Lecture 1
Data assimilation and tropospheric chemistry

H. Elbern

Rhenish Institute for Environmental Research

at the University of Cologne

1

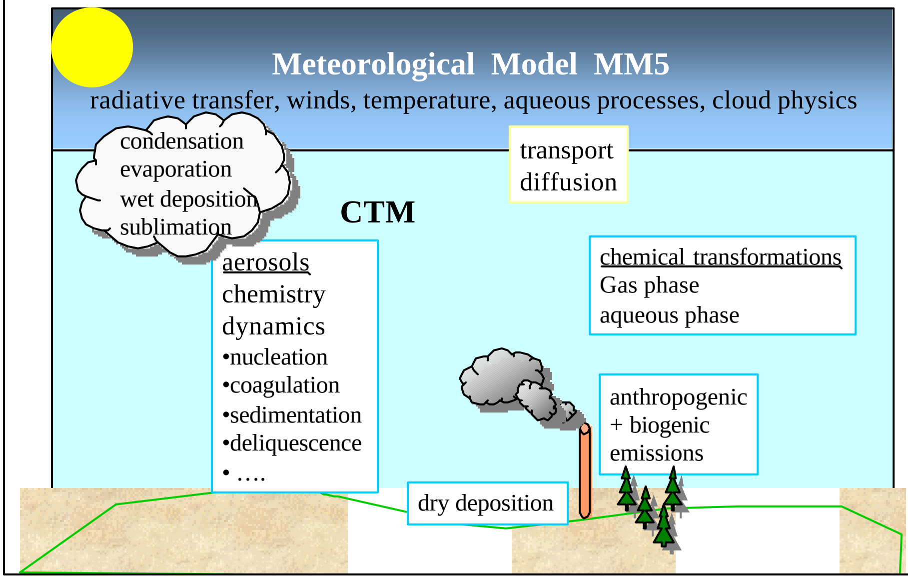
Special challenges of **tropospheric** chemistry data assimilation

The following problems prevail:

1. strong influence of manifold processes including *emissions* and *deposition*
2. *spatially highly variable* "chemical regimes"
3. chemical state observability (= "analyseability") hampered by *manifold hydrocarbon* species
4. consistency with *heterogeneous* data sources: satellite data and in situ observations

2

Processes in a complex chemistry-transport model
(*EURopean Air pollution Dispersion model EURAD)*



ESA summer school on data assimilation, Frascati 2003

General approach for advanced data assimilation systems

- combine *consistently* tropospheric and surface information
  - observations with
  - tropospheric chemistry models and
  - a priory information (climatologies, forcing fields, …)
- to provide an optimal chemical state estimate on a regular grid

This invokes the application space-time data assimilation algorithms preserving the **BLUE** property (**B**est **L**inear **U**nbiased **E**stimator)

4

Advanced spatio-temporal methods used in tropospheric chemistry data assimilation

- 4-dim variational data assimilation (**4D-VAR**, and versions: 4D physical space statistical analysis system (Amodei, 1995; Courtier, 1997); 4D-var incremental (Courtier et al, 1994))
- reduced complexity versions of the Kalman Filter: Reduced Rank Square root KF (**RRSQKF**, Verlaan and Heemink, 1997) , ensemble KF (Evensen, 1994), SEEK, SEIK, (Verron et al., 1999)

