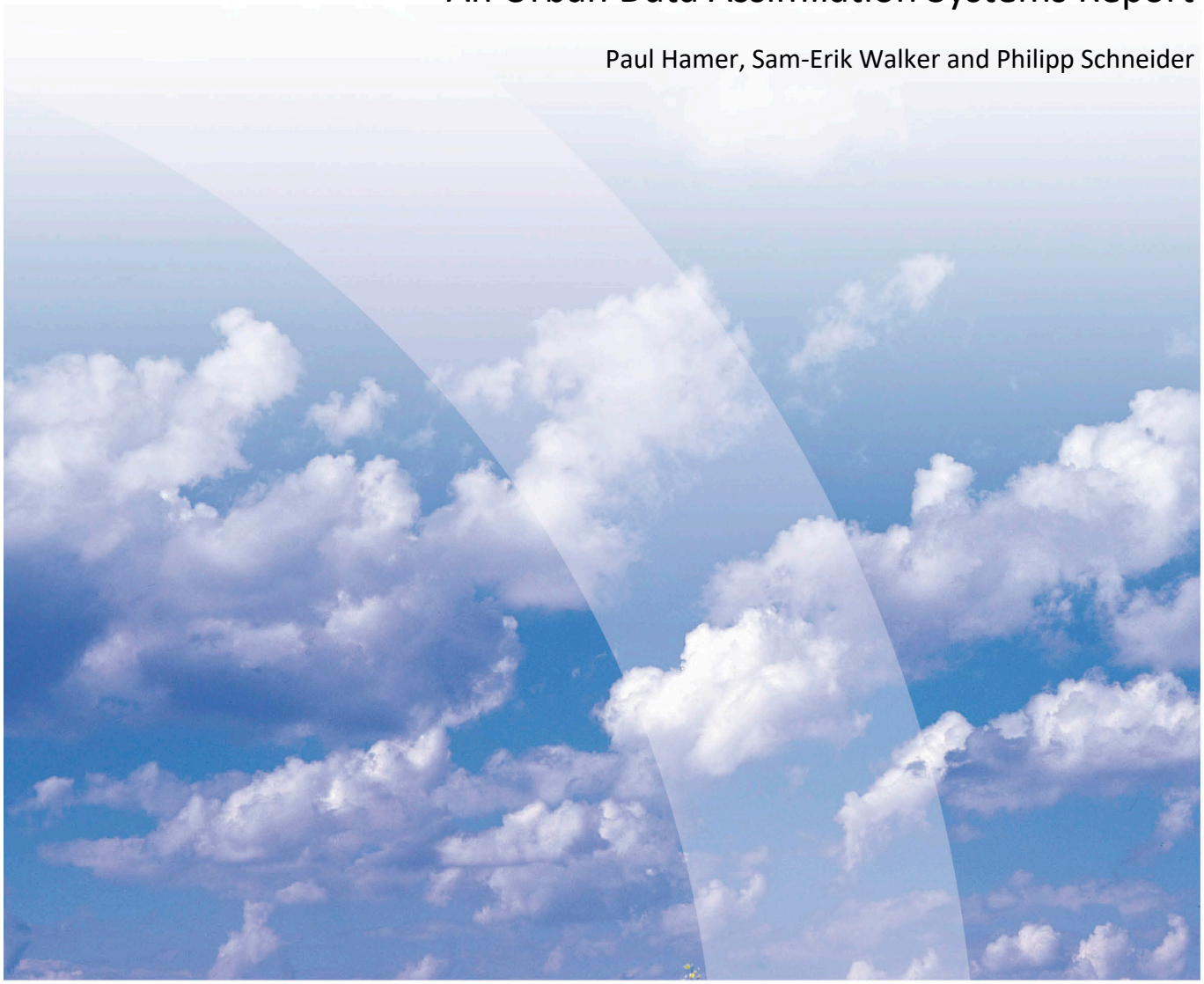
Norsk institutt for luftforskning
Norwegian Institute for Air Research

Appropriate Assimilation Methods for Air Quality Prediction and Pollutant Emission Inversion

An Urban Data Assimilation Systems Report

Paul Hamer, Sam-Erik Walker and Philipp Schneider



NILU report 25/2020

NILU report 25/2020	ISBN: 978-82-425-3027-1 ISSN: 2464-3327	CLASSIFICATION: A - Open report
DATE 2021.01.11	SIGNATURE OF RESPONSIBLE PERSON Ole-Anders Braathen, Deputy director	NUMBER OF PAGES 21
TITLE Appropriate Assimilation Methods for Air Quality Prediction and Pollutant Emission Inversion An Urban Data Assimilation Systems Report		PROJECT LEADER Paul Hamer NILU PROJECT NO. 120008
AUTHOR(S) Paul Hamer, Sam-Erik Walker, and Philipp Schneider		QUALITY CONTROLLER Sam-Erik Walker
REPORT PREPARED FOR NILU		CONTRACT REF. SIS Internal Project

ABSTRACT

This report presents a review of data assimilation methods applicable to air quality. In the introduction, we first describe a brief history of data assimilation method development in the context of numerical weather prediction (NWP), and then we highlight key differences when applying data assimilation methods to air quality prediction from NWP applications. Based on these differences, we outline a set of key requirements for data assimilation when applied to air quality. Following this, we review the available data assimilation algorithms and attempt to identify suitable data assimilation methods that could be applied with air quality models. This review and its findings form the basis of the developments to be carried out in the Urban Data Assimilation Systems project.

NORWEGIAN TITLE

Anbefalte assimileringsmetoder for prediksjon av luftkvalitet og inversjon av utslipp

KEYWORDS

Data Assimilation | Air Quality | Modelling

ABSTRACT (in Norwegian)

Rapporten presenterer en oversikt over dataassimilasjonsmetoder som kan anvendes for luftkvalitetsmodellering. Innledningsvis beskrives kort historisk bakgrunn for bruk av dataassimilasjon i numerisk værvarsling, der vi legger vekt på forskjellene mellom anvendelse av assimilasjon i meteorologiske varslingsmodeller og i spredningsmodeller for luftkvalitet. Basert på disse forskjellene beskrives så ønskede egenskaper til assimilasjonsmetoder for luftkvalitetsmodellering. Deretter gis en oversikt over tilgjengelige assimilasjonsmetoder, der vi søker å identifisere de mest aktuelle for vårt bruk som grunnlag for videre anvendelser i prosjektet.

