly people in urban areas is the major goal of the population exposure to polycyclic aromatic hydrocarbons (EXPAH) LIFE+ Project (www.ispesl.it/expah). To reach these objectives, an integrated approach, based on measurements and modeling techniques, was applied to simulate PAHs levels in the Rome metropolitan area. Field campaigns of indoor/outdoor $PM_{2.5}$ with PAHs content were performed at different sites and living microenvironments between December 2011 and July 2012. The availability of further simulation results and experimental data over Rome domain (see Silibello et al. 2013 for a more detailed description of modeling activities performed within EXPAH project) suggested the use of this dataset to check the potential of DA techniques to reduce model errors and uncertainties in representing the pollutants temporal and spatial variation at the chosen model resolution. The simulation period lasted from June 2011 to May 2012 using the Rome domain depicted in Fig. 1.

Base case and assimilated $PM_{2.5}$ yearly averaged concentrations

Figure 5 reports $PM_{2.5}$ yearly averaged concentrations (microgram per cubic meter) simulated by the model (base case) and obtained by the application of the OI data assimilation

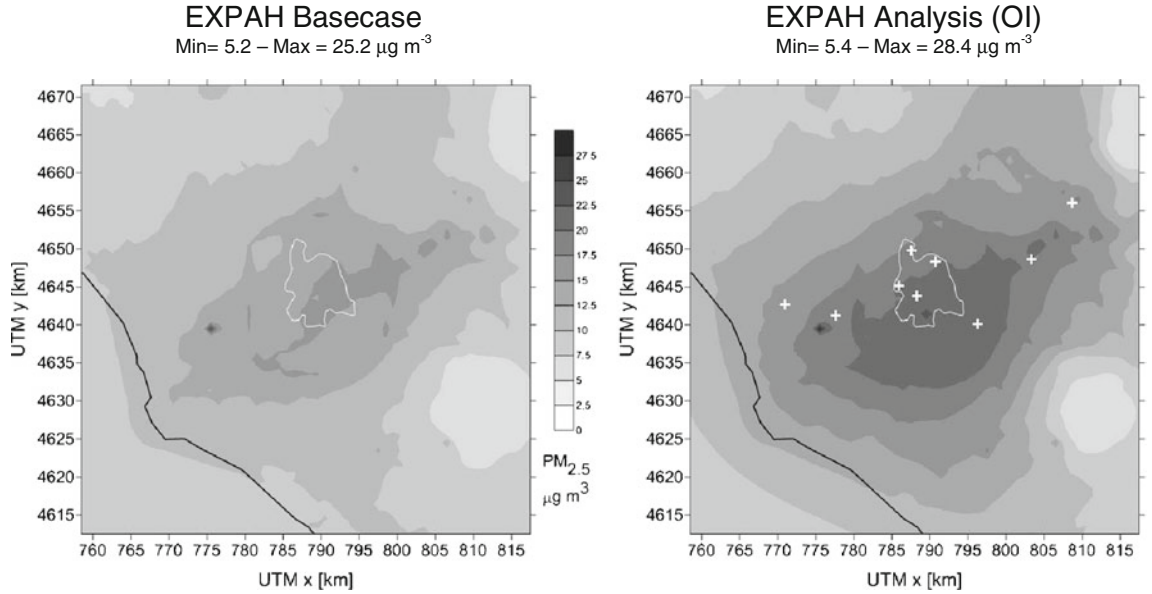


Fig. 5 EXPAH Project: $PM_{2.5}$ yearly averaged concentrations [μg m^{-3}] simulated by the model (Basecase) and obtained by the application of OI data assimilation method. Rome urban area and monitoring sites of the Lazio region network are shown

method. The figure also reports the monitoring stations which routinely measure $PM_{2.5}$ (nine, four of them outside Rome urban area). Due to the limited number of stations which routinely measure $PM_{2.5}$, an analogous optimization

