A Comparison Study of Data Assimilation Algorithms for Ozone Forecasts

CEREA report 2008-6

INRIA-Rocquencourt/ENPC-CEREA, France
Lin Wu, Vivien Mallet, Marc Bocquet, and Bruno Sportisse

{lin.wu,vivien.mallet}@inria.fr;{marc.bocquet,bruno.sportisse}@cerea.enpc.fr

Abstract

The objective of this report is to evaluate the performances of different data assimilation schemes with the aim of designing suitable assimilation algorithms for short-range ozone forecasts in realistic applications. The underlying atmospheric chemistry-transport models are stiff but stable systems with high uncertainties (e.g., over 20% for ozone daily peaks, Hanna et al. [1998]; Mallet and Sportisse [2006b], and much more for other pollutants like aerosols). Therefore the main difficulty of the ozone data assimilation problem is how to account for the strong model uncertainties. In this report, the model uncertainties are either parameterized with homogeneous error correlations of the model state or estimated by perturbing some sources of the uncertainties, e.g. the model uncertain parameters. Four assimilation methods have been considered, namely optimal interpolation, reduced-rank square root Kalman filter, ensemble Kalman filter, and four-dimensional variational assimilation. These assimilation algorithms are compared under the same experimental settings. It is found that the assimilations significantly improve the one-day ozone forecasts. The comparison results reveal the limitations and the potentials of each assimilation algorithm. In our four-dimensional variational method, the low dependency of model simulations on initial conditions leads to moderate performances. In our sequential methods, the optimal interpolation algorithm has the best performance during assimilation periods. Our ensemble Kalman filter algorithm perturbs the uncertain parameters to approximate model uncertainties and has better forecasts than the optimal interpolation algorithm during prediction periods. This could partially be explained by the low dependency on the uncertainties in initial conditions. The sensitivity analysis on the algorithmic parameters is also conducted for the design of suitable assimilation algorithms for ozone forecasts.

1 Introduction

A typical Eulerian atmospheric chemistry-transport model (CTM) computes the concentrations c of a set of chemical species by solving the system of advection-diffusion-reaction equations:

$$\frac{\partial c_i}{\partial t} = -\mathrm{div}(\mathbf{V}c_i) + \mathrm{div}(\rho \mathbf{D} \nabla \frac{c_i}{\rho}) + \chi_i(c,t) + E_i \ , \quad \forall i, \tag{1}$$

where c_i is the concentration of the i-th species, $\mathbf{V}$ is the wind velocity, ρ is the air density, $\mathbf{D}$ is the turbulent diffusion matrix, $\chi_i(c,t)$ stands for the species production and loss due to the chemical reactions, and E_i stands for the elevated emissions. At ground the boundary condition is given by

$$-\rho \mathbf{D} \nabla \frac{c_i}{\rho} \cdot \mathbf{n} = S_i - v_i^{\mathrm{dep}} c_i \ , \tag{2}$$

where $\mathbf{n}$ is the upward unitary vector, S_i stands for the surface emissions and v_i^{dep} is the dry deposition velocity.

In the numerical model (the CTM), the dimension of the discretized system is usually 10^6–10^7. The model computes ozone hourly concentrations over Europe (for instance) given the initial conditions and the input data (also designated herein as parameters).

Data assimilation can be considered as the determination of the initial conditions or of model uncertain parameters by coupling the heterogeneous available information, e.g. model simulations, observations, and statistics for errors. Data assimilation methods are roughly catalogued into variational and sequential ones [Le Dimet and Talagrand, 1986; Evensen, 1994]. The objective of the former can be defined as state estimation by minimizing the quadratic discrepancy between model simulation and a block of observations, usually combined with a priori background knowledge. This can be formalized and solved efficiently with the optimal control theory. The sequential methods make use of observations as soon as they are available. Since this is a filtering process, filter theory (linear or nonlinear) applies.

Both methods have found their applications for CTMs. The pioneering work dates back to Fisher and Lary [1995]. On the variational side, Elbern and Schmidt [2001] use a comprehensive model rather than an academical model in order to assimilate real observations with assessment of ozone forecast. Chai et al. [2007] follow with assimilation of new types of observations, and several practical issues, e.g. background error modeling, are investigated with details. Very few work deals with the assimilation of initial conditions jointly with uncertain parameters [Elbern et al., 2007]. By contrast, Segers [2002] conducts in-depth studies on the applications of efficient filtering methods, in which emissions, photolysis rates and deposition are considered to be uncertain. The model state as well as uncertain parameters are estimated. Constantinescu et al. [2007b] report the filtering results obtained with perturbations on emissions and on boundary conditions, and with distance constraints on the spatial correlations.

All these efforts are part of the recent diffusion of data assimilation expertise from numerical weather prediction (NWP) to air quality community. For a review see Carmichael et al. [2008]. The CTMs are stiff but stable systems; the perturbations on initial conditions tend to be smoothed out rather than amplified. Therefore the conclusions from meteorological experiences [Lorenc, 2003; Kalnay et al., 2007] cannot be applied directly.

The objective of this report is to evaluate different assimilation algorithms for ozone forecasts in the same experimental settings. Hopefully this could serve as a base point for the design of assimilation algorithms suitable for ozone forecasts in realistic applications. Four algorithms, namely optimal interpolation (OI), ensemble Kalman filter (EnKF), reduced-rank square root Kalman filter (RRSQRT) and four-dimensional variational assimilation (4DVar) were implemented.

We note that this comparison study has its limitations in that: i) Only model state is adjusted and uncertain model parameters remain unchanged. ii) The treatment of uncertainties are different. OI parameterizes aggregate uncertainties using the homogeneous Balgovind correlation function. In 4DVar the uncertainties are taken into account, in a way similar to OI (Balgovind correlation), but only at the initial date of the assimilation period. The underlying model is assumed to be perfect, that is, we consider a strongly constrained 4DVar. By contrast, EnKF and RRSQRT represent model uncertainties with ensemble generated by Monte Carlo samplings of uncertain parameters. The reasons for the first limitation are that (1) the adjoint model with respect to model parameters is not available, and (2) correlations between the model state and parameters are unknown. Clearly this should be a research task in near future. The second limitation stems from the unsettled formulation of model error. A novelty of our EnKF and RRSQRT implementation is the perturbation method, originally employed in uncertainty studies for air quality models [Hanna et al., 2001].

The report is organized as follows. Section 2 documents the assimilation algorithms and their implementations. The experiment setup concerning the model and observations is detailed in Section 3. We report the comparison results in Section 4. Therein sensitivity studies with respect to the assimilation algorithm settings are also conducted. Conclusions and discussions can be found in Section 5.

2 Assimilation Algorithms

We rewrite the CTM dynamical differential equation in discrete form from time t_{k-1} to t_k,

$$\mathbf{x}^t(t_k) = M_{k-1}\left(\mathbf{x}^t(t_{k-1})\right) + \epsilon^f(t_{k-1}) \,, \tag{3}$$

where $\mathbf{x}^t$ denotes the true state vector of dimension n, M_{k-1} corresponds to the (nonlinear) dynamical operator from $k-1$ to k, and ϵ^f is the model error vector assumed to have a normal distribution with zero mean and covariance matrix $\mathbf{Q}$. In this study, the state is chosen to be the vector of concentrations for the concerned species. For simplicity we drop t_k to subindex k, e.g. $\mathbf{x}^t_k$ for $\mathbf{x}^t(t_k)$ and $\mathbf{Q}_{k-1}$ for $\mathbf{Q}(t_{k-1})$. At each time t_k, one observes,

$$\mathbf{y}_k = H_k\left(\mathbf{x}^t_k\right) + \epsilon^o_k \,, \tag{4}$$

where ϵ^o_k is the observation error vector assumed to have a normal distribution with zero mean and covariance matrix $\mathbf{R}$, and H_k is the (possibly nonlinear) operator that maps the state to the observation space at time t_k. The vector $\mathbf{y}_k$ is of size p, and usually $p \ll n$. The error vectors ϵ^f_{k-1} and ϵ^o_k are supposed to be independent.

Let $\mathbf{x}^b$ be the a priori state estimation (*background*) with error $\epsilon^b = \mathbf{x}^b - \mathbf{x}^t$ of zero mean and covariance matrix $\mathbf{B}$, and let $\mathbf{x}^a$ be the posterior state estimation (*analysis*) with error $\epsilon^a = \mathbf{x}^a - \mathbf{x}^t$ of covariance matrix $\mathbf{A}$. The data assimilation problem is to determine the optimal analysis $\mathbf{x}^a$ and its statistics $\mathbf{A}$ given background $\mathbf{x}^b$, observation $\mathbf{y}$, and the statistical information in error covariance items $\mathbf{R}$ and $\mathbf{B}$.

2.1 Optimal Interpolation

Optimal interpolation [Daley, 1991] searches for an optimal linear combination between background and *innovation* by minimizing the state-estimation variance. The innovation d is the difference between the observation vector and the state vector, i.e. $d = \mathbf{y} - H(\mathbf{x}^b)$. Under linearity assumption of the observation operator close to the background, i.e. $H(\mathbf{x}) - H(\mathbf{x}^b) = \mathbf{H}(\mathbf{x} - \mathbf{x}^b)$

where $\mathbf{H}$ is the linearized operator, the estimation formulae are given according to best linear unbiased estimator theory as follows,

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - H(\mathbf{x}^b)) \ , \tag{5}$$

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \ . \tag{6}$$

In practice, setting the background error covariance remains problematic. In this study $\mathbf{B}$ is either diagonal or in Balgovind form. In the latter case, the error covariance between two points is given by

$$f(d) = \left(1 + \frac{d}{L}\right) e^{-\frac{d}{L}} v \ , \tag{7}$$

where L is a characteristic length, d is the distance between the two points, and v is the a priori variance [Balgovind et al., 1983].

2.2 Ensemble Kalman Filter

Ensemble Kalman filter [Evensen, 1994, 2003] differs from optimal interpolation in that the error covariance matrix is time-dependent. The assimilation process follows the cycling of two steps of forecast and analysis. At forecast step, the model is applied to the $r-$member ensemble $\{\mathbf{x}_{k-1}^{a,(i)}, i = 1, \ldots, r\}$, and produces the forecast $\{\mathbf{x}_k^{f,(i)}, i = 1, \ldots, r\}$. The forecast error covariance matrix $\mathbf{P}^f$ can be approximated by the ensemble statistics. Whenever observations are available, the cycling enters into analysis step, and each ensemble member $\mathbf{x}_k^{f,(i)}$ is updated to $\mathbf{x}_k^{a,(i)}$ according to the OI formula (5-6), with background error covariance matrix $\mathbf{B}$ replaced by forecast error covariance matrix $\mathbf{P}^f$. Although not necessary in the algorithm, the analysis error covariance matrix $\mathbf{P}^a$ can then be approximated with the analyzed-ensemble statistics.

We summarize the algorithm as follows,

- Initialization: given the probability density function (PDF) of the initial concentrations, an ensemble of initial conditions is generated. In our experiments, except for the cycling context in Section 4.4 where initial ensemble members are ensemble forecasts from the previous cycle, we skip this step: all members in the ensemble start with the same initial condition. The first integration steps are therefore a spin-up period during which the ensemble spread is essentially increasing as a result of the perturbations on uncertain parameters.

- Forecast step:

$$\mathbf{x}_k^{f,(i)} = M_{k-1}\left(\mathbf{x}_{k-1}^{a,(i)}\right) + \epsilon_{k-1}^{f,(i)} \ , \tag{8}$$

$$\mathbf{P}_k^f = \frac{1}{r-1}\sum_{i=1}^{r}\left(\mathbf{x}_k^{f,(i)} - \bar{\mathbf{x}}_k^f\right)\left(\mathbf{x}_k^{f,(i)} - \bar{\mathbf{x}}_k^f\right)^T \ , \tag{9}$$

  where $\bar{\mathbf{x}}_k^f$ is the mean of the forecast ensemble: $\bar{\mathbf{x}}_k^f = \frac{1}{r}\sum_{i=1}^{r}\mathbf{x}_k^{f,(i)}$.

