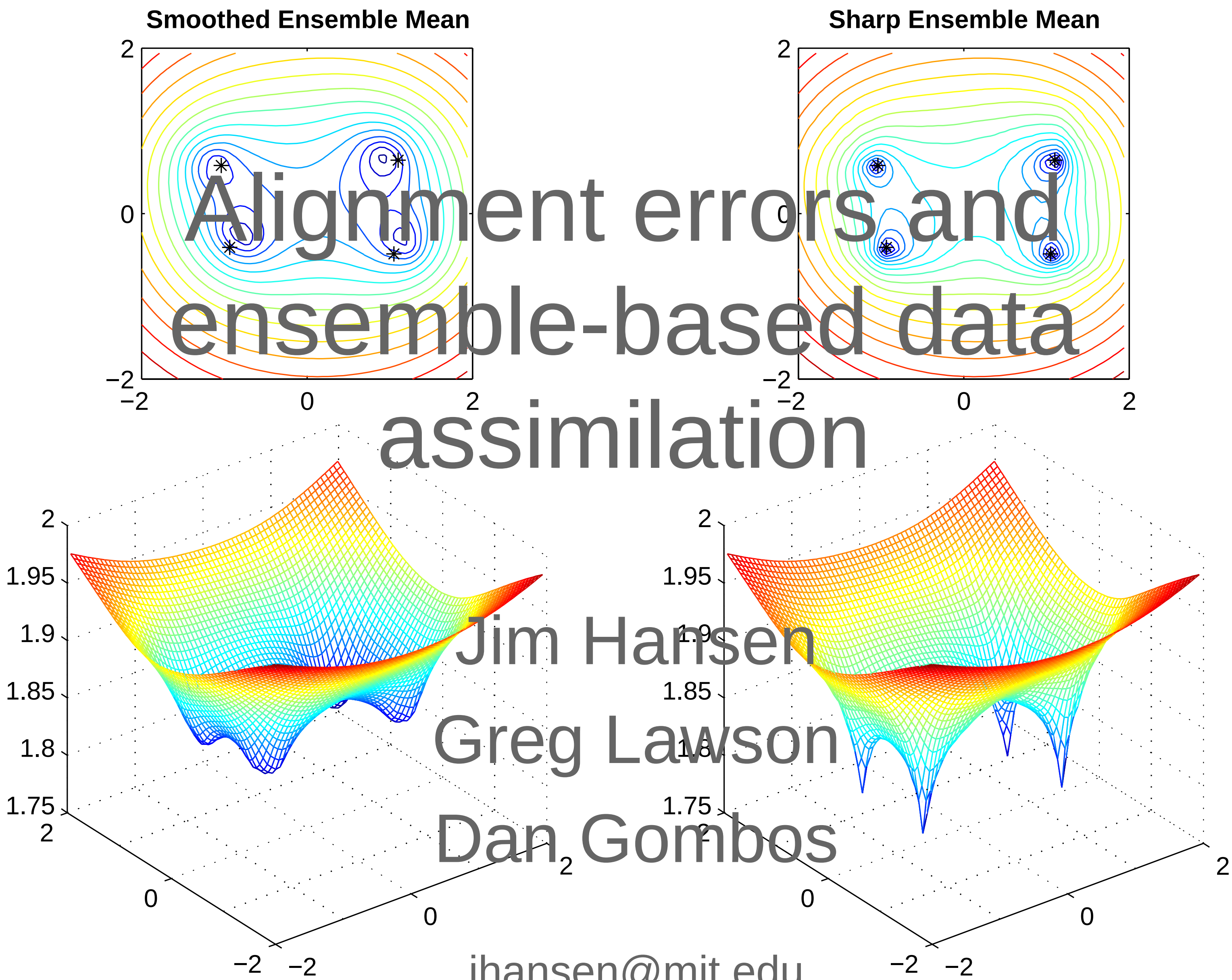


Alignment errors and ensemble-based data assimilation

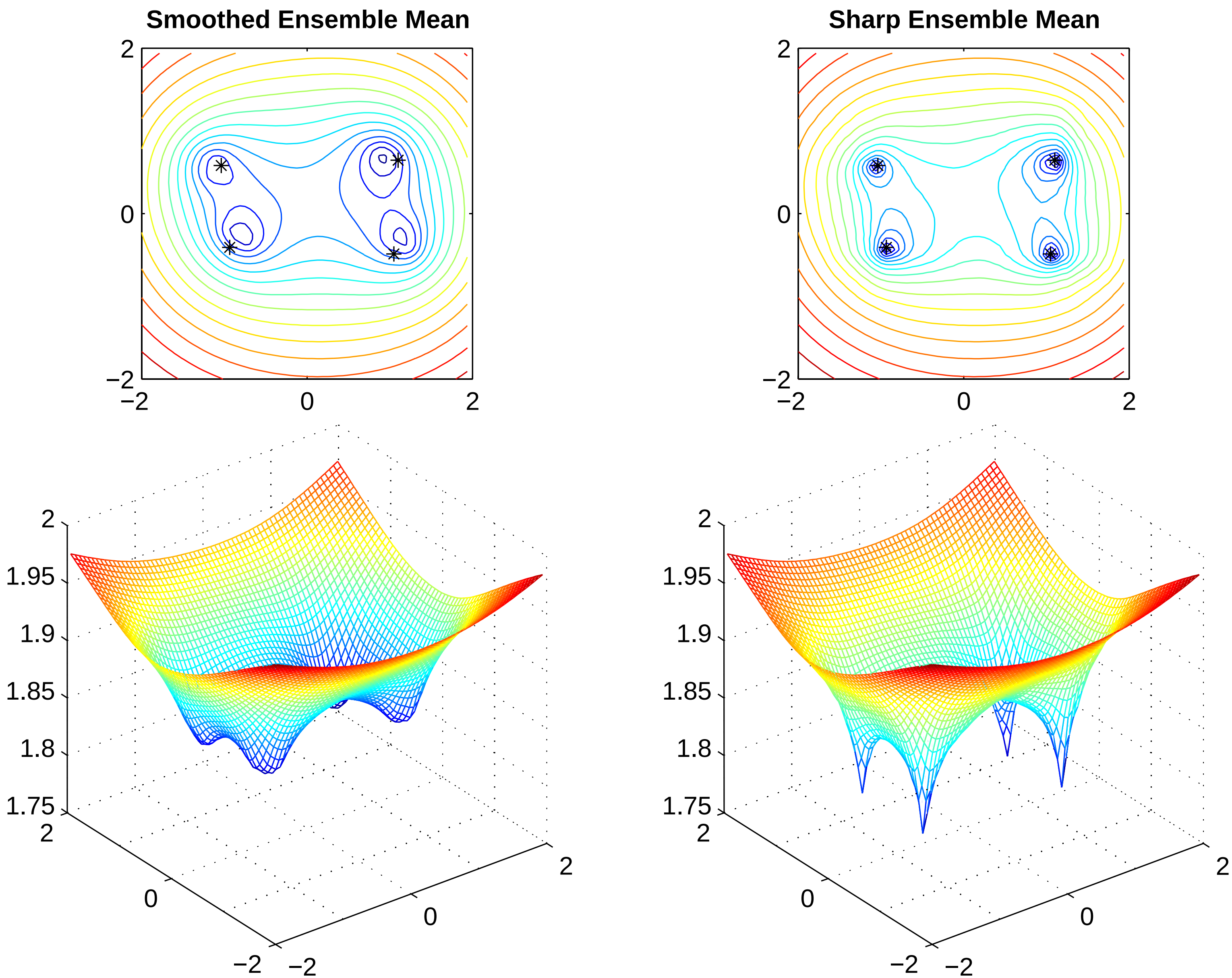
Jim Hansen

Greg Lawson

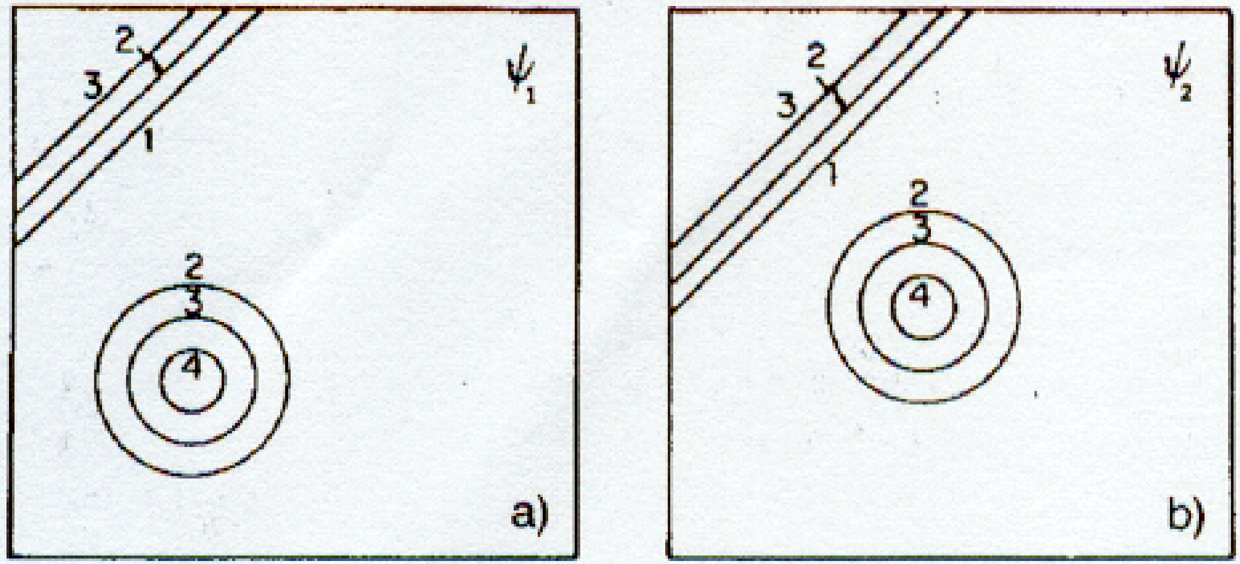
Dan Gombos

jhansen@mit.edu

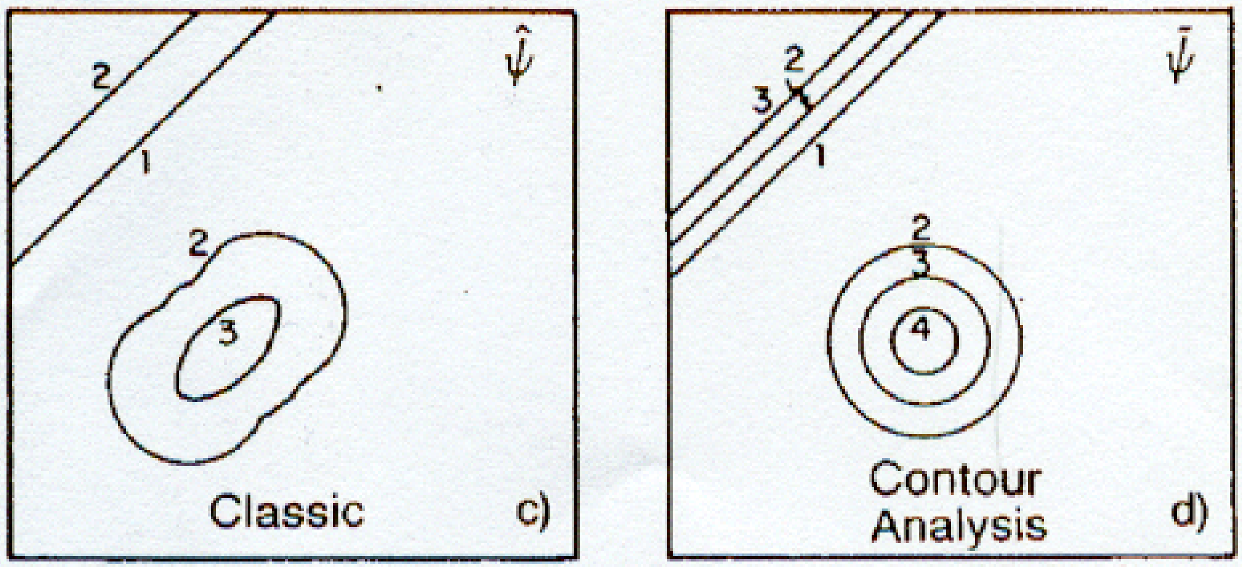
MASSACHUSETTS INSTITUTE OF TECHNOLOGY



Smoothed Ensemble Mean
Sharp Ensemble Mean
2
0
−2
−2
0
2
1.75
1.8
1.85
1.9
1.95



Mariano, JAOT, 1990

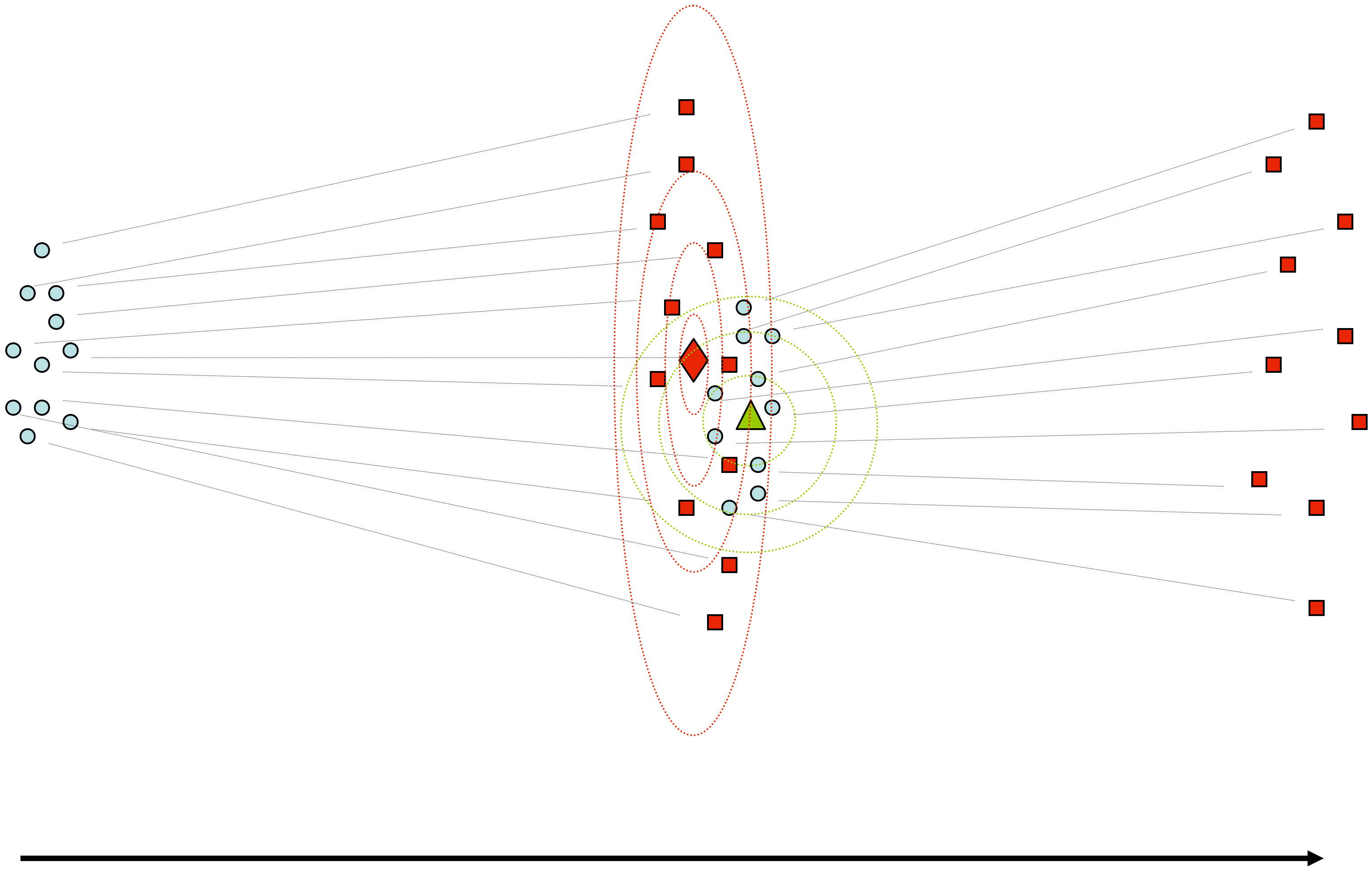


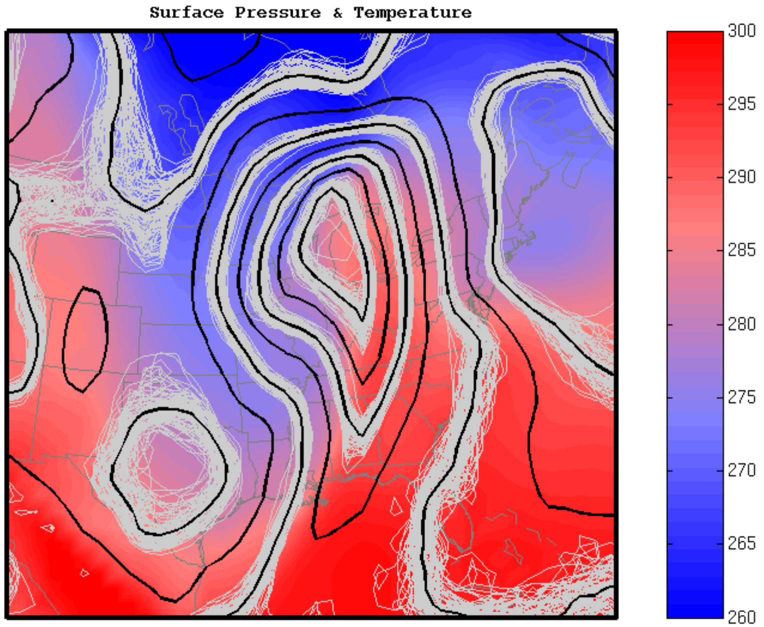
Outline

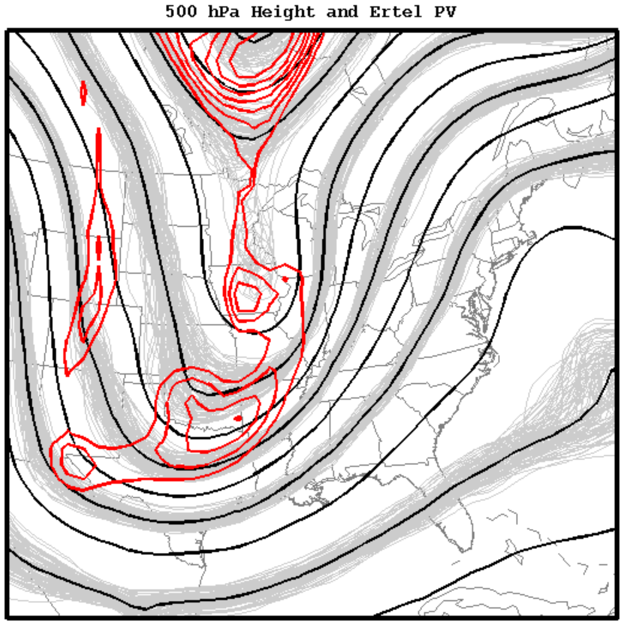
- The joy of ensemble covariances
  - Using ensemble DA as a tool to better understand the system
- Why mis-positioned features are a problem
- Error models to the rescue
- A two-step approach (position, then amplitude) to ensemble data assimilation
- Many examples showing this is a good thing

