Assimilating Data into Models: Nonlinearity vs. Dimension

Christopher Jones, University of Warwick and
University of North Carolina at Chapel Hill

New Directions in Applied Mathematics

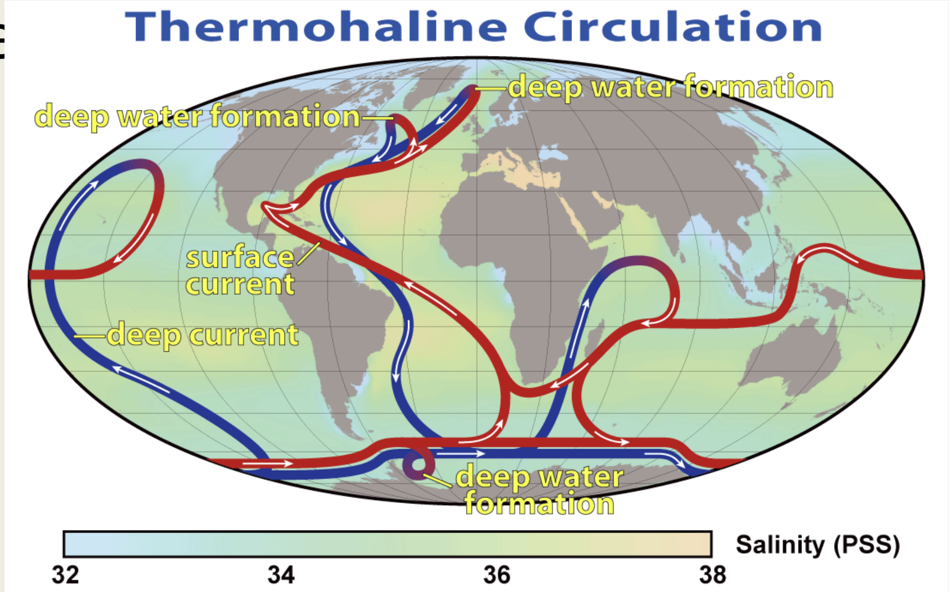
ICTS, Bangalore, India

January, 2010

Supported by ONR and NSF

Climate Change

- Climate issues will drive much of 21st Century science
- Our current understanding of the Earth System and its climate is akin to our knowledge of the human body in the 18th Century (Lovelock)
- What role
take in th



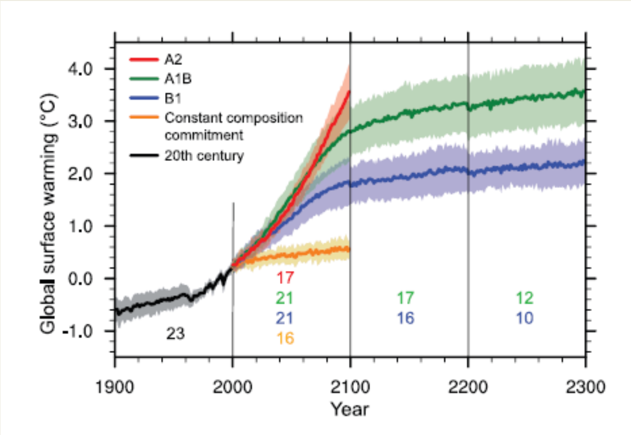
Greenhouse Effect

- Incoming short wavelength radiation does not interact with greenhouse gases,
- Longer wavelength reflected radiation does!

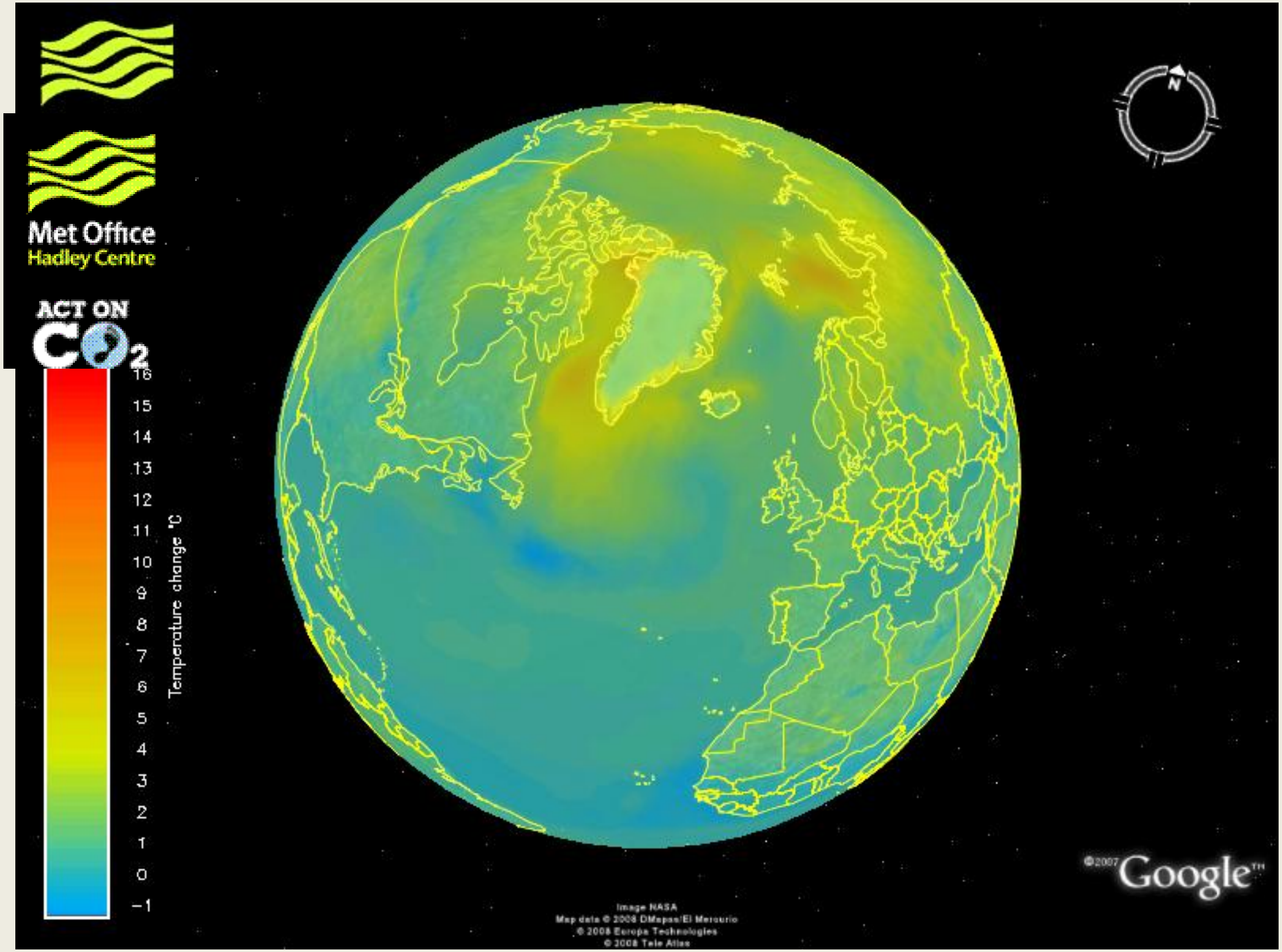
Joseph Fourier, 1824

Back of the envelope:

- 350-400 ppm CO_2 -> 1-2 deg C over 21st C
- 400-450 ppm CO_2 -> 2-3 deg C over 21st C
- 450-500 ppm CO_2 -> 3-4 deg C over 21st C



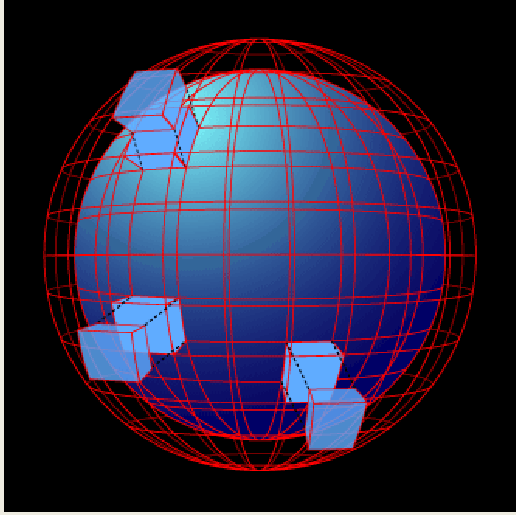
IPCC: 17 Modeling centers (2007) running "big" models. Results are averaged to make predictions



Improve resolution:
Predict down to e.g. 25km in 2050

Improve understanding:
What determines regional temperature distribution?

Role of Applied Mathematics



- Only ONE Earth
- Only ONE realization of Earth System
- Need mathematical models to test hypotheses
- See what happens if …

