



Estimation of volatile organic compound emissions for Europe using data assimilation

M. R. Koohkan^{1,2}, M. Bocquet^{1,2}, Y. Roustan¹, Y. Kim¹, and C. Seigneur¹

¹Université Paris-Est, CEREa, joint laboratory École des Ponts ParisTech and EDF R&D, Champs-sur-Marne, France

²INRIA, Paris Rocquencourt research centre, France

Correspondence to: M. Bocquet (bocquet@cerea.enpc.fr)

Received: 6 October 2012 – Published in Atmos. Chem. Phys. Discuss.: 20 December 2012

Revised: 17 April 2013 – Accepted: 20 May 2013 – Published: 18 June 2013

Abstract. The emissions of non-methane volatile organic compounds (VOCs) over western Europe for the year 2005 are estimated via inverse modelling by assimilation of in situ observations of concentration and then subsequently compared to a standard emission inventory. The study focuses on 15 VOC species: five aromatics, six alkanes, two alkenes, one alkyne and one biogenic diene. The inversion relies on a validated fast adjoint of the chemical transport model used to simulate the fate and transport of these VOCs. The assimilated ground-based measurements over Europe are provided by the European Monitoring and Evaluation Programme (EMEP) network. The background emission errors and the prior observational errors are estimated by maximum-likelihood approaches. The positivity assumption on the VOC emission fluxes is pivotal for a successful inversion, and this maximum-likelihood approach consistently accounts for the positivity of the fluxes. For most species, the retrieved emissions lead to a significant reduction of the bias, which underlines the misfit between the standard inventories and the observed concentrations. The results are validated through a forecast test and a cross-validation test. An estimation of the posterior uncertainty is also provided. It is shown that the statistically consistent non-Gaussian approach based on a reliable estimation of the errors offers the best performance. The efficiency in correcting the inventory depends on the lifetime of the VOCs and the accuracy of the boundary conditions. In particular, it is shown that the use of in situ observations using a sparse monitoring network to estimate emissions of isoprene is inadequate because its short chemical lifetime significantly limits the spatial radius of influence of the monitoring data. For species with a longer lifetime (a few days), successful, albeit partial, emission cor-

rections can reach regions hundreds of kilometres away from the stations. Domain-wide corrections of the emission inventories of some VOCs are significant, with underestimations of the order of a factor of 2 for propane, ethane, ethylene and acetylene.

1 Introduction

Non-methane volatile organic compounds (NMVOCs: further abbreviated VOCs in the following) are of particular environmental concern because they are precursors of secondary pollutants, such as ozone and fine particulate matter (PM_{2.5}), and some VOCs are pollutants in their own right due to their adverse carcinogenic and/or non-carcinogenic health effects. Therefore, it is essential to have accurate emission inventories of VOCs to conduct air quality modelling studies for the development of optimal emission control strategies and for air quality forecasting as well as to follow their temporal emission trends over the years as emission control strategies become implemented. The large number of emission sources, both anthropogenic and biogenic, and processes leading to those emissions (combustion, evaporation and vegetation metabolism) make the development of accurate VOC emission inventories difficult. Furthermore, VOC emissions cannot be derived from mass balances of the emission process (as can be done for example for sulfur or heavy metal emissions) and they must be obtained from measurements conducted at the source of the emissions. Also, emission testing is costly and some emission factors are developed for total VOCs rather than for individual VOCs. Therefore, a chemical speciation must be applied to total

VOC emission factors using speciation data that are limited and uncertain. Uncertainties in anthropogenic emissions have been reduced over the years as a result of better characterisation of major emission sources; for example, tunnel VOC measurements conducted in California have shown underestimates in VOC emissions from vehicles decreasing from a factor of 4 in 1987 (Pierson et al., 1990) to a factor of 2.5 in 1994 (Kirchstetter et al., 1996). Nevertheless, uncertainties still remain (Sawyer et al., 2000). Furthermore, uncertainties of the order of a factor of 2 are associated with isoprene emissions, and greater for other biogenic organic compounds (Guenther et al., 2000, 2006; Warneke et al., 2010), due to the difficulty of estimating the meteorology-dependent emission rates for a large number of vegetation species as well as characterising the land use/land cover of the area of interest. Therefore, several approaches have been used to evaluate the accuracy of VOC emission inventories and, if appropriate, apply some correction.

Uncertainties in emission inventories have been estimated, for example, by comparing ambient air measurements in tunnels with vehicle exhaust emission estimates (e.g. Staehelin et al., 1998; Sawyer et al., 2000; Touaty and Bonsang, 2000; Stemmler et al., 2005; Ho et al., 2007). However, such experiments characterise only one source category (on-road traffic) and focus on a single location and time period. Satellite measurements in combination with various types of models have been used to assess VOC emissions. However, such techniques are limited by the number of VOCs that can be measured via satellite-borne instruments (Vijayaraghavan et al., 2008) and to areas that are specific to a major source category, e.g. the use of formaldehyde (an oxidation product of isoprene) to estimate isoprene emissions in remote areas where biogenic emissions dominate (e.g. Shim et al., 2005; Fu et al., 2007; Millet et al., 2008; Dufour et al., 2009). Measurements of VOC concentrations aloft have also been used to estimate fluxes of VOCs originating from an area (e.g. Hopkins et al., 2009); however, there are uncertainties associated with the mass balance method used to estimate the atmospheric transport flux and relate it to an emission flux. Furthermore, such an approach is limited to the estimation of an emission flux for a given region over a given period. Comparisons of the output concentrations of air quality model simulations with observations aloft (e.g. Xiao et al., 2008) or at ground level (e.g. Harley and Cass, 1995) provide some estimates of uncertainties in VOC emissions; however, such information is typically also limited to a specific region and period for which those measurements are available. Inverse modelling has been conducted using ground-level ambient concentrations to estimate emission inventories for some air pollutants, but such studies (e.g. Quélo et al., 2005; Elbern et al., 2007; Koohkan and Bocquet, 2012) have focused so far on regulated air pollutants with ambient concentration data available from routine monitoring networks and have not yet included a large number of speciated VOC ground-level concentration measurements. Use of glyoxal satellite data (e.g.

Liu et al., 2013) and a combination of formaldehyde and glyoxal satellite data (e.g. Stavrou et al., 2009) have also been used to estimate emission fluxes of isoprene and other biogenic and anthropogenic VOC precursors of those oxidised VOC.

It is, therefore, of great interest to investigate the current status of VOC emissions using an approach that provides information for several major VOCs with spatial distribution over a large domain and for a long period of time. To that end, we present here the first assessment of the emissions of several VOCs measured routinely at several remote sites that covers all VOC sources over western and central Europe for an entire year.

In Sect. 2, the chemical transport model (CTM) used for modelling the VOCs is introduced and its reduced counterpart is described. The source–receptor relationship is built using an approximate adjoint model, which is validated. The control variables (i.e. the emission fluxes), and the inversion modelling method are described. The inversion method crucially depends on the prior statistics of errors in the system, and a method to estimate the so-called *hyperparameters* that parameterise these statistics is introduced. In Sect. 3, the setup of the numerical experiments is described. Details about the observations and the background inventory are provided. The optimal values of the hyperparameters to be used in the inversions are computed. In Sect. 4, the results of the inversions are presented and discussed. A forecast test and a cross-validation test are provided to validate the corrected emission fluxes using independent observations. Some objective results on the uncertainty attached to the emission corrections are provided. Conclusions are presented in Sect. 5.

2 Methodology

2.1 Full chemical transport model and reduced VOC model

The chemical transport model (CTM) POLAIR3D (Sartelet et al., 2007) of the POLYPHEMUS air quality modelling system (Mallet et al., 2007) is chosen to model the atmospheric concentrations of chemical species. The numerical discretisation of the model for chemistry and transport, based on a first-order time-splitting algorithm, can be summarised as follows:

$$c^{k+1} = \mathcal{X}_k \left(M_k(c^k) \right) + \Delta t e_{k+1}, \quad (1)$$

where

- c^k is the field of the concentrations of all simulated species at time step k ,
- M_k is the linear advection-diffusion operator. It also includes the deposition processes,
- \mathcal{X}_k is the chemical reaction operator,

– e_k is the emission field at time step k .

Table 1 lists the 15 VOC species for which measurements are available from the European Monitoring and Evaluation Programme (EMEP) database. In order to simulate the concentrations of these species, the RACM 2 (Regional Atmospheric Chemistry Mechanism, version 2) chemical kinetic mechanism (Goliff et al., 2013) is used within the CTM. The model has been satisfactorily evaluated with a similar configuration by Kim et al. (2009). The chemical reactions considered for these species and their typical lifetime are also presented in Table 1. After undergoing oxidation, these 15 primary species result in secondary species. The latter are not presented in Table 1 because they are not measured in the EMEP network and therefore cannot be assimilated; they are, however, included in RACM 2 either explicitly or via surrogate species. Some of the primary VOC species; isoprene (ISO), acetylene (ACE), ethane (C_2H_6), ethylene (C_2H_4) and benzene (BEN) are treated explicitly in RACM 2. The others are represented through a lumped molecule approach and therefore need a specific treatment to be followed separately. They are added as explicit species with their own oxidation reactions written in a way that does not affect RACM 2. The emissions of the explicit species continue to be treated as part of the lumped species of RACM 2 and the oxidant species involved in the additional oxidation reactions of the explicit species (see Table 1) are included as a product of the oxidation reaction so that their concentrations are not affected by those additional reactions. This approach has been used previously and has been shown not to affect the concentrations of the original mechanism, here RACM 2 (e.g. Pun et al., 2003; Kim et al., 2011).

This chemical mechanism involves more than 300 reactions that result in nonlinear interactions among the chemical species. As a result, the computational burden of inverse modelling studies is very large. To address this issue, we developed a reduced chemical mechanism, denoted X_k , which uses the concentration fields of the oxidants, hydroxyl radicals (OH), ozone (O_3) and nitrate radicals (NO_3), provided as external data. The oxidant concentration fields are pre-computed with RACM 2 and used later in the reduced mechanism. This approximation makes sense if $\delta\mathcal{X}_k(c^k) = \mathcal{X}_k(c^k) - X_k(c^k)$ is small with respect to $X_k(c^k)$. The validation of this approximations is checked a posteriori in Sect. 4.1.1. When replacing \mathcal{X}_k by X_k , Eq. (1) becomes linear with respect to the emission fields and the computational cost of inverse modelling becomes manageable.

2.2 The source–receptor model

The source–receptor model provides the relationship between the emissions and the observations. For species s , this can be written as follows:

$$\mu^s = \mathcal{H}^s e^s + \lambda^s + \epsilon^s, \quad (2)$$

Table 1. The volatile organic compounds, their (indicative) lifetime and reactions.

	species	symbol	lifetime (days)	reactions
	isoprene	ISO	0.07	ISO + OH → ISO + O ₃ → ISO + NO ₃ →
RACM 2 reactions	acetylene	ACE	110	ACE + OH →
	ethane	C ₂ H ₆	60	C ₂ H ₆ + OH →
	ethylene	C ₂ H ₄	1.45	C ₂ H ₄ + OH → C ₂ H ₄ + O ₃ → C ₂ H ₄ + NO ₃ →
	benzene	BEN	11	BEN + OH →
additional reactions to RACM 2	propane	C ₃ H ₈	14	C ₃ H ₈ + OH →
	n-butane	NBUT	7	NBUT + OH →
	isobutane	IBUT	7.5	IBUT + OH →
	n-pentane	NPEN	5	NPEN + OH →
	isopentane	IPEN	4	IPEN + OH →
	propene	C ₃ H ₆	0.625	C ₃ H ₆ + OH → C ₃ H ₆ + O ₃ → C ₃ H ₆ + NO ₃ →
	toluene	TOLU	2.4	TOLU + OH →
o-xylene	OXYL	1.1	OXYL + OH →	
m-xylene*	MXYL	0.625	MXYL + OH →	
p-xylene*	PXYL	1.05	PXYL + OH →	

* Measured jointly in the EMEP monitoring network and represented with the symbol MPXYL.

where $\mu^s \in \mathbb{R}^{d_s}$ represents the vector of the observations (d_s is the number of observations for species s). \mathcal{H}^s is the Jacobian operator with respect to e^s , and $e^s = (e_0^s, e_1^s, \dots, e_k^s, \dots, e_{N_t}^s) \in \mathbb{R}^E$ defines the hourly and spatially discretised emission vector, where $E = N_t \times N_x \times N_y \times N_z$. N_t is the total number of time steps, and N_x , N_y and N_z are the total number of elements (grid cells) in the x , y and z directions. $\lambda^s \in \mathbb{R}^{d_s}$ is the vector of the concentrations induced by the initial and the boundary conditions for species s . If μ^{s,i_k} is the observation of species s at time t_k at station i , then $\lambda_{i_k}^s$ is the concentration at the same time and location, computed with the full CTM – i.e. Eq. (1), with $e^s = \mathbf{0}$. The vector ϵ^s represents the errors: representativeness error, model error and instrumental error of the observations.

