Understanding data assimilation:
how observations and a model are weaved into the analysis via statistics

Tomoko Matsuo
NCAR

Thanks: Doug Nychka, Jeff Anderson,
Kevin Raeder, Alain Caya, Art Richmond, Gang Lu …

NCAR

What is data assimilation?!

Combining Information

prior knowledge of the state of system

empirical or physical models (e.g. physical laws)

complete in space and time x

observations

directly measured or retrieved quantities

incomplete in space and time y_{o}

Bayes Theorem

"Bayesian statistics provides a coherent probabilistic framework for most DA approaches" [e.g., Lorenc, 1986]

prior knowledge $P(x) \sim N(x_\mathrm{f}, \mathbf{P}_\mathrm{f})$ $x = x_\mathrm{f} + \varepsilon_\mathrm{f}$

observations $P(y_\mathrm{o} \mid x) \sim N(H(x), \mathbf{R})$ $y_\mathrm{o} = H(x) + \varepsilon_\mathrm{o}$

Note: observations y conditioned upon the state x

posterior

$$P(x \mid y_\mathrm{o}) \propto P(y_\mathrm{o} \mid x) P(x)$$

$$P(x \mid y_\mathrm{o}) \sim N(x_\mathrm{a}, \mathbf{P}_\mathrm{a})$$

H is linear

where $x_\mathrm{a} = x_\mathrm{f} + \mathbf{K}(y - \mathbf{H}x_\mathrm{f})$

$$\mathbf{P}_\mathrm{a} = (\mathbf{I} - \mathbf{KH})\mathbf{P}_\mathrm{f}$$

Assumption: Normal Distribution

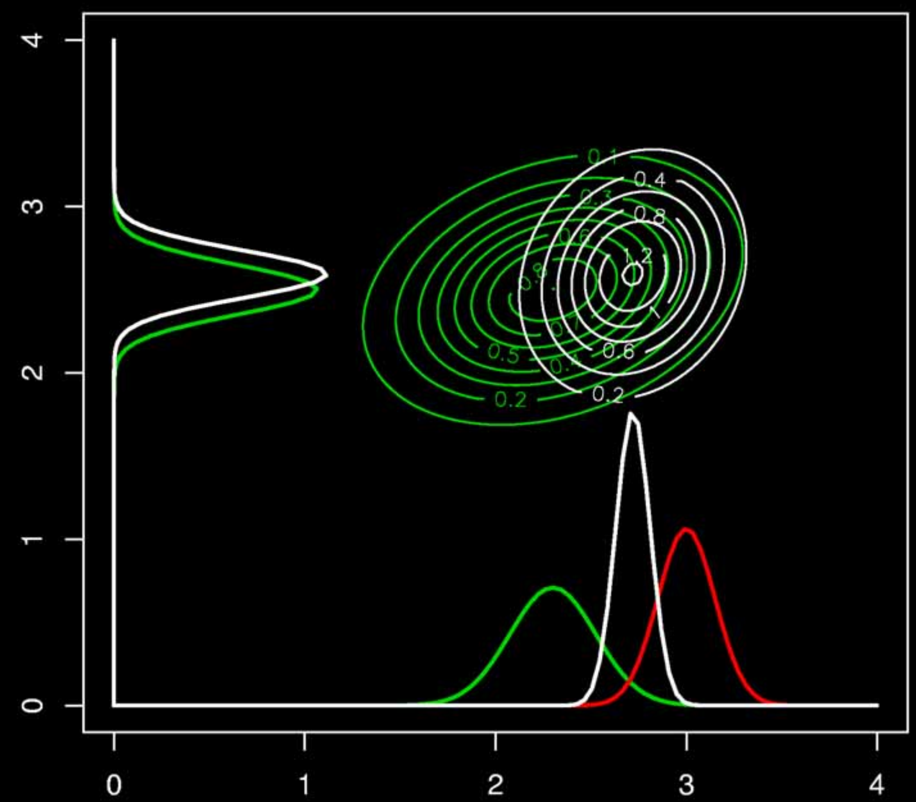
Importance of Covariance [e.g. Rodgers, 2000]