BUT: We get ourselves in deep water

USGS
science for a changing world

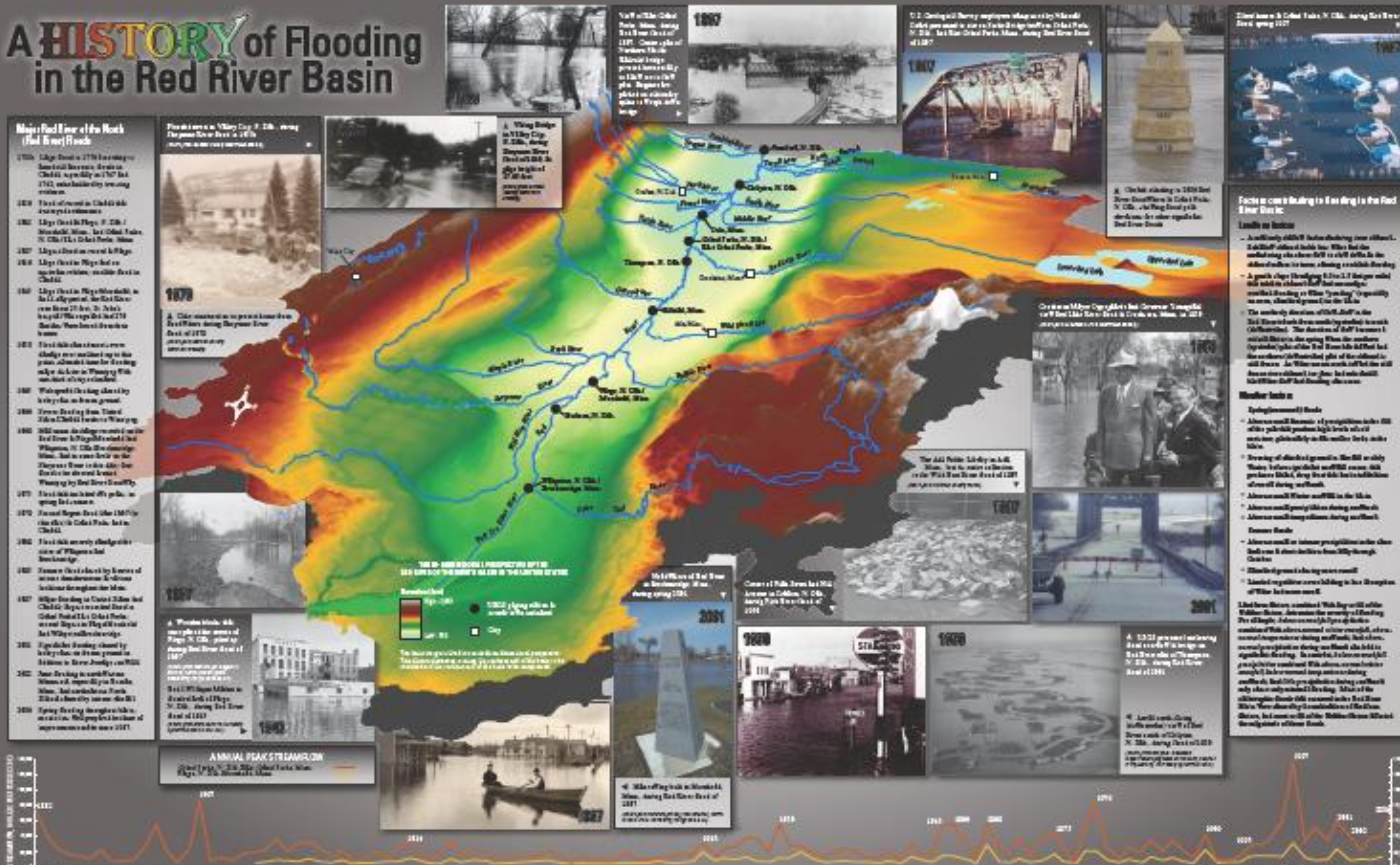
North Dakota State Water Commission

Red River Watershed Management Board

Red River Joint Water Resource District

Upper Sheyenne River Joint Water Resource Board

A HISTORY of Flooding in the Red River Basin



Flood of criticism from 1997 floods: Did faulty forecasts add to disaster?

For six weeks, the National Weather Service had predicted a crest of 49 feet at Grand Forks. Then, over the five days before the river burst through its restraints, forecasters methodically revised it higher, eventually to 54 feet - a difference that spelled disaster in this pancake-flat region.
From evacuation centers to city offices, the same anguished question now arises: How could forecasters have been so far off?

9th April, 2007:

Forecasters are still stung by the spray-painted words, many of them obscene, on what was left of flood-ruined homes after the Red River swamped this city a decade ago.

Mayor of East Grand Forks:
"They blew it big!"

Importance of Data

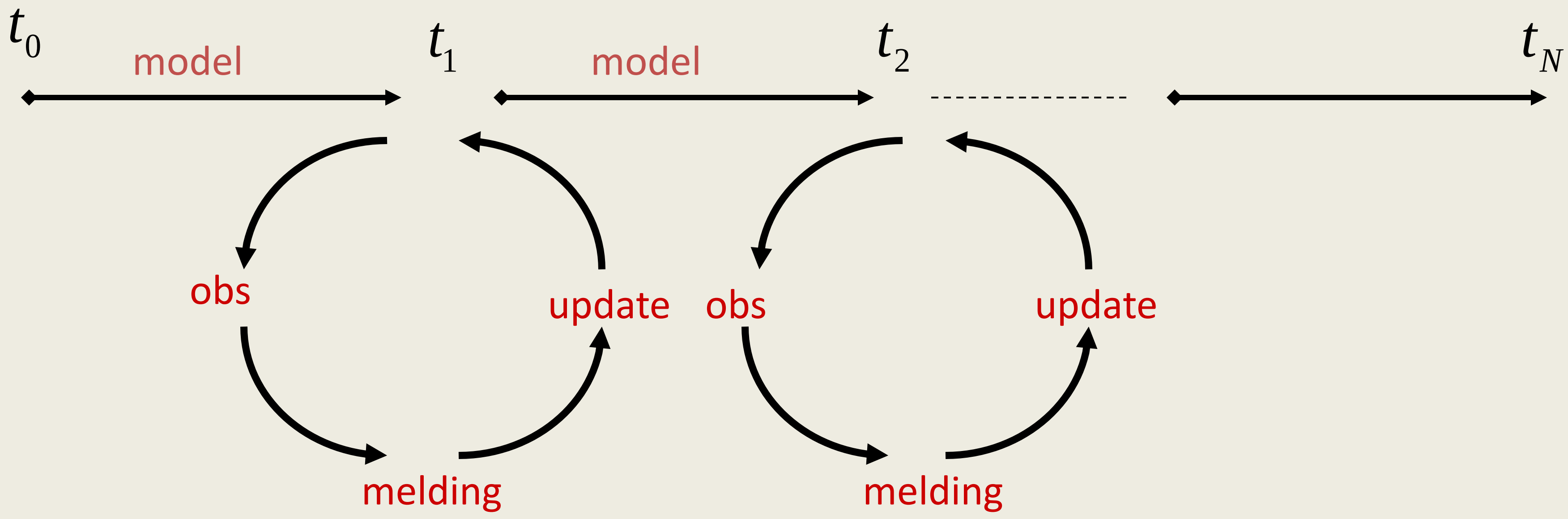
Computer models use data collected over years, translating stream flows into depth predictions for points along the river. But when stream flows are off the chart, as they were along the Red, the models go out the window.

Dean Braatz, then head of the weather service's river-forecasting effort for North Dakota and Minnesota

For accurate predictions, forecasters had to wait to measure actual flood depths at particular points and project them downstream to Grand Forks.

Sequential Data Assimilation

Model + observations ⟶ prediction



Fishkill in Lake Kinneret

Vernieres et al. (2006)

Conjecture: due to “lifting” of lower layer of oxygen-free water

- Occasional “fishkill”
- Feeding of 5,000??

Model:

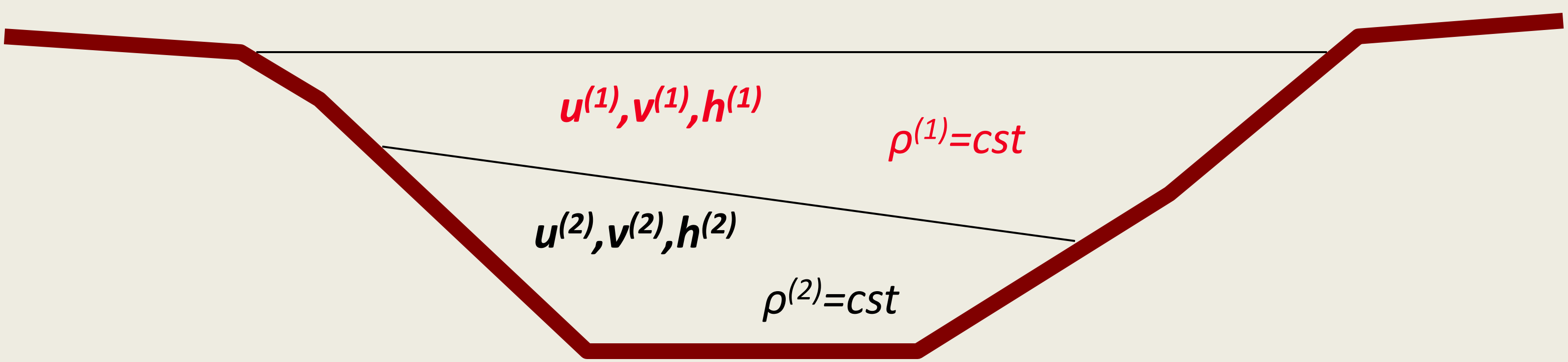
- Stably stratified during summer
- Strong westerly sea breeze

$$\left.\begin{aligned}
&\frac{Du^{(1)}}{Dt} - fv^{(1)} + g\frac{\partial}{\partial x}\left(h^{(1)} + h^{(2)} + D\right) - A_h\nabla^2 u^{(1)} - F_u = 0\\
&\frac{Dv^{(1)}}{Dt} + fu^{(1)} + g\frac{\partial}{\partial y}\left(h^{(1)} + h^{(2)} + D\right) - A_h\nabla^2 v^{(1)} - F_u = 0
\end{aligned}\right\} \text{Top layer momentum}$$

$$\frac{Dh^{(1)}}{Dt} = 0$$

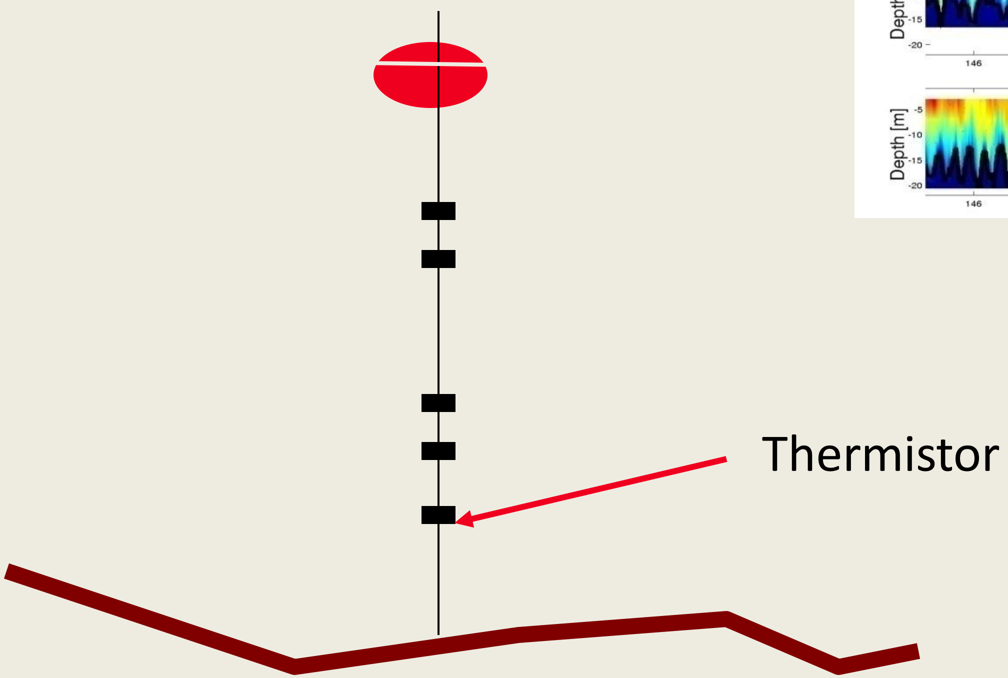
$$\left.\begin{aligned}
&\frac{Du^{(2)}}{Dt} - fv^{(1)} + g\frac{\partial}{\partial x}\left(h^{(1)} + h^{(2)} + D\right) + g'\frac{\partial}{\partial x}\left(h^{(2)} + D\right) - A_h\nabla^2 u^{(1)} = 0\\
&\frac{Dv^{(2)}}{Dt} + fu^{(1)} + g\frac{\partial}{\partial y}\left(h^{(1)} + h^{(2)} + D\right) + g'\frac{\partial}{\partial y}\left(h^{(2)} + D\right) - A_h\nabla^2 v^{(1)} = 0
\end{aligned}\right\} \text{Bottom layer momentum}$$

