Environmental Modelling & Software 74 (2015) 66–74

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A new approach to design source–receptor relationships for air quality modelling

A. Clappier [a], E. Pisoni [b,*], P. Thunis [b]

[a] Université de Strasbourg, Laboratoire Image Ville Environnement, 3, rue de l'Argonne, 67000 Strasbourg, France
[b] European Commission, JRC, Institute for Environment and Sustainability, Air and Climate Unit, Via E. Fermi 2749, 21027 Ispra, VA, Italy

ARTICLE INFO

Article history:
Received 24 March 2015
Received in revised form
9 September 2015
Accepted 15 September 2015
Available online 25 September 2015

Keywords:
Air quality modeling
Integrated assessment modeling
Surrogate models
Source–receptor relationships

ABSTRACT

Air quality models are often used to simulate how emission scenarios influence the concentration of primary as well as secondary pollutants in the atmosphere. In some cases, it is necessary to replace these air quality models with source–receptor relationships, to mimic in a faster way the link between emissions and concentrations. Source–receptor relationships are therefore also used in Integrated Assessment Models, when scenario responses need to be known in very short time. The objective of this work is to present a novel approach to design a source–receptor relationship for air quality modeling. Overall the proposed approach is shown to significantly reduce the number of simulations required for the training step and to bring flexibility in terms of emission source definition. A regional domain application is also presented, to test the performances of the proposed approach.

1. Introduction

Air quality models are complex tools which include detailed representations of the transport, diffusion and chemical processes taking place in the atmosphere. These models work at various horizontal and vertical resolutions and account for the non-linear interactions in each of the processes previously mentioned.

One of the main advantages of AQ models is the possibility to assess the impact of emission changes on concentration levels. The easiest approach is to modify the emissions accordingly, run the model and check the resulting concentrations. This is generally referred to as using a model in "scenario mode". One of the consequences of the high complexity of AQ models is their associated CPU time implying that AQ models can only be run for a limited set of scenarios due to this important constraint. If the number of required scenarios becomes prohibitive, one way out is to design source–receptor relationships (or models), which is a simplified modeling approach that will mimic the full air quality model behavior. Air quality integrated assessment models use this approach when scenario responses need to be known in very short time within an optimization process. In this case a simplified model is constructed to link the emission changes to the concentration changes. The same type of simplified model is also very useful in scenario mode when a user wishes to assess the impact of several possible emission reductions on concentrations without requiring the long computation times that come with running the full AQ model.

Many examples of this approach do exist in literature. Seibert and Frank (2004) developed linear-source receptor relationships to compute the transport of atmospheric traces substances with a Lagrangian particle dispersion model. Simpson et al. (1997) and Tarrason et al. (2004) used the Eulerian EMEP model as basis to compute country-to-country source–receptor relationships over a European domain, considering a multi-annual time-frame. At the national scale, Vedrenne et al. (2014) used a similar approach over Spain with the Atmospheric Evaluation and Research Integrated model for Spain (AERIS). This model allows for assessing the impact of sectorial emission reductions on air quality. All the above mentioned techniques however rely on a large number of computer simulations to identify the source–receptor models.

Alternative approaches also exist to assess the impacts of emission scenarios on air quality. The decoupled-direct method (DDM) (Dunker, 1981, Dunker et al., 2002) or its adjoint sensitivity complementary method (Sandhu et al., 2005; Hakami et al., 2006) provide sensitivity coefficients based on an initial set of nonlinear, partial differential equations. These sensitivity coefficients can then

* Corresponding author.
E-mail address: enrico.pisoni@jrc.ec.europa.eu (E. Pisoni).

http://dx.doi.org/10.1016/j.envsoft.2015.09.007

be used to describe how emissions impact air pollution concentrations. The source-oriented external mixture (SOEM) method (Ying and Kleeman, 2006) or the Particulate Source Apportionment Technology (PSAT) method (Wagstrom et al., 2008) use a source-oriented Eulerian air quality model and monitor the formation of PM2.5 nitrate, sulfate and ammonium ion from primary particles and precursor gases emitted from different sources. However these techniques usually require an a-priori definition of the sources and receptors to be tracked with a requested computer time quickly increasing with the number of these tracked sources/receptors. In this work we will only focus on the formulation of source–receptor relationships on the basis of an Eulerian air quality model.

Given the scope of the source–receptor (SR) model (e.g. focus on yearly averaged model responses), an experiment is designed and assumptions are made to construct the SR model. This experiment consists in the following steps: (1) running several times the full AQ model under selected conditions (training), (2) designing the SR model to mimic at best the source-receptor relations of the full AQ model and (3) validate the SR model on a series of independent simulations. All three steps need to be designed to make sure the SR model becomes a good representation of the full air quality model for the desired scope and that the assumptions made during the derivation are robust.

The objective of this work is to present a novel approach to design a source–receptor model for air quality modeling. The main advantages of the proposed approach lie in the reduced number of simulations required for the training step as well as in the flexibility it brings in terms of source definition. In a first section, we will introduce the overall problem of source–receptor relationships. We then detail possible approaches to simplify the problem both in terms of sources and receptors. In a second section an application on a regional domain is presented, where different configurations are tested.

2. Methodology

2.1. Setting the "source–receptor" problem

The full AQ model operates for a given time period over a given geographical area. Both the emission input and the concentration output are spatially gridded over this geographical area. Although it is the same grid, we make here a distinction between the source (emissions) and the receptor grids (concentrations) for convenience.

Prior to any assumption being made, each receptor cell relates with every source cells, i.e. each grid cell concentration depends on the emissions coming from every grid within the domain. Moreover, different emission precursors can have an effect on the given concentration, i.e. emissions from each precursor in every cell relates to the concentration observed in a single cell (Fig. 1). This relation between precursor and receptor cells can be formalized mathematically using a relation containing several coefficients. For linear relationships, only one coefficient is required to link one receptor cell to one source cell but the number of coefficients increases with the degree of non-linearity characterizing the relations. The exact number of coefficients will depend on the shape of the non-linear function used (or the sum of non-linear function used). These functions which characterize the S/R relationship are identified by performing different scenario simulations with the full AQ model.