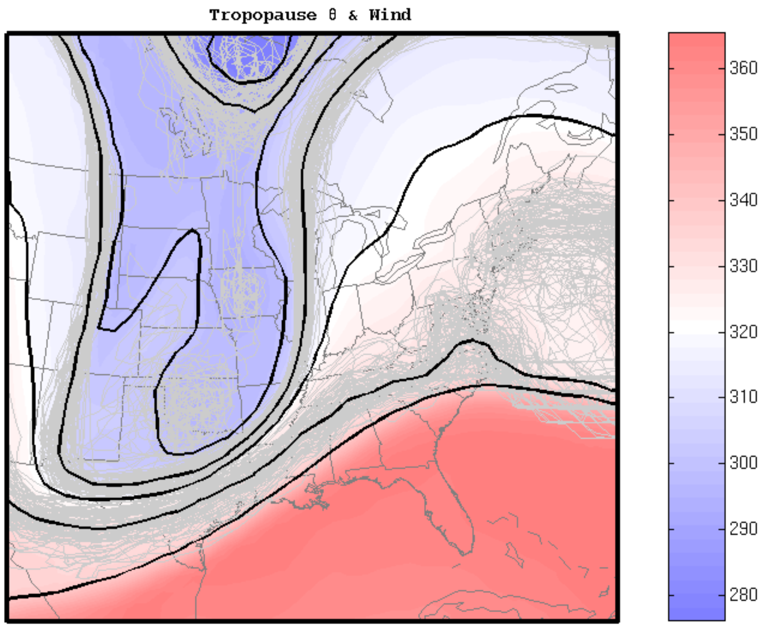
Outline

- The joy of ensemble covariances
  - Using ensemble DA as a tool to better understand the system
- Why mis-positioned features are a problem
- Error models to the rescue
- A two-step approach (position, then amplitude) to ensemble data assimilation
- Many examples showing this is a good thing


time

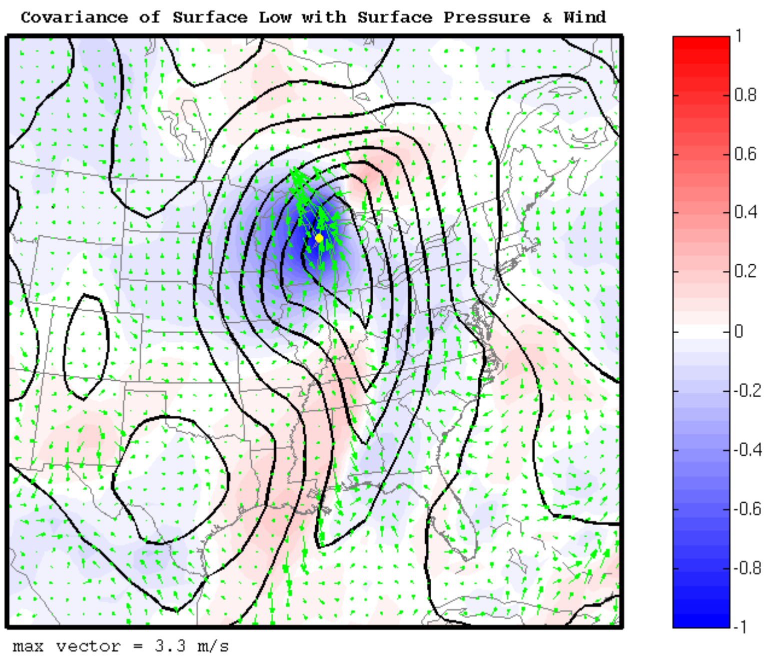


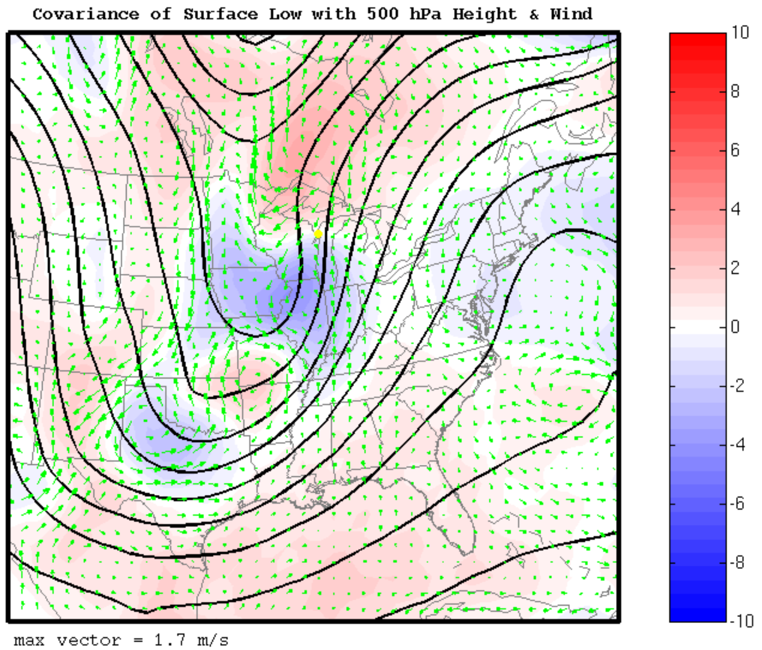




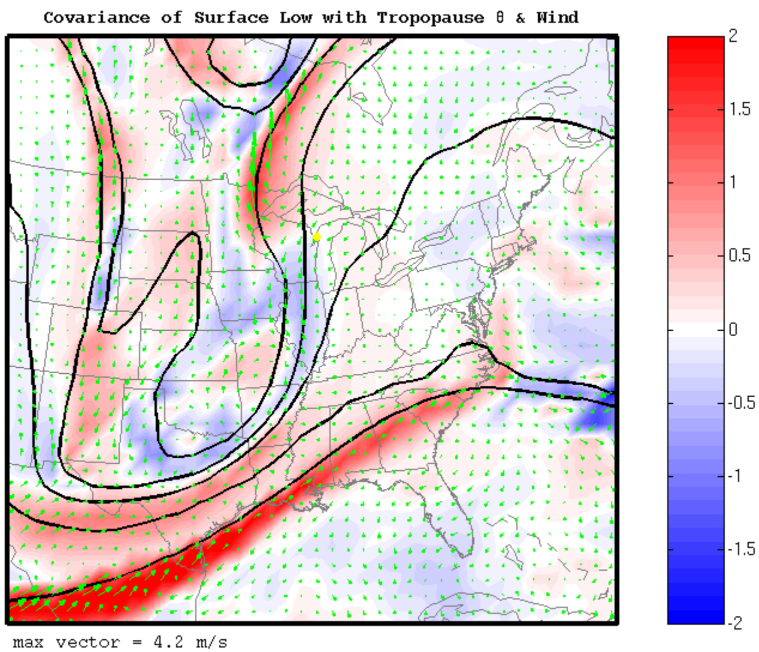
Hakim and Torn

WRF, 100 ensemble members, surface pressure obs





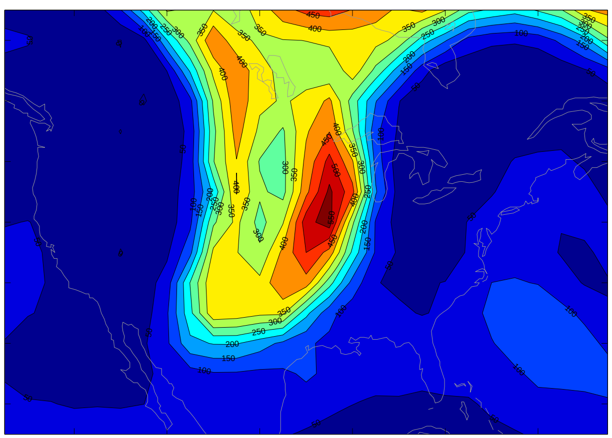
Hakim and Torn



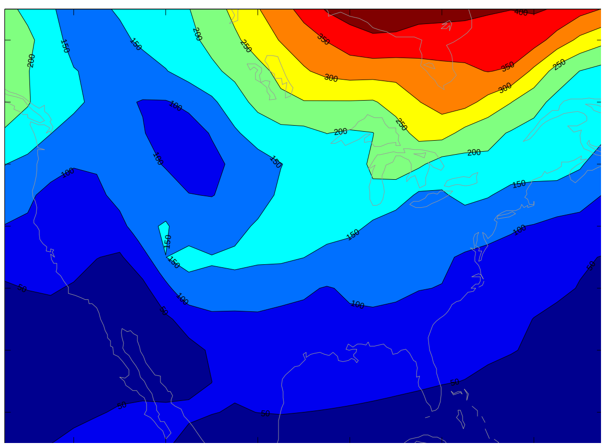
Ensembles make PV inversion fun and easy!

$$\mathbf{PV}_{Ertel} = \mathbf{AX}^{a} \quad \Rightarrow \quad \mathbf{x}^{a} = \mathbf{A}^{-1}\mathbf{pv}_{Ertel}$$

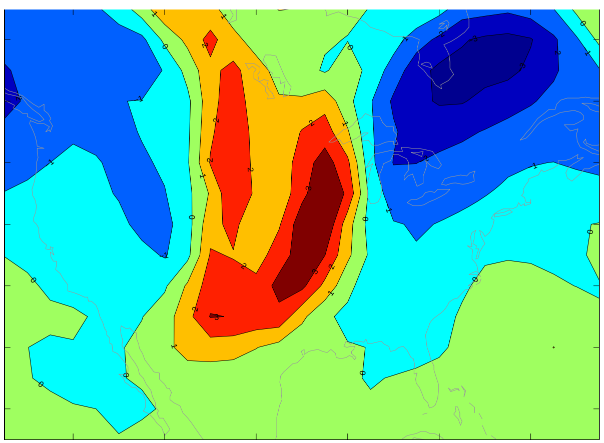
- Approach defined by Hakim and Torn
- Suffers from multicollinearity and overfitting (*k* << *n*)
- Significant improvements are realized by performing the regression in CCA space (Gombos)



Full PV Field (300mb)



Time Mean PV Field



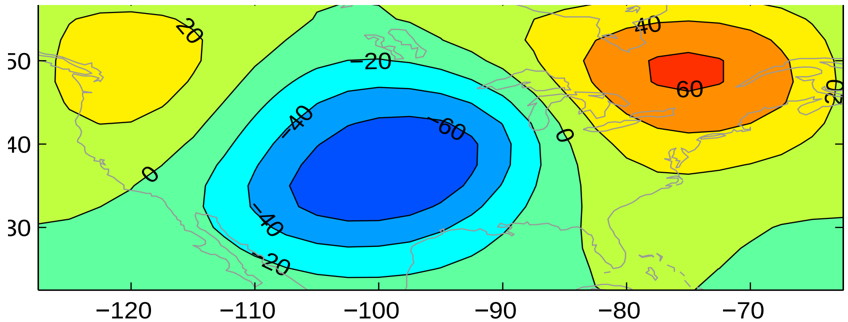
Time Mean Perturbation PV Field

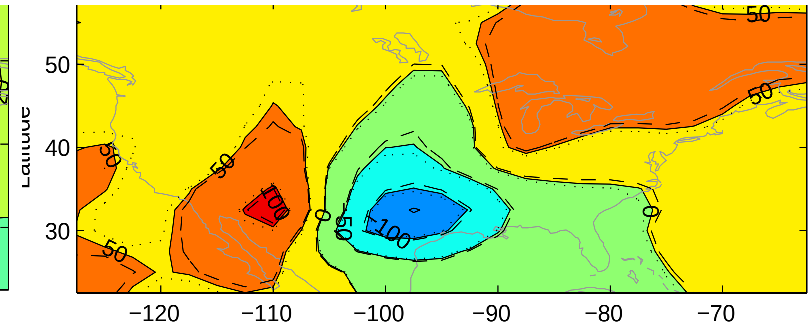
Note, no worries about balance assumptions or boundary conditions

DE Results

ECCR Results

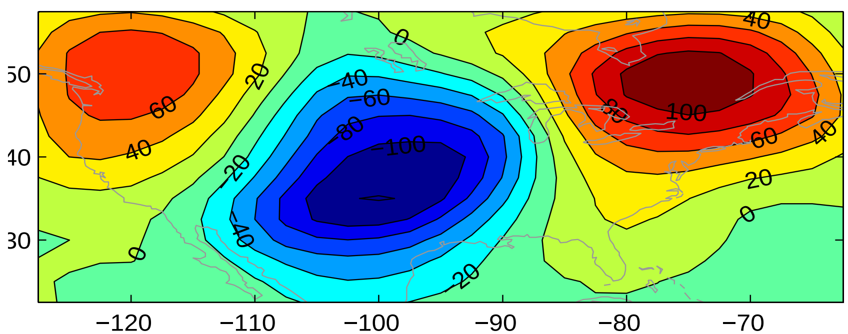
500 mb



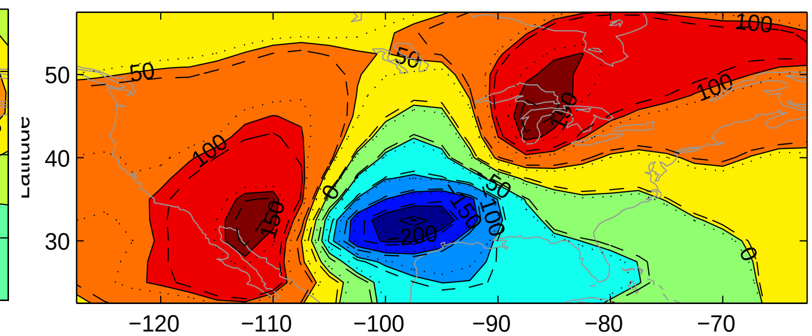


300 mb

c. 300 mb

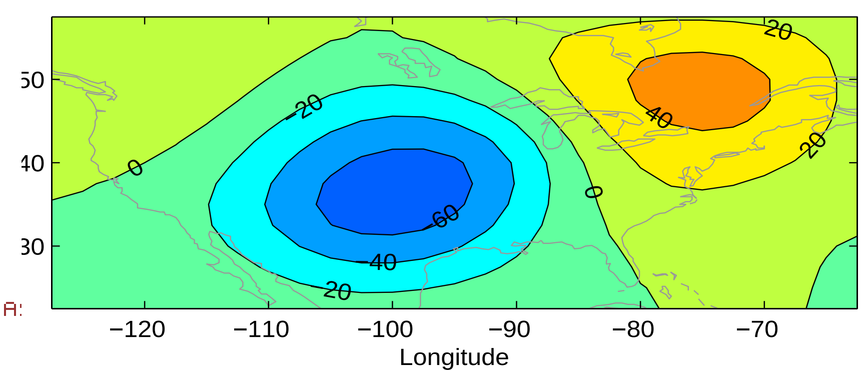


c. 300 mb

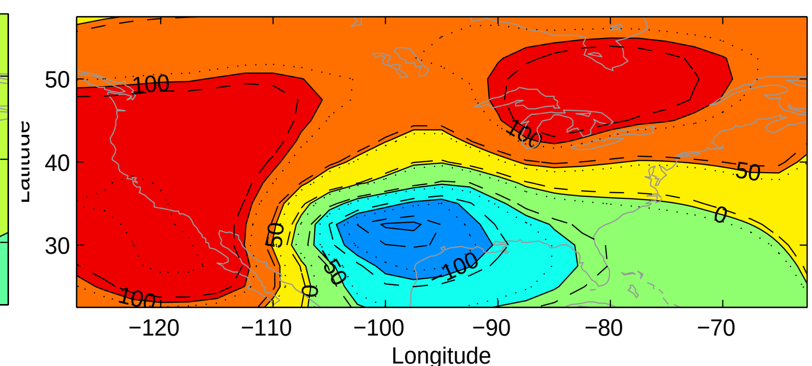


200 mb

e. 200 mb

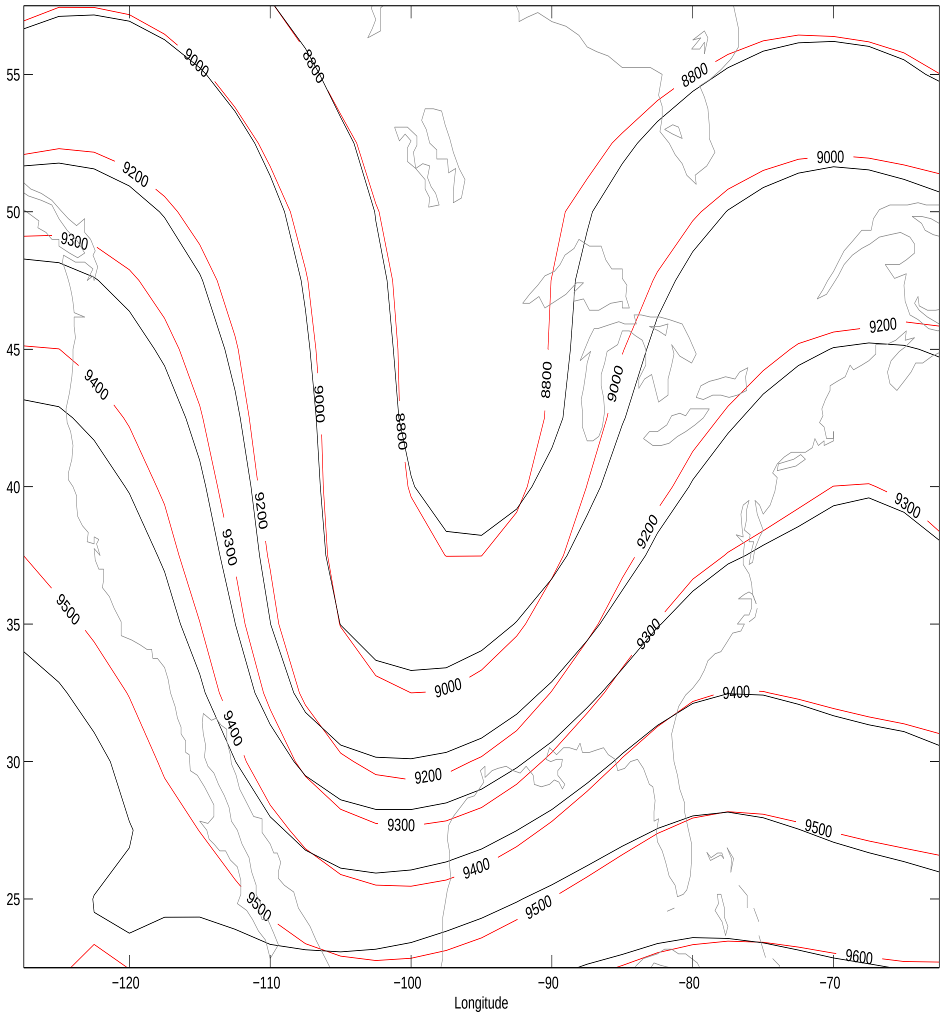


e. 200 mb

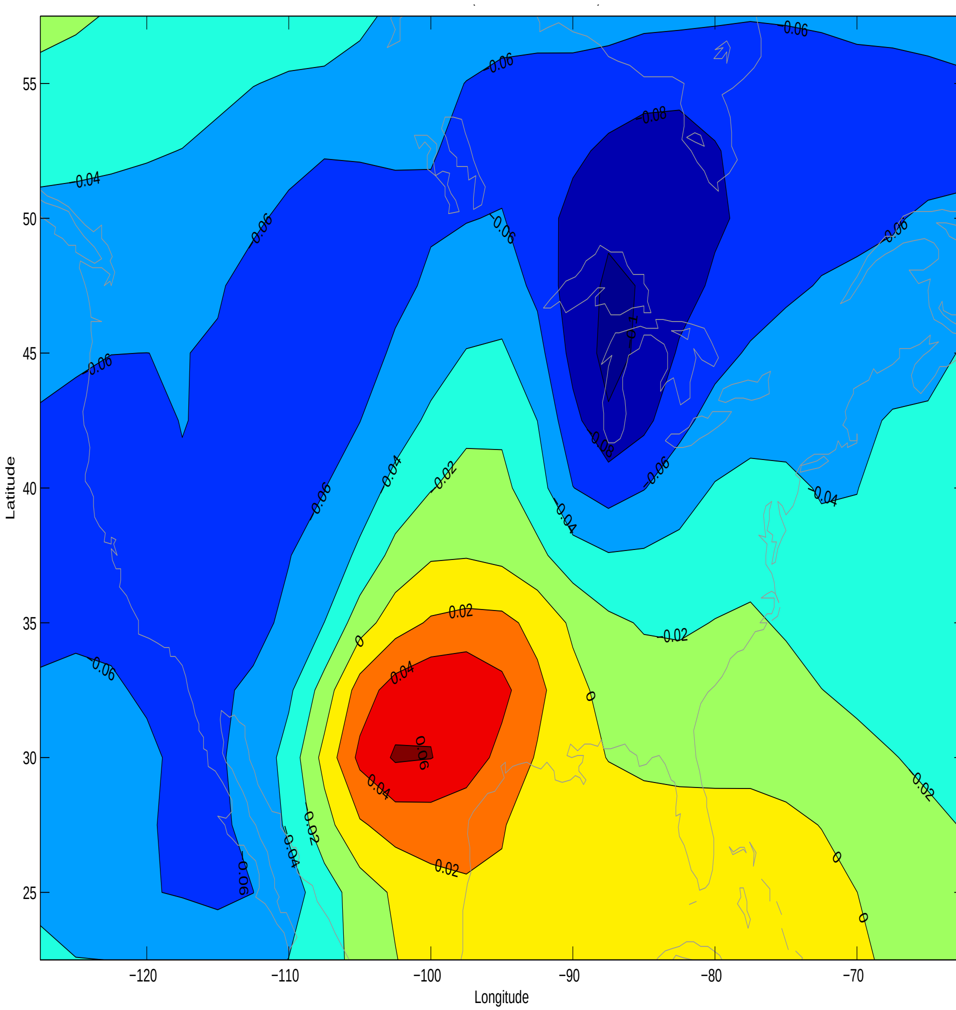


Physical Mechanism: Trough-Ridge Coupling

Two Ensemble Members of the 300 mb Height Field

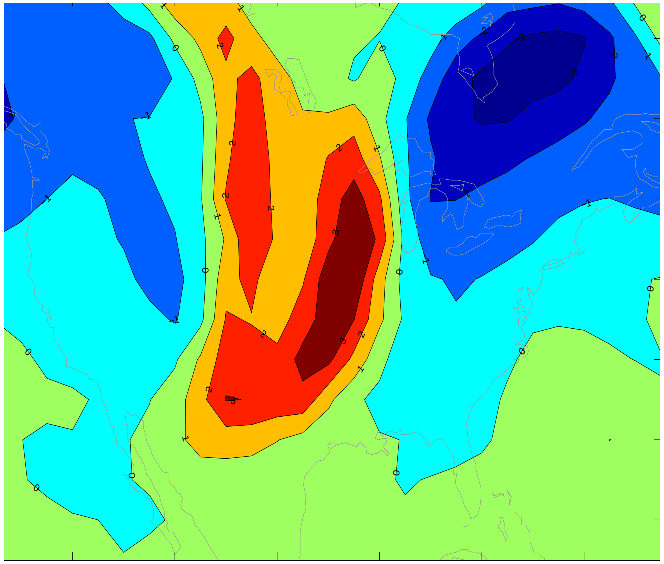


Leading EOF of 300 mb Height Field (77.2% of Variance)