- Analysis formula:

$$\mathbf{x}_k^{a,(i)} = \mathbf{x}_k^{f,(i)} + \mathbf{K}_k\left(\mathbf{y}_k^{(i)} - H_k\left(\mathbf{x}_k^{f,(i)}\right)\right) \ , \tag{10}$$

$$\mathbf{P}_k^a = \frac{1}{r-1}\sum_{i=1}^{r}\left(\mathbf{x}_k^{a,(i)} - \bar{\mathbf{x}}_k^a\right)\left(\mathbf{x}_k^{a,(i)} - \bar{\mathbf{x}}_k^a\right)^T \ , \tag{11}$$

where $\bar{\mathbf{x}}_k^a$ is the mean of analysis ensemble $\{\mathbf{x}_k^{a,(i)}, i = 1, .., r\}$, $\mathbf{y}_k^{(i)}$ is the observation vector, and the Kalman gain is approximated by

$$\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k^T \left(\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^T + \mathbf{R}_k \right)^{-1} . \tag{12}$$

The ensemble initialization and the determination of the model error $\epsilon_{k-1}^{f,(i)}$ are entangled problems. In our implementation, we take identical initial samples, and the model error is approximated by perturbing model input data and model parameters:

$$\epsilon_{k-1}^{f,(i)} \simeq M_{k-1} \left(\mathbf{x}_{k-1}^{a,(i)}, \mathbf{w}^{(i)} \mathbf{d} \right) - M_{k-1} \left(\mathbf{x}_{k-1}^{a,(i)}, \mathbf{d} \right) , \tag{13}$$

where $\mathbf{d}$ is the vector of parameters to be perturbed, and for $i-$th sample, $\mathbf{w}^{(i)}$ is the diagonal matrix whose diagonal elements are perturbation coefficients (see Section 2.5). Let $\mathbf{e}_k^{f,(i)}$ be $\mathbf{x}_k^{f,(i)} - \bar{\mathbf{x}}_k^f$, one (approximate) direction of the forecast error, and let $\mathbf{E}_k^f$ be the matrix $[\mathbf{e}_k^{f,(1)} \mathbf{e}_k^{f,(2)} \dots \mathbf{e}_k^{f,(r)}]$. By formula (9), we have

$$\mathbf{P}_k^f = \frac{1}{r-1} \mathbf{E} \mathbf{E}^T . \tag{14}$$

In this way, the error covariance matrix is approximated by ensemble statistics.

In the original EnKF algorithm, the observation vector $\mathbf{y}_k^{(i)}$ is perturbed for consistent analysis statistics [Burgers et al., 1998]. In this report, we present only the assimilation results without observation perturbations, since the variances of observation errors are in general much smaller than those of model errors. However, in our implementation, the observation perturbation is an option, and preliminary tests showed that, at least for the reference setting in Section 3, there are improvements in forecast performance with this option on.

2.3 Reduced-Rank Square Root Kalman Filter

Reduced-rank square root Kalman filter [Heemink et al., 2001] uses a low-rank representation $\mathbf{L}\mathbf{L}^T$ of error covariances matrix $\mathbf{P}$. $\mathbf{L} = [\mathbf{l}^1, \dots, \mathbf{l}^q]$ is the mode matrix whose columns (modes) are the dominant directions of the forecast error. The evolution of a mode can be approximated by the differences of the forecasts based on the mean (analyzed) state and its perturbation by this mode, that is,

$$\mathbf{l}_k^{f,(i)} = M_{k-1} \left(\mathbf{x}_{k-1}^a + \mathbf{l}_{k-1}^{a,(i)} \right) - M_{k-1} \left(\mathbf{x}_{k-1}^a \right) , \tag{15}$$

where $\mathbf{x}_{k-1}^a$ is given by

$$\mathbf{x}_{k-1}^a = \mathbf{x}_{k-1}^f + \mathbf{K}_{k-1} \left(\mathbf{y}_{k-1} - H_{k-1} \left(\mathbf{x}_{k-1}^f \right) \right) . \tag{16}$$

The forecast $\mathbf{x}_{k-1}^f$ is calculated from previous analyzed state by

$$\mathbf{x}_{k-1}^f = M_{k-2} \left(\mathbf{x}_{k-2}^a \right) . \tag{17}$$

Assuming that at time t_{k-1} the error covariance $\mathbf{P}_{k-1}^a$ has the square root form $\mathbf{L}_{k-1}^a \mathbf{L}_{k-1}^{a,T}$, the propagation of $\mathbf{P}_{k-1}^a$ is tractable. The forecast error covariance matrix at time t_k is calculated by

$$\begin{aligned} \mathbf{P}_k^f &= \mathbf{M_{k-1}} \mathbf{P}_{k-1}^a \mathbf{M_{k-1}}^T + \mathbf{Q}_{k-1} \\ &= \left[\mathbf{M_{k-1}} \mathbf{L}_{k-1}^a \quad \mathbf{Q}_{k-1}^{\frac{1}{2}} \right] \left[\mathbf{M_{k-1}} \mathbf{L}_{k-1}^a \quad \mathbf{Q}_{k-1}^{\frac{1}{2}} \right]^T , \end{aligned} \tag{18}$$

where $\mathbf{M_{k-1}}$ is the tangent linear model, that is the Jacobian matrix of M_{k-1}, $\mathbf{Q}_{k-1}^{\frac{1}{2}}$ is the square root of model error covariance matrix. Considering square root form $\mathbf{L}_k^f \mathbf{L}_k^{f,T}$ for $\mathbf{P}_k^f$, we have the forecast formula for mode matrix $\mathbf{L}^f$:

$$\tilde{\mathbf{L}}_k^f = [\mathbf{M_{k-1}} \mathbf{L}_{k-1}^a \quad \mathbf{Q}_{k-1}^{\frac{1}{2}}], \qquad \mathbf{L}_k^f = \Pi_k^f \tilde{\mathbf{L}}_k^f \,, \tag{19}$$

where Π_k^f projects $\tilde{\mathbf{L}}_k^f$ onto the q leading eigenvectors of $\tilde{\mathbf{L}}_k^f \tilde{\mathbf{L}}_k^{f,T}$ using the singular value decomposition. Recall that analysis error covariance matrix $\mathbf{P}^a$ can be calculated by $(\mathbf{I}-\mathbf{KH})\mathbf{P}^f(\mathbf{I}-\mathbf{KH})^T+\mathbf{KRK}^T$ for arbitrary gain $\mathbf{K}$, then rewriting it in square root form we obtain the analysis formula for mode matrix $\mathbf{L}^a$:

$$\tilde{\mathbf{L}}_k^a = [(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{L}_k^f \quad \mathbf{K}_k \mathbf{R}_k^{\frac{1}{2}}], \qquad \mathbf{L}_k^a = \Pi_k^a \tilde{\mathbf{L}}_k^a \,, \tag{20}$$

where Π_k^a projects $\tilde{\mathbf{L}}_k^a$ onto the q leading eigenvectors of $\tilde{\mathbf{L}}_k^a \tilde{\mathbf{L}}_k^{a,T}$.

We do not use the tangent linear model, but employ (15) to simulate $\mathbf{M_{k-1}} \mathbf{L}_{k-1}^a$ at forecast step. The columns of $\mathbf{Q}_{k-1}^{\frac{1}{2}}$ are obtained in the same manner as the model error formula (13) in EnKF. The above treatments make the RRSQRT implementation similar to our variant of EnKF. The difference is that RRSQRT employs square root formulae. In addition, the error covariance is approximated in dominant eigenvectors in RRSQRT whereas EnKF bears no such process.

2.4 Four-Dimensional Variational Algorithm

Four-Dimensional Variational Algorithm [Le Dimet and Talagrand, 1986] finds the optimal initial condition $\mathbf{x}^*$ by minimizing a cost function:

$$\begin{aligned} J(\mathbf{x}) = & \underbrace{\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)}_{J_b} + \\ & \underbrace{\frac{1}{2}\sum_{k=0}^{N} \left(\mathbf{y}_k - H_k(\mathbf{x}_k)\right)^T \mathbf{R}_k^{-1} \left(\mathbf{y}_k - H_k(\mathbf{x}_k)\right)}_{J_o} \end{aligned} \tag{21}$$

under the constraint $\mathbf{x}_k = M_{0\to k}(\mathbf{x}) = M_{k-1}\left(M_{k-2}\left(\ldots M_1\left(M_0(\mathbf{x})\right)\ldots\right)\right)$. The assimilation period is from t_0 to t_N. The gradient for J_o is calculated by the backward integration of the adjoint model [Bouttier and Courtier, 1999]:

- $\tilde{\mathbf{x}}_N = 0$,
- For $k = N, \ldots, 1$, calculates $\tilde{\mathbf{x}}_{k-1} = \mathbf{M}_{k-1}^T \left(\tilde{\mathbf{x}}_k - \mathbf{H}_k^T d_k\right)$, where $d_k = \mathbf{R}_k^{-1}\left(\mathbf{y}_k - H_k(\mathbf{x}_k)\right)$,
- $\tilde{\mathbf{x}}_0 := \tilde{\mathbf{x}}_0 - \mathbf{H}_0^T(d_0)$ gives the gradient of J_o with respect to $\mathbf{x}$.

Assimilations are performed by model integrations starting from the optimal initial condition $\mathbf{x}^*$. Further integrations from time step N based on the analyzed model state provide the predictions. The inverse of $\mathbf{B}$ is calculated online or, for high dimensional model configurations, $\mathbf{B}^{-1}$ can also be approximated using SVD truncations and saved on disk for later computations. The adjoint operator $\mathbf{M}_{k-1}^T$ is obtained using the automatic differentiation software O∂yssée [Faure and Papegay, 1998]. The forward model simulations are saved for the backward integrations of the adjoint model. No checkpointing technique is employed in our implementation.

2.5 Uncertainties and Model Error

The corrections of the analysis scheme lie in the subspace spanned by covariance matrix of forecast or background errors, i.e., the space induced by the columns of the square root of the matrix $\mathbf{B}$ [Kalnay, 2003]. Unrealistic error structure produces spurious corrections, and probably results in unbalanced physical model state. Therefore the design of the error structure is of great importance.

There are mainly three approaches for error modeling: i) modeling uncertain sources and then perturbing them in the model [Segers, 2002; Constantinescu et al., 2007a]; ii) using the statistics of model states, e.g. NMC method [Chai et al., 2007] and ensemble methods; iii) using parameterizations, e.g. Balgovind correlations for background error covariance [Hoelzemann et al., 2001; Elbern et al., 2007]. Each of the three should be *flow-dependent*, that is, adapting to the "error of the day". The spatial and temporal heterogeneities of the chemistry-transport problem make the last two approaches difficult issues.

The numerical models are usually assumed unbiased. In our case, we assume that the model uncertainties only result from the misspecification of model parameters. In our EnKF and RRSQRT implementation, the ensemble is generated by the model integrations with perturbed parameters. The uncertainties and the distributions are introduced for model parameters that are mainly bidimensional or tridimensional fields under space coordinates. These parameters are modeled as random vectors. In practice, for a field $\hat{\mathbf{p}}$ (a random vector) whose median value is $\mathbf{p}$, a perturbation is applied to the whole field so that every component $\hat{\mathbf{p}}_k^i$ has the prescribed distribution. For instance, for a lognormal distribution, one writes

$$\hat{\mathbf{p}}_k^i = \mathbf{p}_k^i \times \sqrt{\alpha}^{\gamma} \ , \quad \forall k, i \tag{22}$$

where γ is sampled according to a standard normal distribution. The quantity γ is independent of the time index k and of the space index i, so that the perturbations increase the ensemble spread. The same sample of γ is used to perturb all values of the field $\hat{\mathbf{p}}$. The quantity $\sqrt{\alpha}^{\gamma}$ is the perturbation coefficient for the corresponding parameter in matrix $\mathbf{w}^{(i)}$ in formula (13). For normal distributions, perturbations bigger than certain given quantity (by default two times of the standard deviation) are discarded so that no unrealistic parameters are produced, for instance, the negative emissions.

