Model Error Estimation

Its application to chemical data assimilation

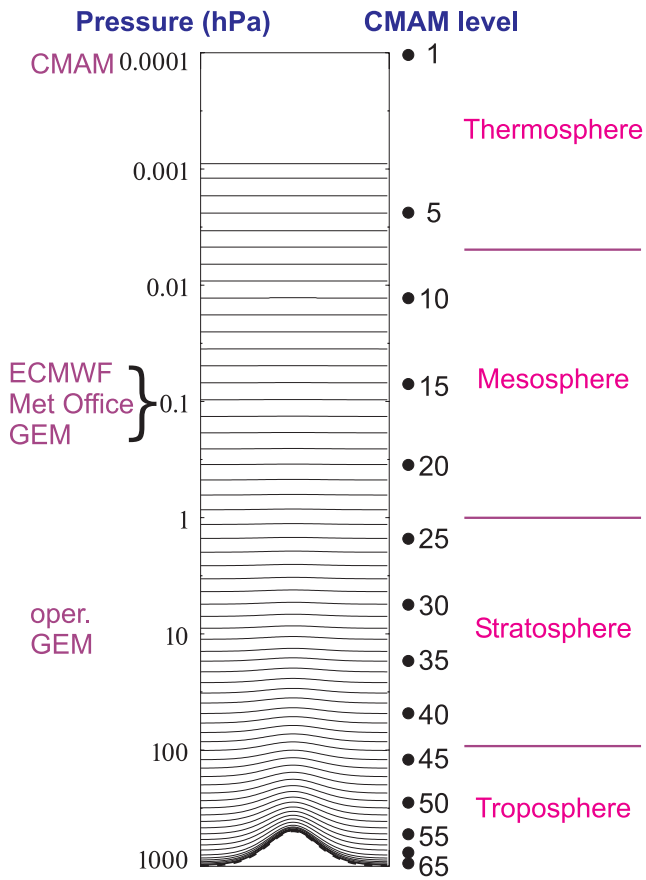
Richard Ménard, Yan Yang, Saroja Polavarapu
Meteorological Service of Canada

ECMWF Workshop on Modeling and Assimilation for the Stratosphere and Tropopause, 23 to 26 June, 2003

Canadian Middle Atmosphere Model = CMAM

CMAM is a complex GCM with interactive chemistry, radiation and dynamics

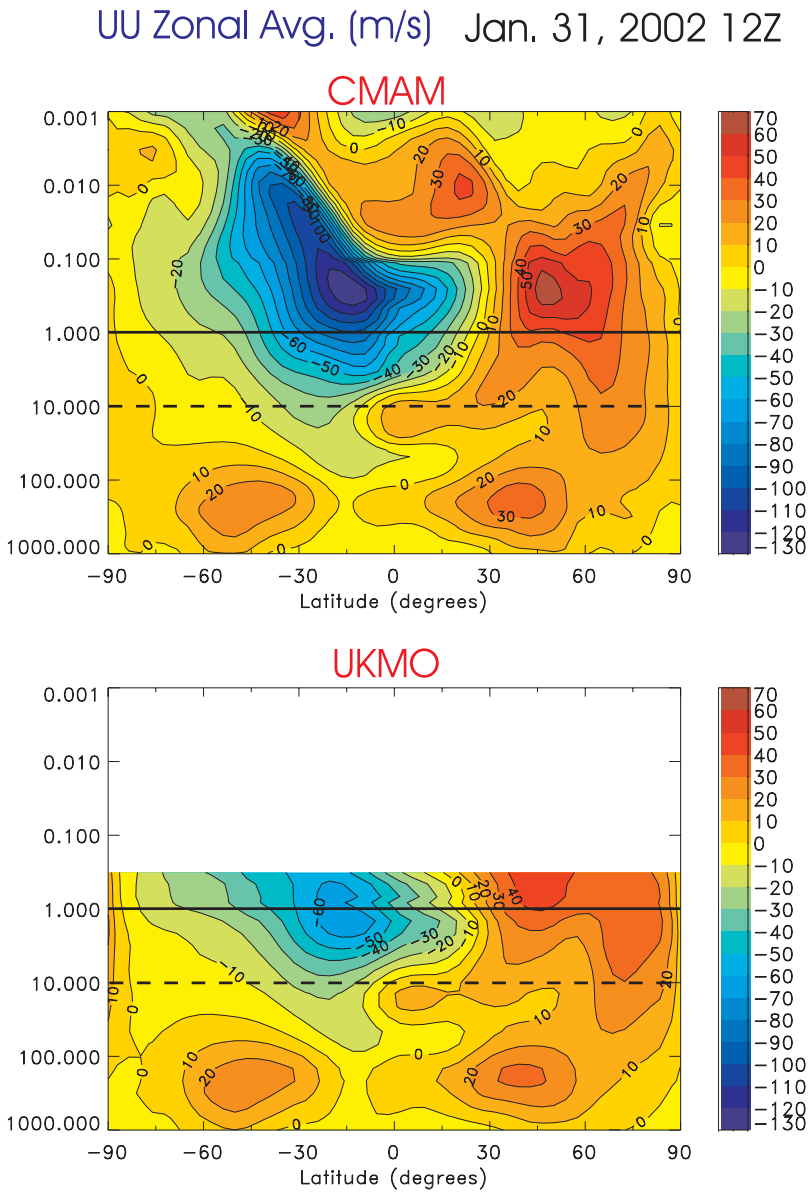
- T47, 65 levels from 0-95 km
- 127 gas-phase chemical reactions
- heterogeneous chemistry
- Hines GWD scheme



