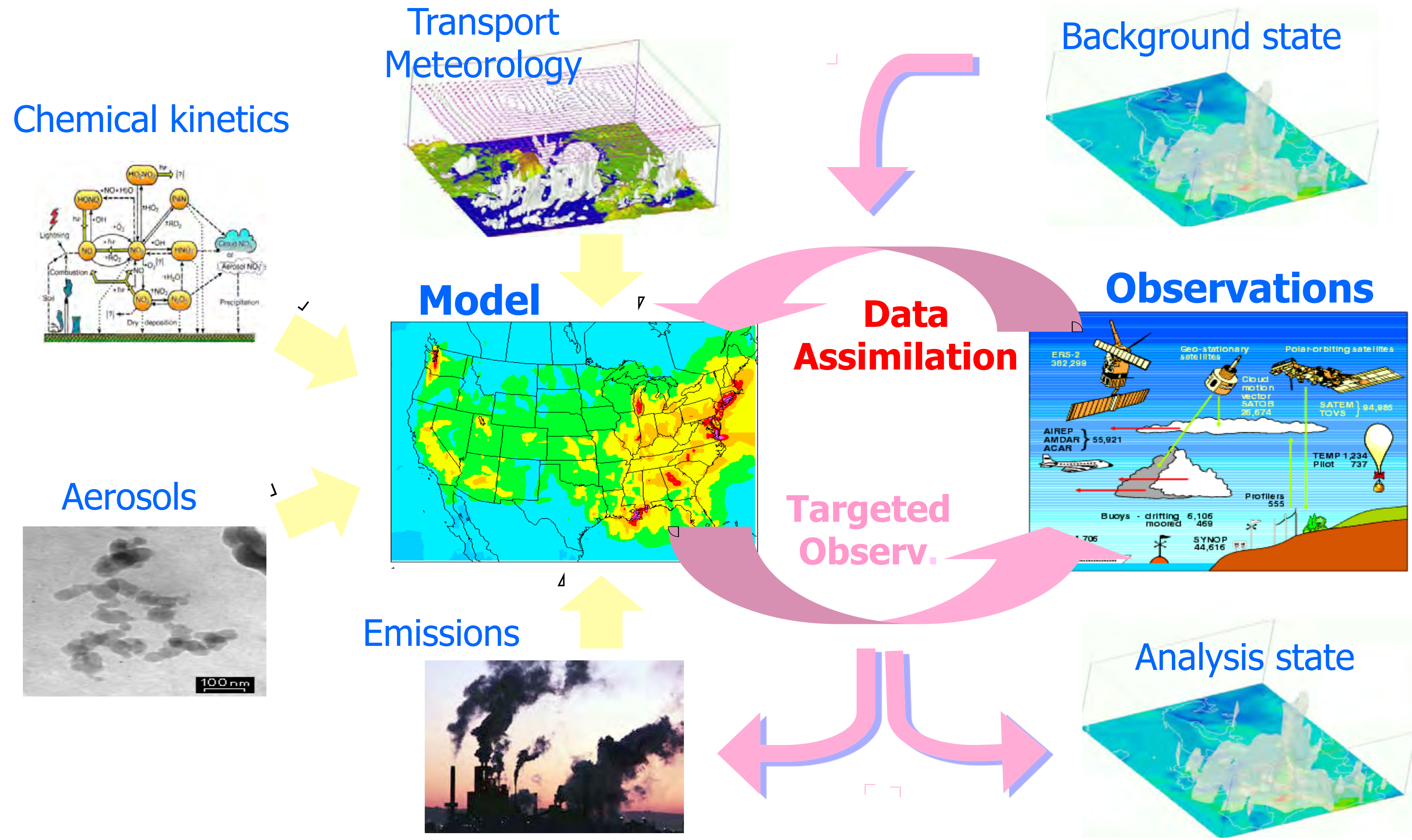
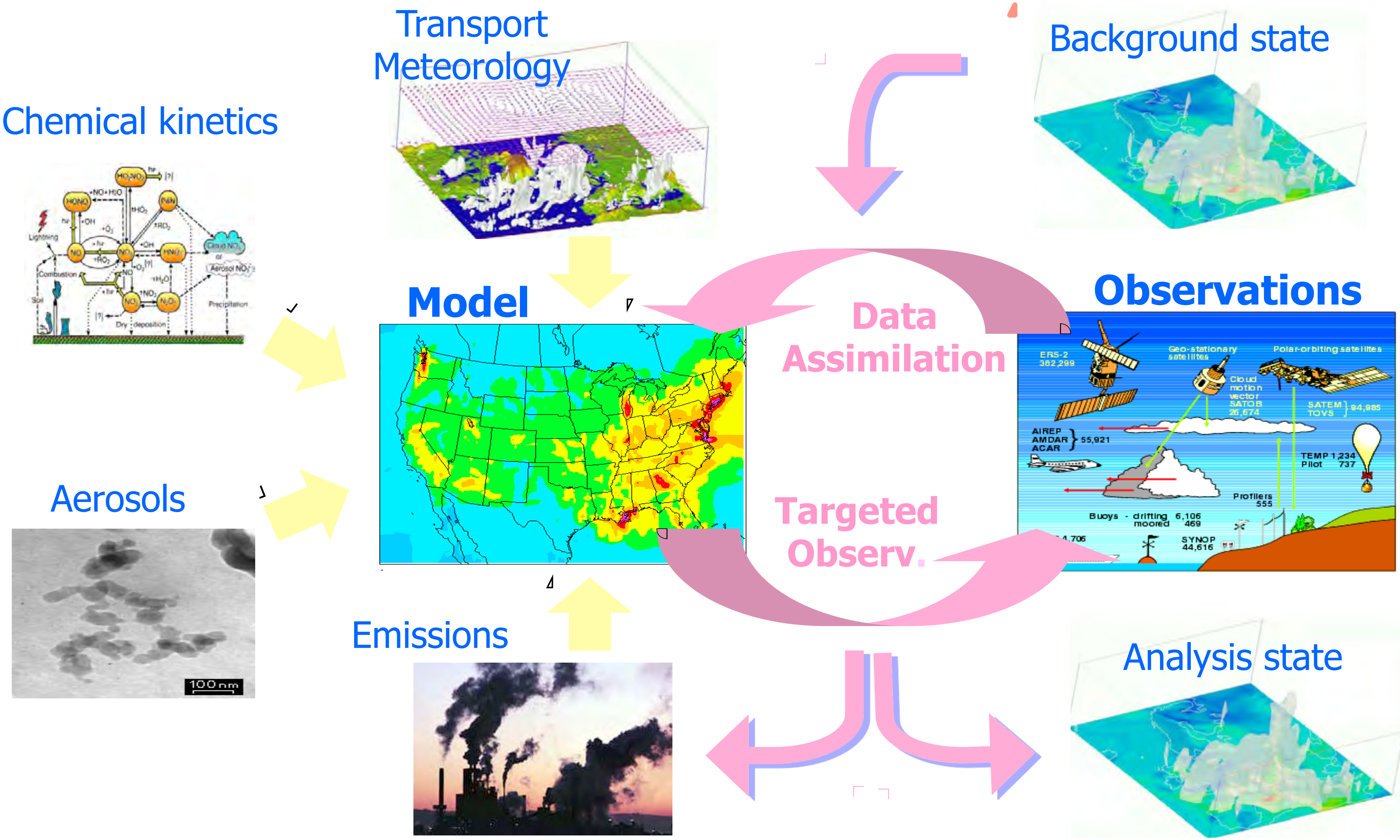
Data Assimilation in Geosciences: *A highly multidisciplinary enterprise*

Adrian Sandu
Computational Science Laboratory
Department of Computer Science
Virginia Tech

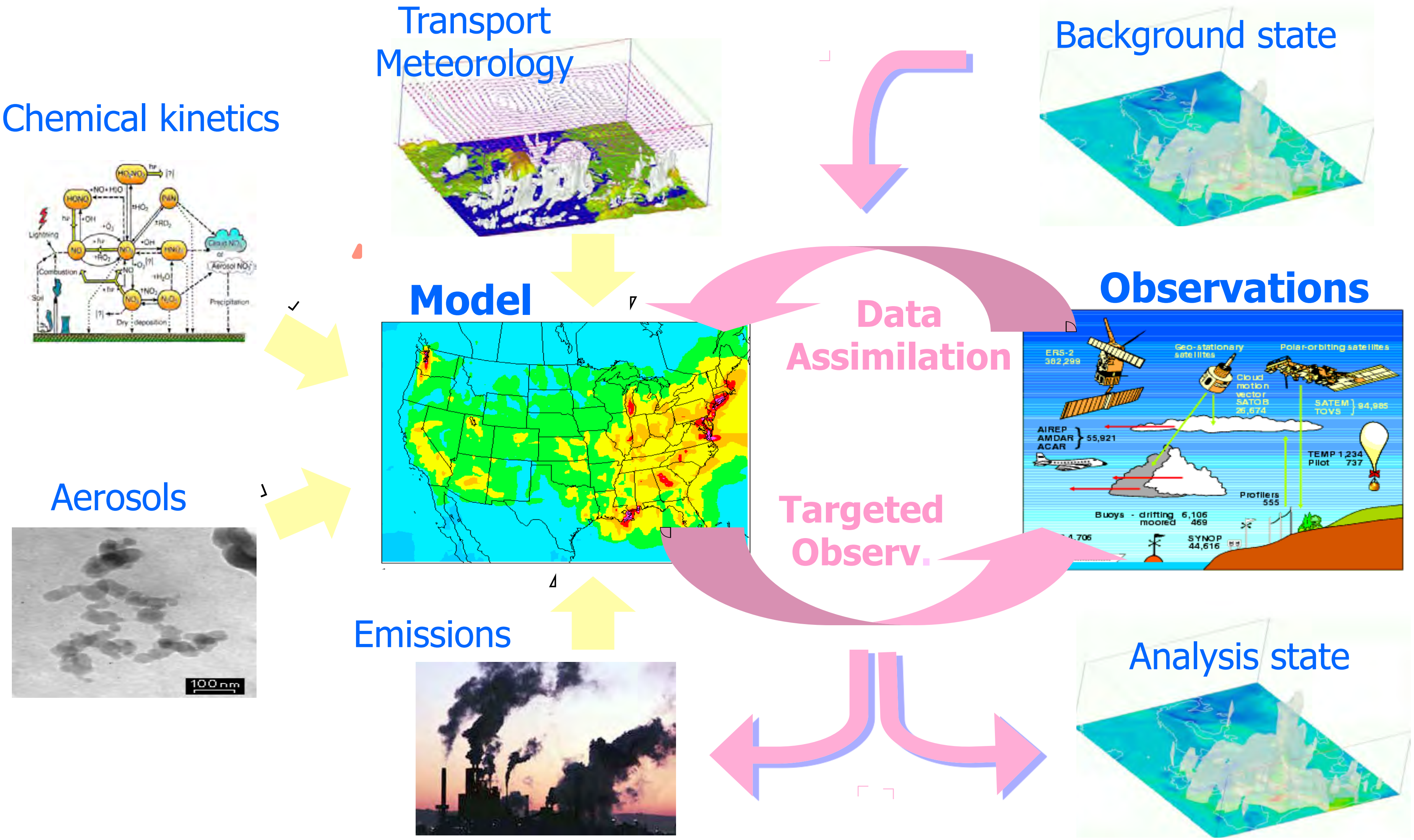
Data assimilation fuses information from prior, model, and observations, to best describe a physical system