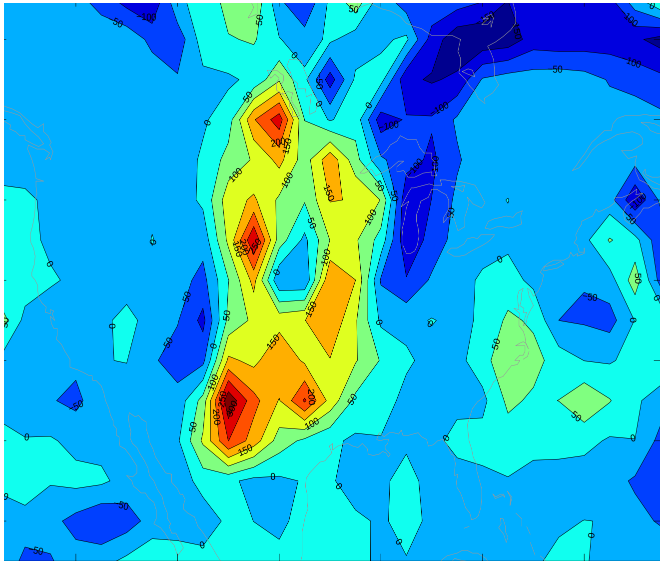


Limitations of ECCR PV inversion

- Null spaces
  - Ensemble size, CCA truncation
- Sampling and model error
- Violation of the linearity assumption

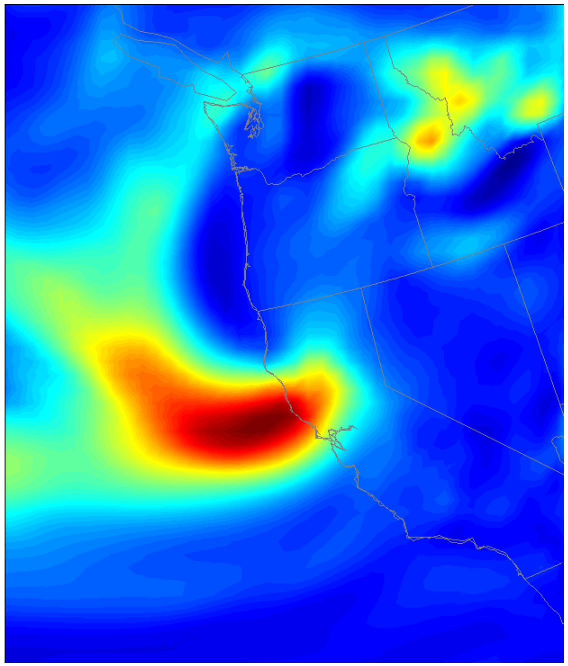


Desired perturbation

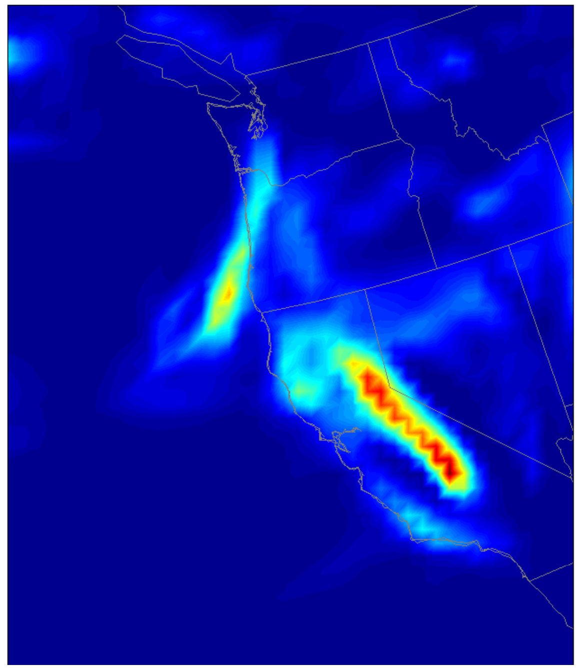


Resolved perturbation

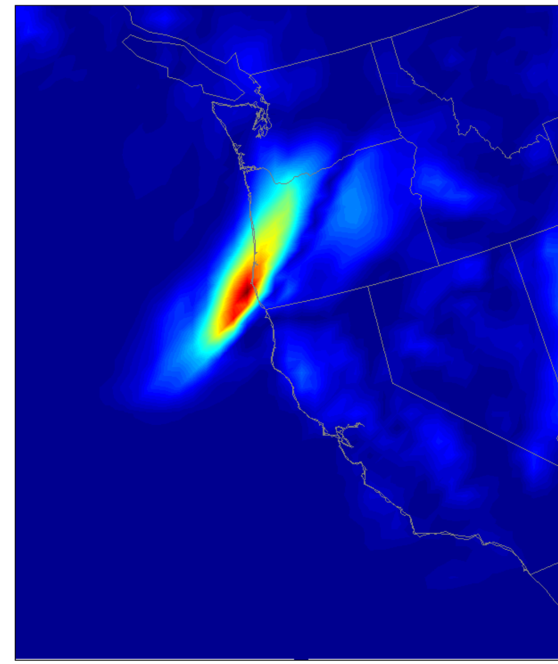
Preliminary PV-Rain Inversion in Time Results



6 Hour Forecast Ensemble Mean 400 mb PV (PVU)



12 Hour Nonconvective Precipitation Forecast (mm)

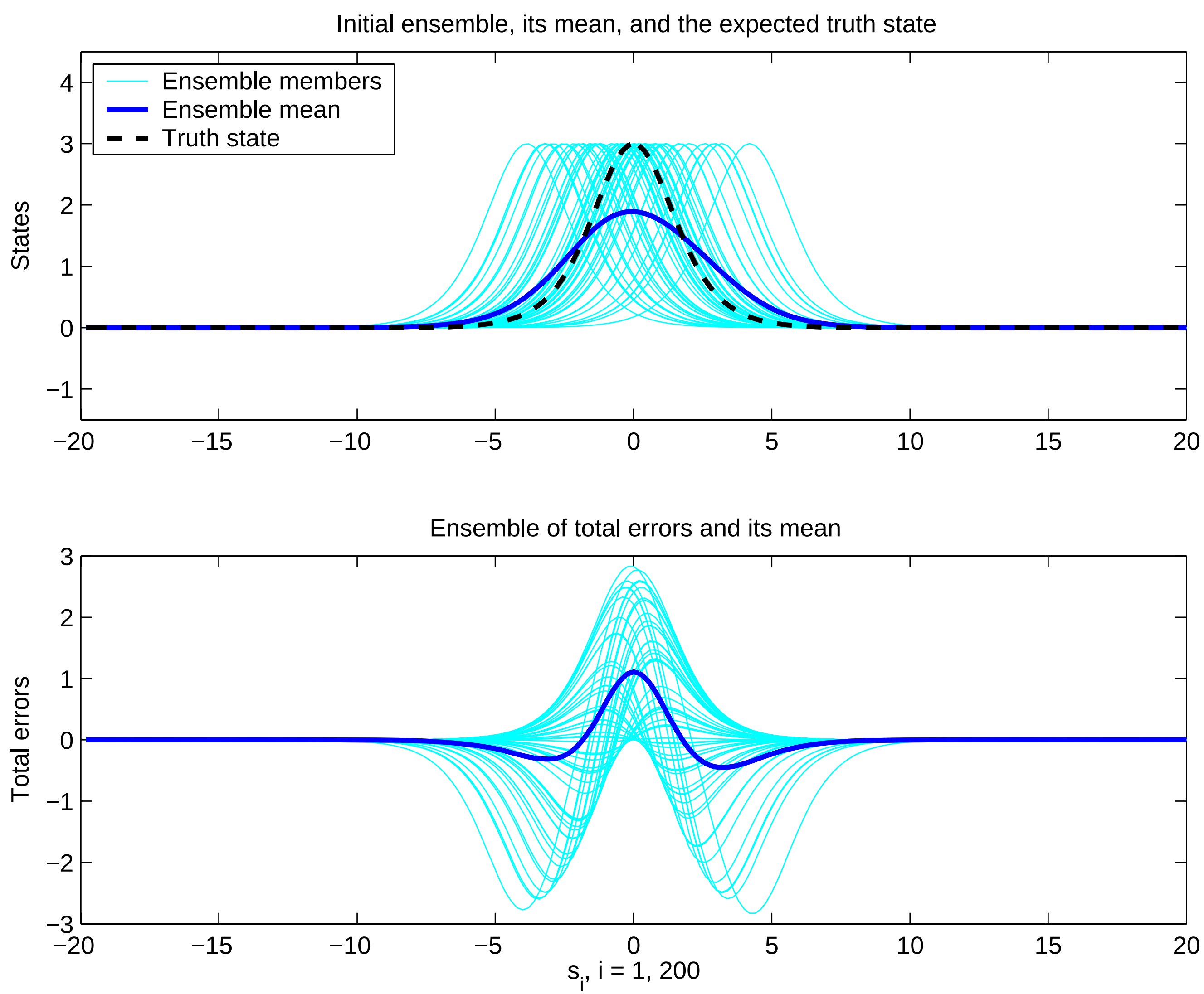


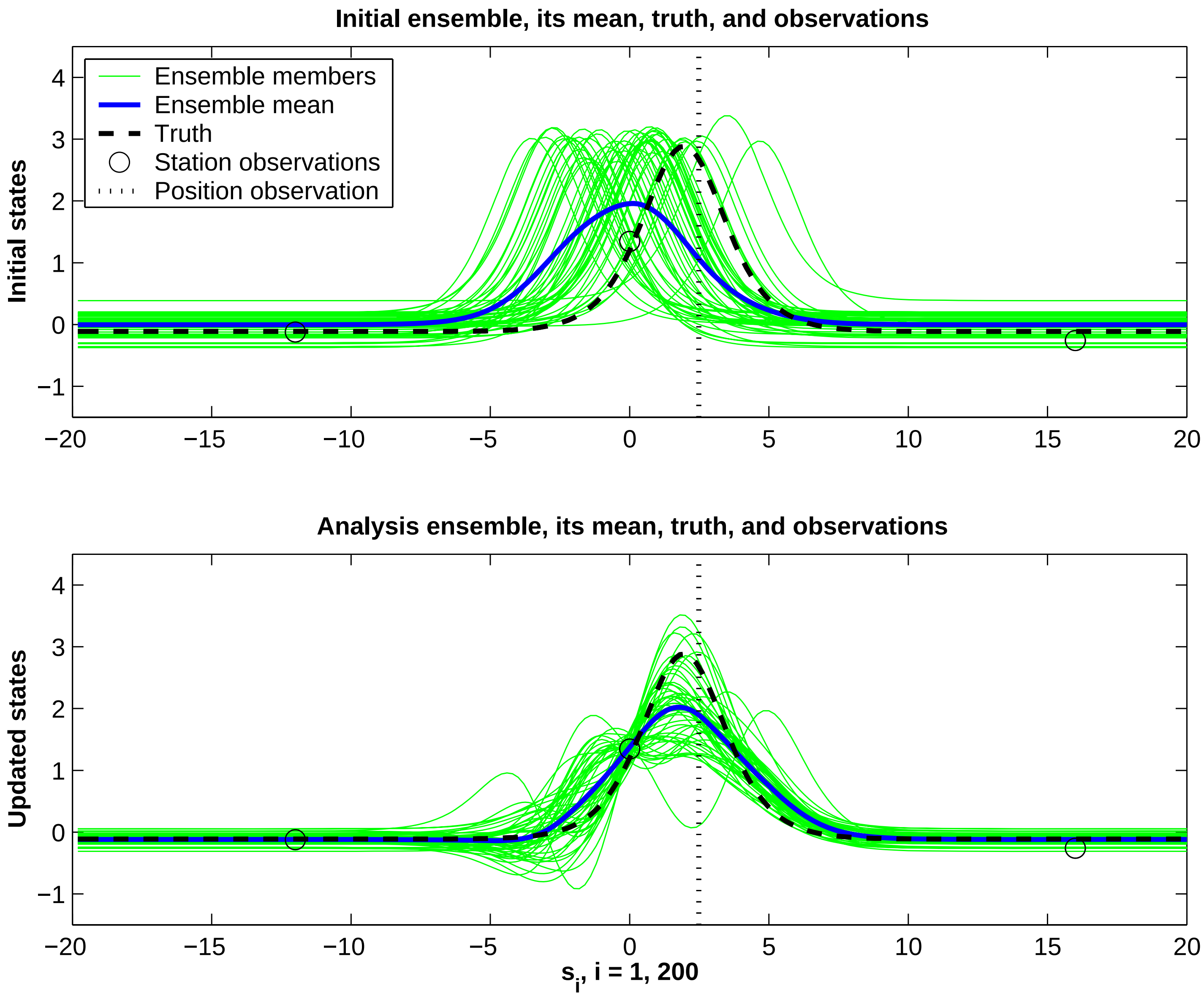
Nonconvective Precipitation (mm) Attributable to the 400 mb PV

Outline

- The joy of ensemble covariances
  - Using ensemble DA as a tool to better understand the system
- Why mis-positioned features are a problem
- Error models to the rescue
- A two-step approach (position, then amplitude) to ensemble data assimilation
- Many examples showing this is a good thing

Gaussian position errors can result in non-Gaussian amplitude errors





Initial ensemble, its mean, truth, and observations
Ensemble members
Ensemble mean
Truth
Station observations
Position observation
Initial states
4
3
2
1
0
−1
−20
−15
−10
−5
0
5
10
15
20
Analysis ensemble, its mean, truth, and observations
Updated states
s_i, i = 1, 200

Outline

- The joy of ensemble covariances
  - Using ensemble DA as a tool to better understand the system
- Why mis-positioned features are a problem
- Error models to the rescue
- A two-step approach (position, then amplitude) to ensemble data assimilation
- Many examples showing this is a good thing

Error Models

- Total errors defined as

$$\boldsymbol{\varepsilon}^f \equiv \mathbf{x}^t - \mathbf{x}^f$$

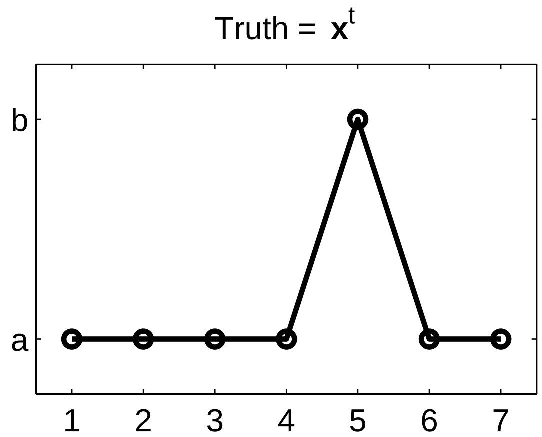
- Many possible error models

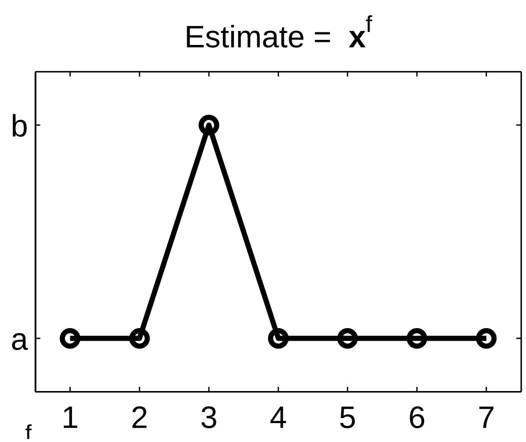
$$\mathbf{x}^t = \mathbf{x}^f(s_i) + \boxed{\boldsymbol{\varepsilon}_A(s_i)}$$ **Additive**

$$\mathbf{x}^t = (1 - \boxed{\boldsymbol{\varepsilon}_M(s_i)}) \circ \mathbf{x}^f(s_i)$$ **Multiplicative**

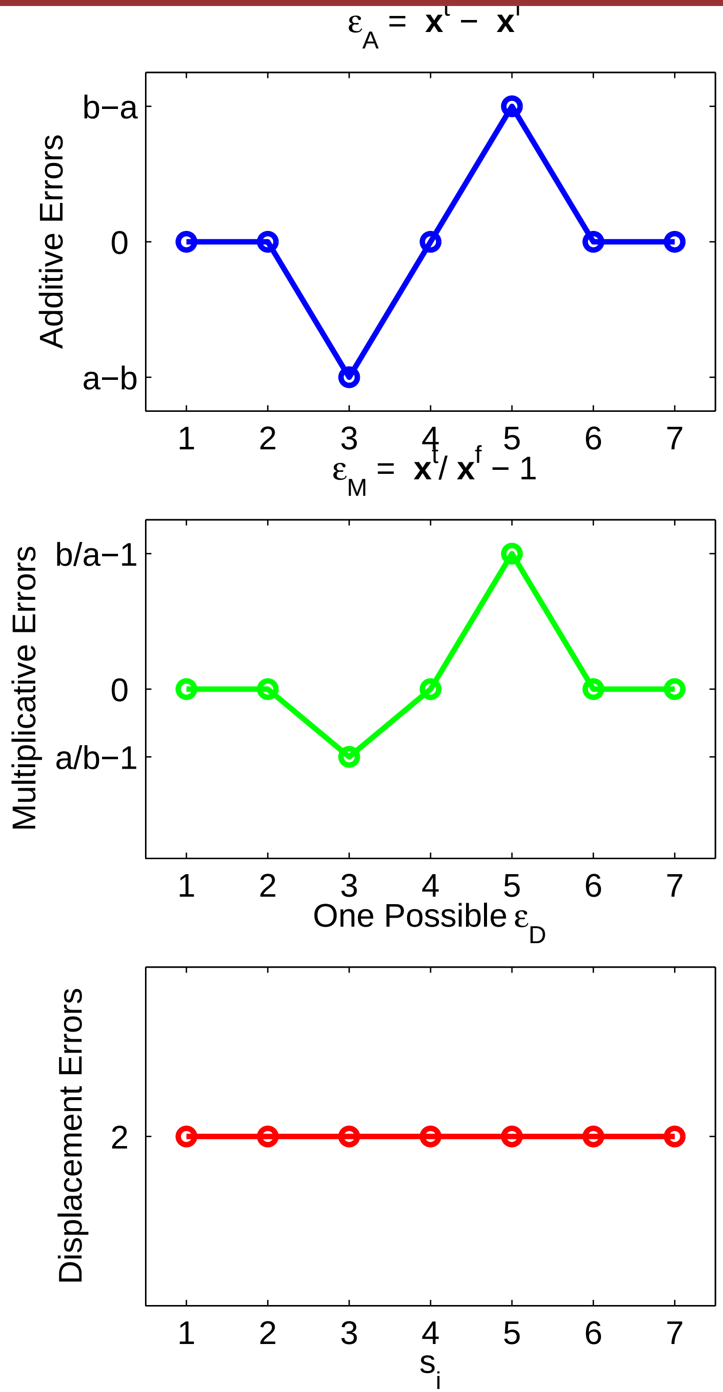
$$\mathbf{x}^t = \mathbf{x}^f(s_i + \boxed{\boldsymbol{\varepsilon}_D(s_i)})$$ **Displacement**

$$\mathbf{x}^t = (1 - \boxed{\boldsymbol{\varepsilon}_M(s_i)}) \circ \mathbf{x}^f(s_i + \boxed{\boldsymbol{\varepsilon}_D(s_i)}) + \boxed{\boldsymbol{\varepsilon}_A(s_i)}$$ **Mixed**


