DATA ASSIMILATION
AND NUMERICAL
MODELLING OF ATMOSPHERIC
COMPOSITION

JULIUS VIRA
Data assimilation and numerical modelling of atmospheric composition

Julius Vira

Academic dissertation in physics

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in the auditorium "Brainstorm" of Finnish Meteorological Institute (Erik Palménin aukio 1, Helsinki) on February 23th, 2017 at 12 o’clock noon.

Finnish Meteorological Institute
Helsinki, 2017
Atmospheric chemistry and transport models are used for a wide range of applications which include predicting dispersion of a hazardous pollutants, forecasting regional air quality, and modelling global distribution of aerosols and reactive gases. However, any such prediction is uncertain due to inaccuracies in input data, model parametrisations and lack of resolution. This thesis studies methods for integrating remote sensing and in-situ observations into atmospheric chemistry models with the aim of improving the predictions.

Techniques of data assimilation, originally developed for numerical weather prediction, are evaluated for improving regional-scale predictions in two forecast experiments, one targeting the photochemical pollutants ozone (O₃) and nitrogen dioxide (NO₂), the other targeting sulphur dioxide (SO₂). In both cases, assimilation of surface-based air quality monitoring data is found to initially improve the forecast when assessed on monitoring stations not used in assimilation. However, as the forecast length increased, the forecast converged towards the reference simulations where no data assimilation was used. The relaxation time was 6-12 hours for SO₂ and NO₂ and about 24 hours for O₃.

An alternative assimilation scheme was tested for SO₂. In addition to the initial state of the forecast, the scheme adjusted the gridded emission fluxes based on the observations within the last 24 hours. The improvements due to adjustment of emissions were generally small but, where observed, the improvements persisted throughout the 48 hour forecast.

The assimilation scheme was further adapted for estimating emission fluxes in volcanic eruptions. Assimilating retrievals of the Infrared Atmospheric Sounding Interferometer (IASI) instrument allowed reconstructing both the vertical and horizontal profile of SO₂ emissions during the 2010 eruption of Eyjafjallajökull in Iceland. As a novel feature, retrievals of plume height were assimilated in addition to the commonly used column density retrievals. The results for Eyjafjallajökull show that the plume height retrievals provide a useful additional constraint in conditions where the vertical distribution would otherwise remain ambiguous.

Finally, the thesis presents a rigorous description and evaluation of a numerical scheme for solving the advection equation. The scheme conserves tracer mass and non-negativity, and is therefore suitable for regional and global atmospheric chemistry models. The scheme is particularly adapted for handling discontinuous solutions; for smooth solutions, the scheme is nevertheless found to perform comparably to other state-of-art schemes used in atmospheric models.
Acknowledgements

This thesis consists of research carried out during 2008–2016 in the Finnish Meteorological Institute. I thank my doctoral advisor, Dr. Mikhail Sofiev for introducing me to the atmospheric sciences and data assimilation, and I am grateful to my supervisors Dr. Ari Karppinen, Prof. Jaakko Kukkonen and Prof. Heikki Lihavainen at the Atmospheric Composition unit for their support. I also acknowledge the funding from Academy of Finland, the European Commission, European Space Agency and NordForsk which made this work possible.

I am honoured to have Dr. Adrian Simmons as my opponent. I thank my custos, Prof. Veli-Matti Kerminen for his support and guidance, and my pre-examiners Dr. Joakim Langner and Prof. Wouter Peters for their thorough and detailed reviews. I thank my coauthors for fruitful collaboration, and my colleagues at FMI for all the discussions.

Finally, I am grateful to my family for their continuing support. And to Marje, for everything.
Contents

List of publications 6

Review of the papers and the author’s contribution 7

1 Introduction 8

2 Numerical modelling of pollutant transport 10

3 Inverse problems and data assimilation 15
   3.1 Theoretical background 15
   3.2 Applications in atmospheric chemistry and dispersion modelling 19

4 Model setup and input data 21
   4.1 The SILAM model 21
   4.2 Emissions, boundary conditions and meteorology 21
   4.3 Observational data 22

5 Results and discussion 23
   5.1 Assimilation of air quality monitoring data 23
   5.2 Volcanic source term inversion using satellite data 27
   5.3 Numerical solution of the advection equation 30

6 Conclusions and future work 33

References 35
List of publications

This thesis consists of an introductory review followed by 4 research articles. In the introductory review, the papers are cited according to their roman numerals.


Review of the papers and the author’s contribution

Paper I discusses an assimilation experiment targeting the sulphur dioxide in central and southern Europe. The conventional assimilation approach where the prognostic variables (concentration in air) are updated was compared to a scheme where the emission forcing was refined in addition to the concentration in air. I wrote the assimilation code, designed and carried out the assimilation experiments and analysed the results. The paper was written in collaboration with Mikhail Sofiev.

Paper II presents a scheme for assimilating surface observations of ozone and nitrogen dioxide, and evaluates its performance in both forecast and reanalysis applications. I implemented the assimilation method, ran the simulations, analysed the results and wrote the paper.

Paper III extends the 4D-Var assimilation method into inverse modeling of volcanic emissions. As an application, the sulphur dioxide emissions in the 2010 eruption of Eyjafjallajökull were reconstructed using data from the IASI satellite instrument. I formulated and implemented the inversion method, ran the simulations, analysed the results and wrote the paper with the exception of section describing the satellite retrievals.

Paper IV evaluates an numerical advection scheme and describes its coupling to the SILAM chemistry transport model. I devised a modification to the original scheme, which significantly improved its performance for realistic flows. I also outlined its theoretical description (Section 2.2), performed the numerical tests discussed in Section 3.5 and contributed to the manuscript preparation.
1 Introduction

The success of numerical weather prediction, along with the increasing computing capability, has made numerical simulations feasible for diverse scientific and practical problems related to the atmospheric composition. Predicting dispersion of a hazardous pollutant, modelling regional air quality or simulating global distribution of aerosols and reactive gases are typical applications for the current chemistry transport models.

However, any prediction given by an atmospheric model is uncertain. In weather prediction, much of the uncertainty has been attributed to errors in the initial state (Magnusson and Källén, 2013). This is a consequence of the nonlinear dynamics (Lorenz, 1963, 1982) that govern the atmospheric flow, and thus, a skillful weather forecast needs to be initialised using current observations in the process known as data assimilation, defined by Talagrand (1997) as “the process through which all the available information is used in order to estimate as accurately as possible the state of the atmospheric or oceanic flow”.

This definition describes the data assimilation from the perspective of numerical weather prediction. In that sense, the first applications of data assimilation in tropospheric chemistry and dispersion modelling date back to late 90s (Elber et al., 1997). Meanwhile, observations have been used in numerous inverse modelling studies aiming at estimation of emissions from point sources (see, for example, the review of Shankar Rao (2007)) or for estimation of greenhouse gas fluxes (Enting and Mansbridge, 1989; Houweling et al., 1999; Enting, 2002). The concept of data assimilation in atmospheric chemistry models is thus somewhat ambiguous. This thesis covers both inverse modelling of emissions and data assimilation in the traditional sense and explores the connection between the two.

The diversity of approaches to chemical data assimilation reflects the diversity of forecast problems in atmospheric chemistry and pollutant dispersion. Modelling the composition of polluted boundary layers is characterised by strong emission forcing and dissipation due to removal processes, and often strong chemical coupling between species. Due to the strong forcing and limited chemical lifetime of pollutants, perturbations in the initial state tend to dissipate rather than grow in time. As demonstrated by this thesis and previous studies discussed in Section 3.2, this sets a fundamental limit on the forecast improvements obtainable with data assimilation taken in its narrow definition.

Contrary to air quality modelling, simulating the dispersion of accidental releases or volcanic plumes usually involves a single, poorly known emission source. The combined uncertainty and importance of the source term makes this type of problem attractive for inverse modelling of emissions. However, the atmospheric lifetime of volcanic plumes is often longer than the lifetime of pollutants emitted near surface, and traditional data assimilation has indeed been found to improve long-range predictions of volcanic plumes (Flemming and Inness, 2013).

Inverse modelling generally requires confidence on the quality of the numerical model, which can be achieved by the appropriate choice of numerical methods and their sufficient evaluation. This aspect is addressed by the final part of this thesis, which describes the development and evaluation of a numerical scheme for
the advection equation, which forms a fundamental component in any chemistry transport model.

In summary, the main objectives of this thesis are as follows:

- To assess the impact of chemical data assimilation on short-term forecasts of gas-phase pollutants
- To evaluate a variational data assimilation scheme based on short-term emission adjustments
- To formulate a variational method for reconstructing temporal and vertical profiles of volcanic emissions using satellite retrievals
- To assess the added value of assimilating plume height retrievals for estimating the vertical profiles of sulphur emissions in explosive volcanic eruptions
- To evaluate a numerical scheme for solving the advection equation using up-to-date benchmarks.

This thesis consists of an introduction and four manuscripts. Paper I and Paper II approach data assimilation and inverse problems from the perspective of air quality forecasting; Paper III discusses inverse modelling of volcanic emissions, and Paper IV is about numerical solution the advection equation.

The introduction, which summarises the findings presented in the manuscripts, is organised as follows. Section 2 presents the theoretical background regarding numerical methods for the advection equation. Section 3 gives a brief overview of inverse problems, their connection to data assimilation, and presents a review of their recent applications in atmospheric chemistry and dispersion models. Sections 5 and 6 present the main results and conclusions of the thesis.
2 Numerical modelling of pollutant transport

Extensive efforts regarding numerical methods in atmospheric chemistry models have been devoted to treatment of the advection equation

$$\frac{\partial \varphi}{\partial t} + \nabla \cdot (\varphi \mathbf{v}) = 0,$$

(1)

where $\varphi$ denotes the species concentration, or equivalently,

$$\frac{\partial \psi}{\partial t} + \nabla \psi \cdot \mathbf{v} = 0,$$

(2)

where $\rho$ denotes air density and $\psi = \varphi/\rho$ is the mixing ratio.

Numerical solution of Eq. (1) is also the main topic of Paper IV, which describes development of the advection scheme used in the SILAM chemistry transport model. This section gives a short introduction to the numerical methods used to solve Eq. (1) in atmospheric chemistry models. The emphasis is on conservative semi-Lagrangian methods – more general reviews of advection schemes have been given by Rood (1987) and Lauritzen et al. (2014).

The advection equation (1) is a component in the system of equations solved by chemistry transport models – namely, the tracer continuity equation

$$\frac{\partial \varphi}{\partial t} + \nabla \cdot (\varphi \mathbf{v}) - \nabla \cdot \mathbf{K} \nabla \varphi = R(\varphi, x, t)$$

(3)

where $\mathbf{v}(x, y, z, t)$ is wind field, $\mathbf{K}$ is the eddy diffusivity tensor and $R$ denotes the sources and sinks. For chemically reactive species, $R$ includes reactions with other species, and in this case, (3) becomes a system of partial differential equations non-linearly coupled by the reaction term.

The equation (3) needs to be solved numerically. An almost universally used approach is the operator splitting, where the terms in Eq. (3) are solved in separate, sequential steps. This uncouples the transport equation of chemically connected constituents from each other, and enables using explicit or implicit time integration depending on the typical time scale of each term separately. The splitting can be extended to solving multidimensional operators sequentially in each dimension; this is called dimensional splitting. Operator splitting introduces additional numerical errors which have been discussed for the advection-reaction-diffusion system by Lanser (1999), Sportisse (2000) and Santillana et al. (2016).

Due to the requirement of strict mass conservation, numerical solution of (1) in atmospheric chemistry models (as reviewed by Kukkonen et al. (2012)) is most commonly based on the integral form of Eq. (1):

$$\frac{d}{dt} \int_{\Omega_i} \varphi(x) d\Omega = - \int_{\partial \Omega_i} \varphi(x) \mathbf{v} \cdot \mathbf{n} dS,$$

(4)

which is readily obtained using the divergence theorem. Here, $\Omega_i$ denotes the $i$th grid cell and $\mathbf{n}$ is the unit normal vector.
If the flux integrals on the right hand side of Eq. (4) are approximated numerically, one obtains a spatially discrete system for the cell-averaged concentration $\bar{\phi}_i$, which in one dimension can be written as

$$\frac{d\bar{\phi}_i}{dt} = \frac{1}{h}(F_{i-1/2} - F_{i+1/2}),$$

(5)

where $h$ is the cell size. The cross-boundary fluxes $F_{i\pm1/2}$ are evaluated by fitting a polynomial or other interpolant to the mean concentrations $\bar{\phi}_i$ of the neighbouring grid cells.

Due to the Courant-Friedrichs-Lewy (CFL) stability condition, schemes in the form (5) are only stable for sufficiently small Courant numbers $C = v\Delta t/h$, which usually needs to be less than one. This is especially problematic in global models using a lon-lat grid, since convergence of the meridians results in high Courant numbers near the poles.

The desire to choose the timestep based on accuracy rather than stability considerations has led to the popularity of the semi-Lagrangian advection schemes. In the semi-Lagrangian schemes, the numerical domain of dependence for each receiving grid cell is not restricted to the adjacent cells, but is instead determined by trajectory integrations. However, the classical gridpoint-based semi-Lagrangian advection methods used in weather prediction models (Staniforth and Côte, 1991) are not conservative, and thus rarely used in chemistry models.

Conservative semi-Lagrangian schemes can be constructed based on the Lagrangian form of Eq. (4),

$$\frac{d}{dt} \int_{\Omega(t)} \varphi(x) d\Omega = 0,$$

(6)

where the Lagrangian volume $\Omega$ follows the flow. The discretisation of Eq. (6) can proceed in several ways, which will be discussed here in single space dimension. A detailed presentation of conservative semi-Lagrangian schemes in two or three dimensions has been given by Lauritzen et al. (2011).

Given the concentration $\varphi(x, t_{k-1})$ at previous timestep, the mean concentration in grid cell $\Omega_i$ at the next timestep can be evaluated by integrating over the departure cell $\Omega_i(t_{k-1})$, as shown in Fig. 1a,

$$\bar{\varphi}_i = \frac{1}{|\Omega_i(t_{k-1})|} \int_{\Omega_i(t_{k-1})} \varphi(x, t_k) d\Omega,$$

(7)

where $|\Omega_i(t_{k-1})|$ denotes the volume of the departure cell, which is determined by integrating the backward trajectories from the cell borders $x_{i\pm1/2}$. The continuous concentration $\varphi(x, t)$ is unknown and needs to be evaluated from an interpolation, $\tilde{\varphi}(x)$, often called a reconstruction function, which has a similar role as the interpolation used in Eulerian schemes represented by Eq. (5). However, contrary to the flux-form schemes, the reconstruction function must be conservative:

$$\int_{\Omega_i} \tilde{\varphi}(x) dx = |\Omega_i| \bar{\varphi}_i.$$

(8)

It would be equally possible to compute forward trajectories from $\Omega_i$ and integrate over intersections of the receiving Eulerian cells and the arrival volume (Fig.
Figure 1: Semi-Lagrangian advection in one dimension: a) a backwards scheme with integration over departure cell; b) a forward scheme with distribution of mass between cells $i$ and $i+1$; c) a flux scheme with integration over swept volumes $S_{i-1/2}$ and $S_{i+1/2}$.

1b). However, the backward integration allows the reconstruction functions $\tilde{\varphi}(x)$ to be defined with respect to the fixed (Eulerian) grid, which is often simpler.

Alternatively, a flux-form scheme can be also constructed following the Lagrangian approach. This follows from the observation that the time integral of a flux through a cell interface can be represented as an integral over the volume which passes through the interface during the timestep:

$$\int_{t_k}^{t_{k+1}} F_{i-1/2}(t) dt = \int_{S_{i-1/2}} \tilde{\varphi}(x, t_k) dV.$$  \hspace{1cm} (9)

The idea is illustrated in Fig. 1c. Since the “swept volume” $S$ can span several grid cells, the scheme is not limited by the typical CFL condition. A flux-form scheme is conservative even if the reconstruction function does not satisfy Eq. (8), or if the integrals are approximated numerically.

Fully multidimensional, conservative semi-Lagrangian schemes have been described by Nair and Machenhauer (2002) and Lauritzen et al. (2010). Tracking the Lagrangian volumes $\Omega(t)$ in a multidimensional flow is considerably more complex than in the one-dimensional case. However, at least for simulations with complex chemistry, the computational cost per tracer can remain reasonable, since the departure volume needs to be evaluated only once for each cell.

A more straightforward way to extend the semi-Lagrangian computations to multiple dimensions is to use dimensional splitting as done in the flux-form semi-Lagrangian scheme of Lin and Rood (1997), and in the scheme described Paper IV. Dimensional splitting is also used in many purely Eulerian schemes (Bott, 1989; Prather, 1986; Colella and Woodward, 1984).

In addition to mass conservation, desirable qualitative features for a numerical advection include positive definiteness (preservation of non-negativity) and monotonicity with regard to the initial data. As discussed by LeVeque (1992), a prac...
Figure 2: The advection scheme of Galperin (1999): a) the continuous function $\varphi(x)$ is represented by box functions $\Pi_i$; b) the box functions are transported with the flow and new centres of mass $X_i$ and total masses $M_i$ are evaluated; c) the box functions $\Pi_i$ are updated according to the new $X_i$ and $M_i$.

A classically useful definition for the monotonicity is that for a monotonously increasing (or decreasing) initial condition, the numerical solution at a later time has to be monotonously increasing (decreasing). This property is sufficient to guarantee that the scheme does not produce spurious oscillations near an isolated discontinuity.

A classical result is the theorem of Godunov (1959), which states that a monotonicity preserving linear finite difference scheme can be at most first order accurate. It is straightforward to show that a positive definite linear scheme will unavoidably be also monotonicity preserving. Due to the excessive diffusiveness of first-order schemes, the advection schemes used in chemistry transport models almost invariably include some nonlinear elements.

The advection scheme of Galperin (1999), developed and evaluated further in Paper IV, is a forward semi-Lagrangian finite volume scheme corresponding to the case b) in Fig. 1. For each grid cell, the scheme tracks the total mass $M_i$ and centre of mass $X_i$ given by

\begin{align}
M_i &= \int_{\Omega_i} \varphi(x) dx \\
X_i &= \frac{1}{M_i} \int_{\Omega_i} x \varphi(x) dx. \tag{11}
\end{align}

The reconstruction function is a rectangular pulse shown in Fig. 2a:

\begin{equation}
\Pi^k_i(x) = \begin{cases} 
\frac{M_i}{2w}, & |x - X_i| \leq w \\
0, & \text{otherwise}
\end{cases}, \tag{12}
\end{equation}

where $w = \min(|X_i - x_{i-1/2}|, |X_i - x_{i+1/2}|)$ and $x_{i\pm1/2}$ are the cell borders. The full continuous distribution is therefore represented by the sum of $\Pi_i$:

\begin{equation}
\bar{\varphi}(x) = \sum_i \Pi_i(x). \tag{13}
\end{equation}
Both $M_i$ and $X_i$ are updated (Fig. 2 panels b and c) using $\Pi_i(x)$, whose form allows the necessary integrals to be evaluated exactly. Since $\Pi_i(x) \geq 0$ for all $x$ and $i$, the scheme conserves non-negative solutions. However, as discussed in Paper IV, the scheme is otherwise not fully monotonicity preserving.

Use of additional prognostic quantities ($X_i$) connects the scheme to Eulerian schemes of Russell and Lerner (1981) and Prather (1986). The reconstruction function given by Eq. (12) distinguishes the scheme from the otherwise similar scheme of Egan and Mahoney (1972), which does not adjust the width $w$, but instead tracks the second moments. Also, the Egan and Mahoney (1972) scheme is formulated in Eulerian context and is only stable conditionally to the Courant number.
3 Inverse problems and data assimilation

Physical models normally predict values of an observable quantity given a set of known input parameters. In mathematical literature (Tarantola, 2005), this is often referred as the forward problem. Conversely, if the input parameters are poorly known and need to be evaluated from a set of observed values, an inverse problem needs to be solved. As it turns out, the concept of inverse problems connects quite naturally with the process called data assimilation in meteorology and related fields. This section aims to offer a unified view to the two concepts.

3.1 Theoretical background

The forward problem can be formalised as

\[ y = \mathcal{H}(x), \]  

(14)

where \( x \) and \( y \) are vectors denoting the input parameters and observations which are connected by the operator \( \mathcal{H} \). In the following discussion, operators which may be nonlinear are denoted with calligraphic letters such as \( \mathcal{H} \), while the boldface letters (eg. \( \mathbf{H} \)) denote matrices and linear operators.

Since the vectors \( x \) and \( y \) usually have different dimensions, inverting Eq. (14) does generally not constitute a well-posed problem. Moreover, in many problems related to atmospheric constituent transport, features of the forward operator \( \mathcal{H} \) render techniques like ordinary least squares ill-conditioned, and to overcome this difficulty, the inverse problem needs to be modified in some sense. One approach is regularisation, where a well-posed “nearby” problem is solved with the hope of arriving at an approximate solution of the original problem. A very common example is the Tikhonov regularisation, which in the simplest form formulates the inversion as a penalised least-squares problem of minimising

\[ J(x) = \| y - \mathcal{H}(x) \|^2 + \alpha^2 \| x \|^2, \]  

(15)

where the parameter \( \alpha^2 \) controls the level of regularisation.

Alternatively, the inversion can be viewed as a Bayesian estimation problem where the ill-posedness is addressed by including additional a priori information, formulated as the prior probability distribution \( p(x) \). To quantify the uncertainty of observations, Eq. (14) is recast into a stochastic form

\[ y = \mathcal{H}(x) + \epsilon, \]  

(16)

where \( \epsilon \) denotes the observation errors. The solution of the inverse problem is then formally the probability distribution of \( x \) conditioned to a realisation of \( y \), \( p(x|y) \), as given by the Bayes theorem

\[ p(x|y) = \frac{p(x)p(y|x)}{p(y)}. \]  

(17)

In principle, the posterior distribution \( p(x|y) \) could be characterised in many ways, including moments or quantiles, but the computational cost of \( \mathcal{H} \) often limits
the choice of estimators. An option which does not require a drawing a large sample from the posterior distribution is the maximum a posteriori (MAP) estimate of \( x \) given by

\[
\mathbf{x}_{\text{MAP}} = \arg \max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}).
\]  

(18)

The MAP estimate can be evaluated using numerical optimisation methods provided that the density function \( p(\mathbf{x}|\mathbf{y}) \) is known. For some probability distributions (most notably Gaussian), \( \mathbf{x}_{\text{MAP}} \) is also equal to the conditional mean \( E(\mathbf{x}|\mathbf{y}) \).

To connect the discussion of general inverse problems with data assimilation, we next assume that \( \mathbf{x} \) in fact represents the state of a stochastic process given by

\[
\mathbf{x}_{k+1} = \mathcal{M}_k(\mathbf{x}_k) + \delta_k,
\]

(19)

where the model operator \( \mathcal{M} \) propagates the state vector between the timesteps \( k \) and \( \delta \) represents the model noise. The problem is now to estimate the state \( \mathbf{x} \) using simultaneously the observations given by Eq. (16) and knowledge of the system evolution encoded into Eq. (19).

The form of Eq. (19) suggests towards a recursive scheme for estimating the distribution of

\[
p(\mathbf{x}_k|\mathbf{y}_0...\mathbf{y}_k)
\]

conditioned to observations until step \( k \). The recursion proceeds in two steps:

1. Analysis step: given the prior (or background) distribution \( p(\mathbf{x}_k|\mathbf{y}_{0...k-1}) \) and the current observations \( \mathbf{y}_k \), characterise the posterior (or analysis) distribution \( p(\mathbf{x}_k|\mathbf{y}_{i\leq k}) \).

2. Forecast step: given \( p(\mathbf{x}_k|\mathbf{y}_{i\leq k}) \) and Eq. (19), characterise the conditional distribution of the forecast \( p(\mathbf{x}_{k+1}|\mathbf{y}_{i\leq k}) \), which becomes the prior distribution for the next analysis step.

Linear observation and model operators \( \mathcal{H} \) and \( \mathcal{M} \) combined with a Gaussian initial distribution \( p(\mathbf{x}_0) \) and noise terms \( \delta \) and \( \epsilon \) form an important special case where all the involved conditional distributions are also Gaussian. The Gaussian distributions are completely determined by their conditional means and covariances, which in turn can be evaluated algebraically. This procedure leads to the well-known Kalman filter, which at each analysis and forecast step updates both the conditional mean and covariance matrix of \( \mathbf{x}_k \).

The requirement to manipulate the covariance matrices of the model state makes the standard Kalman filter suitable only for relatively low-dimensional systems. However, the Ensemble Kalman filter (EnKF, Evensen (1994, 2003)) has proven to be a feasible option even for large-scale geophysical models. Ignoring the covariance propagation and instead using a prescribed covariance matrix for the analysis step leads to another important class of assimilation methods which, in particular, includes the variational methods (Le Dimet and Talagrand, 1986; Lorenc, 1986) used in this work.

The variational methods usually assume that the observation errors are Gaussian and unbiased, and that the background distribution \( \mathbf{x}_k|\mathbf{y}_{i<k} \) is also Gaussian:

\[
\mathbf{x}_k|\mathbf{y}_{i<k} \sim \mathcal{N}(\mathbf{x}_b, \mathbf{B})
\]

\[
\epsilon \sim \mathcal{N}(0, \mathbf{R}),
\]

(20)
Using the background state $x_b$ and the background and observation error covariance matrices $B$ and $R$, the variational methods aim to evaluate a MAP estimate for the analysis state $x_a$. The background error covariance matrix is usually not defined explicitly but instead as an operator acting on $x$, which avoids the computational difficulties associated with its high dimension.

Under the assumptions (20), the posterior probability density is maximised by minimising the cost function
\[
J_{3D}(x) = \frac{1}{2}(y - H(x))^T R^{-1}(y - H(x)) + \frac{1}{2}(x - x_b)^T B^{-1}(x - x_b).
\] (21)

Minimising the cost function $J(x)$ with an iterative, gradient-based method yields the data assimilation scheme known as the three-dimensional variational assimilation, or 3D-Var (Lorenc, 1986). The gradient of the cost function is given by
\[
\nabla J_{3D}(x) = -H^* R^{-1}(y - H(x)) + B^{-1}(x - x_b),
\] (22)
where the adjoint observation operator $H^*$ is related to the linearised observation operator $H$ by the duality relation
\[
\langle H\psi, \varphi \rangle = \langle \psi, H^* \varphi \rangle
\] (23)
for any pair of vectors $\varphi, \psi$ belonging to the observation and model space, respectively.

Algorithms which use observations to update the past in addition to the current model state are called smoothers. The most important data assimilation algorithm in this category is the four-dimensional variational assimilation (4D-Var) method. Normally, 4D-Var (Le Dimet and Talagrand, 1986) is used in the strong-constraint form which estimates the system state within an assimilation window, during which the model noise $\delta_k$ is assumed negligible. This implies that the evolution of $x_k$ is determined by the state $x_0$ in the beginning of the assimilation window. Under the assumptions of Eq. (20), this results in a cost function
\[
J_{4D}(x_0) = \frac{1}{2} \sum_{k=0}^{n} (y_k - H_k(x_k))^T R_k^{-1}(y_k - H_k(x_k)) \\
+ \frac{1}{2}(x_0 - x_b)^T B^{-1}(x_0 - x_b),
\] (24)
with $x_{k+1} = M_k(x_k)$ for each $0 \leq k < n$. By considering a first order perturbation, it is easy to show that the gradient of $J_{4D}(x)$ with respect to the initial state is given by
\[
\nabla J_{4D}(x_0) = -\sum_{k=0}^{n} M_{k,0}^* H_k^* R_k^{-1}(y_k - H(x_k)) + B^{-1}(x_0 - x_b),
\] (25)
where $M_k^*$ is the adjoint model operator defined analogously to the adjoint observation operator $H_k^*$ and $M_{k,0}^*$ is shorthand for $M_0^* M_1^* ... M_k^*$. The first term in the
The gradient is in practice obtained by backwards integration of the adjoint system

$$x_k^* = M_k^* x_{k+1}^* - H_k^* R_k^{-1} (y_k - H_k(x)).$$  \hspace{1cm} (26)$$

Compared to evaluating a 3D-Var analysis separately for each time within the assimilation window, the 4D-Var method uses the prescribed background error covariance matrix $B$ only in the beginning of the window. Within the assimilation window, the forward and adjoint model integrations in 4D-Var result in implicit propagation of the background errors to the later timesteps, which generally makes the 4D-Var analysis more accurate than 3D-Var even in the end of assimilation window.

Another useful consequence of the implied covariance evolution is that 4D-Var can dynamically estimate unobserved components in the state vector, which finally returns us to the inverse problem of estimating an emission forcing given a set of tracer measurements.

In a chemistry model, the state vector $x_k$ consists of tracer concentrations. The emission fluxes are introduced as a forcing $f$,

$$x_{k+1} = M_k(x_k) + f_{k+1},$$ \hspace{1cm} (27)$$

where each $f_k \sim \mathcal{N}(f_k^b, K_k)$ is Gaussian and to be estimated along with $x_0$. The 4D-Var cost function for this system is

$$J_f(x_0, f_1, ..., f_n) = \frac{1}{2} \sum_{k=0}^{n} (y_k - H_k(x_k))^T R_k^{-1} (y_k - H_k(x_k))$$

$$+ \frac{1}{2} \sum_{k=0}^{n} (f_k - f_k^b)^T K_k^{-1} (f_k - f_k^b)$$

$$+ \frac{1}{2} (x_0 - x_b)^T B^{-1} (x_0 - x_b).$$ \hspace{1cm} (28)$$

With a similar derivation as for Eq. (24), it can be shown that the gradient for $J_f$ with respect to $f_k$ is simply given by the adjoint variables $x^*$ as

$$\frac{\partial J_f(x_0, f_1, ..., f_n)}{\partial f_k} = x_k^* + K_k^{-1} (f_k - f_k^b)$$ \hspace{1cm} (29)$$

The variational formulation of the flux inversion was presented by Elber et al. (2000). However, Eq. (27) has the same form as the generic evolution equation (19). If the stochastic forcing $f_k$ is interpreted as the model error, as in Eq. (19), then the problem (27) and its solution are similar to the technique of Derber (1989), later referred by Trémolet (2006) as a weak-constraint 4D-Var algorithm.

