Simulated Experiments with Sequential Importance Resampling on a 1D Advection-Diffusion Model using Pseudo Observations

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. Simulated experiments, defining a true state 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 and Lorentz (Student's t) distributions, are also analysed.

Model description

The 1D model tested is:

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

where c is a space (x) and time (t) varying concentration (μg/m) of some species, u is the wind speed, k_X a turbulent eddy diffusivity coefficient, and q an assumed emission. In (1) boundary conditions and initial conditions are given by c(x,t) = c_B for x = 0 and x = nΔx and c(x,0) = 0. 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, q and c_B, and separate operators for advection (*Bott, 1989*) and diffusion (fully explicit scheme).

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:

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

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 ($\Delta T_{10m\text{-}2m}$) is provided from a meteorological station close to Oslo, Norway. A meteorological preprocessor is used to calculate horizontal turbulence intensities σ_v and diffusion coefficients k_X as $0.1\cdot\Delta x\cdot\sigma_v$ (*Slørdal et al., 2003*). Emissions (q) and background concentrations (c_B) are set equal to 10^{-3} μg/m·s and 10 μg/m respectively for all hours.

The model state vector **x** is defined as the concentration grid vector **c** consisting of 50 state variables $x_i = c_i$ for i = 1,...,50. In order to create the initial ensemble and update the ensemble from one time step to the next, actual input parameters u, k_X, q and 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 of the above parameters are defined using the expectance values and an assumed bias factor f_b as follows: $u^t = E(u)\cdot f_b$, $k_x^t = E(k_X)\cdot f_b$ and $q^t = E(q)/f_b$, where f_b = 1.2 (20% bias). The true background values are always assumed to be unbiased, i.e., $c_B^t = E(c_B)$. Pseudo-observations 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.

Results

Hourly concentration values from grid cell 27 are shown in Figs. 1-4. Only the tests performed with the Lorentz distribution are shown here. Generally it was found that this gave more stable and consistent improvements than using a Gaussian distribution function. The assimilated concentrations (red) lies consistently closer to the true concentrations (blue) than the unassimilated concentrations (green), although the improvement varies with time. When it is small or negative it is due to ensemble collapse, i.e., that there are only a few unique members in the ensemble. 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 did 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 to more than 2 does not improve results significantly. 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 is then apparently contained in a few observations.

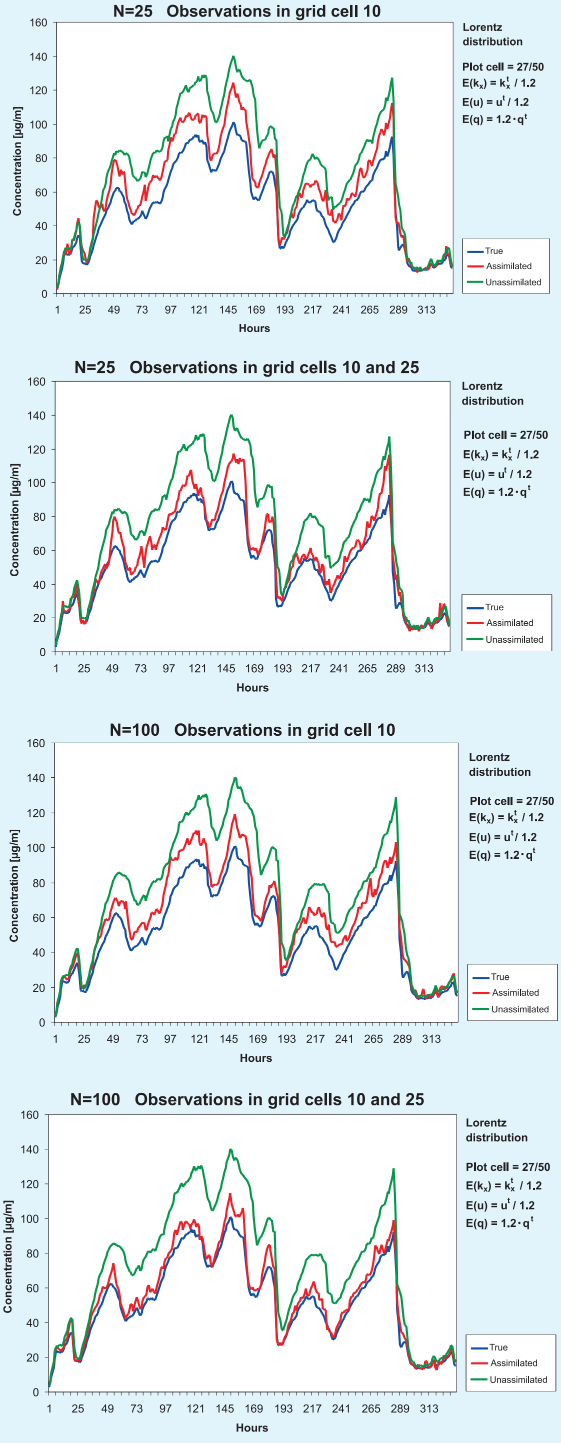
Conclusion

The SIR-method seems to work well on the 1D model (1) reducing both bias and uncertainty if observations are available. The simulated experiments performed indicate that most improvement is achived with a modest ensemble size of between 25 and 100 members and only 1 or 2 observations.

Our other experience of using the SIR-method on this model can be summarized as follows:

- Lorentz (Student's t) likelihood functions give generally better and more consistent results than Gaussian functions.
- If more observations are introduced, a larger ensemble size is needed to obtain improved results and to avoid ensemble collapse.

Fig. 1-4: Plots of hourly model concentrations (grid cell 27)



N=25 Observations in grid cell 10

N=25 Observations in grid cells 10 and 25

N=100 Observations in grid cell 10

N=100 Observations in grid cells 10 and 25

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, de Freitas, Gordon (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.*