Some theoretical aspects of source and parameter estimation in atmospheric transport and chemistry

Marc Bocquet

(bocquet@cerea.enpc.fr)

Victor Winiarek, Mohammad Reza Koohkan, Lin Wu . . .

CEREA, École des Ponts ParisTech and EDF R&D
Université Paris-Est and INRIA

Outline

Context: Atmospheric constituent versus meteorology

- Numerical weather forecast:
  - The global models are weakly non-linear but chaotic.
  - They do not depend on many parameter forcing fields (radiation, friction).
  - Quite accurate at global scale.
  - An inverse modelling problem on the initial condition (short windows).

- [Offline] chemical and transport forecast:
  - They are potentially strongly nonlinear but non-chaotic.
  - They depend on several parameter forcing fields (emissions, boundary conditions) and many uncertain parameters (kinetic rates, species microphysical parameters, transport subgrid parametrisation, etc.).
  - Quite uncertain.
  - An inverse modelling problem on the initial condition and many forcing fields.

Context: Atmospheric constituent versus meteorology

- Atmospheric constituent data assimilation is more of an inverse modelling game because:
  - we may be interested in the forcing/parameters themselves,
  - and successful forecasts rely on an accurate estimation of the forcings.

- Most of the current data assimilation schemes can be applied to either subjects (OI, 3D-Var, EnKF, 4D-Var). However, my vote goes to the smoothers (4D-Var, ensemble Kalman smoothers with weakly nonlinear physics/chemistry, iterative ensemble Kalman smoothers, 4D-En-Var, etc.)

- The background statistics are more uncertain and difficult to build in atmospheric constituent data assimilation.

Successful data assimilation: It's all about controlling the errors

- Problems in atmospheric constituent data assimilation:
    - Our observations are noisy
    - Our models are wrong (biased at the very least)
    - Even when they are fine, observations and models do not tell the same story!

i.e. representativeness errors are especially strong in this field.

- So successful data assimilation and especially inverse modelling is all about errors!

- Need to account for / estimate those errors in order to properly estimate control parameters.

Mathematical tools to correct/estimate the errors

- Statistical methods for hyperparameter estimation (parameters of **R** and **B**):
  - Maximum likelihood [Dee, 1995], [Desroziers and Ivanov, 2001],
  - χ^2 [Tarantola, 1987], [Ménard et al., 2000] ,
  - L-curve [Hansen, 1992], [Bocquet and Davoine, 2007],
  - statistical diagnostics: [Desroziers et al., 2005], [Schwinger and Elbern, 2010],
  - (generalised) cross-validation [Whaba, 1990],
  - online variational estimation [Doicu et al, 2010]

For CO_2 fluxes estimation, discussed in: [Michalak et al., 2005], [Wu et al, 2013]

- Estimating the parameters of model error parametrisations: a powerful paradigm when affordable [Bocquet, 2012], [Koohkan and Bocquet, 2012]
  - Context: A deterministic model full of uncertain parameters
  - Jointly estimate the state variables as well as the uncertain parameters.
  - Overfit is possible. Still might lead to a powerful forecasting tool.

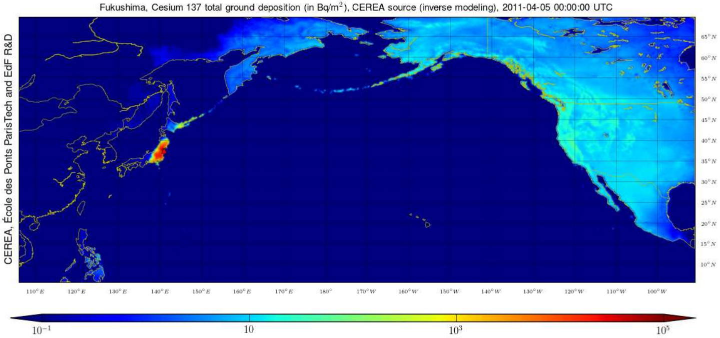
Outline

The Fukushima Daiichi accident

- Chronology: March 12: R 1 venting + explosion; March 13-14: R 3 venting + explosion; March 15: R 2 venting + explosion; March 20-22: R 2 R 3 spraying - smokes.

$\longrightarrow$ Source term of major interest for risk/health agencies, NPP operators

Observations of the Fukushima atmospheric dispersion



- Available data:
  - Very few observations of activity concentrations in the air: A few hundreds of observations over Japan publicly released.
  - Several thousands of observations from the (far away) CTBO IMS network.
  - Activity deposition: a few hundreds, but more difficult to exploit (mainly ^{137}Cs).
  - Hundreds of thousands of gamma dose measurements available.