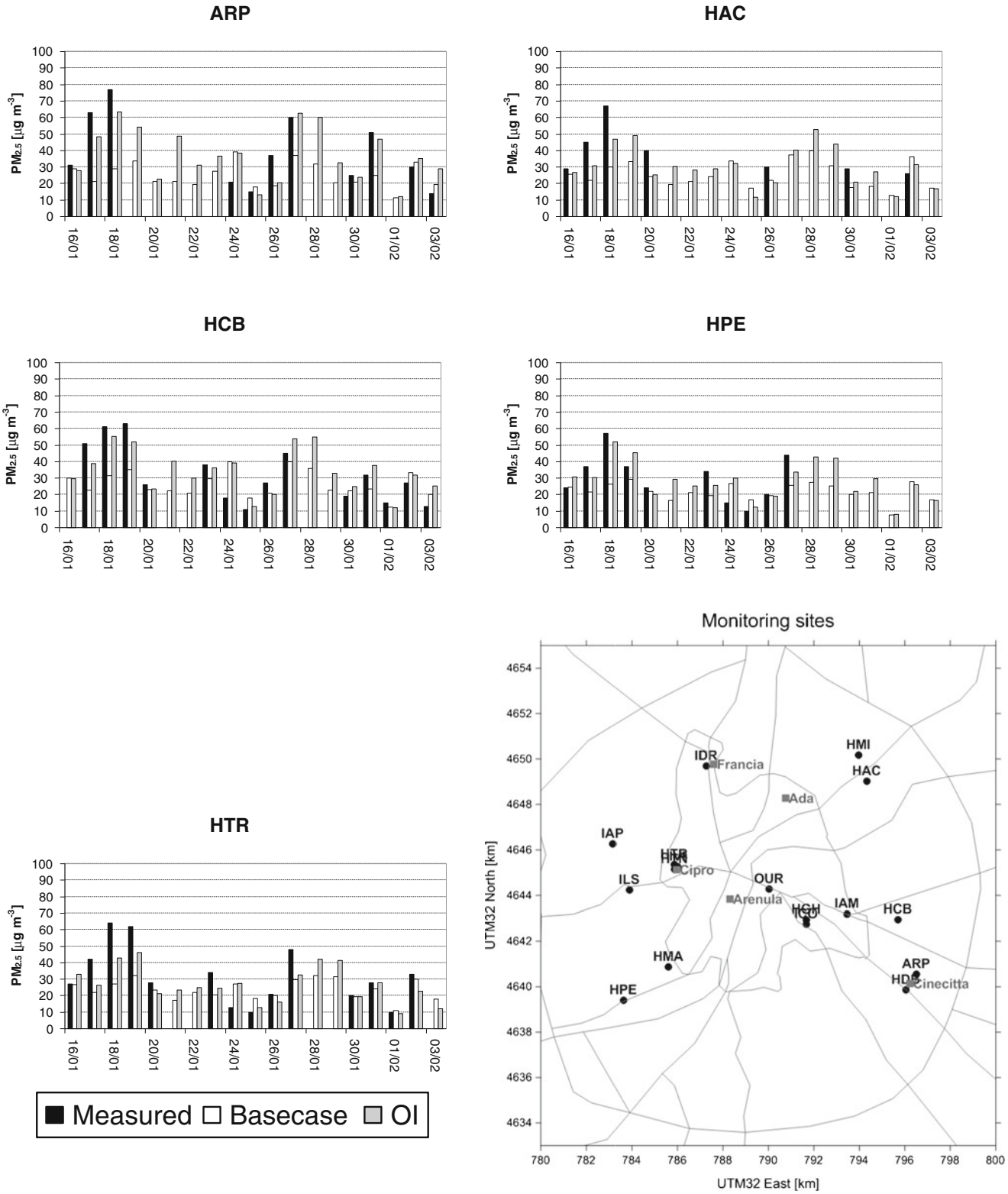


Fig. 6 EXPAH project: comparison between measured, modeled (base case) and OI assimilated concentrations at EXPAH monitoring sites during the period of 16 January–3 February 2012. *Black circles* and *gray squares* correspond to EXPAH and Lazio Region monitoring sites. EXPAH sites (HGH, ICO) and (HTR, HTN, IVI) are partially overlapped

procedure for the scaling distances could not be carried out for this pollutant. Due to its prevalent secondary nature, a L_h value of 20 km has been used (considering the values obtained for ozone and PM_{10} over the Rome domain, respectively, 20 and 10 km). A L_z value of 500 m was considered. Figure 5 reports $PM_{2.5}$ yearly averaged concentration maps (microgram per cubic meter) simulated by the model (base case) and obtained by the application of the OI method. The analysis of this figure clearly evidences the influence of observations on the resulting concentration field produced by OI, leading to a generalized increase of $PM_{2.5}$ levels within the Rome urban area.