A delicate point is having temporal and spatial correlations between the different values of the field. With the perturbation applied in (22), the correlation between two field values $\ln \hat{\mathbf{p}}_k^i$ and $\ln \hat{\mathbf{p}}_l^j$ is equal to 1. But the uncertainty sources at these two points are not the same; hence the perturbation should be different. A fine modeling of the uncertainty should lead to have γ depending on time and position (producing γ_k^i). Such a fine description of uncertainties is mostly beyond available knowledge.

Examples for continental air quality simulations extracted from Hanna et al. [2001] are shown in Table 1. For many fields (associated with $\alpha = 2$), a confidence interval that includes 95% of the probability density integral is $[\frac{m}{2}, 2m]$ if m is the mean of the field. Uncertainty levels must be adjusted to the simulation scale (domain size and temporal discretization). In particular, uncertainties decrease as data is averaged over a larger domain or over a longer period of time.

3 Experiment Setup

3.1 Model and Input Data

The ozone forecasts and the assimilation experiments are performed in the framework of the air quality modeling system Polyphemus [Mallet et al., 2007] whose version 1.2 includes all

Field	Distribution	Uncertainty
Top ozone boundary conditions	log-normal	$\alpha = 1.5$
Top NO_x boundary conditions	log-normal	$\alpha = 3$
Lateral ozone boundary conditions	log-normal	$\alpha = 1.5$
Lateral NO_x boundary conditions	log-normal	$\alpha = 3$
Major NO_x point emissions	log-normal	$\alpha = 1.5$
Wind velocity	log-normal	$\alpha = 1.5$
Wind direction	normal	$\pm 40°$
Temperature	normal	± 3 K
Vertical diffusion (night)	log-normal	$\alpha = 3$
Precipitations	log-normal	$\alpha = 2$
Cloud liquid water content	log-normal	$\alpha = 2$
Biogenic emissions	log-normal	$\alpha = 2$
Photolysis constants	log-normal	$\alpha = 2$

Table 1: Uncertainties associated with several input fields of a chemistry-transport model at continental scale. The uncertainty of a parameter $\hat{p}$ is measured with a confidence interval that includes 95% of the probability density integral. For a log-normal distribution, this interval is defined by a factor α so that $\hat{p}$ is in the interval $[\frac{p}{\alpha}, \alpha p]$ with a probability of 95% (p is the median value of $\hat{p}$). All estimates were derived from Hanna et al. [2001].

algorithms in use in this report and is freely available at `http://cerea.enpc.fr/polyphemus/`.

For this study, the Polyphemus model in use is Polair3D [Boutahar et al., 2004] for which an adjoint version is available (for gas-phase chemistry). The configuration of the model may be roughly described as follows:

1. raw meteorological data: ECMWF[†1] fields (resolution of $0.36° \times 0.36°$, 60 vertical levels, time step of 3 hours, 12 hours forecast-cycles starting from analyzed fields);
2. land use coverage: GLCF[†2] land cover map (14 categories, 1 km Lambert);
3. chemical mechanism: RACM [Stockwell et al., 1997];
4. emissions: the EMEP[†3] inventory, converted according to Middleton et al. [1990];
5. biogenic emissions: computed as proposed in Simpson et al. [1999];
6. deposition velocities: the revised parameterization from Zhang et al. [2003];
7. vertical diffusion: the Troen and Mahrt parameterization [Troen and Mahrt, 1986] (in the unstable boundary layer) and the Louis parameterization [Louis, 1979] (elsewhere);
8. boundary conditions: typical concentrations from the global chemistry-transport model Mozart 2 [Horowitz et al., 2003];
9. numerical schemes: a first-order operator splitting, the sequence being advection–diffusion–chemistry; a direct space-time third-order advection scheme with a Koren flux-limiter; a second-order order Rosenbrock method for diffusion and chemistry [Verwer et al., 2002].

[†1] European Centre for Medium-Range Weather Forecasts

[†2] Global Land Cover Facility

[†3] Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe

The model domain essentially covers Western Europe ($[35.0°\mathrm{N}, 10.5°\mathrm{W}] \times [57.5°\mathrm{N}, 22.5°\mathrm{E}]$). Two meshes are considered. The reference mesh has a 0.5° horizontal resolution, and the altitude of the tops of the vertical layers are 50 m, 600 m, 1200 m, 2000 m and 3000 m. The top layer is high enough to enclose the planetary boundary layer. A time step of 600 s is used. The coarse mesh has a 2° horizontal resolution and it includes three levels whose top heights are 50 m, 600 m and 1200 m. The time step is set to 1800 s. Both models have 72 chemical species (with RACM) mechanism. Hence the dimension of the state vector is about 1.1×10^6 for the full-resolution model and 3.8×10^4 for the coarse-resolution model.

An analysis of the simulations with the coarse mesh demonstrates that the main physical phenomena (at least for ozone) are reasonably reproduced in the context. The model retains good predictive capabilities (see the comparisons with observations in Section 4). The coarse case is used to perform intensive tests. For instance, in the Kalman algorithms that we use, the results depend on random numbers. Thus, they can only be assessed from a large number of trials, which is not tractable with the full resolution model. Nevertheless the full resolution study will be carried out later to verify some key findings in the coarse case.

The horizontal domain and its coarse discretization are shown in Figure 1.

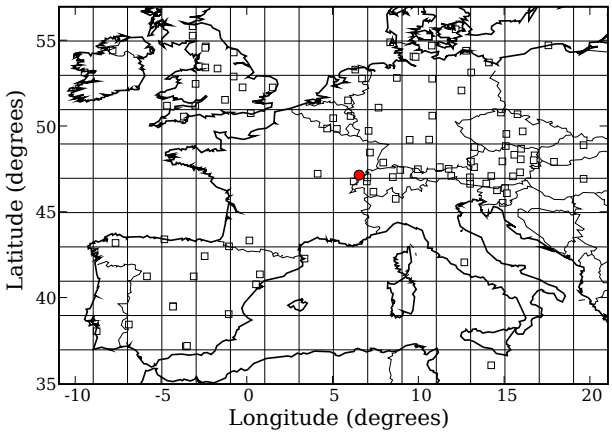


Figure 1: Horizontal coarse discretization of model domain. The squares show the locations of EMEP monitoring stations, and the disc shows the location of the monitoring station Montandon (east of France).

3.2 Observations

The observations to be assimilated are ozone hourly concentrations. These observations are provided by the EMEP[†4] network (Figure 1). The network is made of 151 ground stations among which 80–90 stations are actually available during the assimilation periods. They deliver point measurements integrated over one hour, but we assume that the observations are instantaneous – as it better fits the algorithms implementation.

The EMEP network includes only regional stations, which ensures a proper comparison between the continental-model outputs and the observations. The model outputs are linearly interpolated (on the horizontal, not on the vertical) at the station locations – the observation operator H_k is therefore linear and its adjoint is easily derived.

[†4]Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe

It is assumed that the error covariance matrix for ground ozone observations is diagonal, which is reasonable as the measurements from two stations are performed by distinct instruments. The standard deviation of the observation error is set to 10 $\mu g\,m^{-3}$ [Flemming et al., 2003]. Note that the mean of ozone observations is about 70 $\mu g\,m^{-3}$.

3.3 Reference Settings of the Assimilation Algorithms

In this section, we list our default settings of the assimilation algorithms. Sensitivity studies will be performed by alternating algorithm settings in later sections. The experiments consist of two steps: assimilation and prediction. During the assimilation period, say $[t_0, t_N]$, the observations are assimilated, and during the subsequent prediction period, say $[t_{N+1}, t_T]$, the ozone forecasts are the model simulations starting from the analyzed model state at t_N. In the reference setting the assimilation period is one day from 1st July 2001 at 01:00 UT to 2nd July at 00:00 UT. The prediction period is one day from 2nd July 2001 at 01:00 UT to 3rd July at 00:00 UT.

The model nonlinearity imposes an upper limit on the time length of the assimilation period (hereafter referred to as *assimilation window*). Previous perturbations out of this upper limit are ignored. In fact, driven by the winds, the pollutants may be transported out of the modeling domain after a few days. There is also a lower limit during which the impact of the assimilated observations propagates over the whole model domain. In meteorology, the assimilation time interval is about 6 hours extendable to 12 hours. In this study, the assimilation window is set to one day.

In the reference setting the state vector includes only ozone concentrations of the first two levels in the model domain. The correlations are supposed to expand gradually, through model simulations, to the complete model domain and to the species other than ozone.

In the Balgovind parameterization of the background error covariance matrix, the standard derivation $\sqrt{v}$ is set to 20 $\mu g\,m^{-3}$ (derived from usual RMSE for ozone forecast and from Mallet and Sportisse [2006b]), and the characteristic length is set horizontally to 3° (L_h), vertically to 200 m (L_v). The details about the uncertain parameters to be perturbed for EnKF and RRSQRT will be given in Section 4.2.2.

The EnKF ensemble number r is chosen to be 30. For a comparable computational cost, in RRSQRT, the number of columns q of the mode matrix is set to 20, the number of columns of the square root $\mathbf{Q}^{\frac{1}{2}}$ is set to 10, and the number of columns of the square root $\mathbf{R}^{\frac{1}{2}}$ is set to 10. In 4DVar, we employ the L-BFGS optimization solver [Byrd et al., 1995]. In this study, the computational cost of the adjoint model is about 5–7 times larger than that of the forward model, consequently the number of iterations is set to 6 so that the 4DVar cost may be comparable. Note that less iterations make 4DVar sub-optimal. However, we checked the evolution of the 4DVar cost function against iteration numbers, and found no considerable decrease in cost function values after 6 iterations (results omitted here).

4 Results

4.1 Coarse-Resolution Case

Let $\mathbf{s}$ be the vector of model outputs along space and time and $\mathbf{o}$ the vector of corresponding observations, the performance of ozone forecasts or assimilations can be assessed by the root mean square error (RMSE) calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{s}_i - \mathbf{o}_i)^2} \tag{23}$$

where n is the total number of available observations. The RMSE over a certain period with respect to all available observations is called *score*. In this paer, the RMSE will always be given in $\mu g\ m^{-3}$ · Hereafter this unit will be omitted for convenience.

The four assimilation algorithms provide better forecast scores compared to the reference simulations (model solutions without assimilation of observations), of course during the assimilation period (hourly forecasts for sequential assimilations), but also in the subsequent prediction steps (see Figure 2). OI has the best overall score, probably because the Balgovind parameterization of model error applies well to this coarse test case. From Figure 3, 4DVar has better scores during the early assimilation period but performs worse during the prediction period. The main reason may be that the underlying 4DVar is *strongly constrained*, that is, there is no model error term in its cost function. Only initial concentrations are controlled. The correction on the initial concentrations tends to be forgotten by the stable chemistry-transport system. EnKF provides the best performances during the late prediction period. It might benefit from its manner of perturbations on uncertain parameters. In OI, it can be considered that the model uncertainties are parameterized by the correlation in model states (which will be the initial conditions for following forecasts). The impact of model uncertainties in these initial conditions gradually fades out, when the model uncertainties in uncertain parameters (listed in Section 4.2.2) play an increasingly important role during the prediction period. RRSQRT shows poor performance against EnKF. This is probably due to the projection of mode matrices onto the leading eigenvectors of error covariance matrices. For a comparison of the assimilation performance between EnKF and RRSQRT, we refer to Hanea et al. [2004]. Note that in that paper, colored gaussian noises were added on several uncertain parameters, which is different from our perturbation method. The ozone forecasts at EMEP stations are plotted in Figure 4 and Figure 5. Most forecasts are between the reference simulations and the observations. All forecasts during the predictions period approach to the reference in the end. This shows again the rather low dependency of the short-range predictions on the initial conditions.

Caution has to be paid to the conclusions since the assimilation results can still be improved by optimal tuning of algorithm parameters.

