Application of Sequential Importance Resampling (SIR) for 1D Atmospheric Chemical Data Assimilation

S.E. Walker (*sew@nilu.no*) †, **M. Hatlo** ‡

† *Norwegian Institute for Air Research (NILU)* ‡ *Shell Technology Enterprise Programme (STEP)*

Introduction

The SIR (Sequential Importance Resampling) assimilation method (*Van Leeuwen, 2003; Doucet, 2001*) is tested on a 1D atmospheric advection-diffusion model with photochemistry. Simulated experiments, defining a true set of input parameters and resulting model concentrations, are performed to see if the method can handle both systematic (bias) and unsystematic (random) errors in the input data, and still be able to produce assimilated values close to the true state. The effect on the performance of using different observations likelihood functions, such as Gaussian or Lorentz (Student's t) distributions, are also analysed.

Model description

The 1D model tested is:

$$\frac{\partial \mathbf{c}}{\partial t} = -u\frac{\partial \mathbf{c}}{\partial x} + \frac{\partial}{\partial x}\left(k_x \frac{\partial \mathbf{c}}{\partial x}\right) + \mathbf{R} + \mathbf{q} \quad (1)$$

where $\mathbf{c}$ is a space (x) and time (t) varying concentration vector (µg/m) containing the species NO_2, NO and O_3, u is the wind speed and k_x a turbulent eddy diffusivity coefficient. $\mathbf{R}$ denotes the non-linear fast reaction NO_2-NO-O_3 photochemistry operator, and $\mathbf{q}$ represents emissions of the same three species. Boundary and initial conditions are given by $\mathbf{c}(x,t) = \mathbf{c}_B$ for x = 0 and x = nΔx, and for t = 0, where $\mathbf{c}_B$ denotes a set of background concentrations of the three species. The physical domain [0, nΔx] is divided into n grid cells each with length Δx. For the tests performed here n = 50 and Δx = 1000 m. The equation is discretized and solved on an hourly basis using hourly input data of u, k_x, $\mathbf{q}$ and $\mathbf{c}_B$, and separate operators for advection, diffusion and photochemistry (*Bott, 1989; Slørdal et al, 2003*).

Method description

The SIR-method generates an ensemble of possible model states $\{\mathbf{x}^{(i)}, i = 1,\ldots,N\}$ by randomly drawing selected input parameters to the model. The ensemble represents a discrete approximation of the Bayesian (*Box and Tiao, 1992*) prior and posterior probability density functions (PDFs) of the true model state $\mathbf{x}^t$ given the model forecasts and observations. The number N of ensemble members is kept constant at all time steps.

The assimilated model state is calculated as:

$$\mathbf{x}^a = \sum_{i=1}^{N} w_i \mathbf{x}^{(i)} \quad (2)$$

where w_i = 1/N for i = 1,...,N represents the ensemble weights. Updated weights $\hat{w}_i$ are calculated using a Gaussian or Lorentz shaped likelihood function based on available observations. In the resampling step, ensemble members that correspond well with the observations (high weights) will be kept and copied, while those that correspond poorly with the observations (low weights) will be removed. After the resampling step, all ensemble members again have weights 1/N.

Eq. (2) represents a variance-minimizing estimate of the true model state $\mathbf{x}^t$ even for non-linear models with non-Gaussian error structures. The ensemble size N needed in practice depends on the model, the number of state variables, and the number and position of observations. A trial and error procedure must usually be exercised in order to find the optimal number of ensemble members.

Experimental set-up

The model (1) is run for 2 weeks (336 hours). Realistic hourly values of wind speed (u) and temperature difference (ΔT_{10m-2m}) is provided from a meteorological station close to Oslo, Norway. The station is placed in a relatively flat and homogenous area (z_0 = 0.1 m). A meteorological preprocessor is used to calculate horisontal turbulence intensities σ_y and diffusion coefficients k_x as $0.1 \cdot \Delta x \cdot \sigma_y$ (*Slørdal et al., 2003*). Expected values of emissions ($\mathbf{q}$) and background concentrations ($\mathbf{c}_B$) are set equal to 10^{-3} , $9 \cdot 10^{-3}$ and 0 µg/m·s, and 10, 0 and 50 µg/m respectively for each of the three species and constant for all hours.

The model state vector $\mathbf{x}$ is defined as the concentration grid vector $\mathbf{c}$. In order to create the initial ensemble and to update the ensemble from one time step to the next, actual input parameters u, k_x, $\mathbf{q}$ and $\mathbf{c}_B$ to the model are drawn randomly using lognormal distributions. The hourly observed values are used as mean values in these distributions, and the standard deviations are assumed to be 40% of these values. The values are set equal for all grid cells.

True values $(\cdot)^t$ of the above parameters are defined using the expectance values and an assumed bias factor f_b = 1.2 (20% bias) such that $u^t = E(u) \cdot f_b$, $k_x^t = E(k_x) \cdot f_b$ and $\mathbf{q}^t = E(\mathbf{q})/f_b$. The true background values are always assumed to be unbiased, i.e., $\mathbf{c}_B^t = E(\mathbf{c}_B)$. Pseudo-observations of NO_2 are assumed to be Gaussian or Lorentz-distributed around the true model concentrations using a standard deviation equal to 5% of the true value for each hour. We assume that there are no observations of NO or O_3.