The most important complexity in implementing the 4D-Var method with an existing model lies within developing, testing and maintaining the code corresponding to the adjoint model and observation operators $M^*$ and $H^*$. This can be achieved with either manual or automatic transformation of the program code (Giering and Kaminski, 1998). The resulting code is referred as the discrete adjoint. The alternative approach is to start from the continuous system of equations, such as Eq.
(3), and derive the continuous adjoint system, typically also a partial differential equation. The adjoint system for the flux-form advection equation (1) is an equation similar to its advective form (2) (Marchuk, 1995; Elbern and Schmidt, 1999), which admits a similar numerical solution as the forward equation. Consequently, the continuous adjoint has been adopted for the advection component in several chemistry transport models (Henze et al., 2007; Hakami et al., 2007; Davoine and Bocquet, 2007), and the approach has proven successful even though the gradient obtained this way is only an approximation to Eq. (25). Furthermore, in the numerical tests of Gou and Sandu (2011), the continuous adjoints for advection were found preferable at least in an idealised setting; this was attributed to the nonlinearities in advection schemes mentioned in Section 2.

3.2 Applications in atmospheric chemistry and dispersion modelling

The preceding section showed that inverse modelling of the emission fluxes and chemical data assimilation can be handled by a similar formalism. However, on practical level, the past research has followed several largely distinct lines. The lines of research with influence on this thesis include (i) inverse studies of point sources, (ii) reactive gas flux inversions and (iii) studies on chemical data assimilation for air quality forecasting.

The first experiments with data assimilation in atmospheric chemistry models were with stratospheric chemistry models (Fisher and Lary, 1995) and some years later with tropospheric chemistry and aerosol models (Elbern and Schmidt, 1999; van Loon et al., 2000; Collins et al., 2001). Studies assessing the improvements of short term forecasts in addition to evaluating the analysis fields have been described by Elbern and Schmidt (2001), Blond and Vautard (2004) and Wu et al. (2008) for ozone and by Tombette et al. (2009) and Pagowski et al. (2010) for particulate matter.

For ozone and particulate matter, initialising the forecast from the analysis has been found to improve the forecast mainly within a range of 24 hours with a minor improvement for hours 24–48. Fewer studies have considered the impact of data assimilation on forecasts of nitrogen dioxide (NO₂) or other short-lived pollutants. Due to the shorter chemical lifetime, the forecast improvements can be expected to be transient, which was indeed confirmed by Wang et al. (2011), Silver et al. (2013), and Paper II.

In addition to the mainly regional model studies cited above, chemical data assimilation has been incorporated to global models. Due to the uneven coverage of the surface-based observation networks, the main attention in global applications has been on assimilating satellite retrievals of both aerosol (Benetti et al., 2009; Zhang et al., 2008; Weaver et al., 2007) and gas-phase (Inness et al., 2013) constituents.

Inverse methods for source term estimation were first focused on accidental or intentional releases of radioactive tracers. The 4D-Var method was proposed by Robertson and Persson (1993) for estimating the intensity of a release with a known location and timing. Adjoint methods were later investigated also for locating the
source by Pudykiewicz (1998) and Issartel and Baverel (2003) and, as part of a more strictly Bayesian algorithm, by Bocquet (2007).

In the meantime, algorithms for estimating distributed (as opposed to point) sources were introduced for estimating emission fluxes of greenhouse gases (Kaminski et al., 1999; Peters et al., 2010; Peylin et al., 2013) and short-lived trace gases (Müller and Stavrakou, 2005; Miyazaki et al., 2012) in global scale. Especially in the early studies, the emissions were aggregated into a few (< 100) geographical regions, while the later studies have provided also fully gridded emission estimates.

Updating the emission fluxes as means to improve regional-scale air quality forecasts was investigated by Elbern et al. (2007). In a two-week experiment, adjusting the emission fluxes was found to result in additional and more persistent forecast improvement when compared to adjusting the initial condition only. On the other hand, Curier et al. (2012) assimilated ozone data to adjust fluxes of precursor species, and found the emission adjustments to have little temporal continuity, and consequently were unable to obtain significant forecast improvements by using the adjusted emissions. Paper I considers the approach of Elbern et al. (2007) with a focus on emissions of sulphur dioxide.

Inverse modelling of volcanic emissions usually has to rely on satellite retrievals of aerosols or trace gases. While the geographical location can be assumed known, the emitted amount and vertical distribution are uncertain. Inverse modelling using satellite data has been shown to be useful for estimating the vertical (Eckhardt et al., 2008), temporal (Boichu et al., 2013) or both vertical and temporal (Stohl et al., 2011) emission profiles in explosive volcanic eruptions. The low dimension of the parameter vector (emission as function of time and/or altitude) makes it possible to avoid the adjoint formalism (Eq. 29) and evaluate the required source-receptor matrix elementwise. The aforementioned studies are, nevertheless, based on quadratic cost functions similar to 4D-Var (Eq. 28). Alternative approaches focusing mainly on the vertical distribution have been presented in papers of Flemming and Inness (2013), Zidikheri and Potts (2015) and Heng et al. (2016).

The satellite retrievals previously used in inverse modelling of volcanic emissions have been for the column density (vertical integral) of volcanic ash or sulphur dioxide. Estimation of the vertical profile then depends on resolving the variation of transport patterns with regard to the injection height. This approach has been used successfully for several eruptions (Kristiansen et al., 2010; Moxnes et al., 2014); however, in absence of sufficient vertical wind shear, the inversion relies on a priori profile. In contrast, Paper III derives an observation operator for satellite-retrieved plume height. Using retrievals by the Infrared Atmospheric Sounding Interferometer (IASI), the study demonstrates that complementing the column density retrievals with the retrieved plume heights results in a more realistic vertical emission profile.
4 Model setup and input data

The studies which constitute this thesis use the SILAM chemistry transport model. This section gives a brief overview of the model and the input data used, while detailed descriptions of the model configurations and datasets are given in each publication. The input data described below apply to papers I, II and III. Paper IV is based on two-dimensional synthetic test cases with the main focus on the test suite proposed by Lauritzen et al. (2012).

4.1 The SILAM model

The SILAM (System for integrated modeling of atmospheric composition) model was initially developed as a Lagrangian particle model for emergency applications (Sofiev et al., 2006) and later evolved into an Eulerian chemistry-transport model (Sofiev et al. (2008) and Paper IV). In addition to the horizontal transport described in Paper IV, the model includes the vertical discretisation of Sofiev (2002) and the particle dry deposition scheme of Kouznetsov and Sofiev (2012). Dry deposition of gases as well as wet deposition of gases and particles is described by Sofiev et al. (2006).

The model is normally set up with a regular lon-lat grid. Except for the advection benchmarks in Paper IV, the studies in this thesis use a limited area configuration with resolutions between 0.25° and 0.5°. The vertical grid is flexible; Paper I and Paper II use terrain-following z-levels reaching up to 7-9 km with vertical resolution decreasing from 30-40 meters in lowest layers to 2-3 km in the free troposphere. In Paper III, the maximum layer thickness is limited to 500 meters to better resolve the volcanic plumes injected into upper troposphere.

SILAM includes two chemistry mechanisms: the DMAT scheme of Sofiev (2000) describes formation of inorganic secondary aerosols and partitioning of nitrogen oxides, while all other photochemical processes, including ozone formation, are simulated with the Carbon Bond mechanism (CB4, Gery et al. (1989)). The CB4 mechanism is implemented using a three-stage Rosenbrock solver generated by the Kinetic Pre-Processor software (Sandu and Sander, 2006). The DMAT mechanism is implemented manually following the quasi-steady-state approach.

The SILAM variational data assimilation system is described by the papers I, II and III in this thesis. The adjoint code used in papers I and III uses a continuous adjoint for advective transport and a manually developed adjoint for the DMAT sulphur chemistry; the diffusion and deposition processes are self-adjoint. The background error covariance operators are based on separation of dimensions as described by Singh et al. (2011). The quasi-Newton minimisation code of Gilbert and Lemaréchal (1989) is used with 3D-Var, while the bound-constrained L-BFGSB code of Byrd et al. (1995) is used for emission inversions.

4.2 Emissions, boundary conditions and meteorology

The meteorological fields used in this study originate to the ECMWF IFS weather prediction system; in Paper I and Paper II, operational forecasts are used while
Paper III uses the ERA-Interim reanalysis (Dee et al., 2011).

For anthropogenic emissions, Paper I uses the EMEP inventory valid for the year 2000. In Paper II, the TNO-MACC-II emission inventory (Kuenen et al., 2014) is used. In addition, the biogenic emissions of isoprene as simulated in Paper II using the model of Poupkou et al. (2010).

Paper II uses lateral boundary conditions from the MACC reanalysis (Inness et al., 2013). Paper I, which simulates only oxides of sulphur, does not use lateral boundary conditions; however, this is unlikely to affect the concentrations in the Central European subdomain, where the observations were assimilated.

4.3 Observational data

In Paper I and Paper II, hourly in-situ data of SO$_2$, NO$_2$ and ozone are assimilated. The data are extracted from the AirBase database compiled by the European Environment Agency. In Paper II, only the data from rural (for NO$_2$) or rural and suburban (for O$_3$) stations are used. In Paper I, data from all station types are assimilated. In both studies, part of the stations are withheld from assimilation and used for verifying the analysis fields.

Paper III uses satellite retrievals of Carboni et al. (2012) based on the observations of the Infrared Atmospheric Sounding Interferometer (IASI). The IASI instrument (Clerbaux et al., 2009) onboard the MetOp-A (since 2013, also MetOp-B) satellite is a Fourier transform spectrometer with a spectral range from 3.62 to 15.5 \( \mu \)m and a field of view of about 12 km at nadir. The dataset consists of retrievals of both column density and plume height of sulphur dioxide. For both parameters, the retrievals include rigorously derived error estimates which in turn are included in the inversion.
5 Results and discussion

5.1 Assimilation of air quality monitoring data

Paper I and Paper II study assimilation of in-situ trace gas observations into regional-scale simulations. In Paper I, a specific goal was to assess the usefulness of the gridded emission flux as a control variable in 4D-Var data assimilation. The goal of Paper II was to develop an analysis system with computational performance sufficient for long-term air quality reanalyses as well as operational, near-real time use. While the assimilation method is simpler, attention is given for estimating the statistical parameters affecting its performance.

In the study presented in Paper I, hourly SO$_2$ data at central European stations were assimilated in a two-week experiment, and effectiveness of the assimilation was evaluated in up to 48 h forecasts at stations not used in assimilation. Two approaches to assimilation were compared: the regular adjustment of forecast initial state using the 3D-Var method, and as an alternative, a 4D-Var assimilation scheme which estimates local variations in emissions of SO$_2$ in addition to the airborne concentrations. The refined emission fluxes were used to drive the subsequent forecast.

![Figure 3: Assimilation experiment of Paper I. Left: location of the assimilation (grey) and control (red) stations. Right: emission correction factor after assimilation, average over two weeks.](image)

The relative adjustments (a posteriori divided by a priori) to the SO$_2$ emissions, averaged over the two weeks, are shown in Fig. 3. In most areas, the average change is less than 5%. The strongest positive adjustments are obtained for limited areas in Hungary and Czech republic, while the most prominent reduction is obtained for the degassing emissions of volcano Etna.

The forecast impacts of assimilation were evaluated in terms of root mean squared error (RMSE). As seen from Fig. 4, the overall forecast improvements due to assimilation were mostly minor. The effect of 3D-Var diminished within 12
hours of forecast, while with 4D-Var, a small but noticeable improvement persisted over the 48 hours. However, for a number of stations corresponding to the upper (86th) percentile of RMSE, the reduction was up to 50%.

The upper percentile turned out to include several stations near Etna, where the a priori simulations showed high SO$_2$ levels, and where the a posteriori emissions indicated systematic reductions throughout the two-week period. A feasible explanation to the reduction is that the degassing emissions, which were given on yearly level in the emission inventory, were not representative of the simulated period.

Paper II focuses on the prominent photochemical pollutants: hourly monitoring data of nitrogen dioxide (NO$_2$) and ozone (O$_3$) were assimilated in a European scale domain. The 3D-Var method was chosen mainly due to computational reasons; the system is aimed for reanalysis production in yearly and longer timescales. Special attention was given for estimating the observation and background error statistics, which in Paper I were based on ad-hoc values. The error statistics were calibrated with monthly simulations covering June and December 2011, and the obtained setup was subsequently tested in an assimilation experiment covering the year 2012.

The diagnostic identities presented by Desroziers et al. (2005) were used for estimating the observation error standard deviation $\sigma_{\text{obs}}$ and the background error standard deviation $\sigma_b$, which together control the weighting between the model and observed values in the assimilation. Recently, Ménard (2016) analysed convergence the Desroziers method, and while he showed that estimating the full observation and background error covariance matrices is not possible, estimates of variance parameters, as done in Paper II, were found to converge provided that the assumptions of the parametrisation were satisfied.

The analysis errors were evaluated on stations excluded from assimilation. As a result of the improved diurnal and seasonal profiles for observation and background error standard deviations, the correlation coefficient and RMSE were improved.
Figure 5: Diagnosed background (dashed) and observation error (solid lines) standard deviations ($\mu g m^{-3}$) on rural stations for $O_3$ (left) and $NO_2$ (right). Red lines correspond to the calibration made for June 2011, blue lines correspond to calibration for December 2011. Figure 4 in Paper II.

noticeably for both $O_3$ and $NO_2$. While the standard deviations were adjusted based on two months (June and December) in 2011, the improvement was valid throughout the simulated year 2012.

The obtained observation and background error standard deviations are shown in Fig. 5. The profiles are strikingly different from the first guess standard deviations, which were constant values 11.2 ($\sigma_{obs}$) and 20.6 ($\sigma_b$) $\mu g m^{-3}$ for $O_3$ and 4.0 ($\sigma_{obs}$) and 8.0 ($\sigma_b$) $\mu g m^{-3}$ for $NO_2$. Contrary to the first guess values, the estimated observation errors are larger than the corresponding background errors, which implies that the influence of an individual observation was reduced in comparison to the background field. The estimated observation and background errors also have a clear diurnal variation. For ozone, the variation is different for observation and background errors.

In addition to evaluating the effect of improved statistical parameters, the forecast impact of $O_3$ and $NO_2$ assimilation was evaluated over a three-week period which covered an ozone episode. The results (Fig. 6) demonstrate that for ozone, the effect of initialising the forecast from analysis diminishes within the first 24 forecast hours; the forecast of $NO_2$ relaxes to the background within 6-12 hours. The results regarding the short-lasting forecast improvement due to improved initial conditions are consistent the findings of previous studies: the forecast improvements for $NO_2$ have been limited to the range of a few hours in summer conditions (Wang et al., 2011; Silver et al., 2013); for ozone, improvements have been reported to extend from 24 to 48 h (Curier et al., 2012; Elbern et al., 2007).

Fewer studies have addressed $SO_2$. However, Elbern et al. (2007) found that the model bias was reduced for $\sim$ 24 forecast hours due to improved initial conditions. When also emission fluxes were adjusted similarly to Paper I, the forecast bias and RMSE were substantially reduced throughout the 48 hour forecast window. In the present work, neither of these results could be fully reproduced. The two studies are not immediately comparable due to differences in models, simulation domains...
Figure 6: The model bias ($\mu$gm$^{-3}$) and correlation for O$_3$ (top) and NO$_2$ (bottom) at validation stations as a function of forecast length (blue lines). The corresponding indicators, the analyses (black) and control run (green), are shown averaged by time of day and replicated over the forecast window. Figures 7 and 8 in Paper II.
and the observational networks used. However, in the study of Elbern et al. (2007), the free-running model had a positive bias, which resulted in systematic reduction of a posteriori SO$_2$ emissions. In contrast, the emission adjustments in Paper I showed fewer systematic features. This could explain the different conclusions of the two studies, since using the adjusted emissions in forecasts is based on their assumed persistence.

The observation error estimates obtained in Paper II for NO$_2$ and ozone are somewhat higher than assumed in most previous studies. However, the errors for O$_3$ have similar magnitude as those estimated by Gaubert et al. (2014) using the same diagnostic relations but with the EnKF assimilation system and the CHIMERE chemistry transport model. The magnitude and diurnal variation of the observation errors is hardly explained by instrumental errors, but might be explained by representativeness errors, which arise from the discrepancy between the spatial scales represented by the model and the in-situ observations. The representativeness errors could be assumed to increase during night due to weaker mixing in the boundary layer, which would explain the strong summertime diurnal variation of $\sigma_{obs}$ for O$_3$.

So far, the representativeness of in-situ data has been addressed mainly as a question of characterising the measurement stations (Henne et al., 2010; Joly and Peuch, 2012) and restricting the assimilation to the stations considered representative. Few studies have tried to explicitly quantify the representativeness errors in relation to the model resolution; however, Schutgens et al. (2016) estimated the magnitude of representativeness errors for satellite observations by aggregating data simulated at 10 km resolution, and found RMS differences reaching 30-160% between the high-resolution (“observation”) and aggregated (“model”) values. For air quality monitoring data, the relevant spatial and temporal scales might be smaller and more difficult to reach by modelling. However, representativeness errors can be included explicitly in the analysis scheme as shown by Koohkan and Bocquet (2012), which might provide a practical approach for assimilation studies.

5.2 Volcanic source term inversion using satellite data

Paper III builds on the work on variational assimilation presented in Paper I. However, the study was focused on estimating emission parameters (source term inversion) for the sulphur dioxide released in an individual eruption (Eyjafjallajökull in 2010). The location of eruption is assumed known, but the temporal and vertical variation are to be estimated. Instead of adjusting the emission source in steps of 24 hours, the inversion was performed in a single assimilation window covering 20 days.

The assimilated dataset consisted of satellite SO$_2$ retrievals of Carboni et al. (2012) from the Infrared Atmospheric Sounding Interferometer (IASI) instrument. The retrievals include both the column burden and plume height for SO$_2$, and an observation operator for joint assimilation of both quantities was developed in this work. The observation operator depends only on the vertical centre of mass and column density and thus sets only a partial constrain on the vertical profile. However, this approach avoids making additional assumptions on the shape or
Figure 7: Emission flux (kg m$^{-1}$s$^{-1}$) of SO$_2$ in the Eyjafjallajökull eruption. Left: inversion using column density and plume height retrievals. Right: inversion using only column density retrievals. Figure 7 in Paper III.

In Paper I, the background errors for the emission flux were defined subjectively. In an emission inversion, the assumed background error variance is equivalent to a regularisation parameter, which aims to balance the solution between the bias introduced by the a priori data and the noise introduced by the model and observation errors. Since the regularisation parameter is difficult to determine a priori, a nonparametric method referred as the L-curve (Hansen, 1992) was used in Paper III to estimate the regularisation parameter as a part of the inversion.

In the L-curve method, the inversion needs to repeated with different values of the regularisation parameter, which would be computationally expensive. However, as discussed by Fleming (1990) and Santos (1996), the iterative minimisation schemes used in 4D-Var yield by construction a sequence of solutions with decreasing regularisation. This feature of 4D-Var is rarely exploited in data assimilation, but the synthetic experiments in Paper III demonstrate that the regularised solutions obtained by truncating the 4D-Var iteration are practically equivalent to those obtained with the common Tikhonov regularisation.

The inversion results are shown in Fig. 7 which presents the emission rate of SO$_2$ as function of time and height. The plume height time series of Arason et al. (2011) are plotted together with the emission. The distribution of emissions obtained with and without plume height assimilation are largely similar. However, individual peaks reaching 12-15 kilometres height are strongly suppressed when the plume height retrievals are assimilated, which is consistent with the plume height time series.

The inversion gave an estimate of 0.29 Tg SO$_2$ emitted during the eruption. If only column load was assimilated, the emission increased to 0.33 Tg. The temporal and vertical profiles, integrated over the whole eruption, are shown in Fig. 8. The vertical distribution for SO$_2$ is quite similar to the vertical distribution obtained for volcanic ash by Stohl et al. (2011) despite the differences in the temporal distribution of the ash and SO$_2$ emissions.
The 4D-Var method has previously not been applied in volcanic source term inversions. The 4D-Var inversion is equivalent to the linear algebraic method of Eckhardt et al. (2008), Stohl et al. (2011) and Boichu et al. (2013), which does not require adjoint model integrations. The main advantage of 4D-Var is in lower computational cost, although the difference is less important in short eruptions where the estimated emission has fewer degrees of freedom. Nevertheless, the observation operator for plume height developed in Paper III is applicable regardless of the inversion method, and the operator is also suitable for the traditional data assimilation which does not consider emission sources.

The variational method, as well as its algebraic counterpart, are based on assumption of a deterministic forward model, which makes realistic treatment of model errors difficult. The experiments with synthetic data presented in Paper III show that under fairly realistic assumptions, the model uncertainty has larger impact on the results than the observational noise. When real data are used, ad hoc modifications are frequently needed to obtain realistic inversion results. In Paper III, the retrieval errors were augmented with a constant model error term, while Boichu et al. (2013) kept only 10% of the zero-valued observations to improve the fit to data. However, more rigorous treatment of model uncertainty is difficult in the current inversion approach based on additive Gaussian errors.

During the 2010 Eyjafjallajökull eruption, a major limitation to the ash dispersion forecasts was the unpredictability of the emission source. Contrary to the continuous emissions studied in Paper I, the emissions in explosive eruptions have little persistence, and adjusting the emission fluxes is therefore unlikely to have inherent advantages when compared to adjusting the prognostic variable (airborne concentration). Although inverse modelling has been found to capture the vertical distribution of plumes better than a 4D-Var analysis (Kristiansen et al., 2017), ensemble or hybrid assimilation methods would allow more flexible treatment of the
5.3 Numerical solution of the advection equation

Paper IV evaluates a numerical scheme originally devised by (Galperin, 1999, 2000) for solving the advection equation, and discusses its incorporation to the SILAM. In its original form, the advection scheme produced strong artifacts especially in deformational flows associated with orographic uplifting. Paper I introduces a modification which substantially reduces magnitude of the artifacts. This version of scheme was implemented in the SILAM model and evaluated using a set of synthetic tests which aim to measure the scheme’s performance for both smooth and discontinuous solutions.

As outlined in Section 2, in the Galperin’s advection scheme, the solution is represented by discontinuous, rectangular pulses $\Pi(x)$, which at the beginning of timestep are confined to each grid cell. The scheme uses dimensional splitting, and the slabs are tracked along the one-dimensional flow $u(x)$. In the original version, the slabs were transported rigidly as

$$\Pi_i(x, t + \Delta t) = \Pi_i(x - u_i \Delta t, t),$$

(30)

where $u_i$ is the velocity in the middle of cell $i$. This implies two problems: first, Eq. (30) is inaccurate if $u(x)$ changes quickly; second, if $u(x)$ changes sign within cell $i$, then $\Pi_i(x, t)$ remains almost unchanged. Together, these effects result in spurious accumulation of mass in areas where the atmospheric flow changes suddenly.

The improved version tracks separately the borders of $\Pi_i$, which allows the slab to deform with the flow and avoids the accumulations. The time integration for tracking $\Pi_i$ was first changed to the second order implicit midpoint method, and in the finally published version, to the analytical solution for piecewise linear $u(x)$.

The two-dimensional tests introduced by Lauritzen et al. (2012) were used with various initial data to numerically evaluate the convergence and accuracy of the scheme. For a sufficiently smooth (Gaussian) initial condition, the rate of converge was near second order, and the absolute level of errors was lower when the scheme was run at a higher Courant numbers. Intuitively, the numerical errors in the Galperin scheme are due to the errors both in trajectory integration and in the Eulerian reconstruction (see Section 2). While the trajectory errors increase with increasing Courant number, the reconstruction errors are likely to decrease due to the fewer reconstructions needed, and thus, the optimal performance is achieved at some intermediate Courant number.

In the intercomparison study of Petrova et al. (2008), the Galperin scheme was found to spuriously steepen gradients in initially smooth solutions. This effect was confirmed in Paper IV, however, the presence of such steepening was found to depend nonlinearly on the initial condition. The nonlinearity, dissipation, and effective resolution of the scheme were studied further with spectral analysis.

In a one-dimensional periodic domain with 100 grid points, the scheme was run with sinusoidal initial conditions up to the 25th wavenumber. For each initial condition, the integration was repeated for a range of Courant numbers between 0 and 1. The scheme is then characterised by the amplification factor (ratio of
amplitudes of the initial and final solution) and root mean squared error (RMSE). Intuitively, the amplification factor measures the numerical diffusion which causes a scheme to lose details at smaller spatial scales. The spectrally resolved RMSE complements the amplification factor by including also the impact of phase errors and possible spurious modes.

Fig. 9 shows the amplification factor and spectral RMSE at the Courant number 0.7 as a function of the wavenumber up to $k = 25$. The results for Galperin scheme are shown for two cases: with no constant background (initial range from 0 to 1) and with a large constant background (initial range from 10 to 11). For comparison, the results for a generic, non-conservative semi-Lagrangian scheme with cubic interpolation are shown.

When the Galperin scheme is run without a background term, the amplification factor remains above 0.8 for all wavenumbers. The RMSE increases initially for $k$ up to $\sim 5$, but shows otherwise little connection to the amplification factor, contrary to typical numerical schemes such as the cubic semi-Lagrangian scheme included in comparison. In presence of a background, the behaviour changes radically: all amplification factors are below or equal to one, and frequencies above $k \sim 17$ are almost completely damped. However, for the well-resolved frequencies ($k < 8$), the RMSE is lower in presence of the background.

For linear finite difference schemes, amplification factors above one imply instability. For the Galperin scheme, the amplification factors for a single timestep fluctuate depending on the centres of mass, but the solution remains bounded due to non-negativity and conservation of mass. Based on numerical tests, the asymptotic solution as $t \to \infty$ in a periodic domain is either a constant, or a combination
of rectangular pulses. These cases correspond to the dissipative or non-dissipative behaviour depicted in Fig. 9.

The switch between the dissipative or non-dissipative domains is explained by variation of the centres of mass. Presence of a background constrains the range of variation of the centres of mass, since the changes in centre of mass depend on the relative magnitude of the perturbation that enters the gridcell from upwind side. Paper IV discusses use of a “smoothing factor” which essentially fixes the scheme to the dissipative domain. However, a better solution might be achieved by modifying the reconstruction function (Eq. 12). Since the publication of Paper IV, further research (R. Kouznetsov, personal communication) has shown that by allowing the reconstruction function to depend also on the values of the adjacent grid cells, it is possible to devise a piecewise constant reconstruction function which is invariant under addition of constant. Alternatively, a higher order scheme can be constructed using a piecewise linear reconstruction function.
6 Conclusions and future work

This study has presented applications of variational assimilation methods in air quality forecasting and volcanic source term inversion problems. Furthermore, a numerical scheme for solving the advection equation has been evaluated using recently published benchmarks.

The impact of conventional data assimilation, where observations are used to adjust the initial condition of a forecast, was evaluated for regional-scale predictions of NO$_2$, SO$_2$ and ozone. In Paper II, the forecast correlation and bias were found to improve for up to 12-24 hours for ozone, but for only 6-12 hours for NO$_2$. The impact on SO$_2$ forecasts was evaluated in a separate study (Paper I), where the improvement in RMSE extended for about 12 hours. Thus, especially for short-lived pollutants, this type of data assimilation has limited use in forecasting.

For SO$_2$, the conventional data assimilation approach was compared to a scheme similar to that presented by Elbern et al. (2007), where emission fluxes are adjusted along with the initial conditions. This setup was found to produce forecast improvements which were more persistent, but on average, small. Significant improvements could be obtained for stations where the emission adjustments persisted over the two-week experiment. This suggests that the emission adjustment can be effective assimilation strategy, if the forecast skill is affected by strong and systematic biases in the emission forcing.

For O$_3$ and NO$_2$, the analysis scores were substantially improved by iterative refining of the background and observation error standard deviations. The final values for observation errors are larger than commonly assumed, but similar to a previous study (Gaubert et al., 2014) applying a similar estimation procedure. Better quantification of the representativeness errors of air quality monitoring data would be valuable for both data assimilation and model evaluation based on such data.

In Paper III, the 4D-Var method was adapted for solving the inverse problem of reconstructing the vertical and temporal profile of volcanic SO$_2$ emissions. The 4D-Var based inversion regularised by truncated iteration was shown to produce equivalent results to the more common direct method with Tikhonov regularisation. Depending on the setup, the variational method may require considerably less computation.

The SO$_2$ retrievals from the IASI instrument were assimilated using the 4D-Var inversion scheme. Including the IASI plume height retrievals in the inversion in addition to the column density retrievals, reduced the emissions above 10-12 km. The resulting vertical distribution is more consistent with radar observations of the plume, which supports the conclusion that the SO$_2$ emissions of Eyjafjallajökull were, similarly to ash, confined to the troposphere.

The inverse modelling is established by several studies as a means for incorporating observations into simulations of volcanic plumes. The same approach might turn out useful for simulating other intermittent emissions sources, such as wind blown dust and wildfires. However, the difficulty of accommodating realistic model errors is a weakness of the current inversion approach, and improving this aspect may require significant algorithmic changes.
Evaluation of the advection scheme presented by Galperin (1999) indicated that the scheme is generally comparable in accuracy to the schemes evaluated by Lauritzen et al. (2014), with rate of convergence slightly lower than second order. The scheme is conservative and preserves non-negativity of the tracers, but was found to have a strongly nonlinear response under addition of a constant.