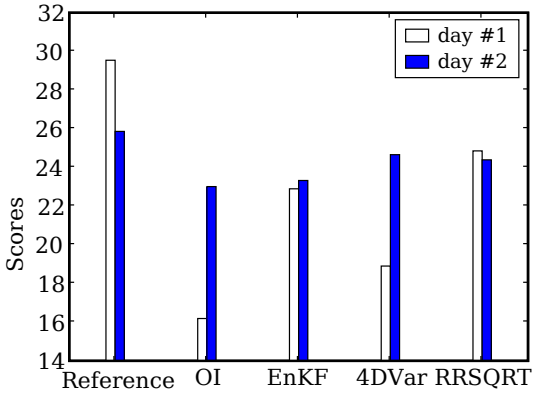


Figure 2: Scores of ozone concentrations during the assimilation period (day #1) and the prediction period (day #2).

4.2 Sensitivity Studies for the Coarse Case

The first set of tests is carried out in the coarse case. Modifications of configurations on each component of the data assimilation systems, i.e. model, observation and algorithm, may influ-

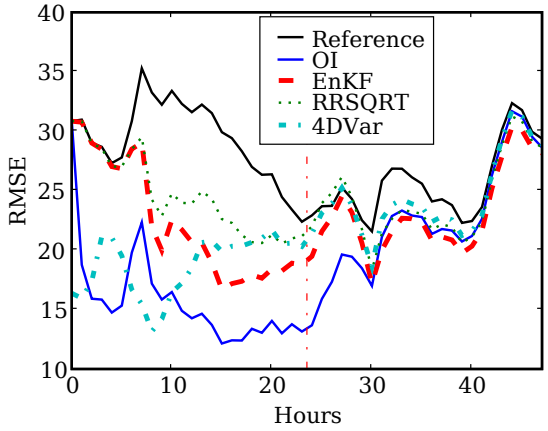


Figure 3: Time evolution of the RMSE for the ozone forecasts. The score over two days is 27.76 for reference, 19.90 for OI, 23.11 for EnKF, 21.98 for 4DVar, and 24.63 for RRSQRT. The vertical lines delimits the assimilation period from the prediction period.

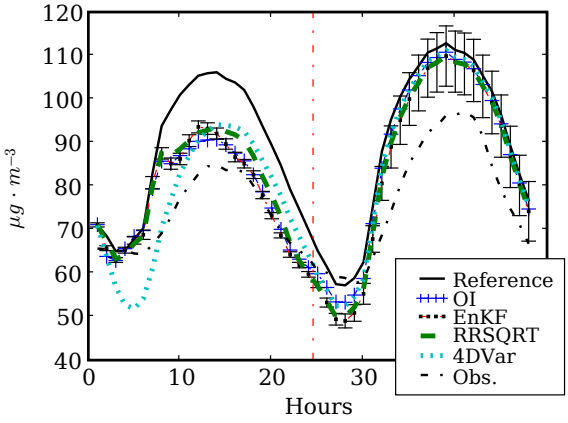


Figure 4: Time evolution of average ozone forecasts over all available stations. The error bar shows the average spread of the EnKF forecast ensemble calculated over these stations.

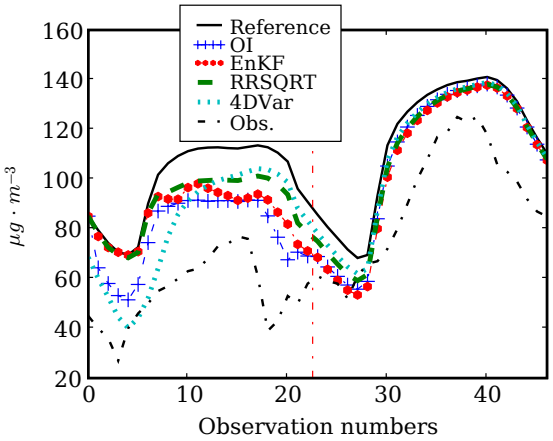


Figure 5: Time evolution of ozone forecasts against available observations over two days at EMEP station Montandon.

ence the assimilation performance. The main difficulty to interpret the results is that the error covariance structures $\mathbf{B}$ and $\mathbf{P}^f$ are unsettled subjects.

As for model component, we examine different state vectors to be controlled, alternative model physical parameterizations, Balgovind parameterization of background error covariance, and various perturbation settings for model error approximations. As for observation component, the error variance ratio between observations and background concentrations is examined. Different observation networks are tested. As for algorithm component, we evaluate the impact of the assimilation time length and EnKF ensemble issues, i.e. ensemble size and randomness. Hopefully some clues can be drawn for error modeling from the results of this sensitivity study. If not mentioned, only one factor is changed in each sensitivity study, and all other algorithm settings remain the same as those in Section 3.3.

4.2.1 Ensemble Randomness and Size

An important parameter for EnKF is its ensemble size. The directions of model error are approximated by the samples deviations from the ensemble mean. Recall formula (13), these directions are related to the outcomes of the parameters perturbations. A key question for this approach is how fast the assimilation results converge as the ensemble size increases. The spread of the model error space is determined not only by the ensemble size but also by the definition of the parameters set to be perturbed. The latter is addressed in the following section.

In this section, we conduct EnKF assimilations with ensembles of sizes 5, 10, 20, 30, 50, 70, 90 and 120. For each ensemble, the randomness of Monte Carlo sampling is accounted for by employing 10 different seeds for the random number generation. This means that 10 ensembles are generated for each ensemble size.

The forecast scores over both the assimilation period and prediction period are shown in Figure 6. Both converge as the ensemble size increases. The influence of the ensemble randomness decreases with ensemble size (see the errorbars). The augmentation of sample numbers improves forecast scores, but the improvements are modest probably due to the fast convergence of the procedure.

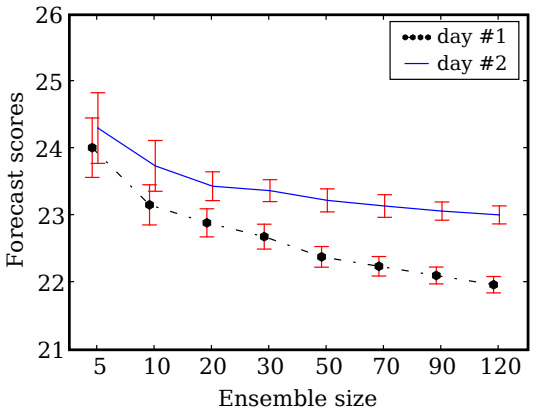


Figure 6: Forecast scores of EnKF against the ensemble size. The score for reference simulation without assimilation is 29.55 over day #1, and 25.87 over day #2. The curve shows mean scores and the errorbar shows the standard derivations over 10 random seed numbers.

The computational cost is proportional to the number of samples in the ensemble. The balance between computational cost and assimilation performance helps the specification of the

ensemble size. For realistic model grids, there is usually a constraint of $30 \sim 100$ ensemble samples due to computational considerations.

4.2.2 Uncertain Parameters Setting

Different parameters sets and perturbation magnitudes are listed in Table 2. The uncertain parameters are input data to the model. The perturbation magnitudes are kept reasonable (within the confidence interval at the 95% level), so that no instabilities may be produced due to the physically unrealistic parameter values. Notice that these parameters are only perturbed in Polair3D which chiefly carries out the numerical time integration. For instance, the perturbation of the temperature has no impact on the deposition velocities which are computed in preprocessing steps. The forecast scores with respect to uncertain parameters settings are shown in Figure 7. The results in the prediction period are slightly sensitive to the different uncertain parameters sets, which is consistent with the finding in Mallet and Sportisse [2006b] that the turbulent closure introduces the highest uncertainty. The results are more sensitive to the uncertain parameters sets than to the perturbation magnitudes. This probably indicates that the dimensionality of the bases of the perturbation parameter space are more important than the lengths of these bases for assimilations.

The perturbation magnitudes are spatiotemporally homogeneous in this study: $\sqrt{\alpha}^{\gamma}$ in equation (22) does not depend on spatial coordinates, and temporal correlations are not taken into account. However this hypothesis is not necessarily true. Refined perturbations might improve the forecast performances. Furthermore, additional uncertain parameters may be included for a larger model error spread.

	Parameter name	α_0	α_1	α_2
Ω_0	Boundary condition	3.	3.	3.
	Deposition velocity	1.5	2.	3.
	Photolysis rate	1.3	1.5	2.
	Surface emission	1.5	2.	3.
	Attenuation	1.3	1.5	2.
	Vertical diff. coef.	1.3	1.5	2.
Ω'	Cloud height	1.3	1.5	2.
	Vertical wind	1.3	1.5	2.
	Meridional wind	1.3	1.5	2.
	Zonal wind	1.3	1.5	2.
	Specific humidity	1.3	1.5	2.
Ω''	Pressure	1.3	1.5	2.
	Air density	1.3	1.5	2.
	Merid. diff. coef.	1.3	1.5	2.
	Zonal diff. coef.	1.3	1.5	2.
	Temperature	0.005	0.01	0.015

Table 2: Definition of uncertain parameters. Let Ω_0 be the set of parameter names for the reference setting in Section 3.3. Let Ω' and Ω'' be the two sets of additional parameter names. We denote Ω_1 as $\{\Omega, \Omega'\}$, and Ω_2 as $\{\Omega, \Omega', \Omega''\}$. The perturbation magnitude is characterized by α defined as in equation (22). The reference magnitudes are listed in α_0 column. In α_1 and α_2 columns, enlarged magnitudes are defined. Note that the distribution of temperature is supposed to be normal, and its magnitude should be interpreted as relative standard derivation.

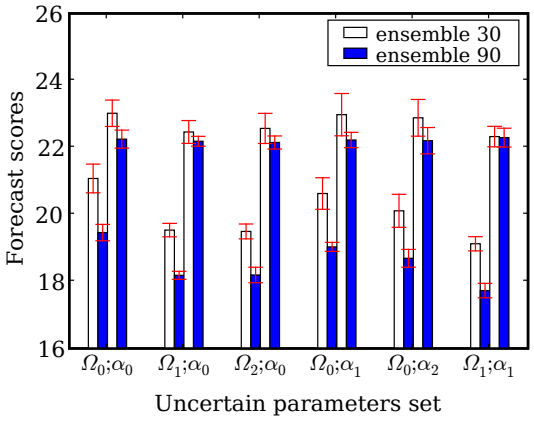


Figure 7: Forecast scores for EnKF against different uncertain parameter definitions. The parameter sets and perturbation magnitudes are defined in Table 2; $\{\Omega_0; \alpha_0\}$ is the reference setting. The EnKF sample number is chosen to be 30 (white columns) and 90 (dark columns) respectively. The two columns of scores for each case show the forecast scores during the assimilation and prediction periods. The bar values are mean scores, and the errorbar shows the standard derivations over 10 random seed numbers.

4.2.3 Assimilation Window

This determination of an optimal assimilation window is essentially linked with model nonlinearity, but should be treated separately according to sequential or variational context. In the variational case, the model nonlinearity makes the cost function nonconvex, and thus the optimization may suffer from the presence of local minima. Clearly there are constraints on the assimilation window, and for better performance the assimilation window has to be short [Pires et al., 1996]. In the sequential case, the observations are assimilated spontaneously. There are corrections on the state vector until the end of the assimilation window, which is an advantage for the subsequent prediction steps. Elegant analysis demands in-depth investigations on how the information (from observation) propagates among state components.