prior $P(x) \sim N(x_f, \mathbf{P}_f)$

observations $P(y_o \mid x) \sim N(\mathbf{H}x, \mathbf{R})$

$\xrightarrow{\text{update}}$

posterior $P(x \mid y_o)$



$$x_f = \begin{pmatrix} 2.3 & 2.5 \end{pmatrix}$$

$$\mathbf{P}_f = \begin{pmatrix} 0.225 & 0.05 \\ 0.05 & 0.15 \end{pmatrix}$$

$$\mathbf{H} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

x_1 : observed

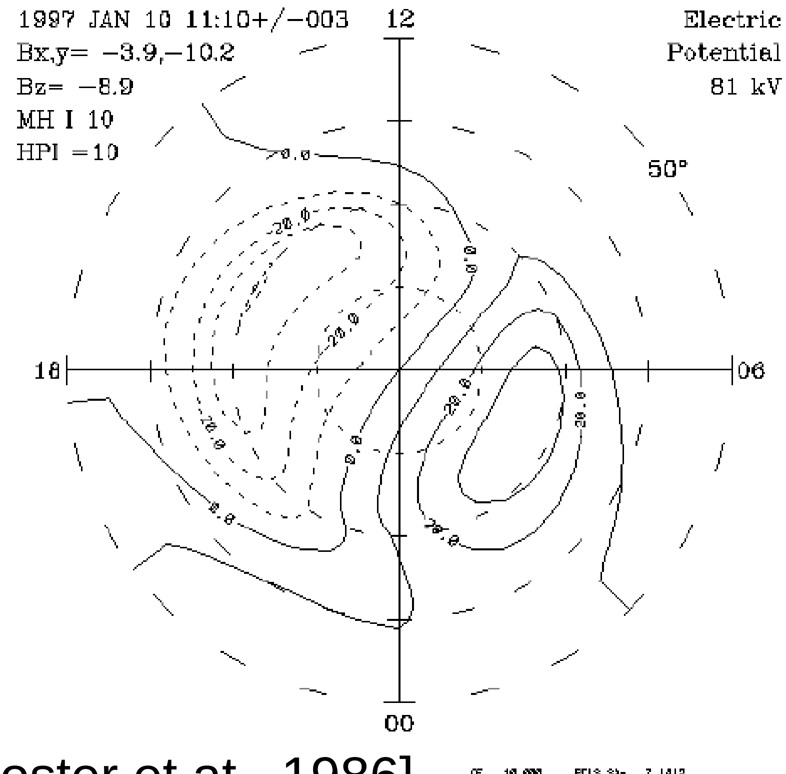
x_2 : unobserved

Assimilative Mapping of Ionospheric Electrodynamics

[Richmond and Kamide, 1988]

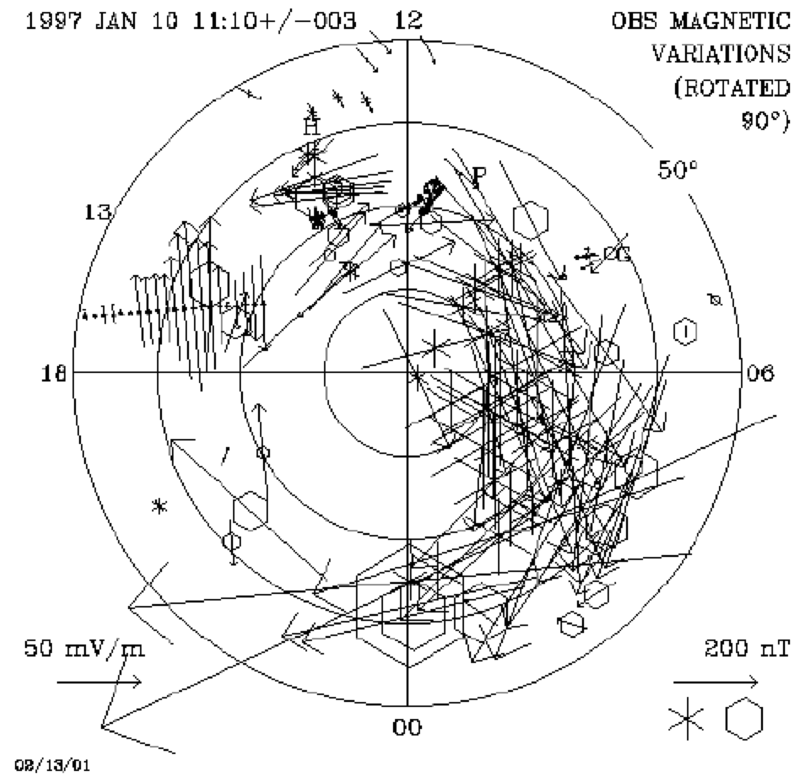
$$x_a = x_b + \mathbf{K}(y - \mathbf{H}x_b)$$