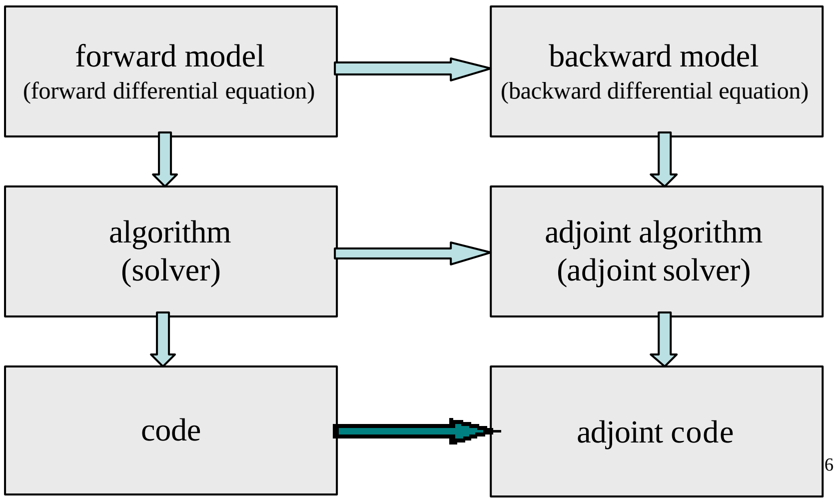
Spacio-temporal BLUEs applied in tropospheric chemistry data assimilation:

4D var: (Elbern and Schmidt, 1999, 2001)

RRSQKF (van Loon et al, 2000)

Remark: *3D BLUE algorithm analyses like those from Optimal Interpolation, once ingested into a model, do not result in a 4D BLUE analysis!*

Construction of the adjoint code



Transport-diffusion-reaction equation and its adjoint

Tendency Equations

direct chemistry transport equation

$$\frac{\partial c_i}{\partial t} + \nabla \cdot (\mathbf{v} c_i) - \nabla \cdot (\rho \mathbf{K} \nabla \tfrac{c_i}{\rho}) - \Sigma_{r=1}^{R} \left(k(r) \left(s_i(r_+) - s_i(r_-) \right) \Pi_{j=1}^{U} c_j^{s_j(r_-)} \right) = E_i + D_i$$

c_i	concentration of species i	c_i^*	adjoint of concentration of species i
$\mathbf{v}$	wind velocity	s	stoichiometric coefficient
$k(r)$	reaction rate of reaction r	$\mathbf{K}$	diffusion coefficient
U	number of species in the mechanism	R	number of reactions in the mechanism
E_i	emission rate of species i (source)	D_i	deposition rate of species i (sink)

adjoint chemistry transport equation

$$-\frac{\partial \delta c_i^*}{\partial t} - \mathbf{v} \nabla \delta c_i^* - \frac{1}{\rho} \nabla \cdot (\rho \mathbf{K} \nabla \delta c_i^*) + \Sigma_{r=1}^{R} \left(k(r) \frac{s_i(r_-)}{c_i} \Pi_{j=1}^{U} \bar{c}_j^{\,s_j(r_-)} \Sigma_{n=1}^{U} \left(s_n(r_+) - s_n(r_-) \right) \delta c_n^* \right) = 0$$

7

1. Strong influence of emissions and deposition

Does it really suffice to optimise the initial concentration values in the troposphere?
No, it does not!

Which other parameters must be optimised to improve analysis and forecast skill?
A rule of thumb: parameter with maximal (paucity-of-knowledge * impact)

With valid Gaußian error characteristics and tangent linearisation a more precise formulation can be given:

8

Singular value analysis

unit constraint (scalar product): $(e, Ce) = 1$

maximise Raleigh quotient:

$$\max_{e_i}\left(\frac{(\mathbf{PM}(t_1,t_0)e_i, \mathbf{EPM}(t_1,t_0)e_i)}{(e_i, \mathbf{C}e_i)}\right)$$

⇔maximise

$$J(e) = (\mathbf{PM}(t_1,t_0)e_i, \mathbf{EPM}(t_1,t_0)e_i) - \lambda(e^T\mathbf{C}e - 1)$$

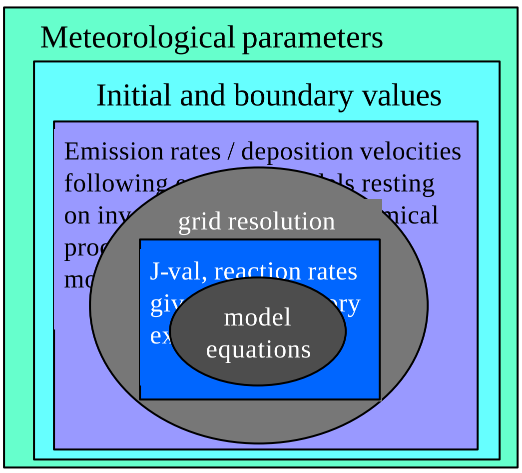
⇒generalised EV problem

$$\nabla_e J(e) = 0 = \mathbf{M}^T(t_1,t_0)\mathbf{P}^T\mathbf{EPM}(t_1,t_0)e = \lambda\mathbf{C}e$$