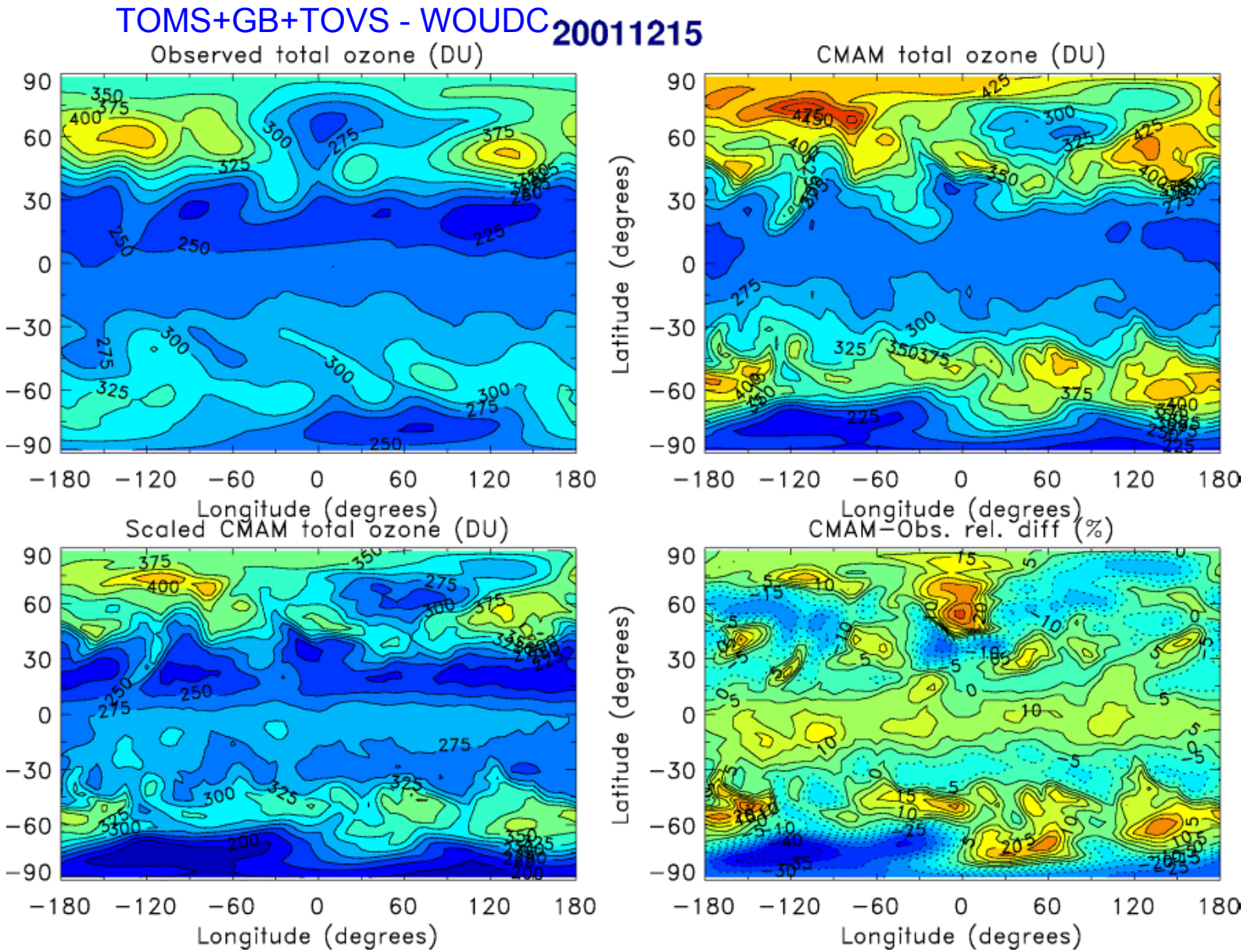
CMAM Data Assimilation

(Polavarapu, Ren, Rochon, Sankey, Yang)

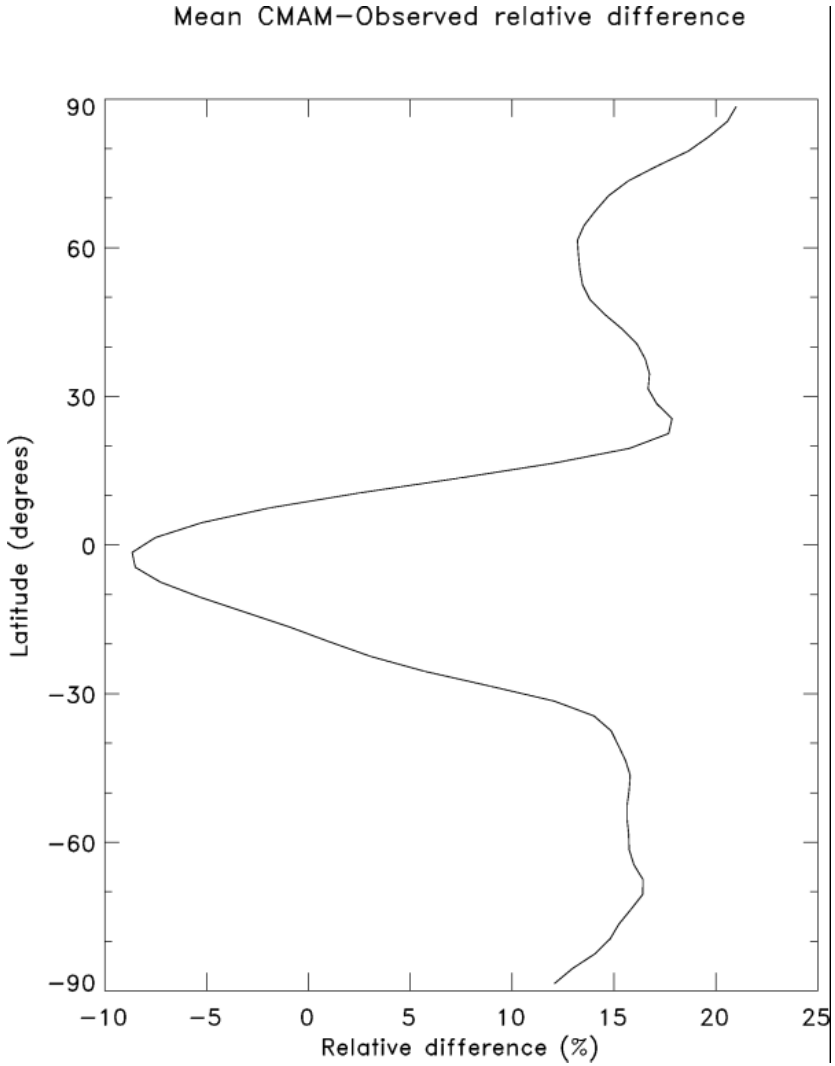
- CMC's 3DVAR on CMAM's coordinates
- obs: conventional, AMSU-A 4-14
- start-up from climate state Dec. 15, 2001

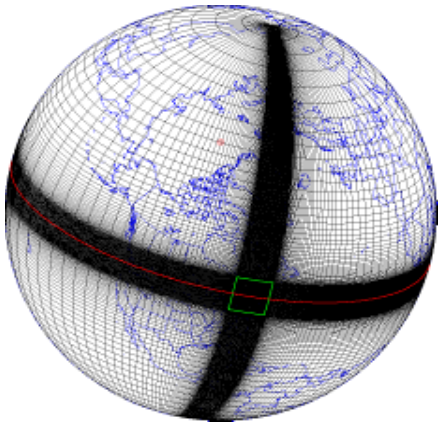

UU Zonal Avg. (m/s) Jan. 31, 2002 12Z
CMAM
0.001
0.010
0.100
1.000
10.000
100.000
1000.000
−90
−60
−30
0
30
60
90
Latitude (degrees)
70
60
50
40
30
20
10
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
−100
−110
−120
−130
UKMO
0.001
0.010
0.100
1.000
10.000
100.000
1000.000
−90
−60
−30
0
30
60
90
Latitude (degrees)
70
60
50
40
30
20
10
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
−100
−110
−120
−130