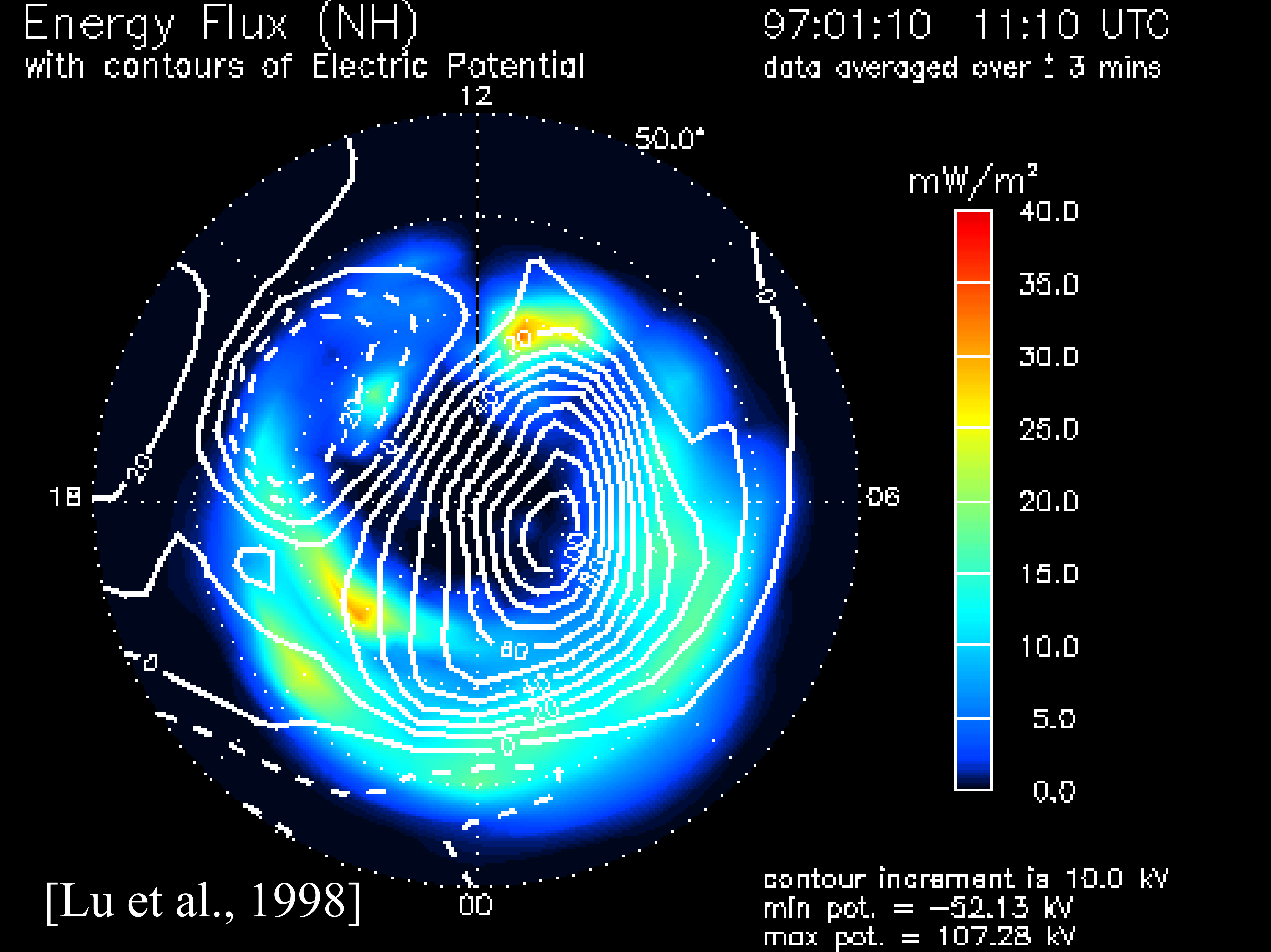
prior knowledge



[Foster et at., 1986]

observations




Energy Flux (NH)
with contours of Electric Potential
97:01:10 11:10 UTC
data averaged over ± 3 mins
12
50.0°
18
06
00
mW/m²
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
contour increment is 10.0 kV
min pot. = −52.13 kV
max pot. = 107.28 kV