Evaluation at EXPAH monitoring sites

The availability of additional $PM_{2.5}$ measurements, collected during the EXPAH field campaigns and not used in the assimilation process, permits the evaluation of the capability and the accuracy of this approach in reducing model uncertainty and obtaining more reliable gridded pollution fields for further epidemiological studies foreseen in the EXPAH project. As an example, in Fig. 6 are compared, observed, modeled (base case) and OI assimilated $PM_{2.5}$ concentrations at EXPAH monitoring sites during the period of 16 January–3 February 2012. The figure also shows the location of EXPAH and Lazio Region monitoring sites within the urban area of Rome. The OI-derived concentrations at EXPAH sites are generally closer to the observed values than those estimated by the AQMS alone (base case). The RMSE values reported in Table 4, considering all EXPAH experimental periods, confirms this result: the use of OI method reduces RMSE values at all EXPAH monitoring sites and produces more reliable concentration fields.

Table 4 EXPAH project: base case and OI RMSE values ($\mu g\ m^{-3}$) at EXPAH monitoring sites

Station	Base case	OI
ARP	23.4	10.7
HAC	18.7	12.1
HCB	15.4	9.0
HDB	6.3	4.2
HGH	8.4	7.0
HMA	5.0	3.2
HMI	3.9	3.4
HPE	14.0	7.8
HTN	5.6	4.3
HTR	15.8	11.0
IAM	8.5	6.8
IAP	12.7	9.3
ICO	11.5	8.9
IDR	15.6	13.9
ILS	8.3	6.5
IVI	9.4	7.8
OUR	6.8	4.4

Conclusions

SCM and OI DA methods, enhanced by introducing the concept of spatial representativeness of air quality measurements, have been used to integrate information from air quality models and monitoring networks. A reference dataset, containing hourly PM_{10}, NO_2, and O_3 concentration fields produced by an AQMS over Central Italy (4-km horizontal resolution) and the city of Rome (1-km horizontal resolution) has been used to assess the capability of DA techniques to improve the spatial distribution and the temporal variation of the studied pollutants over the modeling domains. A preliminary study was conducted to identify the optimal values for the horizontal and vertical scale distances lengths L_h and L_z for these domains. A value of 500 m for the vertical correlation may be considered optimal for both domains. For L_h, the following values were obtained, 20 km for both domains for NO_2, 20 and 10 km for the regional and local domains, respectively, for PM_{10}, and 40 and 20 km for O_3. These values are related to model domain characteristics (dimension and spatial resolution) and the spatial distribution of the monitoring stations (which is different for the different pollutants). The approach adopted in this work to derive optimal values for L_h and L_z can easily extend to other datasets and could also provide useful information in designing a monitoring network. The comparison between SCM and OI results demonstrated better performance of the OI method in terms of lowering RMSE values at the selected monitoring sites. Optimal interpolation is used at major numerical prediction centers around the world and should be preferred to the older SCM. Nevertheless, SCM is still useful because of its low computational demands and its convergences to OI method results. Further, $PM_{2.5}$ modeling results over the Rome urban area and additional measurements collected during experimental campaigns, performed within the EXPAH LIFE+ Project, allowed the evaluation of this approach in reconstructing $PM_{2.5}$ levels at EXPAH monitoring sites, which were not used in the DA process. The results clearly showed the efficacy of this method in improving the accuracy of the spatial distribution of pollution levels simulated by the AQMS dispersion model. This work confirms the capability of DA methods to reduce model errors and uncertainties in large urban areas. In such areas, some critical aspects are represented by sub-grid phenomena not resolved by the model but clear from the observations. Here, station parameters used in the DA process (e.g. number, location, and representativeness) and horizontal/vertical scale lengths, play a key role in obtaining a reliable quantitative spatial description of the actual concentration field, as

demonstrated by this work. The results obtained considering the EXPAH dataset, in which five stations routinely measuring $PM_{2.5}$ levels within Rome urban area were assimilated, evidenced that the number and the location of these monitoring stations are adequate to reconstruct, together with model results, the spatial distribution of $PM_{2.5}$ levels within Rome urban area. The availability of more reliable pollutant concentration fields at higher spatial resolutions permits, moreover, to improve human exposure models simulating the neighborhood-scale population exposure to major air pollutants, and to forecast environmental health impacts. This allows epidemiologic studies to consider spatially inhomogeneous exposure at urban levels and consequently, spatially varying health effects.