The Jacobian operator \mathcal{H}^s can be built following two different methods. The first method consists in using the CTM as it is. Let us assume that $e^s \equiv \delta_{l,h,k'}$ is the unity source at the surface coordinate $l \in [1; N_x \times N_y]$, altitude $h \in [1; N_z]$ and time $k' \in [1; N_t]$, and equals 0 anywhere else. Assuming clean-air boundary and initial conditions, and this source term, the CTM-simulated concentrations at time t_k and station i are stored in $\mathcal{H}_{l,h}^{s,i,k'}$. In order to compute the \mathcal{H}^s operator with this method, the CTM model needs to be run E times, which is computationally intensive. That is why this method is often restricted to point-wise emission sources.

The second method consists in using the adjoint model of the CTM (Marchuk, 1974; Uliasz, 1983, and many more references since the late 90's). For a monitoring site i at time t_k , using the linearity of the CTM for the VOCs, the *adjoint solution* can be written as follows:

$$\phi_i^k = \mathbf{M}_k^\dagger \left(\mathbf{X}_k^\dagger (\phi_i^{k+1}) \right) + \Delta t \pi_i^k, \quad (3)$$

where $\pi_i^k = \delta_{i,k}$ is the *sampling function* related to the concentration measurement at station i and time t_k , \mathbf{X}_k^\dagger is the adjoint of \mathbf{X}_k and \mathbf{M}_k^\dagger is the adjoint of \mathbf{M}_k . At the final time N_t , $\phi_i^{N_t}$ is chosen to be $\mathbf{0}$. The adjoint model is also computed for clean air boundary conditions. Then, the Jacobian matrix is given by $\mathcal{H}_{l,h}^{s,i,k'} = [\phi_{i,s}^{k'}]_{l,h}$. The adjoint model is run d times ($d \leq \sum_s d_s$). This method is of great interest to reduce the computational time when $d \ll E$. It is appropriate for estimating the emissions originating from spatially distributed sources. Because the many adjoint solutions are independent, it is straightforward to parallelise their computation.

The Jacobian matrix of the present study is computed row by row – i.e. using adjoint solutions.

2.3 Control space

In order to reduce the dimension of the control space, which is the space of the fluxes to be estimated via inverse modelling, we introduce a relation between the effective control variables α^s and the emission e^s :

$$[e_k^s]_{l,h} = [\alpha^s]_l [e_k^{s,b}]_{l,h}. \quad (4)$$

In this equation, $e^{s,b}$ is the background vector of emission for species s . Obviously the background value of the scaling factors α^s is $\alpha_b^s = \mathbf{1}$, where $\mathbf{1} = (1, \dots, 1)^T$ and T denotes transposition of a vector or of a matrix. Indices k , l and h are respectively related to the time sequence, the horizontal space grid and the vertical grid. The α^s factors, rather than the full 3D time-dependent emission fields of Eq. (1), will be optimised. This choice of control parameters, which is only a function of l , implies that the correction of the emission fluxes is spatially distributed in the horizontal directions but that the vertical and temporal distributions of the emission fluxes are not modified by the data assimilation analysis. Indeed, the relatively low number of observations available restricts the number of parameters that can be considered for the inversion. The fact that α^s does not depend on time means that the inversion relies on the a priori temporal distribution of the emissions (given by activity sectors), which is not modified.

2.4 Inversion method

Combining Eq. (2) with Eq. (4), one obtains

$$\mu^s = \mathbf{H}^s \alpha^s + \lambda^s + \epsilon^s, \quad (5)$$

where \mathbf{H}^s is the Jacobian matrix that relates μ^s to α^s :

$$\mathbf{H}_{k,l}^s = \sum_k [\phi_{s,i}^k]_l [e_k^{s,b}]_l. \quad (6)$$

In order to optimise the α parameters, the following objective function with a regularisation term is used:

$$\mathcal{L}_s(\alpha^s) = \frac{1}{2} (\mu^s - \mathbf{H}^s \alpha^s - \lambda^s)^T \mathbf{R}_s^{-1} (\mu^s - \mathbf{H}^s \alpha^s - \lambda^s) + \frac{1}{2} (\alpha^s - \mathbf{1})^T \mathbf{B}_s^{-1} (\alpha^s - \mathbf{1}). \quad (7)$$

The vector $\alpha_b^s = \mathbf{1} = (1, \dots, 1)^T$ is the first guess (mean state of the background) of α^s . \mathbf{R}_s is the observation error covariance matrix. \mathbf{B}_s is the background error covariance matrix. For each species, these two matrices are both chosen to be diagonal with uniform variances; that is, $\mathbf{B}_s = m_s^2 \mathbf{I}_{N_x \times N_y}$, $\mathbf{R}_s = r_s^2 \mathbf{I}_{d_s}$. Here, $\mathbf{I}_{N_x \times N_y}$ is the identity matrix of control space and \mathbf{I}_{d_s} is the identity matrix of the observation space. Error standard deviations r_s and m_s respectively give the magnitude of the prior uncertainty on the observation and the background. They will be both discussed further in the next section.

These statistical assumptions imply that we neglect any spatial and temporal correlations between grid cells in the errors. Because of the way the anthropogenic inventory surveys are performed, the anthropogenic emissions of VOCs are not in principle expected to induce long-range correlation (a few tens of kilometres) in the errors. Yet, scaling relationships used to infer emissions of pollutants from others in bottom-up approaches may have biases that would be systematic over large geo-political regions. Moreover, potential important reasons for this diagonal hypothesis to fail are when the biogenic VOC emissions have correlated errors due for some VOCs to common sources and similar emission model formulations, or when transport model errors induce temporal correlation in the error covariance matrix. That is why our diagonal assumption should be understood as a convenient approximation.

In the following, two types of solution for Eq. (7) are considered. The first one assumes that the errors are Gaussian distributed and the analysed parameters α_a^s are given by the best linear unbiased estimator (BLUE):

$$\alpha_a^s = \mathbf{1} + \mathbf{K}^s (\mu^s - \mathbf{H}^s \mathbf{1} - \lambda^s), \quad (8)$$

where \mathbf{K}^s is the gain matrix:

$$\mathbf{K}^s = \mathbf{B}_s \mathbf{H}^{sT} \left(\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{sT} \right)^{-1}. \quad (9)$$

The second solution is obtained by minimising Eq. (7) on α^s , under the positivity constraint of each one of its entry $[\alpha^s]_l$. A bounded quasi-Newton gradient-based minimisation scheme will later be used. As opposed to the Gaussian case, the retrieved scaling factors $[\alpha^s]_l$ cannot be negative. This solution implicitly assumes a truncated Gaussian distribution for the background error statistics.

2.5 Estimation of hyperparameters

The parameters of the prior statistics, such as r_s and m_s , give a measure of the magnitude of the prior uncertainties. For a

reliable emission inverse modelling, these often need to be estimated, in addition to the emission fluxes, because their initial guess is usually inaccurate, while the dependence of the retrieval on these parameters can be dramatic (Davoine and Bocquet, 2007). When those r_s and m_s are considered as parameters to be themselves estimated on top of the flux parameters, they are usually coined *hyperparameters*.

The estimation method for the hyperparameters depends on the statistical assumptions underlying Eq. (7). The observation error vector ϵ^s (in Eq. (5)) is assumed to be Gaussian distributed, $\epsilon^s \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_s)$. A similar assumption applies to the control parameters: $\alpha^s \sim \mathcal{N}(\mathbf{1}, \mathbf{B}_s)$. The probability density function (pdf) of the observation vector can be computed as follows:

$$p(\boldsymbol{\mu}^s | r_s, m_s) = \int p(\boldsymbol{\mu}^s | \boldsymbol{\alpha}^s, r_s, m_s) p(\boldsymbol{\alpha}^s | r_s, m_s) d\boldsymbol{\alpha}^s \quad (10)$$

$$= \int p(\epsilon^s | r_s) p(\boldsymbol{\alpha}^s | m_s) d\boldsymbol{\alpha}^s$$

or analytically,

$$p(\boldsymbol{\mu}^s | r_s, m_s) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}^s - \mathbf{H}^s \mathbf{1} - \lambda^s)^T (\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{sT})^{-1} (\boldsymbol{\mu}^s - \mathbf{H}^s \mathbf{1} - \lambda^s)}}{\sqrt{(2\pi)^{d_s} |\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{sT}|}}. \quad (11)$$

This pdf is also proportional to the likelihood of r_s and m_s . In order to estimate the hyperparameters r_s and m_s , Desroziers and Ivanov (2001) suggested an iterative method to converge towards a fixed point. Chapnik et al. (2006) showed that this approach converges to one local maximum of the likelihood. The maximisation of $\ln(p(\boldsymbol{\mu}^s | r_s, m_s))$ with respect to the two hyperparameters gives the stationary conditions:

$$r_s^2 = \frac{\|\boldsymbol{\mu}^s - \mathbf{H}^s \boldsymbol{\alpha}_a^s\|^2}{\text{Tr}(\mathbf{I}_{d_s} - \mathbf{H}^s \mathbf{K}^s)}, \quad m_s^2 = \frac{\|\boldsymbol{\alpha}_a^s - \mathbf{1}\|^2}{\text{Tr}(\mathbf{H}^s \mathbf{K}^s)} \quad (12)$$

where $\|\cdot\|$ is the Euclidean norm.

However, the Desroziers method relies on Gaussian assumptions, and, for the sake of consistency, one needs another approach to compute the likelihood under the truncated Gaussian assumption (Winiarek et al., 2012). In this case, the prior on the scaling factors is

$$p(\boldsymbol{\alpha}^s) = \frac{e^{-\frac{1}{2}(\boldsymbol{\alpha}^s - \mathbf{1})^T \mathbf{B}_s^{-1} (\boldsymbol{\alpha}^s - \mathbf{1})}}{\sqrt{(2\pi)^{N_x \times N_y} |\mathbf{B}_s| (1 - \Phi_{\mathbf{1}, \mathbf{B}_s}(\mathbf{0}))}} \mathbb{I}_{\boldsymbol{\alpha}^s \geq \mathbf{0}}, \quad (13)$$

where $\Phi_{\mathbf{1}, \mathbf{B}_s}(\mathbf{0})$ is the Gaussian cumulative density function (cdf) of $\mathcal{N}(\mathbf{1}, \mathbf{B}_s)$. $\mathbb{I}_{\boldsymbol{\alpha}^s \geq \mathbf{0}}$ is a function equal to 1 when $[\boldsymbol{\alpha}^s]_l \geq 0$ for each l , and equal to 0 otherwise. This pdf is referred to as $\mathcal{N}(\mathbf{1}, \mathbf{B}_s, \mathbf{0})$ or the truncated Gaussian distribution. From Eq. (10) and Eq. (13), one can derive

$$p(\boldsymbol{\mu}^s | r_s^+, m_s^+) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}^s - \mathbf{H}^s \mathbf{1} - \lambda^s)^T (\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{sT})^{-1} (\boldsymbol{\mu}^s - \mathbf{H}^s \mathbf{1} - \lambda^s)}}{\sqrt{(2\pi)^{d_s} |\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{sT}|}} \times \frac{\int_{\boldsymbol{\alpha}^s \geq \mathbf{0}} e^{-\frac{1}{2}(\boldsymbol{\alpha}^s - \mathbf{1})^T (\mathbf{P}_s^a)^{-1} (\boldsymbol{\alpha}^s - \mathbf{1})}}{\sqrt{(2\pi)^{N_x \times N_y} |\mathbf{B}_s| (1 - \Phi_{\mathbf{1}, \mathbf{B}_s}(\mathbf{0}))}}, \quad (14)$$

where r_s^+ and m_s^+ refer respectively to the standard deviation of the observation error and of the background error (departure of the surface fluxes from their a priori values) according to the truncated Gaussian distribution. \mathbf{P}_s^a is the a posteriori error covariance matrix of the control variables of the Gaussian case,

$$\mathbf{P}_s^a = (\mathbf{I} - \mathbf{K}^s \mathbf{H}^s) \mathbf{B}_s. \quad (15)$$

Even though it is of formal use in the truncated Gaussian case, \mathbf{P}_s^a is not the error covariance matrix of the truncated Gaussian case. A mathematical hardship is that Eq. (15) requires the computation of a 1768-dimensional ($N_x \times N_y$) integral over the space of all positive emission fields. To overcome this difficulty, we resort to the stochastic sampling technique used by Winiarek et al. (2012).

3 Setup of the numerical experiments

3.1 Observations

The in situ observations used in this study are extracted from the EMEP database¹. The EMEP monitoring network covers most of Europe. Eleven stations of western Europe are used in this study, resulting in a rather sparse network. These stations measure the concentrations of 14 different VOCs. Note that the m-xylene and p-xylene are combined in a lumped mp-xylene category. The observations are performed from 11 January 2005 to 29 December 2005. The sample duration ranges from 20 min to 24 h depending on stations and species².