[Lu et al., 1998]

Inhomogeneous / anisotropic covariance

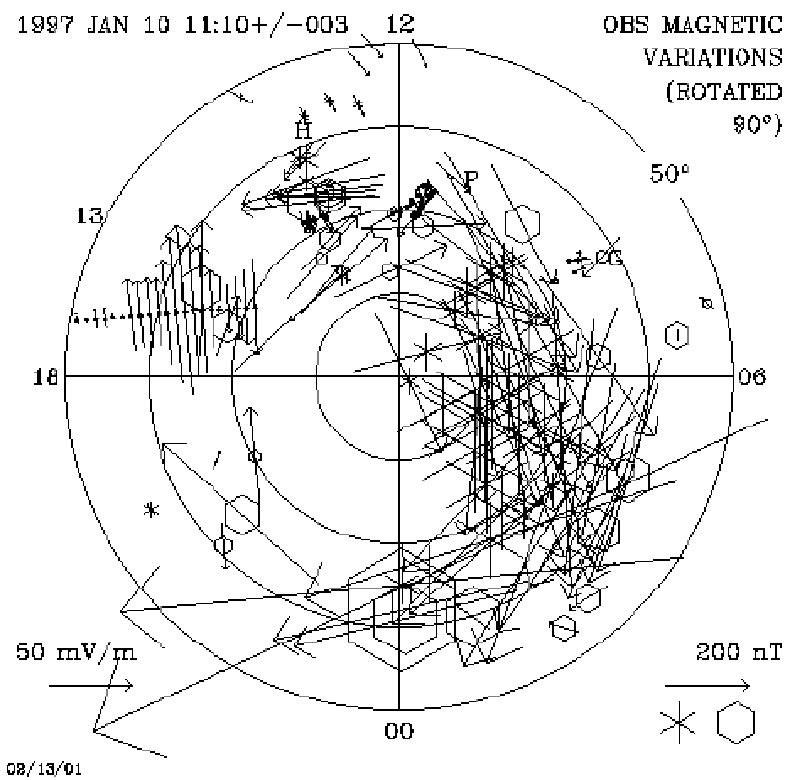
Adaptive Covariance Estimation Using Maximum likelihood Method

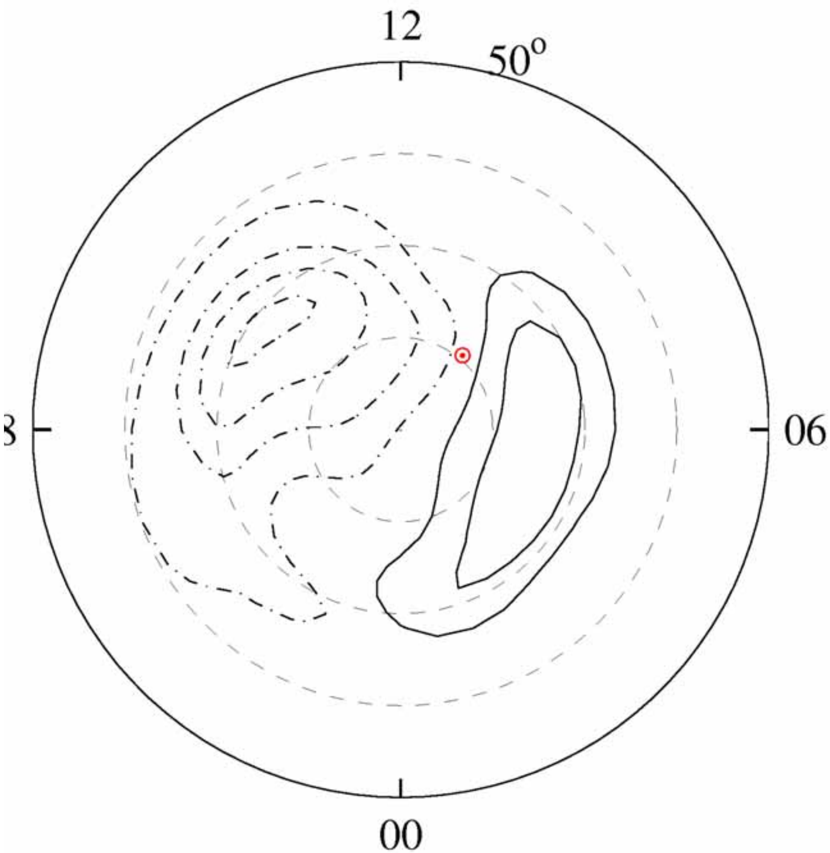
[Dee 1995; Dee and da Saliva 1999]