For a single tracer, the scheme was found to have good computational performance compared the schemes described by Kaas et al. (2013). However, the lack of common benchmarks makes comparing the numerical performance across schemes difficult. A uniform, transparent methodology for assessing the computational aspects of numerical schemes would be helpful for potential users of the numerical schemes.
References


Lauritzen, P. H., Nair, R. D., and Ullrich, P. A. (2010). A conservative semi-
Lagrangian multi-tracer transport scheme (CSLAM) on the cubed-sphere grid. 

A standard test case suite for two-dimensional linear transport on the sphere. 

Lauritzen, P. H., Ullrich, P. A., Jablonowski, C., Bosler, P. A., Calhoun, D., Con-
ley, A. J., Enomoto, T., Dong, L., Dubey, S., Guba, O., Hansen, A. B., Kaas, 
E., Kent, J., Lamarque, J.-F., Prather, M. J., Reinert, D., Shashkin, V. V., 
A standard test case suite for two-dimensional linear transport on the sphere:
results from a collection of state-of-the-art schemes. Geoscientific Model Devel-
opment, 7(1):105–145.

Le Dimet, F.-X. and Talagrand, O. (1986). Variational algorithms for analysis 
and assimilation of meteorological observations: theoretical aspects. Tellus A,
38A(2):97–110.


water model of the sphere. Quarterly Journal of Royal Meteorological Society,
123:2477–2498.


Lorenz, E. N. (1982). Atmospheric predictability experiments with a large numeri-

Magnusson, L. and Källén, E. (2013). Factors influencing skill improvements in the 


Ménard, R. (2016). Error covariance estimation methods based on analysis residu-
als: Theoretical foundation and convergence properties derived from simplified 
observation networks. Quarterly Journal of the Royal Meteorological Society,

derived from an assimilation of OMI tropospheric NO 2 columns. Atmospheric 
Chemistry and Physics, 12(2):2263–2288.
Moxnes, E. D., Kristiansen, N. I., Stohl, A., Clarisse, L., Durant, A., Weber, K.,
and Vogel, A. (2014). Separation of ash and sulfur dioxide during the 2011

the adjoint of the IMAGES model. *Atmospheric Chemistry and Physics*, 5:1157–
1186.


Three-dimensional variational data assimilation of ozone and fine particulate
matter observations: some results using the Weather Research and Forecasting-
Chemistry model and Grid-point Statistical Interpolation. *Quarterly Journal of

Peters, W., Krol, M. C., van der Werf, G. R., Houweling, S., Jones, C. D., Hughes,
Ramonet, M., Schmidt, M., Ciattaglia, L., Apadula, F., Heltai, D., Meinhardt,
F., di Sarra, A. G., Piacentino, S., Sferlazzo, D., Aalto, T., Hatakka, J., Ström,
Rodó, X., Morgui, J. A., Vermeulen, A. T., Popa, E., Rozanski, K., Zimnoch,
M., Manning, A. C., Leuenberger, M., Uglietti, C., Dolman, A. J., Ciais, P.,
Heimann, M., and Tans, P. (2010). Seven years of recent European net terrestrial
carbon dioxide exchange constrained by atmospheric observations. *Global Change
Biology*, 16(4):1317–1337.

ants of TRAP scheme for solving advection equation — comparison with other

Peylin, P., Law, R. M., Gurney, K. R., Chevallier, F., Jacobson, A. R., Maki, T.,
Niwa, Y., Patra, P. K., Peters, W., Rayner, P. J., Rödenbeck, C., Van Der Laan-
Luijkx, I. T., and Zhang, X. (2013). Global atmospheric carbon budget: Results
from an ensemble of atmospheric CO2 inversions. *Biogeosciences*, 10(10):6699–
6720.

Poupkou, A., Giannaros, T., Markakis, K., Kioutsoukis, I., Curci, G., Melas,
Compound emissions: Software development and first validation. *Environmental

Prather, M. (1986). Numerical Advection by Conservation of Second-Order Mo-


On variational data assimilation for estimating the model initial conditions and emission fluxes for short-term forecasting of SOx concentrations

J. Vira, M. Sofiev

Finnish Meteorological Institute, P.O. Box 503, FI-00101 Helsinki, Finland

A R T I C L E   I N F O

Article history:
Received 20 January 2011
Received in revised form
23 September 2011
Accepted 26 September 2011

Keywords:
Data assimilation
Air quality forecasting
Model initialisation
Emission refinement
Numerical modelling

A B S T R A C T

The study analyses the added value of data assimilation for short-term air quality forecasting by means of three modelling experiments with sulphur oxides. Two ways of utilising the observations are considered: determination of the optimal model initial state and adjustment of the emission fluxes of atmospheric pollutants. It is demonstrated that the influence of the initial conditions on the predicted SOx concentrations disappears within less than a day in European-scale applications. Adjusting the emission fluxes had a longer lasting impact on the model results, frequently covering the whole forecast window. The two-week long data assimilation exercise for Southern Europe showed that the largest improvement of the model score with regard to individual monitoring sites was obtained for the stations with the worst initial model-measurement agreement. With the emission adjustment, a major improvement was achieved for the stations near the Etna volcano, the strongest source in the area, where the SO2 emission was reduced by almost 50% as a result of the data assimilation.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The increasing availability of observations of the composition of the lower troposphere opens new possibilities in data assimilation for air quality (AQ) assessment and forecasting. Historically, the data assimilation (DA) technology was developed for the needs of meteorological modelling. One of the first comprehensive descriptions of the data assimilation approach later called Optimal Interpolation (OI) was given by Gandin (1963) based on earlier ideas of A.N. Kolmogorov. Data assimilation was formulated as the problem of simultaneously minimising the differences of the resulting state estimate from a set of observations and from the initial model field.

The means of solving this optimisation problem vary widely. In three-dimensional methods, such as OI, all measurements are taken to represent a single time and the assimilation step does not involve the forecast model. However, the observations can be utilised more efficiently by using the model to connect the estimated variables with the observations at the appropriate times. The simplest approaches are based on nudging (dynamical relaxation, first used in Europe by Davies and Turner, 1977). More elaborate methods are based on the Kalman Filter (see e.g. Kalnay, 2003), which in practice needs to be modified to reduce the computational burden. Recent versions include the Reduced-Rank (Todling and Cohn, 1994) and Ensemble Kalman Filters (EnKF, Evensen, 1994).

In the variational assimilation methods 3D-VAR and 4D-VAR (Lorenc, 1986; Le Dimet and Talagrand, 1986), a state estimate is obtained by iteratively minimising the corresponding cost function in three or four dimensions respectively. A major contribution to the underlying theory was made by G.I. Marchuk, based on the application of optimal control theory to environmental problems (see Marchuk, 1995 and references therein).

The above and other DA methods were developed for the estimation of the state of the atmospheric model, which is then used as a starting point for solving the equations of atmospheric dynamics. The same approach is frequently followed in chemistry transport modelling. Examples include the 3D-VAR and OI studies by Tombette et al. (2009) and Blond and Vautard (2004). However, both studies noted that the effect of initial state adjustment is generally limited to the first 24 h of the forecast. The reason is that the transport and transformation of atmospheric pollutants is strongly controlled by emission fluxes and meteorological fields, which are external parameters. To address this issue, some of the DA methods may be extended beyond the initial state estimation, towards adjusting the input parameters (see Elbern et al. (2000, 2007) for the extension of 4D-VAR to emission fluxes).

Using DA techniques for adjusting emission fluxes has many common features with the source apportionment problem, which has been studied in the context of climate forcing, for analysis of observational campaigns, and in numerous emergency applications (Shankar Rao, 2007). The data assimilation in climate-forcing studies has been mainly utilised for determining the green-house

1352-2310/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
gas emissions (e.g. Meirink et al., 2008; Bergamaschi et al., 2005; Kaminski et al., 1999; Houweling et al., 1999). The analysis of observational campaigns concentrates on interpretation of the obtained time series (Siljamo et al., 2008; Tarasova et al., 2007; Saarikoski et al., 2007).

The source identification problem has been massively studied in the context of accidental releases (Davone and Bocquet, 2007, Bocquet, 2005ab, Sofiev and Atlaskin, 2007; Robertson and Langner, 1998; Thomson et al., 2007; Pudykiewicz, 1998, etc.), especially within the scope of Comprehensive Nuclear Test Ban Treaty (CTBT) (Wotawa et al., 2003 and references therein, Becker et al., 2007; Issartel and Baeverl, 2003). In contrast to the AQ forecasting or climate-forcing studies, only limited prior data regarding the actual source is usually available in the emergency-related problems, but the typical features of the sources are simpler; for instance, the release may be localised in time and space or have an easily identifiable chemical or radiological signature.

The two most common advanced DA methods, EnKF and 4D-VAR have been applied to chemical transport modelling since early 2000s. Applications of EnKF include those by van Loon et al. (2000) and Constantinescu et al. (2007). The 4D-VAR method has been used for model initialisation and state analysis experiments in regional air quality modelling (Elbern and Schmid, 1999, 2001; Chai et al. 2006, 2007). Wu et al. (2008) compared OI and 4D-VAR with variants of the Kalman filter in the context of ozone forecasting and found significant forecast improvements during the first 24 h. However, 4D-VAR suffered from the low sensitivity of the forecast to the initial conditions.

Furthermore, it is possible to use a single set of observations to both establish the model initial state and to refine the input emission information, which connects the method to the source identification problems. This allows for using the DA framework to evaluate emission inventories (Kurokawa et al., 2009; Pan et al., 2007; Henze and Seinfeld, 2009; Yumimoto and Uno, 2006). Barbu et al. (2009) used EnKF to assimilate SO2 and SO4 observations, and pointed out complexities arising from the simultaneous assimilation of species with strong chemical connections. In addition, short-term emission adjustments can potentially improve the model forecast scores more and over a longer period of time than simply adjusting the initial state (Elbern et al., 2007).

The 4D-VAR method allows using virtually any type of the observations. For example, it has been used by Yumimoto et al. (2007) for inverse analysis of a dust episode using LIDAR observations of NO2 by Zhang et al. (2008) and Chai et al. (2009), and for assimilation of ozone profiles from the Microwave Limb Sounder (MLS) by Feng et al. (2008). Adjoint methods have also been used for sensitivity studies, such as the analyses of Sandu et al. (2005) and Quelé et al. (2005). The adjoint formalism is equally applicable within the Eulerian and Lagrangian model formulations (Houarin et al., 2006; Houarin et al., 2006).

A rare example of estimating model parameters via inverse dispersion problem solution can be found in Storch et al. (2007) who used concentration observations to find out the parameters of atmospheric boundary layer. Application of Bayesian Monte-Carlo methodology using trajectory models was demonstrated for sensitivity analysis of PM concentrations to the precursor emission by Derwent et al. (2009). However, Monte-Carlo methods are usually very expensive computationally.

The above-mentioned works have shown ways of utilising the data assimilation methods in various areas related to atmospheric dispersion and chemistry modelling. However, the problem of short time interval when the assimilated information affects the forecasts is largely open with limited research (Elbern et al., 2007; Wu et al., 2008) performed so far.

The current paper analyses the efficiency of two variational approaches to data assimilation for AQ forecasting: 3D-VAR-based estimation of the initial state, and 4D-VAR-based simultaneous estimation of the initial concentration fields and the emission fluxes. The relative advantages and disadvantages of the approaches are illustrated by means of two modelling experiments performed for arbitrarily selected episodes in 2000 and 2006.

2. Methods, the modelling tool, and input data

2.1. Variational data assimilation

In this section, we introduce the basic terms for the variational data assimilation (both 3D- and 4D-VAR) and present an adaptation of the 4D-VAR method for the simultaneous estimation of the model initial state and emission fluxes. Notations follow those of Sofiev et al. (2006).

Let us denote the parameter (such as initial state or emission rate) of interest as $\xi$, and define the model operator $M$ mapping the parameter, or the control variable, to a unique phase-space trajectory $x = M_\xi$ defined over some finite time interval referred as the assimilation window (4D-VAR) or at a specific time referred as the analysis time (OI, 3D-VAR). The vector of observations $y$ corresponds to the model state $x$ via the observation operator $H$: $y = H(x) + \epsilon$, where $\epsilon$ is the observation error, which is commonly assumed to be Gaussian.

The maximum likelihood estimate of the parameter $\xi$ is then the value minimising the cost function.

$$J(\xi) = \frac{1}{2}(y - Hx)^T R^{-1}(y - Hx) + \frac{1}{2}(\xi - \xi_0)^T B^{-1}(\xi - \xi_0).$$ (1)

The first term penalises the deviation from the observations $y$ whose accuracy is described by the covariance matrix $R$. The prior knowledge of $\xi$ is included in the background value $\xi_0$ and the background error covariance matrix $B$. The second term therefore penalises the deviation from the prior $\xi_0$. The cost function is minimised using iterative numerical algorithms.

In the 3D-VAR approach, the vector $\xi$ corresponds to the model state $x$, taken at a specific time. The limitation of 3D-VAR — its restriction to the model state at a single time — is overcome in 4D-VAR by using the forecast model for computing the observation vectors for any time. In such case, the gradient of (1) with respect to $\xi$ is presented as

$$J' (\xi) = M^T H^T R^{-1}(y - Hx) + B^{-1}(\xi - \xi_0),$$ (2)

where $M^*$ and $H^*$ are the tangent linear adjoint model and observation operators, respectively (Marchuk, 1995).

The forward dispersion model corresponding to the operator $M$ and defining the time evolution of the model state is defined by the scalar transport equation

$$\frac{\partial c_n}{\partial t} + \frac{\partial}{\partial x}(u c_n) = \frac{\partial}{\partial x} \left( D_n \frac{\partial c_n}{\partial x} \right) + S_n(x, t) + f_n(x, t),$$ (3)

where $c_n$ is a concentration of the $n$-th species, $f_n(x, t)$ is the emission density, and the chemical sources and sinks are included in $S_n(x, t)$. If the reaction term is linear, i.e., $S_n(x, t) = k c_n(x, t)$, then the adjoint equation to (3) reads as (Marchuk, 1995).

$$\frac{\partial c^*}{\partial t} - \frac{\partial}{\partial x}(u c^*) = \frac{\partial}{\partial x} \left( D_n \frac{\partial c^*}{\partial x} \right) + k c_n(x, t).$$ (4)

Here $c^*(x, t)$ is the first-order sensitivity of the functional (1) to a concentration perturbation at time $t$. Its solution corresponds to $M^*$ in (2).
With the adjoint eq. (4) connecting the model state and the emission input, the vector $\xi$ can include the initial atmospheric concentrations $c(t = 0) = c_0$ and the emission density rate $f$, both having known background values of $c_0$ and $f_0$. If these components are assumed to be uncorrelated, the background term in the cost function (1) can be split into a sum of the two parts:

$$\langle \xi - \xi_0 \rangle^T B^{-1} \langle \xi - \xi_0 \rangle = \langle c_0 - c_b \rangle^T B_{c}^{-1} \langle c_0 - c_b \rangle + \langle f - f_b \rangle^T B_{f}^{-1} \langle f - f_b \rangle$$  \hspace{1cm} (5)

where $B_c$ and $B_f$ are the covariance matrices of the initial concentrations and the emission rate, respectively. Thus, the minimisation of the cost function (1) leads to simultaneous determination of the model initial state and emission. The simultaneous consideration of $c_0$ and $f$, as described by eqs. (1)–(5), is an extension to the approach used in meteorological models, where the initial conditions are the only control variable.

2.2. The SILAM dispersion model and input data

The forward and adjoint eqs. (3)–(4) are solved in the current study with the modelling system SILAM version 4, which has two – Eulerian and Lagrangian – dynamic cores. The previous version v.3.5, based on Lagrangian 3D dynamics, was described by Sofiev et al. (2006). The Eulerian core (Sofiev et al., 2008) used in the current experiment is based on the non-diffusive advection scheme of Galperin (2000) and the adaptive vertical diffusion algorithm of Sofiev (2002).

The study was conducted for the sulphur oxides as a prominent example of slowly-reacting species with moderate-to-small removal intensity. SILAM includes a linear chemical mechanism for the SOx compounds following that of the DMAT model (Sofiev, 2000). The SO2 to SO4 conversion is described as a linear temperature-dependent process (see also Seinfeld and Pandis, 1998) with the reaction term given by

$$S(c, t) = \begin{bmatrix} -k_1 & 0 \\ k_1 & 0 \end{bmatrix} \begin{bmatrix} [SO_2] \\ [SO_4] \end{bmatrix}.$$  \hspace{1cm} (6)

where [SO2] and [SO4] are the molar concentrations of the corresponding species. For the adjoint computations, the conversion matrix is transposed.

3. Setup of the modelling experiments

The efficiency of the extension of data assimilation towards the source apportionment was considered using three modelling experiments. The Model Memory (MM) experiment was set up to estimate a characteristic relaxation time period, after which the initial model fields do not significantly affect the results. That time period would also limit the influence of the assimilated observations. The outcome of the MM experiment was used for setting up the data assimilation experiments: one with the 3D-VAR assimilation adjusting the model concentration fields (DA3), and the one combining the simultaneous adjustment of the model concentration fields and input emission fluxes with the 4D-VAR method (DA4). The DA4 experiment had the assimilation window corresponding to that of the MM relaxation period.

3.1. The model memory experiment

For the MM experiment, the SILAM system was run twice over an arbitrarily picked 2-day period (3–4.01.2000) with different initial states but otherwise identical input data. For both runs, the meteorological forcing was taken from archives of operational weather forecasts of the NWP model HIRLAM (Undén et al., 2002). The meteorological fields were taken with 3-h interval and spatial resolution of 30 km. The emission data for anthropogenic and natural SOx emission were taken from the European Monitoring and Evaluation Programme (EMEP) database (http://www.emep.int) for the year 2000. The data are presented as gridded annual totals, with the time variation introduced via the monthly, daily, and hourly scaling coefficients following Sofiev et al. (1996) and Hongisto et al. (2003). The computational grid covered the whole Europe with 30 km resolution and had nine layers along the vertical in the terrain-following $z$-system reaching up to about 8 km in altitude (Fig. 1).

Two sets of initial fields for the runs were created by cold-start two-day pre-computations of the SILAM system. For the first run, the meteorological fields during pre-computation were taken “as-is”, whereas the second pre-computation run used reverse-wind direction, i.e., the signs of all wind components were changed to the opposite while computing the advection. The continuity equation is an invariant to such transformation, except for the tendency term. However, for a limited time period and near the surface the impact of this term is small, so the resulting wind field can still be used to create the initial field for a numerical experiment. The concentration fields at the end of these two pre-computations (Fig. 1, panels a and b) were used as the initial states for the following 2-day-long model runs with the normal meteorology and emission inputs, same for both simulations. As a result, the experiment emulated the 48-h forecast with identical setup but substantially different initial conditions.

3.2. The data assimilation experiments

The DA experiments were conducted for an arbitrarily picked episode in Central and Southern Europe. The considered species were again SO2 and SO4, but the experiment covered a period of two weeks starting from February 8, 2006. The period was divided into 14 overlapping intervals following each other with one day lag. Each interval was two days long. The computations for each interval consisted of the 3D-VAR analysis for the beginning of the interval (DA3), or of the 4D-VAR data assimilation within a window covering the first 24 h (DA4). Both analyses were followed by a 48-h forecast using the assimilated initial state at the beginning and, in case of DA4, adjusted emission rates throughout the interval. A reference run including 5 days of spin-up with no data assimilation provided the background fields for the first day of assimilation. For the subsequent days, the background state was taken from the previous-day analysis.

The DA computation domain covered the area from 15° W to 40.25° E in longitude and from 30° N to 75.25° N in latitude with horizontal resolution of 0.25° (Fig. 2) and the same vertical structure as the MM runs. The model advection time step was 15 min. The meteorological fields were obtained from the operational ECMWF forecasts with 3 h intervals and 0.25° horizontal resolution. The background emission field $f_0$ was obtained from the EMEP SOx emission inventory for 2003 with the same the time variations as in the MM experiment.

In the DA4 experiment, data assimilation was used to adjust the emission flux density after Eq. (5), in addition to the initial state. However, estimating the complete four dimensional emission distribution was impractical due to under-determination of the problem. Therefore, the approach used in this work and shared by previous studies (Yumimoto et al., 2007; Elbern et al., 2007) was to assume a constant relative deviation of the emission intensity from the background rate throughout the assimilation and forecast windows. The adjusted emission rate is written as $\hat{f}(x, t) = \alpha(x)f(x, t)$, where
\( \alpha(x) \) is to be estimated. The diurnal emission variations are thus not affected by the assimilation. To further simplify the procedure, the correction factor \( \alpha(x) \) was assumed to be constant along the vertical axis (height). The sensitivity, and, consequently, the gradient of \( J(x) \) with respect to \( \alpha(x) \) (see eq. (2)) is obtained by integrating the solution of the adjoint problem (4) over the assimilation window and the vertical extent of the model domain. The DA4 minimisation required 15–35 gradient evaluations to converge, resulting in the cost.

Fig. 1. Initial fields (panels a and b) of sulphates and results of 24- (panels c and d) and 48-h (panels e and f) long model simulations. Panels a, c, e represent the run with normal wind during pre-computation phase, panels b, d, f are for reverse wind during pre-computations. Unit \( \mu g \ \text{m}^{-3} \).

Fig. 2. Setup of the DA modelling experiments. Left: the emission rate of SO\(_2\) (mol s\(^{-1}\) cell\(^{-1}\)) over the model domain. Right: the locations of measurement stations used in assimilation (blue) and of the control set (red).
function decreasing typically by a factor of 2–4, but for some days, by a factor of 10.

3.3. Observational data

Observed hourly concentrations of SO\textsubscript{2} were taken from the AirBase database maintained by European Environmental Agency (http://www.eea.europa.eu). The AirBase dataset includes more than 2000 stations over the whole of Europe, but the density of the network varies.

For the data assimilation experiments, we used the data of five European countries (Austria, Czech Republic, Hungary, Italy and Switzerland), totally 456 stations. The stations were divided randomly into the control set of 50 stations and the remaining 406 assimilation stations. However, the assimilation stations covered only 260 model grid cells. For those cells, where several stations were located, one was picked arbitrarily. Therefore, for the assimilation we used 260 sites, while the 50 control stations were excluded from the assimilation and used only for the model-measurement comparison.

The observation error estimates are not available from AirBase and their construction (e.g. Elbernt et al., 2007) is elaborate and prone to uncertainties. Therefore, for both DA3 and DA4, a factor of 10.

3.4. Background error covariance matrices

Analysis of the eqs. (1) and (5) shows that the relative contribution of the three terms – the deviation of the model state from the observations, the deviation of the model state from its background, and the deviation of the emission from its background – to the cost function are determined by the corresponding covariance matrices. Both the shapes (i.e., diagonality) and the absolute values of the elements of these matrices affect the optimal solution.

The background error covariance matrix \( B \) incorporates the assumed standard deviation of the forecast error as well as the correlations between grid points and, possibly, between the species. In absence of correlations between species or grid points, \( B \) becomes diagonal. In presence of the cross-correlations, \( B \) and its inverse can only be formulated implicitly due to the large size of the problem. In the DA3 experiment, two simplifications are made: first, the standard deviation of the background error is assumed constant, and second, the correlation function is written as

\[
\text{corr}(x, x') = \exp\left(-\frac{(x - x'^2)}{L_x^2} - \frac{(y - y'^2)}{L_y^2}\right) \times \text{corr}(z, z'),
\]

where \( x \) and \( x' \) are the positions of two grid points, \( L_x \) and \( L_y \) are the zonal and meridional correlation lengths and \( \text{corr}(z, z') \) is the vertical correlation function.

The parameters in Eq. (7) were estimated following an NMC type approach (Parrish and Derber, 1992) from a set of differences between 12 and 36 h forecasts obtained from the operational SILAM forecasts (http://silam.fmi.fi) between 1–1-2010 and 1-6-2010. The horizontal length scales \( L_x = 71 \text{ km} \) and \( L_y = 78 \text{ km} \) were estimated as a median from the spatial autocorrelation functions calculated for each sample. The vertical correlation function (Fig. 3) was computed from the dataset as the sample correlation matrix.

The form (7) allows factorising \( B \) as a Kronecker product of the zonal, meridional and vertical components (Chai et al., 2007):

\[
B = X \otimes Y \otimes Z
\]

In practice, \( B \) becomes singular as the correlation radius \( L \) increases. This implies that the background error includes components with vanishing variance, corresponding to the smallest eigenvalues of \( B \). Since the expectation of the increment is zero, these components must also vanish. This is ensured by performing the minimisation using a transformed variable \( x = \text{S}x \), where \( \text{S}^{\text{T}}\text{S} = B \). The factor \( \text{S} \) is easily obtained by diagonalising the \( X \), \( Y \) and \( Z \) components separately.

For the DA4 experiment, we assumed that the components of the vector \( \delta \) are all uncorrelated from each other and homogeneous in space. The background covariance matrix then has the form

\[
B = \begin{bmatrix}
\sigma_0^2 & 0 \\
0 & \sigma_1^2
\end{bmatrix}
\]

where \( \sigma_0^2 \) and \( \sigma_1^2 \) are the variances of the background state and the emission correction factor, respectively, and \( I \) is a unit matrix.

The form (9) of the matrices is a strong simplification. However, it is acceptable for the DA4 experiment because (i) correlated background errors are considered in the DA3 experiment, and (ii) the DA4 experiment has a long assimilation window (24 h). The long assimilation window implies a flow-dependent structure for the increments in both initial state and emission corrections. In contrast, the form (7), which simply requires the analysis increments to be sufficiently smooth, is not easily justifiable for the background errors of the emission corrections.

Regarding the specific values of the variances in (9), it is their ratio that plays a role, not the absolute values (the cost function (1) is invariant to scaling). In this study, values of \( \sigma_0 = 10^{-6} \text{ mol m}^{-3} \) and \( \sigma_1 = 3.16 \text{ (relative units)} \) were used. We also put \( R = \sigma_{\text{obs}} I \), \( \sigma_{\text{obs}} = 10^{-7} \text{ mol m}^{-3} \). The implication of this setup is seen if a typical SO\textsubscript{2} concentration of \( 5 \times 10^{-7} - 10^{-8} \text{ mol m}^{-3} \) in the area are taken into account. Then the standard deviation of the observations is declared to be 10–20% of the value itself (larger for lower concentrations), while the uncertainty of the background field is about 100%, and that of emission is over 300%. Therefore, the 3D- and 4D-VAR procedures rather minimise the difference from the observations than from the background fields. The \( \delta_b \) and \( \delta_e \) priors are only weak regularising terms of the minimisation problem, which are unable to constrain it significantly if the signal from the observational data exists. Similarly, if some discrepancy could be eliminated via both initial condition and emission adjustments, the
current experiment setup promotes the emission modification as the correction with presumably longer temporal correlation period.

4. Results of the modelling experiments

4.1. The model memory experiment

The initial spatial correlation coefficient between the two maps, obtained from the pre-computation with normal and reversed wind, was <0.2 for SO2 and <0.1 for the sulphates. Therefore, the initial fields can be considered as uncorrelated. The forecast runs, however, converge fast (Fig. 1). The spatial correlation coefficients rise to 0.82 (0.87) for SO2 (SO4) within 24 h, and to 0.99 (0.995) for SO2 (SO4) by the 48th hour. The evolution of the coefficients (Fig. 4) shows that the impact of the different initial SO2 distributions largely diminishes already after 12 h. Information older than 24 h has practically no influence on the patterns. The model memory is also spatially inhomogeneous: the impact of initial conditions near the sources becomes negligible after just 2–3 h, while remote regions can be influenced by the aged air masses until the meteorological pattern changes (Fig. 1).

Similar convergence of the simulations is seen from the model-measurement comparison although the data are considerably noisier. As an illustration, the results of both runs were compared with observations of the EMEP network (http://www.emep.int) for three sequential days — the last day of pre-computations, the first-, and the second days of the forecasting runs. The sites report daily-mean values, so the SILAM output was averaged accordingly. The model-measurement agreement was estimated via correlation coefficient across the stations. The comparison at the end of the pre-computation expectedly showed large differences in the model-measurement agreement. For SO2, the cold-start run showed $R_{\text{cold,0}} = 0.4$ (significant correlation, 99%) while the reverse-wind pre-computation scored worse: $R_{\text{reverse,0}} = 0.2$ (statistically insignificant). By the end of the forecasting runs, both simulations showed $R_2 = 0.4$ with significance 99%. The scores become practically indistinguishable after one day of the forecast simulations.

As an overall result of the MM experiment, the emission and meteorological forcing overwhelm the influence of initial SOx concentration fields at regional scale within 12–24 h.

4.2. The assimilation experiments

The optimal initial concentration fields obtained within both DA3 and DA4 cases show generally similar features. However, the DA3 increments (Fig. 5) are much smoother than the ones of DA4. This demonstrates the impact of the prescribed spatial correlation distance that was embedded into the 3D-VAR background error covariance matrix, and resulted in a strong smoothing effect. For the 4D-VAR, a diagonal covariance matrix was used, however, the 24-h backwards integration of the adjoint model propagates the observational information along the wind streamlines. As a result, the analysis increment is smoothed along the streamlines while strong variations occur in other directions. The current experiment is insufficient for conclusions on the preferable approach or their combination but highlights the potentially high impact of the a-priori assumed spatial and temporal correlations.

The emission correction factor in the DA4 experiment tends to concentrate on areas with strong sources of SO2 (Fig. 6) and in a few cases appeared persistent over the considered period. The most

Fig. 4. Evolution of the spatial correlation coefficient between the two runs. $t = 0$ represents the end of pre-computation and start of the unconstrained 2-day run of the model.

Fig. 5. Increments (analysis – background) in initial concentrations of SO2 for the 12th simulation (i.e., the day 12) of the DA experiments. Left-hand panel: 4D-VAR, right-hand panel: 3D-VAR. Unit: mg SO2 m$^{-3}$
persistent emission corrections were found near the volcano Etna where the SO2 levels at the stations in Sicily were strongly over-estimated in the reference run. The 4D-VAR projected this discrepancy to the emission rate of the volcano. Consequently, the total emission (anthropogenic plus natural) of the eastern part of Sicily and the Aeolian islands was reduced by ~46% (mean over the whole period), from 24.9 Gg to 13.4 Gg. However, the emission in some grid cells was frequently scaled down by an order of magnitude.