PUBLICATION TYPE: Digital document (pdf)	COVER PICTURE: Source: NILU

Preface

This report was written as part of the urban data assimilation systems (UDAS) project funded by NILU. The report was commissioned as part of work package 4 (“Development of DA algorithm in a simplistic model”). The specific overriding objective was to justify the selection of a suitable data assimilation algorithm for use with the EPISODE air quality model. To this end, the report carries out a literature review of existing data assimilation methods and then evaluates each one for suitability for implementation in the EPISODE air quality model.

Contents

Summary

This report presents a review of data assimilation methods applicable to air quality. In the introduction, we first describe a brief history of data assimilation method development in the context of numerical weather prediction, and then we highlight key differences when applying data assimilation methods to air quality prediction. Based on these differences, we outline a set of key requirements for data assimilation when applied to air quality. Following this, we review the available data assimilation algorithms. Within the review, we provide a list of the advantages and disadvantages of each algorithm. All of this information is then summarised in a selection process to identify which algorithm should be used within the later stages of the UDAS (urban data assimilation systems) project. This review and its findings form the basis of the developments to be carried out in the Urban Data Assimilation Systems project. The algorithms that we identify for further inspection and development are the optimal interpolation method, the 4D-variational method, and the 4D-Ensemble Kalman Filter.

Appropriate Assimilation Methods for Air Quality Prediction and Pollutant Emission Inversion

An Urban Data Assimilation Systems Report

1 Air Quality Modelling and Data Assimilation – Statement of the Problem

Air quality models consist of a set of discrete non-linear equations that describe the evolution of the simulated three-dimensional field of pollutant concentration state variables, $\boldsymbol{c}$, forwards in time. In addition to the current model state, $\boldsymbol{c}_t$, future model states, $\boldsymbol{c}_{t+n}$, depend on pollutant emissions, $\boldsymbol{q}_t$, the background concentrations external to the model domain, $\boldsymbol{b}_t$, and the three-dimensional fields of meteorology, $\boldsymbol{m}_t$. Since these model input variables are discretized in time and space, they are abstractions and approximations of the actual true state of each variable. We can define these variables as the control variables on the future simulated state, $\boldsymbol{x}_t$, such that

$$\boldsymbol{x}_t = (\boldsymbol{c}_t, \boldsymbol{q}_t, \boldsymbol{b}_t, \boldsymbol{m}_t) \qquad (1)$$

It is then possible to represent an idealized case of the temporal evolution of the model concentrations as

$$\boldsymbol{c}_{t+1} = \boldsymbol{F}_t^{AQ}(\boldsymbol{x}_t), t = 0, \dots, N-1, \boldsymbol{x}_t = (\boldsymbol{c}_t, \boldsymbol{q}_t, \boldsymbol{b}_t, \boldsymbol{m}_t) \qquad (2)$$

where $\mathbf{c}_{t+1}$ represents the model state in the next time step. Here, $\mathbf{F}_t^{AQ}$ is a function representing the discrete non-linear equations that evolve the current model state, $\boldsymbol{c}_t$, according to the control variables, $\boldsymbol{x}_t$, in that time step. Errors in the estimation of the current and future model states are driven by errors in the control variables, i.e., the current state, emissions, meteorology, and background concentrations. Note that this idealized case ignores errors resulting from fundamental aspects of the model formulation and its parameterizations; although not considered here yet we term these errors, $\boldsymbol{\varepsilon}_t$.

We would like to consider here the application of data assimilation as a means to improve the estimation of both the current and future model states, $\mathbf{c}_t$ and $\boldsymbol{c}_{t+n}$ and model parameters of key interest and importance, e.g., the emissions, $\boldsymbol{q}_t$.

Many of the data assimilation algorithms in existence today have been developed specifically to improve numerical weather prediction forecasting (Evensen, 1994; Griffith & Nichols, 2000; Lorenc, 1986; Lorenc et al., 2015; Trémolet, 2007), which has important implications for our considered application in air quality. We should therefore begin by providing a clear description of data assimilation and its applications in meteorology to fully understand the potential and the limitations of existing data assimilation algorithms. We begin with a simple mathematical description of an idealized meteorological model (i.e., one with no model error). We can formulate the temporal evolution within such an idealized model case as,

$$\boldsymbol{x}_{t+1} = \boldsymbol{F}_t^{M}(\boldsymbol{x}_t), t = 0, \dots, N-1, \boldsymbol{x}_t = \boldsymbol{m}_t \qquad (3)$$

where we can reuse the parameters from equations (1) and (2), but in this case, $\boldsymbol{x}_t$ represents the current meteorological atmospheric state, $\boldsymbol{m}_t$. $\mathbf{F}_t^M$ represents the model equations responsible for evolving the meteorological state forwards from one time-step to the next.

Thus, we can see a clear difference in the mathematical formulation of the temporal evolution in air quality and meteorological models, i.e., the future state in meteorological models is assumed to vary only as a function of the current meteorological state, $\boldsymbol{m}_t$. In this idealized case, innate model errors are ignored completely, and the current state is assumed to be the only control variable relevant for the future state. Further to this, it is assumed forecast errors into the future arising within the meteorological model are primarily driven by initial condition errors. In fact, these statements are true to a first-order approximation, and meteorological forecast errors have historically been limited by the error in the estimation of the initial state.

Data assimilation algorithm development in meteorology was therefore done specifically to solve the problem of uncertainties in initial conditions. For example, the 3D-Var algorithm (Lorenc, 1986) can be used to improve estimates of the current state to create better initial conditions for improved future model state estimation. In another example, it is a better strategy to attempt to solve for some past model state or initial condition using strong constraint 4D-Var (Klinker et al., 2000; Mahfouf & Rabier, 2000; Rabier et al., 2000), e.g., $\mathbf{x}_0$. The latter example allows the use of observations of the present to update and correct initial conditions in the past, and thus achieve better prediction of both the current and near-future states.