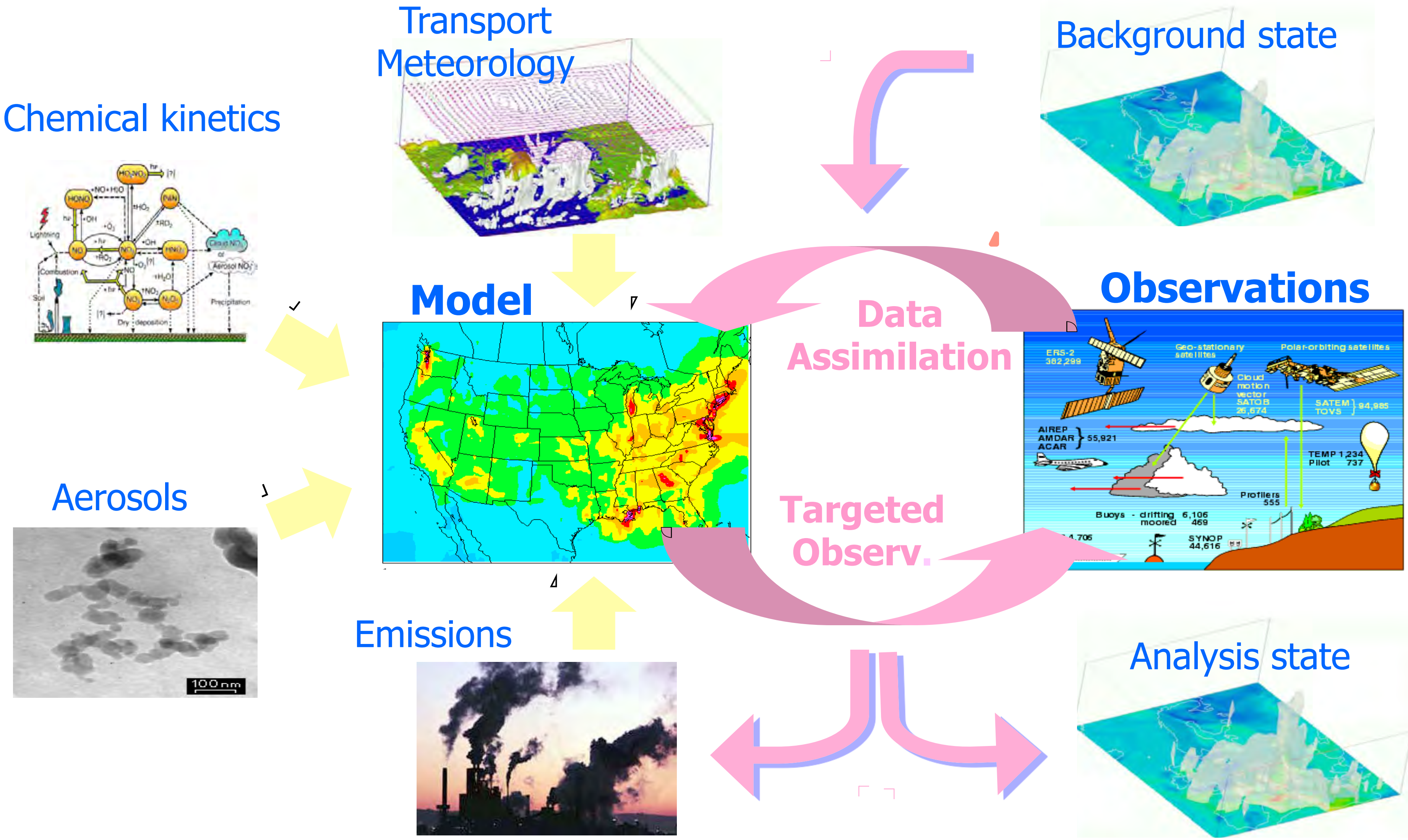
Data assimilation fuses information from prior, model, and observations, to best describe a physical system



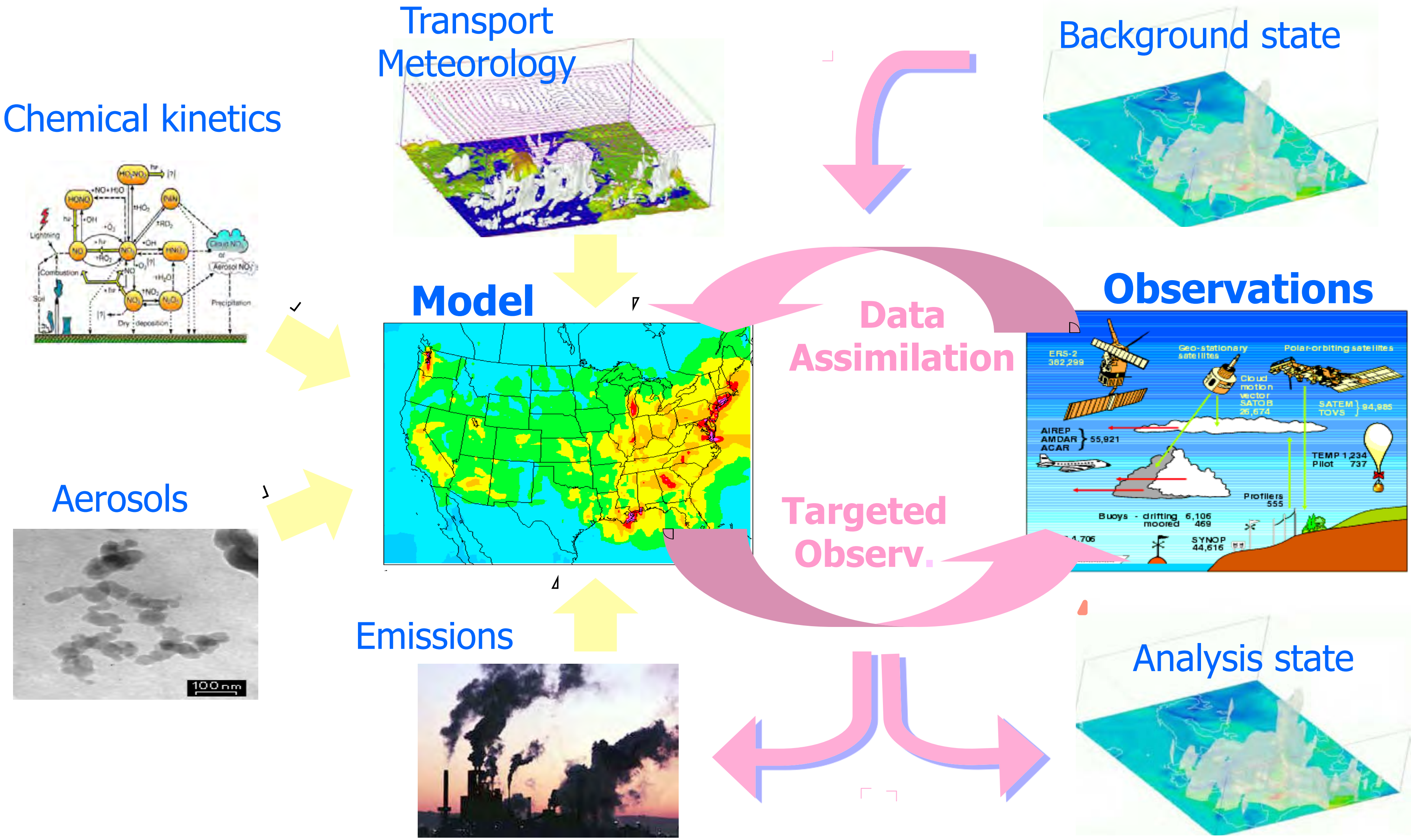
Data assimilation fuses information from prior, model, and observations, to best describe a physical system



Data assimilation fuses information from prior, model, and observations, to best describe a physical system

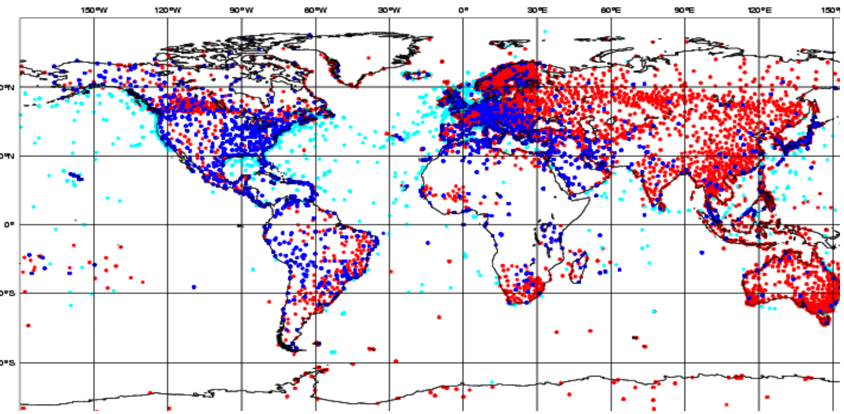


Data assimilation fuses information from prior, model, and observations, to best describe a physical system