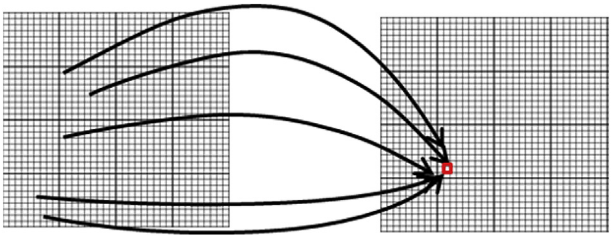

Fig. 1. General configuration of a source–receptor model, separating source (emissions) and receptor (concentrations) grids.

In this work we will focus on particulate matter (PM10) yearly averaged concentrations which are assumed to depend on the following five emission precursors: Nitrogen oxides (NOx), Volatile Organic Compounds (VOC), Sulfur Dioxides (SO2), Primary Particulate Matter (PPM) and Ammonia (NH3). Since Thunis et al. (2015) showed in their application of a full AQ model over three domains in Europe that non-linear effects were marginal and could be neglected for yearly/seasonal averaged PM10 model responses, we will assume linearity in the following derivations and considerations, leading to the following equation for a given receptor cell "j":

$$C_j = C_j^0 + \sum_p^P \sum_i^N a_{ij}^p E_i^p \tag{1}$$

where N is the number of source grid cells within the domain and P is the number of precursors. As seen from this equation, the relation between each precursor (p) in each source cell (i) and a receptor cell (j) is linear, therefore characterized by one constant (a_{ij}^p). Although a linear relation needs only one coefficient to be defined the number of unknowns (P × N+1, i.e. the N × P a_{ij}^p coefficients plus the background C_j^0) remains important and can easily lead to a non-manageable number of simulations to be performed with a full AQ model (e.g. the number of unknowns would reach 12501 for a domain of 50 × 50 cells and 5 precursors).

The coefficients a_{ij}^p in Equation (1) are the absolute potencies described in Thunis and Clappier (2014) defined as the ratio of the concentration change (with respect to the base-case) to the associated emission change, for a given scenario. It is also important to point out that while the approach followed in Equation (1) is precursor driven, the approach could easily be adapted to macro-sectors.

In the next two sub-sections we will analyze how the problem can be simplified by aggregating the source cells and/or receptor cells (i.e. the coefficients a_{ij}^p will be assumed to be constant over a range of source and/or receptor cells).

2.2. Fixed source aggregation and 1 cell receptor window

In this configuration, source grids are aggregated in fixed geographical entities, e.g. countries, regions or set of regions/countries while receptor grids are still considered cell per cell (Fig. 2 top, illustrated for two countries). This approach is used in GAINS-EU (Amann et al., 2011), GAINS-IT (Mircea et al., 2014 and D'Elia et al., 2009) or in the TM5-FASST models.

The number of unknowns is then directly proportional to the number of geographical entities selected. Equation (1) then transforms into:

$$C_j = C_j^0 + \sum_{p=1}^P \sum_{i=1}^{N_A} a_{ij}^p E_i^p \tag{2}$$

where N_A is the number of emission aggregation zones selected and E_i^p is the precursor "p" emission of the fixed entities "i". In this case the minimum number of required scenarios is equal to the number of unknowns (i.e. P × N_A+1).

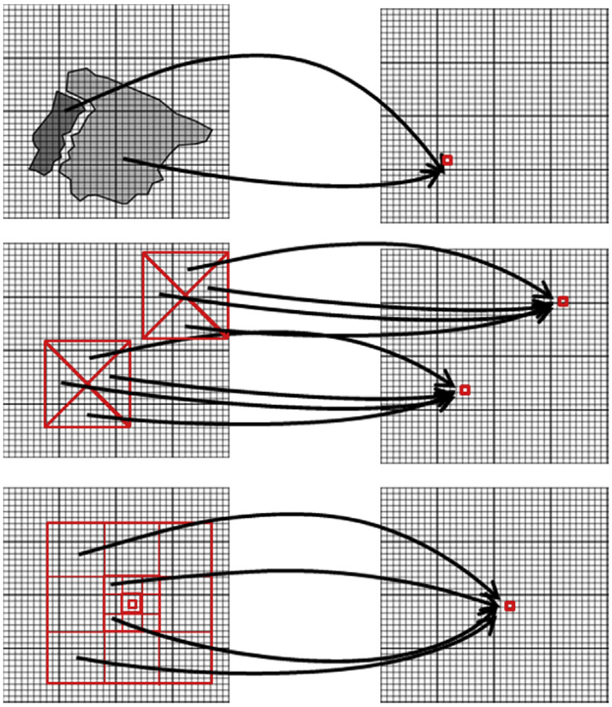

Fig. 2. Various configurations for source (emission) aggregations, considering fixed entitites (top) and sliding entities as quadrants (middle) and multiring (bottom).

The model results (scenarios and base case) are used to write a system of equations to be solved to find the different unknowns. The following system is written for a given receptor cell "j":

$$
\begin{aligned}
C_j^{BC} &= C_j^0 + \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^p E_i^{p,BC} \\
C_j^{SC(1)} &= C_j^0 + \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^p E_i^{p,SC(1)} \\
&\vdots \\
C_j^{SC(N_A\times P)} &= C_j^0 + \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^p E_i^{p,SC(N_A\times P)}
\end{aligned}
\tag{3}
$$

where SC(n) denote the nth scenario. In GAINS these simulations are performed by reducing each precursor and source aggregation entity independently from each other, i.e. by changing one emission type at a time to identify the corresponding unknown coefficient. This approach has the clear advantage of reducing the number of unknowns drastically and consequently the number of required scenarios (now proportional to N_A). If we consider an application of GAINS with 30 countries and 5 precursors, the "entity to cell" approach would lead to 151 unknowns per receptor cell. It has however the disadvantage of pre-defining and fixing the source aggregation areas. Indeed this formulation does not allow analyzing the response of the SR models on other entities than those initially used to design the SR model. In particular, because the SR model formulation only considers relations between countries and receptor cells, source areas smaller than a country cannot be considered. Consequently, the spatial resolution of the SR model is pre-defined by the fixed entities used in the scenario run. The only way to improve the spatial resolution to account for the impact of specific emission sources (e.g. regions rather than countries) is then to redo a set of scenarios with the AQ model. Improving the spatial resolution becomes therefore rapidly time consuming. For example, down-scaling the approach from countries to regions (i.e. about 300 NUTS2 regions within Europe) while keeping an extended coverage for the modeling domain (e.g. Europe) would lead to a very large number of unknowns to be identified, i.e. a significant number of scenarios to be performed (e.g. 1500 scenarios in the case of 300 aggregation zones and 5 precursors).