$$x_a = x_b + \mathbf{K}(y - \mathbf{H}x_b)$$

$$\mathbf{K} = \underline{\mathbf{P}_b(\alpha)}\mathbf{H}^T\left[\mathbf{H}\mathbf{P}_b(\alpha)\mathbf{H}^T + \mathbf{R}\right]^{-1}$$

<u>observations</u>





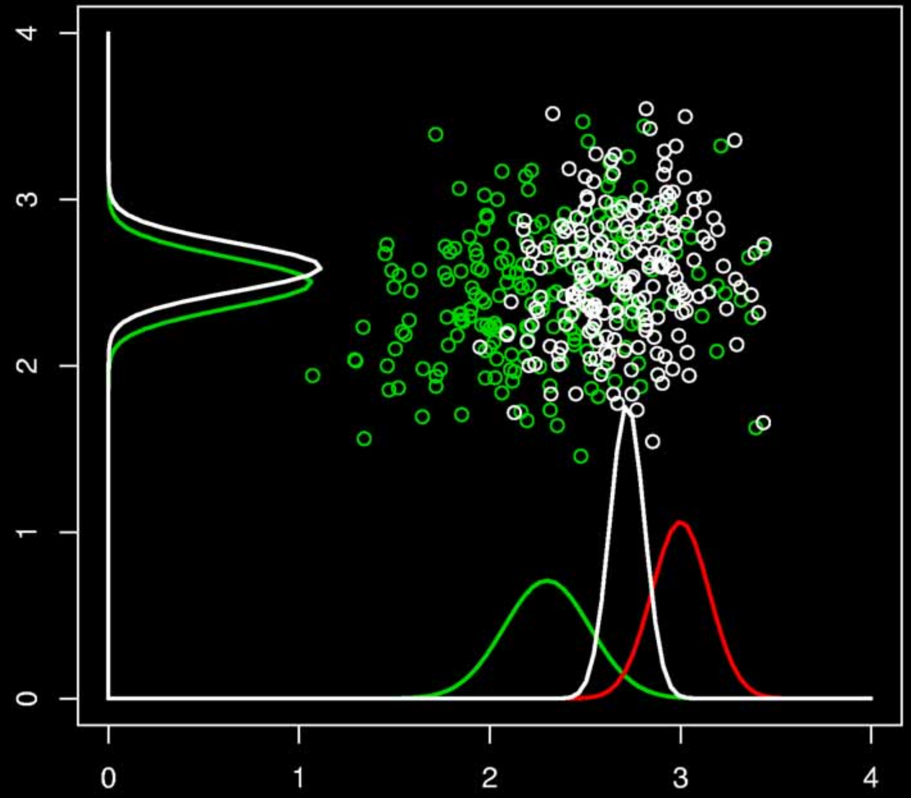
[Matsuo et at., 2002; 2005]

Use of Dynamics $x_{t+1} = M(x_t)$

$P(x_t \mid y_{t-1},\ldots)$ $P(y_t \mid x_t)$ $\xrightarrow{\text{update}}$ $P(x_t \mid y_t, y_{t-1},\ldots)$ $\xrightarrow[\text{forecast}]{M}$ $P(x_{t+1} \mid y_t, y_{t-1},\ldots)$ $P(y_{t+1} \mid x_{t+1})$

Ensemble Kalman FIlter

Let's work with samples!