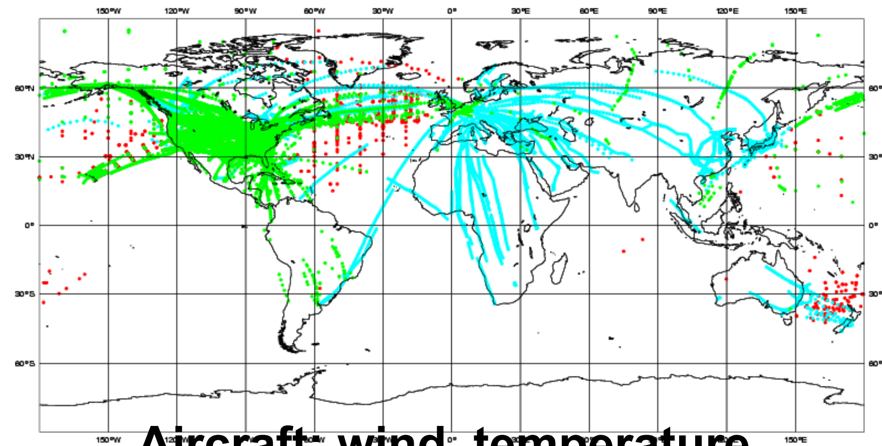


Some conventional and remote data sources used at ECMWF for numerical weather prediction

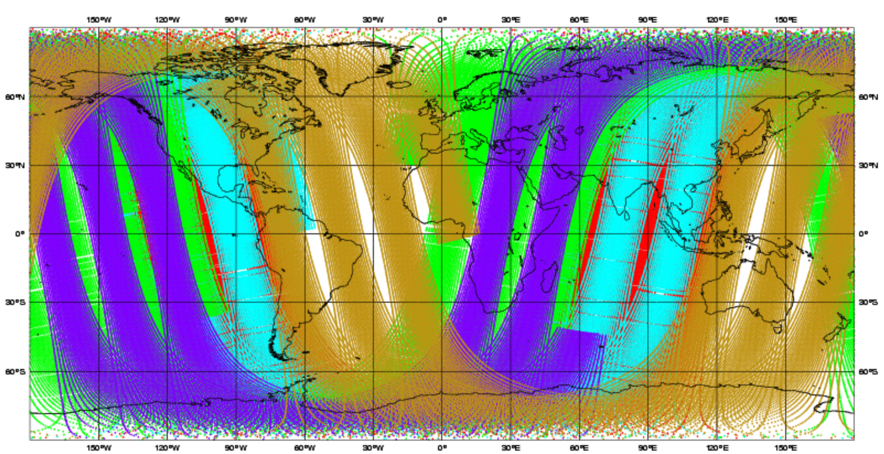
Lars Isaksen (http://www.ecmwf.int)



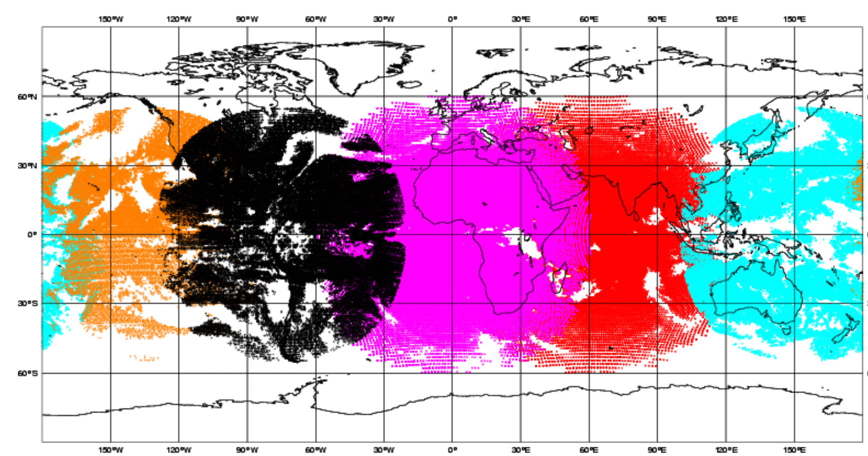
SYNOP/METAR/SHIP: pres., wind, RH



Aircraft: wind, temperature



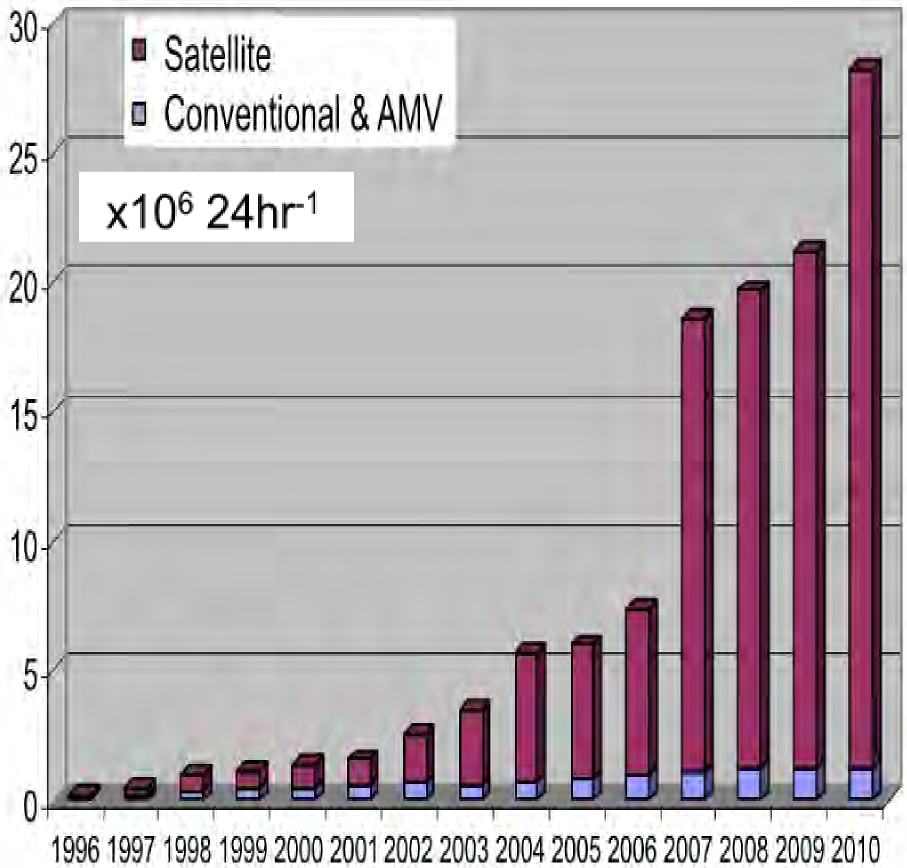
13 Sounders: NOAA AMSU-A/B, HIRS, AIRS, …



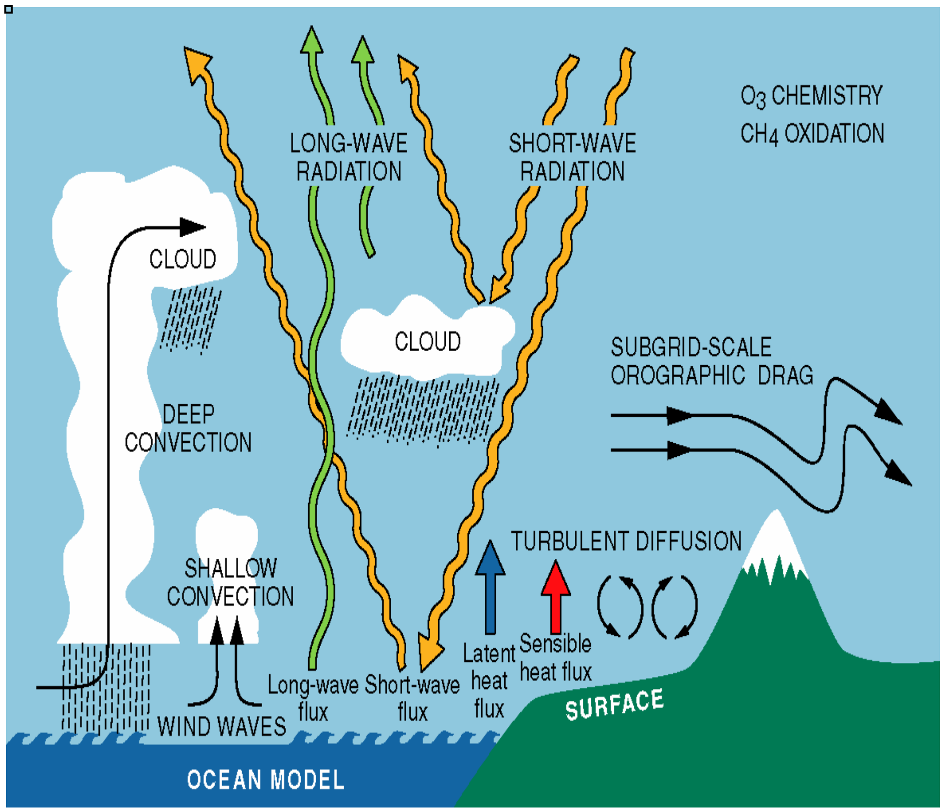
Geostationary, 4 IR and 5 winds

Challenge: data assimilation problems of practical interest are large-scale and computationally intensive

How many observations are being assimilated? All data assimilated at ECMWF 1996-2010