The other area with a stable emission correction covers parts of Czech Republic and Hungary. The emission rates of the sources in these areas were increased by factors of 1.9 and 1.8, respectively (mean over the whole period). Throughout the rest of the domain, the corrections were quite irregular and, if averaged over the considered period, small.

The RMSE for the control stations was reduced by the data assimilation by 20–80% when both the adjustment of initial fields and the emission correction factor were applied. The main improvement was expectedly seen at the stations near Mt. Etna. For other areas, the improvement was highly variable with a general tendency of higher impact at the sites with poor scores in the reference run (Fig. 7). Thus, the upper RMSE percentile (corresponding to the poorly performing sites) almost halved as a result of 4D-VAR. The improvement is quite irregular in time (Fig. 7, right-hand panel), and deteriorates towards the end of the forecast window (Fig. 7, left-hand panel, see also the next section for discussion of the reasons for the deterioration). The effect of assimilation during the considered two weeks varies from almost zero up to about 30% of the daily-mean RMSE. The improvement for the better performing sites (the median and lower percentile curves in Fig. 7, both panels) is
The impact of the emission correction is illustrated in the right-hand panel of Fig. 8. The depicted site is located in the southern Italy with several assimilated stations around, so the impact of the emission correction lasts over the whole period. In particular, the false peak during the forecast run was eliminated by the adjusted emission.

5.2. Variability of the assimilation impact

The reasons for the uneven improvement of the model scores (Fig. 7) include the varying representativeness of the sites in complex terrain, as well as the uncertainty in the distribution of emission sources. However, one of the main reasons for the mixed impact seems to be the dynamic change of transport conditions — primarily, the wind pattern. Gradual changes in the transport and emission conditions are also likely to contribute to the deterioration of forecast scores towards the end of the forecast window.

The effect of the transport conditions can be tested by evaluating the sensitivity distribution (footprint) for each measurement station by solving the adjoint transport eq. (4).

Since the control variable in the DA4 experiment includes the emission scaling factor \(a(x,t)\), the sensitivity of concentration \(c(x_1,t_1)\) at a location \(x_1\) and time \(t_1\) to a unit perturbation of \(a(x_0,t_0)\) will be

\[
\delta c(x_1,t_1) = \int_0^L (x_0, z, t_0) \delta^* c(x_0, z, t_0) \, dz
\]

where \(\delta^* c\) is the solution of the eq. (4) with the source term \(\delta(x - x_1, t - t_1)\).

As follows from eq. (10), adjusting the emission rates can affect a station only if its footprint during the forecast window overlaps with the footprints of the assimilated stations during the assimilation window. The shapes of the footprints then affect the conditioning of the problem: if they are wide and lack a clear overlapping peak, the sensitivity to any single adjustment is low, and the optimal emission distribution is uncertain due to poor signal-to-noise ratio.

Examples of the footprints are shown in Fig. 9 for three sets of the stations: (i) the sites used in the assimilation (panel a), (ii) sites from the control set with the RMSE improvement <5% (small impact of the assimilation, panel b), (iii) sites from the control set with the RMSE improvement >30% (strong impact of the assimilation, panel c). It is seen that the footprints of the control stations with large impact of assimilation are concentrated around some
sources and well correlated with the footprints of the assimilated sites (e.g. around Mt. Etna). The sensitivity areas of the stations with small impact due to assimilation have different shapes and cover larger area than those of the assimilated sites.

There is, however, also a significant overlap between the footprints of low-impact and the assimilation sites, which means that assimilation of some observations did not improve the forecast. Apart from the above-mentioned poor signal-to-noise ratio for wide footprints, this may also indicate a violation of the underlying assumptions of the assimilation method. Firstly, in the standard “strong constraint” 4D-VAR, the modelling errors are assumed to be negligible during the assimilation window: the disagreement with the observations is entirely attributed to emission and initial state fields. Secondly, the emission correction obtained during the assimilation window is kept constant throughout both the assimilation and forecast windows. For some sources, the constant-correction assumption may be too restrictive to be used as a strong constraint, and a weak constraint formulation of 4D-VAR (Trémolet, 2006) might result in a more lasting and homogeneous improvement.

Analysis of the footprints also suggests that in several cases the area covered by the assimilated stations was too small. For example, the sources in the southern France (Fig. 9) were not adjusted but affected some low-impact control sites. Extending the observation network to cover all major source areas would increase the impact of data assimilation.

6. Conclusions

The presented modelling experiments compared two ways of utilising observational information in AQ forecasting: adjusting the initial concentration fields and correcting the emission rates. The initial conditions had practically no influence on the concentration of sulphur oxides after 12–24 h in the modelling experiment. Simultaneously adjusting the initial fields and emission rates using 4D-VAR resulted in substantially longer impact on the model forecasts. Apart from the choice of control variables (initial conditions and/or emission rates), the efficiency of data assimilation strongly depends on the density and coverage of the observational network. For sulphur oxides, it seems to be mandatory to include the most of European continent in order to obtain a long-lasting improvement. However, even for the limited area covered by the presented simulations the model performance improved substantially for the sites with poor initial scores.

Fig. 9. The sensitivity of measurements with respect to emission correction factor for the seventh day of the assimilation experiment. Panel a: the stations used in assimilation cycle. Panels b and c: stations with less than 5% improvement (b) and > 30% improvement (c). The stations are marked by black dots.
Acknowledgements

The work has been performed within the EU projects GEMS, MACC, MEGAPOLI and TRANSForeM. The support of the ISFIRE project of Academy of Finland and COST-EOS062 is kindly acknowledged. The quasi-Newton minimisation code m1np3 was used for the 3D-VAR computations.

References


Assimilation of surface NO$_2$ and O$_3$ observations into the SILAM chemistry transport model

J. Vira and M. Sofiev
Finnish Meteorological Institute, Helsinki, Finland

Correspondence to: J. Vira (julius.vira@fmi.fi)

Received: 17 July 2014 – Published in Geosci. Model Dev. Discuss.: 15 August 2014
Revised: 12 January 2015 – Accepted: 13 January 2015 – Published: 6 February 2015

Abstract. This paper describes the assimilation of trace gas observations into the chemistry transport model SILAM (System for Integrated modelLing of Atmospheric composition) using the 3D-Var method. Assimilation results for the year 2012 are presented for the prominent photochemical pollutants ozone (O$_3$) and nitrogen dioxide (NO$_2$). Both species are covered by the AirBase observation database, which provides the observational data set used in this study.

Attention was paid to the background and observation error covariance matrices, which were obtained primarily by the iterative application of a posteriori diagnostics. The diagnostics were computed separately for 2 months representing summer and winter conditions, and further disaggregated by time of day. This enabled the derivation of background and observation error covariance definitions, which included both seasonal and diurnal variation. The consistency of the obtained covariance matrices was verified using $\chi^2$ diagnostics.

The analysis scores were computed for a control set of observation stations withheld from assimilation. Compared to a free-running model simulation, the correlation coefficient for daily maximum values was improved from 0.8 to 0.9 for O$_3$ and from 0.53 to 0.63 for NO$_2$.

1 Introduction

During the past 10–15 years, assimilating observations into atmospheric chemistry transport models has been studied with a range of computational methods and observational data sets. The interest has been driven by the success of advanced data assimilation methods in numerical weather prediction (Rabier, 2005), as well as by the development of operational forecast systems for regional air quality (Kukkonen et al., 2012). Furthermore, the availability of remote sensing data on atmospheric composition has enabled construction of global analysis and forecasting systems, such as those described by Benedetti et al. (2009) and Zhang et al. (2008). Assimilation of satellite observations into stratospheric chemistry models has been demonstrated, e.g. by Errera et al. (2008).

Data assimilation is defined (e.g. Kalnay, 2003) as the numerical process of using model fields and observations to produce a physically and statistically consistent representation of the atmospheric state – often in order to initialise the subsequent forecast. The main techniques used in atmospheric models include the optimal interpolation (OI, Gandin 1963), variational methods (3D-Var and 4D-Var, Le Dimet and Talagrand, 1986; Lorenc, 1986), and the stochastic methods based on the ensemble Kalman filter (EnKF, Evensen, 2003, 1994). Each of the methods has been applied in air quality modelling. Statistical interpolation methods were used by Blond and Vautard (2004) for surface ozone analyses and by Tombette et al. (2009) for particulate matter. The EnKF method has been utilised by several authors (Constantinescu et al., 2007; Curier et al., 2012; Gaubert et al., 2014) especially for ozone modelling. The 3D-Var method has been applied in regional air quality models by Jaumouillé et al. (2012) and Schwartz et al. (2012), while the computationally more demanding 4D-Var method has been demonstrated by Elbern and Schmidt (2001) and Chai et al. (2007). Partly due to its significance in relation to health effects, the most commonly assimilated chemical component has been ozone.

The performance of most data assimilation methods depends on correctly prescribed background error covariance
matrices (BECM). This is particularly important for 3D-Var, where the BECM is prescribed and fixed throughout the whole procedure, in contrast to the EnKF based assimilation methods, where the BECM is described by the ensemble of states, and to the 4D-Var method, where the BECM is prescribed but evolves implicitly within the assimilation window.

A range of methods of varying complexity have been employed to estimate the BECM in previous studies on chemical data assimilation. The “National Meteorological Centre” (NMC) method introduced by Parrish and Derber (1992) is based on using differences between forecasts, with differing lead times as a proxy for the background error. Kähnert (2008), as well as Schwartz et al. (2012), applied the NMC method for estimating the BECM for assimilation of aerosol observations. Chai et al. (2007) based the BECM on a combination of the NMC method and the observational method of Hollingsworth and Lömberg (1986). This observational method was also used by Kumar et al. (2012) for the assimilation of NO2 and O3 data.

The BECM can also be estimated using ensemble modelling; this approach was taken by Massart et al. (2012) for regional ozone analyses. Finally, Desroziers et al. (2005) presented a set of diagnostics, which can be used to adjust the background and observation error covariances. This method has been previously applied in chemical data assimilation for example by Schwerger and Elbern (2010) and Gaubert et al. (2014).

In contrast to short and medium range weather prediction, the influence of initial condition on an air quality forecast has been found to diminish as the forecast length increases. For ozone, Blond and Vautard (2004) and Wu et al. (2008) found that the effect of the adjusted initial condition extended for up to 24 h. Among other reactive gases, NO2 has been a subject for studies of Silver et al. (2013) and Wang et al. (2011). However, the shorter lifetime of NO2 limits the timescale for forecast improvements especially in summer conditions.

An approach for improving the effectiveness of data assimilation for short-lived species is to extend the adjusted state vector with model parameters. Among the possible choices are emission and deposition rates (Bocquet, 2012; Curier et al., 2012; Elbern et al., 2007; Vira and Sofiev, 2012).

The aim of the current paper is to describe and evaluate a regional air quality analysis system based on assimilating hourly near-surface observations of NO2 and O3 into the SILAM chemistry transport model. The assimilation scheme was initially presented by Vira and Sofiev (2012); in the current study, the scheme is applied to photochemical pollutants and moreover, we discuss how its performance can be improved by introducing statistically consistent background and observation error matrices. The analysis fields are produced for the assimilated species at an hourly frequency using the standard 3D-Var assimilation method (Lorenc, 1986). The diagnostics of Desroziers et al. (2005) are applied in this work for estimating the background and observation error standard deviations, notably resolving their seasonal and diurnal variations. The evaluation is performed for the year 2012 using stations withheld from assimilation. In addition to assessing the analysis quality, the effectiveness of assimilation for initialising the model forecasts is evaluated.

The following Sect. 2 presents the model setup and briefly reviews the 3D-Var assimilation method. The procedure for estimating the background and observation error covariance matrices is discussed in Sect. 3. The assimilation results for O3 and NO2 for the year 2012 are discussed in Sect. 4. Section 5 concludes the paper.

2 Materials and methods

This section presents the SILAM dispersion model, the observation data sets used, and describes the assimilation procedure.

2.1 The SILAM dispersion model and experiment setup

This study employed the SILAM chemistry transport model (CTM) version 5.3. The model utilises the semi-Lagrangian advection scheme of Galperin (2000) combined with the vertical discretisation described by Sofiev (2002) and the boundary layer scheme of Sofiev et al. (2010). Wet and dry deposition were parameterised as described in Sofiev et al. (2006).

The chemistry of ozone and related reactive pollutants was simulated using the carbon bond 4 chemical mechanism (CB4, Gery et al., 1989). However, the NO2 analyses were produced with separate simulations employing the DMAT chemical scheme of Sofiev (2000). This follows the setup used in operational air quality forecasts with the SILAM model, where the two model runs are necessary since the primary and secondary inorganic aerosols are only included in the DMAT scheme. The SILAM model has been previously applied in simulating regional ozone and NO2 concentrations (Huijnen et al., 2010; Langner et al., 2012; Solazzo et al., 2012), for global-scale aerosol simulations (Sofiev et al., 2011), as well as for simulating emission and dispersion of allergenic pollen (Siljamo et al., 2013). The daily, European-scale air quality forecasts contributing to the MACC-II project are publicly available at http://macc-raq.gmes-atmosphere.eu.

In this study, the model was configured for a European domain covering the area between 35.2 and 70.0° N and −14.5 and 35.0° E with a regular long–lat grid. The vertical discretisation consisted of eight terrain-following levels reaching up to about 6.8 km. The vertical coordinate was geometric height. The model was driven by operational ECMWF IFS forecast fields, which were initially extracted in a 0.125° long–lat grid and further interpolated to the CTM resolution. Chemical boundary conditions were provided by the MACC.

Geosci. Model Dev., 8, 191–203, 2015

www.geosci-model-dev.net/8/191/2015/
reanalysis (Inness et al., 2013), which uses the MOZART global chemistry-transport model.

The emissions of anthropogenic pollutants were provided by the MACC-II European emission inventory (Kuenen et al., 2014) for the reference year 2009. The biogenic isoprene emissions, required by the CB4 run, were simulated by the emission model of Poupkou et al. (2010).

Three sets of SILAM simulations were carried out in this study. First, the background and observation error covariance matrices were calibrated using 1-month simulations for June and December 2011. The calibration results were used in reanalysis simulations covering the year 2012. Finally, a set of 72-h hindcasts was generated for the period between 16 July and 5 August 2012, to evaluate the forecast impact of assimilation. The hindcasts were initialised from the 00:00 UTC analysis fields. The timespan included an ozone episode affecting parts of southern and western Europe (EEA, 2013). The reanalysis and hindcasts use identical meteorological and boundary input data, and hence, the hindcasts only assess the effect of chemical data assimilation.

The analysis and forecast runs were performed at a horizontal resolution of 0.2°. The setup for calibrations runs (June and December 2011) was identical except that a coarser horizontal resolution of 0.5° was chosen in order to reduce the computational burden. The model time step was 15 min for both setups.

2.2 Observations

This study uses the hourly observations of NO₂ and O₃ at background stations available in the AirBase database (http://acm.eionet.europa.eu/databases/airbase/) maintained by the European Environmental Agency. Separate subsets are employed for assimilation and evaluation.

Two sets of stations were withheld for evaluation. The first set, referred to here as the MACC set, had been used in the regional air quality assessments within the MACC and MACC-II projects (Rouïl, 2013, also Curier et al., 2012). The second set consisted of the stations reported as EMEP stations in the database. The MACC validation stations included about a third of the available background stations for each species, and were chosen with the requirement to cover the same area as the assimilation stations. The EMEP network is sparser and has no particular relation to the assimilation stations. It can be noted that the EMEP stations included in AirBase do not comprise the full EMEP monitoring network.

The in situ data are used for assimilation and evaluation under the assumption that they represent the pollutant levels in spatial scales resolved by the model. We expect this assumption to be violated, especially at many urban and suburban stations due to local variations in emission fluxes. For this reason, only rural stations were used for evaluation of the 2012 reanalysis. The NO₂ assimilation set also excluded both urban and suburban stations. For ozone, the data from suburban stations were assimilated, however, the observation errors were assessed separately for suburban and rural stations, as outlined in Sect. 3. The station sets are presented on a map in Fig. 1.

The statistical indicators used for model evaluation were correlation, mean bias and root mean squared error (RMSE). Since air quality models are frequently used to evaluate daily maximum concentrations, the indicators were evaluated separately for the daily maximum values.

2.3 The 3D-Var assimilation

In the 3D-Var method, the analysis $x_a$ minimises the cost function:

$$J(x) = \frac{1}{2} (y - \mathcal{H}(x))^T R^{-1} (y - \mathcal{H}(x)) + \frac{1}{2} (x - x_b)^T B^{-1} (x - x_b), \quad (1)$$

where $x_b$ is the background state, $y$ is the vector of observations, and $\mathcal{H}$ is the possibly nonlinear observation operator. The uncertainties of the background state $x_b$ and the observations $y$ are described by the background and observation error covariance matrices $B$ and $R$, respectively. In this study, the control variable $x$ consisted of the three-dimensional airborne concentration for either NO₂ or ozone. The m1qn3 minimisation code (Gilbert and Lemaréchal, 1989) was used for solving the optimisation problem Eq. (1).

For the surface measurements, the operator $\mathcal{H}$ was linear and consisted of horizontal interpolation only, since the surface concentrations were considered to be represented by the lowest model level. Following the hourly observation frequency, the analysis was performed every hour followed by a 1 h forecast. The forecast provides the background field for the subsequent analysis.

In the current study, only a single chemical component was assimilated in each run. Since O₃ is not a prognostic variable in the DMAT scheme, it cannot be assimilated into the NO₂ simulation. Assimilating NO₂ observations into the CB4 simulation would be technically feasible; however, simultaneous assimilation of NO₂ and O₃ would require care due to the strong chemical coupling between the species. The background and observation error covariance matrices would also need to be jointly estimated.

3 Background and observation error covariance matrices

The numerical formulation of the BECM in the current work follows the assumptions made by Vira and Sofiev (2012). We assume that the background error correlation is homogenous in space, and its horizontal component is described by a Gaussian function of distance between the grid points. Furthermore, we assume that the background error standard deviation $\sigma_b$ is independent of location. This allows writing
the BECM as $B = \sigma_b^2 C$, where $C$ is the correlation matrix and $\sigma_b$ is the background error standard deviation.

For estimation of the parameters for the covariance matrices $B$ and $R$, we combined the NMC method, which is used for determining the correlation matrix $C$, and the approach of Desroziers et al. (2005), which is used for diagnosing the observation and background error standard deviations.

In the NMC method, the difference between two forecasts valid at a given time is taken as a proxy of the forecast error. In this work, the proxy data set was extracted from 24 and 48 h regional air quality forecasts for the year 2010. The forecasts are generated with the SILAM model in a configuration similar to the one used in this study. Since no chemical data assimilation was used in the forecasts, the differences were due to changes in forecast meteorology and boundary conditions only. The lead times were chosen to allow sufficient spread to develop between the forecasts. The forecast data were segregated by hour, resulting in separate sets for hours 00:00, 06:00, 12:00 and 18:00 UTC, and the correlations interpolated for all other times of day.

The horizontal and vertical components of the correlation matrix $C$ were estimated separately. The horizontal correlation was determined by the length scale $L$, which was obtained by fitting a Gaussian correlation function to the data set. First, the sample correlation matrix $\tilde{C}$ of the forecast differences was calculated. Then, the Gaussian correlation function was fitted to the empirical correlations $\tilde{C}_{ij}$ by minimising

$$ f(L) = \sum_{|r_i - r_j| < d} |\tilde{C}_{ij} - C_{ij}(L)|^2, $$

where the fitted correlation function is $C_{ij}(L) = \exp(-(|x_i - x_j|^2 + |y_i - y_j|^2)/L^2)$ and $x$ and $y$ are the Cartesian coordinates for each grid point. To reduce the effect of spurious long-distance correlations due to the limited sample size, the fitting was restricted to grid points $r_i$ closer than $d = 1000$ km to each other. The distances, shown in Table 1, were computed for the lowest model layer.

The vertical correlation function was obtained directly as the sample correlation across all vertical columns for each time of day. As an example, the correlation matrix obtained for NO$_2$ at 12:00 UTC is shown in Fig. 2.

Since the NMC data set includes only meteorological perturbations, it is expected to underestimate the total uncertainty of the CTM simulations. Hence, the standard deviations were not diagnosed from the NMC data set, but instead, an approach based on a posteriori diagnostics was taken. The
Table 2. The $\chi^2/N$ consistency indicator and RMSE on rural MACC validation stations during the first and fifth iterations for tuning the observation and background error standard deviations.

<table>
<thead>
<tr>
<th></th>
<th>O3</th>
<th>NO2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2/N$</td>
<td>RMSE</td>
</tr>
<tr>
<td>June</td>
<td>First guess</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Fifth iteration</td>
<td>1.05</td>
</tr>
<tr>
<td>December</td>
<td>First guess</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Fifth iteration</td>
<td>1.05</td>
</tr>
</tbody>
</table>

approach, devised by Desroziers et al. (2005), is based on a set of identities, which relate the BECM and OECM if to expressions that can be estimated statistically from a set of analysis and corresponding background fields.

First, the standard deviation $\sigma^{(i)}_{\text{obs}}$ of the $i$th observation component is equal to

$$E[(y^{(i)} - y^{(i)}_{a})(y^{(i)} - y^{(i)}_{b})] = \sigma^{(i)}_{\text{obs}}^2,$$

where $E$ denotes the expectation, $y$ is the observation vector and $y_a = H(x_a)$ and $y_b = H(x_b)$ are evaluated from the analysis and background fields, respectively.

The background error covariance matrix cannot be uniquely expressed in observation space. However, assuming that each observation only depends (linearly) on a single model grid cell (i.e., horizontal interpolation is neglected), then:

$$E[(y^{(i)}_{a} - y^{(i)}_{b})(y^{(i)} - y^{(i)}_{b})] = \sigma^{(i)}_{b}^2.$$  

The identities (3) and (4) hold for an ideally defined analysis system, provided that the background and observation errors are normally distributed and assuming the observation operator is not strongly nonlinear.

Furthermore, Eqs. (3) and (4) can be used to tune the parameters $\sigma_{\text{obs}}$ and $\sigma_b$ by means of fixed point iteration. First, a set of analyses is produced using initial parameter values. Then, the left-hand sides of Eqs. (3) and (4) are evaluated as averages over the analyses, resulting in new parameter values. The procedure is then repeated using the updated $\sigma_b$ and $\sigma_{\text{obs}}$ to produce a new set of analyses. In this work, we stopped the iteration when the RMSE at validation stations was no longer improving. We chose this criterion to avoid overfitting the parameters to the calibration data.

In this work, the observation error covariance matrix $R$ was assumed diagonal. The initial values for $\sigma_{\text{obs}}$ and $\sigma_b$ were set to 11.2 and 20.6 $\mu g \text{ m}^{-3}$ for O3, and 4.0 and 8.0 $\mu g \text{ m}^{-3}$ for NO2. The values corresponded to typical mean-squared errors for a free-running model, which were attributed to the model and observation error variances in the ratio of 80/20, respectively. The standard deviations, together with the correlation matrices obtained with the NMC procedure, were then employed in the iterations to calculate a set of hourly analyses for the two calibration periods spanning June and December 2011.

The choice of calibration periods representing both winter and summer conditions was motivated by the strong seasonal variations in both O3 and NO2. Both $\sigma_{\text{obs}}$ and $\sigma_b$ were segregated by hour, while for O3 $\sigma_{\text{obs}}$ was also evaluated separately for suburban stations. For the reanalysis of year 2012, the standard deviations, obtained separately for June and December, were interpolated linearly for all other months.

Finally, the overall consistency could be evaluated by checking the identity (Ménard et al., 2000)

$$E(\chi^2) = N,$$

where $\chi^2 = 2J(x_a)$ is twice the value of cost function (1) at the minimum, and $N$ is dimension of the observation vector $y$. The identity (5) tests the overall consistency of the analysis and is affected by both $B$ and $R$.

4 Results and discussion

The SILAM model was run for year 2012 with and without assimilation. Since the 3D-Var analyses require no additional model integrations in form of iterations or ensemble simulations, the hourly analyses increased the simulation runtime by only 10–15%.

The effect of assimilation to the yearly-mean concentrations on the lowest model level is shown in Fig. 3. On average, the ozone concentrations are increased by the assimilation especially around the Mediterranean Sea, which indicates corresponding low bias in the free model run. The main changes in NO2 levels are confined to somewhat more limited areas; in particular, areas near major mountain ranges (Alps and Pyrenees) show enhanced NO2 levels in the analysis run.

4.1 Background and observation error covariance matrices

Refining the background and observation standard deviations iteratively both improves the consistency of the assimilation setup as measured by the $\chi^2$ indicator (Eq. 5), and improves the model–measurement comparison on the validation stations over the calibration period. However, after five iterations (for both June and December), the changes in $\chi^2$ become slow and the validation scores no longer improve. Hence, the values for $\sigma_{\text{obs}}$ and $\sigma_b$ in fifth iterations were taken as the final values for 2012 reanalysis. The changes in $\chi^2$ and model-measurement RMSE are summarised in Table 2.

The diagnosed observation and background error standard deviations for O3 and NO2 are shown in Fig. 4. For June, the standard deviations for ozone range between 11 and 21 $\mu g \text{ m}^{-3}$ for rural stations. For December, the diurnal variation is flatter, but the absolute values are generally not reduced, in contrast to the overall seasonality of O3.
Especially for summertime night conditions, the values are higher than the values adopted in most of the earlier studies (Chai et al., 2007; Curier et al., 2012; Jaumouillé et al., 2012). However, the errors are comparable to the observation errors diagnosed using the CHIMERE model by Gaubert et al. (2014). The main error component is likely to be due to lack of representativeness: using the AIRNOW observation network, Chai et al. (2007) found standard deviations between 5 and 13 ppb for observations inside a grid cell with 60 km resolution. The maximum values occurred during nighttime.

The diagnosed observation and background error parameters are subject to uncertainty, since they are not uniquely determined (Schwinger and Elbern, 2010). Also, the parameters depend on the assumptions made regarding the correlation function. Nevertheless, the relative magnitude of observation errors during nighttime is interesting for interpreting the model-to-measurement comparisons.

The diagnosed background errors for ozone are between 5 and 9 μg m\(^{-3}\) depending on month and time of day. For June, the diagnosed errors are largest between 09:00–10:00 and 21:00–22:00 UTC, which coincides with transitions between stable and convective boundary layers in summertime conditions. For December, only minor diurnal variation is observed.

The observation error standard deviation for NO\(_2\) varies between 2.8 and 5.2 μg m\(^{-3}\) for rural stations. Suburban or urban stations were not assimilated for NO\(_2\). Contrary to ozone, the diurnal variation of background and observation errors both positively correlate with the diurnal variation of the pollutant.

The BECM and OECM were adjusted to optimise self-consistency for 2 months in 2011. To assess the robustness of the obtained formulations, the \(\chi^2\) indicator was also computed for all analysis steps for the 2012 reanalysis simulation.

As seen in Fig. 5, the analyses using the adjusted BECM and OECM generally satisfy the consistency relation better throughout the year, when compared to the first guess values. The yearly-mean values for \(\chi^2\) are 1.05 and 0.97 for ozone and NO\(_2\), respectively.

Overall, the assimilation system is based on rather simplistic assumptions regarding the background and observation error statistics. In addition to computational efficiency, this approach benefits from having few tuning parameters, and the remaining parameters (\(\sigma_{\text{obs}}, \sigma_b, L\)) can be estimated using an automated procedure. As shown in the following
section, the refined background and observation error definitions provide a clear improvement on analysis scores at the control stations, despite the rather limited training data sets.

4.2 Evaluation against independent observations

Tables 3 and 4 present the analysis skill scores for runs with both first guess and final BECM and OECM, and for the free-running model with no assimilation.

In terms of correlation and RMSE, both analysis and free model runs show better performance for predicting the daily maximum than hourly values. This applies to both O3 and NO2, although the difference is more marked for ozone. The opposite holds true for bias, which tends to be higher when calculated for daily maxima.

The comparison reveals a number of contrasts between the “MACC” and “EMEP” validation stations. First, the free-running model shows better performance for NO2 on the EMEP stations, while for ozone, the performance is better on the MACC stations. On the other hand, the data assimilation has a stronger impact on the scores for the MACC validation stations. This is especially visible in the case for NO2, a result that is consistent with the shorter lifetime of NO2 compared to O3.

The differences largely originate from the different representativeness and coverage of the MACC and EMEP station sets. As seen in Fig. 1, the EMEP network covers the computational domain more evenly than the MACC validation stations, which are concentrated in central Europe. Since the coverage of assimilation and MACC validation stations is similar, the average impact of assimilation is stronger on the MACC than EMEP stations.

For the free-running simulations, the better performance for O3 at the MACC stations is consistent with the geographical variations in the model skill; the densest coverage of the MACC validation stations coincides with the parts of Europe where many regional air quality models perform best for ozone (e.g. Vautard et al., 2009). The scores for NO2 also vary by region, however, due to the shorter chemical lifetime, the forecasts of NO2 are more sensitive to unresolved variations in local emissions. This probably explains the better scores for NO2 on the EMEP stations, since the EMEP network is specifically aimed at monitoring the background levels of pollutants.

For ozone, the assimilation had a variable effect on the model bias. While the correlation and RMSE were always improved by assimilation, the analyses have slightly larger negative mean bias (−4.6 vs. −4.0 μg m⁻³ on MACC stations) than the free model. This is confirmed by the average diurnal profile shown in Fig. 6. However, the diurnal variation of analysis errors is flatter, and the strongest bias no longer coincides with the afternoon hours, when the highest O3 concentrations are typically observed.