Significant effort and advances have been made to improve NWP forecasting by focusing on improvements in DA algorithms that reduce errors in the estimation of the initial state (Evensen, 1994; Griffith & Nichols, 2000; Lorenc, 1986; Lorenc et al., 2015; Trémolet, 2007). Algorithms and other NWP improvements are so advanced now to the extent that errors in the estimation of the initial state no longer dominate errors in the forecast (Griffith & Nichols, 2000; Trémolet, 2007). It has become recognized that model errors, e.g., model resolution and discretization errors indeed play a role in limiting the ability to make accurate NWP forecasts (Talagrand, 1998; Trémolet, 2007). Solving only for initial conditions in the presence of model error will lead to imperfect estimates of the initial state, and to forecast errors. Efforts have now turned towards trying to solve the initial condition problem in the presence of an imperfect model. For example, the 4DEnVar (Desroziers et al., 2016; Lorenc et al., 2015) and weak constraint 4D-Var (Griffith & Nichols, 2000; Trémolet, 2007) were both specifically formulated to account for systematic model error in their solutions; we will discuss the details of these and other algorithms in Sect. 2. For example, the weak constraint 4D-Var reformulates equation (3) to include model error, and the algorithm is adapted to account for this more realistic situation.

$$\boldsymbol{x}_{t+1} = \boldsymbol{F}_t^M(\boldsymbol{x}_t) + \boldsymbol{\varepsilon}_t, t = 0, \dots, N-1, \boldsymbol{x}_t = \boldsymbol{m}_t \qquad (4)$$

In the weak constraint 4D-Var, $\boldsymbol{\varepsilon}_t$, is assumed to represent systematic errors that are either constant or evolve or in time, and the algorithm is set up to solve for both $\boldsymbol{x}_t$ and $\boldsymbol{\varepsilon}_t$.

These latest evolutionary steps in data assimilation algorithms are of critical importance for data assimilation in the context of air quality and for atmospheric chemistry models.

Atmospheric pollutants have a limited lifetime in the atmosphere, and this can vary from minutes for NO_x to several days for longer lived particulate matter. This characteristic of the state variable of interest creates a problem for assimilation methods based only on the direct estimation of pollutant concentrations (e.g., 3D-Var) because the data assimilation corrections, termed increments, decay over time as the pollutants are removed from the atmosphere. Thus, in a case where the emissions are underestimated, the increments will decay over time and the model will relax back to the background model state estimate (Emili personal communication). This, therefore, creates the necessity to set the pollutant emissions as an assimilation control variable to correctly estimate the concentrations of short-lived pollutants both present and future. Achieving this will help to remove biases in concentrations (biases are the primary form of error arising in air quality models (Marécal et al., 2015)) arising from systematic emission errors. Aside from this approach being advantageous for concentration state estimation, it also provides the benefit that emissions are themselves an interesting variable to estimate. Emission estimates of pollutants give important information about economic sectors that contribute to pollution and allows the formulation of effective pollution control strategies.

Given the need to consider pollutant emissions as a control variable, this alters the requirements for data assimilation compared to the traditional NWP type problem shown in Eq. (3). As we can see in Eq. (2), the future model state depends on the control variables, $\boldsymbol{x}_t$, which consist of the current state, $\boldsymbol{c}_t$, and the input variables from that timestep consisting of the emissions, $\boldsymbol{q}_t$, the meteorology, $\boldsymbol{m}_t$, and the background concentrations, $\boldsymbol{b}_t$. If we make our control variables the pollutant emissions only, this creates a situation where we potentially ignore other sources of error that influence the prediction of the future pollution concentration state, e.g., those arising from errors in the input meteorology and other types of innate model error. Furthermore, this can also lead to an imperfect solution to the emissions estimate whereby the algorithm overcorrects the emissions by trying to solve for other sources of observation minus model error biases. Lastly, errors in emission estimates are typically more systematic with a bias relative to the true emissions, which can potentially create problems for some data assimilation algorithms which assume errors to be more random and unbiased with a Gaussian distribution.

In practice, while meteorological data are used as an input in air quality models, meteorological models are often run by dedicated teams external to the air quality modelling community. As a result, errors in the meteorological input data are often ignored or overlooked; even if errors in meteorological input data are identified, air quality model users lack the means to take rectifying action and the data themselves are considered immutable. Within this paradigm of thought, and for the purposes of this report, errors on the meteorological input are errors innate to the air quality modelling system.

The latest developments in data assimilation offer the possibility to account for systematic errors arising in air quality models (Marécal et al., 2015) resulting from errors in the input meteorology, model equations, model spatial resolution, and temporal discretization. Thus, instead of trying to formulate air quality model forecast problems as in equation (2), it should also be possible to reformulate the air quality problem as

$$\boldsymbol{x}_{t+1} = \boldsymbol{F}_t^{AQ}(\boldsymbol{x}_t) + \boldsymbol{\varepsilon}_t, t = 0, \dots, N-1, \boldsymbol{x}_t = \boldsymbol{q}_t \qquad (5)$$

such that an error term is now included. Having defined the problem we wish to address, we will now begin a review of the data assimilation methods that are currently available. We will then try to quickly summarise the advantages and disadvantages of each method concerning the application under consideration.

2 Review of Existing Data Assimilation Methods

2.1 Optimal Interpolation

2.1.1 Technical Description

Optimal Interpolation (OI) is one of the earliest assimilation techniques and has seen widespread adoption and use in numerical weather prediction for several decades. First proposed by Gandin (1963), OI allows for combining an *a priori* field (typically from a model) with observations to produce an analysis field that merges the two datasets using objective weights according to the corresponding uncertainties.

OI and geostatistical. techniques such as universal kriging or kriging with external drift (Chiles & Delfiner, 2009; Goovaerts & others, 1997) are very closely linked. While the former originated from the meteorological and atmospheric community (Gandin, 1963; Shiryayev A.N., 1992), the latter was developed within the mining and natural resources community (Matheron, 1963). OI and geostatistics use somewhat different mathematical frameworks but they produce equivalent results given the same inputs (Cressie, 1990; Knudsen & Lefohn, 1988).

OI is equivalent to 3D-Var (Kalnay, 2003), although both use entirely different mathematical techniques.