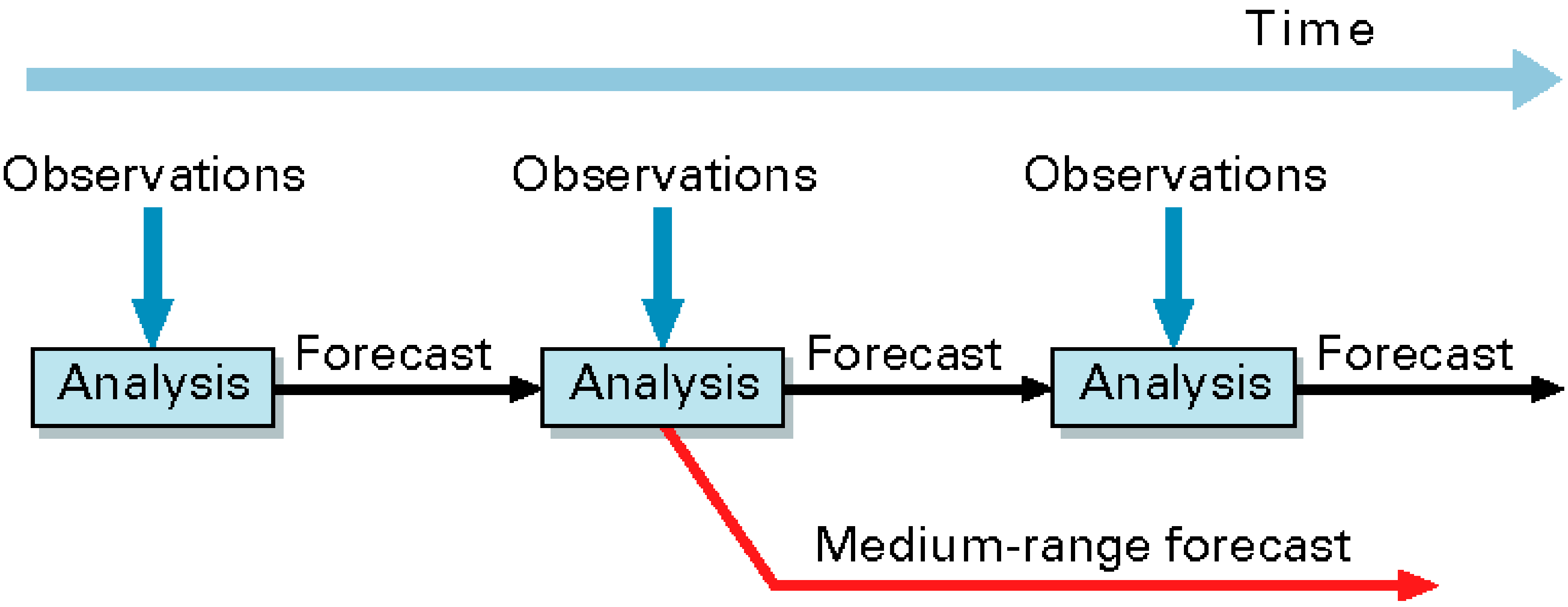
How large are the models? Typically $O(10^8)$ variables, and O(10) different physical processes



Lars Isaksen (http://www.ecmwf.int)

PoI: develop algorithms and implementations for large scale parallel machines, accelerator architectures

How does a data assimilation system work?

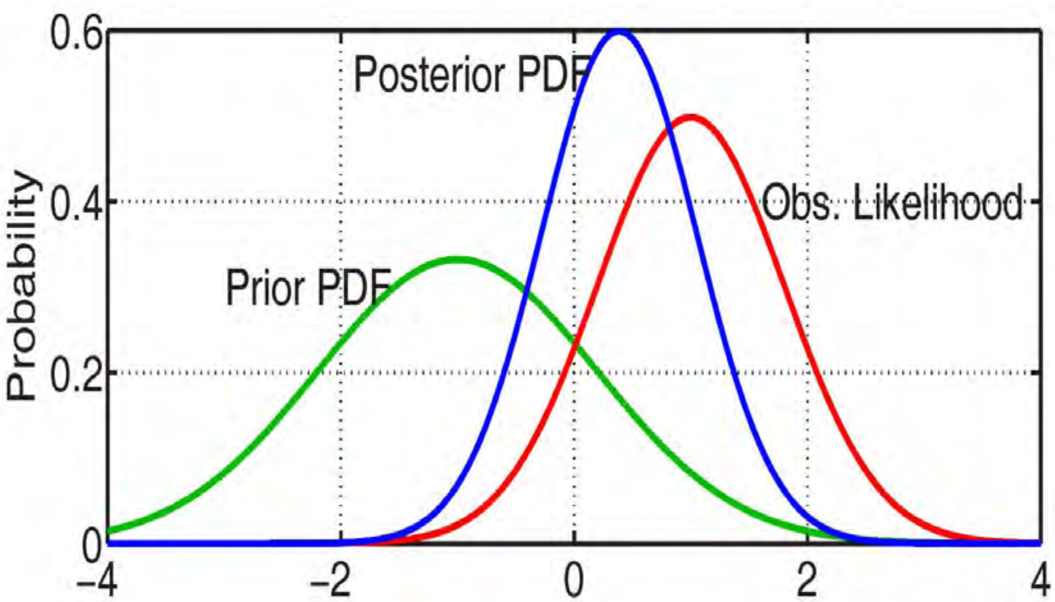


Lars Isaksen (http://www.ecmwf.int)

A **Bayesian framework** is employed to derive the analysis, which encapsulates all our knowledge

- The analysis (posterior) probability density $\mathcal{P}^{\mathrm{a}}(\mathbf{x})$:

$$\text{Bayes:} \qquad \mathcal{P}^{\mathrm{a}}(\mathbf{x}) = \mathcal{P}(\mathbf{x}|\mathbf{y}) = \frac{\mathcal{P}(\mathbf{y}|\mathbf{x}) \cdot \mathcal{P}^{\mathrm{b}}(\mathbf{x})}{\mathcal{P}(\mathbf{y})}\,.$$



[Picture from J.L. Anderson]

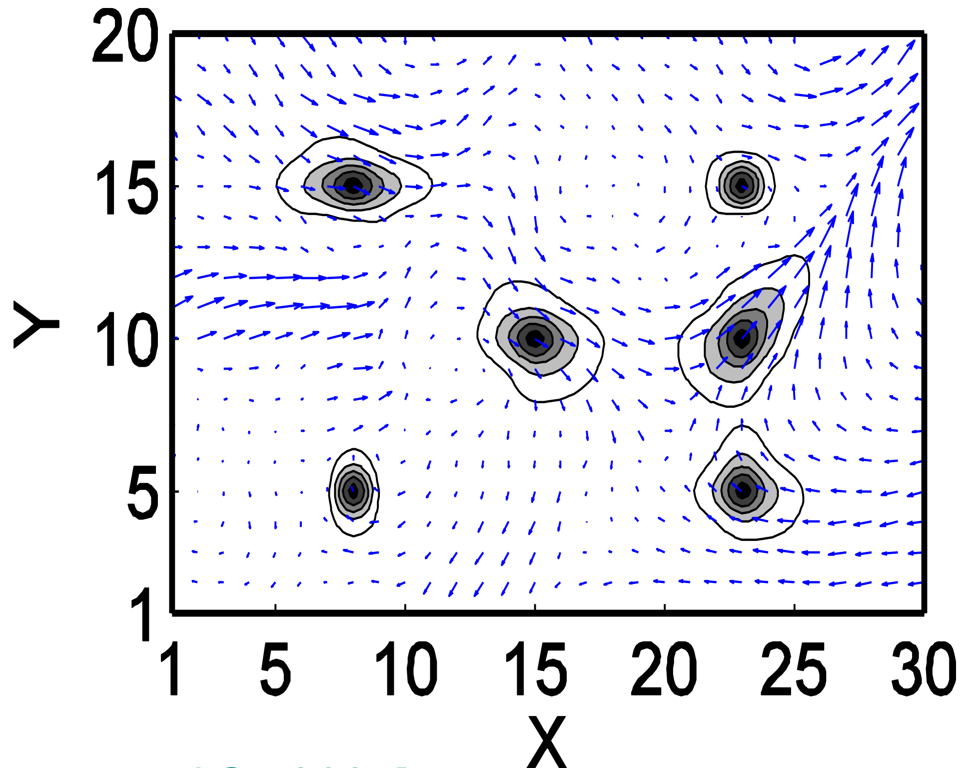
Practical families of methods:

- Suboptimal KF
  (estimation theory;
  min. variance)
- Variational
  (control theory;
  max. likelihood)

PoI: build correct, and computationally efficient, models to quantify background (prior) errors

$$\log P = -\tfrac{1}{2}\left(\mathbf{x}^0 - \mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}^0 - \mathbf{x}^b\right) + \ldots$$

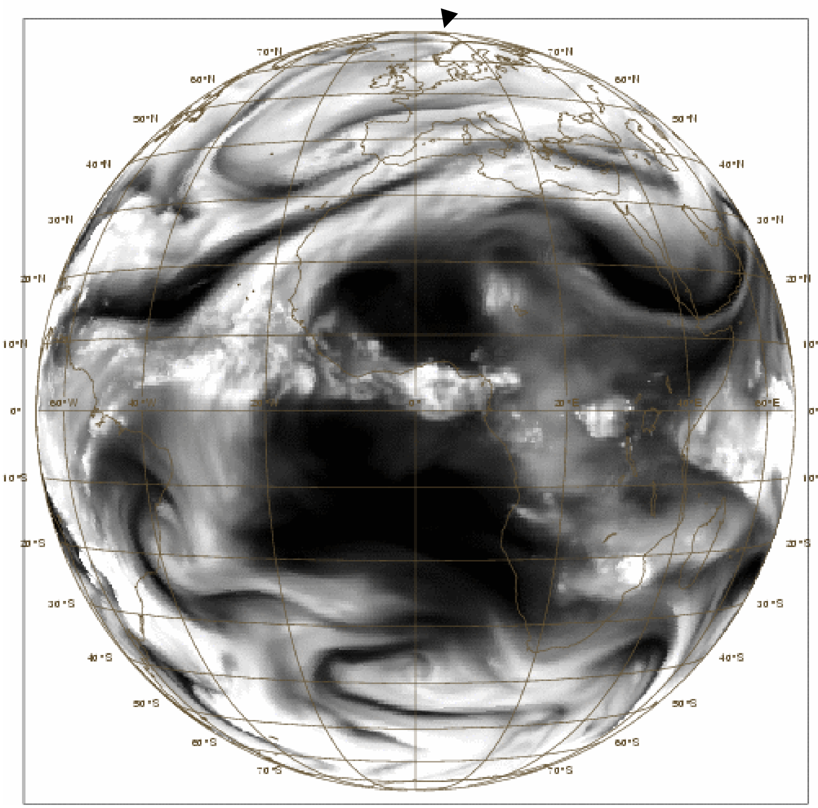
- Background error representation determines the spread of information, and impacts the assimilation results
- Needs: high rank, capture dynamic dependencies, efficient computations
- Traditionally estimated empirically (NMC, Hollingsworth-Lonnberg)