2.3. Sliding source aggregation and zonal receptor window

Another approach is to associate specific emission aggregations to each receptor cell. The aggregation entities are then sliding within the domain in such a way that their locations relative to the receptor cells are always the same. Many types of aggregations are possible but we focus here on two approaches, one that is currently used in regional integrated assessment tools, the second that is developed in the frame of this work to optimize the efficiency of the spatial emission aggregations:

- Quadrants (see Fig. 2 middle). In this case, four quadrants are associated to each receptor cell. This reduces drastically the number of unknowns to $N_A = 4$ but at the cost of aggregating in single entities short and long distance emissions. This is the approach followed in RIAT/RIAT+ (Carnevale et al., 2012).
- Multi-ring (see Fig. 2 bottom). This aggregation is similar to the quadrants but uses aggregated entities arranged in several rings (in our example, $N_A = 25$ distributed on 3 rings) to improve the spatial resolution of the emission impacts.

Similarly to the fixed entities, the emissions within each aggregated entities are equally treated, i.e. all emission cells aggregated within an entity contribute equally to the concentration at the receptor cell. Then, the results of the scenario and base case runs are used to write a system of equations similar to (3) for each receptor cell "j":

$$
\begin{aligned}
C_j^{BC} &= C_j^0 + \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^p E_{ij}^{p,BC} \\
C_j^{SC(1)} &= C_j^0 + \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^p E_{ij}^{p,SC(1)} \\
&\vdots \\
C_j^{SC(N_A\times P)} &= C_j^0 + \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^p E_{ij}^{p,SC(N_A\times P)}
\end{aligned}
\tag{4}
$$

where E_{ij}^p are the precursor "p" emissions of the entity "i". Note that, because source entities are sliding, their emissions do not only depend on the source entity "i" but also on the receptor cell "j". The number of source entities for all receptor cells within the domain is equal to N × N_A (for example 50 × 50 cells and 4 quadrants would lead to 10'000 entities). Obviously, the number of sliding entities becomes far too large to proceed as with fixed entities, i.e. generate one specific scenario for each precursor and each entity, independently.

But, system (4) is solved for each receptor cell "j" independently. For one receptor cell "j", the number of unknowns in system (4) is equal to $P \times N_A+1$ (for example with 5 precursors and 4 quadrants the system has 21 unknowns). One base case run and $P \times N_A$ scenario runs based on emission reductions applied over the whole domain provide enough values of concentrations in each cell. These concentrations can then be used to solve the system of $P \times N_A+1$ Equation (4) for each receptor cell.

In systems (3) and (4) the coefficients a_{ij}^p quantify the relation between the source entity emissions and the receptor cell concentration. One difference between the two approaches is however that the coefficient a_{ij}^{p} in system (3) is identified with simulations characterized by emission reductions applied only over the fixed source entities, whereas in system (4) the a_{ij}^{p} are obtained from scenario where emissions are reduced over the whole domain. Consequently, sliding entities are not attached to the entity areas set in the AQ model scenario runs. They provide flexibility in the application of the source–receptor relationships as the area over which emissions is reduced can freely be defined a posteriori.

On the other hand, the accuracy of this approach depends on the capacity of the sliding entities to capture the spatial distributions of the emissions around the receptor cell. Generally, the more entities within a multi-ring structure, the more quadrants considered or the largest the number of fixed entities, the more accurate the results will be, but a larger number of source aggregation entities directly impacts the number of simulations to be performed. A compromise must therefore be found between these two factors according to the level of quality pursued. In this work we will use a 25 elements multi-ring structure. This structure extends sufficiently far around each receptor cell to cover all emission impacts within the modeling domain. Emission impacts from further distance are accounted through boundary conditions in the AQ model simulations.

In summary, the aggregation of source cells allows reducing the number of unknowns needed to characterize the SR model, i.e. the number of source aggregation determines the size of the problems to be solved. The main advantage of a sliding approach to define the source aggregations is to provide more flexibility (a posteriori) in defining the area where emission reductions take place.

With a sliding approach to define the source entities, an additional assumption can be made that reduces the number of required simulations to be performed. Indeed the links between one receptor cell and its related emission sources are characterized by parameters that can be assumed equal over a given geographical zone (i.e. a set of grid cells around the receptor cell) (Fig. 3 bottom). We will call “receptor window” this geographical zone where all coefficients are assumed constant around a receptor cell. With such an assumption the number of required scenarios significantly reduces because the information retrieved from one scenario can now be used to feed different equations in the system related to one specific receptor cell. Considering a zone around the receptor cell “j”, some equations within system (4) can then be substituted as follows:

$$
\begin{aligned}
C_j^{BC} &= C_j^0 + \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^{p} E_{ij}^{p,BC} \\
C_{j-1}^{BC} &= C_j^0 + \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^{p} E_{ij-1}^{p,BC} \\
C_{j+1}^{BC} &= C_j^0 + \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^{p} E_{ij+1}^{p,BC} \\
&\vdots
\end{aligned}
\tag{5}
$$

where the second and third equations have been fed with base-case concentration and emission values corresponding to the cell “j−1” and “j+1” belonging to the zone where all coefficients are assumed equal (i.e. same a_{ij}^{p} and C_j^0). As we see the number of unknowns remains unchanged, i.e. P × N_A+1 but the number of scenario simulations required to identify these unknowns is now reduced to (P × N_A+1)/N_W where N_W is the number of grid-cells belonging to the receptor window.

In the approach proposed in this work, the receptor window will either contain 9 or 25 cells, including the receptor cell (i.e. the receptor cell itself plus the first or second rings of neighboring cells).

In the RIAT/RIAT + model, the zone is large, generally encompassing the entire domain in the case of regional applications. As a result of the assumptions made on the receptor window and the emission quadrants, few simulations are theoretically necessary to solve the system. In practice a larger number of simulations (around 20) are performed to feed a machine learning (neural network) approach which requires many data as input.

In summary, the assumptions made in terms of source aggregation and receptor window will lead to different impacts on the system to resolve. While the number of aggregated sources (together with the number of precursors) fixes the number of unknowns in the system, the size of the receptor window directly determines the number of scenarios required to identify these unknowns.