Reconstruction of the Fukushima Daiichi source term

- Using three ($d = 3$) heterogeneous datasets:
  - Activity concentrations in the air,
  - Daily measurements of fallout,
  - Total cumulated deposits: densely distributed in space but no information in time.
- Yet, too few observations so that the inversion highly depends on the background.

- Retrieval of the cesium-137 source term $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_{504})$ ($\Delta t = 1\text{h}$) using

$$\mathscr{J} = \frac{1}{2}(\mu - \mathbf{H}\sigma)^{\mathrm{T}}\mathbf{R}^{-1}(\mu - \mathbf{H}\sigma) + \frac{1}{2}\sigma^{\mathrm{T}}\mathbf{B}^{-1}\sigma, \qquad \sigma \geq \mathbf{0} \tag{1}$$

where $\mathbf{R}_i = r_i^2 \mathbf{I}_{d_i}$ is the submatrix of $\mathbf{R}$ related to data set i, $\mathbf{B} = m^2 \mathbf{I}_N$.
$\mathbf{H}$: Jacobian matrix of the atmospheric transport model.

- $N_d + 1$ hyper-parameters to estimate simultaneously.

- Estimation method: maximisation of the non-Gaussian likelihood.

Non-Gaussian maximum likelihood principle

- Non-Gaussian maximum likelihood:

$$p(\mu|r_1,\ldots,r_{N_d},m) = \frac{e^{-\frac{1}{2}\mu^{\mathrm{T}}\left(\mathbf{HBH}^{\mathrm{T}}+\mathbf{R}\right)^{-1}\mu}}{\sqrt{(2\pi)^d|\mathbf{HBH}^{\mathrm{T}}+\mathbf{R}|}} \times \int_{\sigma\geq 0} \frac{e^{-\frac{1}{2}(\sigma-\sigma_{\mathrm{BLUE}})^{\mathrm{T}}\mathbf{P}_{\mathrm{BLUE}}^{-1}(\sigma-\sigma_{\mathrm{BLUE}})}}{\sqrt{(\pi/2)^N|\mathbf{P}_{\mathrm{BLUE}}|}}\mathrm{d}\sigma\,, \quad (2)$$

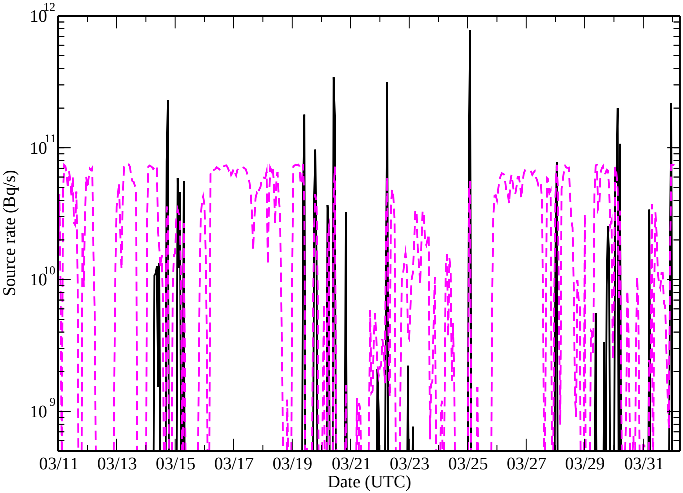
with:

$$\sigma_{\mathrm{BLUE}} = \mathbf{BH}^{\mathrm{T}}\left(\mathbf{HBH}^{\mathrm{T}}+\mathbf{R}\right)^{-1}\mu\,, \quad (3)$$

$$\mathbf{P}_{\mathrm{BLUE}} = \mathbf{B}-\mathbf{BH}^{\mathrm{T}}\left(\mathbf{HBH}^{\mathrm{T}}+\mathbf{R}\right)^{-1}\mathbf{HB}\,. \quad (4)$$