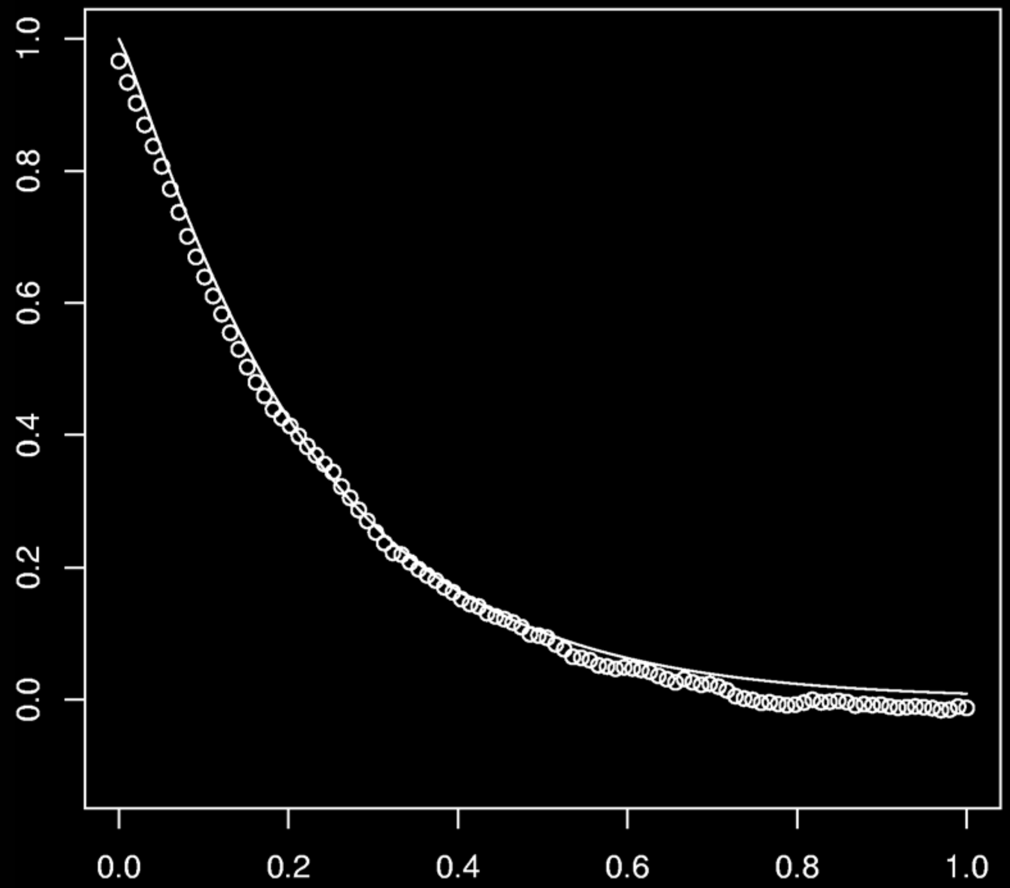


Challenge posed by the size of the covariance matrix (10^{12} –10^{16})

Issues with sampling error

[e.g., Furrer and Bengtsson, 2005]