Results

Hourly concentrations of NO_2 in grid cell 27 are shown in Figs. 1-2. Only the tests performed with the Lorentz distribution are shown here. Generally it was found that this gave more stable and consistent improvements than by using Gaussian distributions.

The assimilated concentrations (red) lies consistently closer to the true concentrations (green) than the free run concentrations (blue), although the improvement varies with time. The yellow and orange curves represents resp. 2.5 and 97.5 percentiles of the assimilated (posterior) concentration distributions based on the ensemble members. Increasing the ensemble size N from 25 to 100 and the number of observations from 1 (cell 10) to 2 (cells 10 and 25) improves the results. Increasing N further does not lead to any great improvements, since the model error statistics seems to be well represented with 100 ensemble members. Increasing the number of observations leads to some improvements in the results, but moderately after two observations have been introduced. This is probably due to the 1D structure of the model, and the fact that the parameters are distributed equal for all grid cells. Most of the information about the true state seems to be contained in a few observations of NO_2.

In Fig. 3, the probability of exceeding 100 ug/m (as an example), and in Fig. 4, the number of unique ensemble members, is shown as a function of time (hours) for the run with N = 25 and obs. of NO_2 in cell 10 only.

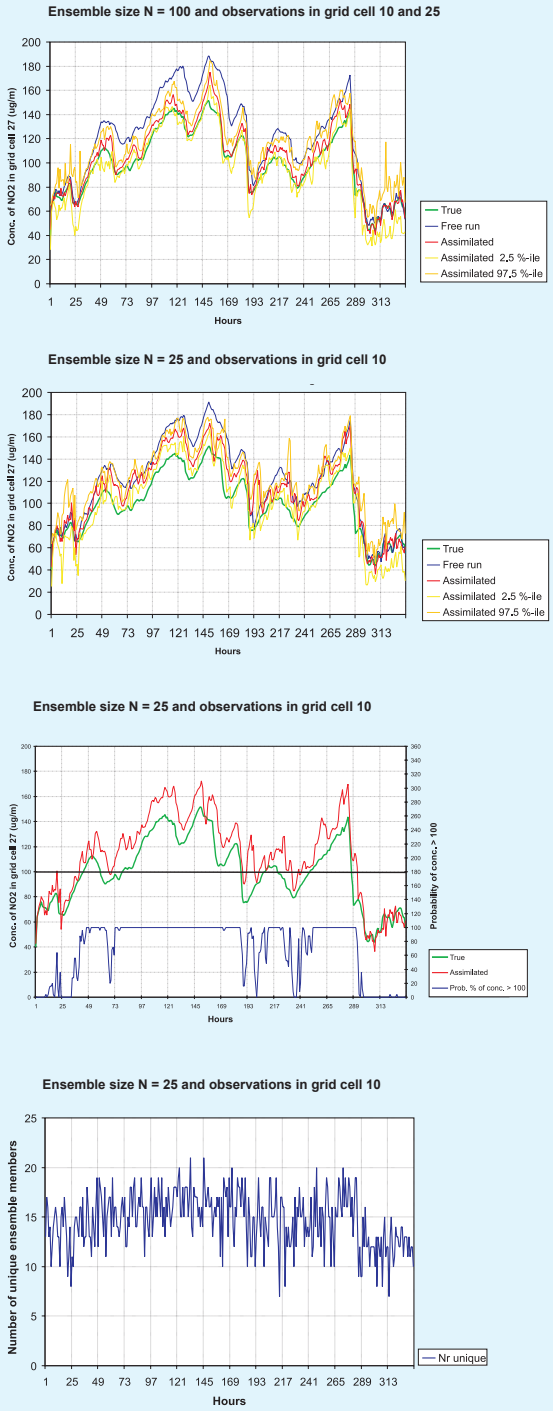
Conclusions

The SIR-method seems to work well on the 1D advection-diffusion model with photochemistry (1) reducing both bias and uncertainty if observations of NO_2 are available.

Lorentz (Student-t) likelihood functions seems to give the best results.

For both ensemble sizes N = 25 and N = 100 there were no ensemble collapses (few unique ensemble members) during the test runs.

Figures 1-4: Results for grid cell 27



References

Bott, A. (1989) A positive definite advection scheme obtained by non-linear renormalization of the advective fluxes. *Mon. Weather Rev. 117, 1006-1015 and 2633-2636.*

Box, G.E.P and G.C. Tiao (1992) Bayesian Inference in Statistical Analysis, *Wiley Classics Library Ed., New York.*

Doucet, A., de Freitas, N., Gordon N. (eds.) Sequential Monte Carlo methods in practice, *Springer Verlag 2001.*

Slørdal, L.H., Walker, S.E., Solberg, S. (2003) The urban air dispersion model EPISODE applied in AirQUIS2003. Technical description. *Kjeller, Norwegian Institute for Air Research (NILU TR 12/2003).*

Van Leeuwen, P. J. (2003) A variance minimizing filter for large scale applications. *Mon. Weather Rev. 131, 2071-2084.*