For NO2, the analyses have only slight negative bias (−0.38 μg m⁻³) on the MACC stations, which turns positive (about 1 μg m⁻³) for the more remote EMEP sites. As seen in Table 4, the difference between the station sets is similar to that of the free-running model. Given the differences between the MACC and EMEP station sets, this suggests that the model overestimates the lifetime of NO2, which in turn results in the positive bias in the analyses. The long lifetime of NO2 in the SILAM DMAT chemistry scheme was also noticed by Huijnen et al. (2010).

The analysis scheme assumes an unbiased model, and hence, the negative bias present in the free-running simulations is reduced but not removed in the analysis fields. The assimilation setup including tuned OECM and BECM produces more biased analyses compared to the first guess setup, as seen in Fig. 6. This is a consequence of the differences between the diagnosed and first guess background and observation error standard deviations. Contrary to the tuned setup, the first guess attributes most of the model-observation discrepancy to the background error, which results in stronger increments towards the observed values. Consequently, the analysis bias is smaller. However, the tuned assimilation setup has consistently better RMSE and correlation than the first guess assimilation setup.
Table 3. Comparison of performance indicators for ozone in the 2012 reanalysis. The scores are given for station sets “MACC” and “EMEP” as defined in Sect. 2.2. For the analysis runs, scores are shown for the different background error covariance matrices discussed in Sect. 3. The unit of bias and RMSE is μg m⁻³.

<table>
<thead>
<tr>
<th></th>
<th>Hourly</th>
<th></th>
<th>Daily maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corr</td>
<td>Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td>MACC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No assimilation</td>
<td>0.67</td>
<td>-4.00</td>
<td>24.91</td>
</tr>
<tr>
<td>Assimilation, first guess B</td>
<td>0.77</td>
<td>-4.62</td>
<td>21.35</td>
</tr>
<tr>
<td>Assimilation, final B</td>
<td>0.8</td>
<td>-4.64</td>
<td>19.2</td>
</tr>
<tr>
<td>EMEP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No assimilation</td>
<td>0.58</td>
<td>-6.32</td>
<td>24.06</td>
</tr>
<tr>
<td>Assimilation, first guess B</td>
<td>0.66</td>
<td>-5.79</td>
<td>21.83</td>
</tr>
<tr>
<td>Assimilation, final B</td>
<td>0.68</td>
<td>-6.00</td>
<td>20.22</td>
</tr>
</tbody>
</table>

Table 4. Comparison of performance indicators for NO₂ in the 2012 reanalysis. The station sets MACC and EMEP and assimilation options are as in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>Hourly</th>
<th></th>
<th>Daily maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corr</td>
<td>Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td>MACC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No assimilation</td>
<td>0.50</td>
<td>-1.18</td>
<td>9.01</td>
</tr>
<tr>
<td>Assimilation, first guess B</td>
<td>0.58</td>
<td>-0.25</td>
<td>8.6</td>
</tr>
<tr>
<td>Assimilation, final B</td>
<td>0.6</td>
<td>-0.38</td>
<td>8.04</td>
</tr>
<tr>
<td>EMEP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No assimilation</td>
<td>0.52</td>
<td>0.47</td>
<td>6.19</td>
</tr>
<tr>
<td>Assimilation, first guess B</td>
<td>0.55</td>
<td>1.17</td>
<td>6.45</td>
</tr>
<tr>
<td>Assimilation, final B</td>
<td>0.57</td>
<td>0.99</td>
<td>5.92</td>
</tr>
</tbody>
</table>

Since the analysis bias is mainly a consequence of a bias in the forecast model, the bias should be addressed primarily by improving the model. As shown by Dee (2005), model biases can, in principle, also be addressed by the assimilation system. However, a possible bias correction scheme should be implemented with care, since due to representativeness errors, also observational biases could arise.

In addition to computing the regular statistical indicators for daily maxima, we evaluated the hit rates (the number of correctly predicted exceedances divided by the number of observed exceedances) for the 180 μg m⁻³ threshold for O₃, with and without assimilation. Assimilation also turns out to improve the hit rate, albeit only slightly: from 0.25 to 0.26 on average for rural MACC validation stations, and from 0.13 to 0.15 for EMEP stations. If the averaging is restricted to the stations with more than 10 exceedances during 2012, the values change from 0.32 to 0.36 for MACC and from 0.21 to 0.43 for the EMEP stations. Obviously, the hit rates are sensitive to the low bias in the daily maxima.

For NO₂, a specific source of observational errors is due to the molybdenum converters used in the chemiluminescence technique, which is the most common measurement technique for monitoring NO₂. As discussed by Dunlea et al. (2007) and Steinbacher et al. (2007), this technique is subject to positive interference by the NO₃ species such as PAN, HNO₃ and HONO.

The interference can lead to overestimation of NO₂ by up to a factor of two, however, the error varies by location and time, and may depend on the features of the instrument (Steinbacher et al., 2007). We estimated the magnitude of this effect from the free-running CB4 simulation. On most continental EMEP sites, the contribution of the NO₂ species to the total NO₂ + NO₃ was about 10–20 % of the simulated yearly mean. However, for a few sites the contribution could reach 50 %.

The O₃ and NO₂ observations were assimilated into separate model runs. Assimilation of O₃ had only a minor influence on NO₂ in the CB4 simulation; however, the mean bias was reduced by about 5 % on average for the MACC validation stations. Because the DMAT simulation does not include ozone as a tracer, the impact of NO₂ assimilation on ozone fields was not evaluated in this study.

4.3 Forecast experiments

In order to quantify the usefulness of data assimilation in forecast applications, a set of simulations without data assimilation were generated using the analysis fields at 00:00 UTC as initial conditions. The forecast experiment covered the time period between 16 July and 5 August 2012. The effect of chemical data assimilation on forecast performance was assessed as a function of the forecast lead time. Figures 7 and 8 present the correlation and bias for the O₃
and NO$_2$ forecasts, respectively, and compare them to the corresponding indicators for the analyses and the control run.

For ozone, the forecast improvements due to data assimilation were largely limited to the first 24 h of forecast. Also, the forecast initialised at 00:00 UTC from the analysis shows a larger negative bias for the daytime than the free model run. This is a result of the corresponding nighttime positive bias of the free model run. The bias is effectively removed in the 00 analysis; however, the subsequent forecast is unable to recover the level observed during daytime. The correlation coefficient during daytime is nevertheless improved slightly (from 0.75 to 0.78) by initialising from the analysis. While the forecast shows somewhat reduced positive bias for the hours between 18 and 30 (i.e. the following night), the subsequent daytime scores are already almost unchanged by assimilation. The results in Fig. 7 are computed for the MACC station network; a similar impact is observed at the EMEP stations.

Due to the shorter chemical lifetime, the effect of initial condition on forecasts of NO$_2$ can be expected to fall away more quickly than for ozone. This has been confirmed in the previous works based on the assimilation of data from the ozone monitoring instrument (OMI). Under summer conditions, Wang et al. (2011) found assimilation to provide no improvement in RMSE with regard to surface observations, while Silver et al. (2013) reported the NO$_2$ concentration to relax to its background values within 3–4 h.

In the forecast experiments performed within this study, the effect of assimilation on NO$_2$ forecast scores was limited to the first 6 forecast hours, which coincides with the nighttime in most of the domain. Hence, at least under the photochemically active summertime conditions, the analyses are only marginally useful for improving forecasts of NO$_2$.

The forecast for short-lived pollutants like NO$_2$ is poorly constrained by the initial condition, because the boundary layer concentrations become driven mainly by local emissions, chemical transformations and deposition. This limits the effectiveness of any assimilation scheme based updating only the initial condition. A possible way to extend the forecast impact is to include more persistent parameters, such as emission rates, into the state vector. This has been demonstrated by Elbern et al. (2007) for forecasting an ozone episode. In general, such an approach requires that the obtained posteriori emission rates can be extrapolated to the forecast window, and that the assimilation scheme is able to correctly attribute the observed discrepancies to the uncertain parameters.

5 Conclusions

An assimilation system coupled to the SILAM chemistry transport model has been described along with its application in reanalysis of ozone and NO$_2$ concentrations for year 2012.
Furthermore, the impact of using the O$_3$ and NO$_2$ analyses to initialise forecasts has been assessed for an ozone episode occurring in July 2012. The assimilation consistently improves the model-measurement comparison for stations not included in the assimilation. For daily maximum values, the correlation coefficient is improved over the free running model from 0.8 to 0.9 for O$_3$ and from 0.53 to 0.63 for NO$_2$ on rural validation stations. The respective biases are also decreased, however, a bias of $-7.4 \mu g m^{-3}$ remains in the O$_3$ analyses due to a negative bias in the free-running model.

During a 3-week forecast experiment, initialising the forecasts from the analysis fields provided an improvement in ozone forecast skill for a maximum of 24 h. For NO$_2$, the improvement was limited to a window of 6 h. The findings for NO$_2$ are similar to the results published in previous studies (Silver et al., 2013; Wang et al., 2011).

The diagnosed observation error standard deviations for ozone have a strong diurnal variation, and reach up to about 21 $\mu g m^{-3}$ at nighttime. These values are higher than usually assumed in chemical data assimilation, but agree well with the results obtained by Gaubert et al. (2014) with similar diagnostics.

The 3D-Var based assimilation has a low computational overhead. This makes it especially suitable for reanalyses in yearly or longer time scales, as well as for high-resolution forecasting under operational time constraints. Future studies will include more accurate characterisation of station representativeness, as well as further investigation of model biases for O$_3$.

**Code availability**

The source code for SILAM v5.3, including the data assimilation component, is available on request from the authors (julius.vira@fmi.fi, mikhail.sofiev@fmi.fi).

**Acknowledgements.** This work has been supported by the FP7 projects MACC and MACC-II and the NordForsk project EmblIA. The authors thank Marje Prank for constructive comments on the manuscript.

Edited by: A. Colette
References


J. Vira and M. Sofiev: Assimilation of surface NO2 and O3 observations


Vira, J. and Sofiev, M.: On variational data assimilation for estimating the model initial conditions and emission fluxes for short-


Variational assimilation of IASI SO2 plume height and total-column retrievals in the 2010 eruption of Eyjafjallajökull using the SILAM v5.3 chemistry transport model

Julius Vira1, Elisa Carboni2, Roy G. Grainger3, Mikhail Sofiev1

1 Finnish Meteorological Institute, Erik Palménin aukio 1, FI-00560 Helsinki, Finland
2 COMET, Atmospheric, Oceanic and Planetary Physics, University of Oxford, Parks Road, Oxford, OX1 3PU, U.K.
3 National Centre for Earth Observation, Atmospheric, Oceanic and Planetary Physics, University of Oxford, Parks Road, Oxford OX1 3PU, U.K.
Correspondence to: J. Vira, julius.vira@fmi.fi

Abstract

This study focuses on two new aspects on inverse modelling of volcanic emissions. First, we derive an observation operator for satellite retrievals of plume height, and second, we solve the inverse problem using the 4D-Var method. The approach is demonstrated by assimilating IASI SO$_2$ plume height and total column retrievals in a source term inversion for the 2010 eruption of Eyjafjallajökull. The inversion resulted in temporal and vertical reconstruction of the SO$_2$ emissions during the 1-20 May, 2010 with formal vertical and temporal resolutions of 500 m and 12 hours.

The plume height observation operator is based on simultaneous assimilation of the plume height and total column retrievals. The plume height is taken to represent the vertical centre of mass, which is transformed into the first moment of mass. This makes the observation operator linear and simple to implement. The necessary modifications to the observation error covariance matrix are derived.

Regularisation by truncated iteration is investigated as a simple and efficient regularisation method for the 4D-Var based inversion. In an experiment with synthetic observations, the truncated iteration was found to perform similarly to the commonly used Tikhonov regularisation. However, the truncated iteration allows the amount of regularisation to be varied a posteriori, without repeating the inversion. For inverting the Eyjafjallajökull SO$_2$ emission at the temporal and vertical resolution used in this study, the 4D-Var method required about 70% less computational effort than commonly used methods based on performing a separate model simulation for each degree of freedom in the estimated source term.

Compared to the inversion using only total column retrievals, assimilating the plume height resulted in a vertical emission profile more closely matching the ash plume heights observed by radar. The a posteriori source term gave an estimate of 0.29 Tg erupted SO$_2$ of which 95% was injected below 11 km.
1 Introduction

Sulphur dioxide (SO$_2$) is one of the major gas-phase species released in volcanic eruptions. Large SO$_2$ releases pose a hazard to aviation, decrease air quality, and as precursors to sulphate aerosols, have a potential impact on the Earth’s radiative balance (Bernard and Rose, 1990; Robock, 2000; Schmidt et al., 2015). Volcanic SO$_2$ plumes can be detected by satellite instruments measuring in either UV or IR wavelengths - however, reliably forecasting the atmospheric transport of volcanic plumes is hindered by the lack of in-situ measurements to characterise the emission fluxes of volcanic species (Carn et al., 2009; Stohl et al., 2011; Zehner, 2012).

While methods based purely on satellite retrievals (Theys et al., 2013 and references therein) exist for inferring the total SO$_2$ flux for a given eruption, a successful prediction of volcanic tracers generally requires information also on the vertical profile of emissions. An important technique for assessing both vertical and temporal distribution of the emission fluxes is provided by inverse dispersion modelling, first demonstrated for volcanic emissions by Eckhardt et al. (2008).

The previous studies on inverse modelling of volcanic emissions have been based on using total column retrievals of SO$_2$ or volcanic ash together with a Lagrangian (Kristiansen et al., 2010; Stohl et al., 2011) or Eulerian (Boichu and Clarisse, 2014; Boichu et al., 2013) dispersion models. In addition, Flemming and Inness (2013) devised a trajectory based scheme to evaluate the vertical emission profile, which was used together with assimilation of SO$_2$ retrievals with the IFS weather prediction system. The previous studies have demonstrated that the vertical distribution of emissions can be inferred from total column retrievals in presence of sufficient vertical wind shear. However, in the case of the Eyjafjallajökull eruption in 2010, both Boichu et al. (2013) and Flemming and Inness (2013) pointed out a lack of wind shear and a subsequent difficulty at estimating the vertical distribution of emissions.

Retrievals of SO$_2$ plume height have been performed with various satellite instruments (Carboni et al., 2012; Rix et al., 2012). Nevertheless, only a few studies have incorporated these data into models. Wang et al. (2013) derived a three-dimensional SO$_2$ distribution from retrievals by the Ozone Monitoring Instrument (OMI), and used the distribution to initialize CTM simulations for the 2008 eruption of Kasatochi. Wilkins et al. (2015) used 1D-Var ash retrievals for initialising dispersion simulations. However, neither of the studies used plume height retrievals in inverse modelling of volcanic emissions.

The first objective of the present paper is to assess the usefulness of assimilating SO$_2$ plume height retrievals from the Infrared Atmospheric Sounding Interferometer (IASI) instrument in a source term inversion. In Section 3.2 we develop an observation operator for the vertical centre of mass. Since the observation operator only depends on the centre of mass and column loading, the vertical profile is only partly constrained. However, in contrast to the previous studies, this approach makes no further assumptions about the shape or thickness of the SO$_2$ layer. This could be advantageous, since volcanic ash or SO$_2$ layers vary considerably in depth (Dacre et al., 2014) and can be emitted in multiple, overlapping layers (Kristiansen et al., 2010). In addition, our approach makes full use of the retrieval error estimates provided with the IASI data for both column mass and plume height, including the estimated correlation between plume height and mass errors.
The second objective of this paper is to explore the connection between the source term inversion and the 4D-Var data assimilation widely used in numerical weather prediction. Elbern et al. (2000) showed that the 4D-Var assimilation method (Le Dimet and Talagrand, 1986) can be easily extended into estimating emission fluxes with a chemistry transport model. Under the assumption of a linear dispersion model and observations, the 4D-Var formulation results in a least squares problem similar to that solved by many existing inversion algorithms. However, the iterative solution employed in 4D-Var favours a different regularisation approach, which is discussed in Section 4.

Finally, we test the variational inversion method and assimilation of plume height retrievals for estimating temporal and vertical distribution of SO\textsubscript{2} emission during the 2010 eruption of Eyjafjallajökull. Results of the inversion, presented in Section 5, indicate that assimilation of plume height retrievals results in more vertically concentrated emission profile. In particular, emissions above 8-10 km are reduced substantially, which is consistent with radar-based estimates of the eruption column height.

2 Model setup and observational data

2.1 Dispersion model

The transport and removal of SO\textsubscript{2} was evaluated using the dispersion model SILAM (System for integrated modelling of atmospheric composition; Sofiev et al., 2015, http://silam.fmi.fi) version 5.3. The model includes chemical removal of SO\textsubscript{2} as described by Sofiev (2000) with the OH climatology of Spivakovsky et al. (2000). The computations were driven by the ERA-Interim meteorological reanalysis (Dee et al., 2011) except for evaluating the simulated satellite retrievals described in Section 4, where operational ECMWF forecasts were used.

SILAM includes a variational data assimilation module, which was previously used for assimilation of air quality monitoring data of SO\textsubscript{2} by Vira and Sofiev (2012). This study uses the same assimilation system, but instead of estimating a refinement for a regional emission inventory, we seek to reconstruct the emissions for a single volcanic eruption as a function of time and injection height.

The model was configured for a domain covering 50°E to 30°W and 30°N to 80°N. Horizontal resolution of 0.5° was used for the inversion, while the a posteriori simulations were run with a higher 0.25° resolution. The vertical discretisation consists of 34 terrain-following z-levels with a 500 m resolution at the top of the domain increasing to 50 m near the surface.

2.2 Observations: the IASI dataset

IASI is an infrared Fourier transform interferometer that measures in the spectral range 645–2760 cm\textsuperscript{-1} with spectral sampling of 0.25 cm\textsuperscript{-1} (apodized spectral resolution of 0.5 cm-1) and has global coverage every 12h. The lev1b dataset from EUMETSAT/CEDA archive is used in this study.
The algorithm and the dataset are explained in more detail by Carboni et al. (2012). The same algorithm has been applied to other volcanic eruptions and successfully compared with other datasets (Carboni et al., 2016; Fromm et al., 2014; Koukouli et al., 2014; Schmidt et al., 2015; Spinetti et al., 2014).

The main points of the retrieval scheme are:

- Retrieval are performed for the pixels that were identified by the SO2 detection scheme (Walker et al 2011, 2012).
- All the channels between 1100-1200 and 1300-1410 cm\(^{-1}\) are used in the iterative optimal estimation retrieval scheme to obtain SO2 column amount and altitude of the plume (in pressure, under the assumption that the vertical concentration of SO2 follows a Gaussian distribution) together with the surface temperature. The scheme determines the column amount and effective altitude of the SO2 plume with high precision (up to 0.3 DU error in SO2 amount if the plume is near the tropopause), and it is well suited for plumes in low troposphere.
- The IASI SO2 retrieval is not affected by underlying cloud. If the SO2 is within or below an ash or cloud layer its signal will be masked and the retrieval will underestimate the SO2 amount. In the case of ash this is discernible a posteriori by the value of the cost function. The altitude retrieved for the Eyjafjallajökull eruption plume (using same dataset as in this paper) in presence of underlying cloud is consistent with the CALIPSO vertical backscatter profile (Carboni et al 2016, Figs. 1,2,3).
- A comprehensive error budget for every pixel is included in the retrieval. This is derived from an error covariance matrix \(S\) that is based on the SO2-free climatology of the differences between the IASI and forward modelled spectra.

Note that the error covariance, \(S\), is defined to represent the effects of atmospheric variability not represented in the forward model, as well as instrument noise. This includes the effects of cloud and trace-gases which are not explicitly modelled. The matrix is constructed from differences between forward model calculations (for clear-sky) and actual IASI observations for wide range of conditions, when we are confident that negligible amounts of SO2 are present. It follows that a rigorous error propagation, including the incorporation of forward model and forward model parameter error, is built into the system, providing quality control and error estimates on the retrieved state. The retrieval state error covariance matrix, used for the assimilation in this work, is directly provided as output of the retrieval pixel by pixel.

### 2.3 Inversion experiments

The IASI data were assimilated in inversion experiments for the Eyjafjallajökull (2010) eruption. The eruption has been described in detail by Gudmundsson et al. (2012). The experiments covered the time between 1 and 21 May, 2010, which as shown by Flemming and Inness (2013) included the most significant SO2 releases.

The emission flux density (kg m\(^{-1}\) s\(^{-1}\)) was estimated for each model level in steps of 12 hours. The inversions were made with three configurations: with assimilation of both column mass and centre of mass, with assimilation of column mass only, and with assimilation of both column mass and centre of mass but with a simplified formulation for the observation error covariance matrix.

Additionally, Section 4 describes a set of inversion experiments with simulated observations. These experiments were performed with similar configuration as the main experiments, but with a lower vertical resolution of 1 km.
3 Assimilation and inversion methods

The forward problem for volcanic tracer transport is defined by the advection-diffusion equation: given the emission forcing $f$, solve

$$\frac{\partial c}{\partial t} + \nabla \cdot (c \mathbf{V}) - \nabla \cdot (K \nabla c) = f(x,t) - s(c,x,t),$$

where $c$ is the tracer concentration, $\mathbf{V}$ is the wind vector, $K$ is the turbulent diffusivity tensor, and $s(c,x,t)$ denotes the chemical and other sinks.

3.1 Variational source term inversion

The inverse problem discussed in this paper is to determine the forcing $f$, given a set of observations depending on $c$. We assume that Eq. (1) has been discretised, and following the common notation in data assimilation literature, we denote the tracer concentrations, collectively for all time steps, with the state vector $x$. The state vector is related to the unknown parameter vector $f$ by the model operator $\mathcal{M}$. Finally, the vector $y$ of observations is given by the possibly non-linear observation operator $\mathcal{H}$ as $y = \mathcal{H}(x) + \varepsilon$, where $\varepsilon$ denotes the observation error.

If the observation errors follow a multivariate normal distribution with covariance matrix $R$, then a solution to the inverse dispersion problem can be sought by maximising the likelihood function, which is equivalent to minimising the cost function

$$J(f) = \frac{1}{2}(y - \mathcal{H}(x))^T R^{-1}(y - \mathcal{H}(x)),$$

where $x = \mathcal{M}(f)$. Model errors are not explicitly included in the cost function.

If the model and observation operators are linear, represented by matrices $M$ and $H$, then (2) becomes a linear least-squares problem. For volcanic eruptions with a known location, the emission vector $f$ is zero almost everywhere, which makes it feasible to evaluate the matrix $HM$ and solve (2) algebraically. This is the basis for inversion methods of Boichu et al. (2013), Eckhardt et al. (2008) and Lu et al. (2016).

As an alternative to the algebraic solution, the minimisation problem (2) can be solved with gradient-based, iterative algorithms, which avoids evaluating the matrix $HM$. In this study, the cost function is minimized using the L-BFGS-B algorithm of Byrd et al. (1995) which allows constraining the solution to non-negative values. Evaluating the gradient requires solving the adjoint problem for Eq. (1).

The iteration is continued until a stopping criterion is satisfied, e.g. until the norm of the gradient is reduced by a prescribed factor. However, in Section 4 we will discuss truncating the iteration before formal convergence in order to control the regularization.
3.2 Assimilation of plume height retrievals

Given the tracer concentration \( c(x, y, z) \) in three dimensions, the observation operator for column integrated mass is given by

\[
y = m_y = \sum_{i=1}^{N} w_i c(x_i, y_j, z_k)
\]

where \( x_i, y_j \) and \( z_k \) are the gridpoint coordinates and \( w_k \) denotes the thickness (in meters) of the \( k \)th model level. The layer concentrations are often weighted with an averaging kernel (Eskes and Boersma, 2003) to account for the vertically varying sensitivity of the satellite retrieval. In this work, weighting is not applied because the IASI retrievals treat the plume height explicitly.

In the retrievals, the plume height is represented by its centre of mass \( Z_{CM} \). It would be possible to develop an observation operator for \( Z_{CM} \), however, the operator would be nonlinear and only defined for nonzero columns. These problems can be overcome by replacing the centre of mass with the first moment of mass \( mZ_{CM} \). Then, the observations consist of pairs \((m_y, m_y Z_{CM})\) given by

\[
\begin{pmatrix}
m_y \\
m_y Z_{CM}
\end{pmatrix} = \begin{pmatrix}
\sum_{i=1}^{N} w_i c(x_i, y_j, z_k) \\
\sum_{i=1}^{N} z_k w_i c(x_i, y_j, z_k)
\end{pmatrix},
\]

where \( z_k \) is the height of the \( k \)th model level and \( i \) and \( j \) refer to the horizontal coordinates. Transforming the observations of \( Z_{CM} \) into \( mZ_{CM} \) changes the magnitudes of observation errors, and introduces a correlation between the observation components \( m \) and \( mZ_{CM} \). However, this effect can be evaluated and included into the observation operator.

Denote the means and standard deviations of \( m \) and \( Z_{CM} \) with \( \mu_m, \sigma_m \) and \( \mu_z, \sigma_z \). Assuming that the errors of \( m \) and \( Z_{CM} \) are normally distributed, it can be shown that the variance of first moment equals

\[
\Var[mZ_{CM}] = \mu_m^2 \sigma_z^2 + \mu_z^2 \sigma_m^2 + \sigma_m^2 \sigma_z^2 + 2 \mu_m \mu_z \Cov[m, Z_{CM}]
\]

Under similar assumptions, the covariance of \( m \) and \( mZ_{CM} \) becomes

\[
\Cov[m, mZ_{CM}] = \sigma_m^2 \mu_z + \mu_m \Cov[m, Z_{CM}].
\]

Finally, the expectation of \( mZ_{CM} \) is shifted due to the correlation between retrievals of \( m \) and \( Z_{CM} \):

\[
\E[mZ_{CM}] = \mu_m \mu_z + \Cov[m, Z_{CM}].
\]
The retrieval errors of different pixels are assumed to be uncorrelated. The observation error covariance matrix $\mathbf{R}$ is therefore block-diagonal, and its entries can be evaluated from the known covariances of $m$ and $Z_{CM}$ using Eqs. (5) and (6).

Assimilation schemes commonly assume uncorrelated and unbiased observation errors. A non-diagonal $\mathbf{R}$ can be introduced with a transformation of variables: define

$$\mathbf{L}^{-1} \mathbf{L} = \mathbf{R}^{-1}$$

$$\tilde{\mathbf{y}} = \mathbf{L}(\mathbf{y} - \mathbf{b})$$

$$\mathbf{H}^T = \mathbf{LH}$$

where $\mathbf{L}^{-1} \mathbf{L}$ is the Cholesky factorisation of the inverse observation error covariance matrix $\mathbf{R}^{-1}$ and $\mathbf{b} = (0, \text{Cov}[m,Z_{CM}])$ corrects for the bias according to Eq. (7). Then, substituting the transformations of Eq. (8) into the cost function (2) shows that assimilation of $\mathbf{y}$ with the original $\mathbf{R}$ is equivalent to assimilation of $\tilde{\mathbf{y}}$ using the transformed observation operator $\mathbf{H}$ with unit matrix in place of $\mathbf{R}$.

The above formulas can be implemented as a preprocessing step for the observations. In summary, the procedure is then as follows:

1. For each available IASI pixel $i$, evaluate the tuple $(m_i, m_i Z_{CM,i}, \text{Cov}[m_i, Z_{CM,i}])$ and the corresponding 2x2 covariance matrix $\mathbf{R}_i$.
2. Factorise $\mathbf{R}_i^{-1} = \mathbf{L}_i^{-1} \mathbf{L}_i$ and transform the observations according to Eq. (8).
3. Store the transformed observations $\tilde{\mathbf{y}}_i$ with their pixel-specific vertical weighting functions given by rows of the matrix $\mathbf{H} = \mathbf{L}_i \mathbf{H}$.

After the transformation, the observations are handled identically to regular column observations with a vertical weighting function.

### 3.3 Observation errors

The IASI retrievals used in this study include pixel-specific error estimates for total column and plume height retrievals. The estimates are derived statistically (Carboni et al., 2012) from differences between the transmission spectra computed by a forward model and those observed by IASI. Together with estimates for the correlation between plume height and total column retrieval errors, this provides the necessary input for equations (5) and (6).

The retrieval error estimates are only provided for pixels with positive SO$_2$ detection. For the non-SO$_2$ pixels, which are assimilated as zero values, a different estimate is used, based on the detection limits estimated by Walker et al. (2012). The detection limit was translated into a standard deviation of a Gaussian random variable assuming, conservatively, a probability of 0.95 for a correct detection.
However, performing the inversions with $R$ defined only by retrieval errors resulted in poor a posteriori agreement with the IASI data, which suggested that the retrieval errors are not sufficient to describe the discrepancy between the simulated and observed values. As found in the synthetic experiments, the impact of model uncertainty is significant compared to the retrieval errors, and it needs to be taken into account. The problem of model errors affecting the inversion is discussed by Boichu et al. (2013), who found the impact to depend strongly on treatment of zero-value observations, and consequently chose to keep only every tenth zero-valued observation.

In this study, the model errors are included by modifying the observation error covariance matrix, which is set to $R = R_{\text{obs}} + R_{\text{model}}$, where $R_{\text{model}}$ is diagonal and corresponds to experimentally determined constant standard deviations of 2 DU for total column and 1 km for the plume height.

The model errors for plume height and total column are assumed uncorrelated and independent of the observation errors. However, their effect is propagated to the covariance matrix for first moment according to Eq. (5). The actual model errors are likely to be variable and correlated in space and between the plume height and total column components; however, a more advanced treatment appears difficult in the current inversion approach.

### 3.4 Regularization

The least squares problem (2) has a unique solution only if the matrix $HM$ is of full (numerical) rank. Furthermore, if $HM$ is close to singular, the problem remains ill-posed, which results in a noisy solution. Consequently, some form of regularisation has been employed in all previous works based on the least-squares approach.

A common option is the Tikhonov regularisation, which introduces a penalty term into the cost function (2), which in the simplest form becomes

$$J(f) = \frac{1}{2}(y - Hx)^T R^{-1}(y - Hx) + \alpha^2 \sum_{k,t} w_k \int f(z,t)^2 \, dt \, dz,$$

where the summation is over levels $k$ and timesteps $n$. The weights $w_k$ in Eq. (9) are set equal to the thickness of each model layer; this makes the penalty term consistent with its continuous counterpart $\int f(z,t)^2 \, dt \, dz$, which in turn ensures that the regularisation term does not depend on the vertical discretisation.