$\mathbf{e}_i$	perturbation vector of potential optimisation parameters: initial values boundary values, emission rates, deposition velocities
C	norm inducing pos. def., sym. operator at initial time t_0 (Mahalanobis)
M	tangent linear model
E	norm inducing pos. def., sym. operator at optimisation time t_1
(**P**	projection operator, extinguishing areas or species outside focus)
λ	Lagrange parameter and generalised eigenvalues

9

Which model parameters are amenable to inverse modelling based optimisation?
(Which of those parameters maximise the Raleigh quotient?)



10

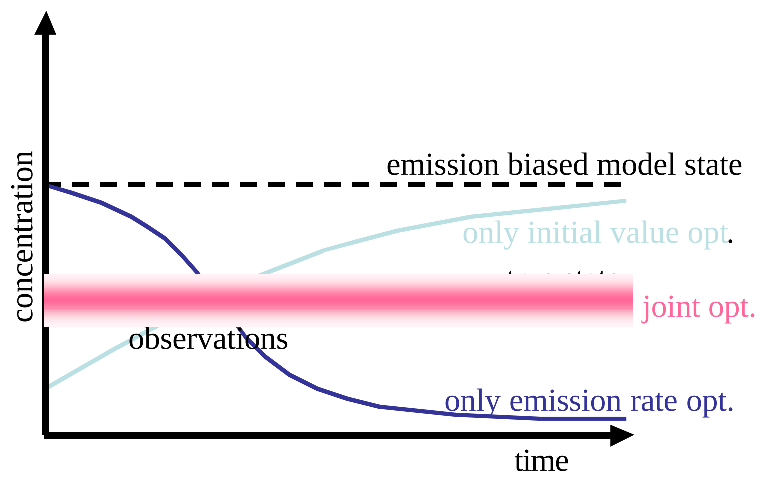
Synthesis: principal error sources to be controlled by inversion

- Meteorological parameters: 2 major error sources are **boundary layer height** (residual layer), and cloud processes (especially **quantitative precipitation**)
- Initial/boundary values: error **impact below 2 days** in the lower troposphere for most pollution conditions, but to be determined anew for each run
- anthropogenic emissions: **large impact**, less variable than initial values, i.e. daily optimisation not justified if not for external reasons (working day, Sunday) or weather changes (heating period). Imposition of reasonable constraints on the optimisation is advisable (diurnal cycle)

Conclusion for tropospheric CTMs: Initial values and emission rates should be joint optimisation parameters, although not necessarily acting on the same time scale. Separate meteorological data assimilation required.

11

Hypothesis: initial state and emission rates are least known



12

In the troposphere for **emission rates** the product (*paucity of knowledge*importance)* is high

Emission Rate Optimization

minimize cost function

$J(\mathbf{x}(t_0),\mathbf{e}) = \frac{1}{2}(\mathbf{x}^b(t_0)-\mathbf{x}(t_0))^T\mathbf{B}_0^{-1}(\mathbf{x}^b(t_0)-\mathbf{x}(t_0)) +$	deviations from background initial state
$\frac{1}{2}\int_{t_0}^{t_N}(\mathbf{e}_b(t)-\mathbf{e}(t))^T\mathbf{K}^{-1}(\mathbf{e}_b(t)-\mathbf{e}(t))dt +$	deviations from a priori emission rates
$\frac{1}{2}\int_{t_0}^{t_N}\left(\mathbf{y}^0(t)-H[\mathbf{x}(t)]\right)^T\mathbf{R}^{-1}(\mathbf{y}^0(t)-H[\mathbf{x}(t)])dt$	model deviations from observations