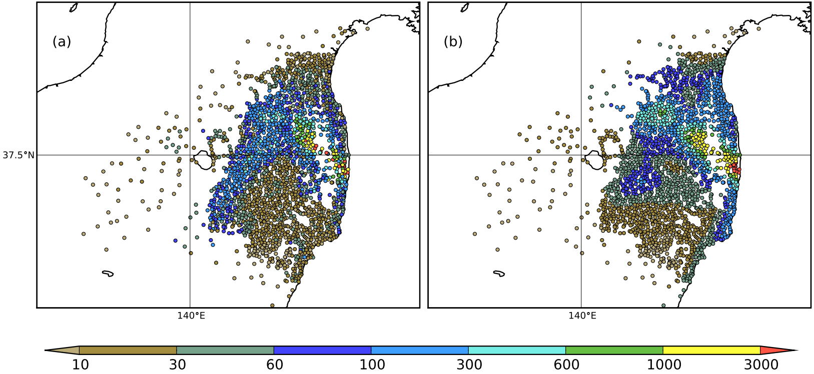
- Integral solved by Geweke-Hajivassiliou-Keane simulator (fine with several thousand variables).

Inversion results (caesium-137)



Total reconstructed activity: 13 PBq

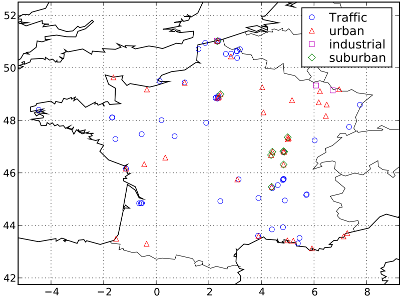
Deposition map reanalysis



Deposition measurements map (June 2011) - Reanalysis using three datasets

Outline

Inverse modelling of carbon monoxide fluxes at regional scale



▶ Using the French 600-stations BDQA network: hourly measurements of CO concentrations at about 80 stations.

▶ Observations highly impacted by representativeness errors (traffic, urban stations).

▶ Great number of observations (about 10^5 assimilated here, 5×10^5 used for validation).

▶ Control space: fluxes and volume sources parameterised with about 70×10^3 variables at $0.25^\circ \times 0.25^\circ$ resolution.

$\longrightarrow$ Even in this linear physics context, 4D-Var is a method of choice.

4D-Var

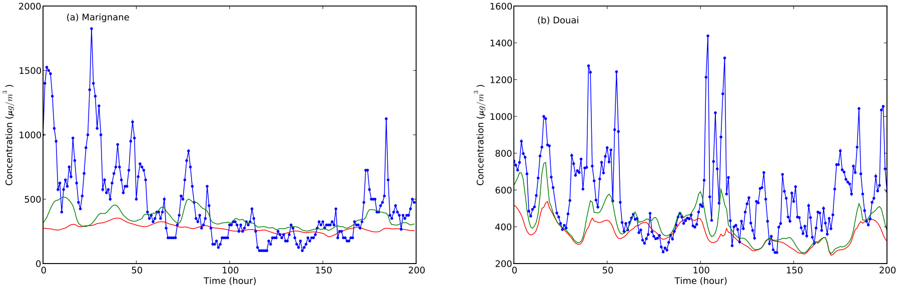
- Gradient obtained from adjoint approximated by the discretisation of the continuous adjoint model [Davoine & Bocquet, 2007; Bocquet, 2012].
- Background: EMEP inventory over Europe with an uncertainty of about 100%.
- Cost function:

$$
\begin{aligned}
\mathscr{J}(\alpha) \quad = \quad & \frac{1}{2}\sum_{h=0}^{N_\alpha-1}(\alpha_h-\mathbf{1})^{\mathrm{T}}\mathbf{B}_{\alpha_h}^{-1}(\alpha_h-\mathbf{1}) \\
& +\frac{1}{2}\sum_{k=0}^{N}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k)^{\mathrm{T}}\mathbf{R}_k^{-1}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k) \\
& +\sum_{k=1}^{N}\phi_k^{\mathrm{T}}(\mathbf{c}_k-\mathbf{M}_k\mathbf{c}_{k-1}-\Delta t\mathbf{e}_k)
\end{aligned}
\tag{5}
$$

- α: control vector of scaling parameters that multiply the first guess.
- Observation (representativeness) errors iteratively re-scaled by χ^2 diagnosis.