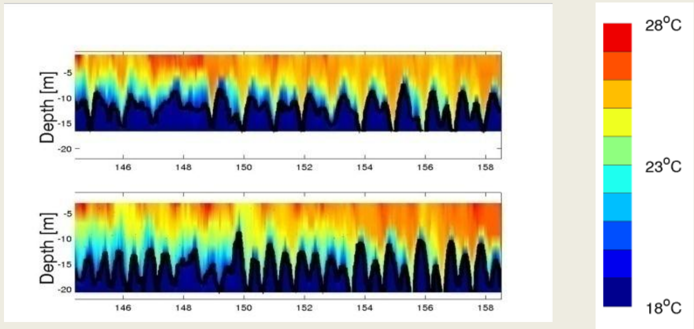
$$\frac{Dh^{(2)}}{Dt} = 0$$



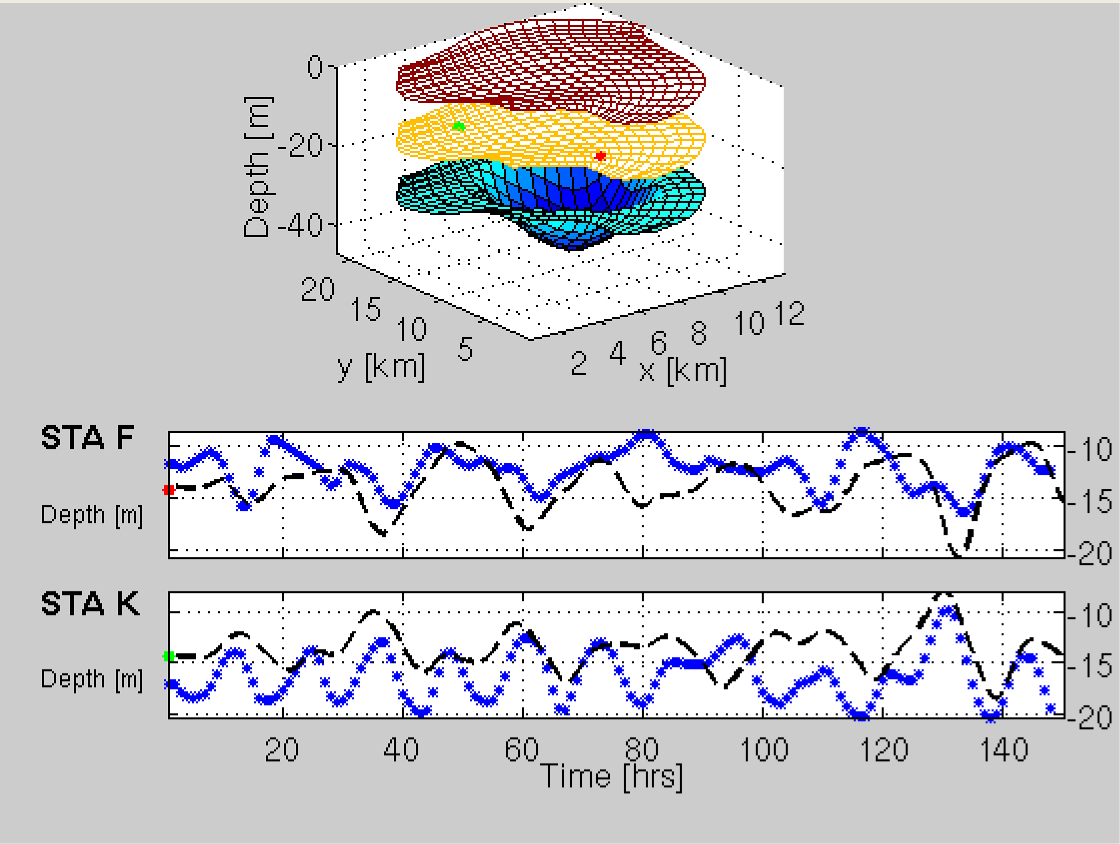
Data

- Thermistor chains

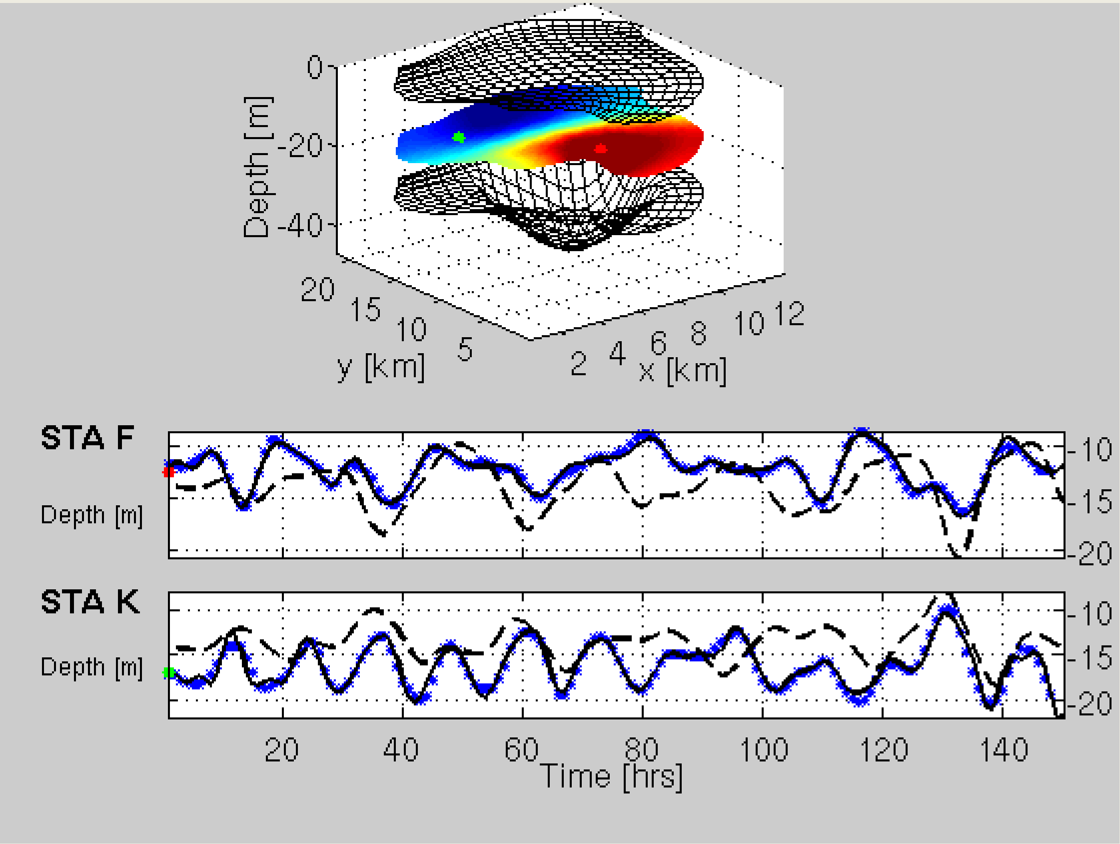




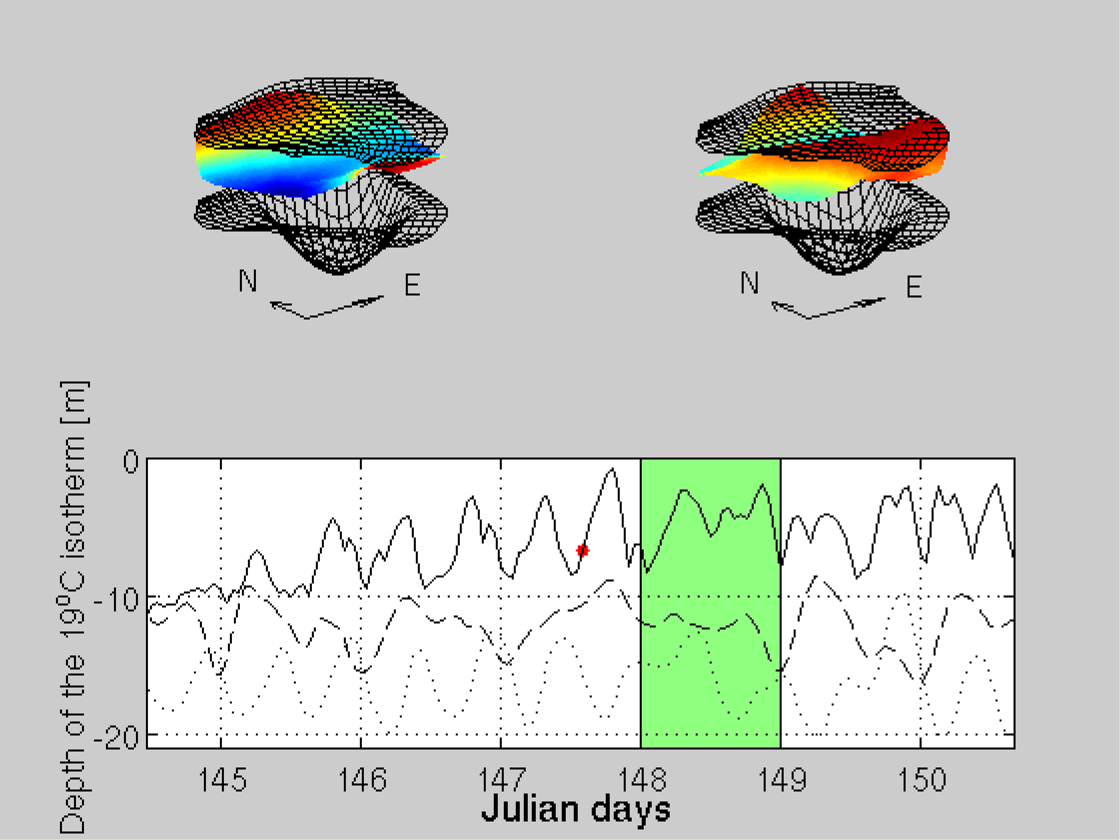
Model running on its own…



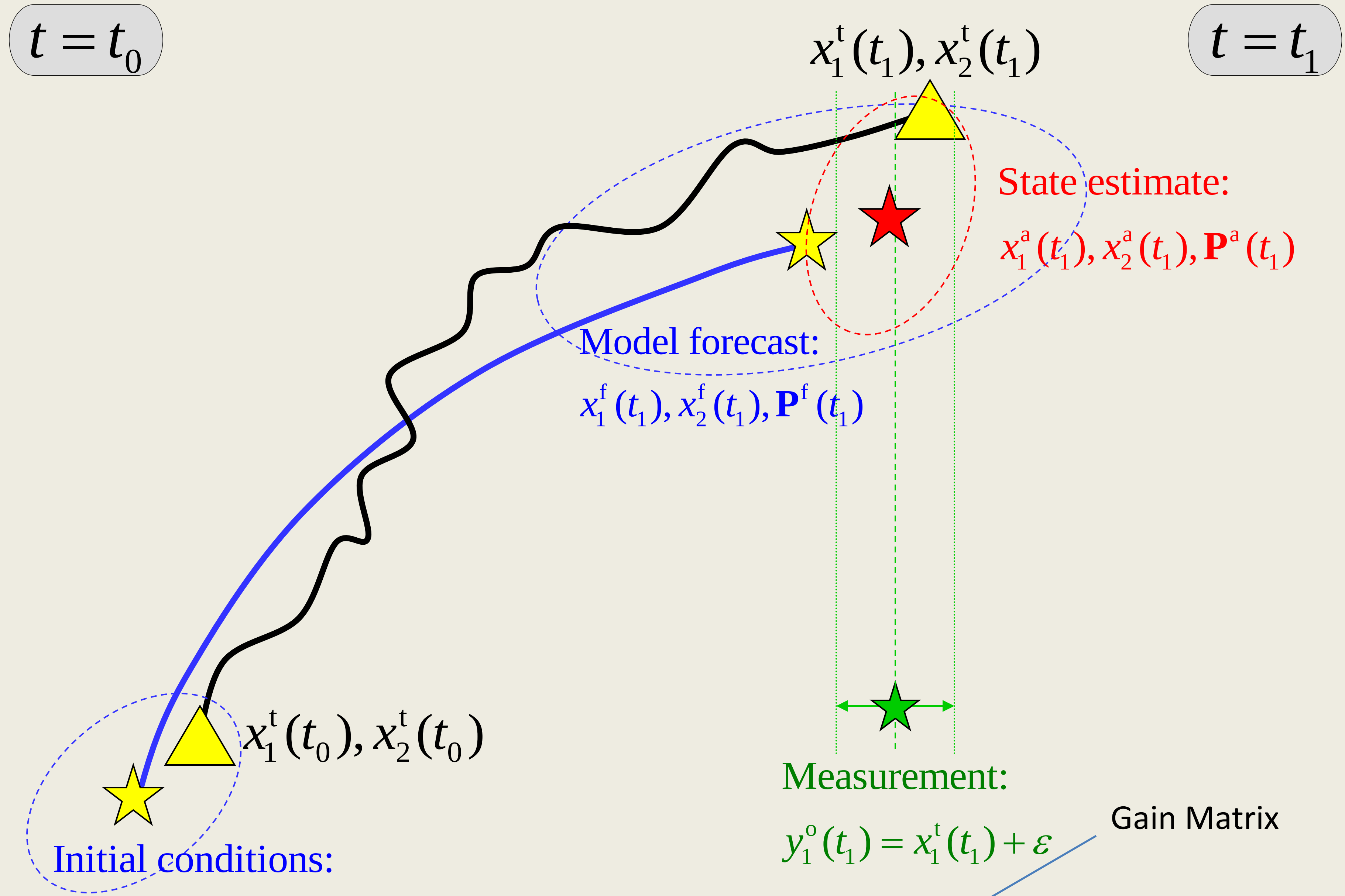
With thermistor data assimilated…



A comparison



Neither model nor data on their own show “fishkill,”
But, together, they do!



$$x^a = x^f + K\left(y - H(x^f)\right)$$

Kalman Filter

- Distributions are Gaussian
- Model is linear-TLM (EKF)
- or fit to Gaussian (EnKF)

Extended Kalman Filter

Forecast model error covariance using tangent linear model:

$$\mathbf{P}^{\mathrm{f}} = E[\Delta\mathbf{x}\Delta\mathbf{x}^{T}]; \qquad \Delta\mathbf{x} \equiv \mathbf{x}^{\mathrm{f}} - \mathbf{x}^{\mathrm{t}}$$

$$\frac{d\mathbf{P}^{\mathrm{f}}}{dt} = \mathbf{M}\mathbf{P}^{\mathrm{f}} + \mathbf{P}^{\mathrm{f}}\mathbf{M}^{T} + \mathbf{Q}(t)$$

$\mathbf{M}_i \equiv \partial M(\mathbf{x}^{\mathrm{f}}, t)/\partial\mathbf{x}$: linearized model operator

Combine model and observations into a new state $\mathbf{x}^{\mathrm{a}}$ minimizing $\mathrm{tr}\,\mathbf{P}^{\mathrm{a}}$

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