Truth = $\mathbf{x}^t$
b
a
1 2 3 4 5 6 7


Estimate = $\mathbf{x}^f$
b
a
1 2 3 4 5 6 7

t f

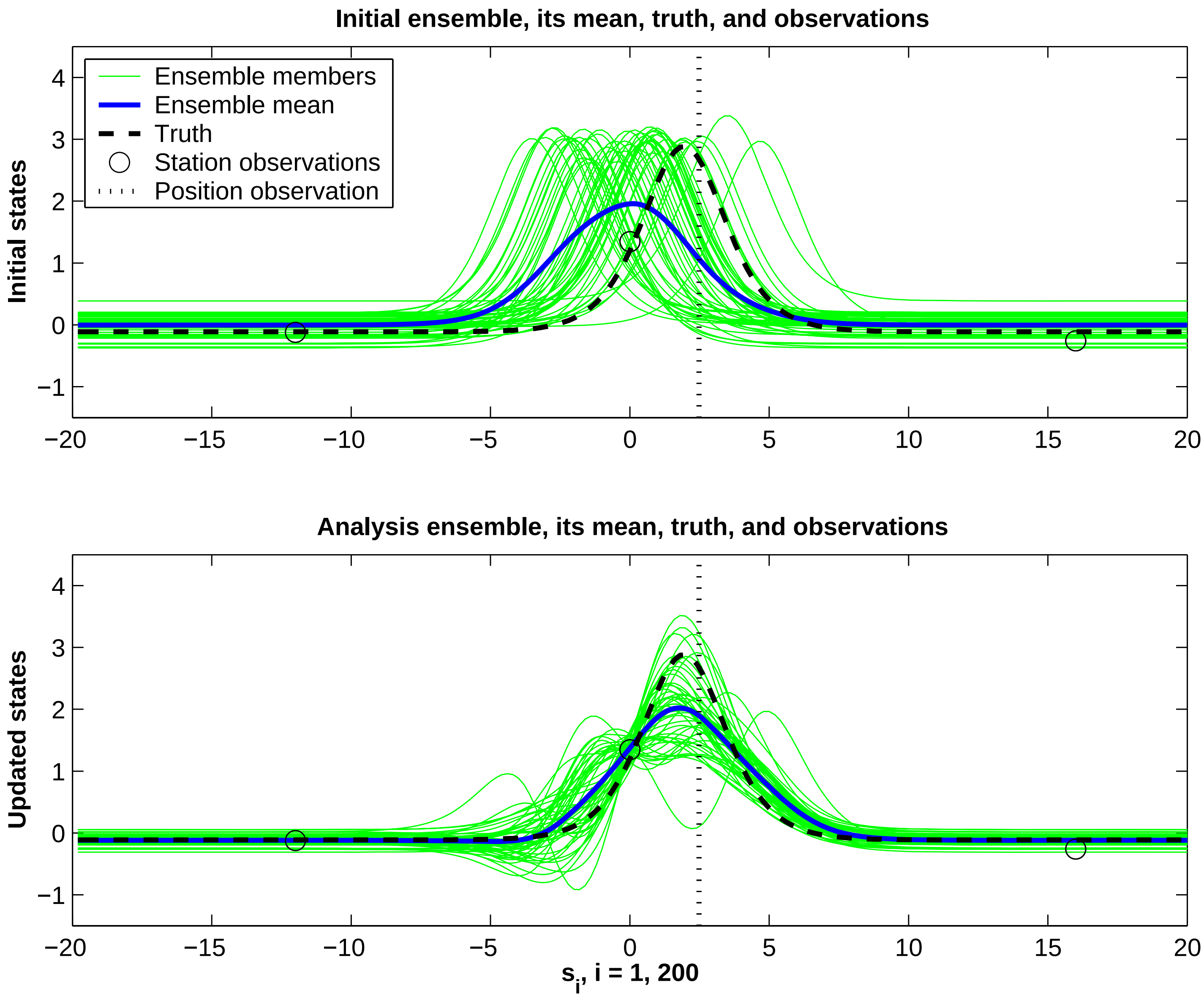

$\varepsilon_A = \mathbf{x}^t - \mathbf{x}^f$
Additive Errors
b−a
0
a−b
1 2 3 4 5 6 7
$\varepsilon_M = \mathbf{x}^t / \mathbf{x}^f - 1$
Multiplicative Errors
b/a−1
0
a/b−1
1 2 3 4 5 6 7
One Possible ε_D
Displacement Errors
2
1 2 3 4 5 6 7
s_i

Example: Korteweg de-Vries

- Mixed displacement/additive error model

$$\mathbf{x}^t = \mathbf{x}^f(s_i + \varepsilon_D) + \boldsymbol{\varepsilon}_A(s_i)$$

- For coherent feature, use a KdV soliton
  - System dynamics: $\mathbf{u}_t + \mathbf{u}\mathbf{u}_s + \mathbf{u}_{sss} = 0$
  - Soliton solution: $\mathbf{u}_{sol}(s,t) = 3A\mathrm{sech}^2\left(\frac{A^{1/2}}{2}(s - At)\right)$
  - "Truth" is a soliton


Initial ensemble, its mean, truth, and observations
Ensemble members
Ensemble mean
Truth
Station observations
Position observation
Initial states
4
3
2
1
0
−1
−20
−15
−10
−5
0
5
10
15
20
Analysis ensemble, its mean, truth, and observations
Updated states
s_i, i = 1, 200

Outline

- The joy of ensemble covariances
  - Using ensemble DA as a tool to better understand the system
- Why mis-positioned features are a problem
- Error models to the rescue
- A two-step approach (position, then amplitude) to ensemble data assimilation
- Many examples showing this is a good thing

A two-step approach

- Mixed displacement/additive error model

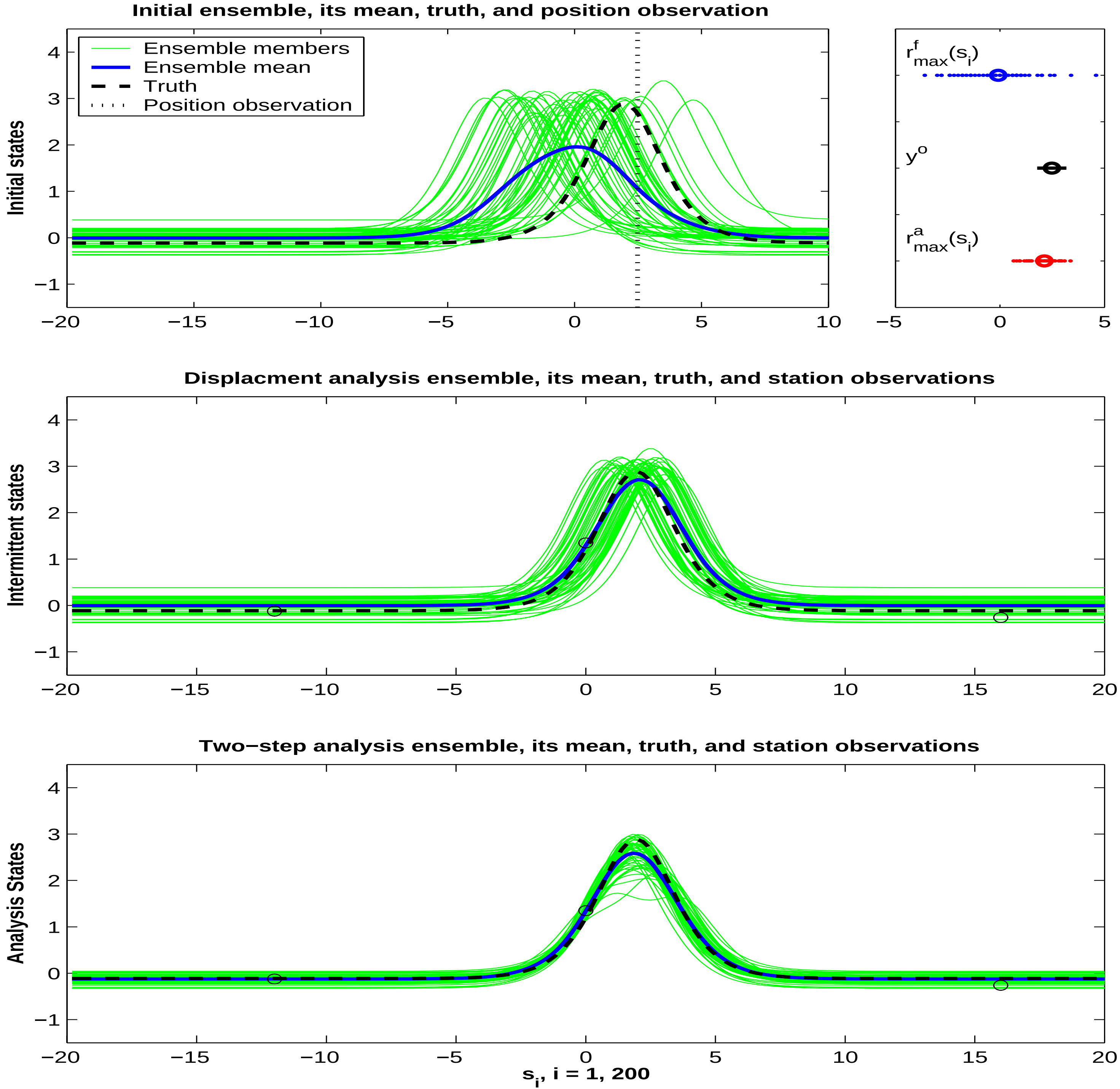
$$\mathbf{x}^t = \mathbf{x}^f(s_i + \varepsilon_D) + \boldsymbol{\varepsilon}_A(s_i)$$

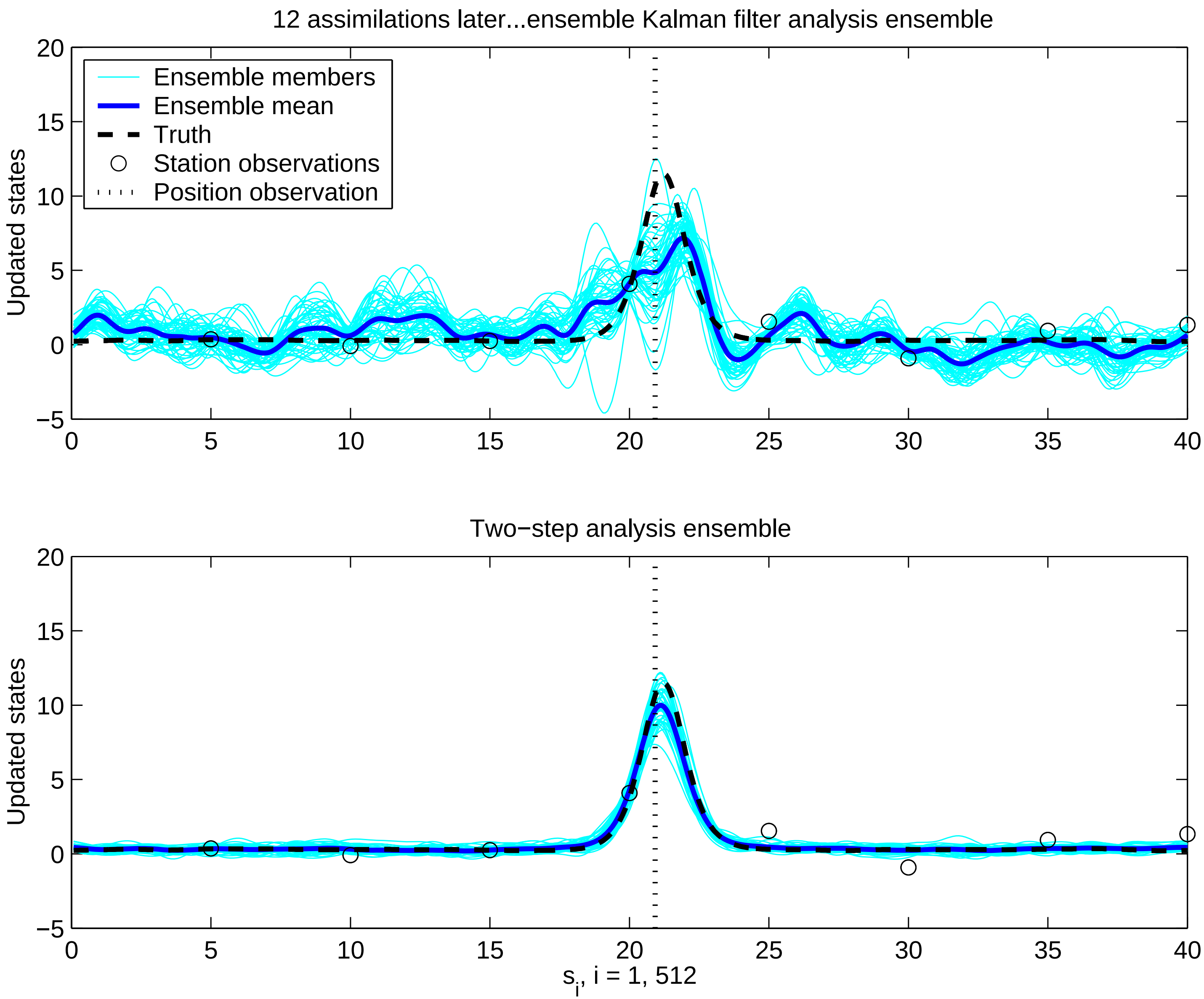
- First assimilate for displacement

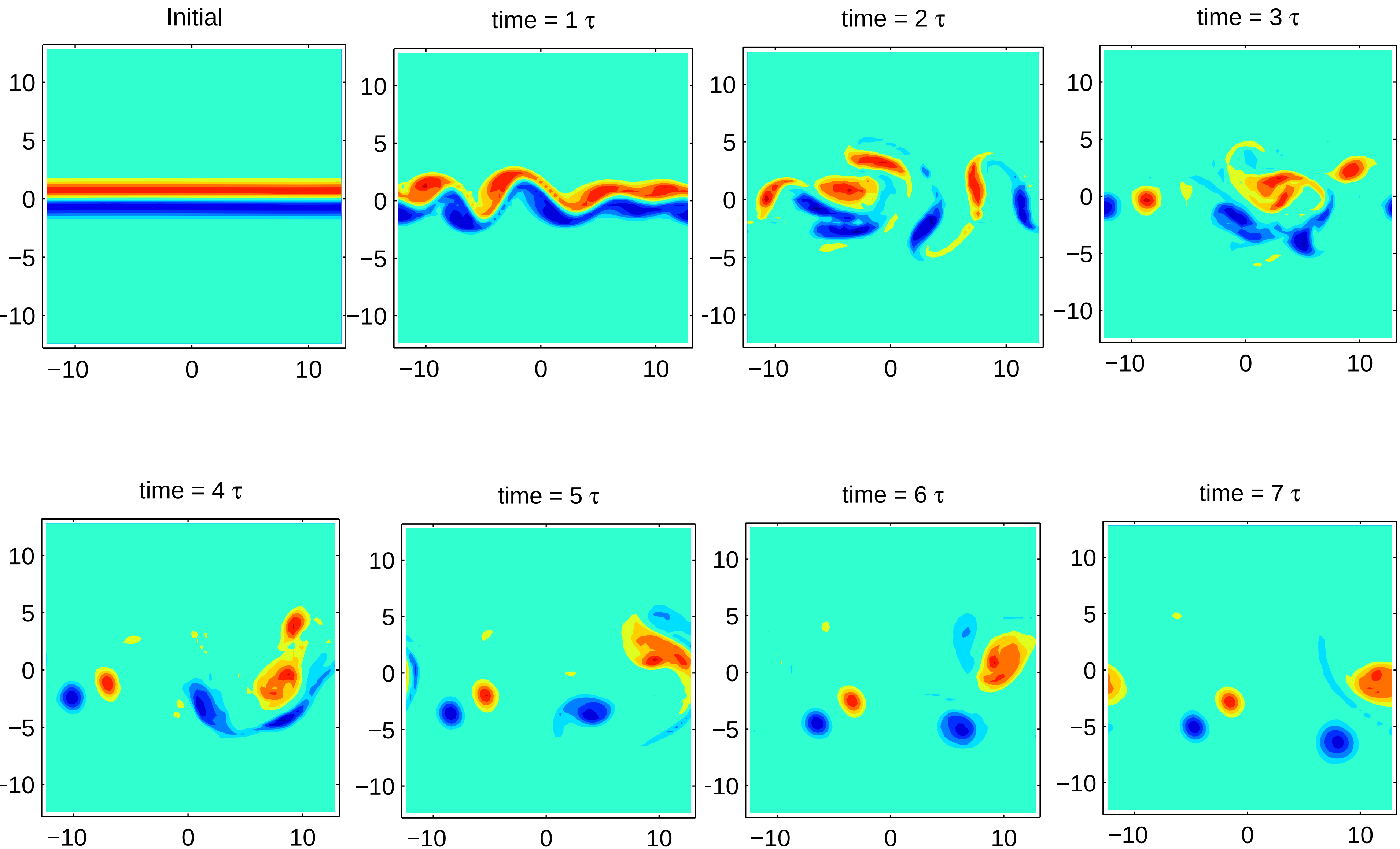
$$\tilde{\mathbf{x}}^f = \mathbf{x}^f(s_i + \boxed{\varepsilon_D})$$

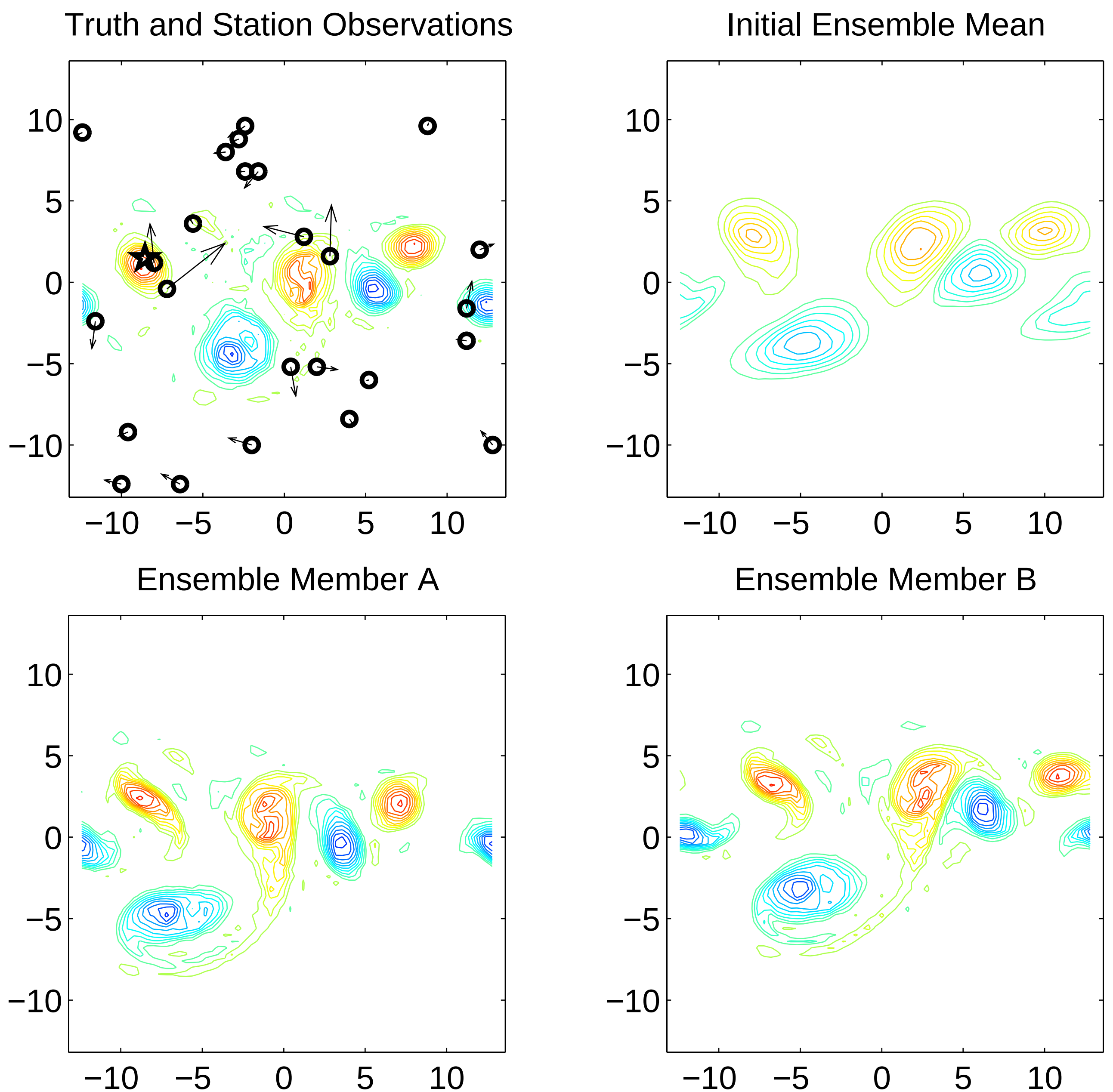
- Then assimilate for amplitude

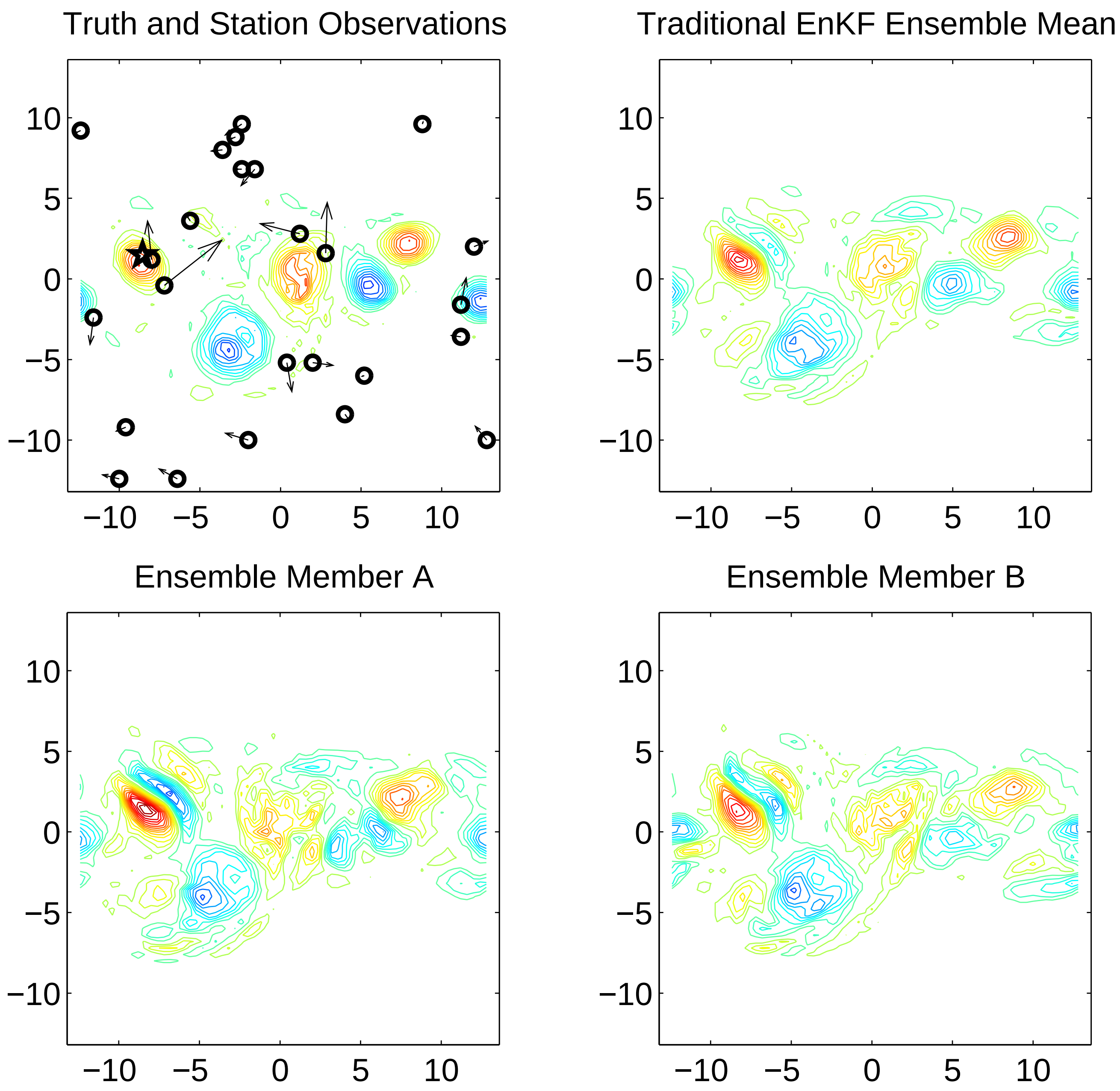
$$\mathbf{x}^t = \tilde{\mathbf{x}}^f(s_i) + \boxed{\boldsymbol{\varepsilon}_A(s_i)}$$