We perform brute-force tests. In the sequential case, the prediction period is fixed from 8$^{\text{th}}$ July 2001 at 01:00 UT to 9$^{\text{th}}$ July at 00:00 UT, whereas the assimilation window varies from 1 day to 7 days and always ends at 00:00 UT 8$^{\text{th}}$ July. The algorithm settings are the same as those for the reference case. For EnKF, the random seed is fixed with an ensemble size set to 30 and 90. The forecast scores over the prediction period are compared in Figure 8. We observe a considerable improvement in forecast scores with a window of 2 days against that with one day window. The first day of assimilation could be interpreted as an ensemble initialization, since the initial conditions of all members are identical in our implementation. We performed ensemble forecasts starting from identical samples and checked the ensemble spread (results not presented here), and we found out that the spread reached its maximum within 10 hours. This explains why a one-day assimilation with EnKF may be unsatisfactory. EnKF with longer assimilation windows (more than 2 days) outperforms OI in this experiment. Larger assimilation windows (more than 5 days) tend to be unnecessary long since the corrections are rapidly forgotten by the model.

In the variational context, we perform an experiment with the setting as that of the sequential case. The assimilation windows varies from 1 to 4 days, followed by one day of prediction. The start date of the prediction period is fixed at 01:00 UT 8$^{\text{th}}$ July 2001. The forecast scores are shown in Figure 9. The performances over prediction period decrease with longer assimilation

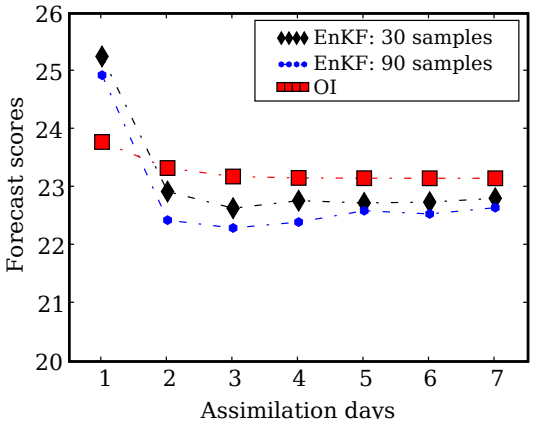


Figure 8: Forecast scores of OI and EnKF (with 30 and 90 members) against the number of assimilation days.

periods and approach the score of reference simulation without assimilation. These results clearly indicate the limitation of strongly constrained 4DVar in which only initial conditions are controlled. Model error has to be taken into account to form the weakly constrained 4DVar for better forecast performance.

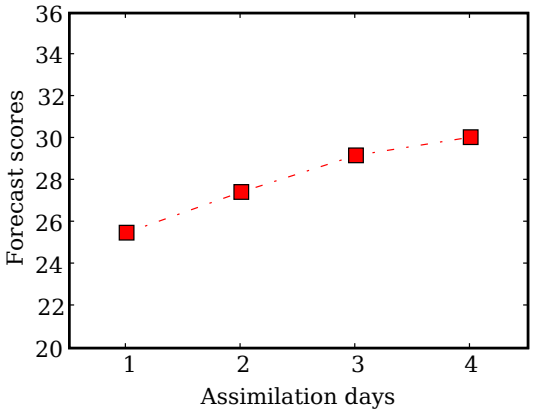


Figure 9: Forecast scores against the number of assimilation days for the two experiments using 4DVar.

4.2.4 Physical Parameterization

In order to assess the robustness of the assimilation, we apply EnKF to modified models which differ in their physical parameterizations. Several alternatives to the reference parameterizations are listed in Table 3, and the corresponding assimilation results are shown in Figure 10. The assimilations are performed using reference EnKF algorithm with 30 samples. The forecast scores are highly sensitive to the chemical mechanism, as in the uncertainty investigations of Mallet and Sportisse [2006b].

Parameterization	Reference	Alternative
Deposition velocities	Zhang	Wesely [Wesely, 1989]
Vertical diffusion	Troen and Mahrt	Louis [Louis, 1979]
Chemistry	RACM	RADM2 [Stockwell et al., 1990]

Table 3: Physical parameterization settings.

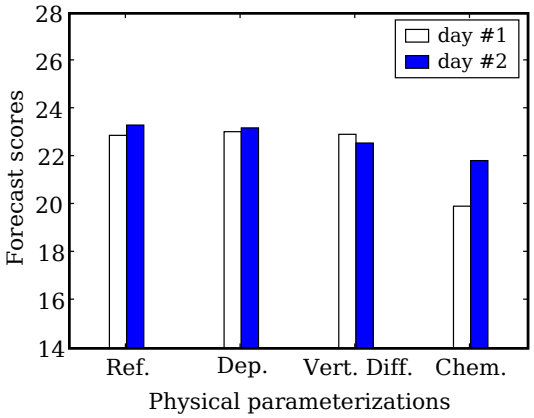


Figure 10: Forecast scores of EnKF, for modified models over the assimilation and the prediction periods.

4.2.5 State Component

It is not straightforward to control all model components, primarily because the correlations among them are unavailable. Investigations are needed to determine the most relevant state vector. In this section, we test the impact when including different vertical levels of ozone concentrations and different chemical species in the state vector. In Figure 11 we show the time evolution of the RMSE when controlling different vertical levels of ozone concentrations. Only including the first model level in the state vector is a fairly limited approach as the vertical transport plays a crucial role in ozone evolution. Consequently it is not surprising that the advantage of assimilating the first two levels over assimilating only ground level is enormous for all assimilation algorithms. However, in OI, the improvement of assimilating all levels over the first two levels is slight. By contrast, in 4DVar the improvement is still considerable during the assimilation period when assimilating all levels. This is probably because OI is a local assimilation algorithm in the sense that observations are assimilated instantaneously, whereas 4DVar searches global optima over the assimilation period that best fit the observations. In EnKF, the improvement is even considerable during the prediction periods when assimilating all levels. This might be due to the difference in error modeling. For OI and 4DVar, the vertical correlations are parameterized by the Balgovind correlation function. For EnKF, the vertical correlation structure is represented by the statistics of an ensemble generated by the perturbation method.

Because the correlation among different species is a priori unknown, only EnKF is employed to test the impact of assimilating different species. The ensemble size is 30, and the same random seed is used for all experiments. Only the first two levels of the domain are controlled. The species included in the state vector are combinations of O_3, NO, and NO_2. The results in Figure 12 show modest impact when assimilating different species. It is hard to interpret these results in depth. For further investigations, the interactions among model components have to

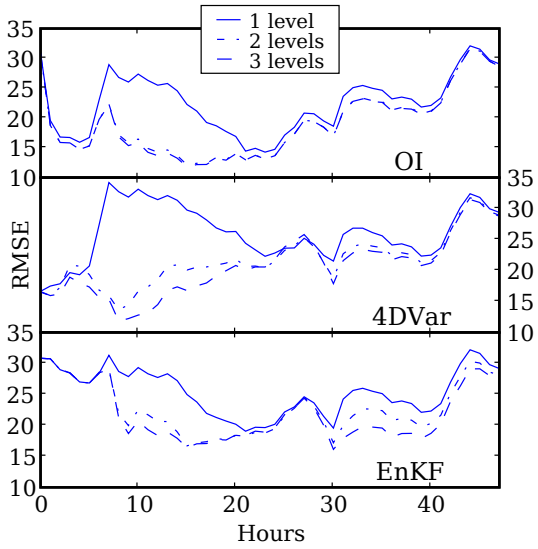


Figure 11: Time evolution of the RMSE against different vertical levels of ozone concentration to be controlled.

be quantified, for instance, by relative entropy [Liang and Kleeman, 2005].

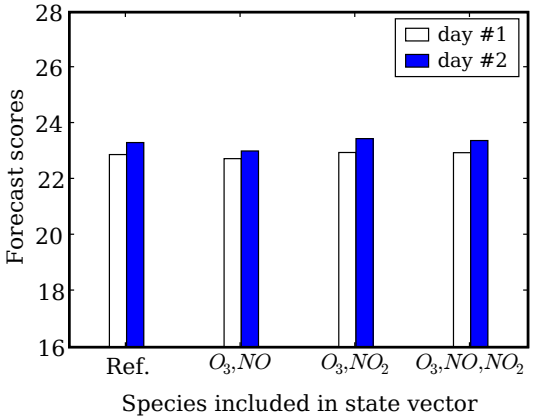


Figure 12: Forecast scores against different state components over assimilation and prediction periods.

4.2.6 Parameters in Balgovind Correlation Function

Balgovind characteristic lengths L_h and L_v determine the spatial structure of the background error covariances. We perform assimilations (OI and 4DVar) with different lengths listed in Table 4. The corresponding covariance structures vary from small to large scale correlations. Other experimental settings are the same as those of Section 3.3. The ozone forecast scores are shown in Figure 13 and Figure 14.

The assimilation is quite sensitive to the Balgovind characteristic lengths. For OI, the worst scores over the prediction period are those with the smallest vertical scale parameter ($L_v = 30$m). In this case, the vertical correlation is too weak. The worst scores over the assimilation period are those with largest horizontal scale parameter ($L_h = 10°$). There might be spurious correlations

	L_h (°)	L_v (m)
a	0.1	30
b	0.1	200
c	0.1	500
d	1.5	30
Reference	1.5	200
e	1.5	500
f	10	30
g	10	200
h	10	500

Table 4: Different configurations for Balgovind scale parameters.

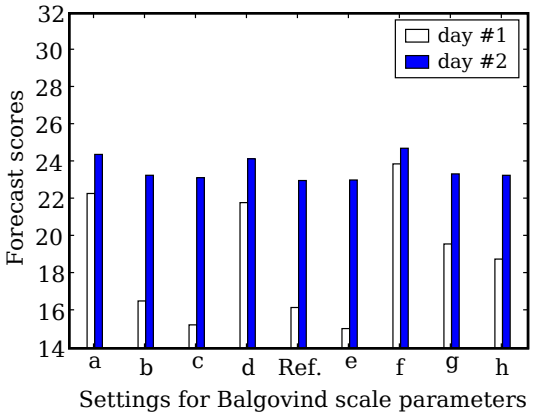


Figure 13: Forecast scores using OI against different configurations in Table 4.

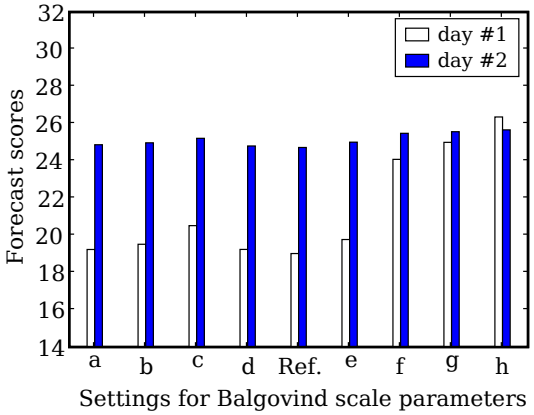


Figure 14: Forecast scores using 4DVar against different configurations in Table 4.

from distant observations. It seems that the medium scale correlation in the reference setting is a proper choice. The forecast scores, especially those over the assimilation periods, deteriorate when vertical scale parameter increases.

4.2.7 Observation Effect

The quantity and the quality of the observations are important factors for assimilation/prediction systems. There might be an optimal relationship between model resolutions and observation availabilities described implicitly by the optimality system [Le Dimet and Shutyaev, 2005; Bocquet, 2005]. Preliminary experiments are designed to address the model-observation relationships.

In a first experiment, we examine the impact of the observation network on ozone forecast performance. The original 151 EMEP stations are catalogued into three partitions: i) the center stations vs. around ones, ii) west stations vs. nonwest ones, and iii) randomly selected stations vs. the unselected ones. The observations from the center, west and randomly chosen stations are assimilated respectively. We take their counterparts as validation stations. Two assimilation algorithms, i.e. OI and EnKF, are employed with the reference settings. The forecast scores over the assimilation and the prediction periods given different observation networks are listed in Table 5. The corresponding scores without assimilations are listed for comparison in Table 6. The score gains are all positive. There is no clear winner for the two algorithms. OI has better overall performance. EnKF shows less disparities and better forecast performance over the prediction period. This may mainly result from the different correlation structures employed by the two methods (detailed in Section 4.3).

	day #1				day #2			
obs. net.	assim. stations		valid. stations		assim. stations		valid. stations	
	OI	EnKF	OI	EnKF	OI	EnKF	OI	EnKF
center	15.80	19.65	26.78	26.67	19.69	20.00	27.79	26.90
	13.74	9.89	2.77	2.88	2.12	1.80	0.88	1.78
west	13.39	20.94	30.63	26.33	24.84	24.02	23.14	22.31
	10.50	2.95	2.77	7.07	2.36	3.18	1.64	2.47
uniform	14.92	23.06	23.50	23.67	22.31	22.90	23.52	23.38
	14.56	6.42	6.11	5.94	3.55	2.95	2.35	2.50

Table 5: Forecast performances for different observation networks. The numerators are the forecast scores with assimilations, and the denominators are the gains in scores over the corresponding simulations without assimilation. The scores without assimilation are shown in Table 6.

	center	around	west	nonwest	uniform	rest
day #1	29.54	29.55	23.89	33.40	29.48	29.61
day #2	21.81	28.67	27.20	24.78	25.85	25.88

Table 6: Forecast scores of reference simulation without assimilation for different observation networks.