The penalty term can be modified to include a non-zero a priori source term. However, this approach is not taken in the present work. Instead, we aim to choose the level of regularisation optimally, so as to avoid excessive bias in the regularised solution. The need for regularisation depends on the coverage of observations, accuracy of the forward model as well as on the meteorological conditions controlling the dispersion. Thus, the regularisation parameter $\alpha^2$ cannot be chosen a priori.

In this work, a criterion known as the L-curve (Hansen, 1992) is used for determining the amount of regularisation. In the L-curve approach, the inversion is performed with various values of $\alpha^2$, and the residual $\|y - Hx\|$ is plotted as a function of the solution norm $\|f\|$. For ill-posed inverse problems, the curve is typically L-shaped. The residual initially reduces quickly...
as the regularization is relaxed, however, for some value of $\alpha^2$, the curve flattens and reducing the regularization further only marginally improves the fit. This point, where L-curve reaches its maximum curvature, is taken to represent the optimal regularization. In the present study, the L-curve is evaluated without the frequently used logarithmic transformation.

The main advantage of the L-curve method is that it does not rely on a priori estimates for the observation error. This is useful, since in practice the discrepancy between simulated observations and the data is also affected by model errors, which are poorly known. The L-curve was, in effect, used in inverse modelling of volcanic SO2 also by Boichu et al. (2013).

Changing the regularization parameter requires the minimisation to be started over, which is costly in the variational inversion scheme where each iteration requires a model integration. However, as noted by Fleming (1990) and Santos (1996), the iteration itself forms a sequence of solutions with decreasing regularisation. Thus, instead of minimising the regularised cost function (9), we iterate to minimise the original cost function (2), and truncate the iteration according to the L-curve criterion. In the following section, we show experimentally that such an approach results in similar solutions as the more common Tikhonov regularisation.

### 4 Experiments with synthetic data

Regularisation by truncated iteration has been studied in detail especially for Krylov subspace based algorithms (Calvetti et al., 2002; Fleming, 1990; Kilmer and O’Leary, 2001). The effect of truncated iteration on quasi-Newton minimisation methods, such as the L-BFGS-B algorithm used in this work, has been studied less extensively. To evaluate the truncated iteration in comparison to Tikhonov regularisation for inverse modelling of volcanic emissions, we performed an experiment with synthetic data generated for the points observed by IASI during the simulated period of 1-21 May, 2010. In addition to the comparison of regularisation methods, the synthetic experiments enable us to evaluate robustness of the L-curve method and to assess how model errors affect the source term estimate.

The inversions were performed for a set of artificial source profiles (cases A to D) shown in the leftmost column of Figure 1. The cases A and B are defined arbitrarily, while cases C and D are realisations of a stochastic process where the total flux (kg/s) is given by a lognormal, temporally correlated random variable and the eruption height follows the relation of Mastin et al. (2009). At each time, a piecewise constant vertical profile is assumed with a transition at 75% of height. The emission rate is distributed evenly between the two sections.

For the sake of computational convenience, the experiments in this section were carried out by pre-evaluating the forward sensitivity matrix $H_M$ by running a separate model simulation for each component of the emission vector $f$. In order to simulate the effect of model errors, the matrix $H_M$ was evaluated with both the ERA-Interim and operational ECMWF forecast fields as meteorological drivers.

The sensitivity matrix for inversions was extracted from the run with ERA-Interim meteorological data. The set of synthetic observations of the SO2 column density, on the other hand, was evaluated from the model run based on the operational meteorological fields and used as the data for the inversion experiments. The simulated observations were
perturbed with Gaussian noise with standard deviation equal to 1 DU + 30% of the true value. The observation error covariance matrix used in the inversion was supplemented with 2 DU “model error” as described in Section 3.3.

The residual and solution norms, which define the L-curves, are evaluated consistently to the penalized cost function (9):

\[ \| \mathbf{Hx} - \mathbf{y} \| = \sqrt{\mathbf{y}^T \mathbf{R}^{-1} (\mathbf{Hx} - \mathbf{y})} \]
\[ \| \mathbf{f} \| = \sqrt{\sum_{i=1}^{n_{\text{w}}} w_i f_{i,\sigma}^2} \]

where \( \mathbf{f} \) denotes the emission, \( \mathbf{x} = \mathbf{Mf} \) and \( w_i \) is the thickness of the \( k \)th model layer. To evaluate the L-curve for Tikhonov-regularisation, the parameter \( \alpha^2 \) was incremented in discrete steps given by \( \alpha^2 = 10^i \cdot 2^{-i} \) for \( i = 0, 1, 2, ... \). The L-BFGS-B minimization method with non-negativity constraint was used for both Tikhonov regularisation and the truncated iteration; in the case of Tikhonov regularisation, the iteration was continued until convergence for each \( \alpha^2 \). With truncated iteration, the weights \( w_i \), required by Eqs. (9) and (10), are not explicitly included in the cost function. Instead, the same effect is achieved by transforming the parameter vector as \( f_{i,\sigma} = w_i^{1/2} f_{i,\sigma} \).

The point where the L-curve flattens, which is taken as the final solution, was determined numerically. First, the points \( (f, \| \mathbf{Hx} - \mathbf{y} \|) \) are sorted according to increasing \( \| \mathbf{f} \| \). Then, the points where the residual increases are removed, and finally, the optimal point is chosen using the “triangle” algorithm of Castellanos et al. (2002).

The inversion results using truncated iteration and Tikhonov regularisation are presented in the middle and left columns of Figure 1. In each test case, the emission timing is well captured within the 12 h resolution. The overall vertical profiles are also recovered, however, spurious features are present especially in cases B and C. The total emitted mass is underestimated by \(< 10\%\) for the solution from truncated and by up to about 15\% for the Tikhonov-regularised solution. The underestimation is expected due to the form of cost function (9). However, the inversion results show that the negative bias is not necessarily large unless the problem is regularised too strongly.

For comparison, Figure 2 presents the solution corresponding to the case B in Figure 1 but evaluated without model errors - that is, using the same sensitivity matrix \( \mathbf{HM} \) for both evaluating the observations and performing the inversion. In this case, regularisation was not needed, and the true solution was recovered almost perfectly despite the noisy observations. Thus, the noise present in the estimated solutions in Figure 1 is mainly due to model error, which affects the elements of matrix \( \mathbf{M} \).

The L-curves corresponding to each case in Figure 1 are shown in Figs. 3 and 4. The root mean squared error (RMSE) of the solution is shown next to each L-curve as a function of the regularisation parameter. When measured by the solution RMSE, an optimal regularisation indeed existed in each case. In case A, where the solution varies smoothly in time and space, the solution error is only moderately sensitive to the regularisation. The L-curve formed by the L-BFGS-B iterates is shallow in this case, which caused the algorithm to choose an unnecessarily high number of iterations. However, the
negative effect on the solution quality was small. For the Tikhonov regularisation, the regularisation parameter was
determined almost optimally.

The choice of regularisation was more critical in the remaining cases. In cases B and C, the L-curve has a clear plateau
after initial decrease, and the chosen corner point is close to optimal for the both regularisation methods. In case D, the
truncated iteration leads to a somewhat under-regularised solution similar to case A.

Outcome of the four experiments indicates that the need for regularisation varies strongly depending on the true source,
whose characteristics also affect how accurately the algorithm determines the optimal regularisation. We used the stochastic
source terms to evaluate this more quantitatively. Figure 5 presents the RMSE as a function of the iteration number for 40
realisations of the stochastic source term used in cases C and D. The optimal iteration numbers chosen from each L-curve are
marked with stars.

The RMS errors shown in Figure 5 are normalised by the minimum error for each inversion, which shows that in most
cases, the inversion was only moderately sensitive to the exact point of truncation. In 34 cases out of 40, the RMSE of the
solution determined from the L-curve was within 20% from the optimally regularised solution. Of the remaining six cases,
two were over-regularised and four were under-regularised.

While the experiments in this section were performed by pre-evaluating the matrix $\mathbf{H} \mathbf{M}$, in 4D-Var, the multiplications
by $\mathbf{H} \mathbf{M}$ and its transpose are replaced by forward and adjoint model evaluations. Although the approaches are formally
equivalent, this change results in a slightly different sequence of iterations from which the L-curve is evaluated. To
investigate this difference, we performed the inversion using the real IASI data using both approaches. The two solutions are
shown in Figure 6. The total released mass differs by less than 1% between the solutions, and the emission patterns are
qualitatively similar. The differences for individual values, although larger, appear small compared to the inversion errors.

In summary, the experiment with synthetic data showed that the truncated iteration resulted in solutions similar to those
obtained with the more common Tikhonov regularisation. This makes the truncated iteration, in combination with the L-
curve, an attractive option for regularising the variational source term inversion. On the other hand, the overall need for
regularisation depended strongly on the assumed source term. No regularisation was needed in absence of model error,
which indicates that the need for regularisation is likely to also depend on quality of the forward model. This emphasizes the
need for a robust method to determine the appropriate regularisation according to the situation at hand.

5 Inversion results for Eyjafjallajökull

Optimising the source term following the regularisation strategy described in Section 3.4 results in satellite-derived
estimates on the temporal and vertical emission profiles, as well as on the total emitted amount. The solutions presented here
correspond to iterates chosen from the L-curve as described in Section 3.4. For assimilation of column mass only, the 9th
iterate was chosen; with column mass and plume height assimilation, the 13th iterate was chosen.
Figure 7 shows the temporal and vertical distribution of the SO\textsubscript{2} emission obtained both with and without assimilation of plume height. The plume height time series estimated from radar and camera observations (Petersen et al., 2012) are plotted on top of the emission distributions. Even if the visible plume does not necessarily coincide with the SO\textsubscript{2} plume, the plume height observations provide an indication of the eruption activity.

The strongest emission occurred during 6th May. However, the vertical distribution of the peak depends on whether the plume height is assimilated. While the maximum occurs at 5-6 km, if plume height is not assimilated, secondary maxima appear at 11 km, reaching 13 km on 9th May. If plume height retrievals are assimilated, the emission above about 8 km is strongly suppressed. Similarly, on 18th May, the isolated emissions at 10 and 15 km are essentially removed when the plume height is assimilated.

Figure 8 shows the vertical profile of emissions integrated over the whole period. The bulk of emissions are between 2 and 8 kilometres even if only column density is assimilated. Assimilating the plume height retrievals further decreases the fraction of emissions above 8 km. When the plume height is assimilated, about 85% of total emission is estimated below 8 km and about 95% below 11 km.

The total released mass of SO\textsubscript{2} is 0.33 Tg when plume height is not assimilated and 0.29 Tg when plume height is assimilated. Figure 9 depicts the emission flux as a function of time and shows that while the largest difference in emission rate is during the peaks of 6th May, the assimilation of plume heights tends to decrease the emission rate throughout the eruption.

The inversion results of Figure 7 can be compared with those in Figure 10, which are obtained by assimilating both total column and plume height but neglecting all off-diagonal observation error covariance matrix elements. The distribution of emissions differs strongly from both cases in Figure 7, and the vertical distribution remains as spread as with assimilation of total column only. The treatment of observation errors as described in Section 3.2 is therefore a necessary step for successful assimilation of the plume height retrievals.

The SO\textsubscript{2} column densities simulated a posteriori are shown for 5-10 May in Figs. 11 and 12, along with the corresponding IASI retrievals. The overall patterns are well reproduced, although the column density is underestimated for some parts of the plume, especially on 7th and 8th of May. Due to the smaller total emission, the column densities are slightly lower when plume height is assimilated, however, the difference is small.

The plume height, evaluated as centre of mass, for 6-9 May is shown in Figure 13. Compared to IASI, the simulation based only on assimilation of total columns tends to overestimate the plume height for all four days. When the plume height retrievals are assimilated, the overestimation is reduced consistently, although not entirely removed.

6 Discussion

No a priori assumptions regarding shape the emission profile were made in this study. If only total column retrievals are used in the inversion, the estimated source term includes isolated emissions reaching up to 15 km. With plume height
assimilation, the vertical distribution becomes more concentrated and also more consistent with the plume observed with the radar, which suggests that the vertical distribution SO$_2$ and ash emissions was mostly similar.

The centre of mass retrievals only partly constrain the vertical distribution, and hence some emission remains between 8 and 12 km, and the overestimation of plume height is reduced but not removed in the a posteriori simulations. However, given the about 1 km uncertainty in the IASI plume height retrievals and the 1 km assumed model uncertainty (Section 3.3) included into the observation errors, the inversion results for plume height are consistent with the assumptions of the inversion.

Previous studies based on Lidar observations (Ansmann et al., 2010), aircraft measurements (Schumann et al., 2011) or inverse modelling (Stohl et al., 2011) do not suggest significant injection above the 10 km altitude. However, these studies were focused on volcanic ash instead of SO$_2$, and as shown by Thomas and Prata (2011), ash and SO$_2$ were not always transported together. In contrast, the SO$_2$ plume height estimates derived from the GOME-2 satellite instrument by Rix et al. (2012) do indicate heights above 10 km and up to 13 km on 5th of May. However, the plume heights retrieved from IASI data are below 6 km for that day, which agrees with the modelled plume heights (not shown) even when only total column retrievals are included in the inversion.

Among the previous emissions estimates for Eyjafjallajökull, Flemming and Inness (2013) estimated a 0.25 Tg total SO$_2$ release using GOME-2 satellite retrievals, and 0.14 Tg using the OMI retrievals. Our estimates of 0.29-0.33 Tg are higher, especially compared OMI, but this is consistent with the higher total SO$_2$ burden estimated (Carboni et al., 2012) from the IASI data used in this study.

The experiments with synthetic data (Section 4) show that the need for regularisation, or in Bayesian terms, the need for a priori information, strongly depends on the situation. In addition, the need for regularisation was strongly affected by uncertainty of the forward model, and the efforts needed to handle zero-valued observations in this and other studies support this conclusion. The errors arising from the dispersion model are likely to be correlated in space, and therefore, introducing the corresponding non-diagonal elements in the error covariance matrix $\mathbf{R}$ could improve the inversion results.

The model errors resulted in noisy temporal and vertical emission profiles in the synthetic experiments and probably also in the real inversions. However, the estimates for total emission were fairly robust regardless of the assumed source term or perturbations to the forward model. Also, halving the vertical resolution of the reconstruction (compare Figs. 6 and 7) resulted in only minimal change in the total emission. Nevertheless, the estimates of the total emission could be affected by biases in the satellite retrievals, or by model errors not exposed by the change of meteorological driver.

While the regularisation used in this work is equivalent to a zero-valued a priori source, a more informative a priori source could be accommodated with a change of variables. Other forms of regularisation proposed for the volcanic source term inversion include second-order temporal smoothing (Boichu et al., 2013), which also could be handled by truncated iteration as discussed by Calvetti et al. (2002).

The variational inversion method is computationally efficient if high temporal or vertical resolution is desired for the reconstruction. In the current configuration, the reconstructed solution had formally 1360 degrees of freedom. Each iteration
consisting of one forward and one adjoint integration, the 25 iterations would require about 1000 days to be simulated. In
comparison, evaluating the matrix $H M$ directly would require 1360 model integrations, and if the sensitivity was evaluated
in windows of eg. 72 hours, almost 4000 simulated days would be required.

A drawback of the 4D-Var inversion method is that the a posteriori error covariance matrix for the source term is difficult
to evaluate. However, Monte Carlo techniques could be used to sample the a posteriori uncertainty.

### Conclusions

We have presented an observation operator for retrievals of the vertical centre of mass of a tracer plume. The operator is
based on transforming the centre of mass into first moment of mass using the retrieval of total column. The approach was
tested by performing a source term inversion using SO$_2$ retrievals from the IASI instrument during the Eyjafjallajökull
eruption in May 2010.

Assimilating the plume height retrievals reduced the vertical spread of the SO$_2$ injection. When the plume height is
assimilated, about 85% of total emission was below 8 km and about 95% was below 11 km. The injection profile obtained
by assimilating the plume height retrievals is more consistent with the radar and camera based observations of the ash plume.

The inverse problem was solved with the 4D-Var method embedded into the SILAM dispersion model. Truncated
iteration is proposed as an efficient regularisation method for the 4D-Var inversion. Using both real and synthetic data, the
4D-Var method was shown to produce a similar solution as the more common algebraic method, but at considerably lower
computational cost.

Experiments with both synthetic and real data suggest that the inversion is sensitive to errors in the forward model, and to
their assumed uncertainty. Methods more robust to model errors are a topic suitable for future research.

### Appendix: moments of products of correlated Gaussian random variables

Let $X$ and $Y$ be scalar random variables with means and variances $\mu_X$, $\mu_Y$, $\sigma_X^2$ and $\sigma_Y^2$. Then, it follows from the
definitions for variance and covariance that

$$\text{Var}[XY] = \sigma_X^2 \sigma_Y^2 + \mu_X^2 \sigma_Y^2 + \mu_X \sigma_X^2 \mu_Y - 2 \mu_X \mu_Y \text{Cov}[X,Y] - \text{Cov}[X,Y]^2 + \text{Cov}[X^2,Y^2]$$

and

$$\text{Cov}[X,XY] = \text{E}[X^2]\text{E}[Y] + \text{Cov}[X^2,Y] - \text{E}[X]\text{E}[XY].$$

To expand $\text{Cov}[X^2,Y^2]$ and $\text{Cov}[X^2,Y]$ we assume that $X$ and $Y$ are normally distributed. We first define normalized
auxiliary variables

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X}, \tilde{Y} = \frac{Y - \mu_Y}{\sigma_Y}.$$
Then, by expressing $\hat{Y}$ as
\begin{equation}
\hat{Y} = c\hat{X} + \sqrt{1-c^2} \hat{Z},
\end{equation}
where $c = \text{Cov}[\hat{X}, \hat{Y}]$ and $\hat{Z} \sim \mathcal{N}(0,1)$ independent of $\hat{X}$, it is simple to verify that
\begin{equation}
\text{Cov}[\hat{X}^2, \hat{Y}] = 2c^2
\end{equation}
\begin{equation}
\text{Cov}[\hat{X}^2, \hat{Y}] = 0.
\end{equation}

For the original random variables $X$ and $Y$, we find by substituting (13) into the definition, expanding the terms, and using identities (15) that
\begin{equation}
\text{Cov}[X^2, Y^2] = 2\text{Cov}[X, Y]^2 + 4\mu_x \mu_y \text{Cov}[X, Y]
\end{equation}
and
\begin{equation}
\text{Cov}[X^2, Y] = 2\mu_x \text{Cov}[X, Y].
\end{equation}
Formulas (5) and (6) now follow by combining Eqs. (16) and (17) with (11) and (12).

Code availability

The source code for SILAM v5.3, including the data assimilation component, is available on request from the authors (julius.vira@fmi.fi, mikhail.sofiev@fmi.fi).

Acknowledgements

This work has been supported by the ESA projects SMASH and VAST and Nordforsk Centre of Excellence EmblA. E.C. and R.G.G. acknowledge funding from the NERC SHIVA (NE/J023310/1) and VANAHEIM (NE/1015592/1) projects. The work of E.C. has been partly funded by the EC-FP7 APhoRISM project. The authors thank Marje Prank for comments on the manuscript.

References


volcanic SO$_2$ flux at high temporal resolution using spaceborne plume imagery and chemistry-transport modelling: the 2010 Eyjafjallajökull eruption case study. Atmos. Chem. Phys. 13, 8569–8584. doi:10.5194/acp-13-8569-2013


Figure 1. Estimated emission flux (kg m\(^{-1}\) s\(^{-1}\)) in source term inversions with simulated data. True source terms for the four cases (A to D) are shown in the left column. Solutions using truncated iteration are shown in the middle column, solutions using Tikhonov regularisation are shown in the right column.
Figure 2. Estimated emission flux with synthetic data: inversion results for the case B in Figure 1 assuming a perfect forward model.
Figure 3. L-curve (left) and RMS error (right) for inversions with simulated data for cases A and B in Figure 1. The iterate (for truncated iteration) or regularisation parameter (for Tikhonov regularisation) chosen from the L-curve is marked with a star.
Figure 4. L-curve (left) and RMS error (right) for inversions with simulated data for cases C and D in Figure 1. The iterate (for truncated iteration) or regularisation parameter (for Tikhonov regularisation) chosen from the L-curve is marked with a star.
Figure 5. The normalised RMSE with respect to iteration number. Each grey line corresponds to an inversion with a randomly generated source term. The colourful stars denote solutions chosen from the L-curve.

Figure 6. Inversion results with real observations: emission flux (kg m\(^{-1}\) s\(^{-1}\)) obtained using 4D-Var (left) and by evaluating the sensitivity matrix (right).
Figure 7. Inversion results for Eyjafjallajökull. Left: emission flux (kg m$^{-1}$ s$^{-1}$) with assimilation of column mass only. Right: assimilation of column mass and plume height with full observation error covariance matrix. White dots denote plume height observations by radar, grey dots denote plume height observations with a camera.

Figure 8. Time-integrated emission of SO$_2$ (kg m$^{-1}$) during the simulated period as function of height (m) for the source term inversions with (green) and without (blue) plume height assimilation.
Figure 9. Estimated SO$_2$ emission flux (kg s$^{-1}$) as function of time with (green) and without (blue) assimilation of plume height retrievals.

Figure 10. Inversion results for Eyjafjallajökull: emission flux (kg m$^{-1}$ s$^{-1}$) with assimilation of column density and plume height but neglecting off-diagonal elements in the observation error covariance matrix. White dots denote plume height observations by radar, grey dots denote plume height observations with a camera.
Figure 11. SO₂ column loading (DU) for the a posteriori simulation with assimilation of plume height (top) and for the IASI column retrievals (bottom row). Results for 5, 6 and 7th May, 2010 are shown in the columns from left to right. The evening overpasses are shown for IASI, the model fields are valid at 22 UTC.
Figure 12. As Figure 11, but 8-10 May, 2010.
Figure 13. Simulated SO$_2$ plume height (centre of mass, km) without (left) and with (middle) assimilation of plume height retrievals for 6-9 (top to bottom row) May, 2010. The corresponding IASI retrievals are shown in the right column.
Construction of the SILAM Eulerian atmospheric dispersion model based on the advection algorithm of Michael Galperin

M. Sofiev1, J. Vira1, R. Kouznetsov1,2, M. Prank1, J. Soares1, and E. Genikhovich3

1Finnish Meteorological Institute, Helsinki, Finland
2A.M. Obukhov Institute for atmospheric Physics, Moscow, Russia
3Main Geophysical Observatory, St. Petersburg, Russia

Correspondence to: M. Sofiev (mikhail.sofiev@fmi.fi)

Received: 28 December 2014 – Published in Geosci. Model Dev. Discuss.: 16 March 2015
Revised: 8 October 2015 – Accepted: 18 October 2015 – Published: 3 November 2015

1 Introduction

One of the key problems in atmospheric composition modelling is the accuracy and reliability of numerical schemes. A less appreciated but important issue is the consistency of the approaches applied in different modules of the modelling system. Usually, model construction follows process-wise split (Yanenko, 1971; Marchuk, 1995; Seinfeld and Pandis, 2006), thus considering separately the advection and diffusion algorithms, physical and chemical transformations, and dry and wet deposition. In practical model developments, features of the transport algorithms, first of all, the advection scheme, largely shape the model and determine its area of application.

1.1 Advection schemes

There are numerous types of advection schemes currently employed in atmospheric dispersion models. Two major categories refer to Lagrangian or Eulerian treatment of tracers: as small-size masses (Lagrangian particles) or as the concentration fields discretised in a prescribed grid. The Eulerian schemes, the primary subject of this paper, can be divided into flux-form finite volume, semi-Lagrangian, or expansion-function schemes. The expansion-function schemes approximate the solution with a given set of basis functions and, in turn, can be divided into spectral, pseudospectral and finite-element approaches. Some classic schemes are also based on finite-difference approximations of the advective term of the transport equation. The basic principles of all these approaches were formulated several decades ago and, with cer-
tain modifications, are still in use. Many modern schemes combine several approaches.

The large diversity of the advection algorithms is explained by a long list of requirements for such schemes. The most important ones are positive definition, minimal numerical diffusion, limited non-monotonicity and non-linearity, stability with regard to high Courant numbers (the number of the model grid cells passed within one advection time step), small phase error, local and global mass conservation, and high numerical efficiency. Some of these requirements contradict each other. For example, numerical diffusion “blurs” the resulting patterns but also makes them smoother, thus improving the monotonicity.

The finite-difference schemes involve direct discretisation of the dispersion equation and various interpolation functions to evaluate derivatives of the concentration fields (see the reviews of Richtmyer, 1962; Leith, 1965; Roach, 1980, as well as Sect. 3.1 in Rood, 1987); specific examples are, for instance, Russell and Lerner (1981), and Van Leer (1974, 1977, 1979). These schemes, being once popular, usually suffer from large numerical diffusion and limited stability, which sets stringent limitations on the Courant number, usually requiring it to be substantially less than 1. Therefore, the interest has gradually shifted towards flux, finite-element, and semi-Lagrangian schemes.

The flux schemes represent the transport via mass fluxes at the grid cell borders, which are calculated from concentrations in the neighbouring cells and wind characteristics. They are inherently mass conservative and have become popular in atmospheric chemistry transport models (Kukkonen et al., 2012). Probably the most widely used flux-type scheme is the one made by A. Bott (Bott, 1989, 1992, 1993), especially if one would count the numerous Bott-type schemes (see examples in Syrakov, 1996; Syrakov and Galperin, 1997; Syrakov and Galperin, 2000; Walcek and Alekseev, 1998; Walcek, 2000; Yamartino, 1993), which are based on the same principle but involve different approximation functions, monotonicity and normalisation procedures, etc.

The semi-Lagrangian schemes have been among the most widely used approaches for decades, with numerous algorithms using its basic concept (Crowley, 1968; Egan and Mahoney, 1972; Pedersen and Prahm, 1974; Pepper and Long, 1978; Prather, 1986; Smolarkiewicz, 1982; Staniforth and Cote, 1991, and references therein; Lowe et al., 2003; Sofiev, 2000, etc.). In the forward form, these schemes consider the transport of mass starting from the grid mesh points (departure points) and calculate the masses at the grid points closest to the arrival point. Backward algorithms start from arrival grid points and find the grid points near the departure point. The schemes can be based on tracking either grid points or grid cells along their trajectories. The grid-point-based schemes are not inherently mass-conserving, whereas the volume-based schemes achieve mass conservation by integrating the mass over the departure volume. They can sometimes be described as a combination of finite-volume and semi-Lagrangian methods (Lin and Rood, 1996, 1997).

Stability of these schemes can be ensured for a wide range of Courant numbers (Leonard, 2002). A general review can be found in Lauritzen et al. (2011), whereas a comparison of 19 modern schemes is discussed in Lauritzen et al. (2014), hereinafter referred to as L14.

Modelling in spectral space is another approach with a long history (Ritchie, 1988; Kreiss and Oliger, 1972; Zlatov and Berkowicz, 1988; Prahm and Christensen, 1977), but is not widely used today.

One of the main problems of the existing schemes is substantial numerical diffusion originating from the finite-step discretisation along space and time. Seemingly inevitable in an Eulerian context, this phenomenon, however, does not exist in Lagrangian advection schemes, which do not contain explicit discretisation of particle movement. The Lagrangian domain is a continuous space rather than a set of pre-defined grid meshes, and the position of the particles can be calculated precisely. As a result, numerical diffusion of purely Lagrangian schemes is always zero – at a cost of strongly non-monotonous concentration fields due to limited spatial representativeness of a single Lagrangian particle and a limited number of particles.

One of the ways to reduce the diffusivity of an Eulerian scheme is to store additional prognostic variables to describe the state of each grid cell with higher spatial resolution than the formal cell size: a sub-grid mass distribution. This can take the form of extra conservation equation(s) for e.g. first- or higher-order moments (Egan and Mahoney, 1972; Prather, 1986). There are other approaches that use different kinds of extra information. For instance, the conservative semi-Lagrangian schemes (Yabe and Aoki, 1991; Yabe et al., 2001) use a semi-Lagrangian step to evaluate the mixing ratio at cell interfaces, and then use the interface values along with the cell integrals to derive an interpolant representing the sub-grid distribution.

In a series of works, Michael Galperin developed a semi-Lagrangian scheme that used the sub-grid information on mass centre location inside the cell. The scheme was made fully non-diffusive for isolated plumes, positively defined, and very efficient computationally (Galperin et al., 1994, 1995, 1997; Galperin, 1999; Galperin and Sofiev, 1998; Galperin and Sofiev, 1995; Galperin, 2000). The early version of this scheme applied in the large-scale modelling by Sofiev (2000) resembled the “moving-centre” approach widely used in aerosol dynamics models (Kokkola et al., 2008) and shared its characteristic weakness – high non-monotonicity. The later developments substantially reduced it without damaging other features (Galperin, 1999, 2000). Further development of this scheme is the subject of the current paper.
1.2 Horizontal and vertical diffusion, dry deposition

Diffusion algorithms are less diverse than advection schemes. The physical reason for one of the common diffusion parameterisations is described in detail by Smagorinsky (1963), who suggested a formula for grid-scale scalar eddy diffusivity based on the model resolution and wind speed derivatives, thus connecting the numerical features of the simulations and hydrodynamics. It is widely used in chemical transport models (Kukkonen et al., 2012).

The dry deposition is usually accounted for as a boundary condition for the vertical advection–diffusion equation. Computation of dry deposition for gases practically always follows the electrical analogy, for which one of the first comprehensive parameterisations was suggested by Wexely (1989). Among the extensions of this approach, one was suggested by Sofiev (2002), which combined it with vertical diffusion and connected it with the Galperin advection scheme. The algorithm used an effective mean diffusion coefficient over thick layers calculated from high-resolution meteorological vertical profiles, the direction also recommended by Venkatram and Pleim (1999). These thick layers were determined using the subgrid information available from the advection scheme, which increased the accuracy of both algorithms (Sofiev, 2002).