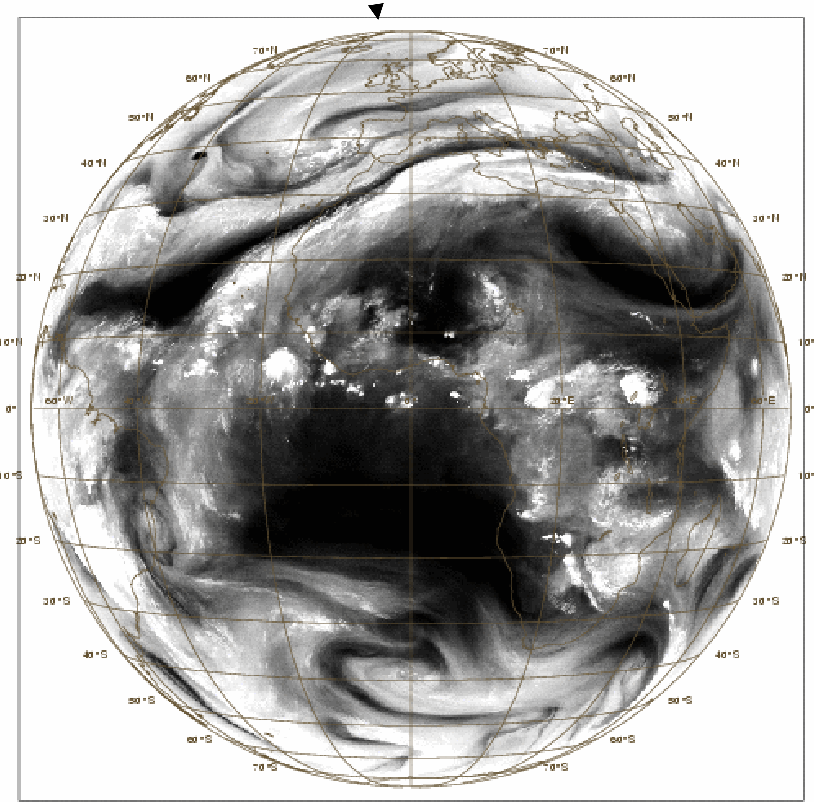


[Constantinescu and S., 2007]

To allow model-data comparison, **observation operators** map the model state space to observation space

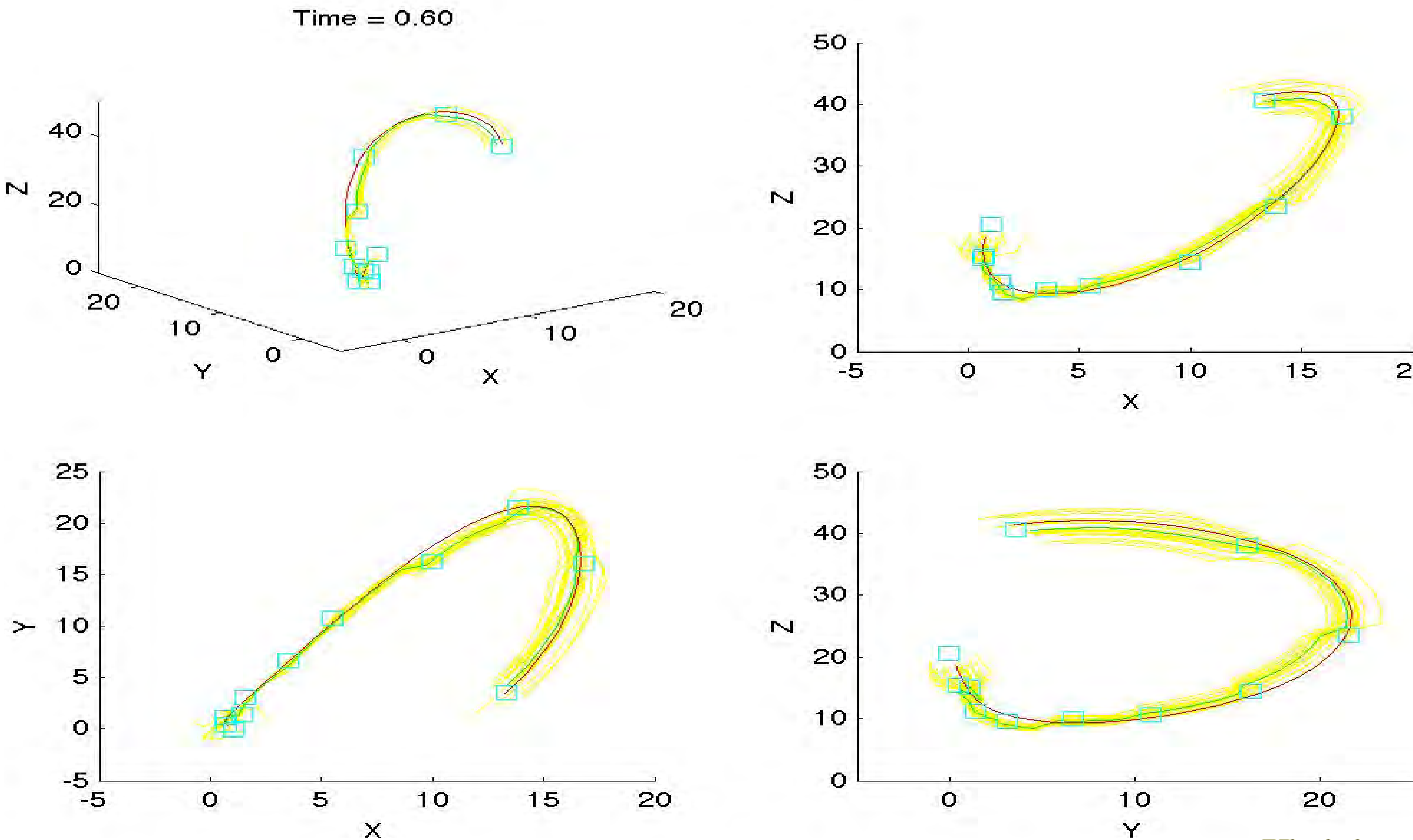
Model T and q → H → Model Radiance ⟷ J_o compare ⟷ Observation Satellite Radiance





Lars Isaksen (http://www.ecmwf.int)

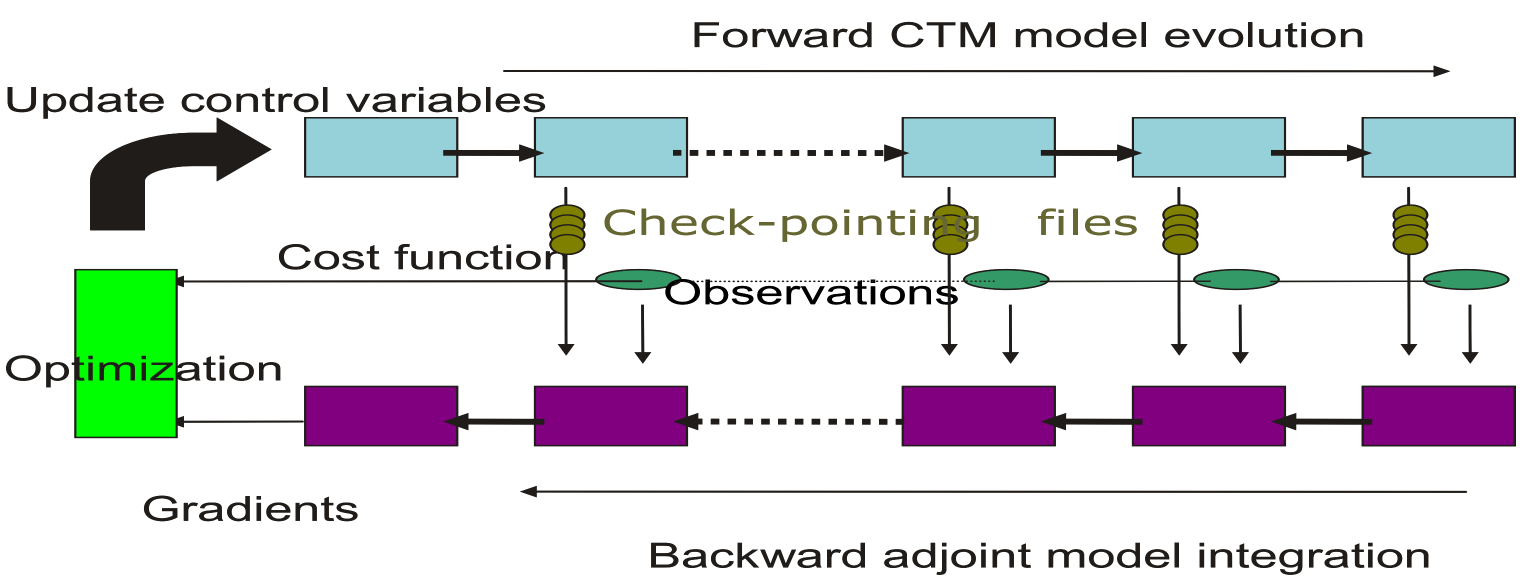
Practical approach: KF too expensive for large scale models; EnKF, PF use MC for covariance equations