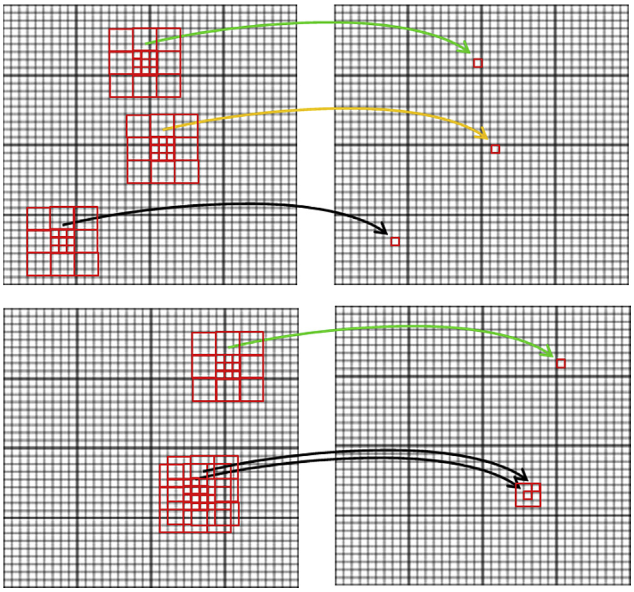

Fig. 3. Receptor windows configurations, with specific links for each single cell (top) and links assumed equal over a given geographical areas (bottom).

2.4. Relative vs. absolute formulation, robustness and co-linearity

In this section we discuss general aspects which apply to the source–receptor relationship model for all possible combination proposed earlier, regardless of the assumptions made in terms of source aggregation and receptor window.

2.4.1. Relative vs. absolute formulation

Systems (3), (4) and/or (5) can be re-formulated in relative terms by subtracting the base-case values from all scenario equations. This leads to a set of equations expressed in terms of delta concentrations $(\Delta C_j^{SC(n)} = C_j^{BC} - C_j^{SC(n)})$ and delta emissions $(\Delta E_{ij}^{SC(n)} = E_{ij}^{BC} - E_{ij}^{SC(n)})$. The system (3) can for example be re-written as:

$$
\begin{aligned}
\Delta C_j^{SC(1)} &= \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^{p} \Delta E_{ij}^{p,SC(1)} \\
&\vdots \\
\Delta C_j^{SC(N_A\times P)} &= \sum_{p=1}^{P}\sum_{i=1}^{N_A} a_{ij}^{p} \Delta E_{ij}^{p,SC(N_A\times P)}
\end{aligned}
\tag{6}
$$

The system expressed in delta values contains one less equation (with respect to the system formulated in absolute values) but also one less unknown (C_j^0). Absolute values can be simply retrieved from: $C_j = C_j^{BC} - \Delta C_j$.

One can either use the absolute approach for the SR relationships solving systems (3) or (5) and retrieve then the delta values or use the relative approach solving system (6) and retrieve then the absolute values. For a single cell receptor window the two approaches are strictly equivalent, whereas this is not the case for a receptor window defined as a zone. Indeed, while the absolute formulation would assume a constant background value (C_j^0) within the zone (see system (5)), nothing can be assumed for C_j^0 in relative terms because the background does not appear in this formulation (see system (6)). The impact of this assumption has been quantified in this work by testing both formulations with different receptor window sizes.

2.4.2. Robustness

In the previous paragraph we have seen that a SR model based on fixed source aggregation associated with a single cell receptor window (e.g. GAINS or TM5-FASST) require $N_A \times N_p+1$ simulations to identify the coefficients. Theses simulations are designed as a series of scenarios in addition to a base case. These scenarios are generally performed by reducing each precursor and source aggregation entity independently from each other. Consequently, system (6) can be written as follow:

$$\begin{aligned}
\Delta C_j^{SC(1)} &= a_{1j}^1 \Delta E_{1j}^{1,SC(1)} \\
\Delta C_j^{SC(2)} &= a_{1j}^2 \Delta E_{1j}^{2,SC(2)} \\
&\vdots \\
\Delta C_j^{SC(P+1)} &= a_{2j}^1 \Delta E_{2j}^{1,SC(P+1)} \\
\Delta C_j^{SC(P+2)} &= a_{2j}^2 \Delta E_{2j}^{2,SC(P+2)} \\
&\vdots \\
\Delta C_j^{SC(N_A \times P)} &= a_{N_A j}^P \Delta E_{N_A j}^{P,SC(N_A \times P)}
\end{aligned} \tag{7}$$

and it becomes trivial to identify the coefficients of the SR relationship:

$$a_{1j}^1 = \Delta C_j^{SC(1)} \Big/ \Delta E_{1j}^{1,SC(1)}; \quad \ldots \; ; a_{N_A j}^P = \Delta C_j^{SC(N_A \times P)} \Big/ \Delta E_{N_A j}^{P,SC(N_A \times P)}$$

But although this set of scenarios is the more straightforward one to solve system (6) it is not the only one. Indeed any set of independent scenarios can be used and will lead to a unique solution for system (6). Independency means that the emission delta from one scenario cannot be expressed as a linear combination of the emission delta of other scenarios. If the relation between the emission deltas and the concentration deltas is fully linear, system (6) will always lead to the same SR coefficients whatever the scenario set chosen. Even though Thunis et al. (2015) showed that the relation between emission deltas and concentration deltas is close to linearity for yearly averaged concentrations, it is never fully linear. Consequently the SR coefficients obtained with different set of scenarios will not be perfectly identical. The variability of the SR coefficients can be quantified by statistical methods which require more scenarios than the exact number requested to solve system (5), leading to an over-defined system (i.e. more equations than unknowns). A regression fit can then be used to find the linear approximation between emissions delta and concentrations delta which will minimize the residual errors (i.e. differences between the concentrations delta provided by the CTM and the concentrations delta obtained with the linear approximation). From the residual errors it is then possible to estimate a statistical distribution and a confidence interval for each SR coefficient. The more linear the relation between emissions and concentration is, the lower the residual errors are, the narrower the SR coefficients confidence intervals are and the more robust the regression estimate is.

As shown earlier an extended receptor window zone reduces the number of requested simulations at the expense of a loss of accuracy. On the other hand, since fewer simulations are necessary, additional simulations can more easily be performed to assess the robustness of the solution. A compromise needs therefore to be found between accuracy and robustness. For a given number of scenarios, a large receptor window zone associated to few source aggregation entities will produce many more equations than unknown coefficients. Consequently, robustness can be high while accuracy might be low. On the contrary, a small receptor window zone associated with a large number of source aggregation entities can lead to a weaker robustness but to a higher accuracy.