OI can only be considered “optimal” when the error characteristics are known perfectly, which is typically not the case. For this reason, OI is often also referred to as objective analysis or Statistical Interpolation (SI). Given the wide use of the term OI in the literature and to avoid confusion with statistical interpolation in geostatistics, we continue to use the term OI in this document for clarity.

Following the nomenclature used in Kalnay (2003), OI calculates the analysis vector $\boldsymbol{x_a}$ as

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \boldsymbol{W}[\boldsymbol{y}_0 - H(\boldsymbol{x}_b)] \qquad (6)$$

where $\boldsymbol{x_b}$ is the background field vector from a model, $\boldsymbol{W}$ is a matrix of weights, $\boldsymbol{y}_0$ is the set of observations, and H is the observation operator that translates the background values into observation space (often using simple bilinear interpolation if the units of the background field and observations are the same). The weight matrix $\boldsymbol{W}$ is calculated as

$$\boldsymbol{W} = \boldsymbol{B}\boldsymbol{H}^{\boldsymbol{T}}\left(\boldsymbol{R} + \boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^{\boldsymbol{T}}\right)^{-1} \qquad (7)$$

where $\boldsymbol{B}$ is the background error covariance matrix, the matrix $\boldsymbol{H}$ is the linear tangent perturbation of H, and $\boldsymbol{R}$ is the matrix of observation error covariances (which is diagonal if the observations errors at different locations are assumed to be uncorrelated).

The analysis error covariance $\boldsymbol{P}_a$ can then be calculated as

$$\boldsymbol{P}_a = (\boldsymbol{I} - \boldsymbol{W}\boldsymbol{H})\boldsymbol{B} \qquad (8)$$

where $\boldsymbol{I}$ is the identity matrix.

2.1.2 Advantages

- Relatively straightforward and quick to implement for small cases (no requirement to derive adjoints or implement cost functions)
- Grounded in many decades of research and application
- Strong ties or even equivalency to similar methods such as geostatistical interpolation and 3D-Var
- Background error covariance matrix can be specified manually in detail (as opposed to kriging-based techniques which typically only use a distance-based covariance function).
- Model-independent (can be run offline).

2.1.3 Disadvantages

- Background error covariance matrix must be specified, and it is often based on crude assumptions.
- Less computationally efficient to implement for large-scale problems compared with e.g. variational methods.

2.1.4 Algorithm Availability

The algorithm has already undergone preliminary testing with the EPISODE model in the context of other projects.

2.2 Extended Kalman Filter

2.2.1 Technical Description

Kalman filter (KF) methods represent a large class of different techniques for performing data assimilation and the equations of the various variants of the Kalman filter are essentially the same as for the OI method as described above. The original Kalman filter (Kalman, 1960) was proposed for a purely linear system of evolution. Later it was extended to non-linear models and is then known as the Extended Kalman filter (EKF).

One of the foremost properties of these methods is that they automatically estimate and update the model error covariance matrix $\boldsymbol{B}$ from one time-step to the next, using the model operator (the dispersion model) itself. The Kalman filter in its original form, however, whether

linear or extended, has too large a computational complexity to be implemented in practice for large model state spaces, so simplifications of the original filter are necessary.

The assumptions behind the different KF methods are essentially the same as the assumptions behind OI and the various variational methods below, namely that the model and observation errors should have a Gaussian probability distribution.

2.2.2 Advantages

- The most important advantage of the Kalman and extended Kalman filters is the automatic estimation and updating of the model error covariance matrix $\boldsymbol{B}$ from one timestep to the next using the model operator (the dispersion model) itself.

- Another advantage of the extended Kalman filter is that the state transition and observation models do not need to be linear functions of the state vector, but rather only non-linear differentiable functions.

2.2.3 Disadvantages

- Too large complexity to be implemented in practice for large model state spaces, so some simplification of the original filter is necessary.
- If the initial state estimate is wrong, or if the process is modelled incorrectly, the filter may quickly diverge, owing to its linearization.
- The estimated covariance matrix tends to underestimate the true covariance matrix and therefore might become statistically inconsistent without the addition of some extra stabilizing noise.
- Unlike its linear counterpart, the extended Kalman filter is generally not an optimal estimator unless the measurement and state transition models are linear.

2.2.4 Algorithm Availability

This algorithm has not been implemented at NILU and nor do we have ready access to a working version of this algorithm.

2.3 Ensemble Kalman Filter

2.3.1 Technical Description

The ensemble Kalman filter (EnKF) was originally proposed in Evensen (1994). The idea behind the method is to apply an ensemble of N model states $\boldsymbol{x_i}$ for i=1,...,N, to represent and estimate the model error covariance matrix $\boldsymbol{B}$. The method thus avoids completely the large computational expense involved in updating this matrix. Instead, the dispersion model is run N times to propagate the ensemble of model states to the next time step, using Monte Carlo random draw procedures to simulate model errors.

The solution converges towards the exact solution of the Extended Kalman filter as N increases. The ensemble size N needed in practical applications depends on the problem to be solved but typically ranges between 25 and 100, and in some cases up to of the order of 1000.

The EnKF-method has been used in numerous realistic applications for both ocean and atmosphere (Evensen, 2003; 2006), and it is able to handle strongly non-linear dynamics and large state spaces. This makes the method clearly applicable on urban and regional scales, where non-linearities in the form of photochemistry and/or aerosol chemistry might occur.

2.3.2 Advantages

- Much simpler update of the model error covariance matrix from one timestep to the next than in the extended Kalman filter.
- Grants flow dependency in the error covariance matrix.
- Correct setup of the ensemble perturbations allows estimation of the background error covariance matrix.
- Can handle non-linear state transition and observation models

2.3.3 Disadvantages

- Need to run the dispersion model N times rather than one, representing a considerable increased computational burden.
- The accuracy of the estimated model error covariance matrix depends on the representativeness of the simulated model errors
- Underlying assumptions are invalidated if the model errors are non-random or non-Gaussian.

2.3.4 Algorithm Availability

Core numerical aspects of the algorithm have been developed at NILU. However, the additional code for perturbing, launching, and updating the analysis and ensembles has not been implemented.