Acknowledgments The LIFE+ EU financial program (EC 614/2007) is acknowledged for the provision of funding for EXPAH project (LIFE09 ENV/IT/082). The authors wish to thanks Paola Radice (Arianet S.r.l.) for her support in preparing emission data and Sandro Finardi and Alessio D'Allura (Arianet S.r.l.) for preparing meteorological fields and managing QualeAria forecast system. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

Adhikary B, Kulkarni S, D'Allura A, Tang Y, Chai T, Leung LR, Qian Y, Chung CE, Ramanathan V, Carmichael GR (2008) A regional scale chemical transport modeling of Asian aerosols with data assimilation of AOD observations using optimal interpolation technique. Atmos Environ 42:8600–8615

Binkowsk FS (1999) The aerosol portion of Models-3 CMAQ. In science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, edited by D.W. Byun, and J.K.S. Ching, EPA-600/R-99/030, 1–23

Binkowski FS, Roselle SJ (2003) Models-3 community multiscale air quality (CMAQ) model aerosol component 1. Model description. J Geophys Res 108(D6):4183

Borrego C, Monteiro A, Payb MT, Ribeiro I, Miranda AI, Basart S, Baldasano JM (2011) How bias-correction can improve air quality forecasts over Portugal. Atmos Environ 45:6629–6641

Bratseth AM (1986) Statistical interpolation by means of successive corrections. Tellus 38A:439–447

Carter WPL (1990) A detailed mechanism for the gas-phase atmospheric reactions of organic compounds. Atmos Environ 24A:481–518

Caserini S, Fraccaroli A, Monguzzi AM, Moretti M, Angelino E (2007) New insight into the role of wood combustion as key PM source in Italy and in Lombardy region. Proc. of 16th Annual International Emissions Inventory Conference "Emission Inventories: Integration, Analysis, and Communications" Raleigh, North Carolina, May 14–17, 2007

Cotton WR, Pielke RA, Walko RL, Liston GE, Tremback CJ, Jiang H, McAnelly RL, Harrington JY, Nicholls ME, Carrio GG, McFadden JP (2003) RAMS 2001: current status and future directions. Meteorog Atmos Phys 82:5–29

Daley R (1991) Atmospheric data analysis. Cambridge University Press, Cambridge

Delle Monache L, Wilczak J, McKeen S, Grell G, Pagowski M, Peckham S, Stull R, McHenry J, McQueen J (2008) A Kalman-filter bias correction of ozone deterministic, ensemble-averaged, and probabilistic forecasts. Tellus B 60:238–249

Denby B (Editor), Brandt J, Elbern H, Frydendall J, Heemink A, Hvidberg M, Kahnert M, Tarrason L, van Loon M, Walker SE, Zlatev Z (2006) Data assimilation in regional scale atmospheric chemical models. NMR Workshop at NILU, Kjeller, Norway, 15 November 2005. NILU: OR 43/2006. ISBN:82-425-1766-5

Denby B, Spang W (2010) The combined use of models and monitoring for applications related to the European Air Quality Directive: A working sub-group of Fairmode. Proc. of 13th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Paris, France

Denby B, Horálek J, Walker SE, Eben K, Fiala J (2005) Interpolation and assimilation methods for European scale air quality assessment and mapping. Part I: Review and recommendations. ETC/ACC Technical Paper 2005/7. December 2005. Final draft

Denby B, Garcia V, Holland DM, Hogrege C (2009) Integration of air quality modeling and monitoring data for enhanced health exposure assessment. Environmental Management. Springer-Verlag, New York, NY, pp 46–49

EC (2008) Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Official Journal L 152, 11(6), 2008, pp. 1–44 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF.