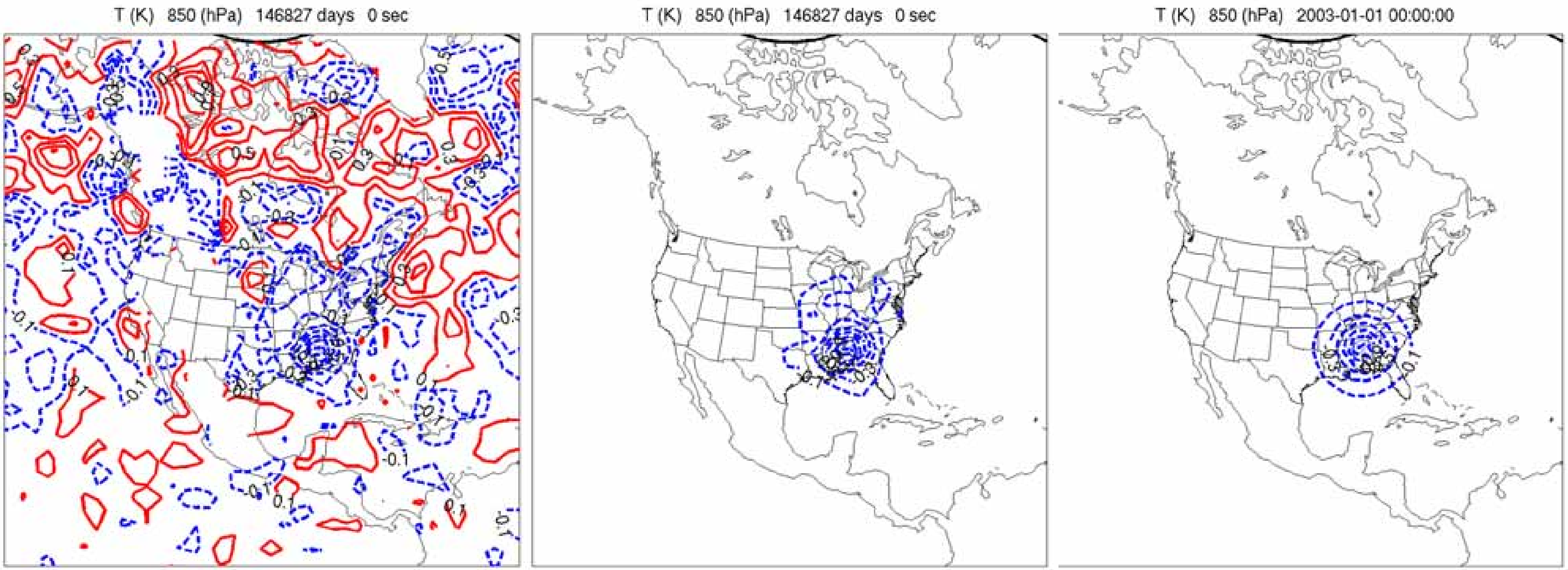
spurious correlations in the area of large lag distance.



EnKF v.s. 3D-Var comparisons [Caya et al., 2005]

$$x_a = x_f + \underline{\mathbf{K}(y - \mathbf{H}x_f)}$$

EnKF | With Localization | 3D-Var



Issue with sampling error: covariance localization (tapering) is necessary to remove spurious correlations in the area far from observation location.

Summary

- Bayesian statistics as an overarching framework.
- By confronting a model with observations via first/second moment statistics, data assimilation
  - *improves the state estimation.*
  - *provides a means to evaluate the quality of the model and the value of observations*.
- Inhomogeneous and anisotropic covariance.
- Ensemble Kalman Filter does not require linearization of forward operator (H) and model (M), and has an advantage in capturing flow-dependent covariance structure.
  - See *http://www.image.ucar.edu/DAReS/DART*

Challenges and Future

- Observation is still sparse... (blessing?!)

- Dissipative system and strongly forced system in comparison with meteorological and oceanic systems.