Comparison of total ozone against data



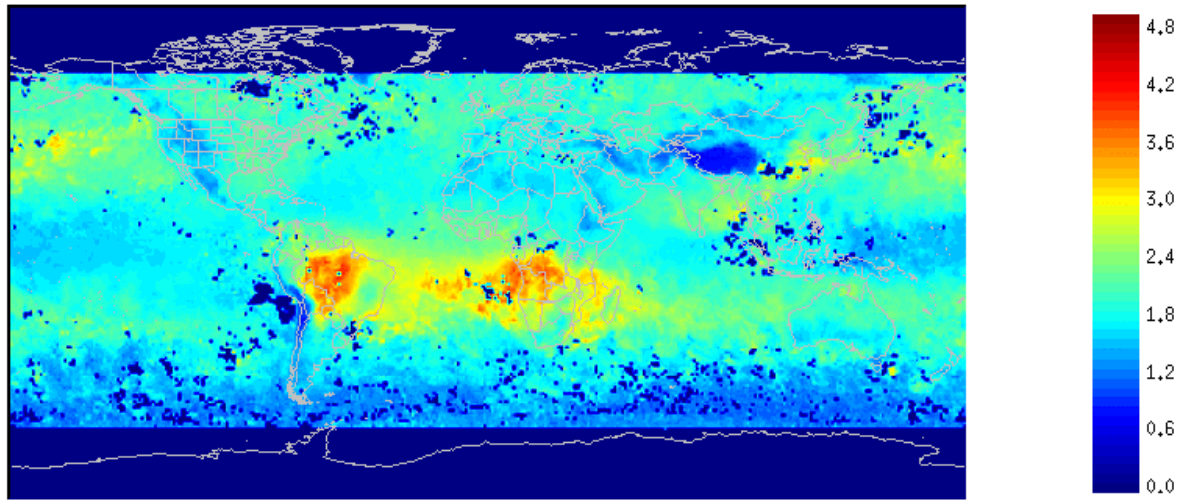
Scaling of ozone by observed values, averaging over one month and over longitudes.



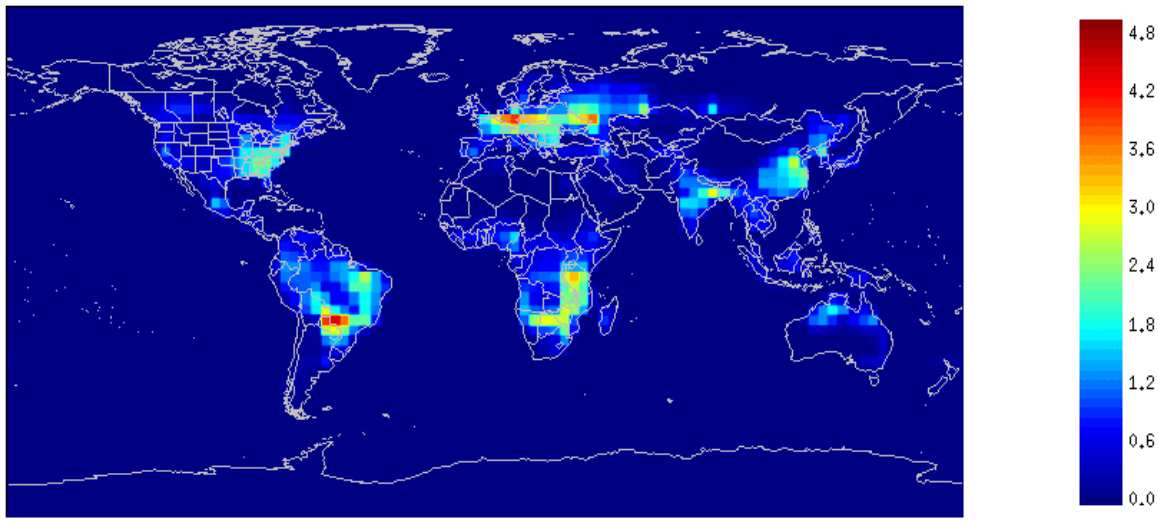


- The Global Environmental Multiscale Model GEM is an online meteorological-chemical transport model with variable resolution.
- Tangent linear and adjoint meteorological core
- Model top 10 mb, upgrade to 1.0 mb

- Tropospheric chemistry. Currently parametrized CO chemistry
  $CO + OH \rightarrow \ldots$ loss of CO using a prescribed zonal OH field
  $CH_4 + OH \rightarrow CO + \ldots$ production of CO with a global value for CH_4 and will be upgraded soon with ADOM gas phase chemistry
- Surface emission based on GEIA and biogenetic modeling
- Tropospheric chemistry based on ADOM (32 advected species), currently adding aerosols based on the CAM. Also possibility to run with the CMAM phase chemistry for stratospheric