Table 2 gives the number of observations used in the inversion per species and per station, for a total of 18 675 observations from 11 stations. A forecast test will also be performed using 19 746 observations of the year 2006. The VOC station Kollumerwaard, in the Netherlands (code NL0009R), does not provide any observation in 2005. However, this station provides 26 732 observations in 2006, and they will be used for cross validation. The locations of the EMEP sampling stations for VOCs are shown in Fig. 1.

The uncertainties associated with the measurements have been analysed in detail by Sauvage (2008) for three stations of the monitoring network considered here (stations 9, 10 and 11 of Fig. 1). For the VOC considered here, the measurement uncertainties were estimated to consist of a heteroscedastic component ranging from 11 % (ethane) to 25 % (m- & p-xylene) of the measured concentration and a constant component of 0.01 ppb associated with the detection limit.

3.2 Inversion and validation setup

Two simulation periods are considered. The first one is the assimilation time window of the study, from 11 January to 29

¹<http://www.nilu.no/projects/ccc/emepdata.html>

²<http://www.nilu.no/projects/ccc/index.html>

Table 2. Number of observations for each species ($N_{\text{species}}^{\text{obs}}$) and number of observation dates for each station ($N_{\text{station}}^{\text{obs}}$). The numbers 1–12 are given to help locate the stations on the map of Fig. 1.

species	$N_{\text{species}}^{\text{obs}}$		station code	$N_{\text{station}}^{\text{obs}}$	
	2005	2006		2005	2006
C ₃ H ₈	1505	1415	1 CZ0003R	1397	1297
NBUT	1503	1414	2 DE0002R	948	1023
IBUT	1033	1453	3 DE0003R	837	990
NPEN	1501	1409	4 DE0007R	942	1029
IPEN	1610	1553	5 DE0008R	883	1054
C ₃ H ₆	1497	1411	6 DE0009R	938	1020
TOLU	1441	1408	7 DE0043G	9206	9283
OXYL	868	1273	8 ES0009R	259	199
MPXYL	1318	1459	9 FR0008R	1215	1285
ISO	936	1309	10 FR0013R	769	1214
ACE	1034	1407	11 FR0015R	1281	1352
C ₂ H ₆	1494	1409	12 NL0009R	0	26732
C ₂ H ₄	1434	1416			
BEN	1503	1410			

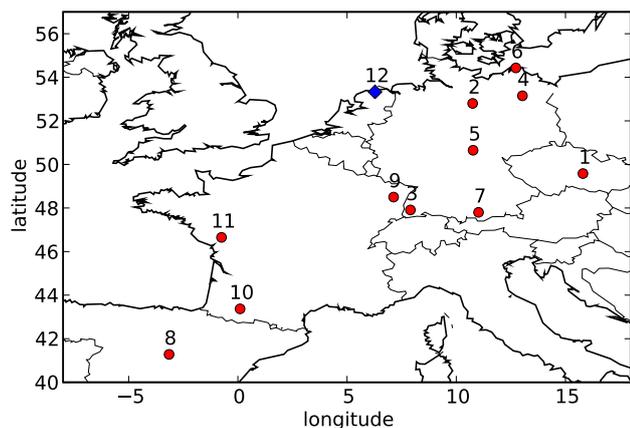


Fig. 1. The 11 monitoring stations of the EMEP network for volatile organic compounds whose observations are assimilated are indicated with a circle. The Kollumerwaard station in the Netherlands used for validation only is indicated by a rhombus.

December 2005. The second one is the whole year 2006, and is meant for a subsequent validation of the inversion results.

The simulation domain over western Europe extends between [40° N, 8° W] (the left bottom corner) and [57° N, 18° E] (the right top corner). The grid resolution is $0.5^\circ \times 0.5^\circ$. Nine levels are considered above the surface: 50, 250, 600, 1000, 1500, 2100, 2800, 3600 and 4500 m a.g.l. The control space discretisation follows the simulation grid, with $N_x = 52$ and $N_y = 34$. As a result, the number of control variable $[\alpha^s]_l$ used in the inversion for each species s is 1768.

The meteorological data are generated from the reanalysis fields of the European Centre for Medium Range Weather

Forecast (ECMWF), delivered in 60 vertical levels and every 3 h, with a horizontal resolution of $0.36^\circ \times 0.36^\circ$.

For anthropogenic emissions, the background emissions over the whole domain are provided by the EMEP inventory for the years 2005 and 2006 (Tarrasón et al., 2007; Fagerli et al., 2008). The disaggregation of the EMEP NMVOC emissions among the many VOC species simulated in the CTM is performed following Passant (2002).

The anthropogenic emissions of EMEP have a resolution of $0.5^\circ \times 0.5^\circ$. These emissions are modulated in time with the help of the hourly, weekly and monthly distribution coefficients, provided by the GENEMIS project (GENEMIS, 1994). The biogenic emissions of isoprene are also taken into account using the model proposed by Simpson et al. (1999). All these emissions are used as a first guess (mean state of the background), e^b , in the data assimilation experiments.

The initial and boundary conditions (top and lateral faces of the domain) concentration fields are obtained from the global chemistry transport model MOZART 2 (Horowitz et al., 2003). Since the species we are interested in are not all explicitly present in MOZART 2, the values for the VOCs not included in MOZART 2 were inferred from the concentration fields of some species present in MOZART 2. The factors applied for this inference are given in Table 3 (see Rudolph and Ehhalt, 1981; Rudolph and Johnen, 1990; Penkett et al., 1993).

Our inverse modelling methodology estimates the magnitude of errors of all types. However, it does not handle a possible systematic bias, for instance in the boundary conditions. To get a rough idea of the uncertainties in the boundary conditions, we have performed a comparison of the results of the MOZART 2 model, from which are extracted the boundary conditions, to the EMEP observations used in the inversion. From this comparison, rough average relative errors are estimated, ranging from 1 % (for C₃H₈) to 116 % (for C₃H₆). For the long-lived species (ACE, C₂H₆ and C₃H₈) rough average relative errors are respectively –60 %, 8 % and 1 %. Thus, we believe that the inversion of acetylene emissions, with a lifetime of about 110 days and presumably biased boundary conditions, must be interpreted with caution.

3.3 Verification of the adjoint solutions

In order to generate the adjoint solutions at a low computational cost, we have used an approximate adjoint model following the construction of Bocquet (2005). The approximate adjoint is built by discretising the adjoint of the continuous transport equation. Since, in general, discretisation and adjointisation are operators that do not commute, this adjoint is only approximate. However, a benefit of this approach is that its implementation re-uses the forward CTM but with reversed wind fields and a backward integration in time.

Moreover, we have assumed the lifetime of the species within the domain to be less than 10 days. After 10 days, the VOCs are assumed to be out of the domain or consumed

Table 3. Factors applied to MOZART 2 explicit species concentrations to determine the initial and boundary conditions of the model species.

species	factors and MOZART 2 species
C ₃ H ₈	C ₃ H ₈
NBUT	0.44 C ₃ H ₈
IBUT	0.22 C ₃ H ₈
NPEN	0.05 C ₃ H ₈
IPEN	0.1 C ₃ H ₈
C ₃ H ₆	C ₃ H ₆
TOLU	0.26 C ₃ H ₈
OXYL	0.03 C ₃ H ₈
MPXYL	0.03 C ₃ H ₈
ISO	0.03 C ₃ H ₈
ACE	0.35 C ₃ H ₈
C ₂ H ₆	C ₂ H ₆
C ₂ H ₄	C ₂ H ₄
BEN	0.44 C ₃ H ₈

by chemical reactions, and as a result the sensitivity of the concentrations within the domain to the emissions is negligible.

In order to check these approximations, the concentration fields from the adjoint model, obtained from the contribution of the source ($\mathbf{H}\alpha$ when $\alpha = \mathbf{1}$), were compared with the concentration fields from the direct simulation (for $e = e^b$, the EMEP inventory first guess, with clean air boundary and initial conditions). This is the so-called duality test (Davoine and Bocquet, 2007).

Figure 2 shows the comparison between the source contribution to the observations estimated with the forward model and with the adjoint model. Table 4 gives the Pearson correlation, the mean values and the root-mean-square error that result from the comparison of the two data sets. These results indicate that, with a view to inverse modelling, the adjoint model is accurate enough for most species. However, the approximation is not as good for isoprene (fractional bias of 17%, a correlation of 0.97 and a normalised root-mean-square error of about 0.25) due to its short lifetime that induces large numerical errors in the adjoint computation because of the resulting weak correlations between the sources and the receptors.

3.4 Values of the hyperparameters

In the Gaussian case, the optimal values of the hyperparameters r_s and m_s are obtained by *value screening* of the pdf Eq. (11). This means that the couples (r_s, m_s) run over all points of a regular grid of the (r_s, m_s) plane, and the corresponding pdf values have been computed. Their optimal values are close to those maximising the likelihood pdf. They are reported in Table 5 for each species. As an example, the left panel of Fig. 3 displays $p(\mu^s | r_s, m_s)$ for the species n-butane (NBUT). The coordinates are normalised with respect

Table 4. Statistics of the discrepancy between the values of the observations as generated by the forward model, or generated with the approximate adjoint model: Pearson correlation (R), mean direct value, mean adjoint-based value and normalised mean square error (NMSE). The mean values are in $\mu\text{g m}^{-3}$.

species	R	mean direct μ^s	mean adjoint-based μ^s	NMSE
C ₃ H ₈	0.9981	0.5325	0.5580	0.0062
NBUT	0.9935	0.6287	0.6776	0.0229
IBUT	0.9967	0.3003	0.3279	0.0389
NPEN	0.9958	0.2369	0.2672	0.0528
IPEN	0.9955	0.2557	0.2825	0.0397
C ₃ H ₆	0.9962	0.1274	0.1404	0.0693
TOLU	0.9927	0.2290	0.2441	0.0387
OXYL	0.9958	0.3167	0.3339	0.0314
MPXYL	0.9901	0.1429	0.1478	0.0526
ISO	0.9728	0.1800	0.1521	0.2540
ACE	0.9995	0.4108	0.4241	0.0037
C ₂ H ₆	0.9986	1.0104	1.0437	0.0029
C ₂ H ₄	0.9953	0.2609	0.2814	0.0236
BEN	0.9979	0.4003	0.4214	0.0081

Table 5. Estimated standard deviations of the observation error (r_s and r_s^+) and background error (m_s and m_s^+), under the Gaussian likelihood and truncated Gaussian likelihood. r_s and m_s are related to case B1 and case B2, while r_s^+ and m_s^+ are related to case C. The units of r_s and r_s^+ are $\mu\text{g m}^{-3}$, while m_s and m_s^+ are unitless.

species	r_s	m_s	r_s^+	m_s^+
C ₃ H ₈	0.48	16.04	0.50	4.03
NBUT	0.36	2.02	0.37	0.64
IBUT	0.24	35.17	0.28	2.72
NPEN	0.19	6.01	0.20	0.76
IPEN	0.30	8.16	0.32	1.63
C ₃ H ₆	0.14	14.22	0.15	2.84
TOLU	0.34	3.16	0.35	0.63
OXYL	0.08	16.75	0.09	1.33
PXYL	0.16	6.52	0.17	0.52
MXYL	0.16	6.52	0.17	0.52
ISO	0.46	82.65	0.60	2.61
ACE	0.35	41.26	0.36	8.23
C ₂ H ₆	0.55	17.40	0.57	4.37
C ₂ H ₄	0.35	31.19	0.40	3.93
BEN	0.24	11.12	0.27	1.11

to the values of the hyperparameters obtained from the fixed-point solutions of Eq. (12). It can be seen that the optimal value of the likelihood Eq. (11) is equal to the result of the fixed-point method.

A comparison between the likelihoods Eq. (11) and Eq. (14) is also shown in Table 5. The optimal value of r_s^+ and m_s^+ computed with Eq. (14) are obviously different from r_s and m_s . Remarkably r_s^+ is always larger than r_s . Indeed, the Gaussian assumption inversion incorrectly

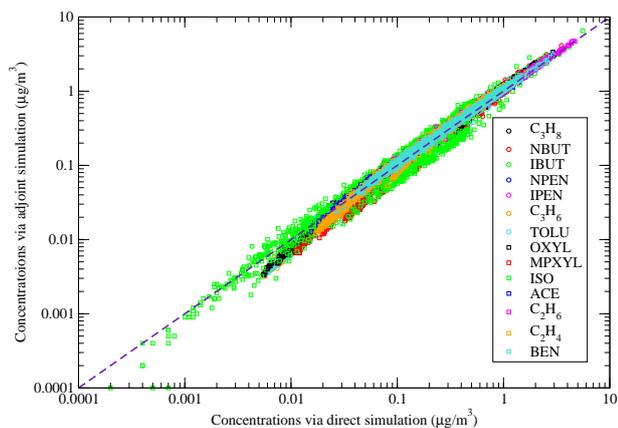


Fig. 2. Comparison of the source contribution to the observations estimated with the direct model and the adjoint model.

interprets part of the noise within the observations as useful information, while the positivity constraint of the truncated Gaussian assumption offers less flexibility. Therefore, the inversion based on Gaussian assumptions underestimates the errors' magnitude. The comparison between m_s^+ and m_s is of less relevance because the two parameters do not represent the same statistical information within the Gaussian or truncated Gaussian assumptions.