$$\mathbf{K} = \mathbf{P}^{\mathrm{f}}\mathbf{H}^{T}\left(\mathbf{H}\,\mathbf{P}^{\mathrm{f}}\mathbf{H}^{T} + \mathbf{R}\right)^{-1} \qquad \mathbf{P}^{\mathrm{a}} = \left(\mathbf{I} - \mathbf{K}\,\mathbf{H}\right)\mathbf{P}^{\mathrm{f}}$$

$\mathbf{H} \equiv \partial H / \partial\mathbf{x}$: linearized observation function

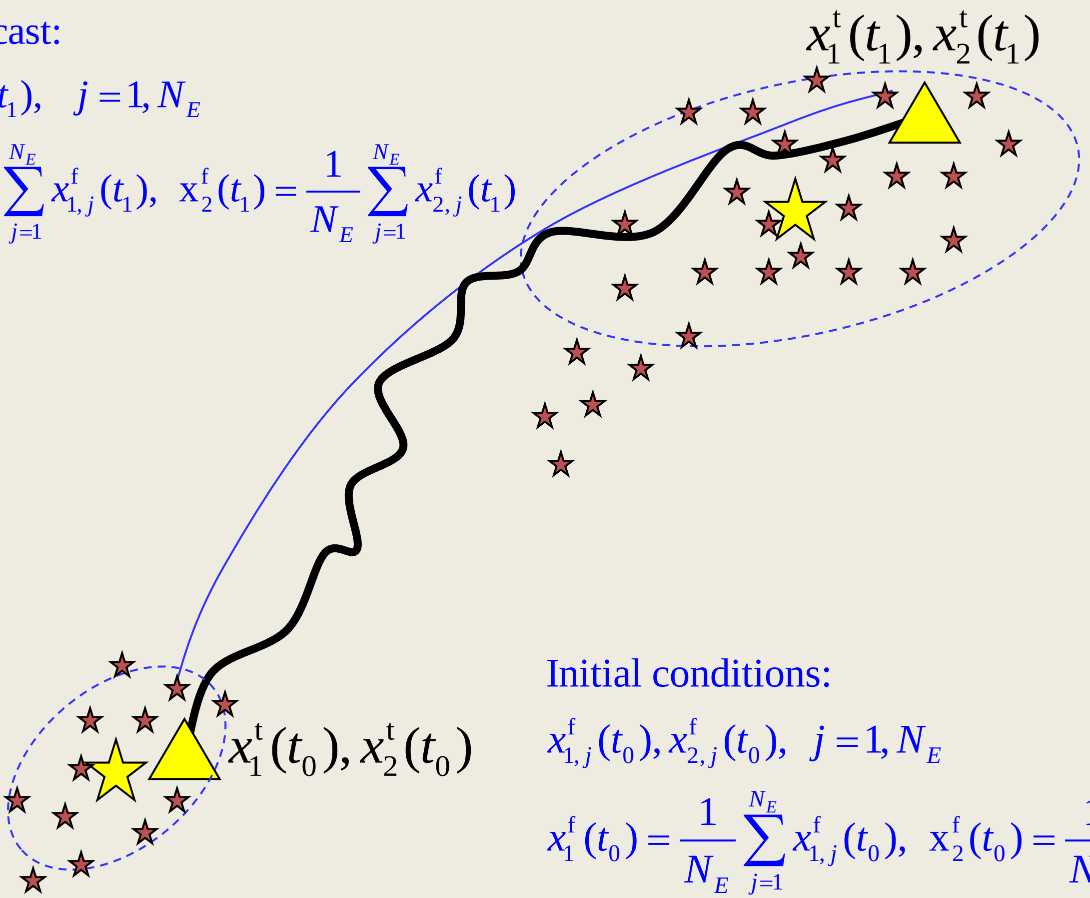
Ensemble Kalman Filter (EnKF)

Error covariance is predicted via solution of full nonlinear system for a Monte-Carlo ensemble of states

Model forecast:

$x^{\mathrm{f}}_{1,j}(t_1), x^{\mathrm{f}}_{2,j}(t_1), \quad j = 1, N_E$

$x^{\mathrm{f}}_1(t_1) = \frac{1}{N_E}\sum_{j=1}^{N_E} x^{\mathrm{f}}_{1,j}(t_1), \ \ x^{\mathrm{f}}_2(t_1) = \frac{1}{N_E}\sum_{j=1}^{N_E} x^{\mathrm{f}}_{2,j}(t_1)$



Initial conditions:

$x^{\mathrm{f}}_{1,j}(t_0), x^{\mathrm{f}}_{2,j}(t_0), \quad j = 1, N_E$

$x^{\mathrm{f}}_1(t_0) = \frac{1}{N_E}\sum_{j=1}^{N_E} x^{\mathrm{f}}_{1,j}(t_0), \ \ x^{\mathrm{f}}_2(t_0) = \frac{1}{N_E}\sum_{j=1}^{N_E} x^{\mathrm{f}}_{2,j}(t_0)$

Update step in EnKF

Kalman gain matrix is computed using error covariance matrix derived from the ensemble. Ensemble members are updated with noisy observations