2.4 3D-Variational Data Assimilation

2.4.1 Technical Description

We can represent an air quality model using the following formulation

$$\boldsymbol{c}_{t+1} = \boldsymbol{F}_t^{AQ}(\boldsymbol{x}_t), t = 0, \dots, N-1, \boldsymbol{x}_{\mathrm{t}} = \boldsymbol{q}_t \qquad (9)$$

for the specific simplified case where the control variables, $\boldsymbol{x}_t$, only consist of the emission variables, $\boldsymbol{q}_t$. Note that in theory the control variables could be set to other input/model variables, but for reasons outlined above, we choose to simplify the problem to this specific case.

In the formulation of the 3D-Var algorithm (Lorenc, 1986) for an air quality model (Emili et al., 2016), the assimilated model state is calculated as $\boldsymbol{F}(\boldsymbol{x})$ that minimizes the following cost function:

$$J(\boldsymbol{x_t}) = \frac{1}{2}\left(\boldsymbol{F}(\boldsymbol{x}_t) - \boldsymbol{F}(\boldsymbol{x}_{a,t}\right)^T \boldsymbol{B}^{-1}\left(\boldsymbol{F}(\boldsymbol{x}_t) - \boldsymbol{F}(\boldsymbol{x}_{a,t})\right) + \frac{1}{2}(\boldsymbol{y} - F(\boldsymbol{x}_t))^T \boldsymbol{R}^{-1}(\boldsymbol{y} - F(\boldsymbol{x}_t)) \qquad (10)$$

with $\boldsymbol{F}(\boldsymbol{x}_a)$ being the background estimate of $\boldsymbol{F}(\boldsymbol{x})$ and $\boldsymbol{y}$ the observation vector. In practice, $\boldsymbol{x}_a$ is the a priori estimate of the pollutant concentration field given either by climatology or from a model simulation from the previous days. Pay special note to the fact that both the control variables and the model are evaluated at the same time step, t.

In 3D-Var, the objective function $J(\boldsymbol{x})$ is typically minimized using standard numerical methods such as Conjugate Gradient methods, Quasi-Newton methods, etc. A minimum value always exists since the objective function is strictly convex due to the matrices $\boldsymbol{B}$ and $\boldsymbol{R}$ being covariance matrices and thus positive definite.

The main difficulty in implementing the 3D-Var-method (as for the OI method) is to specify or estimate the model and observation error covariance matrices $\boldsymbol{B}$ and $\boldsymbol{R}$. These matrices determine how much weight will be put on the model forecasted state and observations, respectively. It is therefore important that these are defined as accurately as possible. The $\boldsymbol{B}$ matrix is usually defined using spatial correlation or covariance functions. The $\boldsymbol{R}$ matrix is usually defined as a diagonal matrix with the observation + representativeness error variances along the diagonal (assuming that these errors are independent), and where the representativeness errors are usually the most difficult to quantify appropriately.

2.4.2 Advantages

- Can operate easily with a non-linear observation operator, i.e. do not need any linearization of the observation model.

2.4.3 Disadvantages

- Difficult to specify or estimate the model error covariance matrix.
- Cannot be used to carry out emission inversion and therefore is unsuitable for short-lived atmospheric pollutants.

2.4.4 Algorithm Availability

The algorithm has not been tested at NILU and nor do we have ready access to an existing implementation.

2.5 Strong Constraint 4D-Variational Data Assimilation

2.5.1 Technical Description

The strong constraint 4D-Var algorithm (Klinker et al., 2000; Mahfouf & Rabier, 2000; Rabier et al., 2000) combines four components:

1. The forward model, in this case, an air quality model, which we can represent as in Eq. 9.

2. The adjoint of this air quality model, which defines the sensitivity of the current model state to changes in the control parameters. These adjoint sensitivities are used to construct the adjoint of the cost function.

3. A cost function that evaluates the observation minus model differences at time window, t, while considering the relative observation and model covariances. Note that the cost function evaluates the present model state using control variables from the model past, $\boldsymbol{x}_{t-n}$. Similarly to the 3D-var algorithm, in practice we need an additional operator, H, the observation operator must act upon $F(\mathbf{x})$ to transform the model state into observation space (not included in the notation here). Here we give an example as used in an air quality model context (Hamer et al., 2015)

$$J(\boldsymbol{x_t}) = \frac{1}{2}\Big(\big(\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{x}_t)\big)\Big)^T \boldsymbol{S}_n^{-1}\big(\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{x}_t)\big) + \frac{1}{2}\big(\boldsymbol{x}_{t-n} - \boldsymbol{x}_{a,t-n}\big)^T \boldsymbol{S}_a^{-1}(\boldsymbol{x}_{t-n} - \boldsymbol{x}_{a,t-n}) \quad (11)$$

4. An iterative gradient minimization algorithm setup to minimize cost function by optimizing $\boldsymbol{x}$ following

$$\hat{\boldsymbol{x}} = min_x \; J(\boldsymbol{x}) \; (12)$$

In practice, we formulate the minimization problem so that the adjoint sensitivity of the cost function is computed and used to find the minimum of $J(\boldsymbol{x})$ and find $\hat{\boldsymbol{x}}$

$$\nabla_{\boldsymbol{x}} J = \boldsymbol{K}^T \boldsymbol{S}_n^{-1}\big(\boldsymbol{y} - \boldsymbol{F}(\hat{\boldsymbol{x}})\big) - \boldsymbol{S}_a^{-1}(\hat{\boldsymbol{x}} - \boldsymbol{x}_a) = 0 \quad (13)$$

where K is the Jacobian matrix describing the forward model sensitivity to perturbations in the emission parameters, or, stated another way, the sensitivity of $\boldsymbol{c}_{t+1}$ to changes in $\boldsymbol{x}$. The strong constraint 4D-Var is formulated with the implicit assumption that errors in the control variable, $\boldsymbol{x}$, are the only source of errors in the estimation of the current and future states of $\boldsymbol{c}$.

2.5.2 Advantages

- Capable of carrying out emission inversion.
- Can solve non-linear problems via the iterative method.
- The assimilation system can cope with bias in the control variable errors.

2.5.3 Disadvantages

- Requires the model adjoint.
- Difficult to define the background errors and thus to create the background error covariance matrix.
- Cannot separate model and emission errors.

2.5.4 Algorithm Availability

A working example of the 4D-Var algorithm is available. This is the same algorithm that was used in (Hamer et al., 2015).