$\mathbf{x}^b(t_0)$	background state at $t=0$
$\mathbf{x}(t)$	model state at time t
$\mathbf{e}_b(t_0)$	background emission rate at $t=0$
$\mathbf{e}(t)$	emission rate field at time t
$\mathbf{K}$	emission rate error covariance matrix
$H[\ \ ]$	forward interpolator
$\mathbf{y}^0(t)$	observation at time t
$\mathbf{B}_0$	background error covariance matrix

ESA summer school on data assimilation, Frascati 2003

Kalman filter, basic equations

(see lecture H. Eskes)

$$\mathbf{x}^f(t_i) = \mathbf{M}(t_i,t_{i-1})\mathbf{x}^a(t_{i-1}) + \eta \qquad (1)$$

$$\mathbf{P}_i^b = \mathbf{M}(t_i,t_{i-1})\mathbf{P}_i^a\mathbf{M}^T(t_i,t_{i-1}) + \mathbf{Q}$$

$$\mathbf{x}^a(t_i) = \mathbf{x}^b(t_i) + \mathbf{K}_i\mathbf{d}_i. \qquad (1)$$

$$\mathbf{K}_i := \mathbf{P}_i^b\mathbf{H}_i^T(\mathbf{H}_i\mathbf{P}_i^b\mathbf{H}_i^T + \mathbf{R}_i)^{-1} \quad \in \mathcal{R}^{n\times p}, \qquad (2)$$

and

$$\mathbf{P}_i^a = (\mathbf{I} - \mathbf{K}_i\mathbf{G}_i)\mathbf{P}_i^b. \qquad (3)$$

14

Reduced rank Kalman filter (basic idea)

Appoximate covariance matrices $P^{b,a}$ (n x n) by a product of suitably low ranked matrix $S^{b,a}$ (n x p), and p << n . Same procedure for the system noise matrix Q with T (n x r).

$$\mathbf{P} \approx \mathbf{S}\mathbf{S}^T, \qquad \mathbf{Q} \approx \mathbf{T}\mathbf{T}^T$$

The forecast step remains unchanged.

$$\mathbf{x}^f = \mathbf{M}\mathbf{x}^a$$

The forecast error covariance matrix rests on 2 x p model integrations only!

$$\mathbf{S}^f\mathbf{S}^{fT} = \mathbf{M}\mathbf{S}^a\mathbf{S}^{aT}\mathbf{M}^T + \mathbf{T}\mathbf{T}^T$$

The enlargement of p by r enforces periodic reductions of columns (with lowest ranked eigenvalues)

$$\mathbf{S}^f = [\mathbf{M}\mathbf{S}^a, \mathbf{T}]$$

ESA summer school on data assimilation, Frascati 2003

Reduced rank Kalman filter (calculus)

In practice, all calculations can be performed without actually calculating matrices P!

Positive semidefinitenes is maintained!

$$\Psi := \mathbf{H}\mathbf{S}$$

$$\mathbf{K} = \mathbf{S}^f\Psi^T(\Psi\Psi^T + \mathbf{R})^{-1}$$

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^f)$$

$$\begin{aligned} \mathbf{S}^a\mathbf{S}^{aT} &= (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{S}^f\mathbf{S}^{fT} \\ &= \mathbf{S}^f\left[\mathbf{I} - \Psi^T(\Psi\Psi^T + \mathbf{R})^{-1}\psi^T\right]\mathbf{S}^{fT} \end{aligned}$$

$$\mathbf{S}^{aT} = \mathbf{S}^f\left[\mathbf{I} - \Psi^T(\Psi\Psi^T + \mathbf{R})^{-1}\psi^T\right]^{1/2}$$