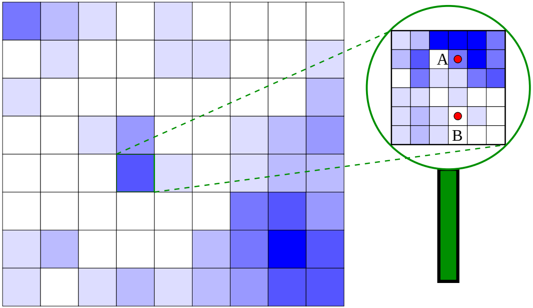
Results of (traditional) 4D-Var

	$\overline{C}$	$\overline{O}$	RMSE	C.Pear.	FA2	FA5
Simulation (01/01–02/26 2005)	303	662	701	0.16	0.52	0.90
Forecast (02/26–03/26 2005)	267	642	648	0.13	0.47	0.88
Optimisation of α	396	662	633	0.36	0.59	0.92
Forecast with optimal α	343	642	589	0.33	0.53	0.90



- Tremendous impact of representativeness errors!

Coupling 4D-Var with a simple statistical subgrid model



▶ We would like to take into account the impact of nearby sources that generate peaks on the CO concentration recordings:

$$\varepsilon_{\mathrm{rep}} \simeq \xi \cdot \mathbf{\Pi e} \quad \longrightarrow \quad \mathbf{y} = \mathbf{Hc} + \xi \cdot \mathbf{\Pi e} + \widehat{\varepsilon}. \tag{6}$$

ξ: set of statistical coefficients (influence factors).

Coupling 4D-Var with a simple statistical subgrid model

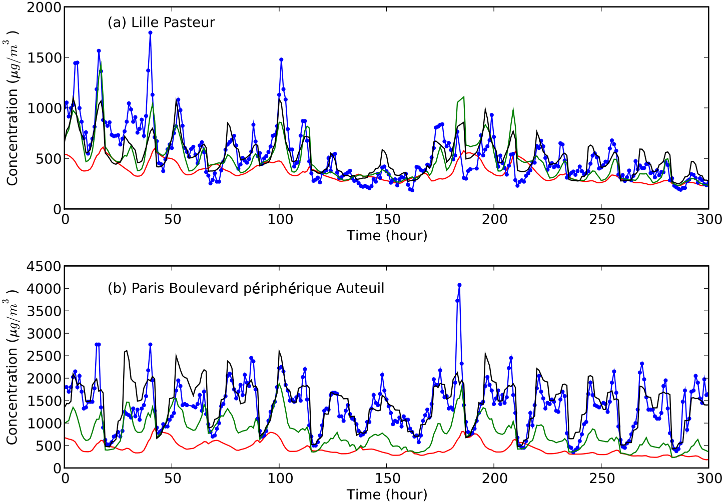
- Cost function of 4D-Var-ξ:

$$\begin{aligned}
\mathscr{J}(\alpha,\xi) &= \frac{1}{2}\sum_{h=0}^{N_\alpha-1}(\alpha_h-\mathbf{1})^{\mathrm{T}}\mathbf{B}_{\alpha_h}^{-1}(\alpha_h-\mathbf{1}) \\
&+\frac{1}{2}\sum_{k=0}^{N}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k-\xi\cdot\mathbf{\Pi}\mathbf{e}_k)^{\mathrm{T}}\widehat{\mathbf{R}}_k^{-1}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k-\xi\cdot\mathbf{\Pi}\mathbf{e}_k) \\
&+\sum_{k=1}^{N}\phi_k^{\mathrm{T}}(\mathbf{c}_k-\mathbf{M}_k\mathbf{c}_{k-1}-\Delta t\mathbf{e}_k)\,. \qquad (7)
\end{aligned}$$

- $\widehat{\mathbf{R}}$ is residual error covariance matrix (smaller than $\mathbf{R}$).

$$\mathbf{R}=\mathrm{E}\left[\varepsilon\varepsilon^{\mathrm{T}}\right]=\xi\cdot\mathbf{\Pi}\mathrm{E}\left[\mathbf{e}\mathbf{e}^{\mathrm{T}}\right]\mathbf{\Pi}^{\mathrm{T}}\cdot\xi^{\mathrm{T}}+\widehat{\mathbf{R}}. \qquad (8)$$

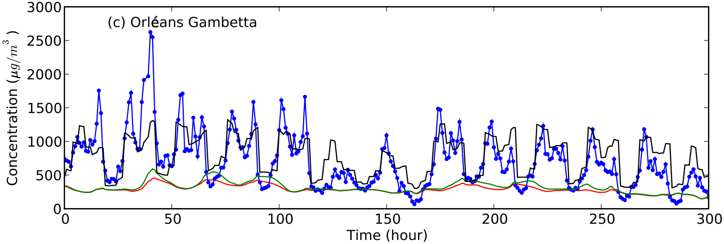
Results of 4D-Var-ξ: Profiles (1/4)



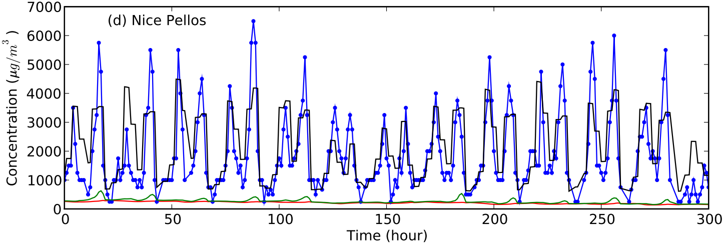
$\xi_i = 0.6$ h.