Initial ensemble, its mean, truth, and position observation
Ensemble members
Ensemble mean
Truth
Position observation
Initial states
$r^f_{max}(s_i)$
y^o
$r^a_{max}(s_i)$
Displacment analysis ensemble, its mean, truth, and station observations
Intermittent states
Two−step analysis ensemble, its mean, truth, and station observations
Analysis States
s_i, i = 1, 200

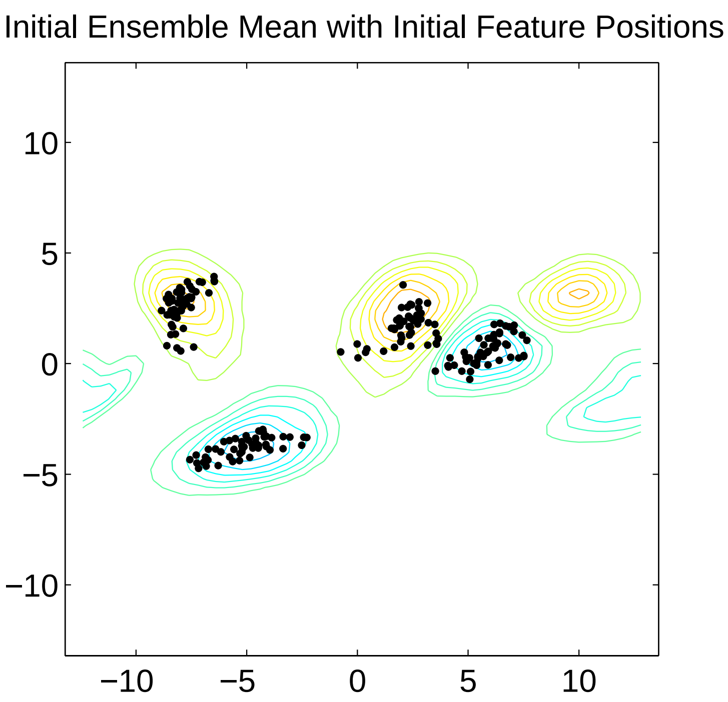

12 assimilations later...ensemble Kalman filter analysis ensemble
Ensemble members
Ensemble mean
Truth
Station observations
Position observation
Updated states
Two−step analysis ensemble
Updated states
s_i, i = 1, 512
20
15
10
5
0
−5
0
5
10
15
20
25
30
35
40

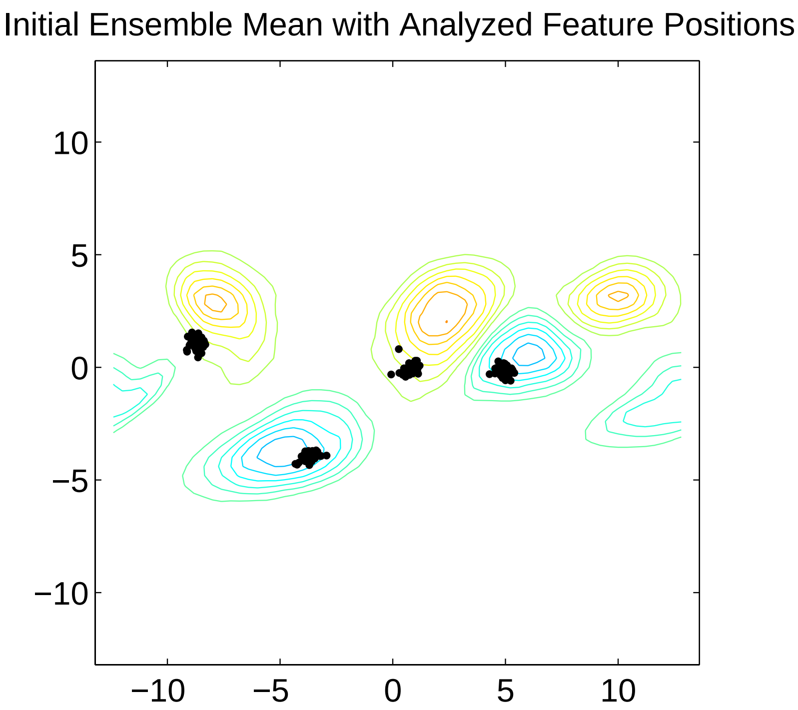

Initial
time = 1 τ
time = 2 τ
time = 3 τ
time = 4 τ
time = 5 τ
time = 6 τ
time = 7 τ
10
5
0
−5
−10
−10
0
10

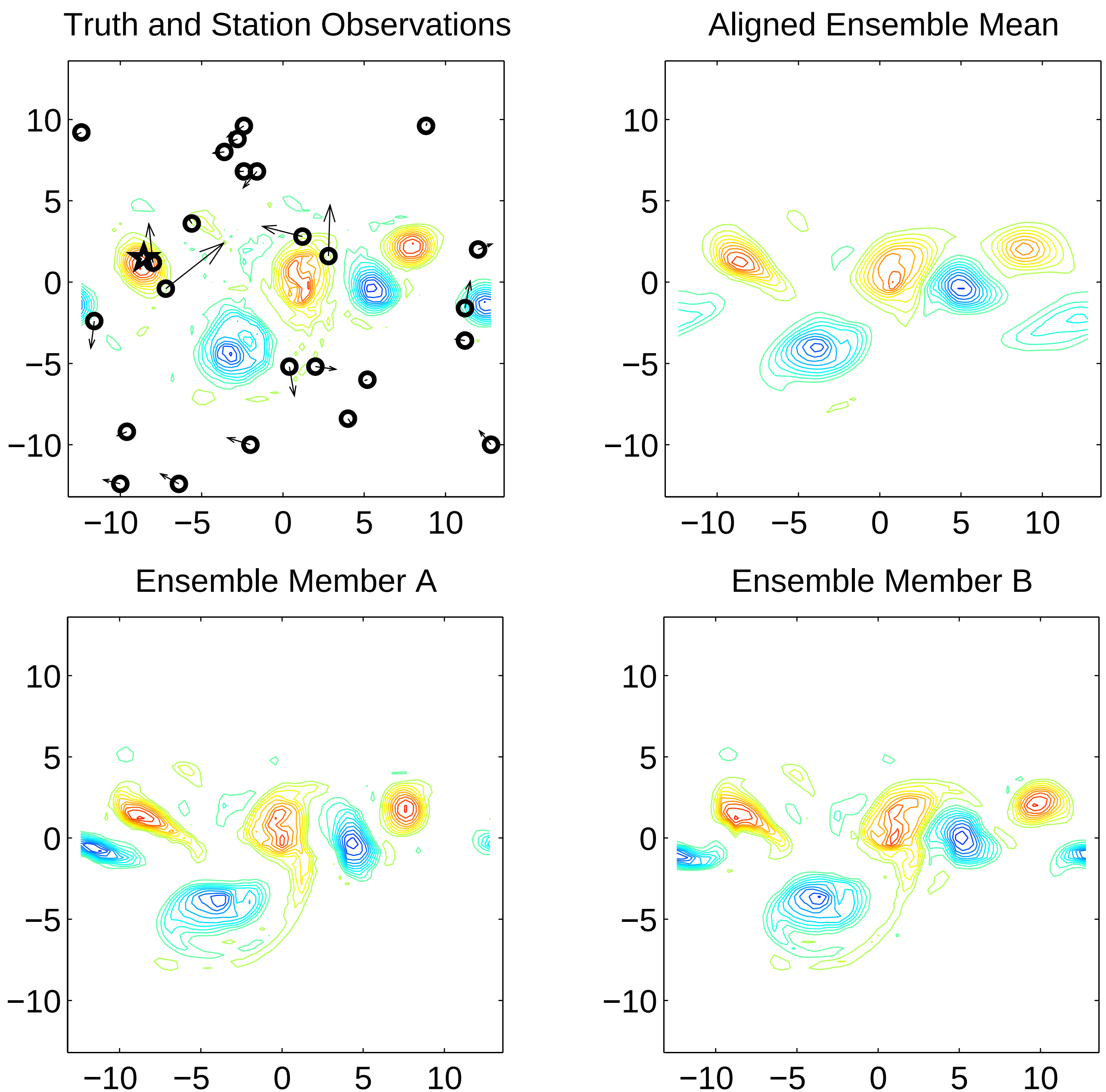

Truth and Station Observations
Initial Ensemble Mean
Ensemble Member A
Ensemble Member B
10
5
0
−5
−10
−10 −5 0 5 10



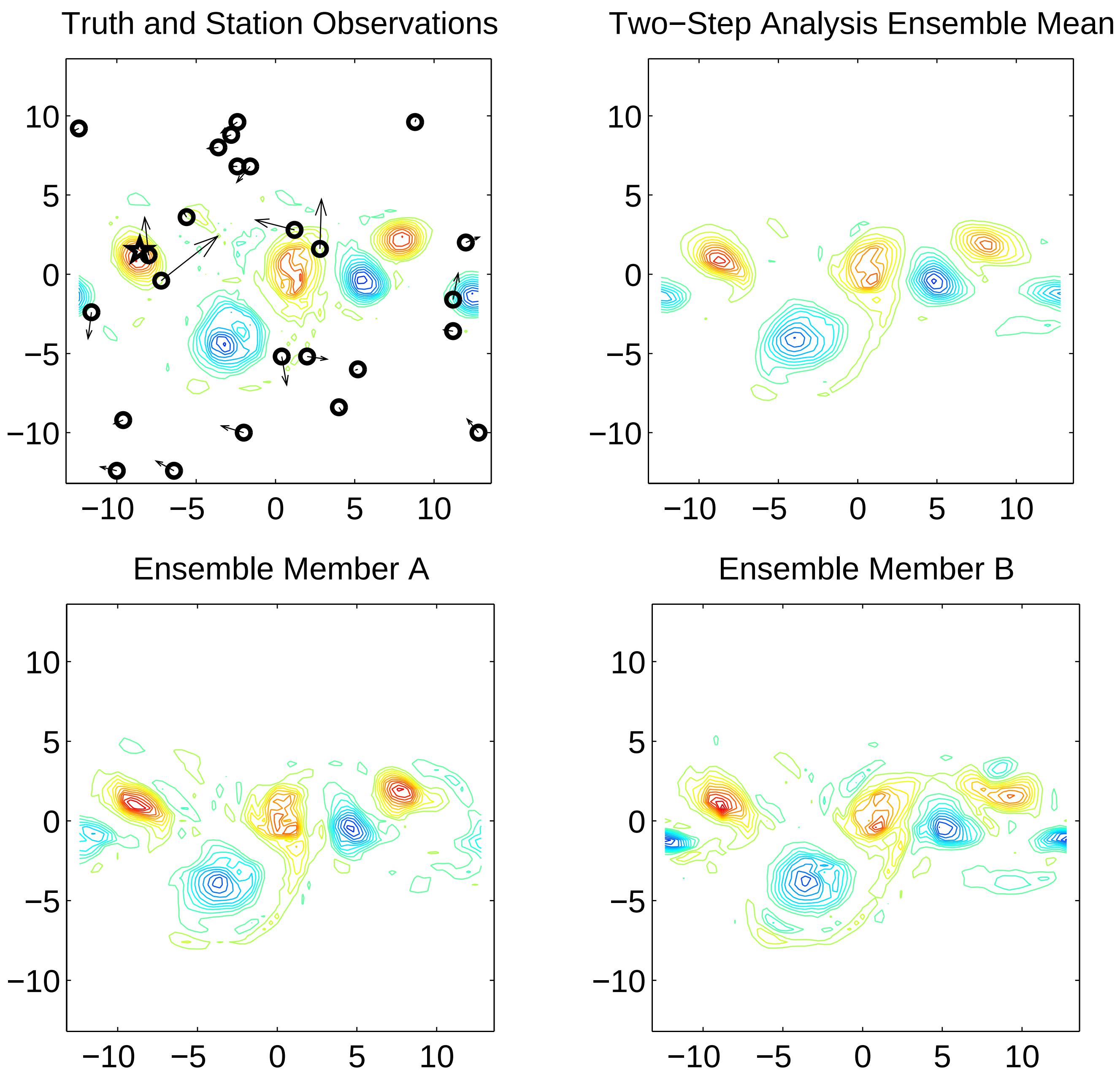
Truth and Station Observations
Traditional EnKF Ensemble Mean
Ensemble Member A
Ensemble Member B
10
5
0
−5
−10


Initial Ensemble Mean with Initial Feature Positions
10
5
0
−5
−10
−10
−5
0
5
10

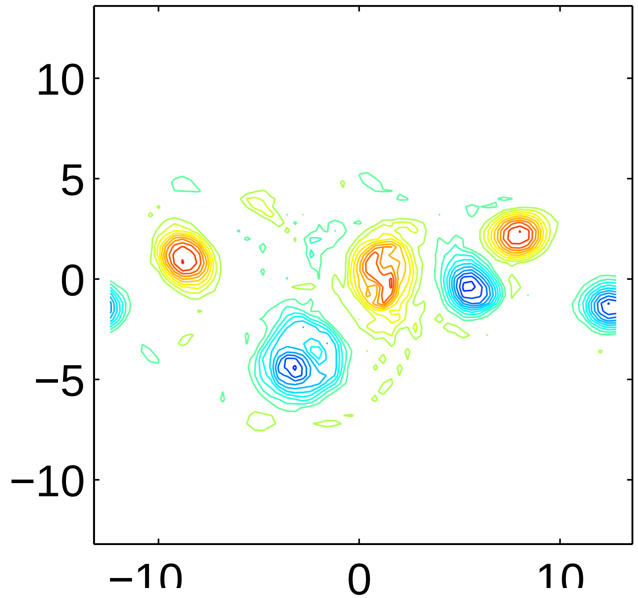

Initial Ensemble Mean with Analyzed Feature Positions
10
5
0
−5
−10
−10
−5
0
5
10



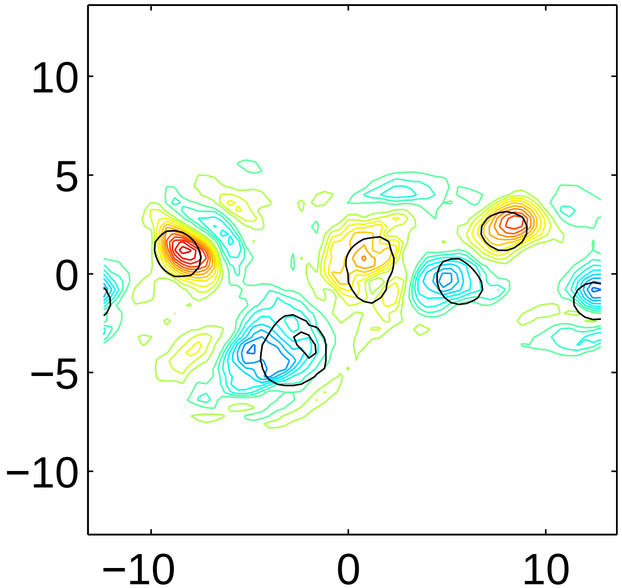
MASSACHUSETTS INSTITUTE OF TECHNOLOGY


Truth and Station Observations
Two−Step Analysis Ensemble Mean
Ensemble Member A
Ensemble Member B
10
5
0
−5
−10
−10 −5 0 5 10