As an example, the right panel of Fig. 3 displays the likelihood of the hyperparameters for ethane in the truncated Gaussian case. Again, the variables are normalised with respect to the values obtained from the Desroziers method.

4 Inversion results

Four types of simulations conducted for the year 2005 are reported here:

- In case A, the VOC concentrations are simulated using the EMEP inventories, $e_k = e_k^b$ in Eq. (1). This is a free run serving as a reference since the observations are not assimilated.
- In cases B1 and B2, the concentrations are simulated using the emissions obtained from Eq. (4). In case B1, the scaling factors α^s used in this equation are obtained from the BLUE matrix formula Eq. (8). By construction, case B1 can lead to negative emission fluxes. This may be considered unphysical but it might have some potential use for air quality forecast.

In case B2, the scaling factors α^s are obtained from the minimisation of Eq. (7). The L-BFGS-B minimisation tool (Byrd et al., 1995) is used under the constraints that $\alpha^s \geq 0$. However, it is assumed in both cases that the errors are essentially Gaussian, so the hyperparameters used in the inversion are computed with the Gaussian

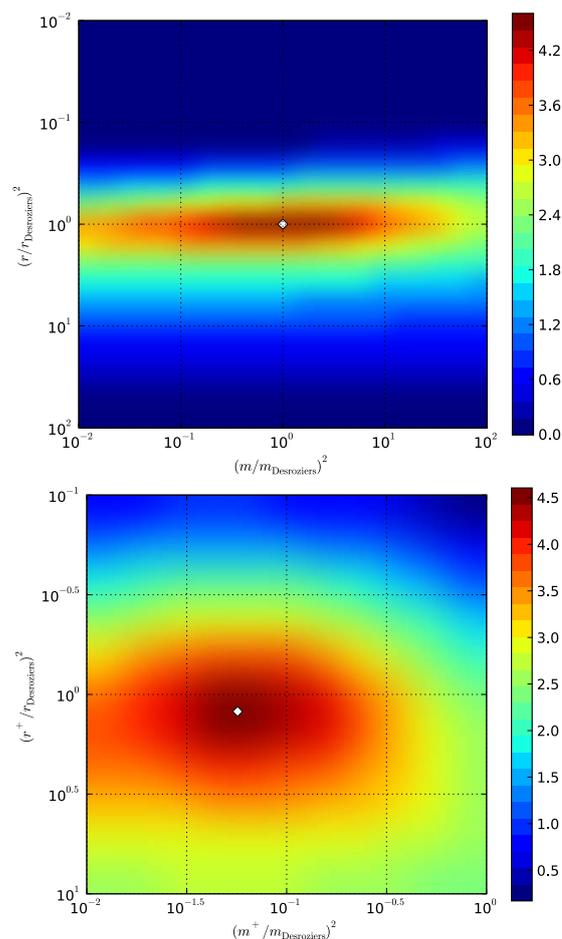


Fig. 3. These density plots display a monotonic transform of the likelihood of the hyperparameters for NBUT in the Gaussian case (left) and C_2H_6 in the truncated Gaussian case (right). The monotonic transform is used to obtain a better contrast in the density plot. The abscissa and ordinate are normalised according to the optimal parameters obtained from the fixed-point method.

likelihood following Sect. 2.5. In case B2, the statistical assumptions are different in the estimation of the fluxes and the estimation of the hyperparameters, which may lead to inconsistencies. However, the fact that the r^+ and r hyperparameters are not too different proves that these inconsistencies are small.

We have also considered a third case B3, which takes the results of B1 and artificially sets all negative values to 0. However, we found that because it is not a minimum of Eq. (7), it leads to a poorly performing estimation. Therefore, case B3 was ruled out and is not reported here.

- Case C is similar to case B2 except that the hyperparameters are obtained using the truncated Gaussian likelihood following Sect. 2.5. In this case the estimation of

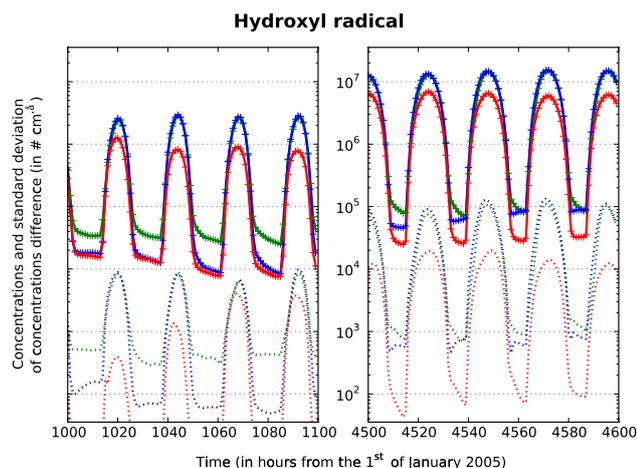


Fig. 4. Time evolution of the horizontal mean over the whole domain of OH hourly concentrations for different levels (ground level in green, third level in blue and top level in red). Full lines are used for a priori simulations and crossed ones are used for a posteriori simulations. The dotted lines represent the standard deviation computed for the corresponding levels.

the emission fluxes and the estimation of the hyperparameters are statistically fully consistent.

4.1 Analysis of the inversion results

4.1.1 A posteriori verification of the model linearisation

In the full CTM, which is used in the present study, the chemical kinetics of the reactions can be written as follows:

$$\mathcal{X}(c) = \left(\frac{\delta \mathcal{X}(c)}{\mathcal{X}(c)} + 1 \right) \mathbf{X}(c), \quad (16)$$

where $\delta \mathcal{X}(c)$ denotes the variation of the VOC concentrations field with respect to the variation of the oxidant concentration field. Until now, we assumed that $\delta \mathcal{X}(c) \ll \mathbf{X}(c)$, and that the reduced model was linear.

In order to check that hypothesis, the a priori and the a posteriori (case C) oxidant concentration fields calculated by the full CTM are compared. Because the inversion of isoprene emission will be later deemed unreliable, the EMEP emissions (a priori emissions) have then been kept for ISO to perform this test. The mean value of the OH concentration is $1.882 \times 10^6 \text{ # cm}^{-3}$ for the a priori fields and $1.877 \times 10^6 \text{ # cm}^{-3}$ for the a posteriori fields. The Pearson correlation between the a priori and a posteriori concentration fields is about 1.00. The standard deviation of the difference between the a priori and the a posteriori fields (computed for all the time steps and cells) is $2.705 \times 10^4 \text{ # cm}^{-3}$. Figure 5 shows the comparison of the concentration fields of OH before and after data assimilation for two representative periods of hundreds of hours: the first one in winter, the second one in summer. For NO_3 and O_3 , the average values of the a priori concentrations are $1.418 \times 10^8 \text{ # cm}^{-3}$ and 1.090×10^{12}

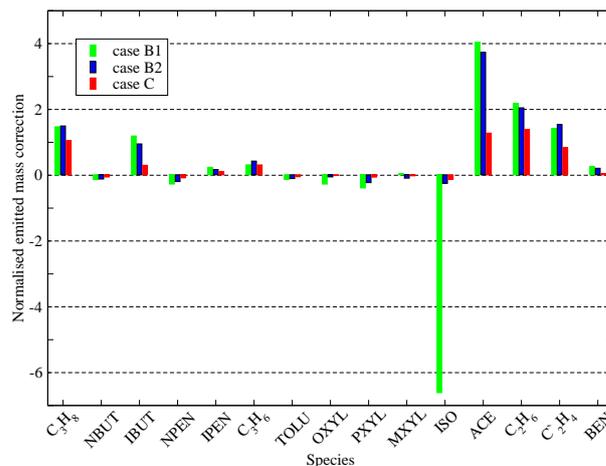


Fig. 5. The total emitted mass correction, normalised with respect to the total emitted mass of the EMEP inventory for cases B1, B2 and C. The normalised emitted mass correction represents the correction to the emitted mass ΔE divided by the original emitted mass E_0 . Therefore, there is no correction if $\Delta E/E_0$ is equal to 0. If it is equal to 1, the corrected emitted mass is twice the original emitted mass. If it is equal to -1 , the corrected emitted mass is 0. If it is less than -1 , then the corrected emitted mass is negative.

molec cm^{-3} , respectively. They are $1.418 \times 10^8 \text{ # cm}^{-3}$ and $1.092 \times 10^{12} \text{ molec cm}^{-3}$, respectively, for the a posteriori concentrations. The Pearson correlations between the a priori and a posteriori concentration fields are about 1.00 for both NO_3 and O_3 . The standard deviation of the difference between a priori and a posteriori fields is $2.627 \times 10^4 \text{ # cm}^{-3}$ for NO_3 and $4.218 \times 10^9 \text{ molec cm}^{-3}$ for O_3 . One additional iteration of the linearisation and inversion cycle barely changes these results.

Furthermore, an examination of the a posteriori VOC concentration fields shows that the results obtained with the reduced linear model are very close to those obtained with the complete model. The relative bias between the two sets of concentrations for all of the VOC species and over the entire spatial domain is 1% for the year 2005 and the correlation is 1.00. Therefore, since the oxidant concentration fields are little affected by the VOC data assimilation, we consider that the hypothesis that the reduced model is about linear is verified.

4.1.2 Comparison to the observations

The four runs A, B1, B2 and C are compared with the observations of the analysis period (year 2005). Statistical indicators for this comparison are reported in Table 6. For most species the bias between the concentrations and the observations decreases with data assimilation, except when the mean of the simulation is already close to the measurement mean. The root-mean-square errors (RMSEs) and the normalised mean square errors (NMSEs) are systematically improved

Table 6. Scores from the comparison between the observations and the simulated concentrations for four simulations. For each species, the first line represents the scores for the simulations with the background fluxes (case A). The scores of the second line and third lines are related to the simulations with the a posteriori emissions from Gaussian hyperparameters estimation (case B1 and case B2, respectively). The scores of the fourth line are related to the simulations performed with the a posteriori emissions under the truncated Gaussian assumption (case C). The means and the RMSE are in $\mu\text{g m}^{-3}$. Bold numbers correspond to the best agreement with the observations.

species	\bar{O}	case	\bar{C}	RMSE	NMSE	R	FA ₂	FA ₅
C ₃ H ₈	1.13	A	0.54	0.83	1.14	0.77	0.40	0.946
		B1	1.08	0.46	0.18	0.85	0.91	0.999
		B2	1.07	0.49	0.20	0.83	0.91	0.998
		C	1.04	0.51	0.22	0.82	0.90	0.998
NBUT	0.62	A	0.64	0.41	0.43	0.68	0.80	0.989
		B1	0.59	0.36	0.35	0.75	0.85	0.996
		B2	0.59	0.36	0.36	0.74	0.85	0.995
		C	0.58	0.37	0.38	0.72	0.84	0.995
IBUT	0.33	A	0.18	0.37	2.21	0.55	0.59	0.960
		B1	0.33	0.23	0.46	0.82	0.79	0.972
		B2	0.33	0.28	0.69	0.71	0.78	0.978
		C	0.31	0.30	0.86	0.65	0.77	0.984
NPEN	0.30	A	0.25	0.25	0.86	0.48	0.57	0.939
		B1	0.26	0.18	0.41	0.70	0.77	0.987
		B2	0.25	0.20	0.51	0.65	0.71	0.974
		C	0.24	0.21	0.63	0.60	0.66	0.972
IPEN	0.51	A	0.45	0.42	1.39	0.42	0.44	0.875
		B1	0.46	0.29	0.36	0.65	0.82	0.993
		B2	0.45	0.31	0.42	0.60	0.78	0.989
		C	0.43	0.32	0.47	0.56	0.77	0.989
C ₃ H ₆	0.18	A	0.06	0.21	4.37	0.45	0.27	0.653
		B1	0.16	0.13	0.61	0.72	0.78	0.970
		B2	0.15	0.15	0.77	0.66	0.71	0.964
		C	0.15	0.18	0.85	0.64	0.70	0.963
TOLU	0.46	A	0.46	0.39	0.73	0.47	0.65	0.953
		B1	0.43	0.34	0.56	0.58	0.72	0.957
		B2	0.43	0.35	0.60	0.56	0.70	0.959
		C	0.43	0.36	0.64	0.53	0.68	0.960
OXYL	0.09	A	0.055	0.11	2.21	0.37	0.49	0.879
		B1	0.086	0.07	0.65	0.74	0.68	0.962
		B1	0.083	0.09	0.97	0.60	0.63	0.950
		C	0.080	0.09	1.06	0.57	0.62	0.946
MPXYL	0.19	A	0.17	0.20	1.25	0.33	0.53	0.909
		B1	0.17	0.15	0.71	0.59	0.65	0.956
		B2	0.16	0.17	0.89	0.52	0.62	0.940
		C	0.16	0.17	1.01	0.47	0.57	0.919
ISO	0.31	A	0.18	0.86	13.47	0.63	0.35	0.708
		B1	0.31	0.46	2.22	0.91	0.45	0.716
		B2	0.36	0.59	3.18	0.82	0.46	0.789
		C	0.35	0.59	3.22	0.82	0.46	0.795
ACE	0.52	A	0.12	0.57	4.95	0.69	0.11	0.723
		B1	0.49	0.34	0.45	0.67	0.72	0.995
		B2	0.49	0.35	0.49	0.64	0.71	0.992
		C	0.47	0.38	0.59	0.57	0.69	0.989
C ₂ H ₆	1.93	A	1.17	0.99	0.43	0.78	0.76	1.000
		B1	1.83	0.53	0.08	0.86	0.99	1.000
		B2	1.81	0.55	0.09	0.85	0.99	1.000
		C	1.76	0.59	0.10	0.83	0.99	1.000
C ₂ H ₄	0.64	A	0.20	0.68	3.50	0.63	0.22	0.677
		B1	0.61	0.33	0.28	0.85	0.81	0.987
		B2	0.59	0.39	0.40	0.79	0.77	0.984
		C	0.58	0.43	0.49	0.74	0.76	0.983
BEN	0.47	A	0.42	0.31	0.47	0.69	0.74	0.994
		B1	0.46	0.23	0.25	0.81	0.84	0.977
		B2	0.46	0.26	0.31	0.77	0.82	0.992
		C	0.46	0.28	0.36	0.73	0.82	0.993