The inverse of the observation variance can be interpreted as the measurement accuracies. The assimilation is an optimization process with the objective weighted by the relative accuracies between observations and model simulations. In a second experiment, we examine the forecast performance using OI and 4DVar with a range of ratios between observation and background

error variances. The observation/background ratios, $\mathbf{R}/\mathbf{B}$ ratio in short, are shown in Table 7. The results are plotted in Figure 15. The assimilation performance is improved by increasing observation accuracies. However, there are little gains when decreasing the $\mathbf{R}/\mathbf{B}$ ratio below 0.05. 4DVar is less sensitive to the $\mathbf{R}/\mathbf{B}$ ratio, since the background errors are only considered at initial conditions. For OI, when the observations are supposed to be extremely accurate ($\mathbf{R}/\mathbf{B}$ ratio at 0.01) there are artificial fluctuations at some stations where the observations are not compatible with the chemical state of the model.

	$\mathbf{R}/\mathbf{B}$	$\mathbf{R}$	$\mathbf{B}$
$\mathbf{R}++$	0.01	100	10000
$\mathbf{R}+$	0.05	100	2000
Reference	0.25	100	400
$\mathbf{B}+$	1	400	400
$\mathbf{B}++$	10	4000	400

Table 7: Different ratios between observation ($\mathbf{R}$) and background ($\mathbf{B}$) error variances; '+' means more accuracy.

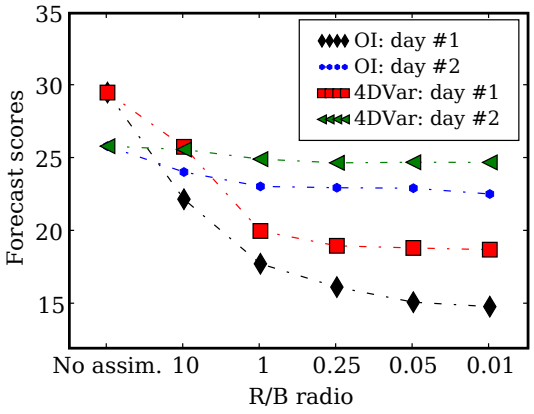
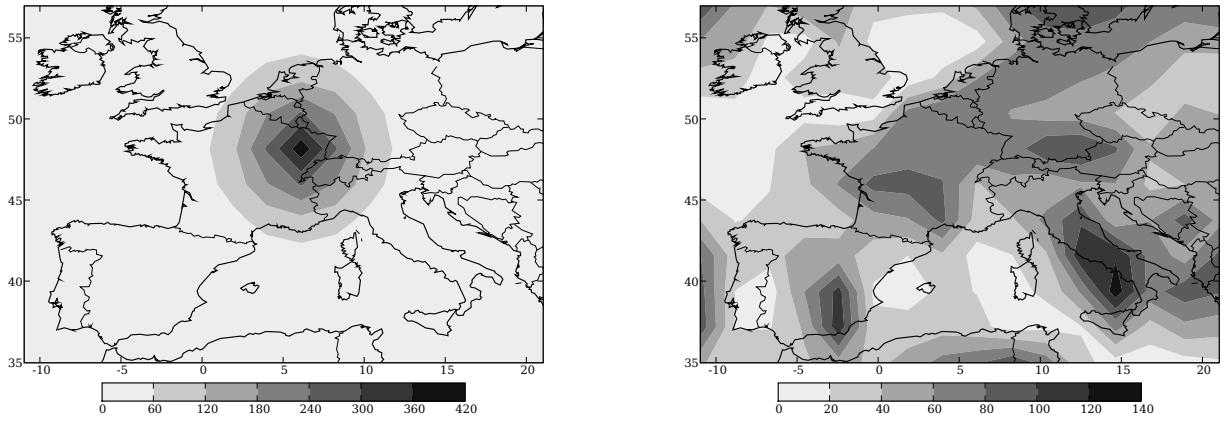


Figure 15: Forecast scores against different $\mathbf{R}/\mathbf{B}$ ratios shown in Table 7.

4.3 Model Error Covariance Structure for Coarse Case

The model error covariance structure ($\mathbf{B}$ or $\mathbf{P}$) is decisive to the assimilation behavior. In many occasions we resort to them for explanations of our results. The covariance between the error at the station Montandon and the error in all ground cells is shown in Figure 16. The covariance field obtained by the statistics of the EnKF forecast ensemble shows an irregular structure which brings detailed information compared to the isotropic Balgovind parameterization. However, spurious correlations may be produced by the homogeneous perturbations (see equation (22)).

In Figure 17, we show the assimilation/prediction performances at several randomly chosen stations. Except for three stations (St. Koloman, Heidenreichstein, and Bottesford), the ensemble predictions fail in the sense that the observations (with their standard deviations set to 10 $\mu g\,m^{-3}$) are not within the range of model errors represented by the EnKF ensemble spread. The ensemble spread during the assimilation period is dramatically decreased after assimilating observations. In our EnKF implementation, no additional inflations are conducted on the state error covariance $\mathbf{P}_k^f$ (see equation (14)) as in Constantinescu et al. [2007b].



(a) Balgovind parameterization

(b) Approximations by EnKF forecast ensemble

Figure 16: The covariance between the error at the station Montandon and the error in all ground cells at 13:00 UT, 2nd July 2001.

The ensemble relative standard derivations averaged over time are shown in Figure 18. As expected, the ensemble spread decreases when the assimilation is applied. One may notice high uncertainties around the coasts, although is not as clear as in Mallet and Sportisse [2006b] where the uncertainties were generated by statistics over several months. Our ensemble spread describes an accidental uncertainty configuration that depends not only on the chemistry-transport processes, e.g. turbulence, but also on the meteorological scenarios.

4.4 Cycling Forecast for Coarse Case

It is expected that the findings in the previous sections are independent of the assimilation dates. To this end, we perform assimilation/prediction processes consecutively for one week. The length of the assim./predict. periods is chosen to be one day. The first assimilation period is from 1st July 2001 at 01:00 UT to 2nd July at 00:00 UT, followed by the prediction from 2nd July at 01:00 UT to 3rd July at 00:00 UT. The assimilation period of the second assim./predict. process is the same as the prediction period of the first assim./predict. process. The second prediction period is from 3rd July at 01:00 UT to 4th July at 00:00 UT. The cycling continues until the final assim./predict. process. The last prediction period is from 8th July at 01:00 UT to 9rd July at 00:00 UT. The initial conditions for the subsequent assimilation periods are the hourly forecasts based on previously controlled states after assimilations. The performance of OI, EnKF and 4DVar forecasts over the prediction periods are shown in Figure 19.

The improvement of forecast performance with assimilation is obvious. The Balgovind parameters for OI and 4DVar are constant, so are the perturbation magnitudes for the EnKF samples. The lesser performance of 4DVar is probably due to the absence of model error during assimilation. EnKF surpasses OI in forecasts longer than twelve hours. The approximation of model error by refined perturbations is promising for longer forecasts.

4.5 Full-Resolution Case

The previous results are obtained with coarse models at a resolution of $2^\circ \times 2^\circ \times 1800$s with 3 vertical levels. In this section, the reference resolution at $0.5^\circ \times 0.5^\circ \times 600$s with 5 levels is

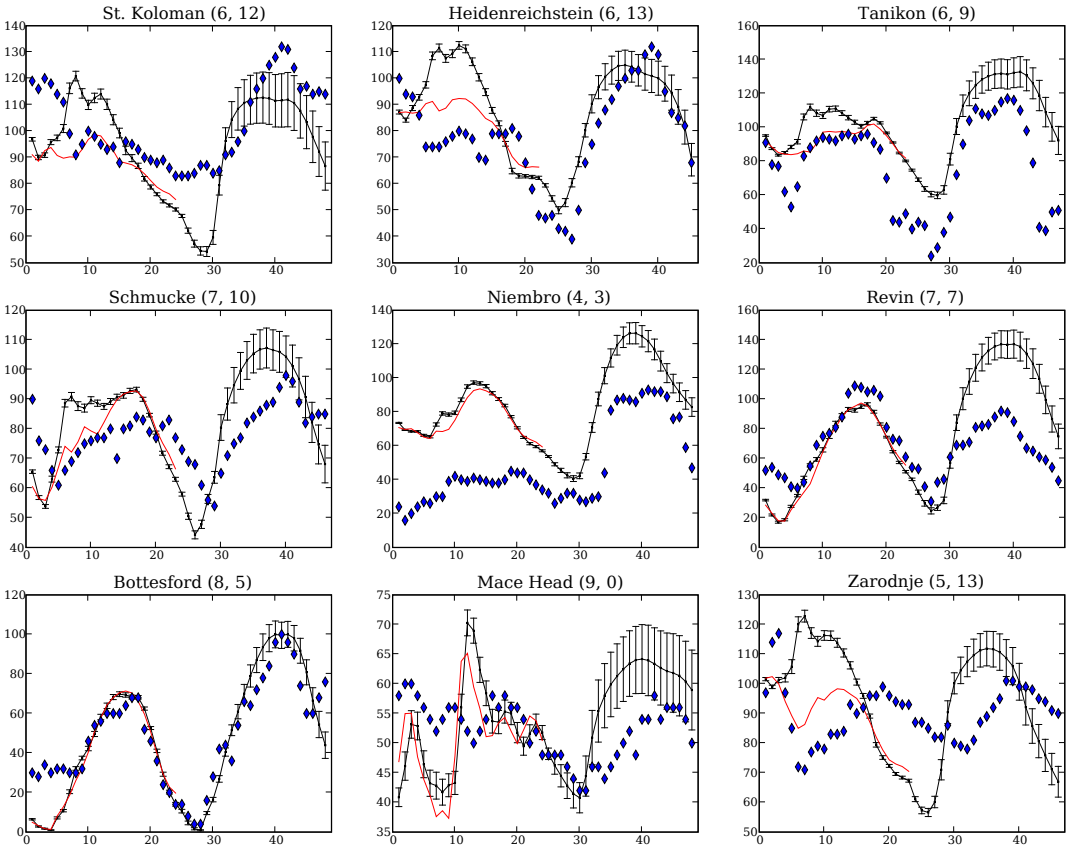
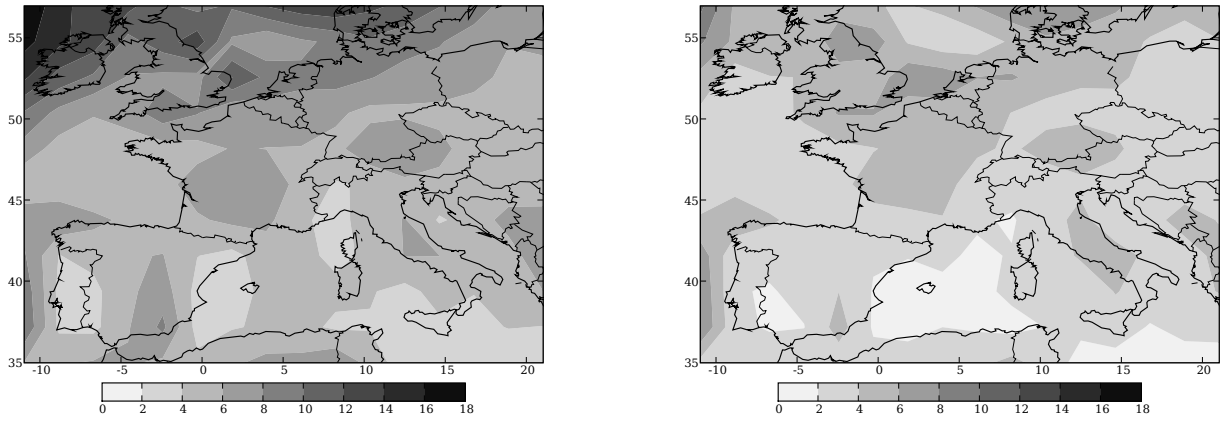


Figure 17: EnKF assimilation/prediction performances at nine stations. The titles read station names and their coordinates in model grid indices. The horizontal axis shows the accumulated available observation numbers along time. Along vertical axis, the ozone concentrations are plotted in $\mu g\,m^{-3}$. The diamond points show the observations, the short lines plot assimilated concentrations, and the lines with error bars are the means of the forecast ensemble. The error bars are the relative standard derivations calculated with the forecast ensemble.