2.6 Weak Constraint 4D-Variational Data Assimilation

2.6.1 Technical Description

The weak constraint 4D-Var algorithm (Griffith & Nichols, 2000; Trémolet, 2007) has been specifically developed to address a key weakness of the strong constraint 4D-Var, i.e., the inability of that algorithm to account simultaneously for both model errors and other error sources beyond the selected control variables. This is a critical weakness in cases where there are multiple competing sources of errors and/or model errors. In such cases, the model errors would impinge on the algorithm's ability to provide accurate state estimation and parameter estimation.

The temporal evolution of the model errors can be formulated as (Griffith & Nichols, 2000)

$$\boldsymbol{e}_{t+1} = \boldsymbol{g}_t(\boldsymbol{c}_t, \boldsymbol{e}_t) \qquad (14)$$

where $\boldsymbol{e}$ is the model error at any one point in time, e is a vector describing different components of the systematic model error, and $\boldsymbol{g}_t$ is a function that evolves the error state forward in time. The model error can be represented in different ways, i.e., as a constant error, a temporally evolving error, or with a cyclic or spectral variability (Griffith & Nichols, 2000). The weak constraint 4D-Var accounts for the model error by redefining the model state equation following Eq. (4).

Accordingly, it redefines the cost function of the strong constraint 4D-Var in Eq. (11) to include an error term both for the model versus observation evaluation and as an additional penalty term (Trémolet, 2007)

$$J(\boldsymbol{x}_t) = \frac{1}{2}\big((\boldsymbol{y} - (\boldsymbol{F}(\boldsymbol{x}_t) + \boldsymbol{e}_t))\big)^T \boldsymbol{S}_n^{-1}(y - (\boldsymbol{F}(\boldsymbol{x}_t) + \boldsymbol{e}_t)) + \frac{1}{2}\big(\boldsymbol{x}_{t-n} - \boldsymbol{x}_{a,t-n}\big)^T \boldsymbol{S}_a^{-1}\big(\boldsymbol{x}_{t-n} - \boldsymbol{x}_{a,t-n}\big) + \frac{1}{2}\big(\boldsymbol{e}_{t-n} - \boldsymbol{e}_{a,t-n}\big)^T \boldsymbol{Q}^{-1}(\boldsymbol{e}_{t-n} - \boldsymbol{e}_{a,t-n}) \qquad (15)$$

where $\boldsymbol{Q}$ is the model error covariance matrix. In this case, the model error covariance matrix differs from the background error covariance matrix. The background error matrix specifically describes errors on the control variables of interest, i.e., emissions in this case.

The model error covariance describes the errors arising from intrinsic model errors and must be based on the statistics of the model behavior independent of other control parameters. With the introduction of the model error terms into the cost function (Eq. 15), the formulation of the adjoint sensitivity in Eq. 13 concerning $\boldsymbol{x}$ and $\boldsymbol{e}$ becomes

$$\nabla_{\mathbf{x}} J = \boldsymbol{K}^T \boldsymbol{S}_n^{-1}(\boldsymbol{y} - (\boldsymbol{F}(\hat{\boldsymbol{x}}) + \boldsymbol{e})) - \boldsymbol{S}_a^{-1}(\hat{\boldsymbol{x}} - \boldsymbol{x}_a) \quad (16)$$

and

$$\nabla_e J = \boldsymbol{G}^T \boldsymbol{S}_n^{-1}(\boldsymbol{y} - (\boldsymbol{F}(\hat{\boldsymbol{x}}) + \boldsymbol{e})) - \boldsymbol{Q}^{-1}(\boldsymbol{e} - \boldsymbol{e}_t^a) = 0 \qquad (17)$$

where ***G*** is the Jacobian matrix relating the sensitivity of the cost function to changes in ***e***. This means the optimization algorithm can now minimize the cost function either by adjusting ***x*** or ***e***, and the emphasis of one over the other is done according to the relative weighting of the $\boldsymbol{S}_a$ and ***Q*** matrices concerning $\mathbf{S}_n$. Note that in these formulations of the adjoint sensitivities, the model is represented in its error-augmented state following Eq. 5.

The novel aspects of the weak constraint 4D-Var algorithm bring with it the requirement to calculate or pre-determine new parameters. Most notably, the dynamically varying model error, the a priori estimate of that error, and the covariance matrix, ***Q***, describing the error between ***e*** and $\boldsymbol{e}_a$. Thus, this requires a good characterization of both the expected errors and the expected deviations from those a priori model error estimates. The challenge facing this algorithm in the context of an atmospheric chemical model is how to distinguish model error from errors resulting from emission errors.

2.6.2 Advantages

- Capable of carrying out emission inversion, of solving non-linear problems, and can cope with bias.
- Can account for model and emission errors separately in the assimilation solution. Model errors will therefore not propagate into the emission estimation solution

2.6.3 Disadvantages

- Requires the model adjoint.
- Requires enough knowledge of the model errors to be able to mathematically describe them. This could potentially be difficult to achieve.

2.6.4 Algorithm Availability

No working version of this algorithm is currently available. It would require adaptation of the 4D-Var algorithm employed in (Hamer et al., 2015).

2.7 4D-Ensemble Variational Data Assimilation

2.7.1 Technical Description

The 4D-EnVar algorithm (Emili et al., 2016; Lorenc et al., 2015) shares many similarities with the 3D-Var and 4D-Var algorithms but was developed to address key weaknesses of both. The 4D-EnVar algorithm builds upon the 3D-Var algorithm by expanding the dimensionality of the variables represented within the cost function from just the three spatial dimensions to also include the time dimension (Emili et al., 2016; Lorenc et al., 2015).

A key difference concerning the 4D-Var algorithm, however, is that there is no need to compute either tangent linear model or adjoint models. Furthermore, the 4D-EnVar uses an ensemble of perturbed simulations to construct the background error covariance matrix (Emili et al., 2016; Lorenc et al., 2015). In both the 3D-Var and 4D-Var algorithms it is typically difficult to correctly estimate the background error covariance matrix.