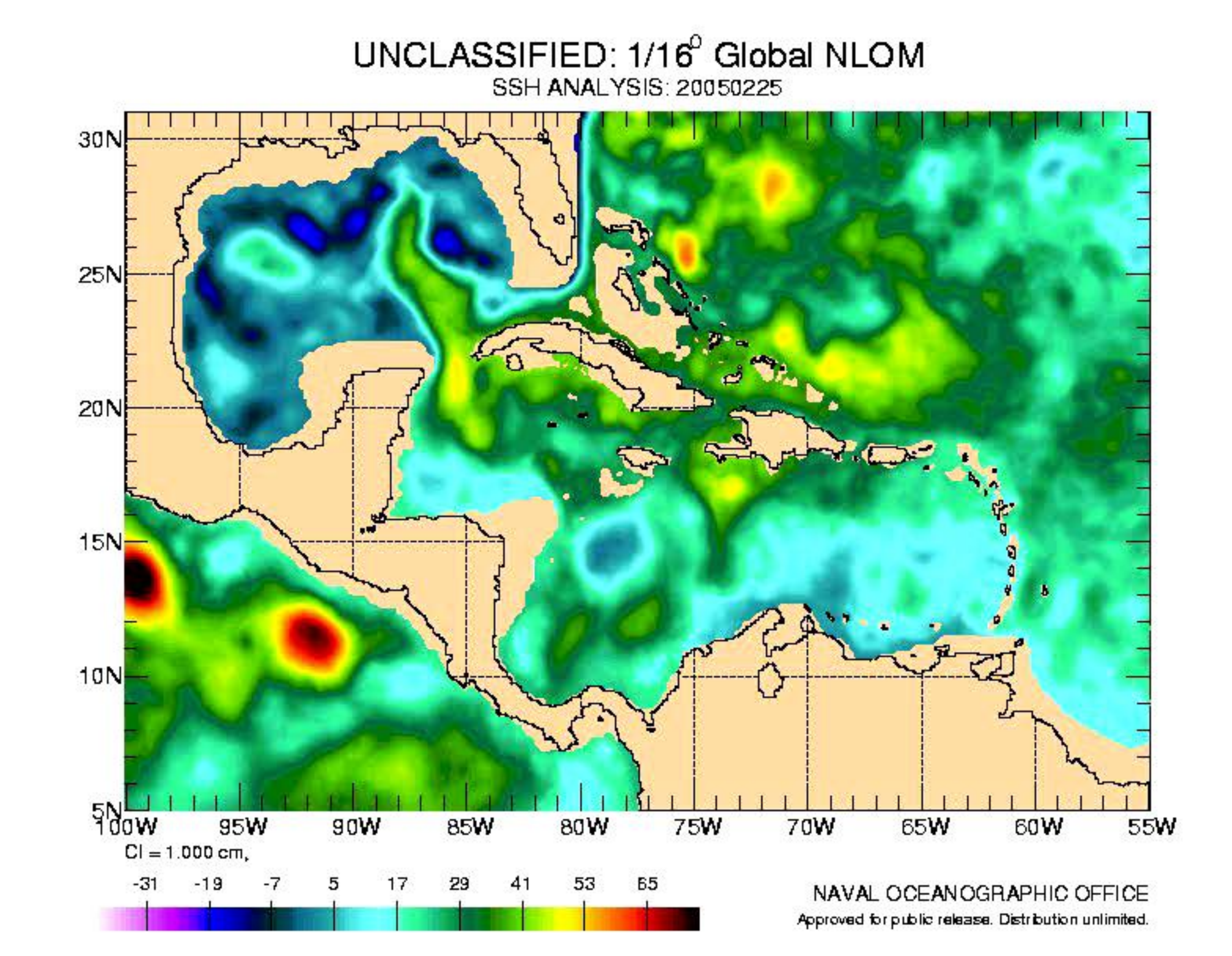
$$\overline{\mathbf{x}}^{\mathrm{f}} = \frac{1}{N_E}\sum_{j=1}^{N_E}\mathbf{x}_j^{\mathrm{f}} \qquad \mathbf{P}^{\mathrm{f}} = \frac{1}{N_E - 1}\sum_{j=1}^{N_E}\left(\mathbf{x}_j^{\mathrm{f}} - \overline{\mathbf{x}}^{\mathrm{f}}\right)\left(\mathbf{x}_j^{\mathrm{f}} - \overline{\mathbf{x}}^{\mathrm{f}}\right)^{\mathrm{T}}$$

Ensemble of observations: $\mathbf{d}_j = \mathbf{y}^{\mathrm{o}} + \tilde{\varepsilon}_j - H(\mathbf{x}_j^{\mathrm{f}})$ $\qquad E[\tilde{\varepsilon}_j \tilde{\varepsilon}_j^T] = \mathbf{R}$

Update ensemble members:

$$\mathbf{x}_j^{\mathrm{a}} = \mathbf{x}_j^{\mathrm{f}} + \mathbf{K}\mathbf{d}_j \qquad \mathbf{K} = \mathbf{P}^{\mathrm{f}}\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^T + \mathbf{R}\right)^{-1}$$

Gulf of Mexico/Caribbean



Thriving on Heat

As Hurricanes Katrina and Rita moved over the Gulf of Mexico, they harvested energy from the warm currents flowing into the gulf from the Caribbean Sea.

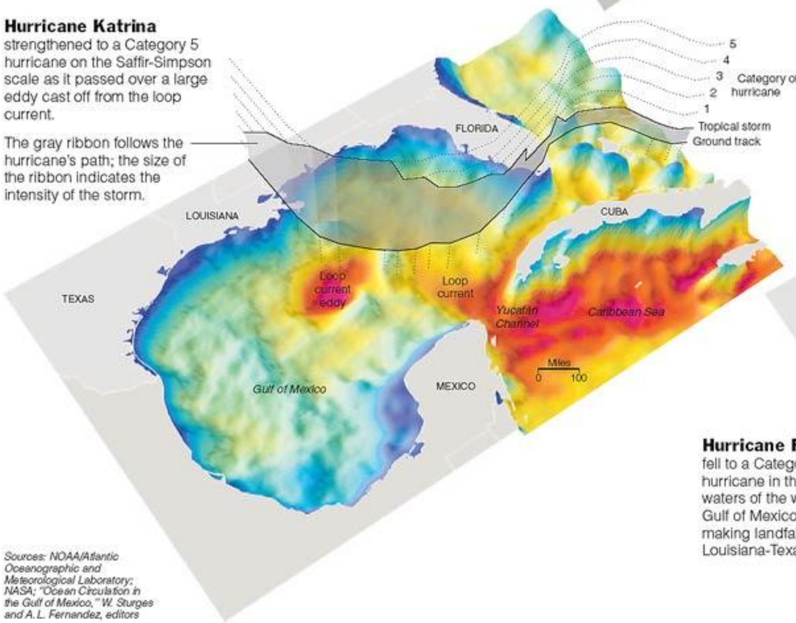
These maps show *tropical cyclone heat potential*, the amount of heat stored in the upper levels of the ocean before each hurricane made landfall on the Gulf Coast. The deeper the warm water, the more heat was available to fuel each hurricane.

The loop current drives the circulation of water in the Gulf of Mexico. The warm water of the loop current enters the gulf through the Yucatán Channel and meanders toward the tip of Florida, eventually helping to form the Gulf Stream.

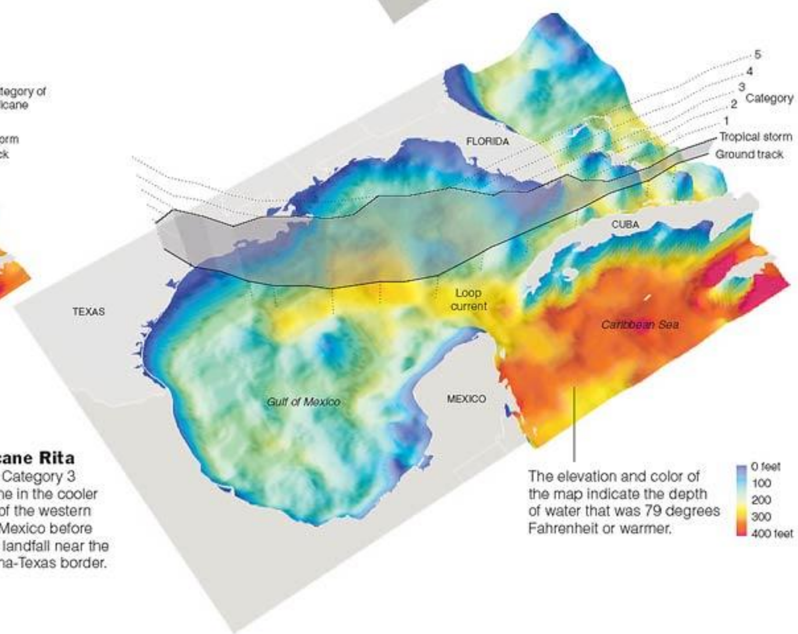
Loop current eddies are rings of warm water that occasionally break off from the loop current. The eddies can be more than 100 miles across and can persist for months, rotating clockwise as they move slowly westward.

Hurricane Katrina strengthened to a Category 5 hurricane on the Saffir-Simpson scale as it passed over a large eddy cast off from the loop current.

The gray ribbon follows the hurricane's path; the size of the ribbon indicates the intensity of the storm.



Hurricane Rita fell to a Category 3 hurricane in the cooler waters of the western Gulf of Mexico before making landfall near the Louisiana-Texas border.

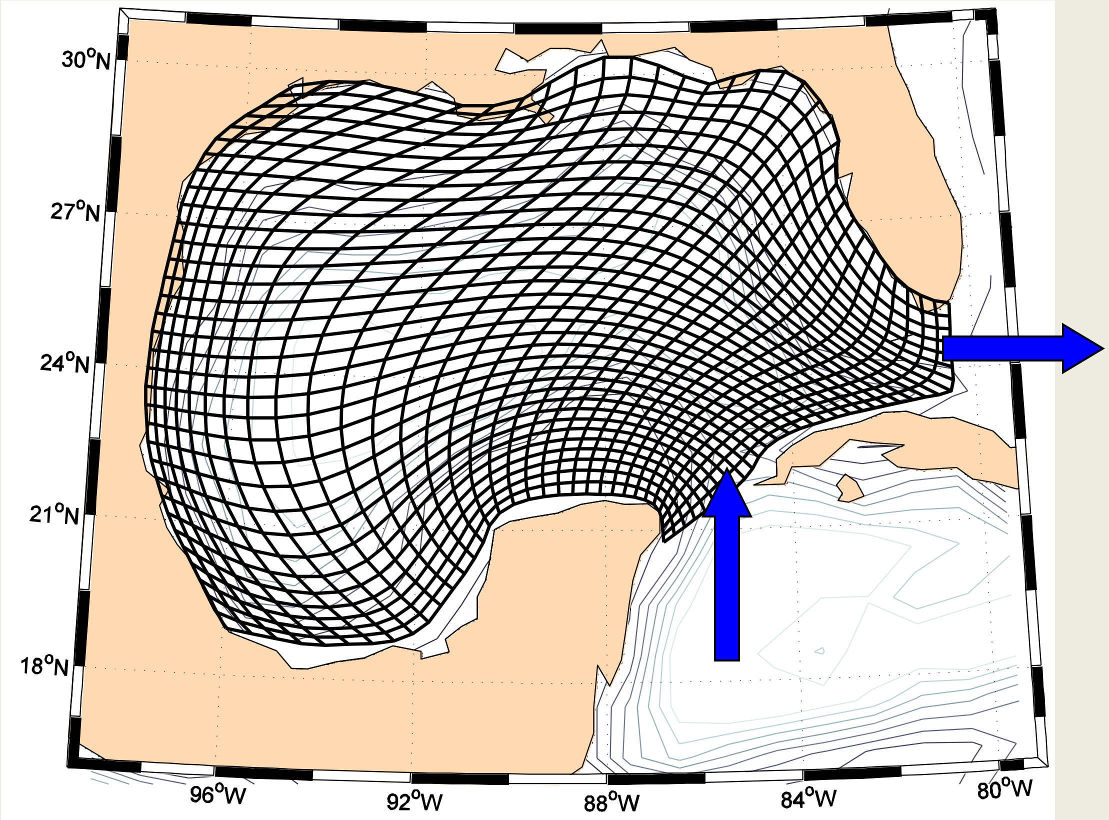


The elevation and color of the map indicate the depth of water that was 79 degrees Fahrenheit or warmer.