(a) Calclulated by Monte Carlo forecast ensemble

(b) Calculated by EnKF forecast ensemble

Figure 18: Maps of the averaged relative standard derivation during both assimilation and prediction periods. The ensemble spread is decreased by assimilating observations.

employed (detailed in Section 3.1). There are 33 cells along latitude and 65 cells along longitude.

All available assimilation algorithms are tested for this full-resolution setting. The assimilation experiments are similar to those in Section 3.3, but conducted at three different dates. The assimilations are performed at 1^{st}, 3^{rd}, 7^{th} July respectively, and the corresponding prediction dates are 2^{nd}, 4^{th}, 8^{th} July. The Balgovind parameters for $\mathbf{B}$ are the same as those in coarse-resolution case for OI and 4DVar. The sample number of EnKF ensemble is set to 30. For RRSQRT, the number of columns in the mode matrix is 30, the number of column in $\mathbf{Q}^{\frac{1}{2}}$ is 20, and the number of columns in $\mathbf{R}^{\frac{1}{2}}$ is 10.

In general, the magnitude and structure of model error vary with respect to the model resolution. The assimilation results are shown in Figure 20 and Figure 21. The assimilations improve the forecast scores. The poor performance of EnKF forecasts at 4^{th} July might be the consequence of excessive perturbations. Comparing with the forecast scores at 2^{nd} July for the coarse case in Figure 2, one can find that the Balgovind parameterization of model error are stable (OI and 4DVar results), whereas the perturbation methods are sensitive to the changing of model resolutions (EnKF and RRSQRT).

Better results can be obtained via tuning the algorithm parameters in perturbation methods for the full-resolution case. The sensitivity of assimilation performance to the ensemble size and the assimilation window are presented in Figure 22 and Figure 23. The aim of this simple sensitivity study is not to find the optimal algorithm parameters for the full-resolution model, but to verify the main findings in the coarse case. For instance, the forecast scores are improved by augmenting ensemble samples. We observe spin-up process in the first day of EnKF assimilation, and we have satisfactory results with assimilation window set to three days.

5 Conclusion

In order to design suitable assimilation algorithms for short-range ozone forecasts in realistic applications, four algorithms, namely optimal interpolation, reduced-rank square root Kalman filter, ensemble Kalman filter, and four-dimensional variational assimilation, have been implemented and compared in the same benchmark settings.

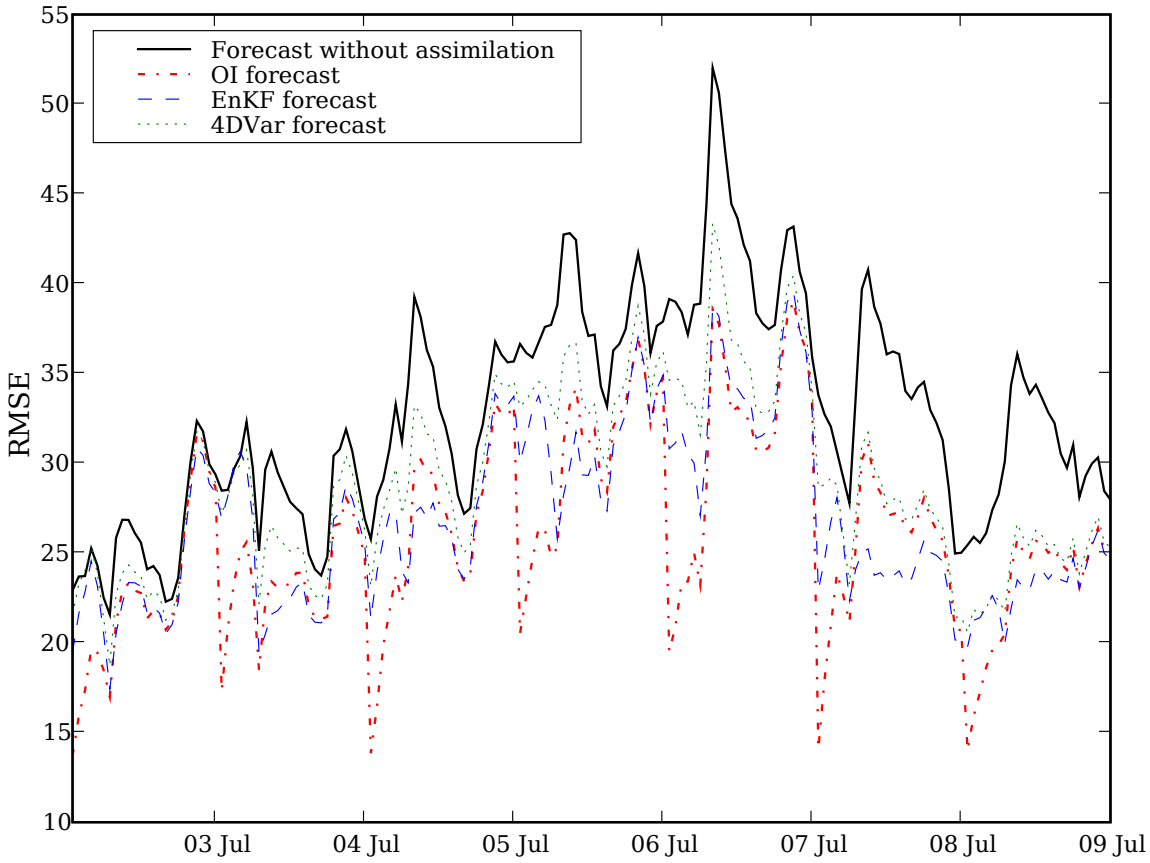


Figure 19: The one-day forecast performances based on model simulations with/without assimilations in the context of cycling assimilation/predictions.

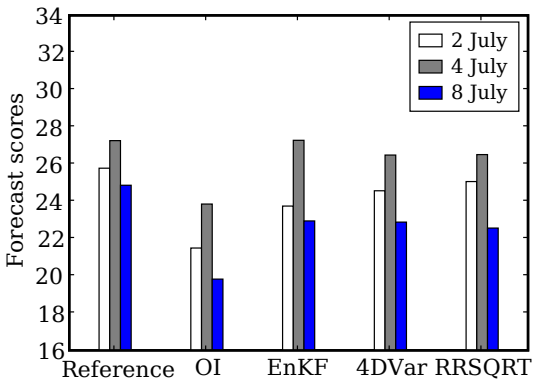


Figure 20: Forecast scores of ozone concentrations during prediction dates.

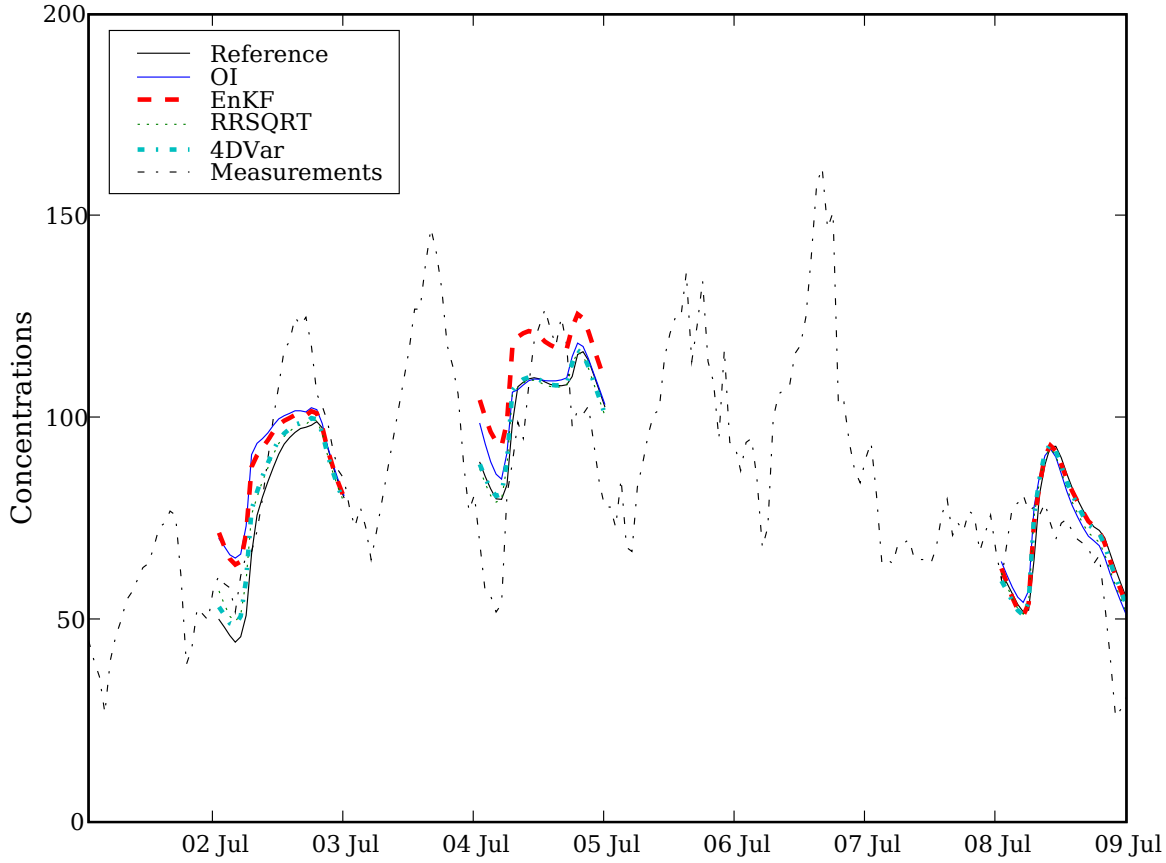


Figure 21: Time evolution of ozone forecasts against available observations at Montandon station.

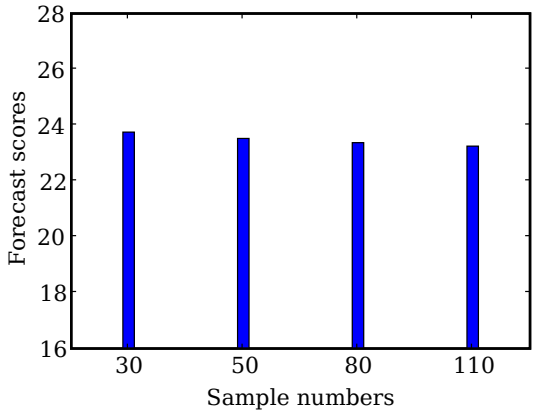


Figure 22: Forecast scores at 2nd July against EnKF ensemble size.