In this work the robustness of the results has been estimated using a Monte-carlo analysis which generated a set of 1000 SR coefficients varying within their respective 95% confidence intervals and assuming uniform distributions. Each set of coefficients has been used to calculate 1000 values of concentration deltas.

2.4.3. Co-linearity

As seen earlier (Equation (5)) the concentration delta at one receptor cell j is expressed as a linear combination of the emission deltas from all precursors p in all aggregation sources i. In this linear relation, the coefficients a_{ij}^p result from a linear fitting involving a large number of input (i.e. ΔC_j and ΔE_{ij}^p from different scenarios and cells inside a receptor window, see systems (5) and (6)). As a result of this fitting process, each coefficient a_{ij}^p quantifies directly the impact of ΔE_{ij}^p (the delta emissions of the aggregation source i and the precursor p) on ΔC_j(the concentration delta in cell j). A high coefficient value stems from a high correlation between ΔE_{ij}^p and ΔC_j indicating a strong impact of ΔE_{ij}^p on ΔC_j in the CTM simulation. One of the issues is that different source aggregations associated to a receptor cell are generally not independent from each other (e.g. ΔE_{ij}^p is correlated to ΔE_{i+1j}^p). If ΔC_j are strongly impacted by a first source aggregation (AgS1) and weakly affected by a second one (AgS2) in the CTM simulation, the expected values should be high for a_{ij}^p (AgS1) and low for a_{ij}^p (AgS2). However, the strong correlation between AgS1 and AgS2 indirectly generates a high correlation between AgS2 and the ΔC_j (because ΔC_j and AgS1 are strongly correlated). In this particular case, the a_{ij}^p coefficients for AgS1 and AgS2 will both be high, resulting in a wrong representation of the CTM simulation. The dependency between AgS1 and AgS2 is related to the way the emission scenarios are sampled. Indeed another scenario set might lead to totally different relations between AgS1 and AgS2 indicating a low robustness.

The tests performed in this study have shown that co-linearity decreases while robustness increases with the number of scenario used for the regression fit.

Another way of reducing co-linearity is to apply a Principal Component Regression (PCR). In a first step a Principal Component Analysis (PCA, Eder et al., 2014) is applied to the input data (i.e. the delta emissions mentioned in system (6)). The Principal Component Analysis is a procedure that converts a set of possibly correlated variables into a set of linearly uncorrelated ones. This leads to a new set of input data, called Principal Components, expressed as a linear combination of the original ones. The next step is to apply the PCR (Rajab et al., 2013), implementing a regression analysis technique based not on the initial explanatory variables, but on a subset (i.e. explaining 95% of the total variance) of the principal components resulting from the PCA application. This two-steps approach reduces the co-linearity and improves the robustness of the results.

3. Practical application

In this section the different options presented above to design a

Table 1
Set of emission reduction scenarios, used for the CTM simulations, and needed to prepare surrogate models. Scenario 1 represents the base case, while the other scenarios are computed applying the % emission reductions as shown in Table (see text of the paper for more details).

	NO_x	VOC	NH_3	PPM	SO_2
1					
2	−37%	−33%	−28%	−28%	−7%
3	−66%	−60%	−50%	−51%	−14%
4	−66%	−33%	−28%	−28%	−7%
5	−37%	−60%	−28%	−28%	−7%
6	−37%	−33%	−50%	−28%	−7%
7	−37%	−33%	−28%	−51%	−7%
8	−37%	−33%	−28%	−28%	−14%
9	−66%	−60%	−28%	−28%	−7%
10	−66%	−33%	−50%	−51%	−14%
11	−66%	−33%	−50%	−28%	−8%
12	−66%	−33%	−50%	−28%	−14%

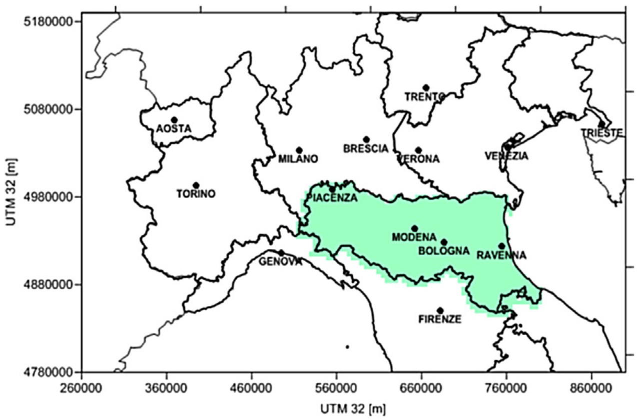

Fig. 4. Emilia Romagna regional domain, located in Northern Italy.

SR model are tested on a practical application performed over a regional scale modeling domain. These tests mainly focus on the different assumptions possible in terms of source aggregation and receptor window but the impact of formulating the problem in relative or absolute terms will also be addressed. At first, a brief description of the modeling set-up is provided.

3.1. Case set-up

The Chemical Transport Model CHIMERE has been applied by ARPA-EMR to perform simulations in the frame of the LIFE + OPERA project (see www.operatool.eu for more details). These simulations, available over northern Italy for the period October–March to reconstruct PM concentrations, are then used in this study. In addition to the base case, 12 emission scenarios have been designed in the frame of the OPERA project (Table 1) to cover the potential range of reduction levels. The minimum reduction level corresponds to the application of Current Legislation (2010 CLE) while the maximum level is obtained after application of all feasible emission abatement measures (2020 MFR). The emission reductions are only applied within the Emilia Romagna region (Fig. 4) while emission reductions outside this region are kept to the CLE (2020) level. The selected combinations of precursor emission reduction (i.e. for each scenario) are obtained through application of a factor analysis approach (Carnevale et al., 2010). SR models are developed for winter averaged PM10 concentrations as a function of the 5 emission precursors previously mentioned. The approach remains valid and could be applied to other pollutants as long as the relations between sources and concentrations can be assumed linear. This is the case for example of NO_2 and O_3 which were shown to behave linearly for relatively long time averages (Thunis et al., 2015).