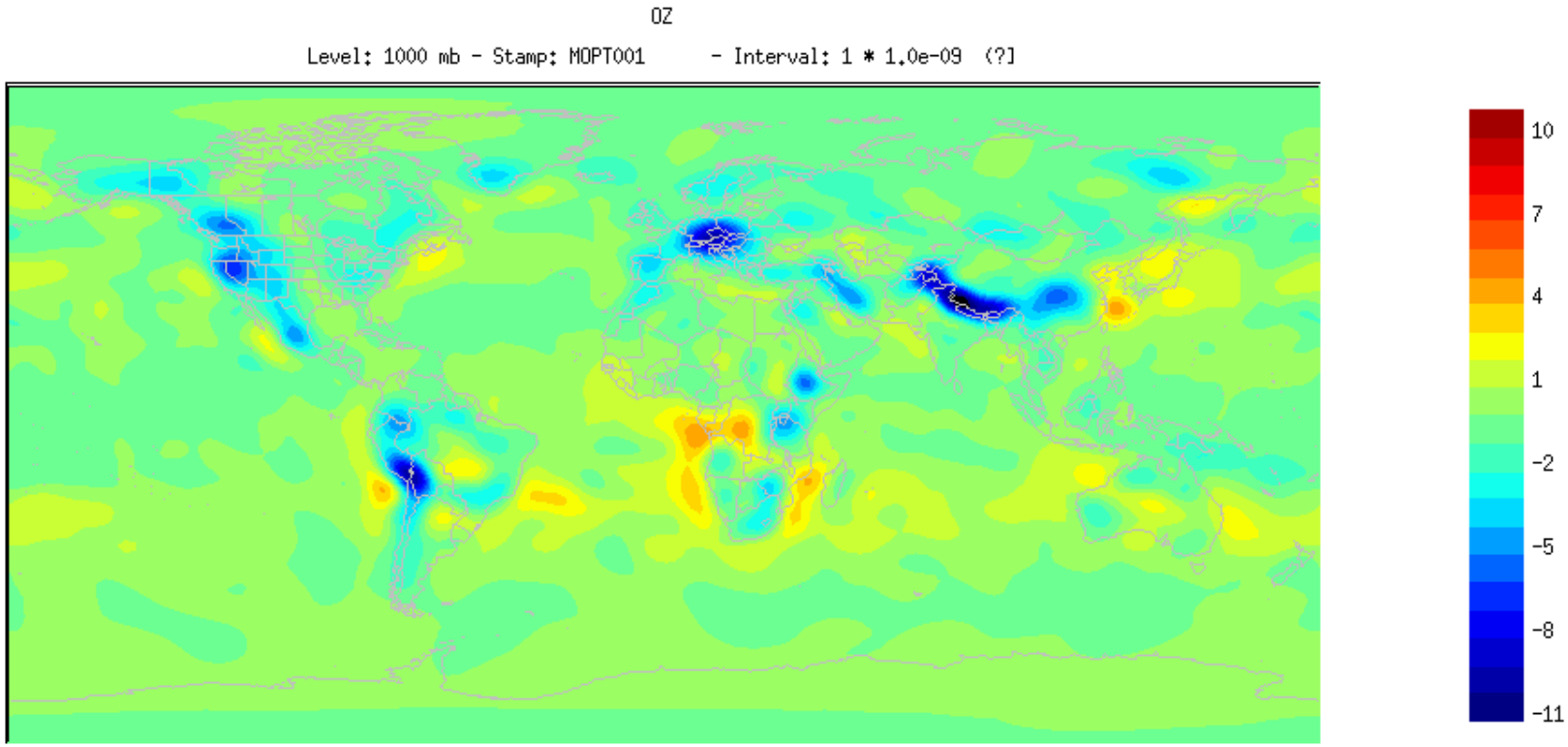


Column CO observations from MOPITT 10/12/00 – 10/27/00



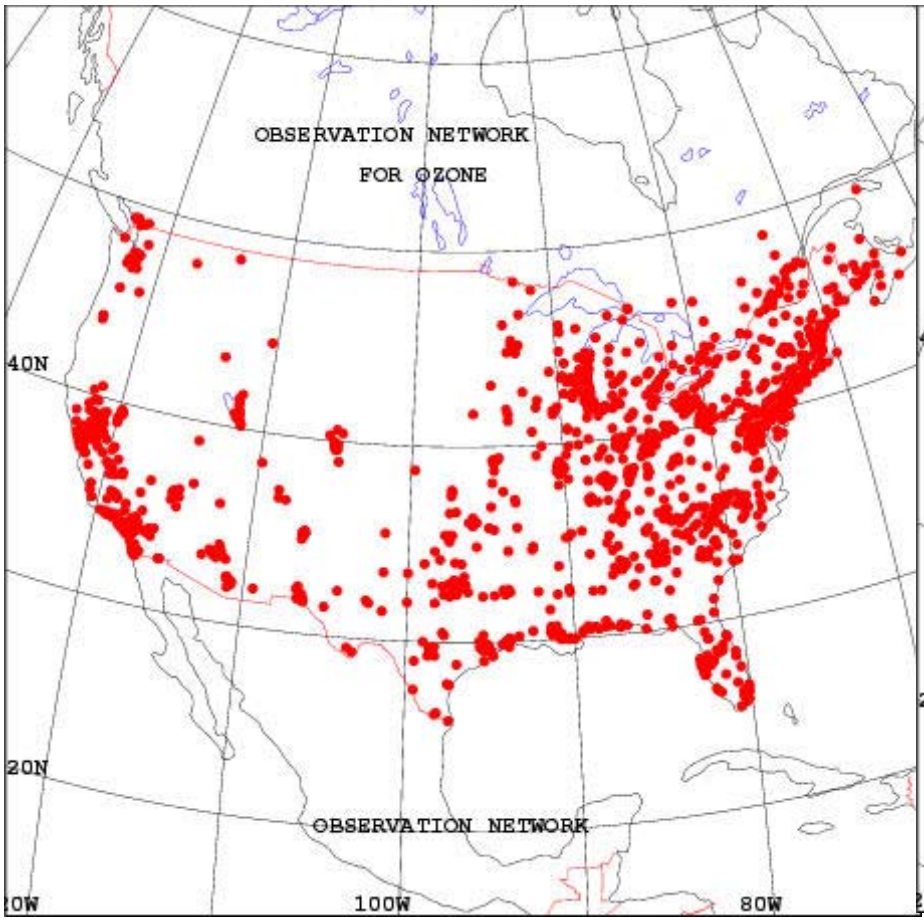
CO emissions for October in molecules cm^{-2} s^{-1} x 10^{+11}