The 4D-EnVar cost function can be defined as

$$J(\boldsymbol{x_t}) = \frac{1}{2}(\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x_a})^T \boldsymbol{B_e}^{-1}(\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x_a})) + \frac{1}{2}(\boldsymbol{y} - F(\boldsymbol{x}))^T \boldsymbol{R}^{-1}(\boldsymbol{y} - F(\boldsymbol{x})) \quad (18)$$

where the $\boldsymbol{B_e}$ matrix is the model error covariance matrix that is estimated from an ensemble of perturbed model trajectories. The construction of $\boldsymbol{B_e}$ follows this general specification

$$\boldsymbol{B_e} = \frac{1}{L-1}(\boldsymbol{F}(\boldsymbol{x}'_1), \dots, \boldsymbol{F}(\boldsymbol{x}'_L))(\boldsymbol{F}(\boldsymbol{x}'_1), \dots, \boldsymbol{F}(\boldsymbol{x}'_L))^T \quad (19)$$

where L is the number of ensemble members. In practice, the solution to $J(\boldsymbol{x})$ is found via conjugate gradient or Quasi-Newton methods.

2.7.2 Advantages

- Can reconstruct the model error covariance matrix using the ensemble approach and therefore does not require specific knowledge of the model errors.
- Does not require specific knowledge of model errors.
- Can solve problems in the case where model and emission errors exist together.
- Can reduce bias errors arising from systematic biases in model inputs, e.g., emissions.
- No need to determine the adjoint or tangent linear models.

2.7.3 Disadvantages

- Complex algorithm.
- The technical point relevant to NILU: the current libraries that exist from ECMWF/CERFACS are written in C++.
- Cannot carry out emission inversion.

2.7.4 Algorithm Availability

A working version of this algorithm used in an atmospheric chemistry model is available as a library from ECMWF that we already have access to. This is the algorithm used in Emili et al., (2016).

3 Discussion and Selection of Appropriate Data Assimilation Methods

We must first define the criteria with which we will select the appropriate data assimilation algorithms for use in an in urban air quality modelling context. We can then evaluate each algorithm and arrive at recommendations for which algorithms to test and develop. The criteria are:

- Can it achieve the objectives we have highlighted, i.e., the dual aims of improving the estimation of pollutant concentrations and of pollutant emissions?
- Can the method cope with biases in either the model error or emissions inputs?
- Is it an appropriate method for problems where model errors exist?
- Can it account for model error?
- Feasibility of implementation: this includes the availability of the algorithm code, the complexity of the code, experience of using the algorithm, etc. Is it feasible to use this algorithm?

To help identify the algorithms that achieve the most criteria, we present in Table 1 summarising and ranking of each algorithm against each of the criteria laid out above.

Algorithm	Control Variable Objective		Are model biases problematic?	Is model error problematic?	Estimation and accounting for model error?	Feasibility
	Concentrations of long-lived pollutants	Emission estimation				
Optimal Interpolation	Yes	No	Yes	Yes	No	Very feasible
EKF	Yes	No	Yes	Yes	No	Less feasible
EnKF	Yes	Qualified Yes	Yes	Qualified Yes	Qualified Yes	Feasible
3D-Var	Yes	No	Yes	Yes	No	Less feasible
Strong constraint 4D-Var	Yes	Yes	No	Yes	No	Feasible
Weak constraint 4D-Var	Yes	Yes	No	No	Yes	The most difficult
4D-EnVar	Yes	Qualified No	No	No	Yes	Feasible but complex

Table 1 Showing summarizing the capabilities, advantages, and disadvantages of each method. The objectives list the ability of each algorithm to assimilate with either long-lived pollutant concentrations or emissions as control variables. The assimilation of short-lived pollutant concentrations is ignored. The colour coding from dark green, to green, then, yellow, orange, red denote the algorithms' decreasing suitability.

We will quickly discuss the qualified responses for the EnKF algorithm. The EnKF can account for model errors and estimate emissions by using model parameter perturbations under specific conditions to estimate background error. These conditions are that the model errors or errors on the emissions have the properties of being random, with zero bias, and with a Gaussian distribution. In cases where the model and emissions errors have these properties, this algorithm can represent and resolve the model error and estimate emissions, respectively. However, emission estimates are typically biased high or low, and air quality model errors are typically then dominated by large persistent bias terms (Marécal et al., 2015). Thus, the EnKF becomes an invalid tool due to the common nature of the emission and model errors present.

The 4D-EnVar algorithm receives a qualified no for being able to estimate emissions. While it cannot be used for emission parameter estimation, it can still estimate concentrations of short-lived compounds even when there are bias terms present in their emissions. This is because the algorithm can estimate and correct for bias terms from all types of model errors including errors in emissions.

We will now outline the applicability of each algorithm for different use cases. We start with the relatively simple examples and then step up in complexity.

1. The control variables are the concentrations of long-lived pollutants where model and emission errors are random with no bias. In this special case, errors in emissions can be ignored because assimilation increments will have a long lifetime, and we avoid the situation where increments decay due to a pollutants' loss. All algorithms can be used in this case and feasibility alone may be sufficient to select a suitable algorithm.

2. The control variables are the emissions of short-lived pollutants where the emission errors are random with no bias and where other model errors can be ignored. Here, emission errors cannot be ignored since increments to the concentrations alone will decay rapidly. Thus, the algorithm must be able to update the emissions. Here the EnKF, strong and weak 4D-Var, and 4D-EnVar can all be used to optimize pollutant concentrations through estimation of pollutant emissions.

3. The control variables are the emissions of short-lived pollutants where the emission errors are random or non-random but with bias and where other model errors can be ignored. In this case, the EnKF no longer becomes an appropriate tool since its fundamental assumptions rely on errors being random with a Gaussian distribution. The strong/weak constraint 4D-Var algorithms can be used in this case.

4. The control variables are the emissions of short-lived pollutants where the emission errors are non-random with bias and where other model errors cannot be ignored. Systematic model errors arising from temporal and spatial discretization, errors in model parameterizations, and input meteorology compete with errors in emissions. Furthermore, such errors are probably correlated in time and thus will create bias terms that could appear as emission errors. In this case, the strong constraint 4D-Var algorithm will encounter problems and the weak constraint 4D-Var algorithm is the only algorithm fully suitable for this situation.

5. The control variables are the concentrations of short-lived pollutants where the emission errors are non-random with bias and where other model errors cannot be ignored. In this case, the 4D-EnVar algorithm in its current form cannot explicitly estimate emissions, but it can be used to estimate the concentrations of short-lived species in the presence of both emission and model bias errors.