Table 7. For all species, the total emitted mass (in Gg) for the EMEP inventory run (case A), the a posteriori emissions under Gaussian assumption (cases B1 and B2) and the a posteriori emissions under the truncated Gaussian assumption (case C).

species	case A	case B1	case B2	case C
C ₃ H ₈	7.4	18.3	18.4	15.1
NBUT	19.0	16.5	16.8	18.0
IBUT	4.7	10.3	9.1	6.1
NPEN	9.7	7.1	7.8	8.9
IPEN	9.2	11.4	10.8	10.1
C ₃ H ₆	4.1	5.4	5.9	5.4
TOLU	12.7	11.1	11.4	12.1
OXYL	2.6	1.9	2.4	2.5
PXYL	5.7	3.5	4.4	5.4
MXYL	2.3	2.4	2.1	2.2
ISO	165.0	-925.5	124.9	143.4
ACE	2.1	10.6	9.9	4.8
C ₂ H ₆	7.2	22.9	22.0	17.2
C ₂ H ₄	7.9	19.1	20.1	14.5
BEN	5.6	7.1	6.7	5.8

in the reanalysis runs, which is consistent with the fact that our inversion scheme minimises the quadratic error. For all species, except acetylene (ACE), the Pearson correlation coefficients R , FA_2 and FA_5 are remarkably improved in the reanalyses. FA_x is the fraction of the simulated concentrations within a factor x of the corresponding observations. In the very few cases where an indicator is not improved, other indicators are improved. The decrease of the correlation in the ACE case is due to a very large bias, which is compensated by a very significant improvement of the other indicators, particularly the bias.

Considering all species (14) and all statistical indicators (6) together, we have counted how often the forecast runs B1, B2 and C beat the free run A: 81 times out of 84 in case B1, 80 times out of 84 in case B2 and 79 times out of 84 in case C.

The fact that run B1 is slightly closer to the observations than B2 and C is consistent with the fact that the optimisation on which B1 relies is less constrained than that of B2 or C (emission fluxes can be negative in case B1). However, it does not prove that method B1 is better than method B2 or C since a comparison with the (already assimilated) observations is merely a check of consistency, not a validation.

4.1.3 Estimated inventories

For each of the 15 species, the total emitted mass of the EMEP inventory is compared to the a posteriori emission obtained from data assimilation following Gaussian assumption (B1 and B2) and truncated Gaussian assumption (C). The results are reported in Table 7. Caution must be used on the interpretation of the results of the statistically consistent Gaus-

sian case B1 since fluxes are allowed to be negative in this case (e.g. isoprene emissions).

The results of case B2 and case C indicate that the EMEP inventories may underestimate the true emissions for propane (C₃H₈), isobutane (IBUT), isopentane (IPEN), propene (C₃H₆), acetylene (ACE), ethane (C₂H₆), ethylene (C₂H₄) and benzene (BEN). They may overestimate the true emissions for the other VOC species. For all the species, the total mass obtained from the inversion based on the truncated Gaussian assumption (case C) is between the EMEP inventory and the total mass estimated with the Gaussian assumption (case B2). Comparison between cases A and C shows a strong correction for C₃H₈, ACE, C₂H₆, and C₂H₄. It is less than 20 % for n-butane (NBUT), n-pentane (NPEN), isopentane (IPEN), toluene (TOLU), o-xylene (OXYL) and m-xylene (MXYL). Figure 5 presents the ratio between the correction of the emission (i.e. the total posterior emission minus the total prior EMEP emissions) and the total prior emission for the Gaussian cases (B1 and B2) and the truncated Gaussian case (C).

4.1.4 Spatial distribution

The spatial extent of the corrections from the EMEP network depends on the nature of the species. As an example, Fig. 6 displays the spatial ratio between the posterior emissions and the prior EMEP emissions for acetylene (ACE, with an average lifetime of 110 days), ethane (C₂H₆, with an average lifetime of about 60 days), propane (C₃H₈, with an average lifetime of about 14 days), isobutane (IBUT, with an average lifetime of about 7.5 days), ethylene (C₂H₄, with an average lifetime of about 1.45 days) and isoprene (ISO, with an average lifetime of about 1.7 h). Obviously, the corrections extend much farther from the monitoring stations for the long-lived species, such as ACE and C₂H₆, than for the short-lived species.

As can be seen, with the B2 approach based on a Gaussian estimation of the prior errors, the magnitude of the corrections is significantly higher than when using the fully consistent non-Gaussian approach C. One way to understand this is that, with the Gaussian assumption, part of the error (noise) is misinterpreted as valuable information (signal), such that the Gaussian assumption leads to over-corrections. A similar phenomenon was put forward by Koohkan and Bocquet (2012) in the inversion of carbon monoxide emission fluxes: the proper identification of representativeness errors leads to smaller corrections of the emission fluxes.

Several studies have performed inverse modelling of isoprene emissions over Europe. To do so, they assimilate satellite observations of formaldehyde (e.g. Curci et al., 2010; Dufour et al., 2009). Specifically, they exploit the observations of SCIAMACHY (instrument operating onboard the sun-synchronous Envisat satellite) with a resolution of 30 × 60 km². The study of Curci et al. (2010) shows an increase of about 5 % for the MEGAN (Guenther et al., 2006) emissions. Our

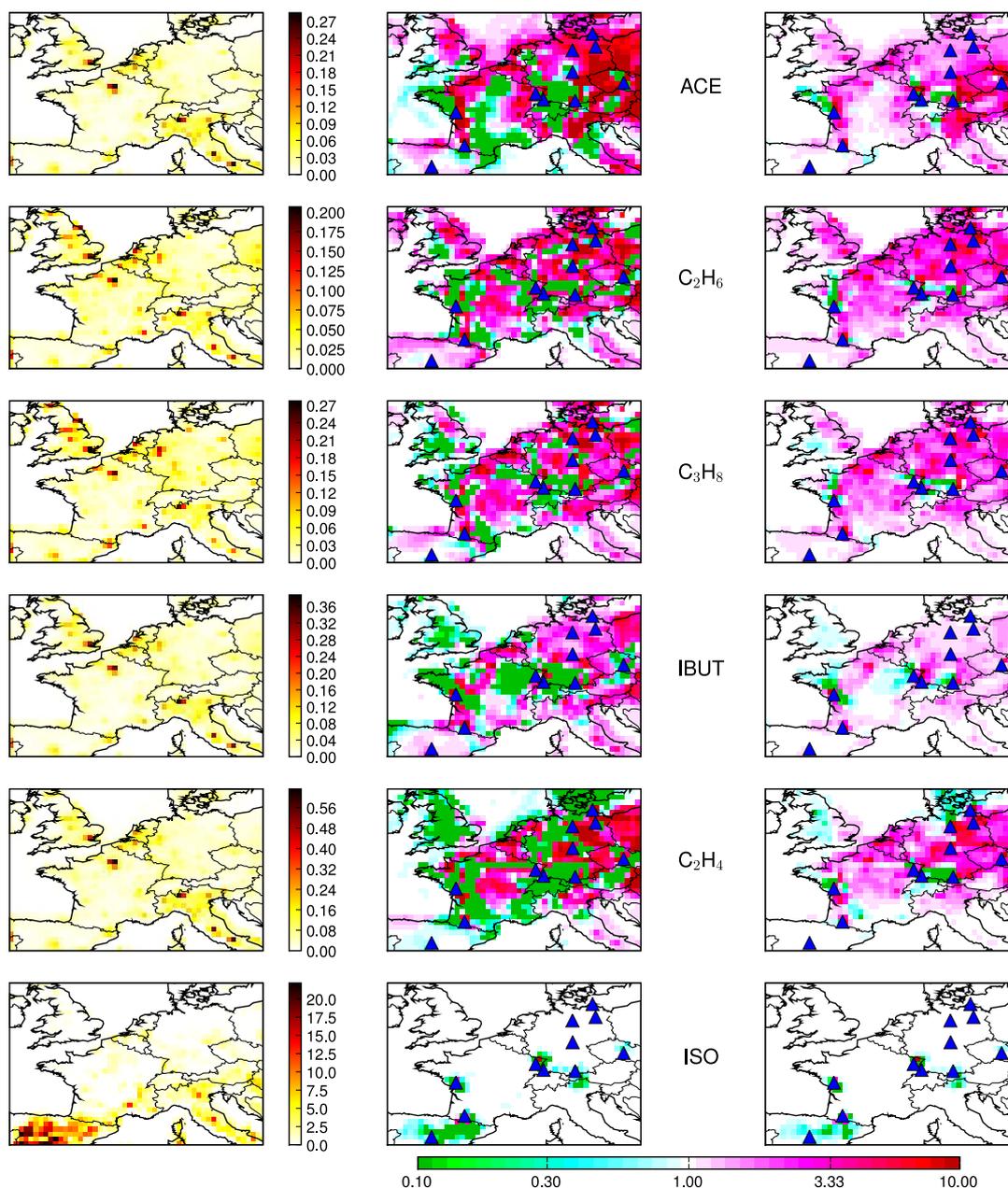


Fig. 6. Gridded EMEP time-integrated flux (in $\text{t km}^{-2} \text{yr}^{-1}$, left column) and gridded ratios of time-integrated retrieved flux to EMEP time-integrated flux for, from top to bottom, acetylene (ACE), ethane (C_2H_6), propane (C_3H_8), isobutane (IBUT), ethylene (C_2H_4) and isoprene (ISO). The species are ordered by decreasing lifetime. The central column corresponds to case B2 and the right column corresponds to case C. Green and blue colours correspond to reduction of the emission fluxes, whereas red and pink colours correspond to an increase of the emission fluxes.

results show that the emissions of isoprene decrease by about 24 % in case B2 and by 13 % in case C for the emissions obtained from Simpson et al. (1999). This discrepancy is explained in part by the fact that the emission inventories of isoprene by Simpson et al. (1999) and MEGAN differ significantly (Bessagnet et al., 2008; Sartelet et al., 2012),

with the former leading to greater isoprene emissions on average over Europe – for example by a factor of about 2.5 for the July–August 2006 period (Sartelet et al., 2012). This discrepancy is also explained by the very short lifetime of isoprene (1–2 h). For this species, assimilation of in situ observations seems quite inoperative because the information

is not spread far enough by the (adjoint) model. Our results are only valid near the five stations measuring isoprene. On the contrary, satellite observations are well suited for this short-lived species as remote sensing offers a (indirect) spatially well-resolved snapshot of the concentrations, although uncertainties are associated with using oxidation products (formaldehyde, glyoxal) of isoprene to retrieve the isoprene emissions. In addition, the errors made by the approximate adjoint model (or automatic adjoint up to some numerical precision) are larger for short-lived species (Bocquet, 2012). As a consequence, the case of isoprene in this study is somehow singular because its lifetime is so short. Nevertheless, most results of isoprene are given in this study for the sake of comparison and to document the inadequacy of the in situ observations for regional inverse modelling in such a case.

4.2 Forecast test

In order to test and possibly validate the corrected emissions obtained from inverse modelling, one needs observations that have not been assimilated. One stringent test is to perform a forecast using the corrected emissions over a period of time different from the data assimilation window, assuming some time persistence of the VOC inventories.

Four inventories are generated with Eq. (4), using the background EMEP emission e_b^s , or $\alpha_b^s = \mathbf{1}$ (case A), and the scaling factors of cases B1, B2 and C. In each case, a forecast is performed for the year 2006. These simulations are then compared with the independent observations of year 2006.