Time mean analysis increments : lowest model level

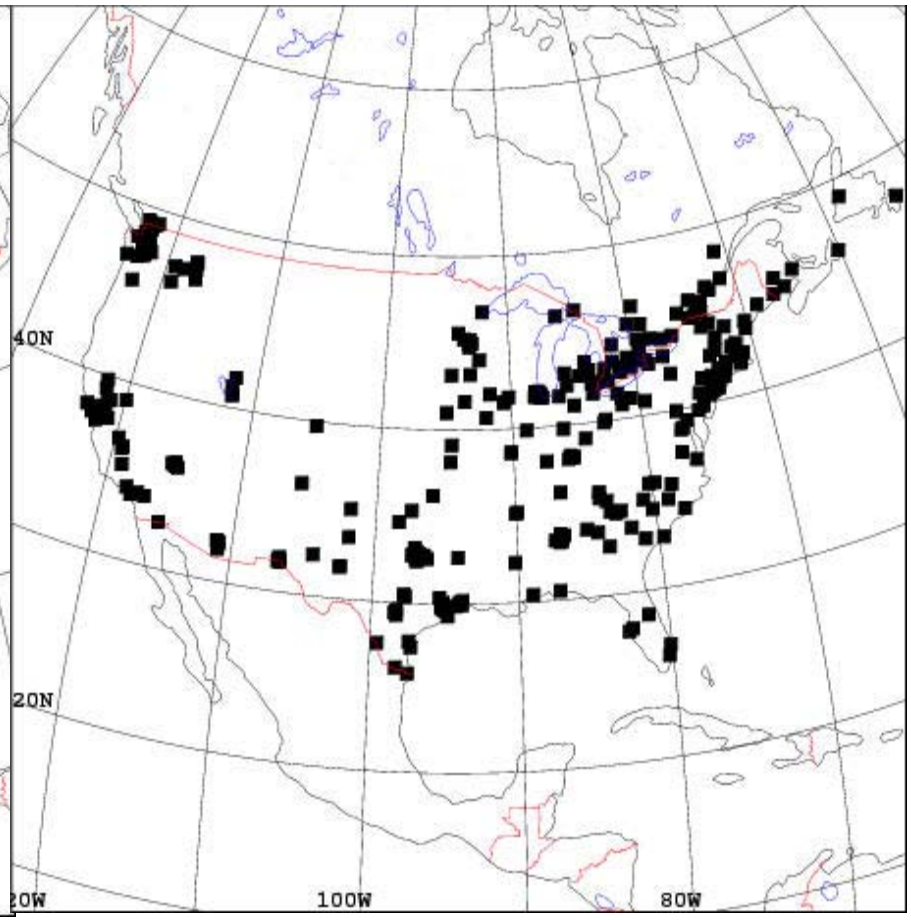


Real-time data

Ozone, ~ 1500 hourly



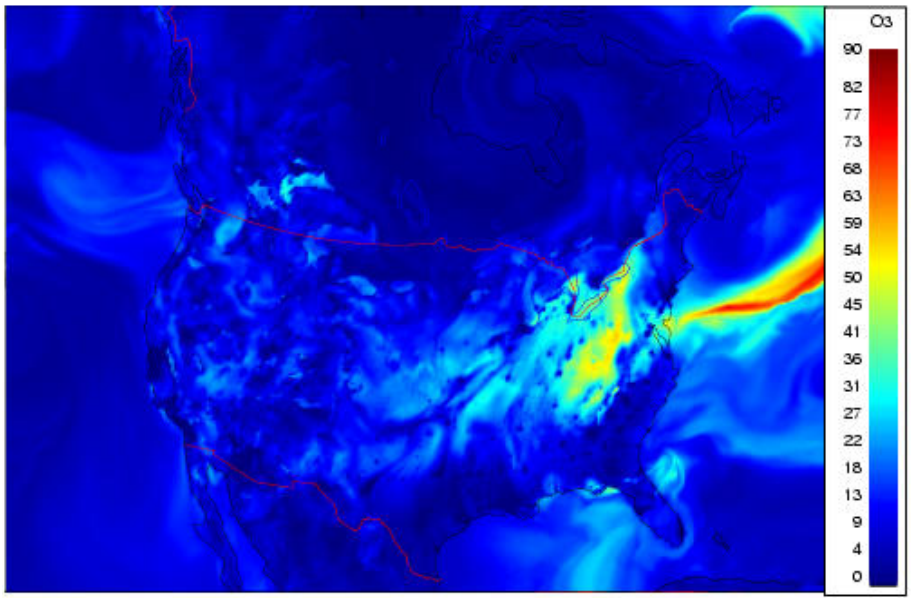
PM 2.5, ~ 300 hourly



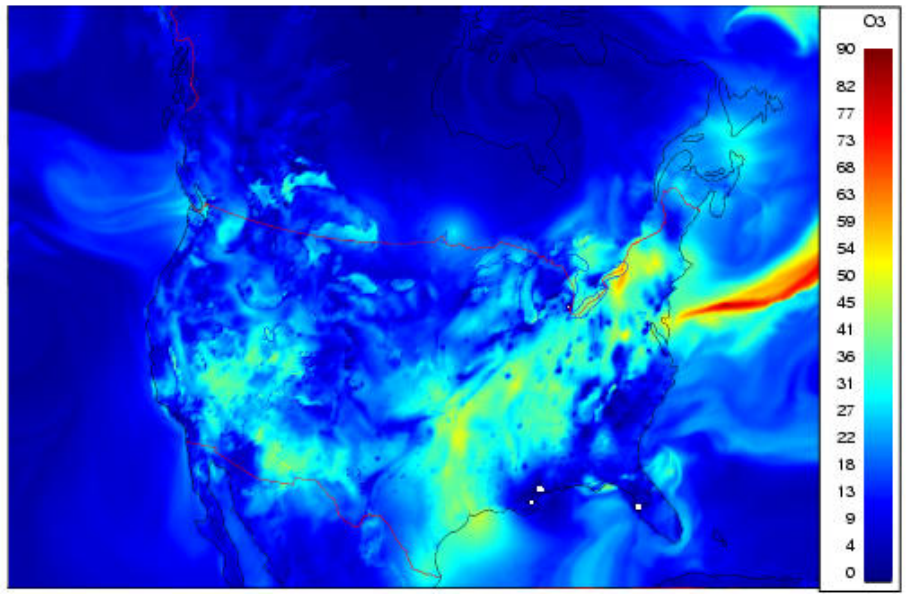
Objective analysis v1.0

Regional air quality model

Surface ozone



Model simulation



Objective analysis

Some of the difficulties with chemical data assimilation

- Large systematic errors, and in some cases lack of sensitivity to initial conditions.
  Observations should be used to estimate model error and chemical state.

- Large control space for the analysis scheme.
  e.g. gas phase chemistry may have 32 advected species,
  aerosols ; 4 types, 12 bins correspond to 64 advected species

Some approach for simultaneous state and model error estimation

- Dee and daSilva (1998), Dee and Todling (2000)

$$J(\mathbf{x},\mathbf{u}) = \frac{1}{2}\left[(\mathbf{x}+\mathbf{u})-\mathbf{x}^f\right]^T \mathbf{P}^{-1}\left[(\mathbf{x}+\mathbf{u})-\mathbf{x}^f\right]$$
$$+\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})$$
$$+\frac{1}{2}(\mathbf{u}-\mathbf{u}^f)^T \mathbf{Q}^{-1}(\mathbf{u}-\mathbf{u}^f)$$

- Griffith and Nichols (2001)

$$J(\mathbf{x}_0,\mathbf{u}_0) = \frac{1}{2}\left[\mathbf{x}_0-\mathbf{x}_0^f\right]^T \mathbf{P}_0^{-1}\left[\mathbf{x}_0-\mathbf{x}^f\right]$$
$$+\frac{1}{2}\sum(\mathbf{y}-\mathbf{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})$$
$$+\frac{1}{2}(\mathbf{u}_0-\mathbf{u}_0^f)^T \mathbf{Q}_0^{-1}(\mathbf{u}_0-\mathbf{u}_0^f)$$

- Most schemes do not contain an error cross-covariance $\mathbf{P}_{ux}$

  - In Dee and daSilva, $\mathbf{P}_{ux}$ is implicit and in fact equal to $\mathbf{Q}$. This arises because of the use of the bias-blind approach
  - In Griffith and Nichols, the error cross-covariance is generated implicitly because of the propagation in time, although is set to zero at initial time.

- But $\mathbf{P}_{ux}$ is the *most important error statistic* for model error estimation

  - In a bias-aware algorithm the gain matrix for the model error is

$$\mathbf{K}_u = \mathbf{P}_{ux}^f \mathbf{H}^T \left(\mathbf{H} \mathbf{P}_{xx}^f \mathbf{H}^T + \mathbf{R} \right)^{-1} = \mathbf{P}_{ux}^f \left(\mathbf{P}_{xx}^f \right)^{-1} \mathbf{K}_x$$

Forecast of a chemical tracer with source

- Lets continue with the scalar model, and parametrized the model evolution by the simple equation

$$\frac{dC}{dt} + LC = S$$

where L is a constant that represent chemical life time and diffusion effects

- Also let consider a constant source

$$\frac{dS}{dt} = 0$$

- In order to compute the evolution of error, a model of the truth is needed. To simplify, lets assume that the main uncertainty arises from modeling of the source dynamics rather than from the modeling of the transport,

$$\frac{dC^t}{dt} + LC^t = S^t$$

$$\frac{dS^t}{dt} = \varepsilon^q$$

where the superscript ^{t} denotes the true value

The error thus follows

$$\frac{d\tilde{C}}{dt} + L\tilde{C} = \tilde{S}$$

$$\frac{d\tilde{S}}{dt} = \varepsilon^q$$

where the tilde ~ denotes the departure from the truth

- We get the error covariance system, and can solve analytically

$$f_{ss}^2(t) = f_{ss}^2(0)\,\exp(+Qt)$$

$$f_{cs}^2(t) = \left[f_{cs}^2(0) - \frac{f_{ss}^2(0)}{L+Q} \right] \exp(-Lt) + \frac{f_{ss}^2(0)}{L+Q} \exp(+Qt)$$

$$f_{cc}^2(t) = \exp(-2Lt) \left[f_{cc}^2(0) - \frac{2}{L} \left(f_{cs}^2(0) - \frac{f_{ss}^2(0)}{L+Q} \right) - \frac{2\, f_{ss}^2(0)}{(L+Q)(2L+Q)} \right]$$
$$+ \exp(-Lt) \left[\frac{2}{L} \left(f_{cs}^2(0) - \frac{f_{ss}^2(0)}{L+Q} \right) \right]$$
$$+ \exp(+Qt)\, \frac{2\, f_{ss}^2(0)}{(L+Q)(2L+Q)}$$