Model simulations are necessary for two purposes. A first set of simulations is required to derive the SR model (i.e. identify the unknown coefficients) while a second set serves for validation (comparison between the Chemical Transport Model and the SR model). The first and second set should be as distinct as possible to guarantee a proper validation. In this practical application, 8 scenarios (referred as training) are used to derive the SR model whereas 4 are used for validation. As the quality of the results is similar for all 4 validation scenarios only one is used in the following sections.

3.2. SR formulation: absolute vs. relative

The aim of this section is to analyze the impact of using either a relative (relating emissions delta to concentrations delta) or an absolute (relating emissions to concentrations) formulation for the training and validation phases.

Figs. 5 and 6 clearly show that better performances are obtained with a relative approach. As mentioned the absolute and relative formulations would lead to strictly equivalent results if the receptor window is restricted to one cell. When the receptor window is larger, however, results could differ. This results from the fact that the background concentration C_j^0 is assumed to be similar in all grid cells belonging to the zone in the absolute formulation while this is not the case in relative terms. The difference between the two formulations will tend to decrease when the receptor window zone size decrease and the two formulations are strictly equivalent for a

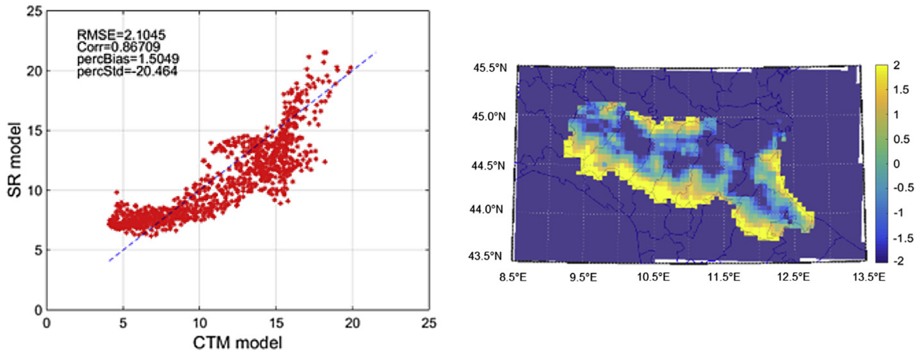

Fig. 5. Comparison between the CTM and S/R model, for scenarios 10, left: scatter plot, right map of absolute differences. Configuration of S/R model: Training on "absolute" values, R-window: "all domain", S-aggregation:"quadrants".

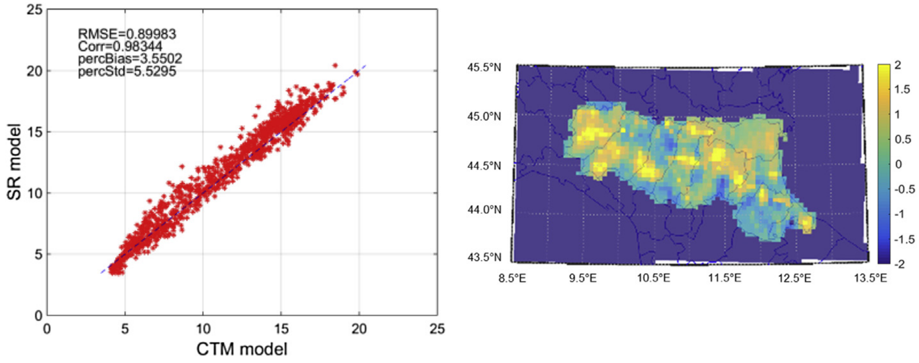

Fig. 6. Comparison between the CTM and S/R model, for scenarios 10, left: scatter plot, right map of absolute differences. Configuration of S/R model: Training on "delta" values, R-window: "all domain", S-aggregation:"quadrants".

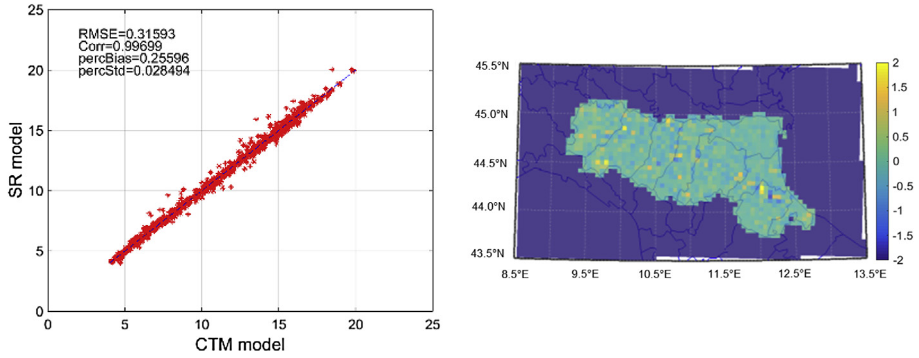

Fig. 7. Same as Fig. 6 but the S/R model uses a 5 × 5 R-window.

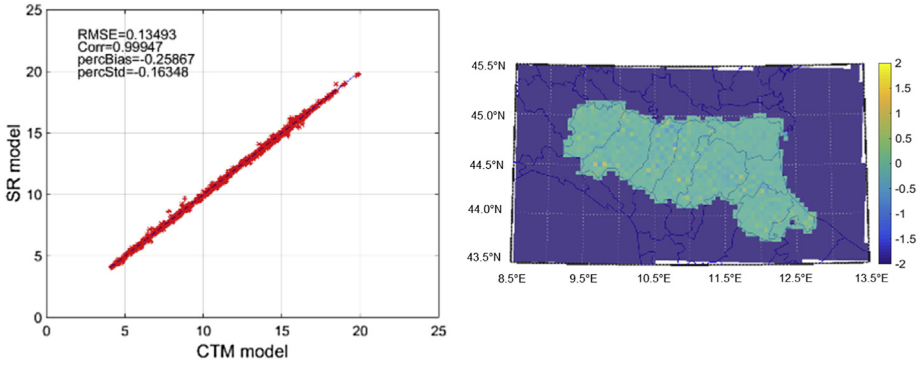

Fig. 8. Same as Fig. 6 but the S/R model uses a 3 × 3 R-window.