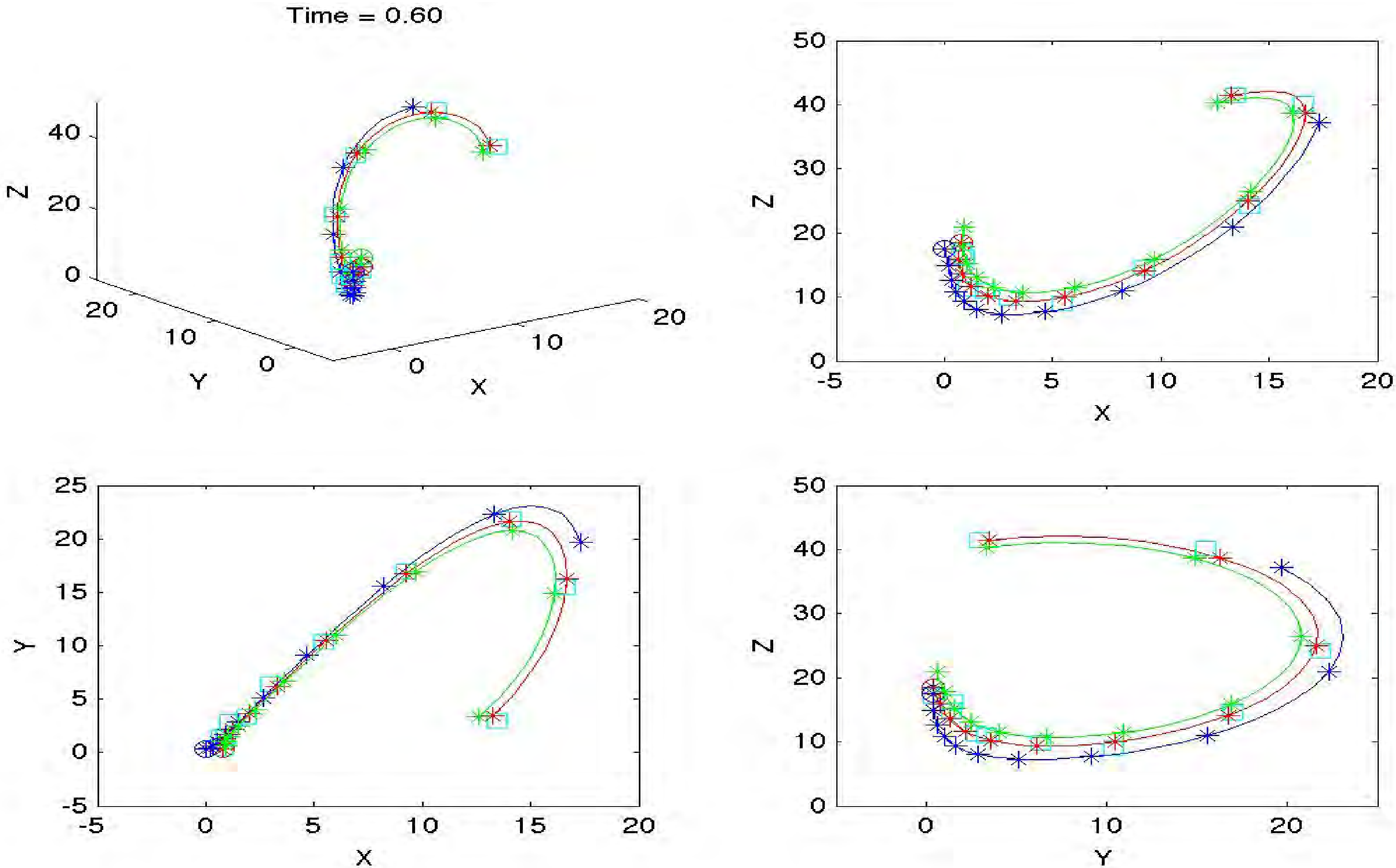
Practical approach: MAP estimator calculates the most likely state conditioned by observations

- 4D-Var MAP estimate via model-constrained optimization problem

$$
\begin{aligned}
\mathcal{J}(\mathbf{x}_0) &= \frac{1}{2}\left\|\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}}\right\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\sum_{i=1}^{N}\left\|\mathcal{H}(\mathbf{x}_i) - \mathbf{y}_i\right\|^2_{\mathbf{R}_i^{-1}} \\
\mathbf{x}_0^{\mathrm{a}} &= \arg\min \mathcal{J}(\mathbf{x}_0) \\
&\quad \text{subject to: } \mathbf{x}_i = \mathcal{M}_{t_0 \to t_i}(\mathbf{x}_0)\,, \quad i = 1, \cdots, N
\end{aligned}
$$



Example: The Lorenz three-variable system. 4D-Var solution, 2 optimization iterations



PoI (4D-Var DA): constructing adjoints is work-intensive, error-prone. Automatic implementation (KPP)

Chemical mechanism

```
#INCLUDE atoms

#DEFVAR
O  =  O;  O1D  =  O;
O3  =  O + O + O;
NO  =  N + O;
NO2  =  N + O + O;

#DEFFIX
O2  =  O + O;  M  =  ignore;

#EQUATIONS  { Small Stratospheric }
O2  + hv = 2O          :  2.6E-10*S;
O   + O2 = O3          :  8.0E-17;
O3  + hv = O   + O2    :  6.1E-04*S;
O   + O3 = 2O2         :  1.5E-15;
O3  + hv = O1D + O2    :  1.0E-03*S;
O1D  + M  = O   + M    :  7.1E-11;
O1D  + O3 = 2O2        :  1.2E-10;
NO   + O3 = NO2 + O2 :  6.0E-15;
NO2  + O  = NO  + O2   :  1.0E-11;
NO2  + hv = NO  + O    :  1.2E-02*S;
```

K P P

Simulation code

```
SUBROUTINE FunVar ( V,  F, RCT, DV )
    INCLUDE 'small.h'
     REAL*8 V(NVAR), F(NFIX)
     REAL*8 RCT(NREACT), DV(NVAR)
C A - rate for each equation
     REAL*8 A(NREACT)
C Computation of equation rates
     A(1) = RCT(1)*F(2)
     A(2) = RCT(2)*V(2)*F(2)
     A(3) = RCT(3)*V(3)
     A(4) = RCT(4)*V(2)*V(3)
     A(5) = RCT(5)*V(3)
     A(6) = RCT(6)*V(1)*F(1)
     A(7) = RCT(7)*V(1)*V(3)
     A(8) = RCT(8)*V(3)*V(4)
     A(9) = RCT(9)*V(2)*V(5)
     A(10) = RCT(10)*V(5)
C Aggregate function
     DV(1) = A(5)-A(6)-A(7)
     DV(2) = 2*A(1)-A(2)+A(3)-A(4)+A(6)-&A(9)+A(10)
     DV(3) = A(2)-A(3)-A(4)-A(5)-A(7)-A(8)
     DV(4) = -A(8)+A(9)+A(10)
     DV(5) = A(8)-A(9)-A(10)
     END
```

[Damian et.al., 1996; S. et.al., 2002]

Challenge: sensitivity, optimization carried out with the discrete model, approximate continuous solutions?

Sensitivity analysis: how well does the derivative of the numerical solution approximate the continuous derivative?

$$\text{Compute} \quad \nabla J^h \quad \text{to represent} \quad \nabla J$$

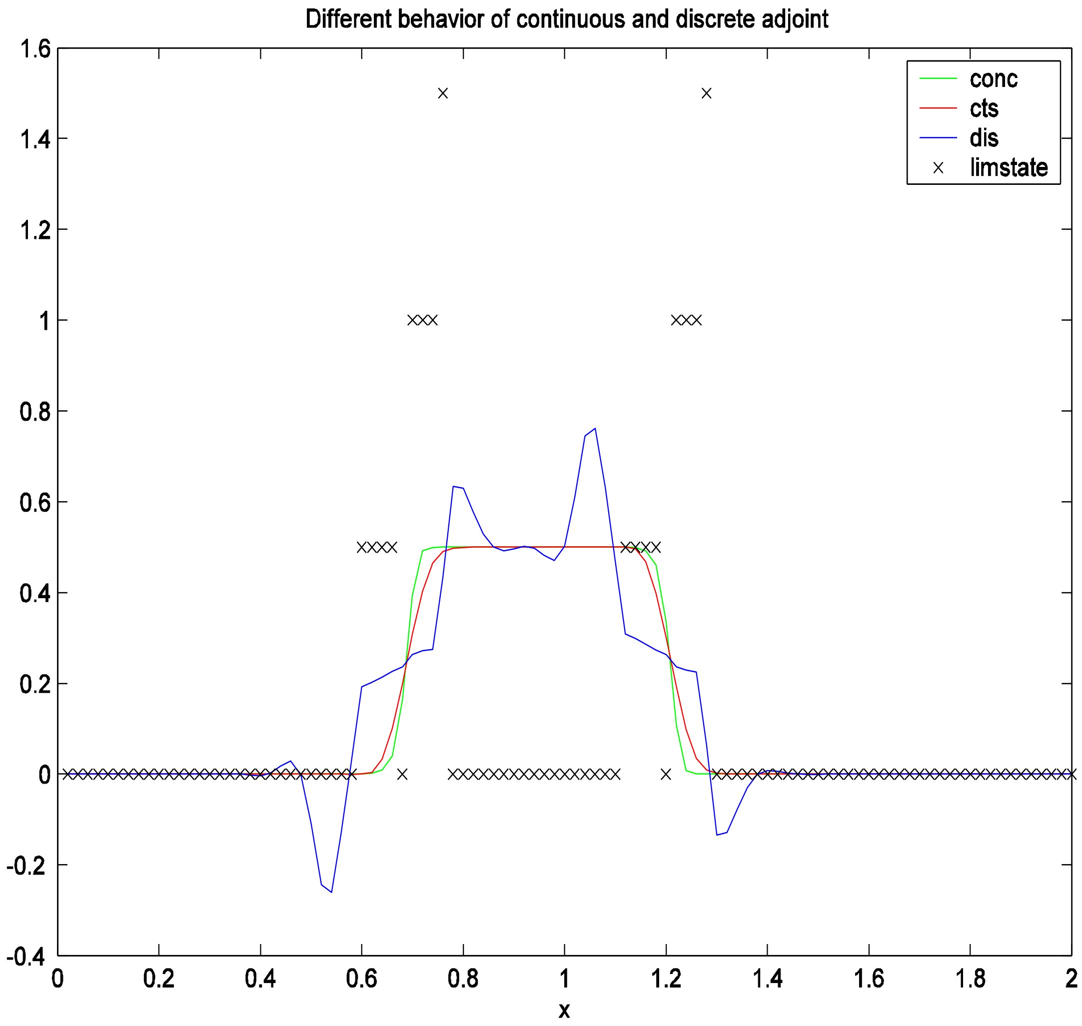
Inverse problems: how well does the discrete optimum approximate the continuous optimum?