- In the limit $t \to \infty$

$$f_{ss}^2(t) = f_{ss}^2(0)\,\exp(+Qt)$$

$$f_{cs}^2(t) \to \frac{f_{ss}^2(t)}{L+Q}$$

$$f_{cc}^2(t) \to \frac{f_{ss}^2(t)}{(L+Q)(2L+Q)}$$

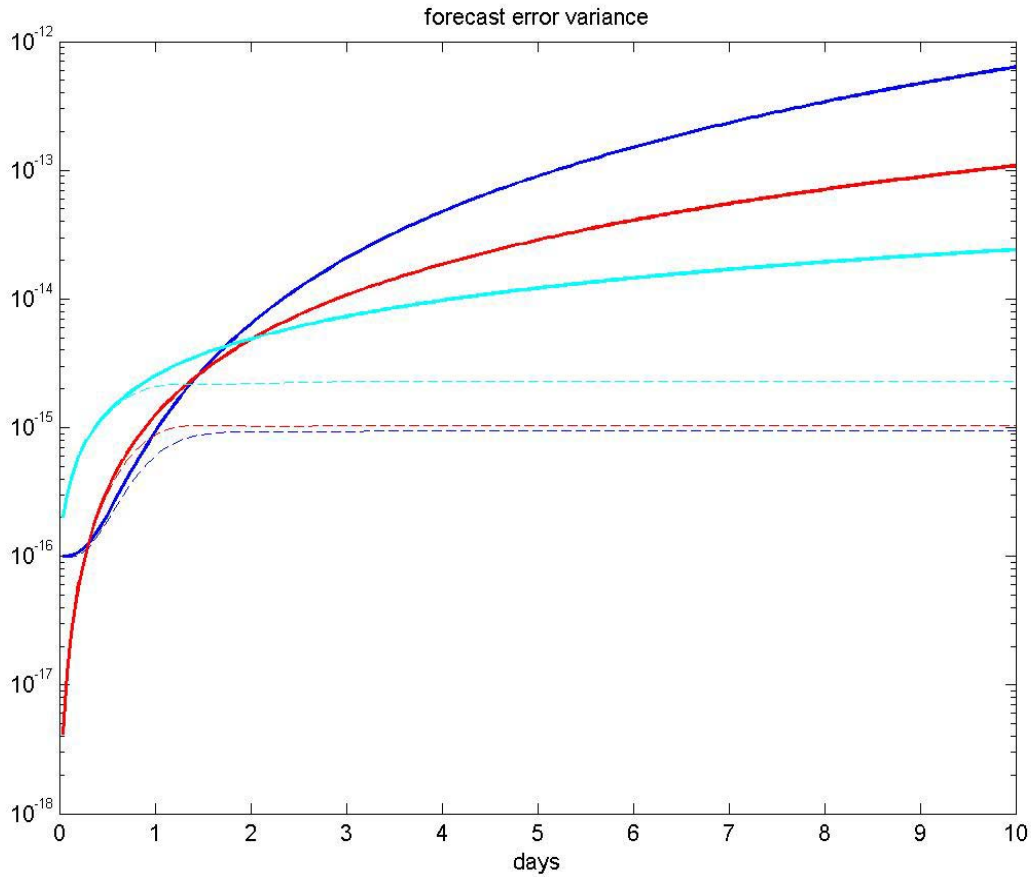
Numerical experiment

$$C \sim 100\ \text{ppmv}\ ;\ \widetilde{C} \sim 10\%\ ;\ L = \frac{1}{30\ \text{days}}\ ;\ f_{cc}^2 \sim \widetilde{C}^2\ ;\ f_{ss}^2 \sim f_{cc}^2 \sim Q\ ;\ f_{cs}^2 = 0$$

f_{cc}^2 : blue f_{cs}^2 : red f_{ss}^2 : cyan

No assimilation solid line

With assimilation R= Q*100 dashed line



The importance of time lag statistics

- In synthetic inversion of the source long time integration period are used. Lets examine how longer time integrations may carry more information about the source. For this we will examine the time correlations

$$\left\langle \widetilde{C}(t)\widetilde{S}(0)\right\rangle$$

$$\left\langle \widetilde{S}(t)\widetilde{S}(0)\right\rangle$$

$$\left\langle \widetilde{C}(t)\widetilde{C}(0)\right\rangle$$

we get,

$$\left\langle \widetilde{C}(t)\widetilde{S}(0)\right\rangle = \exp(-Lt)\left\langle \widetilde{C}(0)\widetilde{S}(0)\right\rangle + \frac{1}{L}\left(1-\exp(-Lt)\right)\left\langle \widetilde{S}^2(0)\right\rangle$$

$$\left\langle \widetilde{S}(t)\widetilde{S}(0)\right\rangle = \left\langle \widetilde{S}^2(0)\right\rangle$$

$$\left\langle \widetilde{C}(t)\widetilde{C}(0)\right\rangle = \exp(-Lt)\left\langle \widetilde{C}^2(0)\right\rangle$$