Sources: NOAA/Atlantic Oceanographic and Meteorological Laboratory; NASA; "Ocean Circulation in the Gulf of Mexico," W. Sturges and A. L. Fernandez, editors

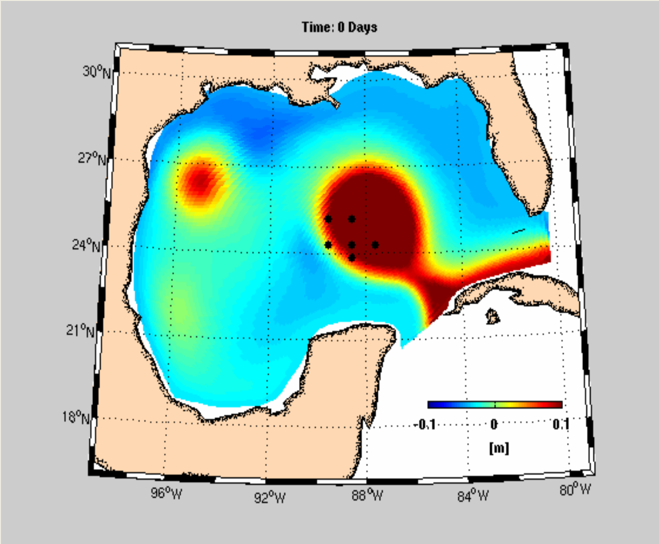
Gulf of Mexico

- 3 active layer, reduced gravity
- Modeling of the loop current in the GoM
- Limited area model
- 12-20 km

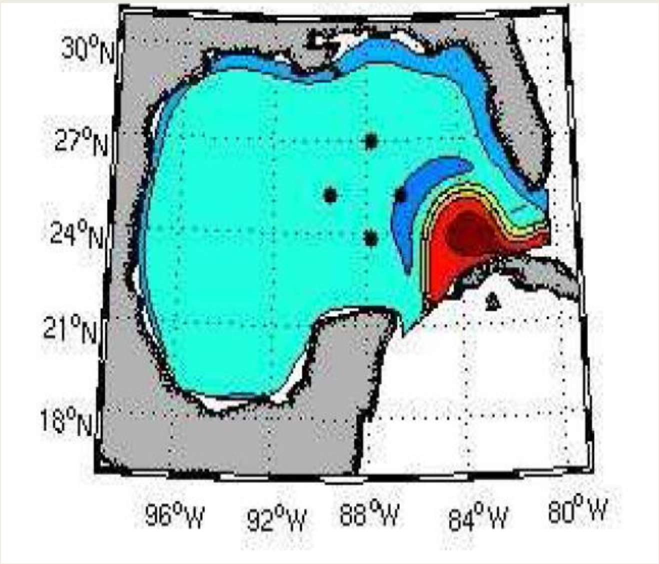


Identical twin experiment

TRUTH



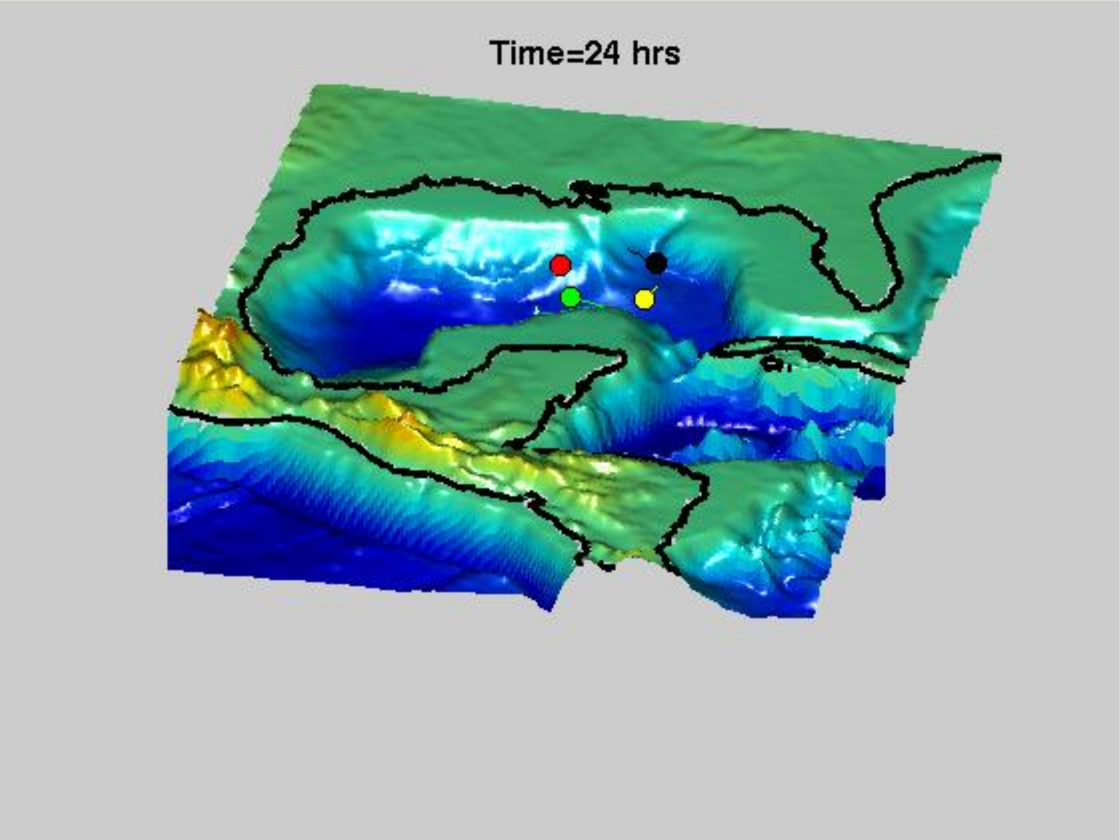
ESTIMATE



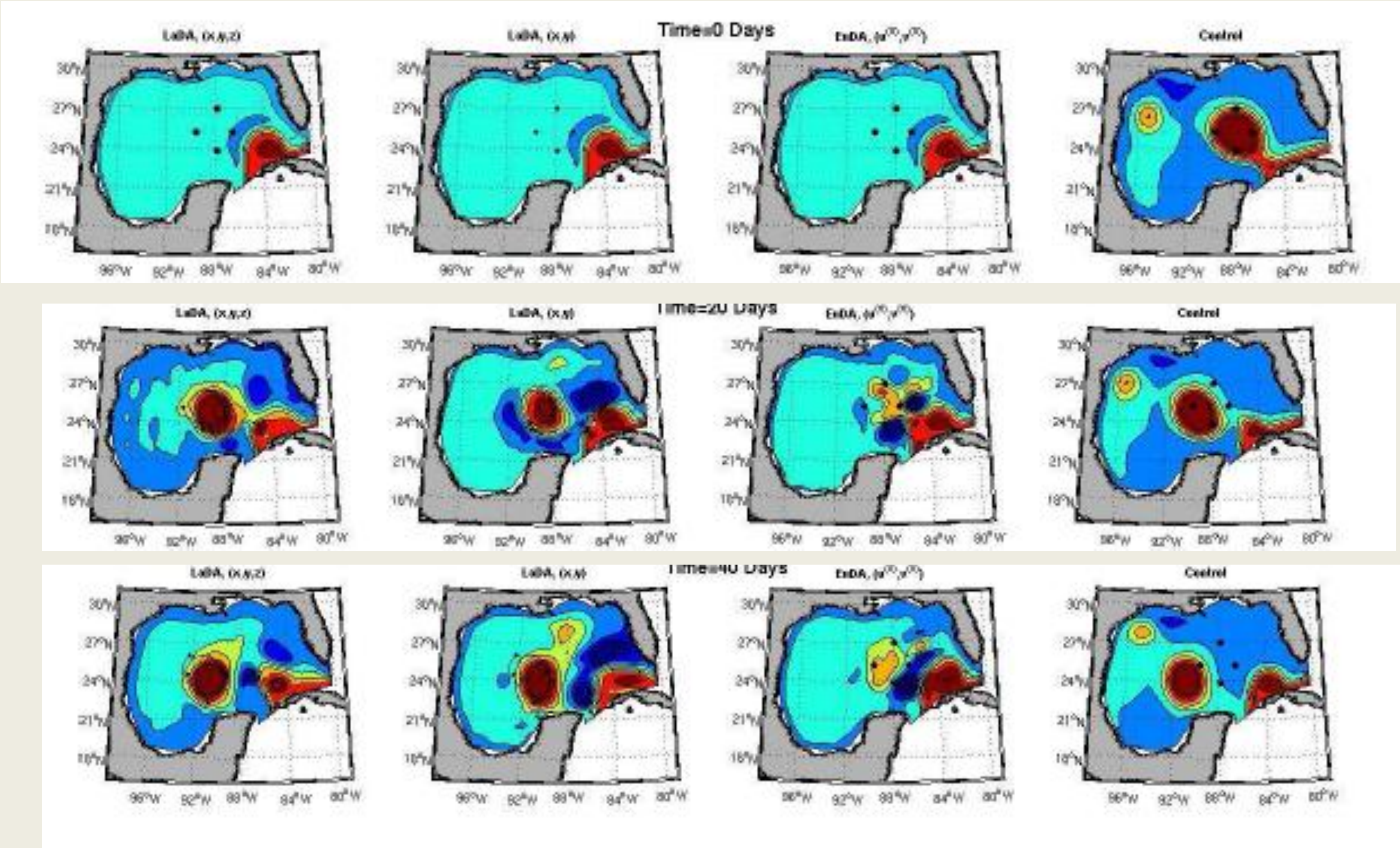
Synthetic observations:

- **Fixed stations(u,v)**
- **Surface drifters (x,y)**
- **Isopycnal floats (x,y,z)**

Recapturing the eddy



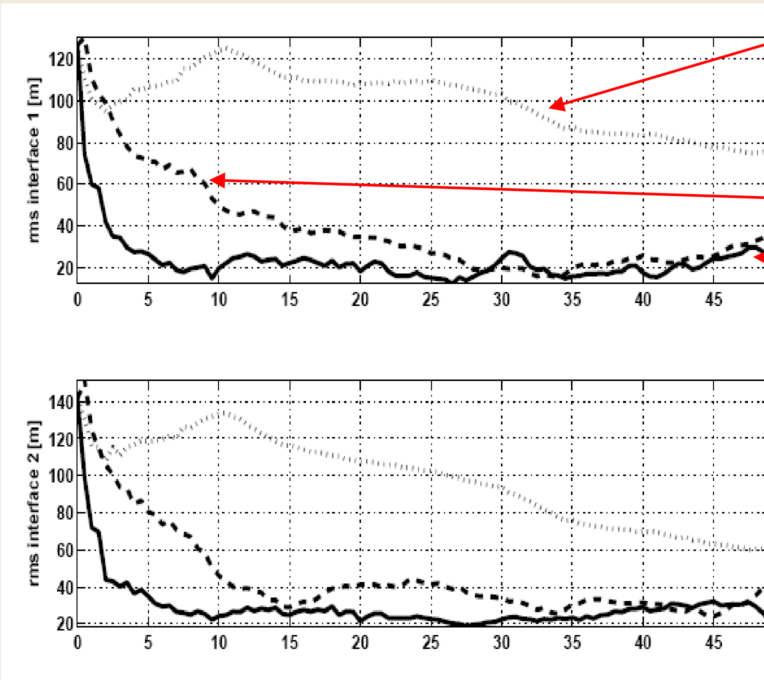
Eddies in GoM



Work with Guillaume Vernieres (NASA) and Kayo Ide (MD)

Results: rms(truth-analysis) of interface's depths

Interface between layer 1 and 2



EuDA (u,v)

LaDA (x,y)

LaDA (x,y,z)

Interface between layer 2 and 3

Techniques of Data Assimilation

Deterministic techniques

- Kalman filter
- Ensemble Kalman filter
- Variational methods (3DVAR, 4DVAR)

Requirements:

1. Gaussian
2. Close to linear

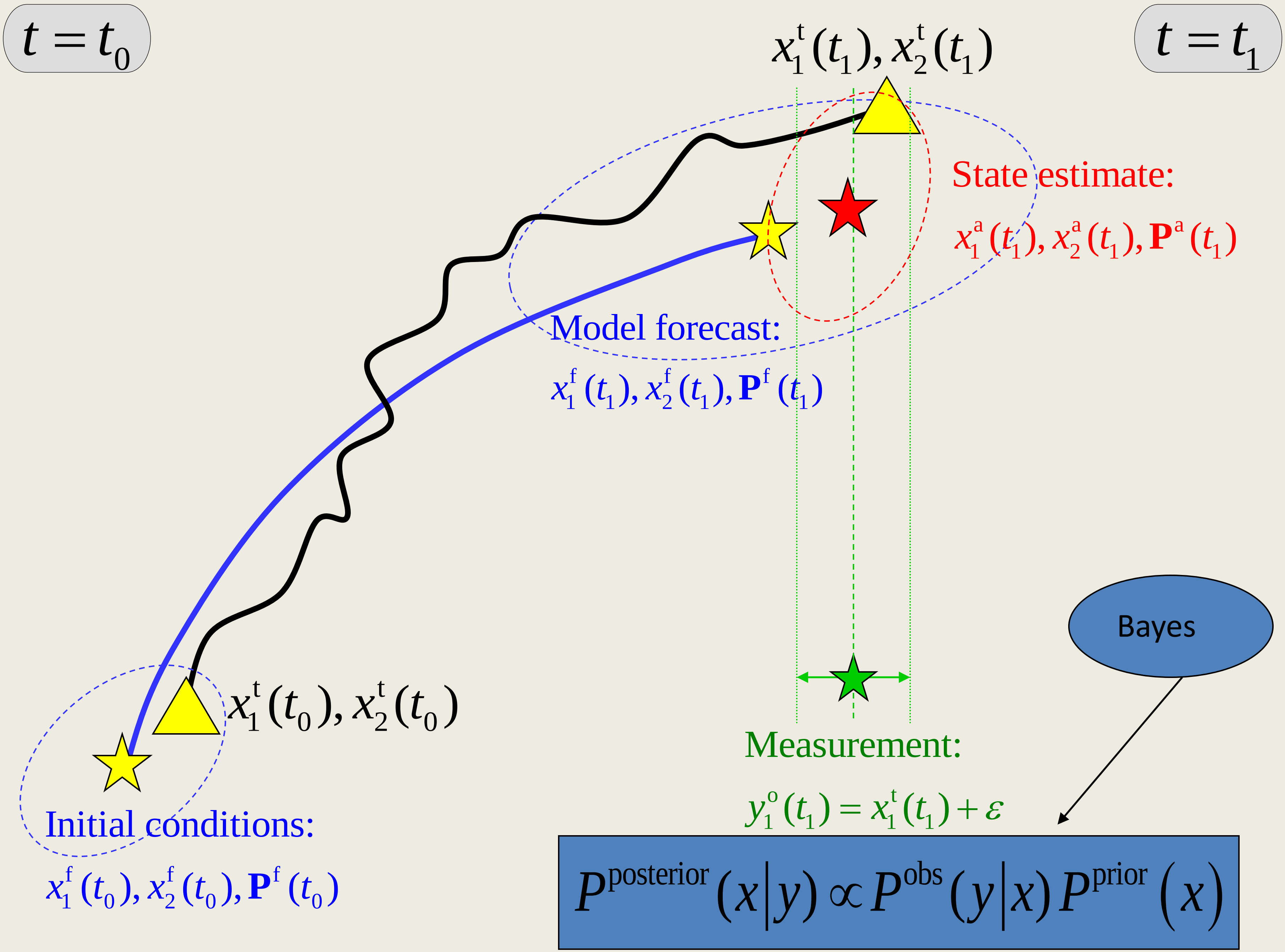
Statistical techniques

- Particle filtering
- Hybrid Monte-Carlo
- Metropolis-Hastings
- Langevin sampling

Requirement:

Low dimension

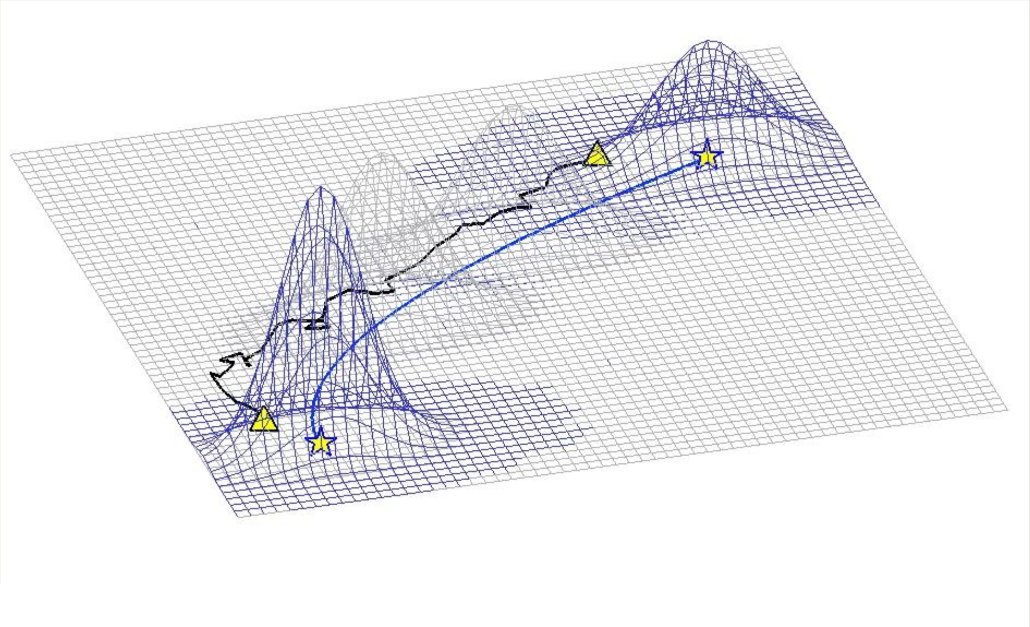
NONLINEARITY vs. DIMENSION

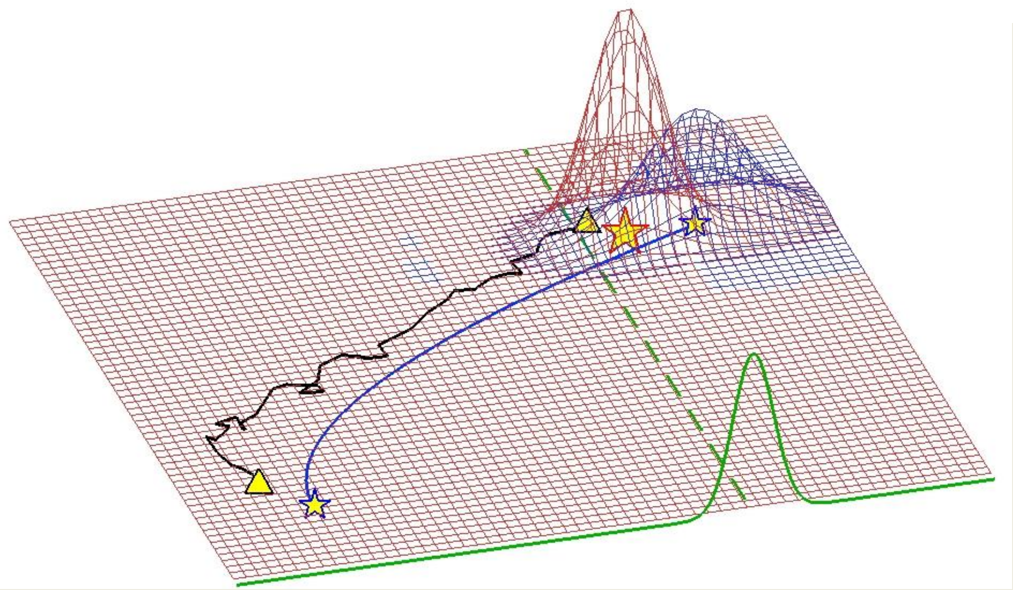

$t = t_0$
$t = t_1$
$x_1^{\mathrm{t}}(t_1), x_2^{\mathrm{t}}(t_1)$
State estimate:
$x_1^{\mathrm{a}}(t_1), x_2^{\mathrm{a}}(t_1), \mathbf{P}^{\mathrm{a}}(t_1)$
Model forecast:
$x_1^{\mathrm{f}}(t_1), x_2^{\mathrm{f}}(t_1), \mathbf{P}^{\mathrm{f}}(t_1)$
$x_1^{\mathrm{t}}(t_0), x_2^{\mathrm{t}}(t_0)$
Bayes
Measurement:
$y_1^{\mathrm{o}}(t_1) = x_1^{\mathrm{t}}(t_1) + \varepsilon$
Initial conditions:
$x_1^{\mathrm{f}}(t_0), x_2^{\mathrm{f}}(t_0), \mathbf{P}^{\mathrm{f}}(t_0)$
$P^{\mathrm{posterior}}(x|y) \propto P^{\mathrm{obs}}(y|x)\, P^{\mathrm{prior}}(x)$