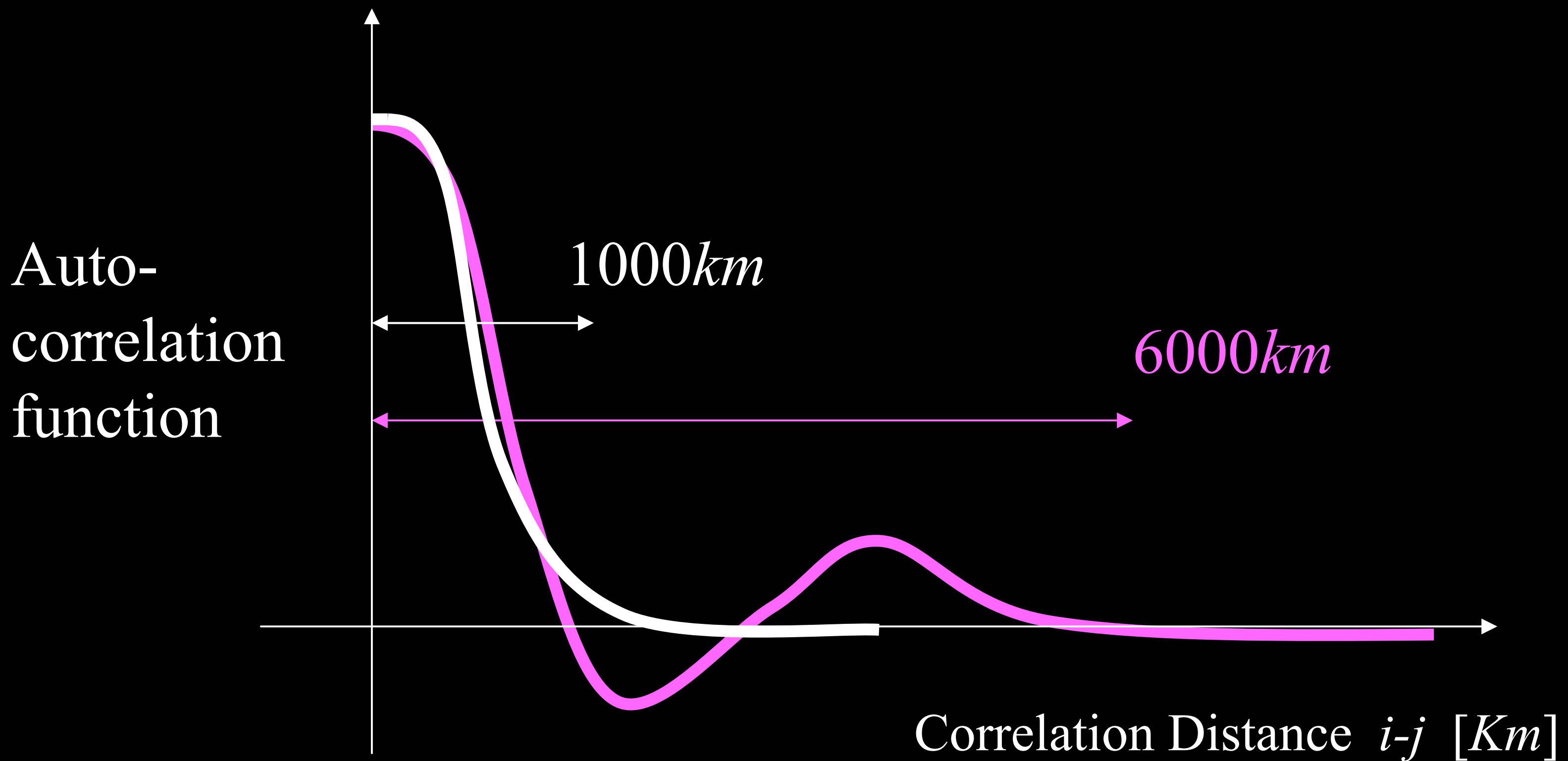
  (forcing prediction is key to forecasting)

- Observing system design analysis or adaptive observation [e.g., Bishop et al., 2001]. (feedback to the design of observational campaigns)

Why is data assimilation in a data sparse region challenging?

<u>Large Correlation Distance</u>

Inhomogeneous & Anisotropic



Adaptive covariance estimation using maximum likelihood method

OI analysis: optimal estimation of α

$$\alpha_a = \mathbf{K}^{OI}\mathbf{y}', \quad \text{where} \quad \mathbf{y}' = \mathbf{y}_o - \mathbf{H}(\mathbf{x}_b)$$

$$\mathbf{K}^{OI} = \left[(\mathbf{EOF})^T\mathbf{R}^{-1}\mathbf{EOF} + \mathbf{P}_b^{-1}\right]^{-1}(\mathbf{EOF})^T\mathbf{R}^{-1}$$

Background error covariance:

$$\mathbf{P}_b = \left\langle \alpha \cdot \alpha^T \right\rangle$$

$$\approx \mathrm{diag}(\mathbf{P}_b)_{\nu\nu} \approx \zeta_1 \nu^{-\zeta_2} \qquad \nu = 1, \ldots, 11$$

Observational error covariance:

$$\mathbf{R} \approx \mathrm{diag}(\mathbf{R}) \approx f(\zeta_3, \zeta_4)$$

Maximum-likelihood method: optimal estimation of ζ

[Dee 1995; Dee and da Silva, 1999]

Innovation covariance:

$$<\mathbf{y}' \cdot \mathbf{y}'^T> = \mathbf{R} + \mathbf{EOF}\,\mathbf{P}_b\,(\mathbf{EOF})^T \approx \mathbf{S}(\zeta) \quad \{\zeta_k \mid k = 1 \rightarrow 4\}$$

Cost function:

$$J(\zeta) = \log\det \mathbf{S}(\zeta) + \mathbf{y}'^T\mathbf{S}^{-1}(\zeta)\mathbf{y}'$$