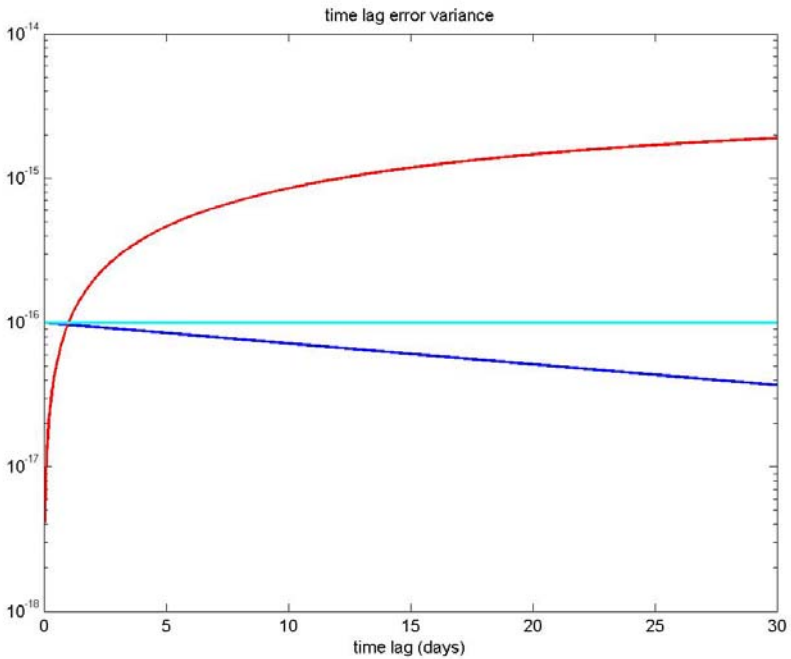
- Using the formal solutions

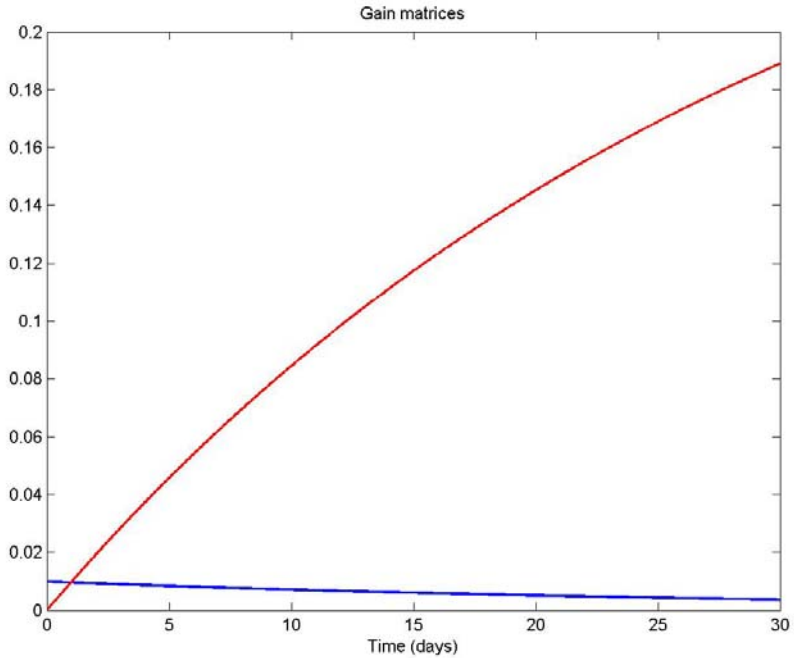
$$\widetilde{C}(t) = \exp(-Lt)\widetilde{C}(0) + \exp(-Lt)\int_0^t \exp(L\tau)\widetilde{S}(\tau)\, d\tau$$

$$\widetilde{S}(t) = \widetilde{S}(0) + \int_0^t \varepsilon^q(\tau)\, d\tau$$

Numerical experiment

(same conditions as before)





$\left\langle \widetilde{C}(t)\widetilde{S}(0)\right\rangle$: red

$\left\langle \widetilde{S}(t)\widetilde{S}(0)\right\rangle$: cyan

$\left\langle \widetilde{C}(t)\widetilde{C}(0)\right\rangle$: blue

red : gain for the source
blue: gain for the state

- The effect of this dynamics is to increase the lag cross-covariance between state error and source (or bias) error.

Sequential estimation of state and model error

If there are some variables (such as model error) that are unobserved, the analysis can be split into two parts: One leaving the 3D Var scheme intact and the other as an additional step for solving for the unobserved variables.

From

$$J(\mathbf{x},\mathbf{u}) = \frac{1}{2}\begin{pmatrix}\mathbf{x}-\mathbf{x}^f\\ \mathbf{u}-\mathbf{u}^f\end{pmatrix}^T \begin{pmatrix}\mathbf{P}^f_{xx} & \mathbf{P}^f_{xu}\\ \mathbf{P}^f_{ux} & \mathbf{P}^f_{uu}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{x}-\mathbf{x}^f & \mathbf{u}-\mathbf{u}^f\end{pmatrix} + \frac{1}{2}\left(\mathbf{y}-H(\mathbf{x})\right)^T \mathbf{R}^{-1}\left(\mathbf{y}-H(\mathbf{x})\right)$$

Decomposing the 4x4 forecast error covariance with a square root of the form,

$$\mathbf{P}^f = \mathbf{S}\mathbf{S}^T \qquad \mathbf{S} = \begin{bmatrix}\mathbf{L} & \mathbf{0}\\ \mathbf{P}_{ux}\left(\mathbf{L}^T\right)^{-1} & \mathbf{D}^{-1/2}\end{bmatrix}$$

$$\mathbf{P}_{xx} = \mathbf{L}\mathbf{L}^T \quad , \quad \mathbf{D}^{-1} = \mathbf{P}_{uu} - \mathbf{P}_{ux}\mathbf{P}_{xx}^{-1}\mathbf{P}_{xu}$$

and then with the change of variable

$$\mathbf{z} = \mathbf{S}\,\varsigma \quad , \quad \varsigma = \begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

We get to solve in sequence,

$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x} - \mathbf{x}^f\right)^T \mathbf{P}_{xx}^{-1}\left(\mathbf{x} - \mathbf{x}^f\right) + \frac{1}{2}\left(\mathbf{y} - \mathbf{H}(\mathbf{x})\right)^T \mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{H}(\mathbf{x})\right)$$

and then

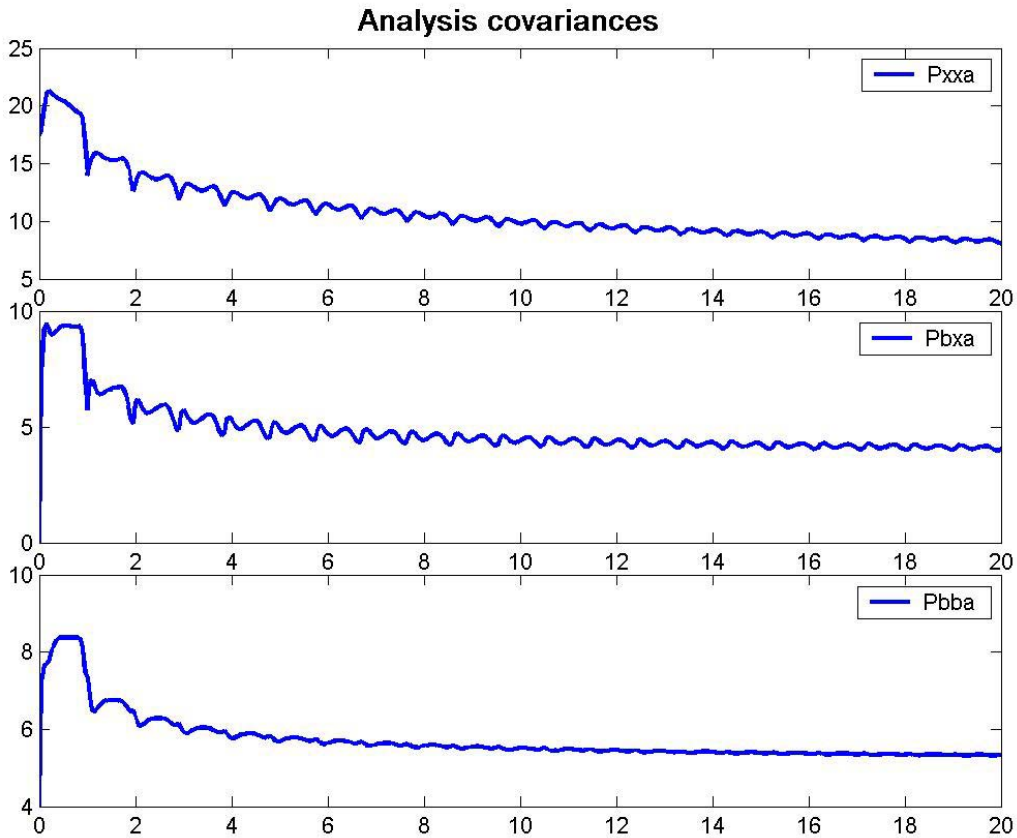
$$\mathbf{u}^a = \mathbf{u}^f + \mathbf{P}_{ux}\left(\mathbf{P}_{xx}\right)^{-1}\left(\mathbf{x}^a - \widetilde{\mathbf{x}}^f\right)$$