$$\mathbf{x}_{\text{opt}} = \text{argmin}_{\mathbf{x}_0} J$$

$$\mathbf{x}_{\text{opt}}^{\text{h}} = \text{argmin}_{\mathbf{x}_0} J^{\text{h}}$$

$$\left\| \mathbf{x}_{\text{opt}}^{\text{h}} - \mathbf{x}_{\text{opt}} \right\| \le \text{cond}\left(\nabla^2 J(\mathbf{x}_{\text{opt}}) \right) \cdot \left\| \nabla J^{\text{h}} - \nabla J \right\|$$

Challenge: continuous and discrete adjoints lead to different computational models



Active forward limiters act as pseudo-sources in adjoint

Example: minmod

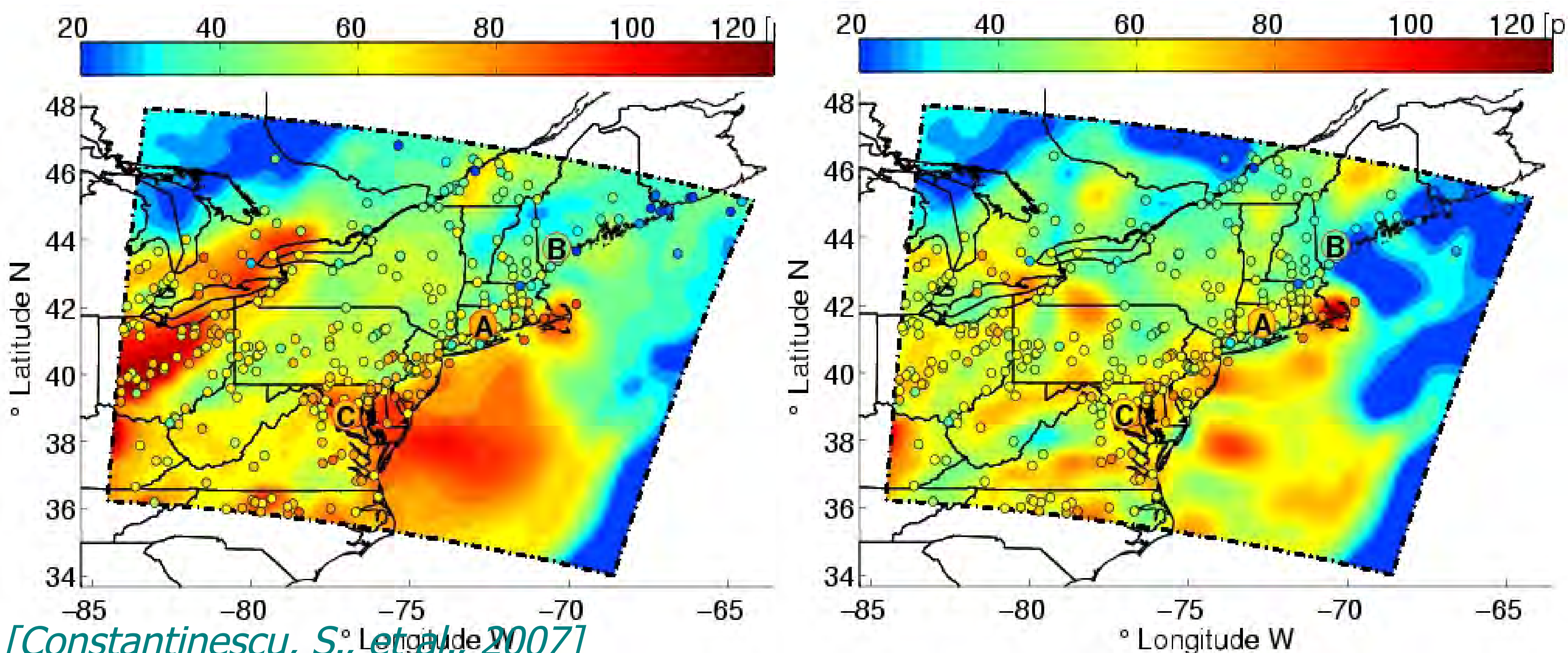
[Liu and S., 2005]

Example: LEnKF assimilation of ozone data from the ICARTT field campaign in Eastern U.S., July 2004

Ground level ozone at 2pm EDT, July 20, 2004

Observations: circles, color coded by O_3 mixing ratio

Forecast (R^2=0.24/0.28) Analysis (R^2=0.88/0.32)



[Constantinescu, S., et al., 2007]

PoI (ensemble DA): represent uncertainty in many dimensions via small ensembles

$$\mathbf{x}_f^k = M\left(t^{k-1}, \mathbf{x}_a^{k-1}\right)$$

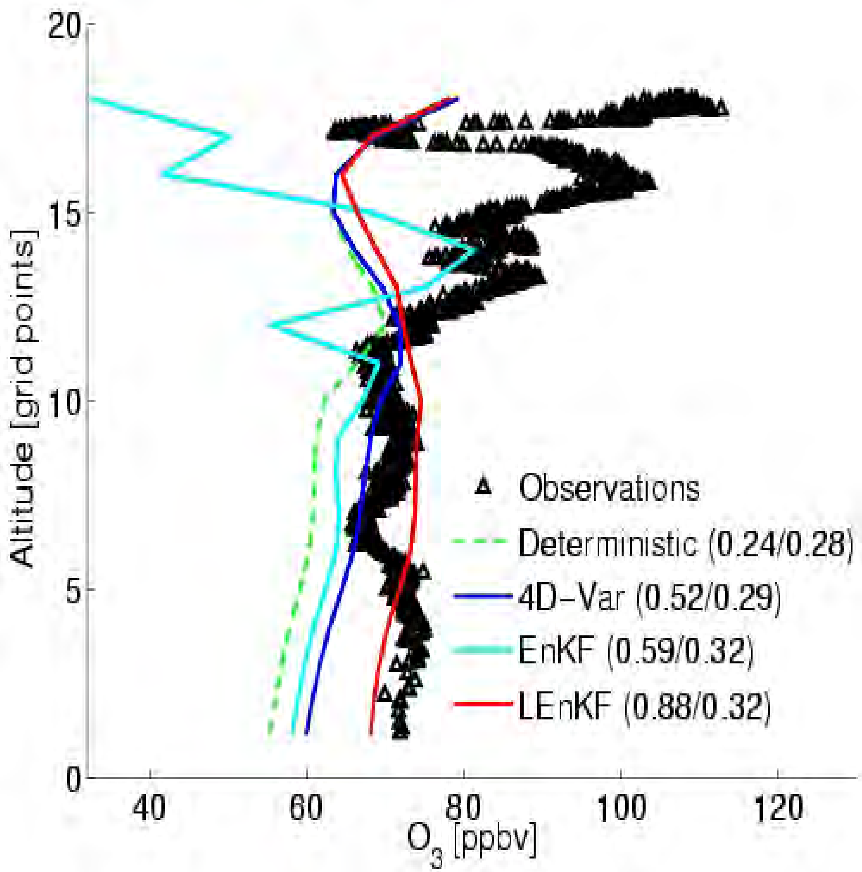
$$\mathbf{x}_a^k = \mathbf{x}_f^k + \mathbf{P}_f^k \mathbf{H}_k^T \left(\mathbf{R}_k + \mathbf{H}_k \mathbf{P}_f^k \mathbf{H}_k^T\right)^{-1} \left(\mathbf{y}_{obs}^k - \mathbf{H}_k \mathbf{x}_f^k\right)$$

Specify initial ensemble (sample B)

Covariance inflation: Prevents filter divergence (additive, multiplicative, model-specific)

Covariance localization (limit long-distance spurious correlations)

Correction localization (limit increments away from observations)