Note that because some of the fluxes retrieved in case B1 are negative, and because numerical schemes of CTMs often rely on the positivity of the concentrations, we had to circumvent the difficulty. One solution consists in decomposing the scaling factors into a positive part and a negative part: $\alpha^s = \alpha_+^s + \alpha_-^s$, such that, invoking the linearity of physics, the concentrations are given by $\mathbf{H}^s(\alpha^s) = \mathbf{H}^s(\alpha_+^s) - \mathbf{H}^s(-\alpha_-^s)$.

The statistical indicators are reported in Table 8. The results indicate that, for most species, the scores are improved using the retrieved scaling factors of case B2 and case C, whereas they are degraded in case B1. Considering all species and all statistical indicators together, we have counted how often the forecast runs B1, B2 and C beat the free run A: 41 times out of 84 in case B1, 69 times out of 84 in case B2 and 77 times out of 84 in case C.

The fully consistent truncated Gaussian approach C performs best and beats the Gaussian-based but positively constrained approach B2. Both positively constrained approaches beat the fully consistently Gaussian approach B1. Since B1 partially leads to unphysical negative fluxes, this could have been expected. However, as was shown by Bocquet (2012), an unconstrained optimisation of parameters that are allowed to take unphysical values may sometimes lead to valuable better forecasts because it compensates for other sources of model error. Obviously, this is not the case here.

According to Table 8, an improvement of all the scores can be seen for C_3H_8 , NBUT, NPEN, C_3H_6 , C_2H_6 , C_2H_4 and BEN using the optimal scaling factors. However, the bias is increased for OXYL (in case B2), MPXYL (in cases B2 and C) and ISO (in case B2), using the corrected emissions. Despite that degraded bias, the other indicators are improved. For IBUT, the RMSE is increased in case B2, and the correlation is decreased in cases B2 and C. The statistical indicators show that the RMSE increases for IPEN using the corrected emissions. For that species, the Pearson correlation coefficient is also deteriorated in case B2. For TOLU, the bias, NMSE and FA_5 are degraded using scaling factors. The RMSE is also increased in case B2. The scores also show a decrease of the Pearson correlation coefficient for ACE when using the corrected emissions.

4.3 Cross-validation test

Since the EMEP Netherlands VOC station Kollumerwaard does not provide any observation for the year 2005, that station is not included in the list of stations used for the analysis of 2005. Yet, for the year 2006, 11 out of 14 VOCs were measured at Kollumerwaard, excluding NBUT, ACE and C_2H_6 . Kollumerwaard is located far from the other stations whose observations lead to the corrections in the emission inventories. Therefore, we do not expect the simulated concentrations of the VOC with a short lifetime, at this station, to be very sensitive to the correction of the emissions. Indeed, as shown in Table 9, the scores obtained from the comparison of the simulated concentrations in case A of ISO (lifetime of about 1.7 h) with the measurements are similar to those obtained from the comparison of the concentrations in cases B and C with the observations.

For species with a longer lifetime (about 1 to 2 days), for C_3H_6 (lifetime of about 15 h), OXYL (lifetime of about 25–26 h), MPXYL (lifetime of about 14–25 h) and TOLU (lifetime of about 2.4 days), the corrections of the inventories are more spread out over the domain. The scores indicate that the results in case C are in a better agreement with the observations. The results also show that for the species with a lifetime between 4 and 7 days (IBUT, NPEN and IPEN), the scores are deteriorated. It is likely that the corrections performed close to the Kollumerwaard station are not reliable enough. For the species with a lifetime longer than 10 days (C_3H_8 and BEN), the scores are remarkably improved. The emission inventories of C_3H_8 are corrected almost all over the domain.

Considering all species and all statistical indicators together, we have counted how often the forecast runs B2 and C beat the free run A at the Kollumerwaard station: 34 times out of 66 in case B2 and 43 times out of 66 in case C. The use of the corrected emissions from case C has a positive impact on the forecast at the station, even though the station is far away from the other stations whose observations were assimilated.

Table 8. Scores of the forecast test (year 2006) from the comparison of the observations and the simulated concentrations for four simulations: case A, case B1, case B2 and case C. The means and the RMSEs are in $\mu\text{g m}^{-3}$. Bold numbers correspond to the best agreement with observations.

species	\bar{O}	case	\bar{C}	RMSE	NMSE	R	FA ₂	FA ₅
C ₃ H ₈	1.10	A	0.55	0.85	1.19	0.76	0.46	0.958
		B1	1.14	0.63	0.32	0.80	0.85	0.992
		B2	1.06	0.61	0.32	0.78	0.89	0.998
		C	1.00	0.59	0.32	0.79	0.88	0.998
NBUT	0.56	A	0.75	0.45	0.49	0.72	0.70	0.977
		B1	0.38	0.39	0.40	0.78	0.83	0.994
		B2	0.62	0.36	0.38	0.76	0.80	0.994
		C	0.65	0.37	0.38	0.76	0.79	0.994
IBUT	0.34	A	0.21	0.33	1.46	0.61	0.67	0.967
		B1	0.36	0.46	1.73	0.49	0.56	0.825
		B2	0.37	0.35	0.93	0.55	0.75	0.975
		C	0.32	0.33	0.98	0.54	0.74	0.975
NPEN	0.29	A	0.31	0.31	1.08	0.45	0.56	0.919
		B1	0.18	0.30	1.72	0.48	0.39	0.692
		B2	0.26	0.28	1.03	0.51	0.61	0.949
		C	0.27	0.27	0.94	0.52	0.62	0.950
IPEN	0.47	A	0.29	0.37	0.99	0.49	0.54	0.906
		B1	0.37	0.62	2.17	0.31	0.43	0.733
		B2	0.46	0.46	0.99	0.42	0.64	0.963
		C	0.42	0.38	0.73	0.49	0.65	0.969
C ₃ H ₆	0.19	A	0.08	0.24	3.86	0.44	0.39	0.674
		B1	0.07	0.24	4.40	0.51	0.24	0.521
		B2	0.13	0.20	1.58	0.60	0.50	0.877
		C	0.13	0.20	1.62	0.60	0.49	0.873
TOLU	0.66	A	0.57	0.69	1.28	0.28	0.55	0.876
		B1	0.47	0.73	1.65	0.25	0.56	0.852
		B2	0.48	0.70	1.55	0.27	0.58	0.864
		C	0.50	0.68	1.42	0.29	0.60	0.875
OXYL	0.07	A	0.07	0.07	1.05	0.42	0.54	0.885
		B1	0.04	0.08	2.35	0.42	0.36	0.693
		B2	0.06	0.07	0.98	0.52	0.54	0.918
		C	0.07	0.07	0.95	0.50	0.56	0.910
MPXYL	0.21	A	0.21	0.25	1.45	0.32	0.51	0.856
		B1	0.10	0.23	2.73	0.49	0.38	0.720
		B2	0.16	0.21	1.38	0.50	0.53	0.885
		C	0.18	0.22	1.33	0.44	0.53	0.863
ISO	0.31	A	0.35	1.04	10.25	0.21	0.26	0.571
		B1	1.39	19.95	939.9	-0.18	0.16	0.287
		B2	0.25	0.81	8.59	0.46	0.34	0.682
		C	0.27	0.81	8.05	0.44	0.33	0.656
ACE	0.68	A	0.14	0.83	7.17	0.57	0.17	0.649
		B1	0.49	0.58	1.00	0.60	0.63	0.966
		B2	0.47	0.62	1.22	0.51	0.54	0.952
		C	0.36	0.70	1.98	0.39	0.52	0.837
C ₂ H ₆	1.89	A	1.15	1.05	0.51	0.78	0.80	0.998
		B1	1.96	0.76	0.16	0.81	0.98	0.999
		B2	1.78	0.76	0.17	0.78	0.97	1.000
		C	1.71	0.75	0.17	0.79	0.97	0.999
C ₂ H ₄	0.66	A	0.26	0.87	4.43	0.59	0.36	0.751
		B1	0.39	0.74	2.15	0.65	0.39	0.713
		B2	0.61	0.62	0.97	0.73	0.64	0.978
		C	0.56	0.66	1.18	0.69	0.65	0.980
BEN	0.47	A	0.44	0.38	0.69	0.65	0.77	0.995
		B1	0.57	0.42	0.65	0.70	0.71	0.966
		B2	0.47	0.35	0.55	0.71	0.81	0.994
		C	0.46	0.36	0.59	0.69	0.80	0.997

Table 9. Scores at the Kollumerwaard station (for the year 2006) from the comparison of the observations and the simulated concentrations for three simulations: case A, case B2 and case C. The means and the RMSEs are in $\mu\text{g m}^{-3}$. Bold numbers correspond to the best agreement with observations.

species	case	\bar{O}	\bar{C}	RMSE	NMSE	R	FA ₂	FA ₅
C ₃ H ₈	A		0.82	2.12	3.24	0.05	0.39	0.799
	B2	1.69	0.97	1.81	2.08	0.43	0.63	0.936
	C		0.81	1.87	2.56	0.47	0.56	0.932
IBUT	A		0.44	0.33	1.32	0.54	0.60	0.868
	B2	0.33	0.41	0.39	1.19	0.34	0.60	0.927
	C		0.36	0.34	0.98	0.37	0.62	0.948
NPEN	A		0.28	0.42	1.53	0.52	0.50	0.899
	B2	0.42	0.18	0.48	2.99	0.48	0.45	0.859
	C		0.23	0.44	2.00	0.53	0.52	0.917
IPEN	A		0.43	0.59	1.32	0.20	0.48	0.868
	B2	0.61	0.37	0.63	2.04	0.13	0.41	0.770
	C		0.33	0.65	1.75	0.12	0.36	0.814
C ₃ H ₆	A		0.08	0.30	6.10	0.65	0.36	0.719
	B2	0.17	0.14	0.26	2.79	0.63	0.44	0.780
	C		0.13	0.26	2.97	0.66	0.39	0.734
TOLU	A		0.52	0.43	0.68	0.48	0.67	0.886
	B2	0.53	0.35	0.35	0.64	0.72	0.61	0.937
	C		0.39	0.33	0.52	0.73	0.67	0.942
OXYL	A		0.06	0.10	1.51	0.52	0.39	0.847
	B2	0.11	0.05	0.10	1.76	0.59	0.36	0.818
	C		0.06	0.09	1.58	0.60	0.38	0.829
MPXYL	A		0.15	0.22	1.44	0.72	0.39	0.881
	B2	0.22	0.11	0.23	2.26	0.75	0.36	0.875
	C		0.14	0.21	1.48	0.75	0.38	0.888
ISO	A		0.05	0.35	14.3	0.33	0.19	0.445
	B2	0.18	0.05	0.35	13.29	0.32	0.20	0.463
	C		0.05	0.35	13.29	0.32	0.20	0.463
C ₂ H ₄	A		0.14	0.44	4.33	0.57	0.24	0.609
	B2	0.18	0.17	0.45	3.58	0.44	0.33	0.691
	C		0.20	0.40	2.42	0.55	0.31	0.662
BEN	A		0.54	0.83	2.92	0.00	0.48	0.772
	B2	0.44	0.24	0.67	4.26	0.23	0.43	0.816
	C		0.30	0.65	3.25	0.27	0.52	0.869

4.4 Uncertainty of the retrieved emissions

Although the forecast and cross-validation tests are sufficient to validate the added value of the retrieved emissions, they do not yield the expected uncertainty of these fields. To do so, one approach would consist in performing sensitivity analysis of the retrieved emissions on several factors such as boundary conditions, oxidant species or transport model. However, that would require knowing the uncertainty of those influential factors, which are not always available for VOC, or that would be difficult to estimate, for instance, for the boundary conditions (Sect. 3.2). Another approach relies on the fact that prior errors were objectively assessed via the estimation of two hyperparameters (Sect. 2.5). It is then possible to estimate the posterior uncertainty. In the B1 case, the uncertainty is given by the a posteriori error covariance ma-

trix of Eq. (15). Due to the positivity constraint, there is no analytical formula in cases B2 and C for the uncertainty of the retrieved emissions. However, it can be computed using a Monte Carlo approach. One first draws a new set of observations and a new emission background according to their prior statistics. Then an inversion is performed and constitutes one draw of the Monte Carlo analysis. A large number of Monte Carlo draws are performed until convergence of the statistics, for instance the variance of the emission field.

In a first experiment, using these Monte Carlo statistics, the standard deviation of the total mass of the emission as well as the arithmetic and quadratic mean of the standard deviation of a flux in a grid cell are computed. From experimenting, the number of draws necessary to reach an acceptable convergence for the total mass and its standard deviation is 300. Because, for most VOCs, the correction does not extend far from the EMEP stations, the retrieved emission relaxes to the background away from them, such that the resulting total mass uncertainty is small. Besides, it does not mean that the estimated uncertainty cannot be locally large. Therefore, in this first experiment, we only perturb the observations, so the uncertainty that is computed is a measure of the uncertainty of the correction to the background, assuming the background is reliable wherever the observation does not lead to a correction. For each VOC species, the resulting standard deviation of the total mass is either normalised by the total retrieved mass or by the absolute difference between the retrieved and the background emission total mass. The second normalisation is more enlightening as it tells how reliable the correction to the background due to the assimilation of observations is. Moreover, the arithmetic and quadratic means of the standard deviation of individual emission fluxes averaged over all grid cells are also computed, and show the expected much stronger uncertainty of individual fluxes. The quadratic mean strengthens even more the heterogeneity of the local correction uncertainties.