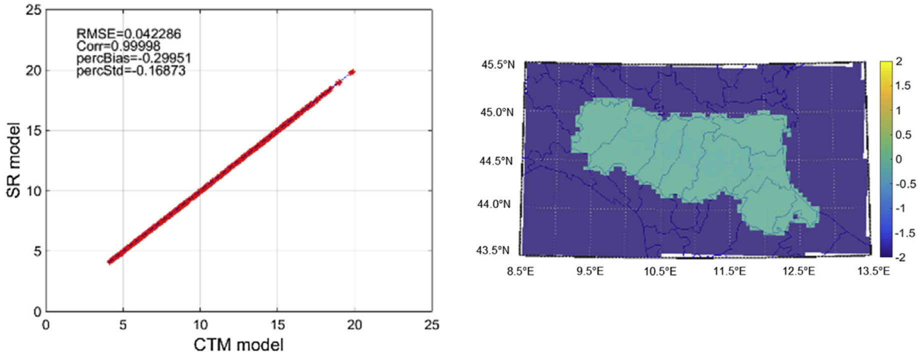

Fig. 9. Same as Fig. 7 but the S/R model uses a 5 × 5 R-window and "multi-ring" S-aggregation.

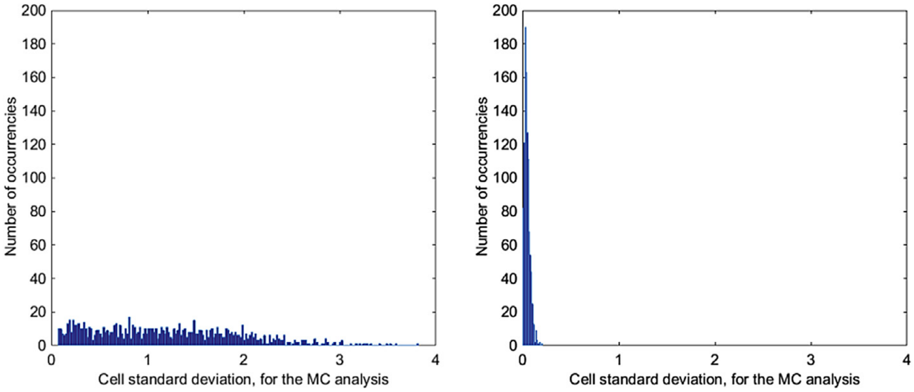

Fig. 10. Computing the standard deviation of the Monte Carlo simulations performed, for the standard regression (left) and Principal Component Regression (right).

one cell receptor window. The relative value approach is kept as reference in the following comparisons.

3.3. *Receptor window size and source aggregation number and shape*

First, the impact of the assumption made on the receptor window on the accuracy of the results is assessed by comparing results obtained with different receptor zone configurations: (1) a zone of 3 × 3 cells, (2) a zone of 5 × 5 cells and (3) a zone covering the entire domain. As clearly shown by Figs. 6–8, results improve with smaller size of receptor zones.

Regarding source aggregation, the impact of the different assumptions (number and shape) is investigated by comparing the quadrants and the multi-ring approaches. The multi-ring approach clearly leads to better results than the quadrants (Fig. 9), as a result of its better spatial resolution and of allowing for differentiating short and long distance emission changes for each receptor cell.

In summary smaller window zone associated to well resolved aggregation sources lead to more accurate results. In the next section we will investigate whether this SR configuration also lead to robust results.

3.4. *Robustness and uncertainty analysis*

The Principal Component Regression (PCR) approach has been used to reduce co-linearity among the input data (combinations of precursors and emission aggregations) to improve the robustness of the regression results.

The robustness of the results is tested by using a Monte-Carlo analysis to generate 1000 different set of regression coefficients. These 1000 set of coefficients are used to calculate 1000 values of concentration deltas which are distributed around the results obtained with the initial regression. For each cell the standard deviation of the results (over the 1000 simulations) has been computed, and shown in Fig. 10. The results obtained with a standard linear regression method are shown on Fig. 10 (left). The standard deviation between the simulations reached also values up to 4 μg/m3 showing a high degree of variability and sensitivity even to small values of the coefficients (as estimated in the Monte-Carlo analysis). Fig. 10 (right) shows the results obtained when the PCR approach is used. The standard deviation is considerably reduced, showing a much smaller uncertainty on the output.

A comparison of the best results obtained with the standard regression and PCR – Principal Component Regression (Figs. 9–11) shows that a PCR implementation provides slightly less accurate (but more robust) results than standard regression. In the proposed methodology the PCR is finally used, as it represents a better compromise between accuracy and robustness than a standard regression.

4. Conclusion

Source–receptor relationships are simplified models that are constructed to mimic the behavior of full air quality models. As a result of their simplicity, they run much faster. While an air quality model simulation will probably require hours or days of CPU for a full year run, depending on domain size and available computational power, the SR model will only requires minutes or seconds. The SR model can therefore be used in real-time in integrated

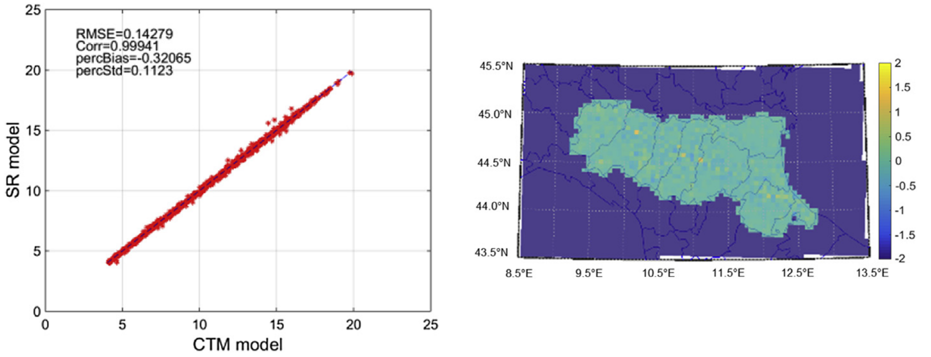

Fig. 11. Same as Fig. 9 but the S/R model applies the Principal Component Regression instead of the simple Regression.

assessment models to relate the impacts of emission changes on concentration levels. Different approaches based on different levels of assumptions exist to design these SR models. In this work the assumptions made in terms of emission aggregation and receptor window are differentiated. These assumptions indeed have important consequences on the resulting size of the problem to be solved (i.e. number of equations and unknowns) but also on the number of simulations to be performed to solve this system.