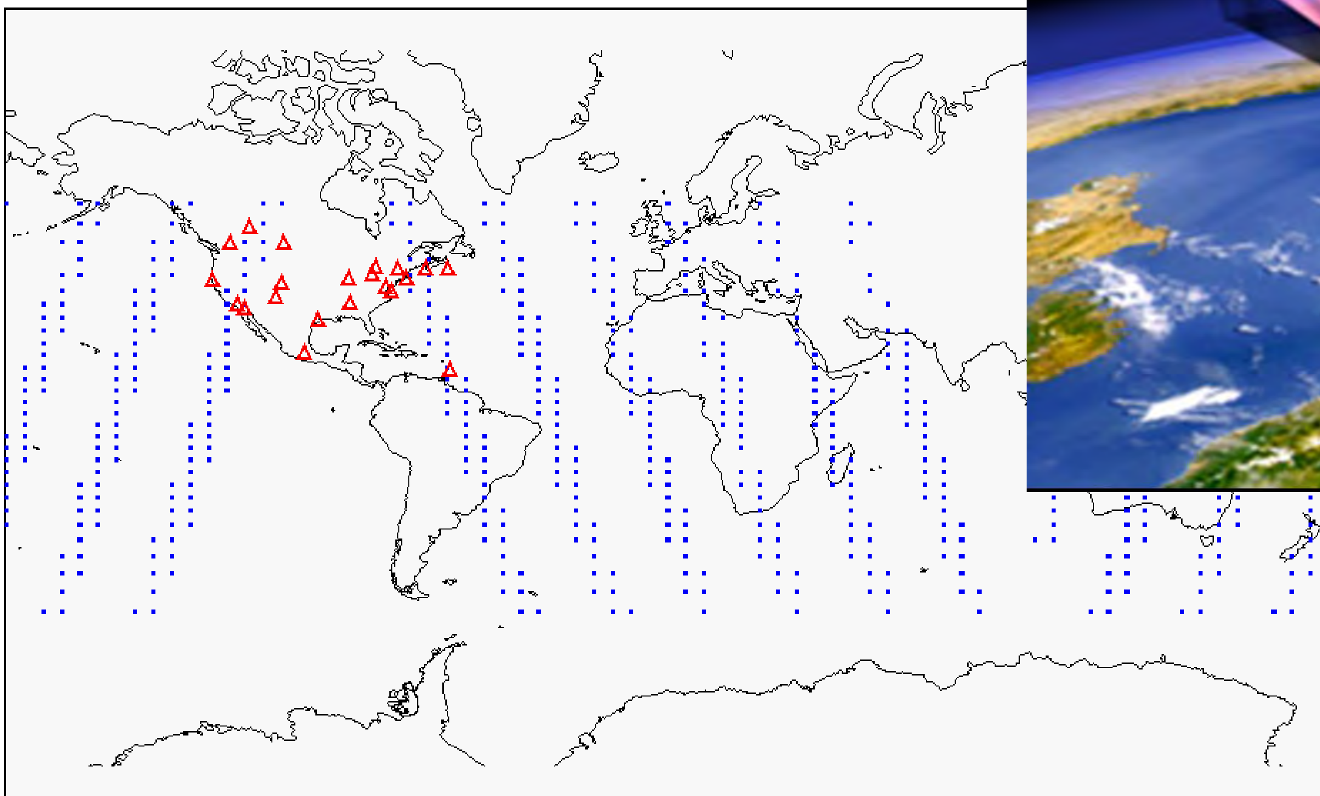


[Constantinescu, S., et al., 2007]

Ozonesonde S2 (18 EDT, July 20, 2004)

Example: 4D-Var assimilation of TES ozone column, Aug. 2006. Validation against IONS-6 ozonesonde.

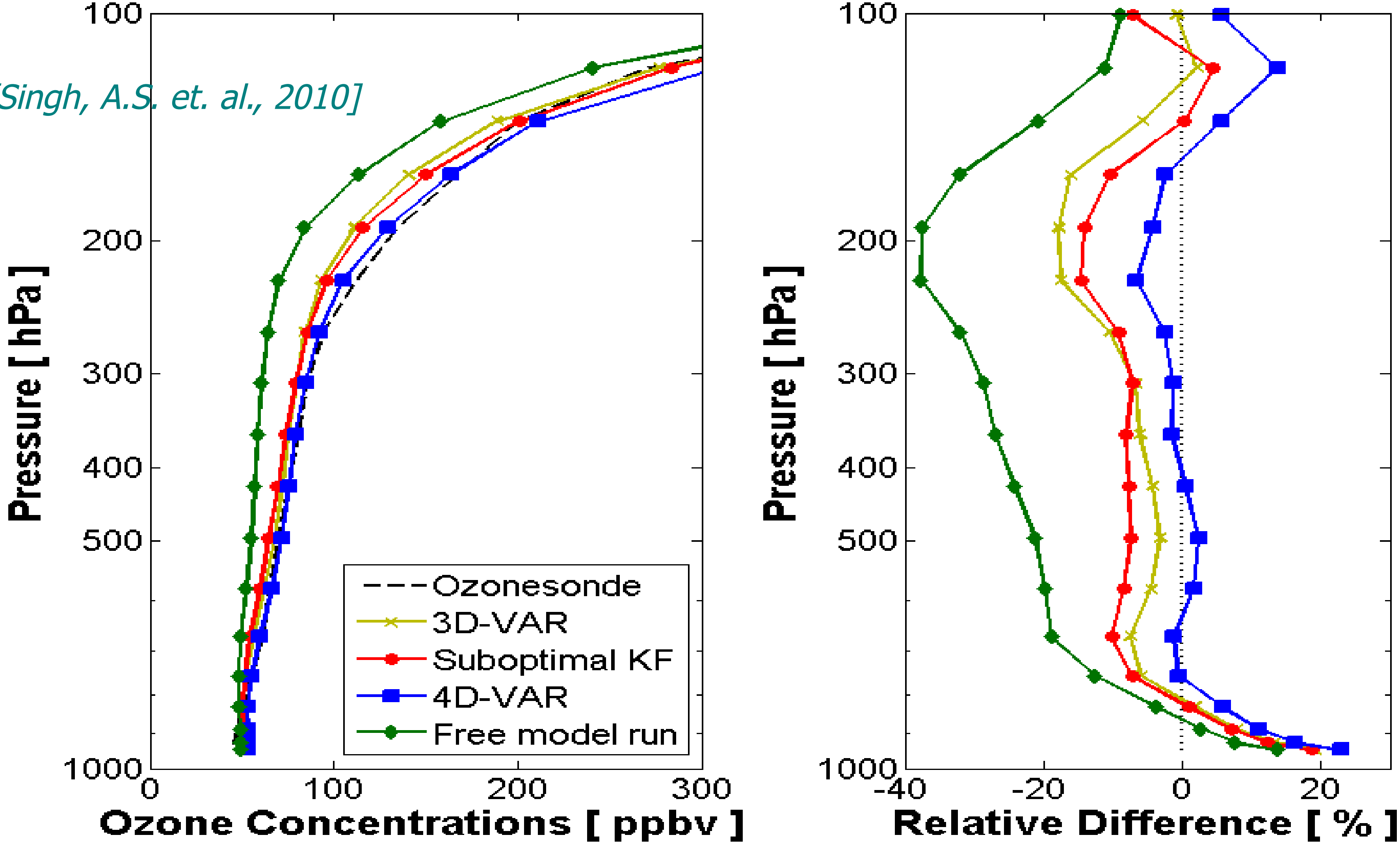
TES is one of four instruments on the NASA EOS Aura platform, launched July 14 2004



[Singh, S., et al., 2010]

Quality of TES ozone column assimilation results for several DA methods (August 1-15, 2006)

[Singh, A.S. et. al., 2010]

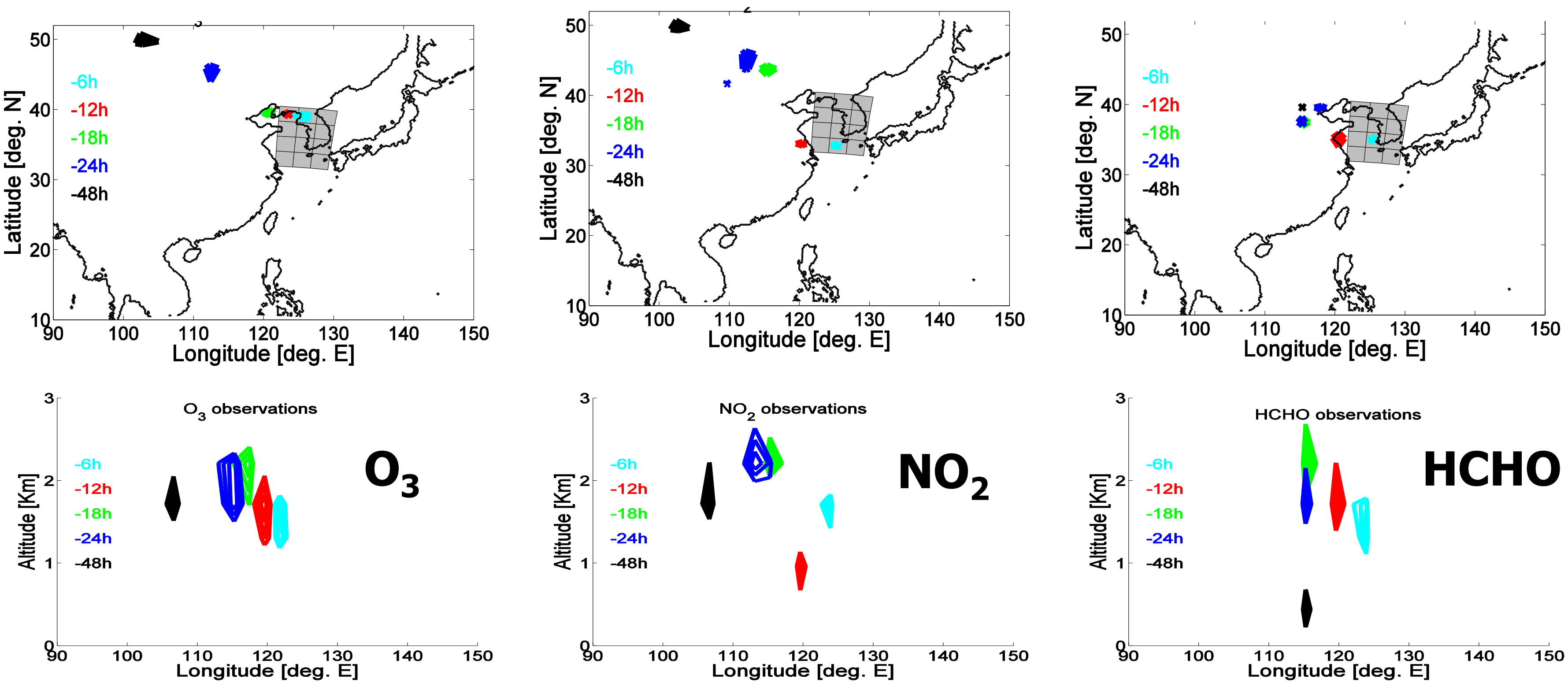


PoI: develop algorithms to configure the sensor network such as to maximize the information benefit

$$T = \sum_{k \geq 1} \frac{\sigma_k^2}{\sigma_{\max}^2} s_k^2$$

(criterion based on SVs)

Verification:
Korea, ground O_3
0 GMT, Mar/4/2001



[S., 2006]