Table 10 gives these standard deviations in case C and for each VOC. Each of these uncertainties needs to be interpreted with care. Let us look at three species. The correction to the total emitted mass of isoprene (ISO) has a large uncertainty of 30 %, which is consistent with the findings of this study. The correction to the total emitted mass of acetylene (ACE) has an uncertainty of 11 %, but the possible large bias that was deemed possible in the ACE because of a possible bias in the boundary conditions case cannot be diagnosed here because our methodology does not handle biases (i.e. systematic errors). Therefore, this value of 11 % does not represent the possible bias in the boundary condition (but it does account for errors in the boundary conditions without bias). Since these corrections extend far away from the stations, the mean standard deviation of individual emission fluxes is as large as 111 %. The correction to the benzene (BEN) total emission mass is 42 %, which although large is an estimation relative to the small correction brought about by the assimilation of EMEP observations.

Table 10. Monte Carlo estimation of the uncertainty of the correction to the EMEP emission inventory yielded by data assimilation in case C. M_{EMEP} : the total 2005 emitted mass for EMEP inventory (in Tg), $\langle M_C \rangle$: the total mean reconstructed emitted mass (in Tg); $\frac{\text{std}(M_C)}{\langle M_C \rangle}$: the standard deviation (std.) of the reconstructed mass normalised by the mean reconstructed emitted mass; $\frac{\text{std}(M_C)}{|\langle M_C \rangle - M_{\text{EMEP}}|}$: the standard deviation of the reconstructed mass normalised by the difference in absolute value between the reconstructed mass and the EMEP emission mass; $\frac{\langle \text{std}(m_C) \rangle}{\langle m_C \rangle}$: the arithmetic mean of the standard deviations of the individual emission fluxes normalised by the mean emitted mass in a grid cell; $\frac{\sqrt{\langle \text{var}(m_C) \rangle}}{\langle m_C \rangle}$: the quadratic mean of the standard deviations of the individual emission fluxes, normalised by the mean emitted mass.

Species	M_{EMEP}	$\langle M_C \rangle$	$\frac{\text{std}(M_C)}{\langle M_C \rangle}$	$\frac{\text{std}(M_C)}{ \langle M_C \rangle - M_{\text{EMEP}} }$	$\frac{\langle \text{std}(m_C) \rangle}{\langle m_C \rangle}$	$\frac{\sqrt{\langle \text{var}(m_C) \rangle}}{\langle m_C \rangle}$
C ₃ H ₈	0.00736	0.01510	0.033	0.064	0.217	0.628
NBUT	0.01901	0.01798	0.010	0.174	0.057	0.191
IBUT	0.00465	0.00606	0.031	0.133	0.155	0.562
NPEN	0.00971	0.00894	0.011	0.132	0.068	0.195
IPEN	0.00917	0.01018	0.021	0.218	0.137	0.446
C ₃ H ₆	0.00413	0.00544	0.026	0.107	0.153	0.477
TOLU	0.01268	0.01210	0.008	0.177	0.042	0.154
OXYL	0.00255	0.00253	0.016	1.883	0.070	0.286
MXYL	0.00572	0.00541	0.005	0.092	0.023	0.089
PXYL	0.00226	0.00222	0.002	0.123	0.009	0.038
ISO	0.16502	0.14335	0.046	0.304	0.083	0.490
ACE	0.00209	0.00483	0.062	0.110	0.282	1.105
C ₂ H ₆	0.00720	0.01726	0.026	0.045	0.206	0.539
C ₂ H ₄	0.00790	0.01456	0.029	0.063	0.198	0.572
BEN	0.00555	0.00584	0.014	0.291	0.070	0.230

In a second experiment, a map of the gridded flux standard deviations obtained from the Monte Carlo analysis is displayed. As opposed to the previous experiment, both the background and the observations are perturbed according to their diagnosed statistics, such that an absolute uncertainty of the estimated emission fluxes is assessed instead of the uncertainty on the correction. Furthermore, the number of draws is chosen to be 3×10^4 instead of 300 because one needs to obtain convergence of the variance in each grid cell. As examples, two maps are displayed in Fig. 7: for propane (C₃H₈) and isobutane (IBUT) chosen among the species of Fig. 6. Far from the stations, the uncertainty relaxes to some finite value driven by the uncertainty of the background, hence the value of hyperparameter m_s^+ . Had we only perturbed the observations, the uncertainty would have relaxed to 0 far from the stations. As expected, close to the stations, the uncertainty decreases. As for the inverse modelling corrections, the spatial distribution of the uncertainty also strongly depends on the lifetime of the species.

Finally, let us mention that unaccounted model errors make these uncertainty estimations overly optimistic. However, the estimation of the diagonal part of the observation error covariance matrix \mathbf{R}_s performed in this study partially accounts for model errors, so this may alleviate the issue.

5 Conclusions

The goal of this study was to estimate the emission inventories of 15 VOC species using ground-based in situ measurements. The concentration observations at 11 stations from the EMEP network over western Europe were assimilated to perform inverse modelling of the emission field for each of the 15 species for the year 2005.

For that purpose, the Jacobian matrix – i.e. the source–receptor relationship – was built using the POLAIR3D CTM. To compute that matrix, a fast version of this CTM as well as its validated approximate adjoint model have been developed. The chemistry module of this fast version only includes the chemical reactions between the VOC species and three oxidants (OH, NO₃ and O₃), the concentrations of which are pre-computed with the full CTM.

For each species and each grid cell, a scaling factor that multiplies the local EMEP emission flux is computed. The uncertainty attached to the prior scaling factors and the covariance matrix of the observation errors, which are crucial statistical components of the inversion, are obtained using the maximum-likelihood principle. The principle was implemented using two different assumptions: (1) the errors attached to the scaling factors follow a Gaussian pdf or (2) the scaling factor follows a truncated Gaussian pdf.

In the Gaussian case, the simulated concentrations for the year 2005 using the corrected emissions lead to a significant improvement in most statistical indicators. However, the fact that the VOC fluxes are positive is not statistically accounted

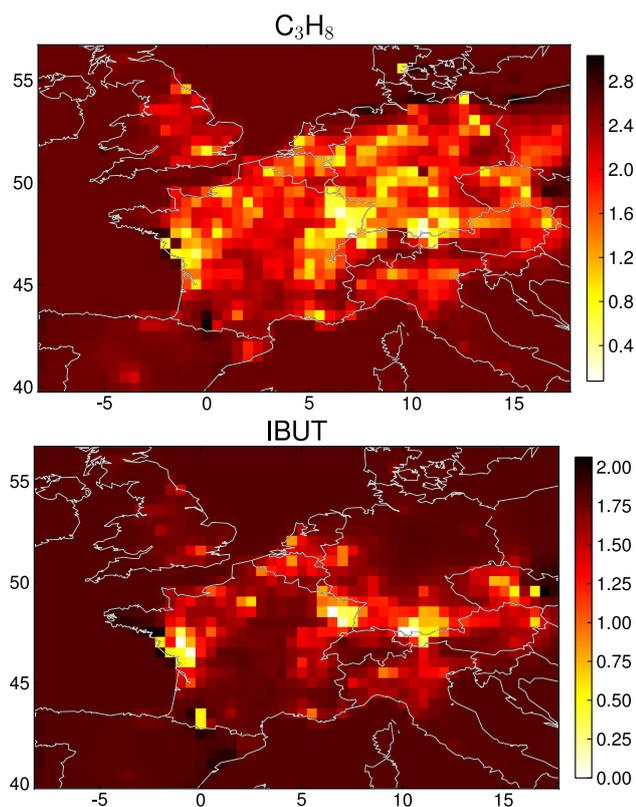


Fig. 7. Plot of the gridded unitless standard deviations of the scaling variables α that represent a posteriori uncertainties of the gridded emission fluxes of propane (C_3H_8 , left) and isobutane (IBUT, right), in case C.

for, and this is shown to lead to a probable over-fitting to the observations and to over-corrections of the EMEP emissions. Using a fully consistent truncated Gaussian assumption for the emission fluxes, including the use of a non-Gaussian likelihood for the estimation of the hyperparameters, the corrections are significantly smaller.

For short-lived species, it is shown that information cannot propagate far from the monitoring stations, such that the corrections are rather local to the stations. That is why we deem the isoprene inversion to be unreliable. That is a typical case where remote sensing assimilation is necessary to offer a satisfying coverage.

Long-lived species modelling may be impacted by bias in the boundary conditions. We found that with a lifetime of 4 months and highly uncertain boundary conditions, the inversion of the acetylene emission should be interpreted cautiously as the inversion may try to correct for a bias in the boundary conditions. In such circumstances, the inversion of the boundary conditions, in addition to the emissions, may be required (Roustan and Bocquet, 2006).

The corrected emissions have been partly validated thanks to a forecast conducted for the year 2006 using independent observations. The simulations using the corrected emissions

often led to significant improvements in the statistical indicators. Considering all statistical indicators, the fully consistent truncated Gaussian approach emerged as the best approach from this test.

The 2006 forecasts have also been compared to the observations at the Kollumerwaard station in the Netherlands. The Kollumerwaard station is not part of the 11 stations used in the analysis of 2005. Even though this station is far away from the 2005 network and its surroundings have emission fluxes that are little affected by the 2005 analysis, some improvements are noticed for several long-lived VOC species using the statistically consistent positively constrained inversion.

The estimation of the uncertainty of the retrievals has been discussed. Using a Monte Carlo technique, an objective estimation of the uncertainty of the retrieval was provided for each species. However, this estimation does not account for possible systematic errors in the boundary conditions.

For anthropogenic VOC (alkanes, alkenes, acetylene and aromatics), the inversion modelling suggests that the current EMEP emission inventories are correct within 33% on average for n-butane, isobutane, n-pentane, isopentane, propylene and aromatics (benzene, toluene and xylenes). However, there are large discrepancies (about 1 order of magnitude) in specific regions. The emission inventories appear to be significantly underestimated (about a factor of 2) for ethane, n-propane, ethylene and acetylene. Revisions to VOC emission inventories appear to be warranted and experimental investigations of speciated VOC emissions from major source categories (e.g. on-road traffic, petrochemical industry) are recommended.

As an extension to this study, we have tried but failed to disaggregate the inversion among the different emissions sectors in this study using a positive matrix factorisation approach. The chemical profiles of the sectors in regard to the considered species were too collinear to get a robust segregation.

It would be of interest to assess the method potential efficiency in more specific situations. The methodology developed here could be applied for specific source categories using tracer measurements such as levoglucosan for biomass burning. Its efficiency could also be investigated for short-lived species but for local emission inventories – for instance at the urban scale. Of course, this extension of the method implies the availability of a large and consistent set of measurements of specific species.

Acknowledgements. We would like to thank three reviewers for their thorough comments as well as for their useful suggestions on improving the manuscript. This paper is a contribution to the CARBOSOR project, which is sponsored by the Primequal research programme of the French Ministry of Ecology and of ADEME, the French Agency of Environmental Protection and Energy Management, and a contribution to the ADOMOCA-2 project, which is sponsored by the INSU/LEFE council. The

authors acknowledge the useful pieces of advice from Stéphane Sauvage on the EMEP data and inventory, and the support of Agnès Borbon, principal investigator of the CARBOSOR project.