16

ESA summer school on data assimilation, Frascati 2003

2. Spatially highly variable “chemical regimes”

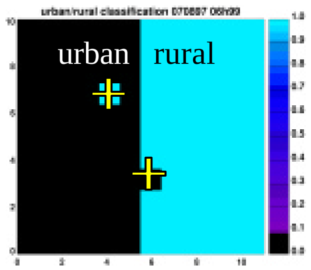
- boundary layer structures are generally fine scale as induced by the surface emission, deposition and land use texture.
- observational coverage from satellites and in situ observations is generally coarser than model resolution
- elements of “feature assimilation” to be introduced to exploit fine grid model information for assimilation

17

ESA summer school on data assimilation, Frascati 2003

Schematic example of anisotropic and inhomogeneous covariances

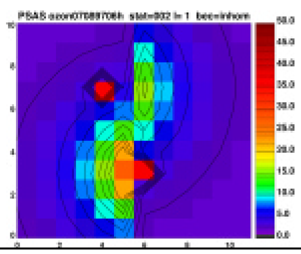
following Hoelzemann et al. (2001)



Anisotropy and inhomogeneity introduced by a generic decorrelation function $f_{\mu\nu}=1-a|u_\mu-u_\nu|$, $0<u<1$ being an index describing a feature strength, here urban emission characteristics vs rural. $0<a<1$ is a sensitivity parameter. The Balgovind formula is selected as homogeneous and isotropic “carrier” function.

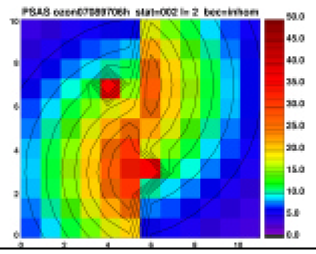
$$B_{\mu\nu} = e_b\left(1-\frac{r_{\mu\nu}}{L}\right)\exp\left(-\frac{r_{\mu\nu}}{L}\right)f_{\mu\nu}$$

2 observation sites

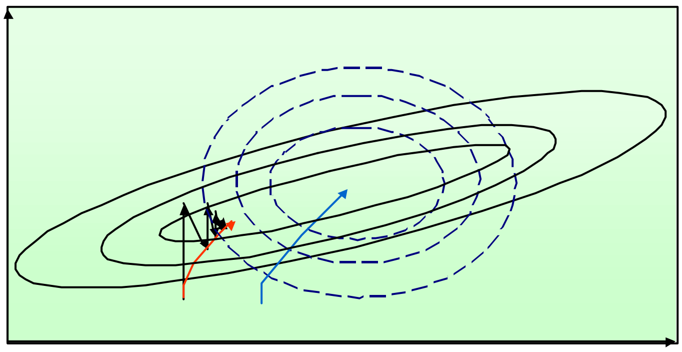


L is radius of influence, r distance between sites μ and ν, e_b is the observation error

correlation with L=1 (left) and L=3 (right)



Isopleths of the cost function and transformed cost function and minimisation steps



Minimisation by mere **gradients,** quasi-Newon method **L-BFGS** (Large dimensional Broyden Fletcher Goldfarb Shanno), and preconditioned (transformed) L-BFGS application

The gradient of the cost function and the processing of background error covariance matrices

direct model	$\frac{d\mathbf{x}}{dt} = \mathcal{M}(\mathbf{x}) + \mathbf{e}(t), \qquad \frac{d\delta\mathbf{x}}{dt} = \mathcal{M}'(\delta\mathbf{x}) + \delta\mathbf{e}(t)$	(1)
tangent linear model	$\delta\mathbf{x}(t_n) = \mathbf{M}(t_n, t_0)\delta\mathbf{x}(t_0) = \Pi_{i=0}^{1} \mathbf{M}(t_i, t_{i-1})\delta\mathbf{x}(t_0)$	(2)
adjoint model	$-\frac{d\delta\mathbf{x}^*(t)}{dt} - \mathcal{M}'^T(\delta\mathbf{x}^*(t)) = \mathbf{R}^{-1}(\mathbf{y}^0(t) - H[\mathbf{x}(t)])$.	(3)

gradient of the cost function

$$\boxed{\nabla_{[\mathbf{x}(t_0),\mathbf{e}]} J = -\mathbf{B}_0^{-1}(\mathbf{x}^b(t_0) - \mathbf{x}(t_0)) - \mathbf{K}^{-1}(\mathbf{e}^b(t) - \mathbf{e}(t)) - \Sigma_{m=0}^{N} \Pi_{i=1}^{m} \mathbf{M}^T(t_{i-1}, t_i)\mathbf{R}^{-1}(\mathbf{y}^0(t_m) - H[\mathbf{x}(t_m)])}$$