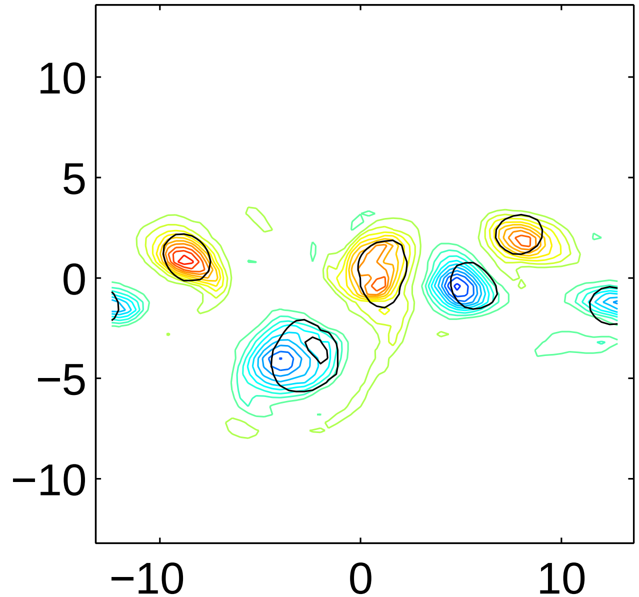
Truth

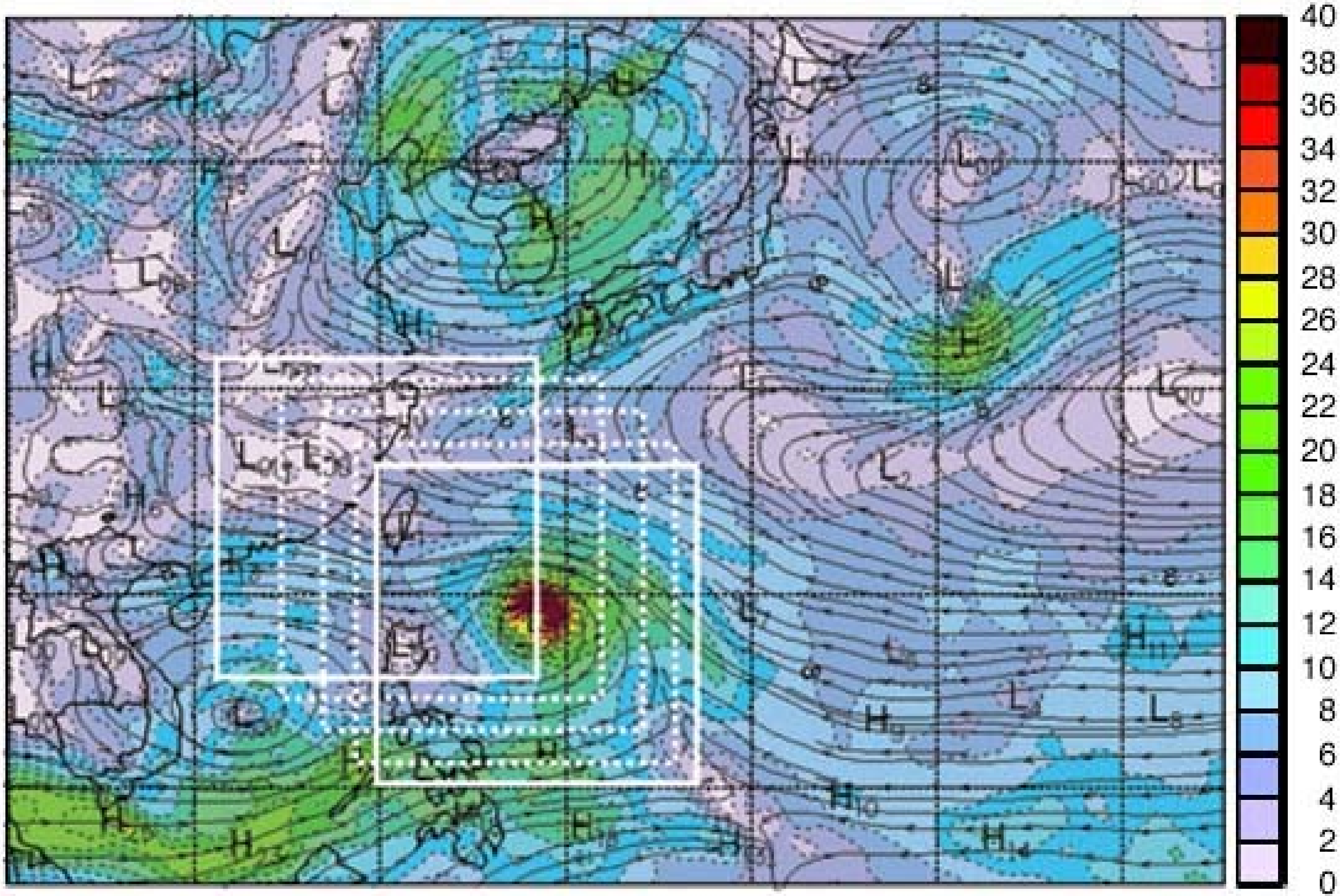


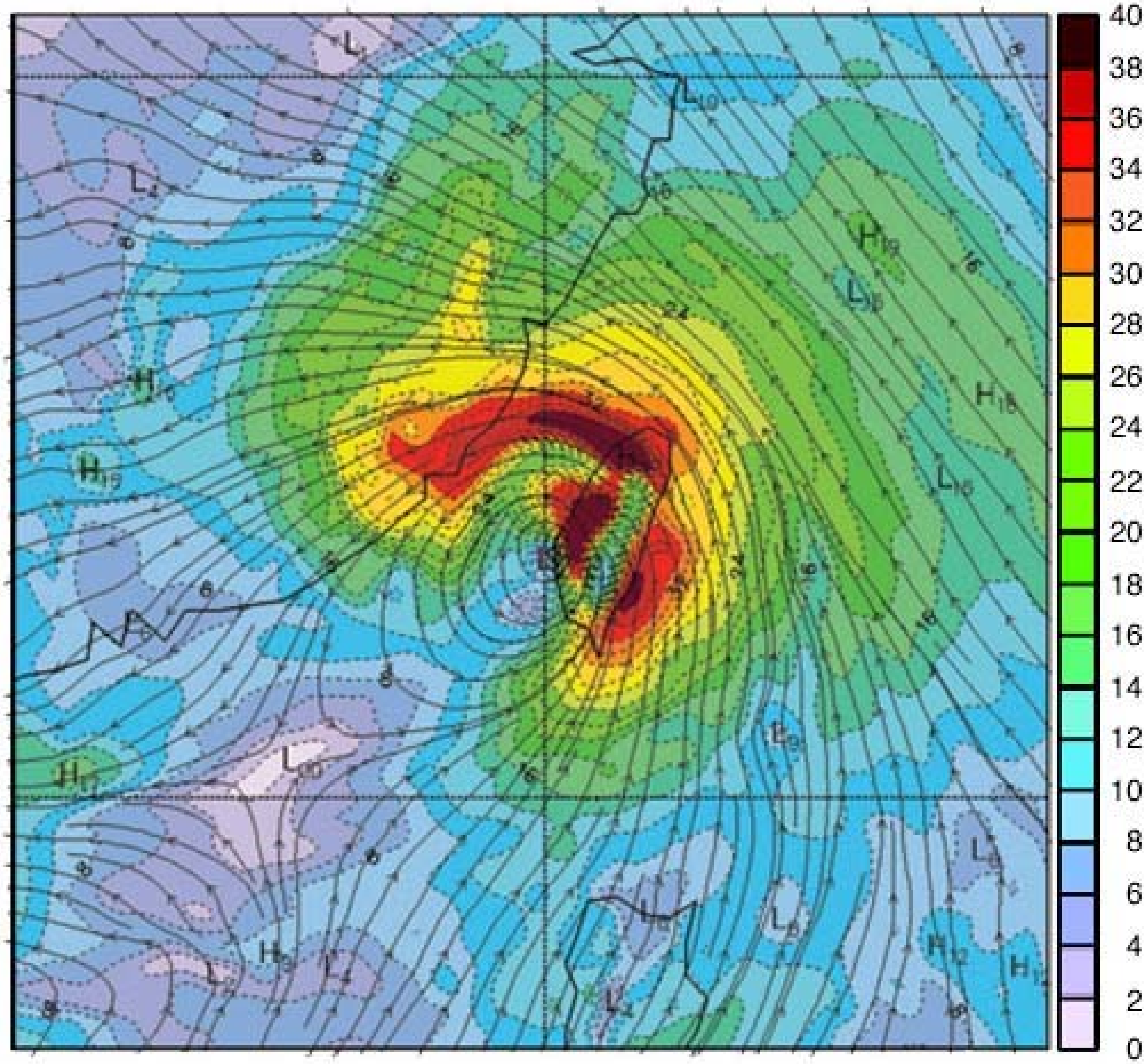
Traditional EnKF Ensemble Mean

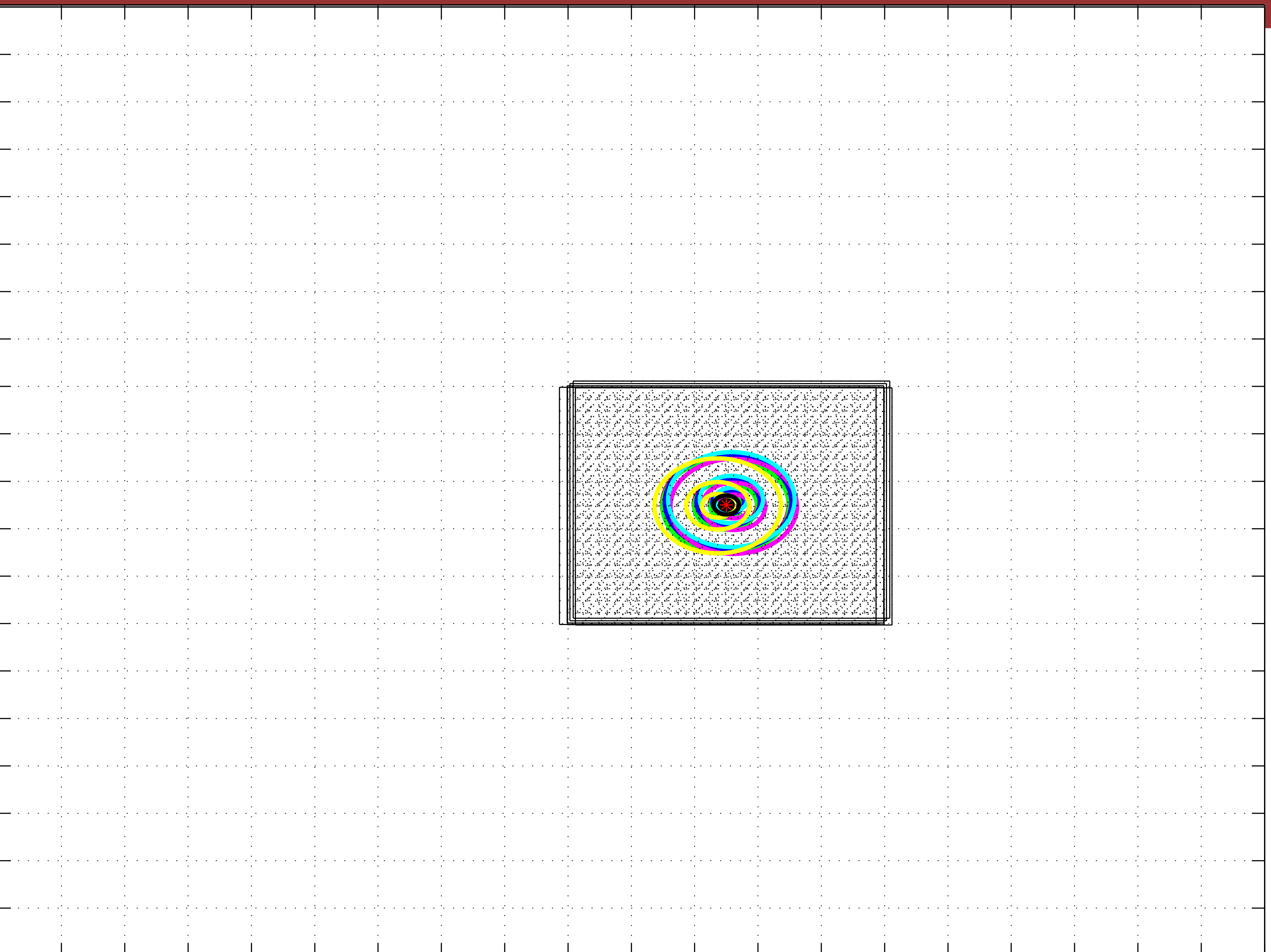


Two−Step Analysis Ensemble Mean



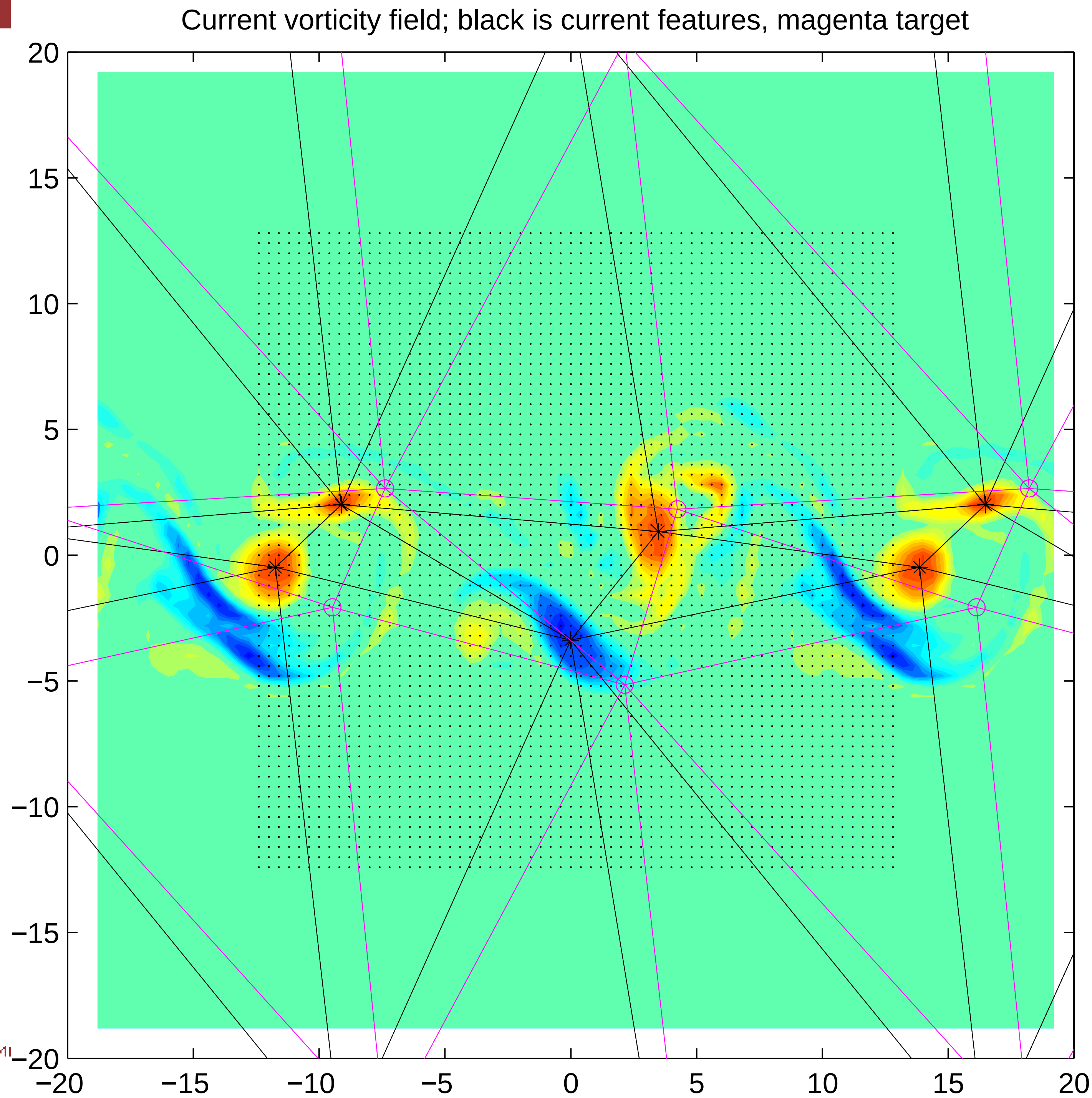

40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0


40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0

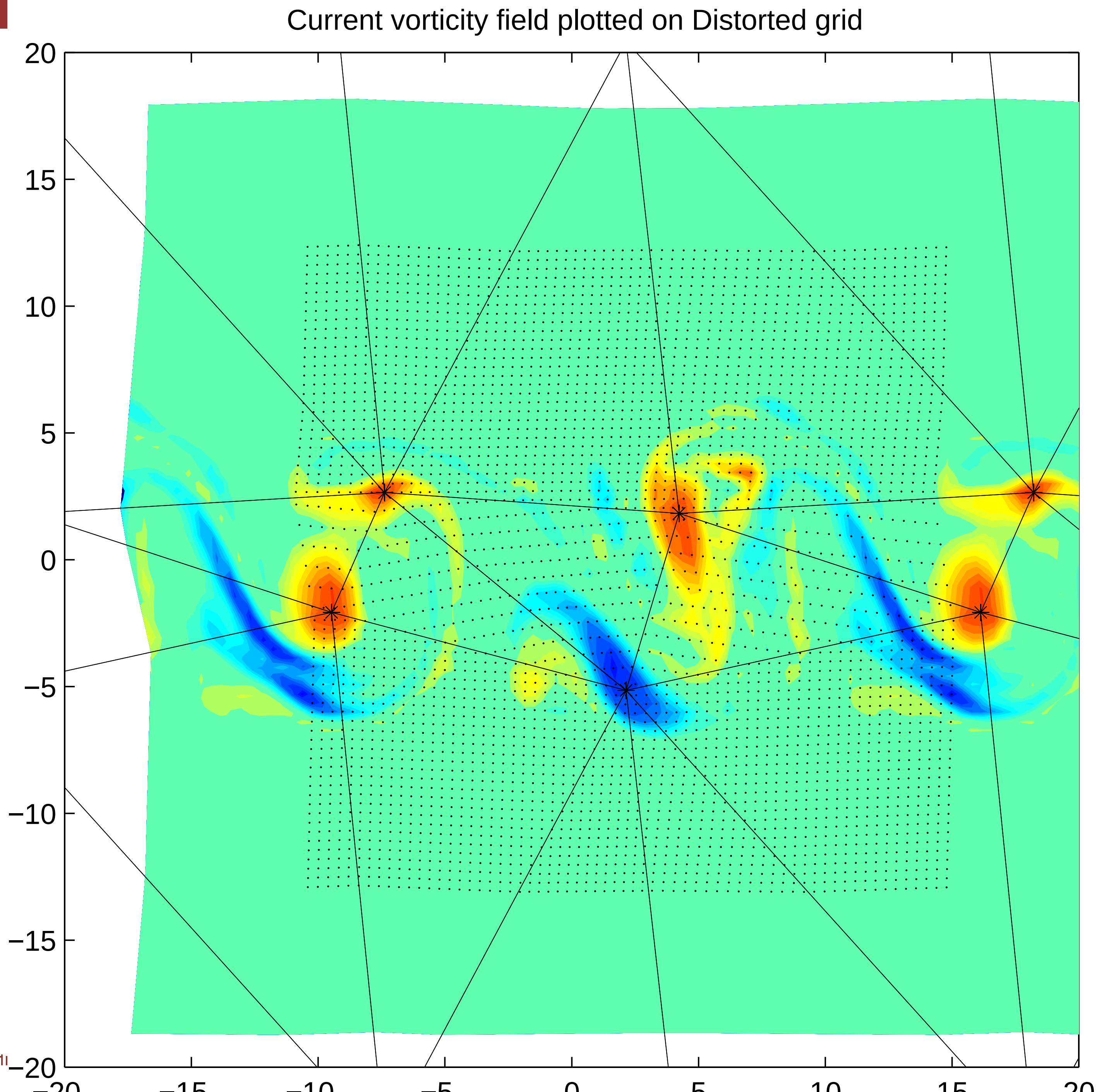


Conclusions

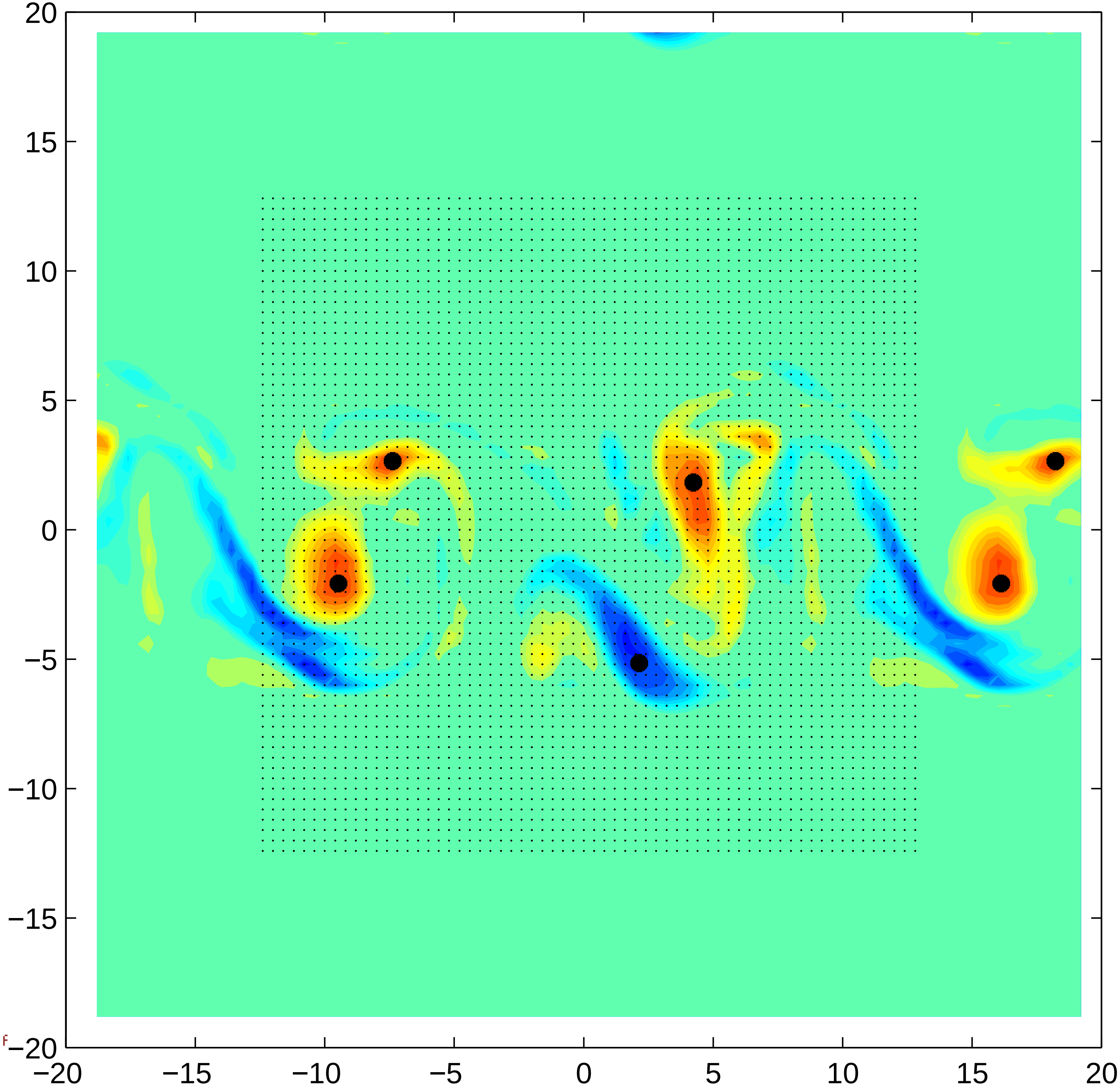
- Ensemble data assimilation is neat
  - covariance for DA
  - covariance for understanding
  - covariance for predictability
- Non-Gaussianity is bad (or at least a problem)
- Gaussian position errors can lead to non-Gaussian amplitude errors
- Alternative error models provide a useful framework for understanding and ameliorating the impact of this type of non-Gaussianity (use physics to inform error model)
- The two-step filter approach is effective
  - primarily due to background covariance information
  - requires an appropriate alignment scheme



Current vorticity field; black is current features, magenta target
20
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
20


Current vorticity field plotted on Distorted grid
20
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
20

Aligned vorticity field after Bicubic Interpolation



"Stochastic" vs. "Deterministic" filters

- Stochastic filter update

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

$$\overline{\mathbf{x}}'^a_i = \overline{\mathbf{x}}'^f_i + \mathbf{K}(\mathbf{y}'^o_i - \mathbf{H}\overline{\mathbf{x}}'^f_i)$$