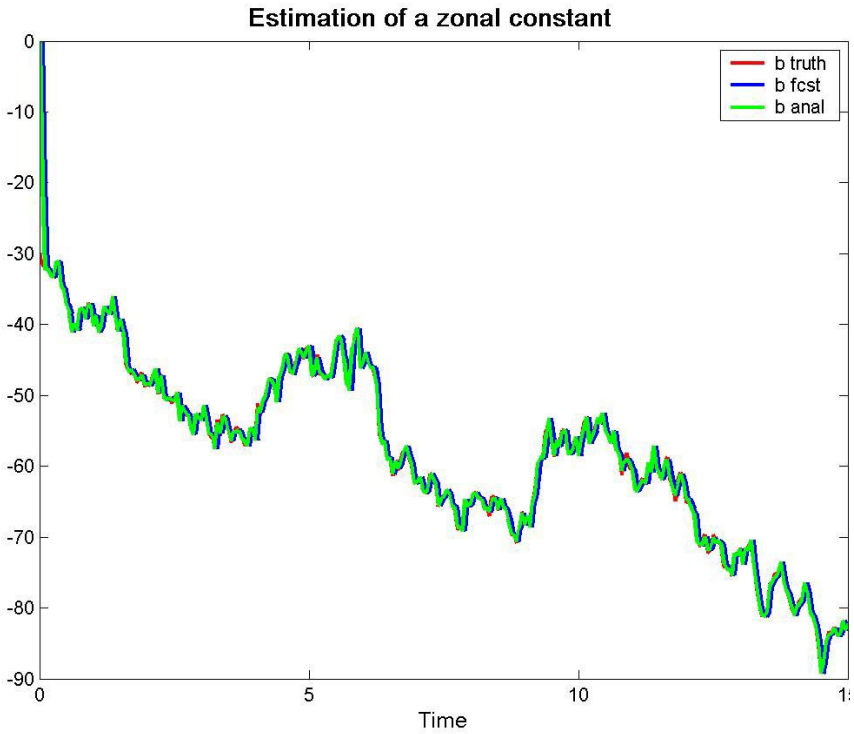
- This splitting may have a computational advantage, as the control variable remain x in the 3D Var numerical code.

- Also if there are several u to be estimated, the **u** update equation can be computed in parallel.

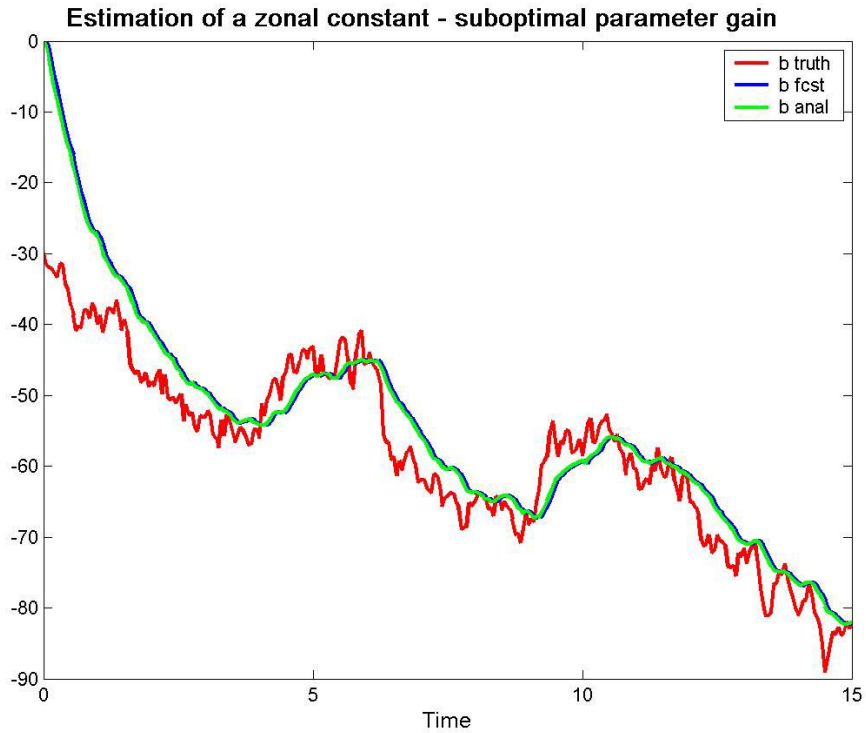
Simple Kalman filtering experiment

One-dimensional advection over a periodic domain.
Estimate a scalar (zonal constant) using a single point observation





optimal $\mathbf{K}^x$ and $\mathbf{K}^u$



$\mathbf{K}^u$ is an order of magnitude smaller than the optimal value

Lagged-innovation covariances and its potential use for estimating P_{ux}

- Since u is not an observed variable, lag-0 innovations, does not provide information about $\mathbf{P}_{ux}$.

- Dynamics is needed to enable information about $\mathbf{P}_{ux}$ to become "visible" to the observations.

$$\mathbf{C}_k^k = \mathbf{H}_k \, \mathbf{P}_{xx,k}^f \, \mathbf{H}_k^T + \mathbf{R}_k \quad : \text{ Lag-0 innovation covariance}$$

$$\mathbf{C}_{k+1}^k = \mathbf{H}_{k+1} \mathbf{M}_k \left[\mathbf{K}_k^x - \widetilde{\mathbf{K}}_k^x\right] \mathbf{C}_k^k + \mathbf{H}_{k+1} \mathbf{G}_k \left[\mathbf{K}_k^u - \widetilde{\mathbf{K}}_k^u\right] \mathbf{C}_k^k$$

Lag-1 is zero when the gain matrices are optimal, i.e.

$$\mathbf{K}^x = \widetilde{\mathbf{K}}^x \;\; ; \;\; \mathbf{K}^u = \widetilde{\mathbf{K}}^u$$

For dense observation network, say $\mathbf{H} \approx \mathbf{I}$, and to estimate model bias, i.e. $\mathbf{G} = \mathbf{I}$, and if $\mathbf{K}^u$ is optimal and $\widetilde{\mathbf{K}}^u$ is small then we get the surprising result that

$$\mathbf{C}_{k+1}^{k} \approx \mathbf{P}_{ux}^{f}$$

Of course in practice, if lag-1 is non zero, then it is not clear we can have the right state Kalman gain.

This specific issue, the development of a method to estimate $\mathbf{P}_{ux}$ is currently investigated.

Conclusions

- Estimation of model error is an important issue in chemical data assimilation
- Model error can be considered as an unobserved variable, and requires the knowledge of cross-error covariance
- In 3D Var, observed and unobserved variables can be estimated in sequence, and this may have computational advantages
- Lag-1 innovation covariances contain information about the cross-covariance error