While the aggregation of emissions has been shown to reduce the size of the system, the size of the receptor window directly influences the number of requested scenarios. Two main options currently used to aggregate emissions have been analyzed, i.e. fixed and sliding. The approach proposed in this work follows the sliding aggregation option as it has the main advantage of allowing flexibility in the definition of the source aggregation a-posteriori. This is seen as a key advantage as we want to use the SR model for emission changes imposed on any given domain. Regarding the receptor side, results clearly show that the smallest the receptor window is, the better the results are. In the proposed approach a receptor window zone of 9 cells is selected as this preserves a good accuracy while in the same time reduces the number of requested simulations by a factor of 9.

This SR model has been tested either in relative (delta concentration vs. emission delta) or absolute forms. The relative formulation has been shown to perform much better in terms of accuracy than the absolute one and has therefore been adopted in the proposed approach. From the point of view of robustness, it has been shown that, to manage co-linearity among input, a viable option is to apply the Principal Component Regression approach. This allows for a robust solution, even if with a slight decrease in terms of accuracy.

The practical application of the proposed approach based on a multi-ring to zone SR model, on a regional domain, has shown good results. As future work, additional emission aggregation types will be explored; and at the same time, non-linear parts will be added to the regression model in case of need (as in the case of an extension of this work to other pollutants, or other time aggregations).

Acknowledgments

The authors would like to acknowledge Marialuisa Volta and Claudio Carnevale (University of Brescia) and Giorgio Guariso (Politecnico of Milano), for interesting exchanges on the issue related to accuracy and uncertainty, and on initial assumptions related to emission aggregations; also, we acknowledge Eriberti de Munari, Marco Deserti and Michele Stortini, from Arpa Emilia Romagna and the whole OPERA project team (http://www.operatool.eu) for sharing the simulations used in this paper.

References

Amann, M., Bertok, I., Borken-Kleefeld, J., Cofala, J., Heyes, C., Höglund-Isaksson, L., Klimont, Z., Nguyen, B., Posch, M., Rafaj, P., Sandler, R., Schöpp, W., Wagner, F., Winiwarter, W., 2011. Cost-effective control of air quality and greenhouse gases in Europe: modeling and policy applications. Environ. Model. Softw. 26, 1489–1501.

Carnevale, C., Pisoni, E., Volta, M., 2010. A non-linear analysis to detect the origin of PM10 concentrations in Northern Italy. Sci. Total Environ. 409, 182–191.

Carnevale, C., Finzi, G., Pisoni, E., Volta, M., Guariso, G., Gianfreda, R., Maffeis, G., Thunis, P., White, L., Triacchini, G., 2012. An integrated assessment tool to define effective air quality policies at regional scale. Environ. Model. Softw. 38, 306–315.

D'Elia, M., Bencardino, L., Ciancarella, M., Contaldi, G., Vialetto, 2009. Technical and non-technical measures for air pollution emission reduction: the integrated assessment of the regional air quality management plans through the Italian national model. Atmos. Environ. 43, 6182–6189.

Dunker, A.M., 1981. Efficient calculation of sensitivity coefficients for complex atmospheric models. Atmos. Environ. 15, 1155.

Dunker, A.M., Yarwood, G., Ortmann, J.P., Wilson, G.M., 2002. Comparison of source apportionment and source sensitivity of ozone in a three-dimensional air quality model. Environ. Sci. Technol. 36, 2953–2964.

Eder, B., Bash, J., Foley, K., Pleim, J., 2014. Incorporating principal component analysis into air quality model evaluation. Atmos. Environ. 82, 307–315.

Hakami, A., Seinfeld, J.H., Chai, T.F., Tang, Y.H., Carmichael, G., Sandu, A., 2006. Adjoint sensitivity analysis of ozone non-attainment over the continental United States. Environ. Sci. Technol. 40, 3855–3864.

Mircea, M., Ciancarella, L., Briganti, G., Calori, G., Cappelletti, A., Cionni, I., Costa, M., Cremona, G., D'Isidoro, M., Finardi, S., Pace, G., Piersanti, A., Righini, G., Silibello, C., Vitali, L., Zanini, G., 2014. Assessment of the AMS-MINNI system capabilities to simulate air quality over Italy for the calendar year 2005. Atmos. Environ. 84, 178–188.

Rajab, J.M., MatJafri, M.Z., Lim, H.S., 2013. Combining multiple regression and principal component analysis for accurate predictions for column ozone in Peninsular Malaysia. Atmos. Environ. 71, 36–43.

Sandhu, A., Daescu, D.N., Carmichael, G.R., Chai, T.F., 2005. Adjoint sensitivity analysis of regional air quality models. J. Comput. Phys. 204 (1), 222–252.

Seibert, P., Frank, A., 2004. Source – receptor matrix calculation with a Lagrangian particle dispersion model in backward mode. Atmos. Chem. Phys. 4, 51–63.

Simpson, D., Olendrzynski, K., Semb, A., Støren, E., Unger, S., 1997. Photochemical Oxidant Modelling in Europe: Multi-annual Modelling and Source-receptor Relationships. EMEP/MSCW Report 3/97. EMEP MSC-W. Norwegian Meteorological Institute, PO Box 43-Blindern, N - 0313 Oslo 3, Norway.

Tarrasón, L., Fagerli, H., Jonson, J.E., Klein, H., van Loon, M., Simpson, D., Tsyro, S., Vestreng, V., Wind, P., 2004. Transboundary Acidification, Eutrophication and Ground Level Ozone in Europe. EMEP Status Report 2004. ISSN: 0806-4520. EMEP/MSC-W – Det Norske Meteorologiske Institutt, Oslo, Norway.

Thunis, P., Clappier, A., 2014. Indicators to support the dynamic evaluation of air quality models. Atmos. Environ. 98, 402–409.

Thunis, P., Clappier, A., Pisoni, E., Degraeuwe, B., 2015. Quantification of non-linearities as a function of time averaging in regional air quality modeling applications. Atmos. Environ. 103, 263–275.

Vedrenne, M., Borge, R., Lumbreras, J., Rodríguez, M.E., 2014. Advancements in the design and validation of an air pollution integrated assessment model for Spain. Environ. Model. Softw. 57, 177–191.

Wagstrom, K.M., Pandis, S.N., Yarwood, G., Wilson, G.M., Morris, R., 2008. Development and application of a computationally efficient particulate matter apportionment algorithm in a three-dimensional chemical transport model. Atmos. Environ. 42, 5650–5659.

Ying, Q., Kleeman, M.J., 2006. Source contributions to the regional distribution of secondary particulate matter in California. Atmos. Environ. 40, 736–752.