Find minimum of $J(\mathbf{x}(t_0), \mathbf{e})$ with $\nabla_{[\mathbf{x}(t_0),\mathbf{e}]} J$ by use of a minimization routine

The Preconditioning Problem

Hessian matrix = (analysis error covariance matrix)$^{-1}$

$$\nabla^2 J = \mathbf{B}_0^{-1} + H^T \mathbf{R}^{-1} H$$

the optimum is (linear case)

$$\mathbf{x}_{opt} - \mathbf{x}_b = (\mathbf{B}_0^{-1} + H^T \mathbf{R}^{-1} H)^{-1} \nabla J$$

practical problem:
$\mathbf{B}_0$ is generally ill–conditioned

Example:
2D grid 25×25
influence radius $L = 6$
gives

$$\mathrm{cond}(\mathbf{B}_0) \sim 10^9$$

21

The Preconditioning of the Minimisation Procedure

3 steps of external preconditioning (1) spatial (horizontal presently) (2) chemical (3) numerical

Procedure:
Transformation by a crudely factorized background error covariance matrix $\tilde{\mathbf{B}}_0$

Define:

$$\mathbf{B}_0 \approx \tilde{\mathbf{B}}_0 := \mathbf{B}_s \mathbf{B}_c \mathbf{B}_n$$

Spatial: $\mathbf{B}_s$ (Balgovind)

$$C(|\mathbf{r}|) = (1 + |\mathbf{r}|/L) \exp(-|\mathbf{r}|/L)$$

Chemical: $\mathbf{B}_c$ (from identical twin experiments)
for example

$$\mathbf{B}_c(O_3, NO) = \mathrm{diag}(10, 1)$$

Numerical: $\mathbf{B}_n$ (scaling experiments) $\mathbf{B}_n = \mathrm{diag}(\mathbf{x}_b^{-2})$
or, to ensure positive definiteness $\mathbf{B}_n = \log^2(\mathrm{diag}(\mathbf{x}_b^{-2}))$

22

Transformed cost function

define $\mathbf{d}_i := \mathbf{y}_i^o - \mathbf{H}\mathbf{M}(t_i, t_0)\mathbf{x}(t_0)$

$$\delta\mathbf{x}(t_0) := \mathbf{x}^b - \mathbf{x}(t_0)$$

transformation $\mathbf{v} := \mathbf{B}^{-1/2}\delta\mathbf{x}_b$

transformed cost function

$$J(\mathbf{v}) = 1/2\mathbf{v}^T\mathbf{v} + 1/2\sum_{m=0}^{N} \mathbf{d}_m^T\mathbf{R}^{-1}\mathbf{d}_m$$

transformed gradient of the cost function

$$\nabla_{\mathbf{v}} J(\mathbf{v}) = \mathbf{v} + \mathbf{B}^{1/2}\sum_{m=0}^{N} \mathbf{M}^T(t_m, t_0)\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}_m$$

Pro transformation:
minimisation problem is better conditioned
Contra:
strictly positive definite approximation to **B** required

Computation of inverse B, square root B, inverse square root B by (Sca)LAPACK eigenpair decomposition