Elbern H, Schmidt H (2000) Ozone episode analysis by four-dimensional variational chemistry data assimilation. J Geophys Res 106(D4): 3569–3590

Elbern H, Strunk A, Schmidt H, Talagrand O (2007) Emission rate and chemical state estimation by 4-dimensional variational inversion. Atmos Chem Phys 7:3749–3769

Finardi S (Editor), Baklanov A, Clappier A, Fay B, Joffre S, Karppinen A, Ødegard V, Slørdal LH, Sofiev M, Sokhi RS, Stein A (2005) Improved interfaces and meteorological pre-processors for urban air pollution models. FUMAPEX Report D5.2e3, Milan, Italy, 55 pp, http://fumapex.dmi.dk

Gariazzo C, Silibello C, Finardi S, Radice P, Piersanti A, Calori G, Cecinato A, Perrino C, Nussio F, Cagnoli M, Pelliccioni A, Gobbi GP, Di Filippo P (2007) A gas/aerosol air pollutants study over the urban area of Rome using a comprehensive chemical transport model. Atmos Environ 41:7286–7303

Horálek J, Kurfürst P, Denby B, de Smet P, de Leeuw F, Brabec M, Fiala J (2005) Interpolation and assimilation methods for European scale air quality assessment and mapping. Part II: Development and testing new methodologies. ETC/ACC Technical Paper 2005/8. December 2005. Final draft

IIASA (2001) RAINS-Europe http://www.iiasa.ac.at/rains/home.html.

Janssen S, Dumont G, Fierens F, Mensink C (2008) Spatial interpolation of air pollution measurements using CORINE land cover data. Atmos Environ 42:4884–4903

Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, Cambridge

Messina P, D'Isidoro M, Maurizi A, Fierli F (2011) Impact of assimilated observations on improving tropospheric ozone simulations. Atmos Environ 45:6674–6681

Nanni A, Radice P (2004) Sensitivity analysis of three EF methodologies for PM10 in use with climatological dispersion modelling in urban Italian study cases. Proc. of 9th Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes. Garmisch-Partenkirchen (Germany) 1:309–314

Pagowski M, Grell GA, McKeen SA, Peckham SE, Devenyi D (2010) Three-dimensional variational data assimilation of ozone and fine particulate matter observations: Some results using the Weather Research and Forecasting – Chemistry model and Grid-point Statistical Interpolation. Q J R Meteorol Soc 136:2013–2024

Petritoli A, Palazzi E, Giovanelli G, DiNicolantonio W, Ballista G, Carnevale C, Finzi G, Pisoni E, Volta M (2011) Combined use of space-borne observations of NO_2 and regional CTM model for air quality monitoring in northern Italy. Int J Environ Pollut 47:158–166

Physick WL, Cope ME, Lee S, Hurley PJ (2007) An approach for estimating exposure to ambient concentrations. J Expo Sci Environ Epidemiol 17:76–83

Silibello C, Calori G, Brusasca G, Giudici A, Angelino E, Fossati G, Peroni E, Buganza E (2008) Modelling of PM10 concentrations over Milan urban area using two aerosol modules. Environ Model Softw 23:333–343

Silibello C, D'Allura A, Finardi S, Radice P (2013) EXPAH-Technical report on FARM model capability to simulate PM2.5 and PAHs in the base case – Action 4.5. ARIANET R2013.06. May 2013 (available at http://www.ispesl.it/expah/documenti/R2013–06_ARIANET_EXPAH_A4.5_final.pdf).

Tombette M, Mallet V, Sportisse B (2009) PM_{10} data assimilation over Europe with the optimal interpolation method. Atmos Chem Phys 9:57–70

Vialetto G, Contaldi M, De Lauretis R, Lelli M, Mazzotta V, Pignatelli T (2005) Emission scenarios of air pollutants in Italy using integrated assessment models. Pollut Atmosphérique 185:71

Walker SE, Schaap M, Slini L (2006) Data assimilation, Air4EU WP6 synthesis. Milestone Report 6:8

Zanini G, Pignatelli T, Conforti F, Vialetto G, Vitali L, Brusasca G, Calori G, Finardi S, Radice P, Silibello C (2005) The MINNI Project: an integrated assessment modeling system for policy making. proceedings of the 15th MODSIM congress: "Advances and applications for management and decision making", Melbourne, Australia

Zhang Y, Bocquet M, Mallet V, Seigneur C, Baklanov A (2012) Real-time air quality forecasting, part II: state of the science, current research needs, and future prospects. Atmos Environ 60:656–676