- Deterministic filter update

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

$$\overline{\mathbf{x}}'^a_i = \overline{\mathbf{x}}'^f_i - \hat{\mathbf{K}}(\mathbf{H}\overline{\mathbf{x}}'^f_i)$$

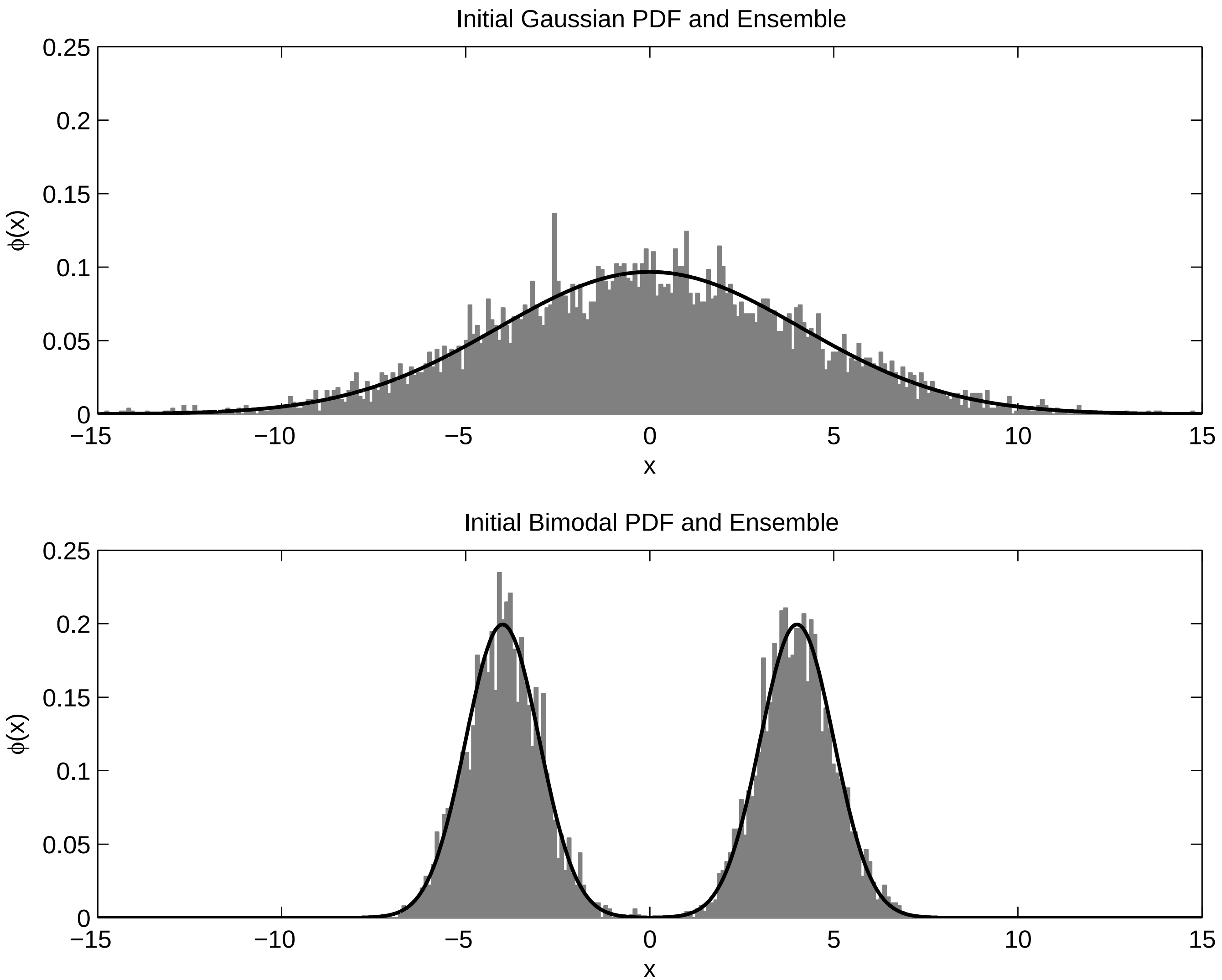
$$\hat{\mathbf{K}} = \left(1 + \sqrt{\frac{\mathbf{R}}{\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}}}\right)^{-1}\mathbf{K}$$

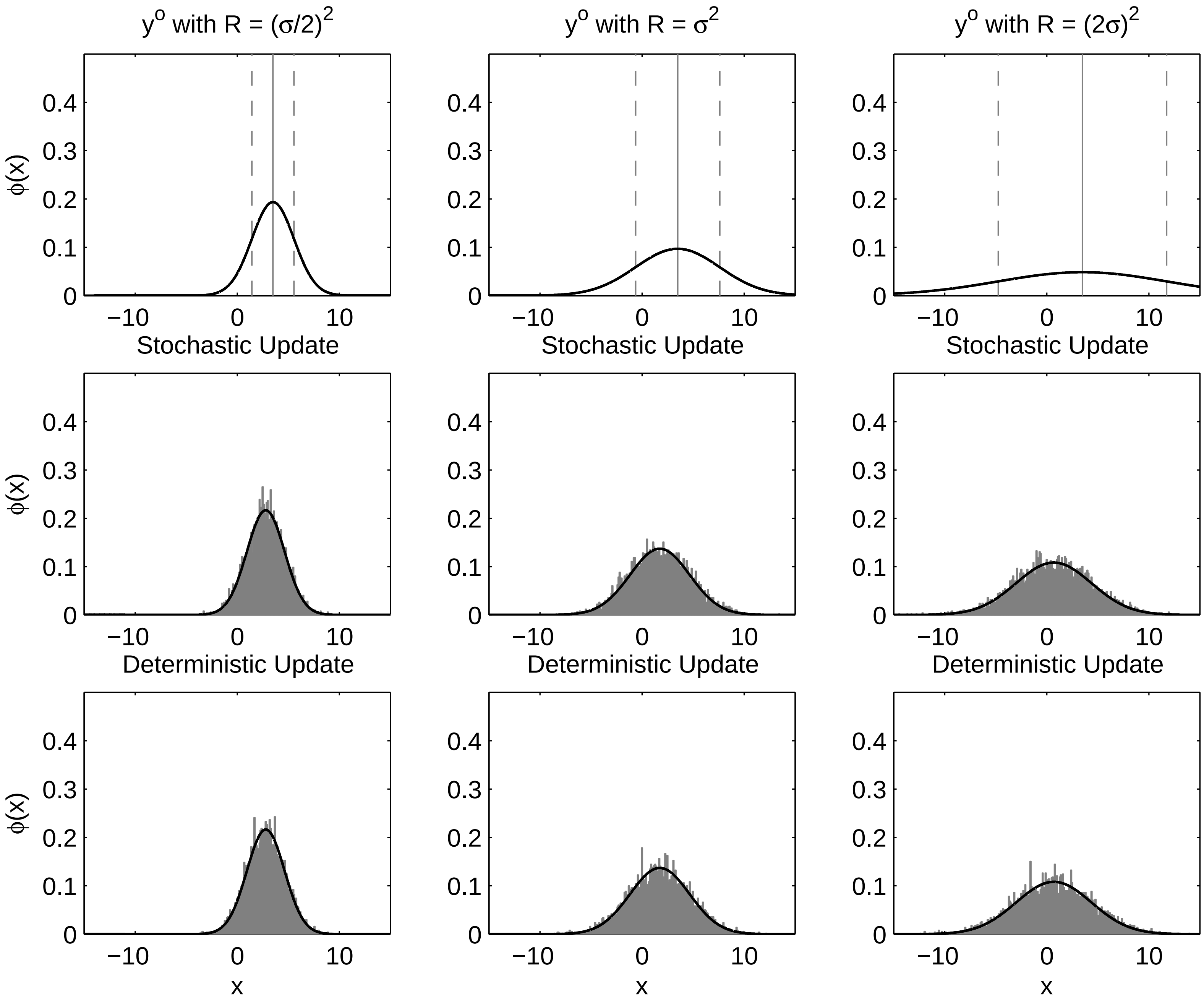
$$\mathbf{P}^a(t) = \mathbf{P}^f(t) - \mathbf{K}(t)\mathbf{H}'(t)\mathbf{P}^f(t)$$

$$= \frac{1}{N_{ens} - 1}[\mathbf{A}^a(t) - \mathbf{M}^a(t)][\mathbf{A}^a(t) - \mathbf{M}^a(t)]^T$$

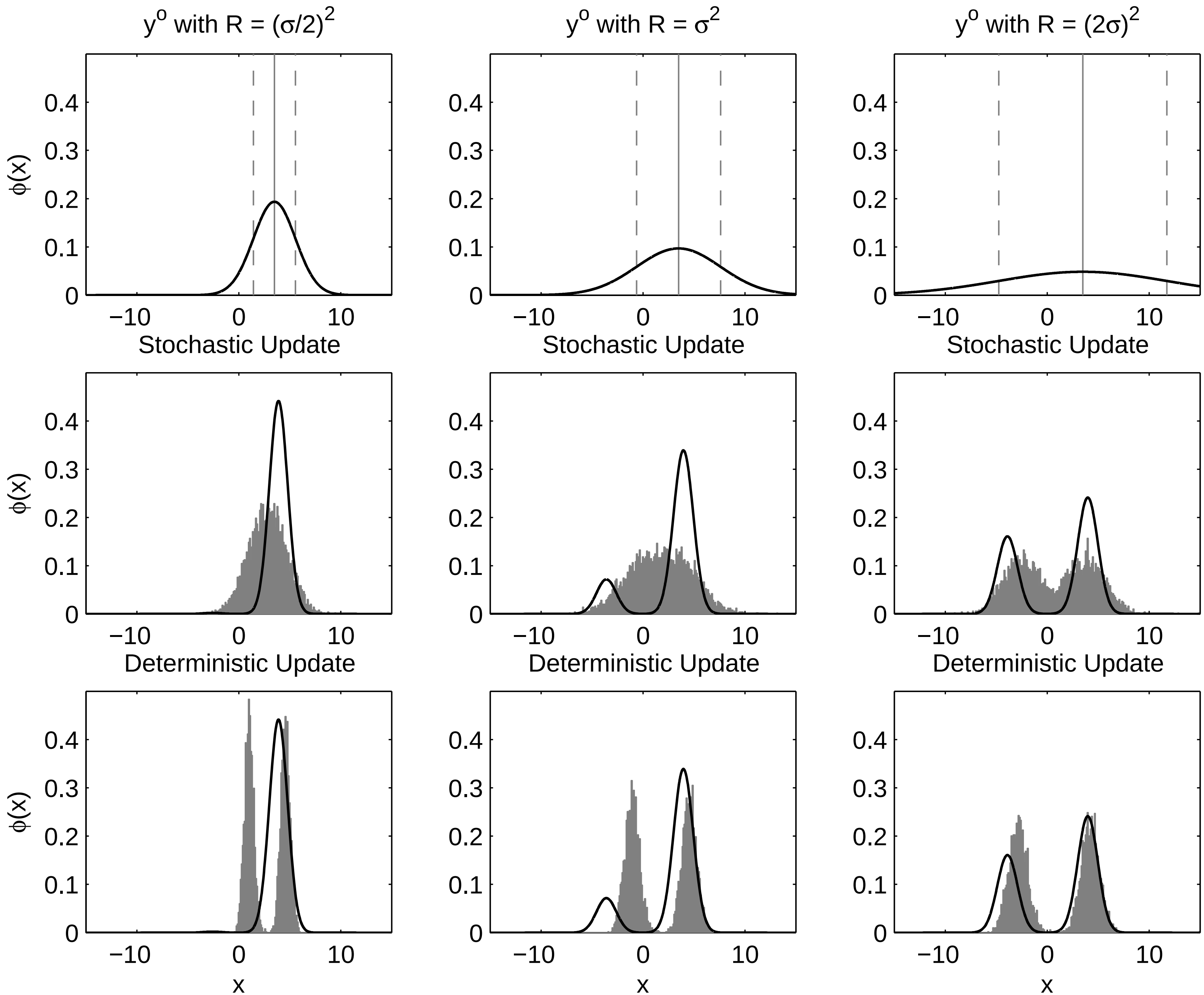
Effects of non-Gaussianity

2 PDFs with identical means and variances





y^o with R = (σ/2)^2
y^o with R = σ^2
y^o with R = (2σ)^2
φ(x)
0
0.1
0.2
0.3
0.4
−10
0
10
Stochastic Update
Deterministic Update
x


y^o with $R = (\sigma/2)^2$
y^o with $R = \sigma^2$
y^o with $R = (2\sigma)^2$
φ(x)
0.4
0.3
0.2
0.1
0
−10
0
10
Stochastic Update
Deterministic Update
x

The ensemble Kalman filter

$$For\ i = 1, N_{ens}$$

$$\mathbf{x}_i^f(t) = \mathbf{F}(\mathbf{x}_i^a(t-\tau))$$

$$\mathbf{P}^f(t) = \frac{1}{N_{ens}-1}[\mathbf{A}^f(t) - \mathbf{M}^f(t)][\mathbf{A}^f(t) - \mathbf{M}^f(t)]^T$$

$$\mathbf{K}(t) = \mathbf{P}^f(t)\mathbf{H}(t)^T[\mathbf{H}(t)\mathbf{P}^f(t)\mathbf{H}(t)^T + \mathbf{R}(t)]^{-1}$$

$$(\mathbf{K} = \frac{\mathbf{P}^f}{\mathbf{P}^f + \mathbf{R}})$$

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

$$\mathbf{P}^a(t) = \frac{1}{N_{ens}-1}[\mathbf{A}^a(t) - \mathbf{M}^a(t)][\mathbf{A}^a(t) - \mathbf{M}^a(t)]^T$$

Ensemble Canonical Correlation Regression (ECCR) Piecewise PV Inversion

- **Project the state and PV ensemble anomalies onto their principal components**
  - Orthogonalizes the predictors, thereby eliminating multicollinearities

$$\mathbf{U}_P = \mathbf{P}_e^{'}\mathbf{E} \qquad \mathbf{U}_Y = \mathbf{Y}_e^{'}\mathbf{F}$$

- **Truncate the principal components**
  - Further reduces the dimensionality, thereby minimizing overfitting and regression instability
  - Determine the numbers of retained principal components via leave-one-out cross validation

- **Project principal components onto bases in which projections are maximally correlated**
  - Identifies the subspace in which predictors and predictands are maximally correlated

$$\mathbf{V} = \mathbf{A}\mathbf{U}_P \qquad \mathbf{W} = \mathbf{B}\mathbf{U}_Y$$

$$\mathbf{A}^{\mathrm{T}} = \mathbf{S}_{U_P U_P}{}^{-1/2}\mathbf{E}_{U_P} \qquad \mathbf{B}^{\mathrm{T}} = \mathbf{S}_{U_Y U_Y}{}^{-1/2}\mathbf{F}_{U_Y}$$

$$\mathrm{SVD}\left(\mathbf{S}_{U_P U_P}{}^{-1/2}\mathbf{S}_{U_P U_Y}\mathbf{S}_{U_Y U_Y}{}^{-1/2}\right) = \mathbf{E}_{U_P}\mathbf{R}_{\mathbf{c}}\mathbf{F}_{U_Y}{}^{\mathrm{T}}$$

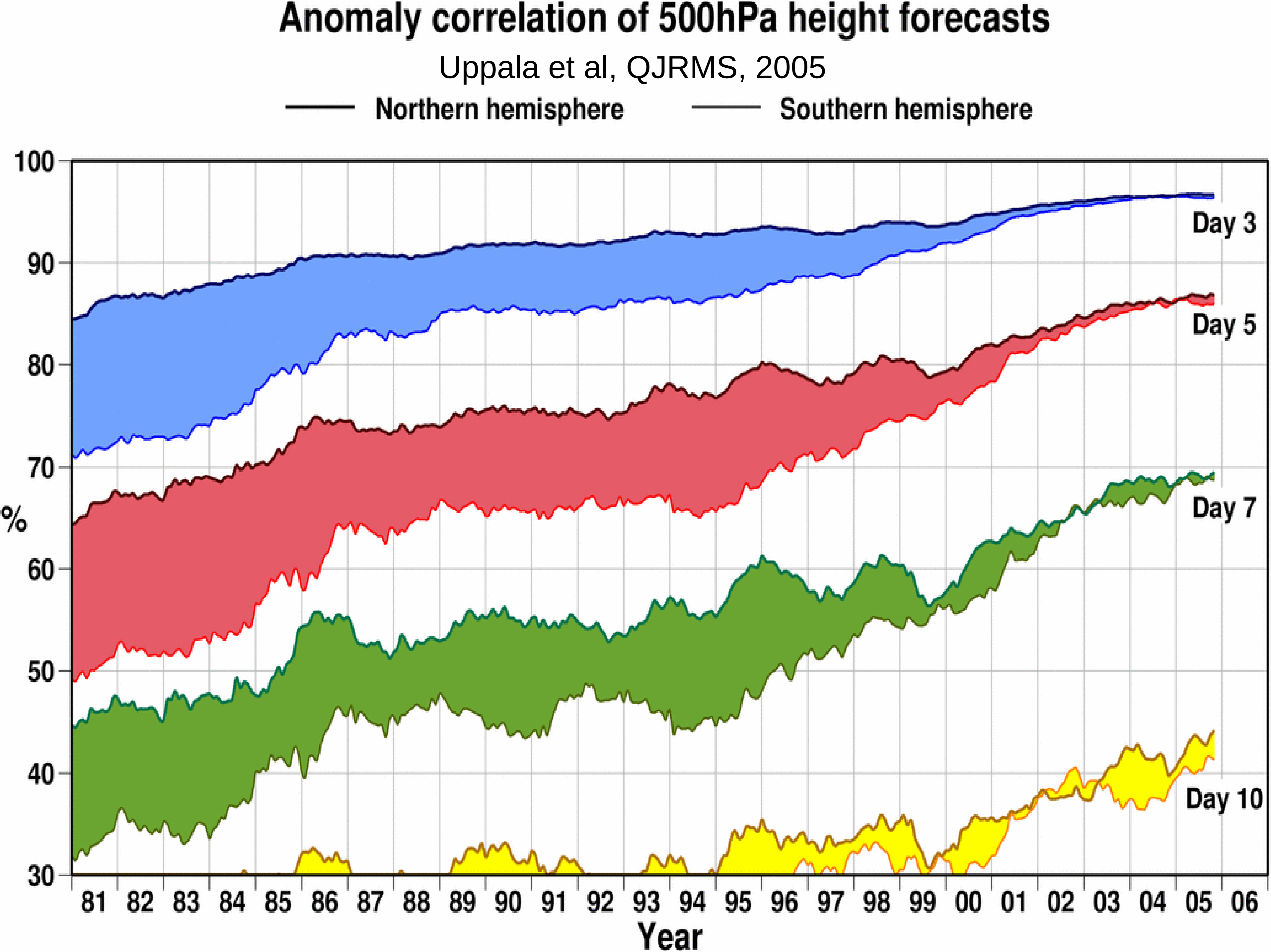
- **Perform the regression in principal component CCA space**

$$\hat{\mathbf{W}} = \mathbf{R}_{\mathbf{c}}\mathbf{V}$$

Augmented control vector sample covariance

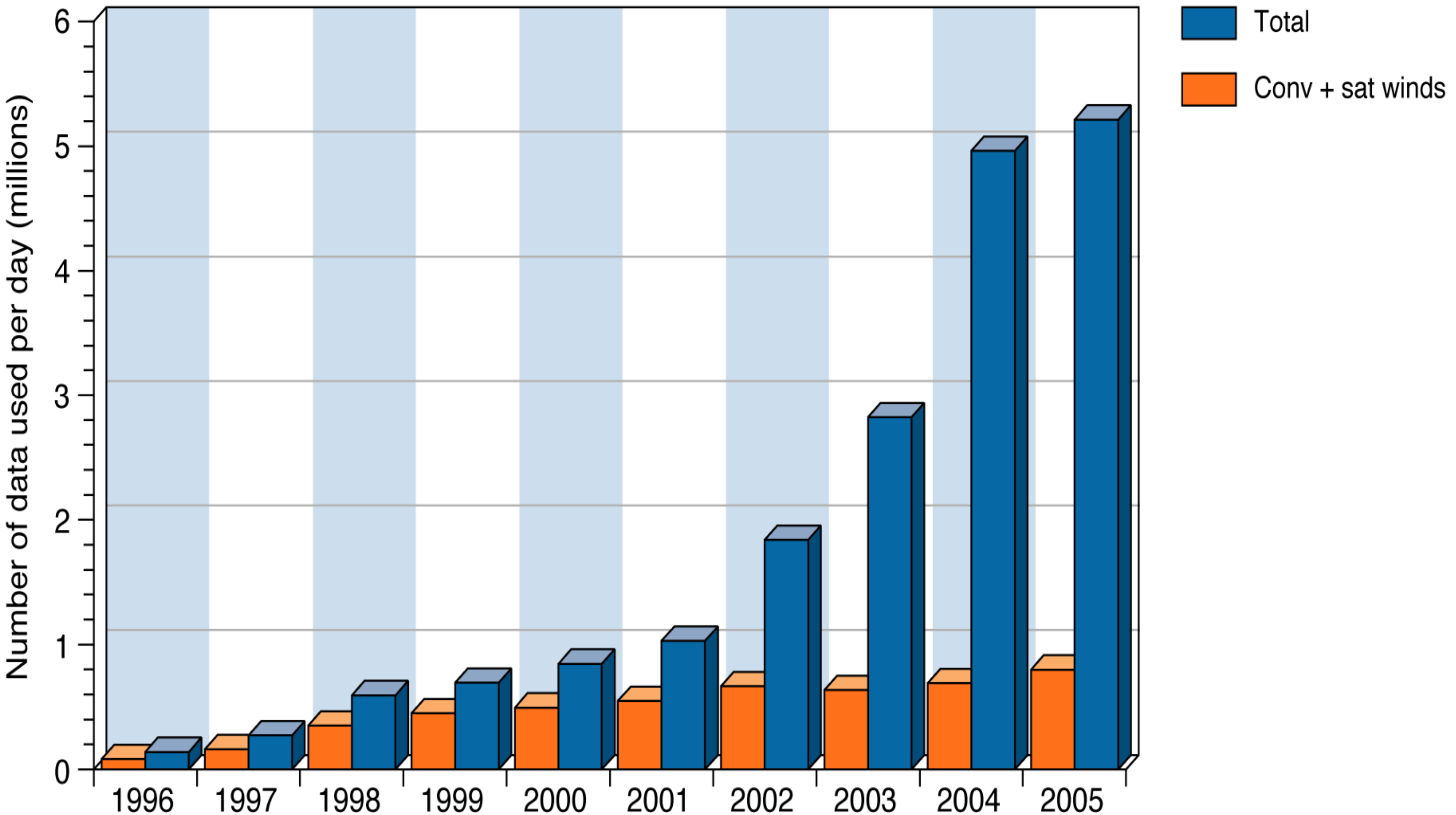
$$\tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\alpha} \end{bmatrix}$$