For aerosol species, the electrical analogy is not correct (Venkatram and Pleim 1999). Compromising approaches suggested by Slinn (1982) and Zhang et al. (2001) and updated by Petroff and Zhang (2010) involve numerous empirical relations, sometimes on thin ground. A more rigorous approach unifying the gas and aerosol deposition parameterisations into a single solution was developed by Kouznetsov and Sofiev (2012).

1.3 Organisation of the paper

The current paper describes the Eulerian transport algorithm of the System for Integrated modelling of Atmospheric composition SILAM v.5, which is based on the advection scheme of Michael Galperin with several updates.

The paper is organised as follows. Section 2 describes the original algorithm of Michael Galperin and positions the scheme among other approaches. Section 3 presents the improvements made during its implementation and testing in SILAM. Section 4 outlines the scheme interconnections with other model parts. Standard and advanced model tests are shown in Sect. 5. Finally, the discussion in Sect. 6 includes an analysis of the scheme performance in the tests, as well as comparison with other algorithms.

Following the SILAM standards, all units throughout the paper are the basic SI: (mole) for chemicals, (kg) for aerosols without chemical speciation, (m) for distance and size, (s) for time, etc. The model operates with concentrations (mol m⁻³) or (kg m⁻³). Some of the tests below are formulated in mixing ratios (mol mol⁻¹) or (kg kg⁻¹).

2 Background

2.1 Basic equations

We consider the forward dispersion equation with the first-order K-theory-based closure for diffusion:

\[ L\varphi = E, \quad \text{where} \quad L = \frac{\partial}{\partial t} + \frac{\partial}{\partial \xi_i} (u_i) \]

\[ \frac{\partial}{\partial \xi_i} \rho \mu_{ij} \frac{1}{\partial \xi_j} + \zeta. \]  \( (1) \)

Here \( \varphi \) is the concentration of the pollutant, \( t \) is time, \( E \) is the emission term, \( \xi_i, i = 1..3 \) denote the three spatial axes, \( u_i \) are components of the transport velocity vector along these axes, \( \mu_{ij} \) are components of the turbulent diffusivity tensor, \( \rho \) is air density, and \( \zeta \) represents the transformation source and sink processes.

Equation (1) is considered on the time interval \( t \in (t_0, t_\infty) \) in the domain \( \{\xi_i\} \in \Sigma = [h_1, H] \times \Omega \), where the heights \( h_1 \) and \( H \) are the lower and upper boundaries of the computational domain and \( \Omega \) is the horizontal computational area with border \( \partial \Omega \). The initial conditions are

\[ \varphi |_{t=t_0} = \varphi_0(\xi). \]  \( (2) \)

Boundary conditions depend on the type of the simulations. In a general form, they constrain concentration and/or its first derivative:

\[ \varphi |_{\xi_j \in \partial \Sigma} = \gamma. \]  \( (3) \)

Here the values of \( \alpha, \beta, \) and \( \gamma \) depend on the type of the boundary. In particular, dry deposition at the surface \( \xi_3 = h_1 \) is described via \( \alpha = \mu_{33}, \beta = -v_3 \) (dry deposition velocity), and \( \gamma = 0 \); prescribed concentration \( \varphi_l \) at the lateral boundaries \( \xi_{1,2} \in \partial \Sigma \) implies \( \alpha = 0, \beta = 1, \gamma = \varphi_l \), etc. A global domain would require periodic longitudinal conditions.

2.2 Advection scheme of Michael Galperin

The current section presents the advection algorithm suggested by Michael Galperin in the 2000s as a contribution to the Eulerian transport scheme of SILAM. The idea of the scheme can be found in a short methodological publication of Galperin (2000) (in Russian) and conference materials (Galperin, 1999; Sofiev et al., 2008). It is very briefly outlined by Petrova et al. (2008) (hereinafter referred to as P08), but no systematic description exists so far.

For the 1-D case, let us define the simulation grid, \( \xi_1 = x \), as a set of \( I \) grid cells \( i = 1..I \). Let the coordinate of the centre of the \( i \)-th grid cell be \( x_i \), and the coordinates of the cell left- and right-hand borders be \( x_{i-0.5} \) and \( x_{i+0.5} \), respectively. The 1-D cell size is then \( V_i = x_{i+0.5} - x_{i-0.5} \). The advected field \( \varphi \), in each grid cell \( i \), is defined via the total
mass in the cell $M_i$ and the position of the centre of mass $X_i$, $X_i \in [x_{i-0.5}, x_{i+0.5}]:$

$$M_i = \int_{x_{i-0.5}}^{x_{i+0.5}} \varphi(x)dx,$$

$$X_i = \frac{1}{M_i} \int_{x_{i-0.5}}^{x_{i+0.5}} x \varphi(x)dx.$$  

(4)

Let us represent the mass distribution in each grid cell via the rectangular slab:

$$
\pi_i^n(x) = \begin{cases} 
\frac{1}{\omega_i^n}, & |x - X_i^n| \leq \omega_i^n \\
0, & \text{otherwise}
\end{cases},
$$

(5)

where $n$ is the time step and $\omega_i^n = \min\left(\left|X^n_i - x_{i-0.5}\right|, \left|X^n_i - x_{i+0.5}\right|\right)$ is the distance from the centre of mass $X^n_i$ to the nearest border of the cell $i$. Equation (5) defines the widest unit-volume slab that can be confined inside the cell (Fig. 1) for the given centre of mass.

The advection scheme consists of a transport step and a re-projection step. At the transport step, each slab $\pi_i$ is moved along the velocity field $u(x)$. Advection of the slab does not change its shape within the time step $\delta t = t_{n+1} - t_n$, but can move it anywhere over the domain or bring it outside. In essence, the slab transport is replaced with advection of its mass centre, which during this time step becomes analogous to a Lagrangian particle:

$$X_i^{n+1} = X_i^n + \int_{t_n}^{t_{n+1}} u(X_i^n, t) dt,$$

(6)

where $u(X_i^n, t)$ is the wind speed at the mass centre location.

The original Galperin scheme employed wind at the cell mid-point $x_i$ and used explicit first-order time discretisation: $u(x_i^n, t_n) = u^n_i$. Then the transported slab is given by

$$\pi_i^{n+1}(x) = \pi_i^n(x - u^n_i \delta t).$$

(7)

Following the transport step (7), the masses $M_k$ and centres of mass $X_k$ of the receiving set of cells $k \in K$ are updated using the transported slabs $\pi_i^n$:

$$M_k^{n+1} = \sum_{i=1}^{N_k} \alpha_{i,k} M_i^n,$$

$$X_k^{n+1} = \frac{1}{M_k^{n+1}} \sum_{i=1}^{N_k} \beta_{i,k} M_i^{n+1},$$

(8)

where $\alpha_{i,k} = \int_{x_{i-0.5}}^{x_{i+0.5}} \pi_i^n(x)dx$ and $\beta_{i,k} = \int_{x_{i-0.5}}^{x_{i+0.5}} \tilde{\pi}_i^n(x)dx$ correspond to the mass and the first-moment fractions arriving from the cell $i$ into cell $k$. The integrals are easy to evaluate due to the simple form of $\pi^n_i(x)$ in Eq. (5). In essence, Eq. (8) describes a mass-conservative projection of the advected slab to the computation grid.

The coefficients $\alpha_{i,0} = \int_{-\infty}^{0.5} \pi_i^n(x)dx$ and $\alpha_{i,1+1} = \int_{0.5}^{\infty} \pi_i^n(x)dx$ determine the transport outside the domain through the left and right borders, respectively; that is, the scheme is fully accountable and mass-conservative since $\sum_k \alpha_{i,k} = \int_{-\infty}^{\infty} \pi_i^n(x)dx = 1$ for each $i$. Also, since the functions $\pi_i^n(x)$ are nonnegative, the coefficients $\alpha_{i,k}$ are nonnegative, and consequently $M_k^{n+1} \geq 0$ as long as $M_k^n \geq 0$ for all $i$. It means that the scheme is positively defined for any simulation set-up: $u$, $\Delta t$, and discretisation grid.

In multiple dimensions, the slabs are described by the total mass in multidimensional cells and centres of mass along each dimension. In two dimensions, an analogue of Eq. (5)

will be

$$\pi_{i,j}^n(x,y) = \pi_{i,j}^n(x)\pi_{i,j}^n(y),$$

(9)

where the functions $\pi_{i,j}(x)$ and $\pi_{i,j}(y)$ depend on $X_{i,j}$ and $Y_{i,j}$, respectively. The advection step in the form of Eq. (7) and the slab projection integrals Eq. (8) are then defined in 2-D space.

However, a simpler procedure used in the original scheme is obtained with dimensional splitting, where the transport in each dimension is processed sequentially with the grid projection applied in between. For an $x-y$ split, the intermediate distribution for each row $j$ is obtained by setting

$$\pi_{i,j}^{n+1/2}(x,y) = \pi_{i,j}^n(x)\pi_{i,j}^n(y),$$

(10)

evaluating $\alpha_{i,k}$ and $\beta_{i,k}$ from $\pi_{i,j}^n(x)$ and updating $M_{i,j}$, $X_{i,j}$ and $Y_{i,j}$. Since $\int_{y_{j-0.5}}^{y_{j+0.5}} \pi_{i,j}(y)dy = 1$ and $\int_{x_{i-0.5}}^{x_{i+0.5}} \pi_{i,j}(x)dx = Y_{i,j}$, the 2-D slab projection simplifies...
The advection is then performed by taking the transport step for $n_{i,j}^{n+1/2}(y)$ starting from $Y_i^{n+1/2}$, evaluating $\alpha_{i,k}$ and $\beta_{i,k}$ from $\tilde{n}_{i,j}^{n+1/2}(y)$, and applying the re-projection Eq. (11) with $X$ and $Y$ inverted. The generalisation to three dimensions is analogous.

### 2.3 Relation of the Galperin scheme to other approaches

The Galperin scheme shares some features with other approaches (see the reviews of Rood, 1987, and Lauritzen et al., 2011). Arguably the closest existing scheme is the finite-volume approach of Egan and Mahoney (1972), hereinafter referred to as EM72, and Prather (1986), hereinafter P86. The main similarity between these schemes is the representation of the mass distribution as a set of slabs (rectangular in EM72 and continuous polynomial distributions in P86), one per grid cell, with the mass centre identified via the slab first moment, plus additional constraints. During the EM72 and P86 advection steps, mass and the first moment are conserved, similarly to the re-projection step (8). However, the similarity ends here.

There are several principal differences between the EM72/P86 and Galperin algorithms.

Firstly, in EM72 the slab width is computed via the second moment (variance) of the mass distribution in the grid cell. P86 uses the second moments to constrain the shape of the polynomials. As a result, this moment has to be computed and stored for the whole grid, and the corresponding conservation equation has to be added, on top of those for the mass and the first moment. Galperin’s approach does not require the second moment, instead positioning the slab against the cell wall. In fact, EM72 pointed out that the second moment can be omitted, but did not use the wall-based constraint in such a “degenerated” scheme, which severely affected its accuracy.

Secondly, EM72 uses the movements of the slabs in adjacent grid cells to calculate the mass flows across the border. Such local consideration requires the Courant number to be less than 1: the so-called “portioning parameter” (a close analogy to the Courant number in the scheme) is limited between 0 and 1. The same limitation is valid for the P86 approach. Galperin’s scheme can be applied at any Courant number and its re-projection step can rather be related to Lin and Rood (1996).

### 3 Updates of the scheme in SILAM v.5

There are several features of the original scheme which make it difficult to use in large-scale chemical transport simulations. These are listed here and the corresponding improvements are introduced in the following sub-sections.

- The scheme is formulated with zero inflow boundary conditions.
- Real-wind tests have shown that the scheme has difficulties in complex-wind and complex-terrain conditions, similar to EM72 (Ghods et al. 2000).
- The explicit forward-in-time advection (Eq. 7) is inaccurate.
- The scheme, being very good with individual sharp plumes over zero background, noticeably distorts the smoother fields with a non-zero background – see P08.

In addition, the accuracy of the dimension split was increased via symmetrisation: the order of dimensions in SILAM routines is inverted each time step: $x-y-z-z-y-x$ (Marchuk, 1995).

### 3.1 Lateral and top boundary conditions

The open boundary for the outgoing masses is kept in SILAM regional simulations. The inflow into a limited-area domain is defined via prescribed concentration at the boundary (Eq. 3), $\alpha = 0$, $\beta = 1$, $\gamma = \psi_t$. The mass coming into the domain during a single time step is equal to

$$
M_{1}^{\text{in}} = \psi_t(x_{0.5}, u(x_{0.5})) N(u(x_{0.5})) \delta t, \\
M_{1}^{\text{in}} = \psi_t(x_{1.0}, u(x_{1.0})) |u(x_{1.0})| N(-u(x_{1.0})) \delta t. 
$$

Here $N(u)$ is a Heaviside function ($= 1$ if $u > 0$, = 0 if $u \leq 0$). Assuming the locally constant wind, we find that $M_{1}^{\text{in}}$ is distributed uniformly inside the slab, similar to that of Eq. (5). For e.g. the left-hand border, the continuous form will read

$$
\Pi_{\text{in}}^{n+1}(x) = \left\{ \begin{array}{ll}
\psi_t(x_{0.5}) N(u(x_{0.5}, t_n)) \delta t, & x \in [x_{0.5}, x_{0.5} + u(x_{0.5}, t_n) N(u(x_{0.5}, t_n)) \delta t] \\
0, & \text{otherwise}
\end{array} \right.
$$

It is then projected to the calculation grid following Eq. (8).

The top boundary follows the same rules as the lateral boundaries. At the surface, the vertical wind component is zero, which is equivalent to closure of the domain with regard to advection.

Global-domain calculations require certain care: SILAM operates in longitude–latitude grids; that is, it has singularities at the poles and a cut along the 180° E line. For longitude, if a position of a slab part appears to be west of $-180^\circ$ E or east of $180^\circ$ E, it is increased or decreased by

\[
M_{k,j}^{n+1/2} = \sum_{i=1}^{N_x} \alpha_{i,k} M_{i,j}^n, \\
X_{k,j}^{n+1/2} = \frac{1}{M_{k,j}^n} \sum_{i=1}^{N_x} \beta_{i,k} M_{i,j}^n, \\
Y_{k,j}^{n+1/2} = \frac{1}{M_{k,j}^n} \sum_{i=1}^{N_x} \alpha_{i,k} M_{i,j}^n Y_{i,j}^n.
\]
360°, respectively. Resolving the pole singularity is done by
reserving a cylindrical reservoir over each pole. The radius of
the reservoirs depends on the calculation grid resolution but
is kept close to 2°. The calculation grid reaches the borders
of the reservoirs, whose mean concentrations are used for the
lateral boundary conditions:

\[
\begin{align*}
\varphi|_{y_2=y_{2,0.5}} &= \varphi^*_{\text{pole}}(t, z); \\
\varphi|_{y_2=y_{2,1.0.5}} &= \varphi^*_{\text{pole}}(t, z).
\end{align*}
\]  

(14)

Vertical motion in the pole cylinders is calculated separa-
ately using the vertical wind component diagnosed from the
global non-divergence requirement.

### 3.2 Extension of the scheme for complex wind patterns

The original Galperin scheme tends to accumulate mass at
stagnation points where one of the wind components is small.
Similar problems were reported by Ghods et al. (2000) for
the EM72. Ghods et al. (2000) also suggested an explanation
Similar problems were reported by Ghods et al. (2000) for
stagnation points where one of the wind components is small.

The original Galperin scheme tends to accumulate mass at
stagnation points where one of the wind components is small.
Similar problems were reported by Ghods et al. (2000) for
stagnation points where one of the wind components is small.
Similar problems were reported by Ghods et al. (2000) for
stagnation points where one of the wind components is small.

The way this smoother works becomes clearer if one no-

\[
x_{i-0.5} \leq x \leq x_{i+0.5}.
\]  

(17)

Then, the trajectory equation (6) can be piece-wise inte-
 grated analytically for each slab edge. Let us denote \( \Delta u =
\frac{u_{i+0.5} - u_{i-0.5}}{\Delta t}, \quad \Delta t = t_n+1 - t_n, \quad \alpha = \Delta u/\Delta t \)
and consider the transport starting at e.g. \( x_{i-0.5} \). Then, the time needed for
passing through the entire cell, \( \Delta x = x_{i+0.5} - x_{i-0.5} \), is

\[
T_{\text{cell}} = \log(1 + \alpha \Delta x u_i)/\alpha.
\]  

(18)

If \( \Delta t < T_{\text{cell}} \), the point will not pass through the whole cell
and stop at

\[
x_{\Delta t} = x_{i-0.5} + u_i(\exp(\alpha t - 1))/\alpha.
\]  

(19)

Applying sequentially Eqs. (18) and (19) until completing
the model time step \( \Delta t \), one obtains the analytical solution
for the final position of the slab edges.

This approach neglects the change of wind with time.
However, the integration method is robust, since the linearly
interpolated wind field is Lipschitz-continuous everywhere,
which in turn guarantees the uniqueness of the trajectories
of \( X_L \) and \( X_R \). Therefore, using the analytic solution Eqs. (18)
and (19), the borders of the slabs will never cross.

### 3.4 Reducing the shape distortions

The original scheme tends to artificially sharpen the plume
edges and to gradually redistribute the background mass in
the vicinity of the plume towards it (Fig. 2, blue shapes).
Similar “antidiffusive” distortions were also reported by P08
and by EM72 – for their scheme.

The reason for the feature can be seen from Eq. (8): if a
large mass is concentrated at one of the grid cell sides, the
centre of mass becomes insensitive to the low-mass part of
the cell; that is, a small mass that appears there from the
neighbouring cell is just added to the big slab with little effect
on its position and width.

A cheap, albeit not rigorous, way to confront the effect is
to compensate for it via an additional small pull of the mass
centre towards the cell mid-point before forming the slab for
advective:

\[
\tilde{\chi}_i^n = x_i + (X_i^n - x_i)(1 - \varepsilon),
\]  

(20)

where \( \varepsilon \) is the smoothing factor. The adjusted mass centre \( \tilde{\chi}_i^n \)
is then used to form the slab Eq. (5).

The way this smoother works becomes clearer if one no-

\[
u(x) = u(x_{i-0.5}, t_n) \frac{(x_{i+0.5} - x)}{(x_{i+0.5} - x_{i-0.5})} + u(x_{i+0.5}, t_n) \frac{(x - x_{i-0.5})}{(x_{i+0.5} - x_{i-0.5})}
\]  

Geosci. Model Dev., 8, 3497–3522, 2015

www.geosci-model-dev.net/8/3497/2015/
Figure 2. Shape preservation tests: (a) step, (b) triangle peak, (c) sine-shaped dip, and (d) sine-shaped peak. Sequential positions are shown: “r” denotes the scheme without a smoother, “r_diff” with it. The legend includes the number of times steps made. Wind is from left to right; Courant = 0.4.

Figure 3 presents the amplification factor and RMSE for the Galperin scheme without the smoother (panels a, d) and with it, $\varepsilon = 0.08$ (panels c, f). Furthermore, the impact of doubling the background component to twice the wave amplitude is shown (panels b, e). In the case of $B = 1$, the scheme without the smoother shows only minor damping of all considered wave numbers ($k$ up to 25). The RMSE has a maximum for $k$ of between 5 and 10 but stays almost constant from $k = 10$ to $k = 25$. This shows the scheme’s ability to resolve sharp gradients when there is no significant background. The cumulative amplifying factors for some wavelengths exceed 1, but this does not imply instability, since the single-step amplifying factors fluctuate depending on the positions of the centres of mass. If the integration is continued over a large number of time steps (not shown), the solution converges to a combination of rectangular pulses (a similar feature was mentioned in EM72 for that scheme).

Introducing the smoothing $\varepsilon = 0.08$ resulted in strong attenuation of high-frequency components and increased the RMSE for wave numbers above $\sim 10$. Since the smoothing factor effectively damps the fluctuations of the centres of mass, the amplification factors are below 1 for all wave numbers. Adding a background term also reduces the responsiveness of the mass centres to newly coming amounts (see

3.5 Analysis in frequency space

The non-linearity introduced by the coupling of cell mass and centre of mass in Eq. (8) makes formal stability and convergence analysis after Charney et al. (1950) difficult. However, the features of the scheme can be investigated numerically following the approach of Kaas and Nielsen (2010).

The scheme was run for 200 time steps in a 1-D domain with 100 grid points. For each wave number up to 25, the scheme was initialised with the corresponding sine function and run with Courant numbers ranging from 0.05 to 0.95 in steps of 0.05. This allowed evaluation of the spectrally resolved root mean squared error (RMSE) and, after a Fourier transform, the spectral amplification factor (cumulative for the 200 steps) for each wave number. The amplification factor quantifies the scheme’s ability to resolve the corresponding harmonics, while the RMSE additionally includes the effect of phase errors and possible spurious modes. Since the scheme is formulated for nonnegative concentrations, a constant background $B = 1$ is added to each waveform.

$\varepsilon \sim 0.08$ (Fig. 2, red shapes). This behaviour and the value were similar for various Courant numbers and tests. It is also seen from the spectral features of the scheme in the next section – and further discussed in relation to scheme tests.
Eq. 8), which leads to a similar damping of the higher frequencies as in Fig. 3c, f.

To further investigate the spectral response of the scheme, it was evaluated with a broadband input:

$$f(x) = \sin(2\pi x \cos(20\pi x)) + B.$$  \hspace{1cm} (21)

Figure 4, right panel, depicts the power spectral densities for the exact and numerical solutions after a single revolution with $CF = 0.7$ and 100 grid points. The corresponding solutions are shown in the lefpanel. For the comparison, results are also shown for the fourth-order Bott (1989) scheme without shape preservation, and for a generic non-conservative upstream semi-Lagrangian scheme with cubic spline interpolation.

With $B = 1$, all schemes capture the first spectral peaks around $k = 10$ and therefore resolve most of the spectral content. Without a smoother, the Galperin scheme that follows the spectrum of the true solution also resolves the spectral features around $k = 30$. Application of the smoother leads to a damping effect throughout the spectrum, including the spurious high-frequency components, such as the peak at $k = 40$. This illustrates the use of the smoother for reducing over- and under-shoots, as discussed in Sect. 3.4.

Similarly to the single-harmonic tests, the situation changes in the presence of a significant background ($B = 2$). Regardless of smoothing, the Galperin scheme damps the spectral peaks starting around $k = 10$, which corresponds to the reduction of amplitude visible in the numerical solution.

4 Connection of the advection scheme with other SILAM modules

Construction of the dispersion model using the Galperin advection scheme as its transport core is not trivial because all other modules should support the use of the sub-grid information on positions of the mass centres. In some cases it is straightforward, but in others one can only make the module to return them undamaged back to advection.

4.1 Vertical axis: combined advection, diffusion, and dry deposition

For the vertical axis, SILAM combines the Galperin advection with the vertical diffusion algorithm following the extended resistance analogy (Sofiev, 2002), which considers the air column as a sequence of thick layers. Vertical slabs within these layers are controlled by the same 1-D advection, which is performed in absolute coordinates – either $z$- or $p$-, depending on the vertical (height above the surface or hybrid). Settling of particles is included in advection for all layers except for the first one, where the exchange with the surface is treated by the dry deposition scheme.

The centres of masses are used but not modified by diffusion: the effective diffusion coefficient between the neighbouring thick layers is taken as an inverse of aerodynamic resistance between the centres of mass of these layers (Sofiev, 2002):

$$< K_{i,i+1} > = \frac{\Delta z_{i,i+1}}{Z_{i+1} \int_{Z_i}^{Z_{i+1}} \frac{dz}{K(z)}}.$$  \hspace{1cm} (22)
The effective thickness $\Delta z_{i,i+1}$ is taken to be proportional to the pressure drop between the centres of masses, which ensures equilibration of mixing ratios due to diffusion.

In the lowest layer, the dry deposition velocity is calculated at the height of the centre of mass $Z_1$ following the approach of Kouznetsov and Sofiev (2012).

The advantages of using the mass centres as the vertical diffusion meshes are discussed in detail by Sofiev (2002), where it is shown that the effect can be substantial if an inversion layer appears inside the thick dispersion layer. Then the location of the mass centre above/below the inversion can change the up/down diffusion fluxes by a factor of several times.

4.2 Emission-to-dispersion interface

Emission data are the only source of sub-grid information apart from the advection itself: location of the sources is transformed into the mass centre positions of their emission.

For point sources, the mass is added to the corresponding grid cell and centres of masses are updated:

$$
\hat{M}_{ijk} = M_{ijk} + M_{ems} \\
\hat{X}_{ijk} = (X_{ijk}M_{ijk} + M_{ems}X_{ems})/\hat{M}_{ijk} \\
\hat{Y}_{ijk} = (Y_{ijk}M_{ijk} + M_{ems}Y_{ems})/\hat{M}_{ijk} \\
\hat{Z}_{ijk} = (Z_{ijk}M_{ijk} + M_{ems}Z_{ems})/\hat{M}_{ijk}
$$

where $M_{ems}$ is the mass emitted to the cell $(i, j, k)$ during the time step, $X_{ems}, Y_{ems}$ are the coordinates of the source in the grid and $Z_{ems}^k$ is the effective injection height in the layer $k$, equal to the middle of the layer if no particular information is available.

For area sources, the approach depends on the source grid. If it is the same as the computational one, the mass centre is put to the middle of the cell (no extra information can be obtained). If the grids are different, the source is re-projected.

For each computational grid cell $(i, j)$, the centre of mass of emission is

$$
X_{em,ij} = \frac{\iint (x,y) \in (i,j) xM(x,y) \, dx \, dy}{\iint (x,y) \in (i,j) M(x,y) \, dx \, dy},
$$

$$
Y_{em,ij} = \frac{\iint (x,y) \in (i,j) yM(x,y) \, dx \, dy}{\iint (x,y) \in (i,j) M(x,y) \, dx \, dy},
$$

where $M(x, y)$ denotes the original source distribution. After that, the procedure is the same as in the case of point source Eq. (23).

4.3 Meteo-to-dispersion interface

Modifications described in Sect. 3 require staggered wind fields, which have to be provided by the meteo pre-processor.
(unless they are directly available from the input data). Moreover, the pre-processor needs to ensure consistency between the flow and air density fields (Prather et al. 1987; Rotman et al., 2004; Robertson and Langner, 1999). This is particularly important with the present advection scheme, since mixing ratio perturbations caused by the mass-flow inconsistency are not suppressed by numerical diffusion.

The wind pre-processing follows the idea of a “pressure fixer”, which means adding a correction $\delta V$ to the original horizontal wind field $V_0$ such that for their sum, the vertical integral of mass flux divergence corresponds to the surface pressure tendency:

$$\int_0^p \nabla \cdot (V_0 + \delta V) dp = -\frac{dp_s}{dt},$$  \hspace{1cm} (25)

where the surface pressure tendency $\frac{dp_s}{dt}$ is evaluated from the meteorological input data. The correction $\delta V$ is not uniquely determined, and SILAM adopts the algorithm of Heimann and Keeling (1989), where the correction term is given by the gradient of a 2-D potential function:

$$\delta V = \nabla \psi(x, y).$$  \hspace{1cm} (26)

Substituting Eq. (26) into (25) yields a Poisson equation for $\psi(x, y)$, which is solved to subsequently recover $\delta V$. The obtained correction flux is then distributed within the column proportionally to the air mass in each layer, ending up with the corrections to the horizontal winds. The vertical wind is then evaluated in each column to enforce the proper air-mass change in each cell.

4.4 Chemical module interface

This interface is implemented in a very simple manner: the mass centres are not affected by the transformations. The chemical module deals exclusively with concentrations in the grid cells. The newly created mass is added to the existing one, thus accepting its centre position in the cell. If some species did not exist before the transformation, the new mass centre is put to the middle point of the cell.

5 Testing the Galperin advection algorithm

5.1 Standard tests

A set of basic tests and comparison with some classical approaches has been presented by Galperin (1999) and P08 for the original scheme, along with Bott, Holmgren, and several other schemes. Their main conclusions were that the scheme is very good for sharp-edge patterns: in particular, it transports delta functions without any distortions. It had, however, issues with long slopes, smooth shapes, etc., where the tendency to gradually convert them to a collection of rectangles was noticeable.

Addressing these concerns, tests used during the scheme improvements and implementation in SILAM included puff-over-background, conical and sine-shaped peaks and dips, etc. (some examples are shown in Fig. 2), a divergent 1-D high-Courant wind test in the 1-D divergent wind field (Fig. 5), a constant-level background field in eight vortices with stagnation points (Fig. 6), and rotation tests for various shapes (Fig. 7).

The scheme stays stable at arbitrarily high Courant numbers and handles the convergence and divergence of the flows (Fig. 5).

Transport and rotation tests of the improved scheme maintain low distortions of the shapes; the $L_2$ norm of the error varies from 0.1% up to 3.8% of the initial-shape norm – for the most challenging task in Fig. 7. The effect of the improvements in comparison with the original scheme is demonstrated in Fig. 2, where the blue contours show the results of the original scheme. In particular, application of the smoothing Eq. (20) reduced the distortions of smooth shapes (red curves), largely resolving the concerns of P08: Fig. 2b presents the same test as one of the P08 exercises. However, the smoother also leads to a certain numerical viscosity of the scheme, so its use in problems requiring non-diffusive schemes (e.g. narrow plumes from accidental releases) should be avoided.