Edited by: J. West

References

- Bessagnet, B., Menut, L., Curci, G., Hodzic, A., Guillaume, B., Lioussé, C., Moukhtar, S., Pun, B., C., S., and Schulz, M.: Regional modeling of carbonaceous aerosols over Europe-focus on secondary organic aerosols, *J. Atmos. Chem.*, 61, 175–202, 2008.
- Bocquet, M.: Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory, *Q. J. Roy. Meteor. Soc.*, 131, 2191–2208, 2005.
- Bocquet, M.: Parameter field estimation for atmospheric dispersion: Application to the Chernobyl accident using 4D-Var, *Q. J. Roy. Meteor. Soc.*, 138, 664–681, doi:10.1002/qj.961, 2012.
- Byrd, R. H., Lu, P., and Nocedal, J.: A Limited Memory Algorithm for Bound Constrained Optimization, *SIAM Journal on Scientific and Statistical Computing*, 16, 1190–1208, 1995.
- Chapnik, B., Desroziers, G., Rabier, F., and Talagrand, O.: Properties and first application of an error-statistics tuning method in variational assimilation, *Q. J. Roy. Meteor. Soc.*, 130, 2253–2275, 2006.
- Curci, G., Palmer, P. I., Kurosu, T. P., Chance, K., and Visconti, G.: Estimating European volatile organic compound emissions using satellite observations of formaldehyde from the Ozone Monitoring Instrument, *Atmos. Chem. Phys.*, 10, 11501–11517, doi:10.5194/acp-10-11501-2010, 2010.
- Davoine, X. and Bocquet, M.: Inverse modelling-based reconstruction of the Chernobyl source term available for long-range transport, *Atmos. Chem. Phys.*, 7, 1549–1564, doi:10.5194/acp-7-1549-2007, 2007.
- Desroziers, G. and Ivanov, S.: Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation, *Q. J. Roy. Meteor. Soc.*, 127, 1433–1452, 2001.
- Dufour, G., Wittrock, F., Camredon, M., Beekmann, M., Richter, A., Aumont, B., and Burrows, J. P.: SCIAMACHY formaldehyde observations: constraint for isoprene emission estimates over Europe?, *Atmos. Chem. Phys.*, 9, 1647–1664, doi:10.5194/acp-9-1647-2009, 2009.
- Elbern, H., Strunk, A., Schmidt, H., and Talagrand, O.: Emission rate and chemical state estimation by 4-dimensional variational inversion, *Atmos. Chem. Phys.*, 7, 3749–3769, 2007, <http://www.atmos-chem-phys.net/7/3749/2007/>.
- Fagerli, H., Gauss, M., Jonson, J., Nyíri, Á., Simpson, D., Tarasón, L., Tsyro, S., and Wind, P.: Transboundary acidification, eutrophication and ground level ozone in Europe. Part I: Unified EMEP model description, Tech. rep., EMEP, 2008.
- Fu, T.-M., Jacob, D., Palmer, P., Chance, K., Wang, Y., Barletta, B., Blake, D., Stanton, J., and Pilling, M.: Space-based formaldehyde measurements as constraints on volatile organic compound emissions in east and south Asia and implications for ozone, *J. Geophys. Res.*, 112, D06312, doi:10.1029/2006JD007853, 2007.
- GENEMIS: Generation of European Emission Data for Episodes (GENEMIS) project, Tech. rep., EUROTRAC annual report 1993, Garmisch-Partenkirchen, Germany, 1994.
- Goliff, W. S., Stockwell, W. R., and Lawson, C. V.: The regional atmospheric chemistry mechanism, version 2, *Atmos. Environ.*, 68, 174–185, 2013.
- Guenther, A., Geron, C., Pierce, T., Lamb, B., Harley, P., and Fall, R.: Natural emissions of non-methane volatile organic compounds, carbon monoxide, and oxides of nitrogen from North America, *Atmos. Environ.*, 34, 2205–2230, 2000.
- Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C.: Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), *Atmos. Chem. Phys.*, 6, 3181–3210, doi:10.5194/acp-6-3181-2006, 2006.
- Harley, R. and Cass, G.: Modeling the atmospheric concentrations of individual volatile organic compounds, *Atmos. Environ.*, 29, 905–922, 1995.
- Ho, K., Ho, S., Cheng, Y., Lee, S., and Yu, J.: Real-world emission factors of fifteen carbonyl compounds measured in a Hong Kong tunnel, *Atmos. Environ.*, 41, 1747–1758, 2007.
- Hopkins, J. R., Evans, M. J., Lee, J. D., Lewis, A. C., H Marsham, J., McQuaid, J. B., Parker, D. J., Stewart, D. J., Reeves, C. E., and Purvis, R. M.: Direct estimates of emissions from the megacity of Lagos, *Atmos. Chem. Phys.*, 9, 8471–8477, doi:10.5194/acp-9-8471-2009, 2009.
- Horowitz, L. W., Walters, S., Mauzerall, D. L., Emmons, L. K., Rasch, P. J., Granier, C., Tie, X., Lamarque, J. F., Schultz, M. G., Tyndall, G. S., Orlando, J. J., and Brasseur, G. P.: A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version 2, *J. Geophys. Res.*, 108, 4784, doi:10.1029/2002JD002853, 2003.
- Kim, Y., Sartelet, K., and Seigneur, C.: Comparison of two gas-phase chemical kinetic mechanisms of ozone formation over Europe, *J. Atmos. Chem.*, 62, 89–119, 2009.
- Kim, Y., Sartelet, K., and Seigneur, C.: Formation of secondary aerosols over Europe: comparison of two gas-phase chemical mechanisms, *Atmos. Chem. Phys.*, 11, 583–598, doi:10.5194/acp-11-583-2011, 2011.
- Kirchstetter, T. W., Singer, B. C., and Harley, R. A.: Impact of oxygenated gasoline use on California light-duty vehicle emissions, *Environ. Sci. Technol.*, 30, 661–670, 1996.
- Koohkan, M. R. and Bocquet, M.: Accounting for representativeness errors in the inversion of atmospheric constituent emissions: Application to the retrieval of regional carbon monoxide fluxes., *Tellus B*, 64, 19047, 2012.
- Liu, Z., Wang, Y., Vrekoussis, M., Richter, A., Wittrock, F., Burrows, J. P. P., Shao, M., Chang, C. C., Liu, S. C., Wang, H., and Chen, C.: Exploring the missing source of glyoxal (CHOCHO) over China, *Geophys. Res. Lett.*, 0, 0–0, doi:10.1029/2012GL051645, in press, 2013.
- Mallet, V., Quélo, D., Sportisse, B., Ahmed de Biasi, M., Debry, É., Korsakissok, I., Wu, L., Roustan, Y., Sartelet, K., Tombette, M., and Foudhil, H.: Technical Note: The air quality modeling system Polyphemus, *Atmos. Chem. Phys.*, 7, 5479–5487, doi:10.5194/acp-7-5479-2007, 2007.
- Marchuk, G. I.: Numerical Solution of the Problems of the Dynamics of the Atmosphere and Ocean, *Gigrometeoizdat*, in Russian, 1974.
- Millet, D., Jacob, D., Boersma, K., Fu, T.-M., Kurosu, T., Chance, K., Heald, C., and Guenther, A.: Spatial distribution of isoprene emissions from North America derived from formaldehyde col-

- umn measurements by the OMI satellite sensor, *J. Geophys. Res.*, 113, D02307, doi:10.1029/2007JD008950, 2008.
- Passant, N. R.: Speciation of UK Emissions of NMVOC, Tech. Rep. AEAT/ENV/0545, AEA Technology, 2002.
- Penkett, S. A., J., B. N., Lightman, P., Marsh, A. R. W., Anwyl, P., and Butcher, G.: The Seasonal Variation of Nonmethane Hydrocarbons in the Free Troposphere Over the North Atlantic Ocean: Possible Evidence for Extensive Reaction of Hydrocarbons With the Nitrate Radical, *J. Geophys. Res.*, 98, 2865–2885, 1993.
- Pierson, W. R., Gerlter, A. W., and Bradow, R. L.: Comparison of the SCAQS tunnel study with other on-road vehicle emission data, *J. Air Waste Manage. Assoc.*, 40, 1495–1504, 1990.
- Pun, B., Wu, S.-Y., Seigneur, C., Seinfeld, J. H., Griffin, R. J., and Pandis, S. N.: Uncertainties in modeling secondary organic aerosols: three-dimensional modeling studies in Nashville/Western Tennessee, *Environ. Sci. Technol.*, 37, 3647–3661, 2003.
- Quélo, D., Mallet, V., and Sportisse, B.: Inverse modeling of NO_x emissions at regional scale over northern France: Preliminary investigation of the second-order sensitivity, *J. Geophys. Res.*, 110, D24310, doi:10.1029/2005JD006151, 2005.
- Roustan, Y. and Bocquet, M.: Inverse modelling for mercury over Europe, *Atmos. Chem. Phys.*, 6, 3085–3098, 2006, <http://www.atmos-chem-phys.net/6/3085/2006/>.
- Rudolph, J. and Ehhalt, D.: Measurements of C2-C5 hydrocarbons over the North Atlantic, *J. Geophys. Res.*, 86, 11959–11964, 1981.
- Rudolph, J. and Johnen, F. J.: Measurements of Light Atmospheric Hydrocarbons Over the Atlantic In Regions of Low Biological Activity, *J. Geophys. Res.*, 95, 20583–20591, 1990.
- Sartelet, K., Debry, E., Fahey, K., Roustan, Y., Tombette, M., and Sportisse, B.: Simulation of aerosols and gas-phase species over Europe with the Polyphemus system. Part I: model-to-data comparison for 2001, *Atmos. Environ.*, 41, 6116–6131, 2007.
- Sartelet, K., Couvidat, F., Seigneur, C., and Roustan, Y.: Impact of biogenic emissions on air quality over Europe and North America, *Atmos. Environ.*, 53, 131–141, 2012.
- Sauvage, S.: Origine et comportement des composés hydrocarbonés non-méthaniques (HCNM) en zone rurale, Ph.D. thesis, Université des Sciences et Technologies de Lille, France, in French, 2008.
- Sawyer, R., Harley, R., Cadle, S., Norbeck, J., Slott, R., and Bravo, H.: Mobile sources critical review: 1998 NARSTO assessment, *Atmos. Environ.*, 34, 2161–2181, 2000.
- Shim, C., Wang, Y., Choi, Y., Palmer, P., Abbot, D., and Chance, K.: Constraining global isoprene emissions with Global Ozone Monitoring Experiment (GOME) formaldehyde column measurements, *J. Geophys. Res.*, 110, D24301, doi:10.1029/2004JD005629, 2005.
- Simpson, D., Winiwarter, W., Börjesson, G., Cinderby, S., Ferreira, A., Guenther, A., Hewitt, C., Janson, R., Khalil, M., Owen, S., Pierce, T., Puxbaum, H., Shearer, M., Skiba, U., Steinbrecher, R., Tarrason, L., and Oquist, M.: Inventorying emissions from nature in Europe, *J. Geophys. Res.*, 104, 8113–8152, 1999.
- Staehelin, J., Keller, C., Stahel, W., Schläpfer, K., and Wunderli, S.: Emission factors from road traffic from a tunnel study (Gubris tunnel, Switzerland), Part III: results of organic compounds, SO₂ and speciation of organic exhaust emission, *Atmos. Environ.*, 32, 999–1009, 1998.
- Stavrakou, T., Müller, J.-F., De Smedt, I., Van Roozendaal, M., Kanakidou, M., Vrekoussis, M., Wittrock, F., Richter, A., and Burrows, J. P.: The continental source of glyoxal estimated by the synergistic use of spaceborne measurements and inverse modelling, *Atmos. Chem. Phys.*, 9, 8431–8446, doi:10.5194/acp-9-8431-2009, 2009.
- Stemmler, K., Bugmann, S., Buchmann, B., Reimann, S., and Staehelin, J.: Large decrease of VOC emissions of Switzerland's car fleet during the past decade: results from a highway tunnel study, *Atmos. Environ.*, 39, 1009–1018, doi:10.1016/j.atmosenv.2004.10.010, 2005.
- Tarrasón, L., Fagerli, H., Jonson, J., Simpson, D., Benedictow, A., Klein, H., and Vestreng, V.: Transboundary acidification, eutrophication and ground level ozone in Europe in 2005, Tech. rep., EMEP, 2007.
- Touaty, M. and Bonsang, B.: Hydrocarbon emissions in a highway tunnel in the Paris area, *Atmos. Environ.*, 34, 985–996, doi:10.1016/S1352-2310(99)00195-8, 2000.
- Uliasz, M.: Application of the Perturbation Theory to the Sensitivity Analysis of an Air Pollution Model, *Z. Meteor.*, 33, 355–362, 1983.
- Vijayaraghavan, K., Snell, H., and Seigneur, C.: Practical aspects of using satellite data in air quality modeling, *Environ. Sci. Technol.*, 42, 8187–8192, 2008.
- Warneke, C., de Gouw, J. A., Del Negro, L., Brioude, J., McKeen, S., Stark, H., Kuster, W. C., Goldan, P. D., Trainer, M., Fehsenfeld, F. C., Wiedinmyer, C., Guenther, A. B., Hansel, A., Wisthaler, A., Atlas, E., Holloway, J. S., Ryerson, T. B., Peischl, J., Huey, L. G., and Case Hanks, A. T.: Biogenic emission measurement and inventories determination of biogenic emissions in the eastern United States and Texas and comparison with biogenic emission inventories, *J. Geophys. Res.*, 115, D00F18, doi:10.1029/2009JD012445, 2010.
- Winiarek, V., Bocquet, M., Saunier, O., and Mathieu, A.: Estimation of Errors in the Inverse Modeling of Accidental Release of Atmospheric Pollutant: Application to the Reconstruction of the Cesium-137 and Iodine-131 Source Terms from the Fukushima Daiichi Power Plant, *J. Geophys. Res.*, 117, D05122, doi:10.1029/2011JD016932, 2012.
- Xiao, Y., Logan, J., Jacob, D., Hudman, R., Yantosca, R., and Blake, D.: Global budget of ethane and regional constraints on U.S. sources, *J. Geophys. Res.*, 113, D21306, doi:10.1029/2007JD009415, 2008.