It is worth mentioning a special case for the offline assimilation of short-lived pollutants in the presence of emission errors. Here 'offline' means assimilation carried out as a post-processing step after the model has been run as in (Schneider et al., 2017). Thus, in this case, there is no flow dependency of increments: increments are not evolved forward in time and therefore cannot decay. Thus, methods such as optimal interpolation, 3D-Var, and the EKF could be used in this way to avoid problems associated with being unable to estimate emissions. The disadvantage of this approach is that the lack of flow dependency prevents increments from being transported by model dynamics, which means that observations can only update the model field in their immediate spatial vicinity. For this offline assimilation, the feasibility should be the leading criteria, and therefore the optimal interpolation algorithm would be the favored choice.

For the case of emission estimation, the most readily available tool that can cope with bias terms in the emissions is the strong constraint 4D-Var. After that, the weak constraint 4D-Var becomes the only tool that can estimate emissions in the presence of model error.

For the estimation of short-lived pollutant concentrations in the presence of model error, the 4D-EnVar is the most readily available and appropriate algorithm since we already have access to the ECMWF OOPS library. The weak constraint 4D-Var could be considered as a backup option to the 4D-EnVar for this case.

A further general comment is that if we assimilate observations on the grid, we either need to carefully select urban background stations, or we need to adjust for representativity errors. Otherwise, assimilation on the model receptor points should be considered.

References

Chiles, J.-P., & Delfiner, P. (2009). Geostatistics: modeling spatial uncertainty (Vol. 497). John Wiley & Sons.

Cressie, N. (1990). The origins of kriging. Mathematical Geology, 22(3), 239–252.

Desroziers, G., Arbogast, E., & Berre, L. (2016). Improving spatial localization in 4DEnVar. Quarterly Journal of the Royal Meteorological Society, 142(701), 3171–3185.

Emili, E., Gürol, S., & Cariolle, D. (2016). Accounting for model error in air quality forecasts: An application of 4DEnVar to the assimilation of atmospheric composition using QG-Chem 1.0. Geoscientific Model Development, 9(11), 3933–3959.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research: Oceans, 99(C5), 10143–10162.

Gandin, L. S. (1963). Objective Analysis of Meteorological Fields. translated from Russian in 1965 by Israel Program for Scientific Translations. Jerusalem, 242pp.

Goovaerts, P., & others. (1997). Geostatistics for natural resources evaluation. Oxford University Press on Demand.

Griffith, A. K., & Nichols, N. K. (2000). Adjoint methods in data assimilation for estimating model error. Flow, Turbulence and Combustion, 65(3–4), 469–488.

Hamer, P. D., Bowman, K. W., Henze, D. K., Attié, J.-L., & Marécal, V. (2015). The impact of observing characteristics on the ability to predict ozone under varying polluted photochemical regimes. Atmospheric Chemistry and Physics, 15(18), 10645–10667.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45.

Kalnay, E. (2003). Atmospheric modeling, data assimilation and predictability. Cambridge university press.

Klinker, E., Rabier, F., Kelly, G., & Mahfouf, J.-F. (2000). The ECMWF operational implementation of four-dimensional variational assimilation. III: Experimental results and diagnostics with operational configuration. Quarterly Journal of the Royal Meteorological Society, 126(564), 1191–1215.

Knudsen, H. P., & Lefohn, A. S. (1988). The use of geostatistics to characterize regional ozone exposures. In Assessment of crop loss from air pollutants (pp. 91–105). Springer.

Lorenc, A. C. (1986). Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society, 112(474), 1177–1194.

Lorenc, A. C., Bowler, N. E., Clayton, A. M., Pring, S. R., & Fairbairn, D. (2015). Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. Monthly Weather Review, 143(1), 212–229.

Mahfouf, J.-F., & Rabier, F. (2000). The ECMWF operational implementation of four-dimensional variational assimilation. II: Experimental results with improved physics. Quarterly Journal of the Royal Meteorological Society, 126(564), 1171–1190.

Marécal, V., Peuch, V. H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Van Der Gon, H. A. C. D., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., … Ung, A. (2015). A regional air quality forecasting system over Europe: The MACC-II daily ensemble production. Geoscientific Model Development, 8(9), 2777–2813.

Matheron, G. (1963). Principles of geostatistics. Economic Geology, 58(8), 1246–1266.

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F., & Simmons, A. (2000). The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. Quarterly Journal of the Royal Meteorological Society, 126(564), 1143–1170.

Schneider, P., Castell, N., Vogt, M., Dauge, F. R., Lahoz, W. A., & Bartonova, A. (2017). Mapping urban air quality in near real-time using observations from low-cost sensors and model information. Environment International, 106, 234–247.

Shiryayev A.N. (1992). Interpolation and extrapolation of stationary random sequences. In: Shiryayev A.N. (Ed.), *Selected Works of A. N. Kolmogorov*. Springer, Dordrecht.

Talagrand, O. (1998). A posteriori evaluation and verification of analysis and assimilation algorithms. Proc. of ECMWF Workshop on Diagnosis of Data Assimilation System, 17–28.

Trémolet, Y. (2007). Model-error estimation in 4D-Var. Quarterly Journal of the Royal Meteorological Society: A Journal of the Atmospheric Sciences, Applied Meteorology and Physical Oceanography, 133(626), 1267–1280.

NILU – Norwegian Institute for Air Research

NILU – Norwegian Institute for Air Research is an independent, non-profit institution established in 1969. Through its research NILU increases the understanding of climate change, of the composition of the atmosphere, of air quality and of hazardous substances. Based on its research, NILU markets integrated services and products within analysing, monitoring and consulting. NILU is concerned with increasing public awareness about climate change and environmental pollution.

NILU's values: Integrity - Competence - Benefit to society
NILU's vision: Research for a clean atmosphere

NILU – Norwegian Institute for Air Research
P.O. Box 100, NO-2027 KJELLER, Norway

E-mail: nilu@nilu.no
http://www.nilu.no

ISBN: 978-82-425-3027-1
ISSN: 2464-3327

Norsk institutt for luftforskning
Norwegian Institute for Air Research