The test with eight vortices was difficult for the original scheme (Fig. 6a) due to its insufficient sub-grid resolution, but the improvement Eqs. (15)–(16), Sect. 3.2, resolved the problem (Fig. 6b). This refinement is instrumental in complex-topography domains.
5.2 Global 2-D tests

Performance of Galperin’s advection scheme in the global spherical domain was assessed with the collection of demanding tests of Lauritzen et al. (2012). The cases are designed to evaluate the accuracy of transport schemes at a wide range of resolutions and Courant numbers. The tests used a prescribed non-divergent 2-D velocity field defined on a sphere and consisting of deformation and rotation, so that the initial concentration pattern is reconstructed at the end of the test, \( t = T \), providing the exact solution \( \phi(t = 0) = \phi(t = T) \).

Four initial concentration distributions were used (Fig. 8): “Gaussian hills” with unity maximum value, “cosine bells” with a background of 0.1 and maxima of 1, “slotted cylinders” – a rough pattern with a 0.1 background and 1 maximum level, and “correlated cosine bells” – the distribution obtained from “cosine bells” with a function

\[
\phi_{\text{ccb}} = 0.9 - 0.8 \phi_{\text{cb}}^2.
\]  

The tests were run with SILAM on a global regular non-rotated lon–lat grid, with \( R = 6400 \) km and \( T = 12 \) h. Spatial resolutions were \( 6, 3, 1.5, 0.75, 0.375, \) and \( 0.1875^\circ \), each run with mean Courant numbers of \( \sim 5.12, \sim 2.56, \) and \( \sim 0.85 \) (for a \( 6^\circ \) grid they correspond to the model time step of \( T/12 = 1 \) h, \( T/24 = 30 \) min, and \( T/72 = 5 \) min), and a total of 18 runs for each initial pattern.

Examples of the most challenging runs with slotted cylinders at \( t = T/2 \) and at \( t = T \) are shown in Figs. 9 and 10, respectively. The corresponding error fields are collected in Fig. 11 as decimal logarithms of the absolute difference between the corresponding field in Fig. 10 and the slotted-cylinder initial shape of Fig. 8. The main complexity of the test was in reproducing the very tiny sharp-edge structures obtained from the cylinder cut at \( t = T/2 \) – and then returning them back by \( t = T \). The pictures, together with the error field at \( t = T \) (Fig. 11), show that already 24 time steps allow the scheme to make the shape recognisable (\( 3^\circ, C = 5.12 \) pattern), whereas 48 time steps allow for the main details to show up. Expectedly, certain deviations at the cylinder edge remain at any resolution – as is visible from the error fields.

Deviation of the resulting field \( \psi_T = \psi(t = T) \) from the initial shape \( \psi_0 = \psi(t = 0) \) was considered in three spaces: \( L_2, L_\infty, \) and \( L_1 \). The corresponding distance metrics are defined as follows:

\[
\begin{align*}
L_2 &= \left[ \frac{\int (\psi_T - \psi_0)^2}{\int \psi_0^2} \right]^{1/2}, \quad L_\infty = \frac{\max |\psi_T - \psi_0|}{\max \psi_0}, \quad L_1 &= \left[ \frac{\int |\psi_T - \psi_0|}{\int \psi_0} \right],
\end{align*}
\]

where \( \int [\bullet] \) is an area-weighted sum over latitude and longitude. The values of these three metrics for all model runs are presented in Fig. 12. The main interest of these curves is that they show the rate of the scheme convergence (straight grey lines correspond to the first- and second-order convergence rates). Expectedly, the rates depend on the transported shape (the smoother the shape, the faster the convergence) and on the norm used. Thus, the scheme converges in \( L_1 \) faster than in \( L_2 \), whereas in \( L_\infty \) no convergence in the case of sharp edges is an expected result. The rate in the \( L_2 \) norm is in between the first and the second order, whereas in \( L_1 \) it is close to the latter one.

Advection should also keep the local ratio of the tracer’s concentrations. Such a ratio between “cosine bells” and “correlated cosine bells” was calculated at \( t = T/2 \) and \( t = T \).
Figure 7. Double-vortex rotation tests for a rectangular split between the vortices (upper panels); three single-cell peaks and two connected rectangles (middle panels); and sin- and cone-shaped surfaces (lower panels). A series of time steps is shown in the left panels, except for the low panel (shown $t = 361$). Right panels: error field after one full revolution (obs 10-fold more sensitive scale and relative $L_2$ norm given above each plot). Max Courant $\sim 1.5$. Grid dimensions $400 \times 200$. Without smoother.
Since these initial patterns are related by Eq. (27), the concentration fields during the tests should maintain the same relation. The scatter plots of the concentrations in these two tests give an indication of how the ratio is kept. Ideal advection would keep all points on a line given by Eq. (28). The results of the tests for \( t = T/2 \) are shown in Fig. 13, where the results with and without the smoother in Eq. (20) are presented. The smoother improves the scheme mixing preservation; that is, it can be recommended to chemical composition computations, which usually also tolerate some numerical viscosity.

5.3 Global 3-D test with real wind

Testing the scheme with real-wind conditions has one major difficulty: there is no accurate solution that can be used as a reference. An exception is simulations of the constant-mixing-ratio 3-D field, which, once initialised, must stay constant throughout the run. Deviation from this constant is then the measure of the model quality. Such a test verifies both the scheme and the meteo-to-dispersion interface, which has to provide the consistent wind fields.

The constant-vmr test was set with winds taken from the ERA-Interim archive of ECMWF, for the arbitrarily selected month of January 1991 (Fig. 15). The model was initialised with \( \text{vmr} = 1 \) and run with 3° of lon–lat resolution and a time step of 30 min (max Courant number exceeding 13 in the stratosphere and reaching up to 2–3 in the troposphere). The model top was closed at 10 Pa, which corresponds to the top level of the ERA-Interim fields. The procedure described in Sect. 4.3 was used to diagnose the vertical wind component.

The results of the test are shown in Fig. 15, which depicts the model state after 240 h of the run, panel a) showing the boundary-layer vmr, and panel b) presenting it in the stratosphere. The zonally averaged vertical cross section is shown in panel c). Green colours in the pictures correspond to less than 1 % of the instant-field error.

An important message is that the limited distortions about 1–2 % are visible in a few places, but they are not related to topography, rather being associated with the frontal zones and cyclones. The comparatively coarse spatial and temporal resolution of the test makes the associated changes of the wind quite sharp, so that the dimension-split errors start manifesting themselves. Smoother flows in the stratosphere posed minor challenges for the scheme. The L2 error (not shown) is approximately proportional to the model time step.

6 Discussion

The presented SILAM v.5 transport module is based on the semi-Lagrangian advection scheme of Michael Galperin with subgrid information available through the positions of centres of masses. It poses certain challenges in implementation. Firstly, one has to organise the sub-grid information use and transmission between the advection and other model units. Secondly, the scheme requires storage of four full fields for each transported species (mass and moments) and care should be taken to maintain an efficient exchange between the processors and the computer memory. Thirdly, the possibility to run with high Courant numbers and MPI parallelisation via horizontal domain split can be utilised only if the MPI split allows for sufficient buffer zones. Finally, the better performance of the advection at a Courant number greater than 1 challenges the implementation of other modules, first of all, chemical and emission. Indeed, introduction of emitted mass once per long time step would result in a broken plume unless the mass is spread downwind over the corresponding distance. Similar problems show up in chemical...
Figure 9. Half-period ($t = T/2$) shapes for the 2-D global test with slotted cylinders for different spatial and temporal resolutions. Without smoother.
Figure 10. Final shapes ($t = T$) for the 2-D global tests with slotted cylinders for different spatial and temporal resolutions. Without smoother.
Figure 11. The error fields for the final shapes of Fig. 10 as compared with the slotted cylinder initial shape in Fig. 8. Without smoother.
Figure 12. Dependence of the performance metrics $l_1$, $l_2$, and $l_\infty$ for the spherical 2-D tests with initial shapes of Fig. 8. Dashed straight lines mark the slope for the first and second order of convergence. Without smoother.

transformation calculations. At present, the actual SILAM applications are performed with Courant close to but mostly smaller than 1 to avoid such problems.

The above challenges are mostly technical and their solution allows the scheme to demonstrate strong performance with low computational costs.

In particular, by attributing the release from point source to its actual location, one can reduce the impact of the common problem of Eulerian models: point release is immediately diluted over the model grid cell. This substantially improves the transport but does not solve the problem completely: (i) the chemical module still receives the diluted plume concentration, and (ii) the slab size in the case of the source near the centre of the grid cell will still be as large as the grid cell itself. A more accurate solution would be the plume-in-grid or similar approaches, which is being built in SILAM. Another example of the sub-grid information usage is utilization of full meteorological vertical resolution to calculate effective values of meteorological variables for thick dispersion layers (Sofiev, 2002).

The model can operate at any Courant number (Fig. 5). Its time step is limited not by grid cell size, but by a spatial scale of the wind-shear field; that is, it has to satisfy a much less restrictive Lipschitz criterion, which relates spatial and temporal truncation errors (Pudykiewicz et al., 1985). It follows from the advection step Eq. (6) and the reprojection step Eq. (8), which do not restrict new positions of the slabs: they can find themselves anywhere in the grid or outside it after the time step is made.
Figure 13. Mixing preservation test for cosine bells and correlated cosine bells Eq. (27) at \( t = T/2 \). Every two lines show the tests without (upper line) and with (lower line) a smoother (20).

SILAM heavily relies on such features of Galperin scheme as mass conservation and accountability: the scheme provides complete mass budget including transport across the domain boundaries. In particular, nesting of the calculations is straightforward and does not need the relaxation buffer at the edges of the inner domain: the inflow through the boundaries is described by the same slabs as the main advection. The scheme is also shape-preserving – in the sense this term is used by L14 – that is, it does not result in unphysical solutions, such as a negative mixing ratio. Some distortions are
still possible (Fig. 2), which can be reduced by the smoother described in Sect. 3.4, Eq. (20).

6.1 Standard advection tests

Evaluating the Galperin scheme with the simple tests (Figs. 2–7), one can point out the known issues of the classical schemes resolved in the Galperin approach: high-order algorithms suffer from numerical diffusion, oscillations at sharp gradients (require special efforts for limiting their amplitude), high computational costs and stringent limits to Courant number. None of these affects the Galperin scheme.

The main issue noticed during the implementation of the original scheme was the unrealistically high concentrations near the wind stagnation points. Thus, the concentration pattern at the test Fig. 6a resembles the situation of a divergent wind field. However, it is not the case: the 2-D wind pattern is strictly solenoidal. The actual reason is insufficient resolution of the advection grid: one centre of mass point is not enough if the spatial scale of the wind variation is comparable with the grid cell size. Tracking the edges of the slab rather than its centre resolves the problem (Fig. 6b).

The other challenging tasks for Galperin algorithm were those with smooth background and soft gradients, a frequent issue for semi-Lagrangian schemes, which is easily handled by more diffusive approaches. This feature was visible in the P08 tests where the scheme noticeably distorts the Gaussian and conical plumes. For the puff-over-background pattern, the scheme makes a single low-mass dip in the vicinity of the puff, which receives this mass (Fig. 2). From a formal point of view, the scheme does not conserve the higher moments inside the grid cell, which becomes a problem when the pattern changes at a spatial scale shorter than the grid cell size. The smoothing step (20) may be advised despite it having no rigorous basis and, as in L14 evaluation of other schemes, may damage some formal quality scores (adding this step introduces numerical viscosity – Fig. 2).

Figure 14. A histogram of the mixing diagnostic (stacked) for the same resolutions, Courant number and smoother factor as in Fig. 13. Metrics are the following (see text and Lauritzen et al. 2012 for more details): $l_r$ is “real mixing”, $l_u$ is “range-preserving unmixing”, and $l_o$ is “overshooting”. Values are relative to the reference CSLAM performance in L14 tests. The picture is comparable with panel (b) of Fig. 15 in L14.

Figure 15. Constant-vmr test with real-wind conditions after 122 h. (a) vmr within the boundary layer, (b) vmr above the tropopause, and (c) zone-average vertical cross section of vmr. Without smoother.
6.2 Global 2-D and real-wind advection tests

The application of the scheme to the highly challenging tests of Lauritzen et al. (2012) allowed its evaluation in a global 2-D case and comparison with the state-of-the-art schemes evaluated by L14 and Kaas et al. (2013).

Performing these tests with different spatial and temporal resolutions, as well as Courant numbers, suggested that the scheme has an “optimal” Courant number for each spatial resolution where the error metrics reach their minimum, so that the increase in temporal resolution is not beneficial. Indeed, in Fig. 12 the low Courant runs are by no means the most accurate. This is not surprising: for an ideal scheme, increasing the grid resolution and reducing the time step should both lead to gradual convergence of the algorithm; that is, the error metrics should reduce. For real schemes, higher temporal resolution competes with accumulation of the scheme errors with increasing number of steps. Convergence in L14 tests was still solid for all fixed Courant number series (Fig. 12), but excessive temporal resolution (specific to each particular grid cell size) was penalised by higher errors. Similarly, the most accurate representation of the correlated patterns is obtained from the runs with the intermediate Courant numbers (Fig. 13). This seems to be a common feature: the same behaviour was noticed by L14 for several schemes.

High optimal Courant numbers, however, should be taken with care. For L14, the smooth wind fields reduced the dimension-split error and made the long time steps particularly beneficial.

It is also seen (Fig. 11) that the best performance, in case of a near-optimal Courant, is demonstrated by the high-spatial-resolution simulations, which have reproduced both the sharp edges of the slotted cylinders, the flat background and the cylinder’s top planes.

The scheme demonstrated a convergence rate higher than 1 for all metrics and all tests with smooth initial patterns. Even for the most stringent test with the slotted cylinders, the scheme showed the first-order convergence rate in the $L_1$ norm (Fig. 12).

Among the other features of the solution, one can notice a certain inhomogeneity of the background field away from the transported bodies. The error is very small ($< 10^{-4}$) for high-resolution cases (Fig. 11) and $< 0.1$% for inexpensive set-ups, such as $\Delta \lambda = 0.75, C = 2.56$. For coarser resolutions, it grows. The inhomogeneity also grows with Courant number, which is opposite to the decreasing error of representation of the shapes themselves. The issue originates from the dimension-split error in polar areas, where the spatial scale of wind change becomes comparable with the distance passed by the slabs within one time step.

Similar non-monotonicity of background is visible for some schemes tested by L14. Unfortunately, no error fields are given there, but Figs. 7–10 there are comparable with our Fig. 9 (results without a smoother). With few exceptions (schemes TTS-I and LPM, notations of L14), all algorithms manifested such patterns unless filters are applied. For some schemes (SFF-CSLAM3, SFF-CSLAM4, UCISOMCS, CLAW, and CAM), these inhomogeneities are visible also for the tests with shape-preserving filters. One should note however that the 0.1 level, which distinguishes between the two violet colours in Figs. 9 and 7–10 of L14, corresponds to the background level in the slotted-cylinder test. As a result, even a very small deviation leads to the appearance of such shapes in the plots (note the stripes in the background of Fig. 8).

Comparing the so-called “minimal resolution” threshold for $L_2$, the norm of cosine bells to reach 0.033 (Fig. 3 of L14) for SILAM was about $0.75^\circ$, which puts it in the middle of that multi-model chart (the specific place depends on whether the shape preservation is considered or not).

Another criterion can be the optimal convergence of $L_2$ and $L_\infty$ norms for Gaussian hills: about 1.7–1.8 for SILAM – this is again in the middle of the L14 histograms, in the second half if the unlimited schemes (without shape-preservation filters) are considered and in the first half if the unphysical negative concentrations are suppressed (since the Galperin advection is strictly positively defined, no extra efforts are needed to satisfy this requirement).

Interestingly, the L14 tests were limited with $3^\circ$ as the coarsest resolution, and it was pointed out that the schemes start converging only when a certain limit, specific to each scheme, is reached. The SILAM results show similar behaviour only for the lowest Courant number (red lines in Fig. 12), which indeed required appropriate resolution to start working. Higher Courant set-ups were much less restrictive (the errors decrease with growing resolution also for coarse grids) and, as already pointed out, often worked better than the low Courant runs (similar to many L14 schemes).

The scheme demonstrated limited distortion of pre-existing functional dependence – see the cosine bells and correlated cosine bells tests in Eq. (27) (Fig. 13). Formal scores suggested by Lauritzen et al. (2012) calculated for the Galperin scheme are shown in Fig. 14. Notations are the following: $l_o$, “overshooting”, describes the values that fell outside the rectangular $[0.1:1]$ (Fig. 13), $l_u$, “shape-preserving unmix”, describes the values inside that rectangular but outside the “lens” formed by its diagonal $(0.1, 1)$–$(1, 0.1)$ and the curve, and $l_t$, “real mixing”, describes the values inside the “lens”. Comparison with L14 (Fig. 15, middle panel) shows that the Galperin scheme outperforms CLAW, SLFV-ML, SLFV-SL, and all set-ups of ICON schemes, being close to CAM-SE, MIPAS, and HOMME, and trailing behind the runs with CSLAM, HEL, SFF, and UCISCOM schemes.

A peculiarity of the mixing diagnostic scores is that they are significantly affected by the background areas far from the advected bells, which occupy only a small fraction of the domain (Fig. 8). As a result, small background fluctuations discussed above in application to slotted cylinders (see the error field in Fig. 11) contribute significantly to the mixing.
diagnostic scores too. In particular, the high Courant simulations, which accurately reproduce the bells themselves (the dots are close to the curve in the scatter plots in Fig. 13), still show poor formal scores due to non-zero width of the cloud near the location (0.1, 1), where all background dots should arrive. This issue contributes most significantly to the “overshooting” part of the error, but also to the other two components.

Expectedly, the smoother improves the mixing diagnostic scores, mainly affecting the representation of the bells themselves (Fig. 13). This is in contrast with the schemes tested in L14, where the shape-preservation filters mostly removed the penalty for overshooting the background but rarely improved the other two components, sometimes worsening them.

Following the conclusions of Sect. 3.4 and the 1-D tests, we used the smoothing factor of 0.08, which is a compromise between the scheme diffusivity and distortion reduction. As a result, some non-linearity exists also in the smoothed solution. The test showed that a simple increase in temporal resolution leads to an increase in the number of steps and related re-projections, which then worsen the representation of the bells – but improve the background field by reducing the dimension-split errors. A synchronous rise of the resolution in time and space with the same Courant number (columns in Fig. 13) showed better results for higher-resolving set-ups.

Further investigating the flat-field behaviour in complex wind patterns, the simulations with the constant-vmr initial conditions (Fig. 15) were performed, showing that the model has no major problem in keeping the homogeneous distribution: deviations do not exceed a few %, with no relation to topography. The existing ups and downs of the vmr are related to cyclones and atmospheric fronts, which challenge the dimension-splitting algorithm rather than the core 1-D advection (it transports the homogeneous field perfectly – no distortion was found after 105 steps regardless of the Courant number). Increasing the resolution leads to a lower “unmix” of the pattern (not shown). This experiment refines the “optimal Courant” recommendation of the L14 test, which had smoother wind fields and, consequently, a higher optimal Courant number. For real-life applications, especially with coarse grids, it may be necessary to choose a time step short enough to ensure comparable levels of time- and space-wise truncation errors (Pudykiewicz et al., 1985). This case also argues for developing the 2-D implementation of the Galperin scheme, which would eliminate the horizontal dimension split.

6.3 Where to use the smoother

When deciding whether to apply the smoother Eq. (20), one has to keep in mind that the Galperin scheme is always positively defined and does not need a shape-preserving filter to provide a “physically meaningful” solution, i.e. without negative values. It is free from this caveat. The purpose of the smoother is only to reduce the non-linear distortions of fields.

The smoother has both a positive and negative impact on the scheme performance. Among the positive ones are that (i) it damps the distortions of smooth shapes and gradients (Sect. 3.4), (ii) it reduces the amplification factor, precluding it from exceeding 1 even for a few time steps (Sect. 3.5), and that (iii) it reduces the unmixing problem (Fig. 14). Its negative features are that (i) the obtained solution is diffusive (Sect. 3.4), (ii) moderate and high frequencies in the solution spectrum are damped (Sect. 3.5), and that (iii) formal scores and convergence rates are lower in some tests (Sects. 5.2 and 6.2). The smoother has little impact on background inhomogeneity.

Most of the positive and negative features coincide with the impact of shape-preserving filters (e.g. L14), despite the different idea and formulations.

Since the smoother computational cost is negligible, one can decide whether to apply it depending only on the problem at hand. Strict interconnections between the species, smooth patterns and tolerance to diffusion form a case for the smoother. Conversely, sharp plumes over zero background (e.g. the accidental release case) argue against it.

The smoother impact grows monotonically with its parameter $\epsilon$. Numerous tests showed that the distortions and above 1 amplification factor essentially disappear at $\epsilon \sim 0.08$, where the diffusivity also becomes significant. This value appeared stable with regard to Courant number and set-up of the tests.

6.4 Efficiency of the Galperin advection scheme

Evaluation of the scheme efficiency is always very difficult as it strongly depends on the algorithm implementation, but also on computers, parallelisation, compiler options, etc. Nevertheless, basic characteristics of the scheme can be deduced from comparison of its original version with several classical schemes made by Galperin (2000). It included, in particular, EM72 and Bott, which appeared stable with regard to Courant number and set-up of the tests. The updated scheme version, however, is bound to be heavier. It is also worth putting it in line with modern approaches.

In this section, the efficiency of the updated Galperin scheme is evaluated from several points of view: (i) the scalability with regard to the number of transported species, spatial and temporal resolution, specific to the problem at hand, (ii) comparison with “standard implementation” of the Bott algorithm and the semi-Lagrangian scheme, and (iii) comparison of the run time in the L14 tests with the HEL and CSLAM schemes.

www.geosci-model-dev.net/8/3497/2015/
6.5 SILAM run time vs. number of species, temporal and spatial resolution

The scalability of the scheme and the whole SILAM model was tested in real-wind global simulations for an arbitrarily taken 3 days (15–17 May 2012). The reference run was set with 0.5° resolution, six vertical layers, a time step of 30 min, and one aerosol species. Two types of emission were considered: an artificial 1 h long source filling up the whole 3-D domain, and the SILAM own wind-blown dust emission model, which created dust plumes from sandy areas of the Sahara. Vertical diffusion, which is coupled with vertical 1-D advection, was turned off for artificial source tests but turned on for dust sources in order to allow the model to quickly populate the upper layers of the domain. Then, the number of aerosol species, spatial and temporal resolutions were repeatedly doubled (one change at a time).

The model was run in a single-processor mode but compiled with O3 optimisation and OMP code pre-processing. Runs were made in a notebook with an Intel Core i7 processor and repeated in a workstation with an Intel Xeon E5. The scaling differed by 10–20 %, which was considered to be negligible.

The results (Fig. 16) highlight the scalability of the scheme and its implementation in SILAM. The species-unrelated time of horizontal 2-D advection (Fig. 16a, offset in regression line) is ~30 % of a single-species computation time (represented via the slope). This “overhead” is, in fact, the transport-step integrals Eqs. (17)–(19), which are computed only once and used for all species. Higher overhead of the vertical advection is due to the necessity to handle the uneven vertical layers, which makes its scaling just 20 % better than the 2-D horizontal one. It also has larger species-independent overhead.

With the chemical module turned off, advection constitutes ~85 % of the total model run time.

Since the scheme operates with the source grid cells, it can check that $M^f = 0$ before going into computations, which gives a very substantial speed-up in the case of limited-volume plumes (Fig. 16b). In the Saharan dust run, the horizontal advection time is about twice lower, whereas the vertical advection, even together with diffusion, becomes all but negligible, owing to efficient filtering of zero columns in comparison with lon or lat stripes.

A faster-than-proportional growth of the horizontal advection time with increasing spatial resolution (Fig. 16c, normalised run time) is a result of a growing Courant number: for a 4-times smaller grid cell (0.25° lon–lat resolution), the time step of 30 min means $C \geqslant 1$ over a large part of the domain. As a result, transport integrals Eqs. (17)–(19) have to be analysed over longer paths. Still, the growth is much smaller than the cost of 4-fold reduction of the time step, which makes the high-C computations attractive. Vertical advection is not affected and its time is proportional to the number of columns to analyse.

The time spent by advection is practically proportional to the temporal resolution (Fig. 16c); that is, it follows the number of times the advection is computed in the run.

6.6 Comparison with efficiency of other schemes

Comparison with other schemes is arguably the most uncertain part of the exercise: the scheme efficiency is strongly dependent on the quality of the implementation (note the different results for the Bott scheme obtained by Galperin, 2000, and Petrova et al., 2008). To obtain reproducible results, we made this comparison against the “standard implementation” of the Bott code available from the Internet (http://www2.meteo.uni-bonn.de/forschung/gruppen/tgwww/people/abott/fortran/fortran_english.html, visited 28 September 2015). Since our code is also available, this comparison is reproducible.

The test with $10^4$ time steps, 2000 grid points in a 1-D periodic grid, Courant number = 0.1, and one species took 0.92 s for the Galperin scheme (~0.3 s for cell border advection, ~0.6 s for slab reprojection) and 0.85 s for the Bott scheme. This confirms the expectation that the updates of the Galperin scheme from its initial version about tripled its run time, which is now similar to that of the Bott scheme. However, the Galperin scheme still scales better with the number of species: as shown in the previous section, only reprojection is multiplied by the number of species, whereas the Bott scheme does not have such a saving possibility.

The above numbers should be considered as indicative only since the environment for the tests was completely artificial: the schemes were used as a stand-alone code applied in 1-D space. The Galperin scheme needed only one moment instead of three, which would be the case for 3-D advection. Despite very limited extra computations, this would still raise the memory exchange. The Bott scheme was taken without a shape-preservation filter, which would be needed for any real-life applications.

The tests were also made for our own implementation of the semi-Lagrangian scheme (took ~50 % longer than the above timing), but its efficiency was not carefully verified.

The L14 tests allowed rough benchmarking of the SILAM implementation of the scheme in 2-D tasks. In particular, the run with 0.75° resolution and 120 time steps can be related to the performance of the HEL and CSLAM schemes, which were tested against the same test collection by Kaas et al. (2013). Extrapolating the charts of Fig. 13 of Kaas et al. (2013) to one species (the range given there is 2–20 species), the test takes about 190 s for HEL and 300 s for CSLAM, but only 47 s for SILAM; i.e., the difference was about 4 and 6 times, respectively.

Formal benchmarks of the computers, the main uncertainty in this comparison, are essentially the same: Kaas et al. (2013) used an Intel Core2 Duo E6550 processor (Intel Linpack 20 GFlops, http://www.techpowerup.com, visited 8 October 2015). Our tests were run on a simple notebook
with a mobile Intel Core i5-540M Duo (Intel Linpack 18.5 GFlops). These CPUs were also compared in http://www.cpubenchmark.net (visited 8 October 2015), which also put them within 20% of each other, albeit that the i5-540M was put forward. The memory bandwidth of our notebook, as always for compact computers, was modest: 7.2 GB s$^{-1}$ (STREAM test, http://www.cs.virginia.edu/stream/ref.html accessed 5 October 2015). We used a GNU compiler with –O3 optimisation without parallelisation, similar to Kaas et al. (2013).

6.7 Further boosting the scheme efficiency: parallelisation

In SILAM applications, advection is parallelised using the shared-memory OMP technology, whereas the MPI-based domain split is being developed. The OMP parallelisation is readily applicable along each dimension, thus exploiting the dimensional split of the advection scheme. For MPI, care should be taken to allow for a sufficient width of the buffer areas to handle the Courant > 1 cases.

The original scheme was formulated for the bulk mass of all transported tracers, thus performing the advection step for all species at once: the tracer’s mass in the slab definition Eq. (5) was the sum of masses of all species. This is much faster than the species-wise advection and reduces the number of the moments per dimension down to 1 regardless of the number of tracers. It is also useful in the case of strong chemical interconnections between the species because the bulk advection keeps all pre-existing relations between the species. However, transport accuracy diminishes if the species have substantially different lifetimes in the atmosphere, are emitted from substantially different sources, or otherwise decorrelated in space.

7 Summary

The current paper presents the transport module of the System for Integrated modeLling of Atmospheric coMposition SILAM v.5, which is based on the improved advection routine of Michael Galperin combined with separate developments for vertical diffusion and dry deposition.

The cornerstone of the advection scheme is the subgrid information on distribution of masses inside the grid cells, which is generated at the emission calculation stage and maintained in a consistent way throughout the whole model, including chemical transformation, deposition, and transport itself. This information, albeit requiring substantial storage for handling, allows for accurate representation of transport.

The scheme is shown to be particularly efficient for point sources and sharp gradients of the concentration fields, still showing solid performance for smooth patterns. The most challenging task was found to be the puff-over-plain test, where the scheme showed noticeable distortions of the conform.
centration pattern. Application of a simple smoother efficiently reduces the problem at a cost of non-zero viscosity of the resulting scheme.

Advanced tests and comparison with state-of-the-art algorithms confirmed the compromise between the efficiency and accuracy. SILAM performance was fully comparable with the other algorithms, outperforming some of them.

Among the future developments, implementation of the scheme in 2-D space and replacement of the smoother with extensions of the core advection algorithm are probably the most pressing ones.

Code availability

SILAM is a publicly available model. Our experience shows however that its successful modeling and understanding of the model concepts. Therefore, SILAM is available on an on-request basis from the authors of this paper, who also provide support in the initial model installation and set-up. The model description, operational and research products, as well as reference documentation, are presented at http://silam.fmi.fi (accessed 5 October 2015). The model user’s guide is available at http://silam.fmi.fi/doc/SILAM_v5_userGuide_general.pdf (accessed 5 October 2015). Potential model users are also encouraged to refer to the SILAM Winter School material at http://silam.fmi.fi/open_source/SILAM_school/index.htm (accessed 5 October 2015).


Acknowledgements. The development of the SILAM model was supported by the ASTREX project of the Academy of Finland, as well as by the ESA-ATILA and FP7-MACC projects. The authors thank the late M. Galperin for the original version of the scheme, A. Bott for the publicly available version of his scheme, and E. Kaas and an anonymous reviewer for detailed comments on the manuscript.

Edited by: V. Grewe

References


www.geosci-model-dev.net/8/3497/2015/