$\xi_i = 2.7$ h.

Results of 4D-Var-ξ: Profiles (2/4)



$\xi_i = 11.9$ h.



$\xi_i = 45.8$ h.

Results of 4D-Var-ξ: Scores (3/4)

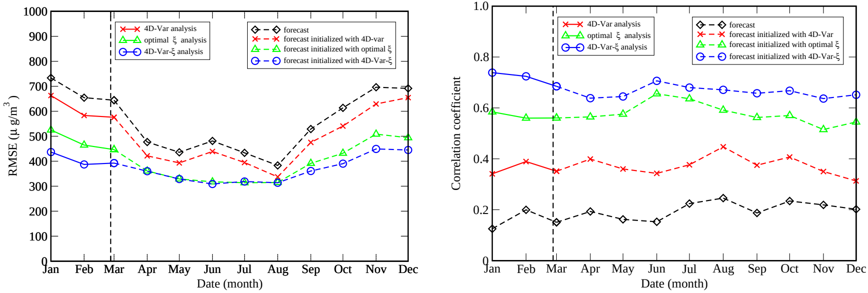
- Skills:

	$\overline{C}$	$\overline{O}$	RMSE	C.Pear.	FA2	FA5
Simulation (01/01–02/26 2005)	303	662	701	0.16	0.52	0.90
Forecast (02/26–03/26 2005)	267	642	648	0.13	0.47	0.88
Optimisation of α	396	662	633	0.36	0.59	0.92
Forecast with optimal α	343	642	589	0.33	0.53	0.90
Optimisation of ξ	615	662	503	0.57	0.73	0.96
Forecast with optimal ξ	574	642	451	0.56	0.76	0.97
Coupled optimisation of ξ, α	671	662	418	0.73	0.79	0.97
Forecast with optimal ξ, α	631	642	340	0.68	0.81	0.98

- We found an increase of 9% in the French CO total emission. Consistent with satellite retrieval for Western Europe.

Results of 4D-Var-ξ: Forecast (4/4)

- Validation of a 10-month forecast after the 8-week assimilation window (2005)



- Skills almost as good in the forecast period as in the assimilation time window!
- Seasonal effects impacting scores.

Outline

Future plans

- Development of an EnVar method, the iterative ensemble Kalman smoother (IEnKS, [Bocquet and Sakov, 2013]) that
  - + performs a variational analysis over a time data assimilation window
  - + has flow-dependent error estimation
  - + does not use an explicit tangent linear/adjoint
  - - - requires localisation (no free lunch)
  - +/- weak-constraint formalism under development

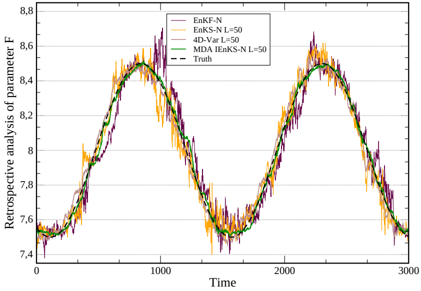
- Solves the Bayesian problem with minimal Gaussian assumptions (has the potential to outperform 4D-Var and EnKF in all regimes)

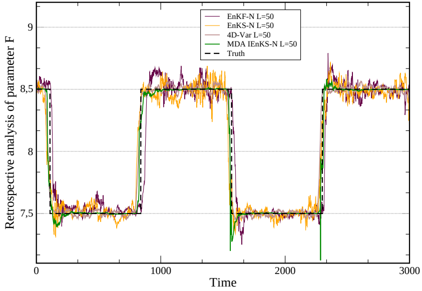
- Potentially well suited for joint state and parameter estimation, with nonlinear dependencies.

Future plans

- The augmented state formalism is convenient for the IEnKS, and offers an easy implementation of technically challenging data assimilation problems.
- Lorenz '95 with joint estimation of the forcing parameter F (41 variables): RMSEs.