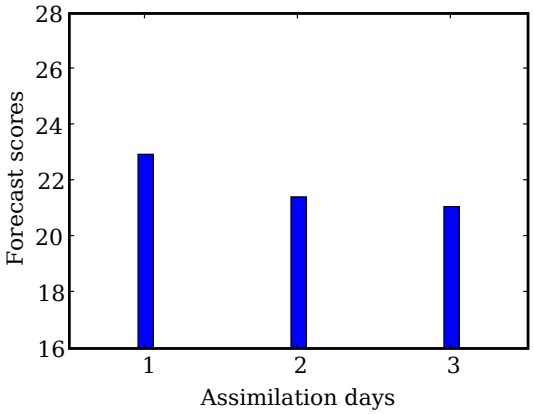


Figure 23: EnKF forecast scores at 8^{th} July against the number of assimilation days.

Although the forecasts beyond one day tend to approach the model simulations without assimilation (because of the low dependency of model simulations on initial conditions), it has been shown that the assimilation algorithms significantly improve the ozone forecasts. Data assimilation would be an indispensable part of practical ozone forecast systems as in NWP.

The comparison results have illustrated the limitations and potentials of different assimilation algorithms. OI provides overall better performances. It benefits from the Balgovind parameterization of model uncertainties during the assimilation periods. In EnKF, the model uncertainties were approximated by the statistics of the ensemble generated by perturbing uncertain model parameters. This perturbation method shows good potential to alleviate the constraint of the low dependency on initial conditions in ozone forecasts. EnKF produces best forecasts during the end of prediction periods. The strongly constrained 4DVar does a moderate job, because uncertainties are taken into account only at the initial date of the assimilation. Further studies are needed, e.g. the estimation of ozone concentrations jointly with the emission rates [Elbern et al., 2007]. We remark that there are no final conclusions because of the unsettled formulation of model error. We paid less attention to RRSQRT, since, in our implementation, it is quite similar to EnKF.

We have also conducted sensitivity analysis on algorithm parameters, e.g. ensemble randomness and size, assimilation window, perturbation fields, and diverse model settings. Further refinements of the assimilation algorithms can thus be tested by tuning these algorithm parameters for better forecast performances. This is especially necessary for the case of the full-resolution model.

It is the complexity of the chemistry-transport phenomena and the limited observations that make it difficult for the modeling and assimilation. The approximations of the complex phenomena make CTMs imperfect, and uncertainties arise. If a deterministic model is employed, or if the uncertainties are not realistic, the forecasts of the stable system approach to the reference simulations without assimilation. Therefore all assimilation algorithms have to be adaptive, in one way or another, so that the uncertainties should be better represented.

For the design of better assimilation algorithms, serious investigations on error modeling are needed. The ensemble could be obtained from more uncertain sources, e.g. numerical approximations and subgrid physical parameterizations. The statistics of the enlarged ensemble are expected to be more accurate approximations of the model error.

Spatially heterogeneous perturbations, which smooth out the spurious correlations of long distance, would certainly produce more realistic model errors. Another concern is that the

lognormal perturbations might result in non-Gaussian model errors. In this case, assimilation methods deviating from normal may have to be accounted for. Other methods rely on hybridizations between sequential and variational methods which essentially use the assimilation results from both methods for error modeling. The comparison of ensemble forecast techniques [Mallet and Sportisse, 2006a] and assimilation algorithms would also be an interesting topic for future studies.

References

R. Balgovind, A. Dalcher, M. Ghil, and E. Kalnay. A stochastic-dynamic model for the spatial structure of forecast error statistics. *Mon. Wea. Rev.*, 111(4):701–722, 1983.

M. Bocquet. Grid resolution dependence in the reconstruction of an atmospheric tracer source. *Nonlinear Proc. Geoph.*, 12(2):219–233, 2005.

Jaouad Boutahar, Stéphanie Lacour, Vivien Mallet, Denis Quélo, Yelva Roustan, and Bruno Sportisse. Development and validation of a fully modular platform for numerical modelling of air pollution: POLAIR. *Int. J. Env. and Pollution*, 22(1/2):17–28, 2004.

F. Bouttier and P. Courtier. Data assimilation concepts and methods. Meteorological Training Course Lecture Series, ECMWF, 1999.

Gerrit Burgers, Peter Jan van Leeuwen, and Geir Evensen. On the analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, 126:1719–724, 1998.

R. H. Byrd, P. Lu, and J. Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM J. on Sci. and Stat. Comp.*, 16(5):1,190–1,208, 1995.

Gregory R. Carmichael, Adrian Sandu, Tianfeng Chai, Dacian N. Daescu, Emil M. Constantinescu, and Youhua Tang. Predicting air quality: Improvements through advanced methods to integrate models and measurements. *J. Comp. Phys.*, 227:3540 – 3571, 2008. doi: 10.1016/j.jcp.2007.02.024.

T. Chai, G. R. Carmichael, Y. Tang, A. Sandu, M. Hardesty, P. Pilewskie, S. Whitlow, E.V. Browell, M.A. Avery, P. Nédélec, J.T. Merrill, A.M. Thompson, and E. Williams. Four-dimensional data assimilation experiments with international consortium for atmospheric research on transport and transformation ozone measurements. *J. Geophys. Res.*, 112, 2007. doi: 10.1029/2006JD007763.

E.M. Constantinescu, T. Chai, A. Sandu, and G.R. Carmichael. Autoregressive models of background errors for chemical data assimilation. *J. Geophys. Res.*, 112:D12309, 2007a. doi: 10.1029/2006JD008103.

E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. Ensemble-based chemical data assimilation I: General approach. *Quart. J. Roy. Meteor. Soc.*, 133(626):1229–1243, July 2007b. ISSN 0035-9009. doi: 10.1002/qj.76.

R. Daley. *Atmospheric data analysis*. Combridge University Press, 1991.

H. Elbern and H. Schmidt. Ozone episode analysis by four-dimensional variational chemistry data assimilation. *J. Geophys. Res.*, 106(D4):3,569–3,590, 2001.

H. Elbern, A. Strunk, H. Schmidt, and O. Talagrand. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.*, 7(14):3749–3769, 2007.

Geir Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynam.*, 53:343–367, 2003.

Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, 99(C5):10,143–10,162, 1994.

C. Faure and Y. Papegay. O∂yssée user's guide – version 1.7. Technical Report 0224, INRIA, 1998.

M. Fisher and D. J. Lary. Lagrangian four-dimensional variational data assimilation of chemical species. *Quart. J. Roy. Meteor. Soc.*, 121:1681–1704, 1995.

J. Flemming, M. van Loon, and R. Stern. Data assimilation for CTM based on optimum interpolation and Kalman filter. paper presented at 26th NATO/CCMS International Technical Meeting on Air Pollution Modeling and Its Application, NATO Comm. on the Challenges of the Mod. Soc., Istanbul, 2003.

R. G. Hanea, G. J. M. Velders, and A. Heemink. Data assimilation of ground-level ozone in Europe with a Kalman filter and chemistry transport model. *J. Geophys. Res.*, 109, 2004.

Steven R. Hanna, Joseph C. Chang, and Mark E. Fernau. Monte Carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. *Atmos. Env.*, 32(21):3,619–3,628, 1998.

Steven R. Hanna, Zhigang Lu, H. Christopher Frey, Neil Wheeler, Jeffrey Vukovich, Saravanan Arunachalam, Mark Fernau, and D. Alan Hansen. Uncertainties in predicted ozone concentrations due to input uncertainties for the UAM-V photochemical grid model applied to the July 1995 OTAG domain. *Atmos. Env.*, 35(5):891–903, 2001.

A. W. Heemink, M. Verlaan, and A. J. Segers. Variance reduced ensemble Kalman filtering. *Mon. Wea. Rev.*, 129:1,718–1,728, 2001.

J. Hoelzemann, H. Elbern, and A. Ebel. PSAS and 4DVar data assimilation for chemical state analysis by urban and rural observation sites. *Phys. Chem. Earth*, 26:807–812, 2001.

Larry W. Horowitz, Stacy Walters, Denise L. Mauzerall, Louisa K. Emmons, Philip J. Rasch, Claire Granier, Xuexi Tie, Jean-François Lamarque, Martin G. Schultz, Geoffrey S. Tyndall, John J. Orlando, and Guy P. Brasseur. A global simulation of tropospheric ozone and related tracers: description and evaluation of MOZART, version 2. *J. Geophys. Res.*, 108(D24), 2003.

E. Kalnay. *Atmospheric modeling, data assimilation and predictability.* Cambridge Univ. Press, 2003.

E. Kalnay, H. Li, T. Miyoshi, S.-C. Yang, and J. Ballabrera. 4DVar or ensemble Kalman filter. *Tellus A*, 59A:758–773, 2007.

F.-X. Le Dimet and V.P. Shutyaev. On deterministic error analysis in variational data assimilation. *Nonlinear Proc. Geoph.*, 12(4):481–490, May 2005.

François-Xavier Le Dimet and Olivier Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38A:97–110, 1986.

X. S. Liang and R. Kleeman. Information transfer between dynamical system components. *Phys. Rev. Lett.*, 95(24):244101, 2005.

Andrew C. Lorenc. The potential of the ensemble Kalman filter for NWP - a comparison with 4DVar. *Quart. J. Roy. Meteor. Soc.*, 129(595):3,183–3,203, 2003.

Jean-François Louis. A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteor.*, 17:187–202, 1979.

Vivien Mallet and Bruno Sportisse. Ensemble-based air quality forecasts: A multimodel approach applied to ozone. *J. Geophys. Res.*, 111(D18), 2006a.

Vivien Mallet and Bruno Sportisse. Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: An ensemble approach applied to ozone modeling. *J. Geophys. Res.*, 111(D1), 2006b.

Vivien Mallet, Denis Quélo, Bruno Sportisse, Meryem Ahmed de Biasi, Édouard Debry, Irène Korsakissok, Lin Wu, Yelva Roustan, Karine Sartelet, Marilyne Tombette, and Hadjira Foudhil. Technical Note: The air quality modeling system Polyphemus. *Atmos. Chem. Phys.*, 7 (20):5,479–5,487, 2007.

Paulette Middleton, William R. Stockwell, and William P. L. Carter. Aggregation and analysis of volatile organic compound emissions for regional modeling. *Atmos. Env.*, 24A(5):1,107–1,133, 1990.

C. Pires, R. Vautard, and O. Talagrand. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus A*, 48A:96–121, 1996.

Arjo Segers. *Data assimilation in atmospheric chemistry models using Kalman filtering.* PhD thesis, Delft University, 2002. www.library.tudelft.nl.

D. Simpson, W. Winiwarter, G. Börjesson, S. Cinderby, A. Ferreiro, A. Guenther, C. N. Hewitt, R. Janson, M. A. K. Khalil, S. Owen, T. E. Pierce, H. Puxbaum, M. Shearer, U. Skiba, R. Steinbrecher, L. Tarrasón, and M. G. Öquist. Inventorying emissions from nature in Europe. *J. Geophys. Res.*, 104(D7):8,113–8,152, 1999.

W. R. Stockwell, P. Middleton, J. S. Chang, and X. Tang. The second generation regional acid deposition model chemical mechanism for regional air quality modeling. *J. Geophys. Res.*, 95 (D10):16,343–16,367, 1990.

William R. Stockwell, Frank Kirchner, Michael Kuhn, and Stephan Seefeld. A new mechanism for regional atmospheric chemistry modeling. *J. Geophys. Res.*, 102(D22):25,847–25,879, 1997.

I.B. Troen and L. Mahrt. A simple model of the atmospheric boundary layer; sensitivity to surface evaporation. *Boundary-Layer Meteor.*, 37:129–148, 1986.

J. G. Verwer, W. Hundsdorfer, and J. G. Blom. Numerical time integration for air pollution models. *Surveys on Math. for Indus.*, 10:107–174, 2002.

M. L. Wesely. Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models. *Atmos. Env.*, 23:1,293–1,304, 1989.

L. Zhang, J. R. Brook, and R. Vet. A revised parameterization for gaseous dry deposition in air-quality models. *Atmos. Chem. Phys.*, 3:2,067–2,082, 2003.