23

Transformed Cost Function

define

$$\mathbf{d}_i := \mathbf{y}_i^0 - H\mathbf{M}^T(t_i, t_0)\mathbf{x}(t_0)$$
$$\delta\mathbf{x}(t_0) := \mathbf{x}^b(t_0) - \mathbf{x}(t_0)$$
$$\mathbf{v} := \sqrt{\mathbf{B}_0}\,\delta\mathbf{x}_b$$

transformed cost function

$$J(\mathbf{v}) = 1/2\mathbf{v}^T\mathbf{v} + 1/2\,\Sigma_{m=0}^{N}(\mathbf{d}_i)^T\mathbf{R}^{-1}\mathbf{d}_i dt,$$

transformed gradient

$$\nabla_{\mathbf{v}} J(\mathbf{v}) = -\mathbf{v} - \sqrt{\mathbf{B}_0}\,\Sigma_{m=0}^{N}\Pi_{i=1}^{m}\,\mathbf{M}^T(t_{i-1}, t_i)\mathbf{R}^{-1}\mathbf{d}_i dt,$$

Pro:
Minimisation problem is better conditioned

Contra:
strictly positive $\tilde{\mathbf{B}}_0$ required

Computation of inverse B, square root B, inverse square root B by (Sca)LAPACK eigenpair decomposition

4

3. Chemical state observability (="analysability")

Contrary to the troposphere, a few key species must be observed in the stratosphere to approximately identify the chemical state

Results from Observability Tests:

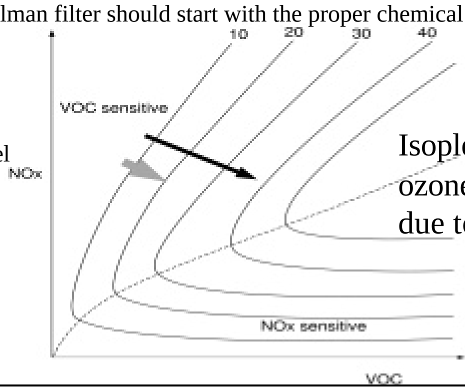
- To better analyse Chlorine species it was sufficient to add one of ClO, $ClONO_2$, $HOCl$ or $OClO$ to the standard data set.
- The BrO observation is needed to analyse Bromine species.
- Assimilation results are fully satisfying with the seven species of standard data set (N_2O, CH_4, HNO_3, H_2O, O_3, NO_2, BrO)
- With extended data set the algorithm manages to recover the reference run almost perfectly in the majority of cases.

25

Problem in the troposphere:
A huge number of volatile organic hydrocarbons (VOCs) exist, only very few, if any, are observed.

- Nitrogene oxides and numerous hydrocarbons act highly nonlinearly as precursors of ozone.
- Chemical conditions are either controlled by NOx or VOC deficit, delineating the "chemical regime".
- Both 4D-var and Kalman filter should start with the proper chemical regime.

EKMA diagram
(Empirical Kinetic Model Approach)



Isopleths of ozone production, due to NO2 and VOC

26

Cost function and gradient for relative background information of volatile organic compounds

Exploit the fact that

relative composition of emitted species from technical devices (motor cars) is much better known than absolute values

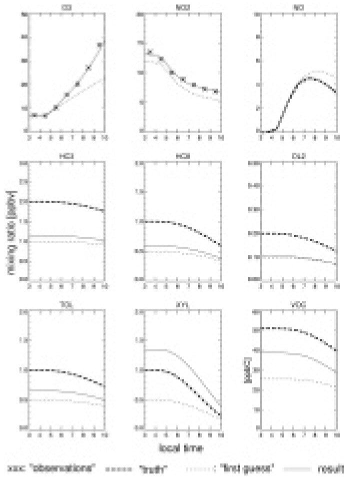
$\mathbf{c}(t_0)$ subset (VOC) of components of $\mathbf{x}(t_0)$
$\mathbf{c}^b(t_0)$ estimate of the background
$\bar{c}_0$ total sum of subset (VOC)

$$J_c(\mathbf{c}(t_0)) = \frac{1}{2}\left(\frac{\mathbf{c}(t_0)}{\bar{c}_0} - \frac{\mathbf{c}^b}{\bar{c}^b}\right)^T \mathbf{C}^{-1}\left(\frac{\mathbf{c}(t_0)}{\bar{c}_0} - \frac{\mathbf{c}^b}{\bar{c}^b}\right) \quad (1)$$