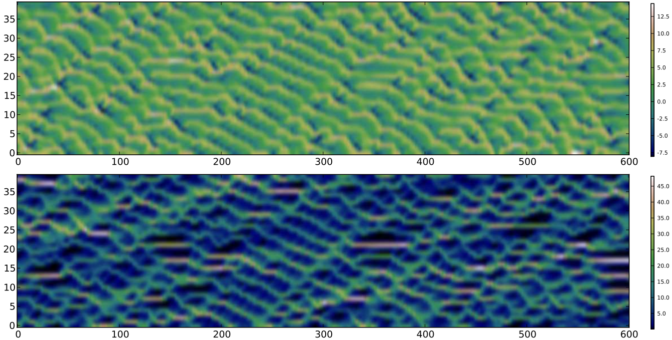
Method / F profile	Sinusoidal	Step-wise
EnKF	0.063	0.079
EnKS L=50	0.040	0.063
4D-Var L=50	0.030	0.045
MDA IEnKS L=50	0.020	0.031





Future plans

- Development of low-order models that couple a Lorenz model and a chemical model.



Lorenz '95 coupled to a tracer model.

- The goals of this study will be:
  - to probe the added value of online/coupled models DA vs offline models DA,
  - to probe the added value of joint state and parameter estimation, integrated data assimilation,
  - to assess the nonlinearity and the numerical cost of these games.

References I

- Bocquet, M., 2012a. An introduction to inverse modelling and parameter estimation for atmospheric and oceanic sciences. In: Blayo, E., Bocquet, M., Cosme, E. (Eds.), Advanced data assimilation for geosciences. Oxford University Press, Les Houches school of physics.
- Bocquet, M., 2012b. Parameter field estimation for atmospheric dispersion: Application to the Chernobyl accident using 4D-Var. Q. J. Roy. Meteor. Soc. 138, 664–681.
- Bocquet, M., Sakov, P., 2013. Joint state and parameter estimation with an iterative ensemble Kalman smoother. Nonlin. Processes Geophys. 0, 0–0, in press.
- Davoine, X., Bocquet, M., 2007. Inverse modelling-based reconstruction of the Chernobyl source term available for long-range transport. Atmos. Chem. Phys. 7, 1549–1564.
- Dee, D. P., 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. Mon. Wea. Rev. 123, 1128–1145.
- Desroziers, G., Berre, L., Chapnik, B., Poli, P., 2005. Diagnosis of observation, background and analysis-error statistics in observation space. Q. J. Roy. Meteor. Soc. 131, 3385–3396.
- Desroziers, G., Ivanov, S., 2001. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. Q. J. Roy. Meteor. Soc. 127, 1433–1452.
- Doicu, A., Trautmann, T., Schreier, F., 2010. Numerical Regularization for Atmospheric Inverse Problems. Springer and Praxis publishing.
- Elbern, H., Strunk, A., Schmidt, H., Talagrand, O., 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. Atmos. Chem. Phys. 7, 3749–3769.

References II

- Hansen, P. C., 1992. Analysis of discrete ill-posed problems by means of the L-curve. SIAM Review 34, 561–580.
- Koohkan, M. R., Bocquet, M., 2012. Accounting for representativeness errors in the inversion of atmospheric constituent emissions: Application to the retrieval of regional carbon monoxide fluxes. Tellus B 64, 19047.
- Saunier, O., Mathieu, A., Didier, D., Tombette, M., Quélo, D., Winiarek, V., Bocquet, M., 2013. An inverse modeling method to assess the source term of the Fukushima nuclear power plant accident using gamma dose rate observations. Atmos. Chem. Phys. 0, 0–0, in press.
- Tarantola, A., 1987. Inverse Problem Theory. Elsevier.
- Vogel, C. R., 2002. Computational Methods for Inverse Problems. SIAM, Frontiers in Applied Mathematics.
- Wahba, G., 1990. Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics 59. SIAM, Philadelphia.
- Winiarek, V., Bocquet, M., Duhanyan, N., Roustan, Y., Saunier, O., Mathieu, A., 2013. Estimation of the caesium-137 source term from the Fukushima Daiichi nuclear power plant using a consistent joint assimilation of air concentration and deposition observations. Atmos. Env. 0, 0–0, in press.
- Winiarek, V., Bocquet, M., Saunier, O., Mathieu, A., 2012. Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant: Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant. J. Geophys. Res. 117, D05122.