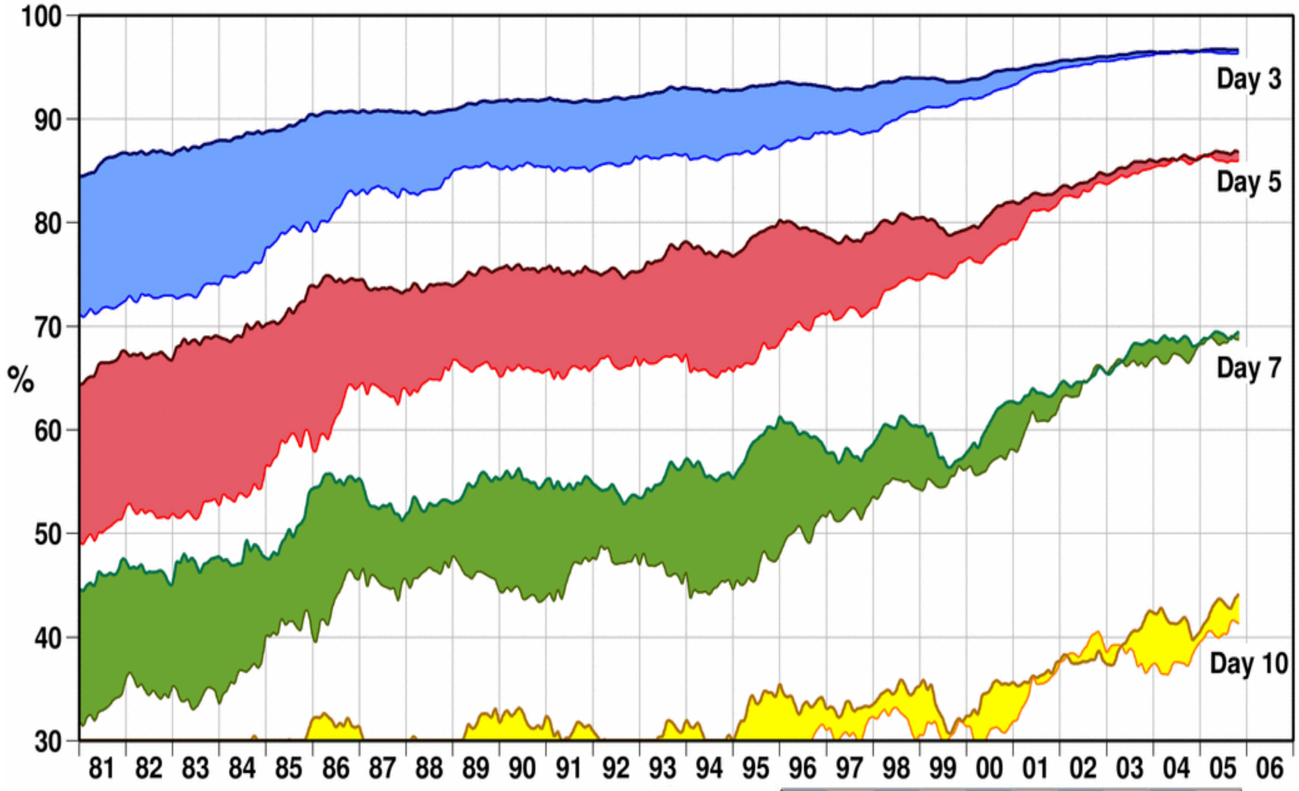
$$\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T = \left[\begin{array}{c:c} \mathbf{x}\mathbf{x}^T & \mathbf{x}\boldsymbol{\alpha}^T \\ \hdashline \boldsymbol{\alpha}\mathbf{x}^T & \boldsymbol{\alpha}\boldsymbol{\alpha}^T \end{array}\right]$$



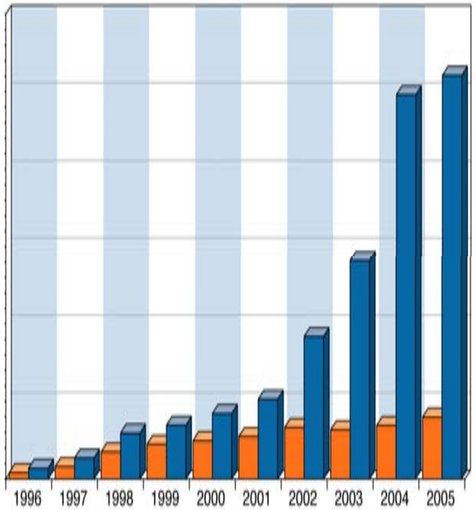
Anomaly correlation of 500hPa height forecasts
Uppala et al, QJRMS, 2005
Northern hemisphere
Southern hemisphere
100
90
80
70
60
50
40
30
%
Day 3
Day 5
Day 7
Day 10
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
00
01
02
03
04
05
06
Year

Number of data used per day



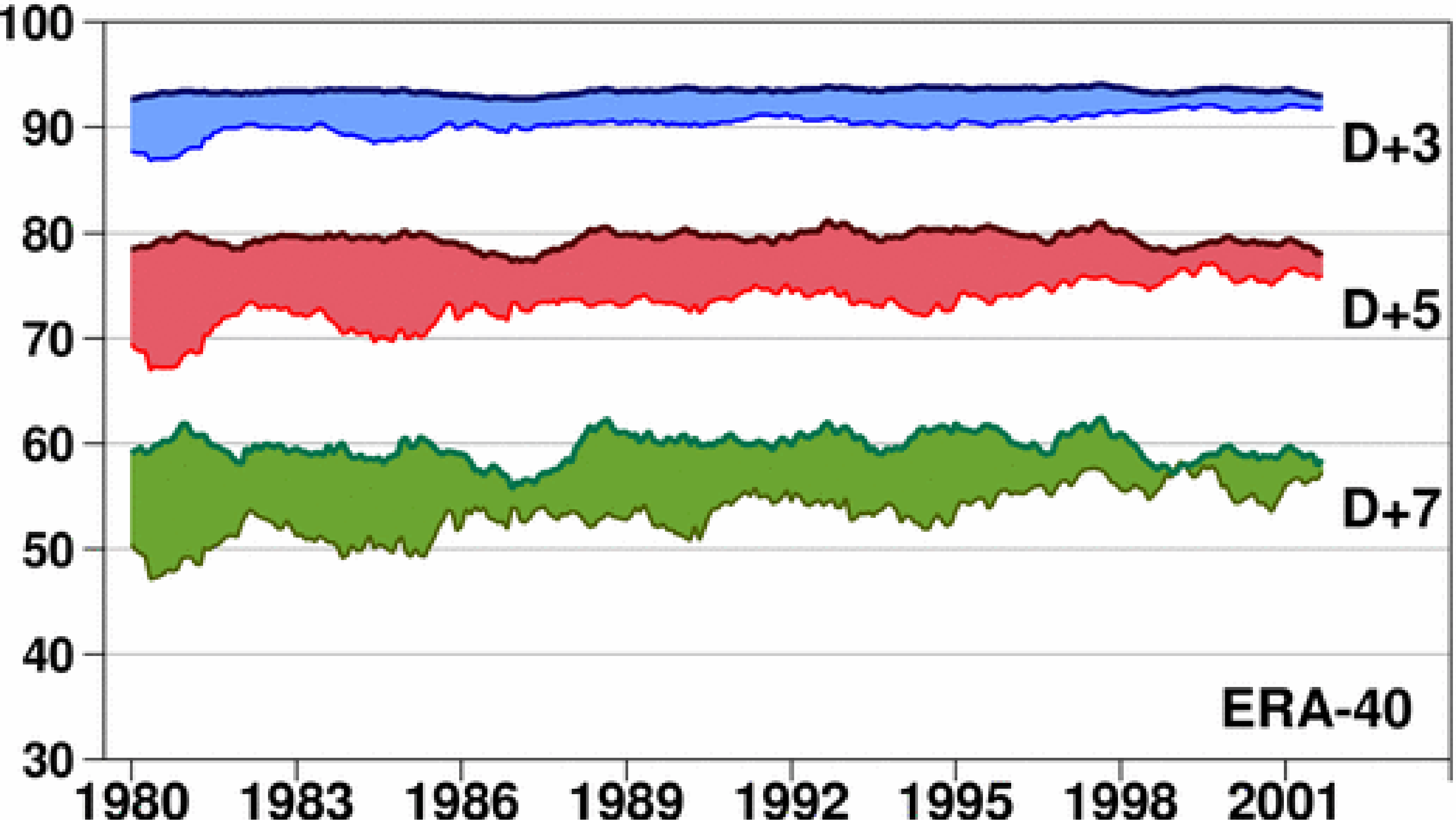


100
90
80
70
60
50
40
30
%
Day 3
Day 5
Day 7
Day 10
81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06

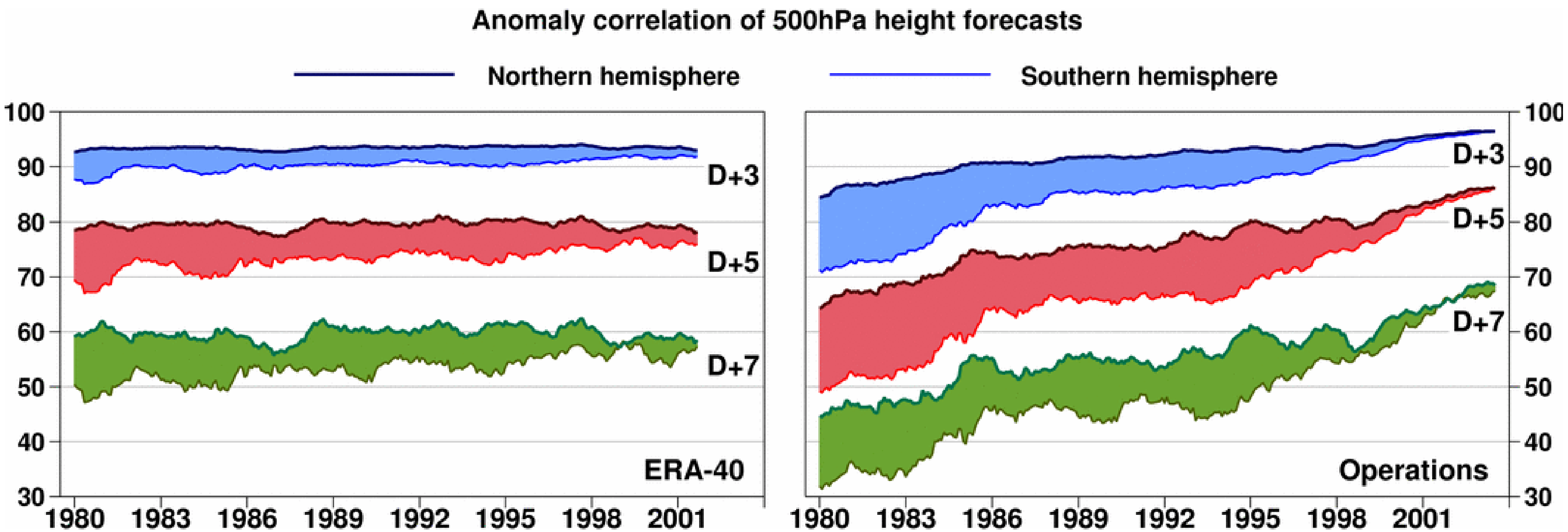


1996 1997 1998 1999 2000 2001 2002 2003 2004 2005

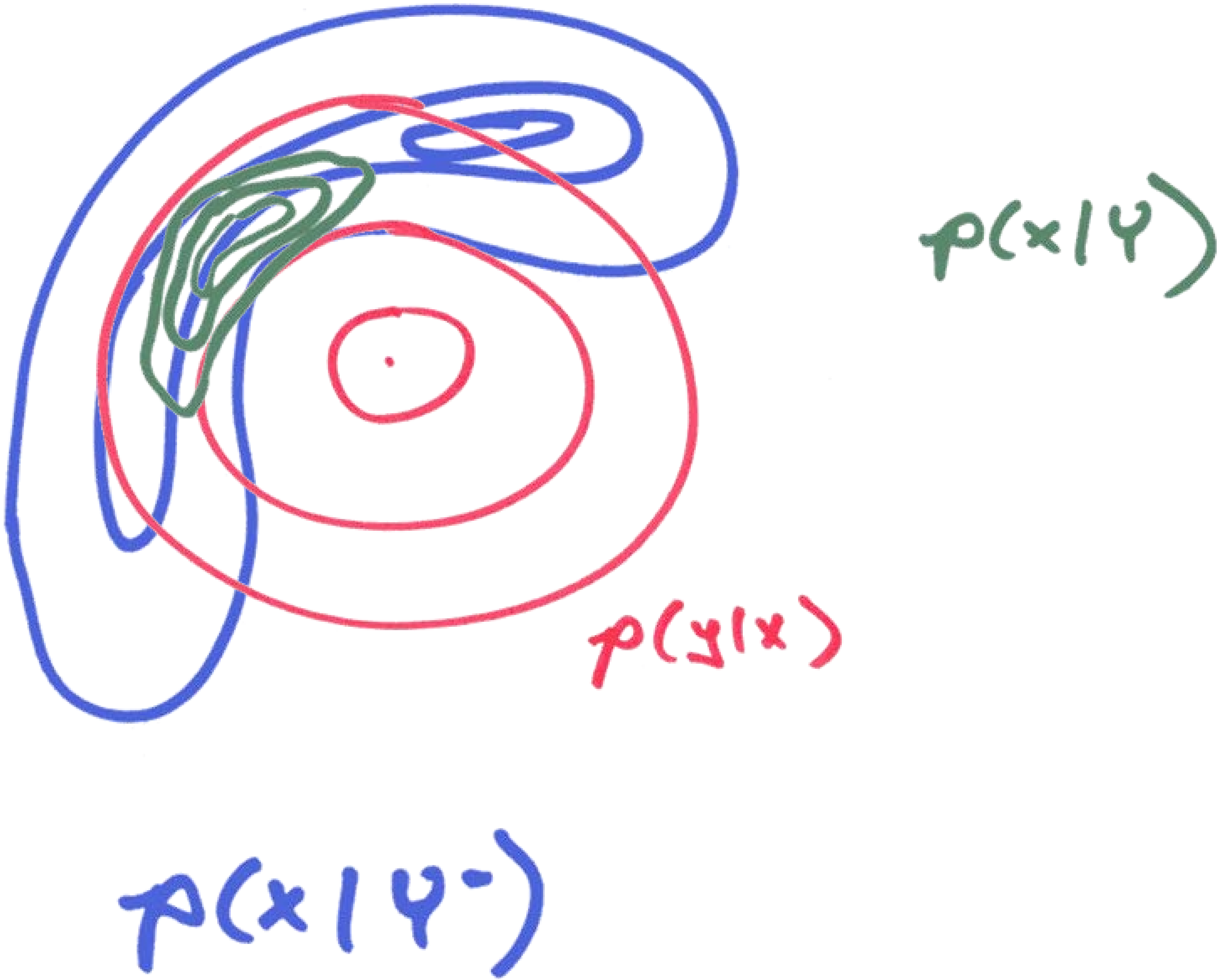
Fixing the data assimilation scheme (3d-Var) and model (T159L60)

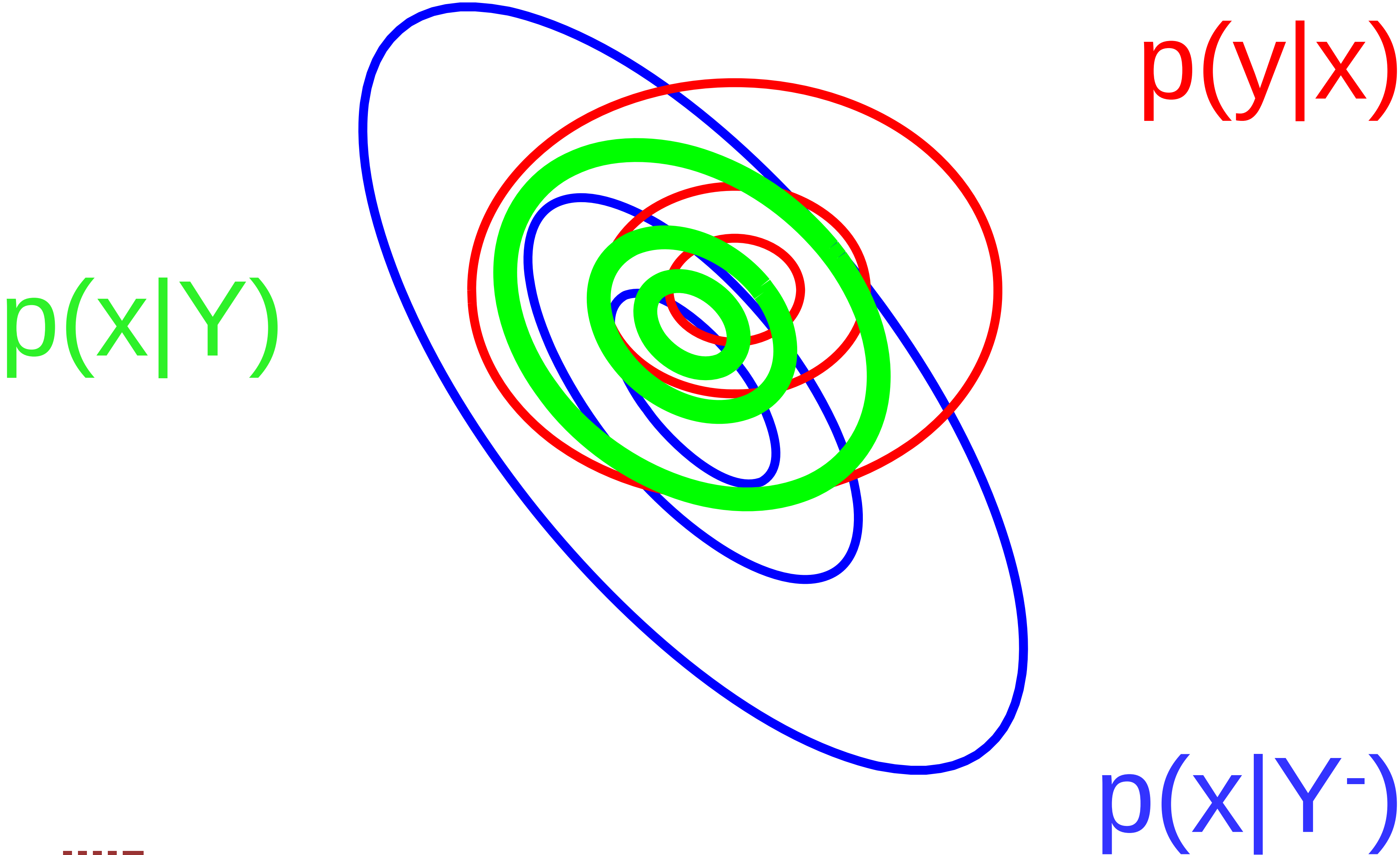


Uppala et al, QJRMS, 2005



- Increasing observations without improving DA or model has minimal impact (NH).
- DA research needs to continue in order to utilize information available in our current (and future) observing system (at least that's the story I tell my students!)


$p(x|y)$
$p(y|x)$
$p(x|y^-)$


$p(y|x)$
$p(x|Y)$
$p(x|Y^-)$