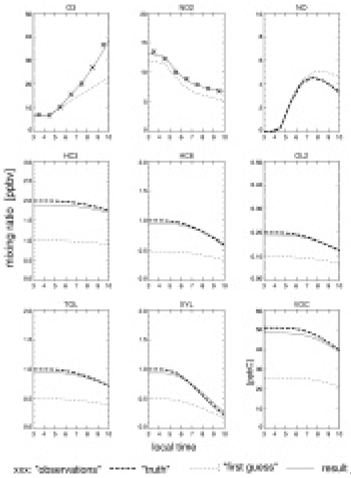
$$\frac{\partial J_c}{\partial c_i} = \mathbf{C}^{-1}_{(i,i)} \frac{(c_i - \frac{c_i^b}{\bar{c}^b})(\bar{c}_0 - c_i)}{\bar{c}_0^3} - \sum_{j \neq i} \mathbf{C}^{-1}_{(j,j)} c_j \frac{(c_j - \frac{c_j^b}{\bar{c}^b})}{\bar{c}_0^3} \quad (2)$$

Relative a priori information: the VOC problem, only O3 and NO2 observed

traditional approach,
no relative a priori information



novel approach, **with relative** a priori information



28

4. Consistency with heterogeneous data sources: satellite data and in situ observations

following Talagrand (1998) and Desroziers and Ivanov (2001)

Including background information, in situ observations, ozone profile retrievals and NO2 column retrievals, tropospheric information sources are especially heterogeneous.

Can consistency of the assimilation result ("analysis") be identified?
What is an improvement in the analysis?
Is an independent validation possible?
What is an improved forecasts?
What is the contribution of individual observation sources?

ESA summer school on data assimilation, Frascati 2003

Assumptions:

- Gaussian error distribution assumption sufficiently valid
- First guess not too far from "solution" (tangent-linear approximation must hold)
- A priori defined error covariances (background, observations)

Necessary condition for a posteriori validation: adjust B and R such that:

at the minimum:

$$J_{min} = 1/2 d^T (\mathbf{HBH^T} + \mathbf{R})^{-1} d$$

$d := y - Hx^a$

p number of observations

Expectation $\mathcal{E}[J_{min}] = p/2$

Variance $\mathcal{V}[J_{min}] = p/2$

The partition of costs in terms of all observations and the background at the final analysis

How do the partial costs in the cost function divide?

$$J(x^a) = J^o(x^a) + J^b(x^a)$$

partial costs of observations
(I_p identity matrix in observation space with p observations)

$$\mathcal{E}[J^o(x^a)] = 1/2\ \mathrm{tr}(\mathbf{I_p} \quad \mathbf{HK})$$

partial costs of background:

$$\mathcal{E}[J^b(x^a)] = 1/2\ \mathrm{tr}(\mathbf{KH})$$

The partition of costs in terms of individual classes of information sources z_j

objective function as a sum of various data / information sources

$$J_j(x) = \Sigma_{j=1}^{k} J_j(x)$$

with

$$J_j(x^a) = 1/2(\Gamma_j x - z_j)^T \mathbf{S}_j^{-1}(\Gamma_j x - z_j)$$

$\Gamma_j = \mathbf{H_j M}(t_j, t_0)$ combined observation - model operator
$\mathbf{S}_j$ information error covariance of class j with m_j elements of information

$$\mathcal{E}[J(x^a)] = 1/2\left(m_j - \mathrm{tr}(\mathbf{\Gamma_j^T S_j^{-1} \Gamma_j P^a})\right)$$

Summary

- an outline of special problems of tropospheric chemistry data assimilation is given, which are not prevalent in the stratosphere,
- problems are far from being solved, rather early attempts for solution are presented, or only theoretical access is shown,
- treatment of aqueous phase and aerosol phase data assimilation is still in its infancy, up to now resting on crude assumptions and ignoring system proclivity to non-Gaußian errors