Forecast step:

$$p(\mathbf{x}, t_0) \rightarrow p(\mathbf{x}, t_1)$$

$$\frac{\partial p}{\partial t} + \frac{\partial (M_i p)}{\partial x_i} = \frac{1}{2} \frac{\partial^2 (Q_{ij} p)}{\partial x_i \partial x_j}$$





Bayes step (update/analysis):

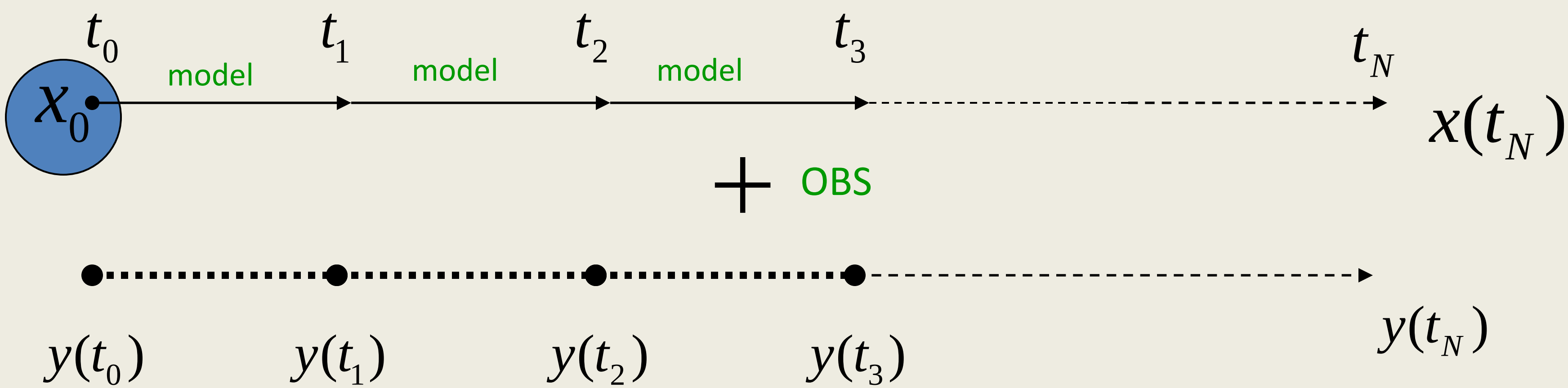
$$p(\mathbf{x}, t_1) \rightarrow p(\mathbf{x}, t_1 \mid \mathbf{y}^{\mathrm{o}})$$

$$p(\mathbf{x}, t_1 \mid \mathbf{y}^{\mathrm{o}}) = \frac{p(\mathbf{y}^{\mathrm{o}} \mid \mathbf{x})\, p(\mathbf{x}, t_1)}{\int p(\mathbf{y}^{\mathrm{o}} \mid \mathbf{z})\, p(\mathbf{z}, t_1) d\mathbf{z}}$$

But: computationally prohibitive, state $\approx 10^6$

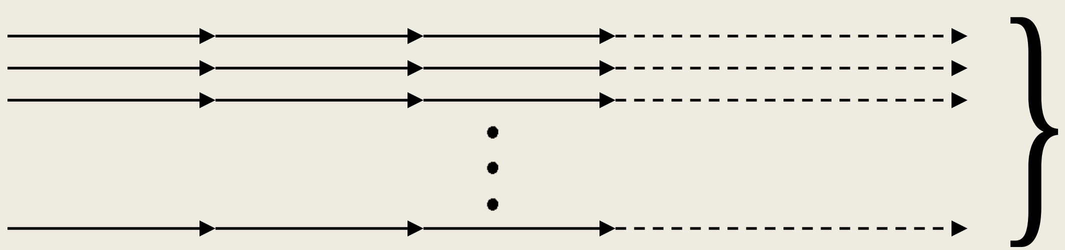
State Estimation

Model runs + observations $\longrightarrow$ state estimate



Bayes: $$P^{\text{posterior}}(x|y) = P^{\text{obs}}(y|x)\,P^{\text{prior}}(x)$$

sample posterior



Model runs

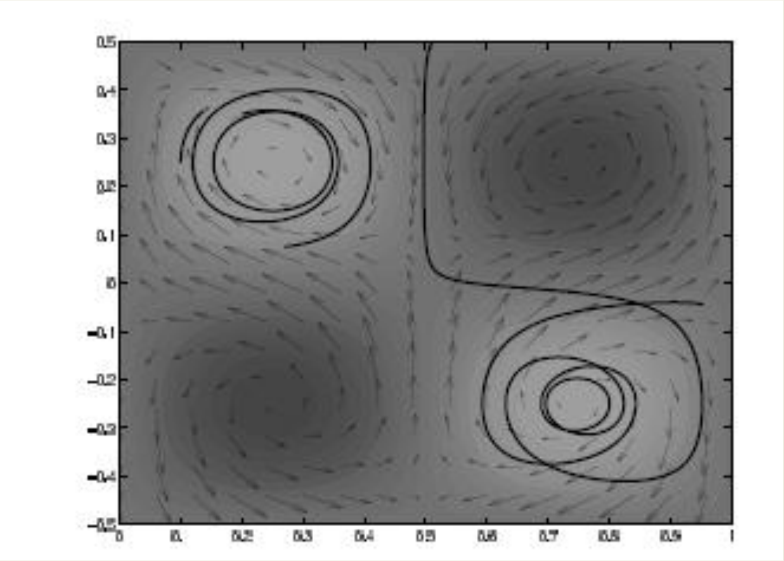
Perturbed Cellular Flow Field

$$\frac{\partial u}{\partial t} = v - \frac{\partial h}{\partial x},$$
$$\frac{\partial v}{\partial t} = -u - \frac{\partial h}{\partial y},$$
$$\frac{\partial h}{\partial t} = -\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y},$$

$$u(x,y,t) = -2\pi l \sin(2\pi kx)\cos(2\pi ly)u_0 + \cos(2\pi my)u_1(t),$$
$$v(x,y,t) = 2\pi k \cos(2\pi kx)\sin(2\pi ly)u_0 + \cos(2\pi my)v_1(t),$$
$$h(x,y,t) = \sin(2\pi kx)\sin(2\pi ly)u_0 + \sin(2\pi my)h_1(t),$$

$$\dot{u}_0 = 0,$$
$$\dot{u}_1 = v_1,$$
$$\dot{v}_1 = -u_1 - 2\pi m h_1,$$
$$\dot{h}_1 = 2\pi m v_1,$$

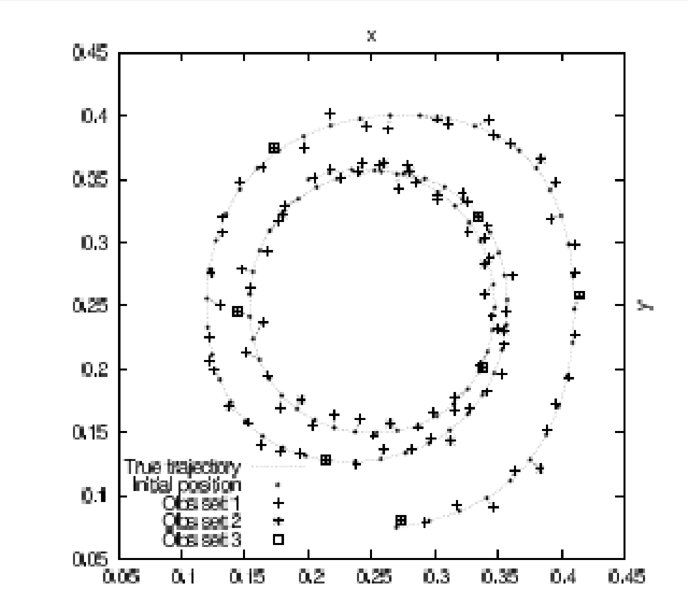
$$\dot{x} = u(x,y,t)$$
$$\dot{y} = v(x,y,t)$$

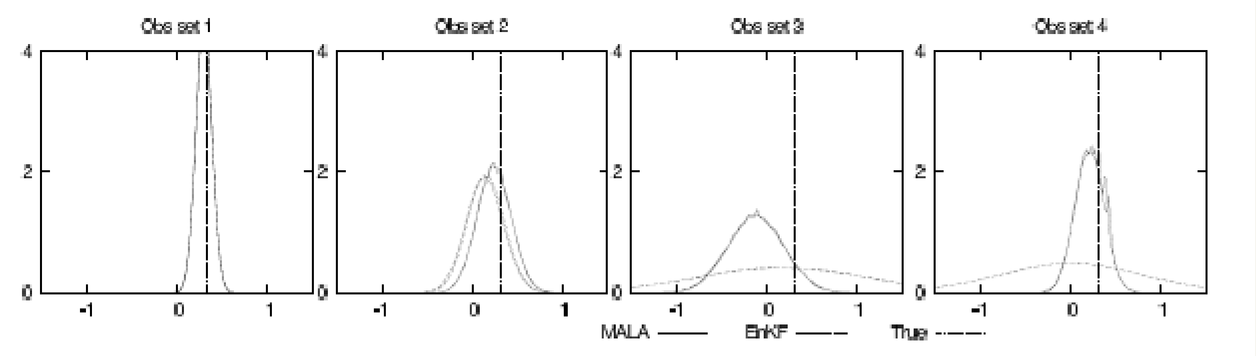


Apte, Stuart and J., Tellus A 2008

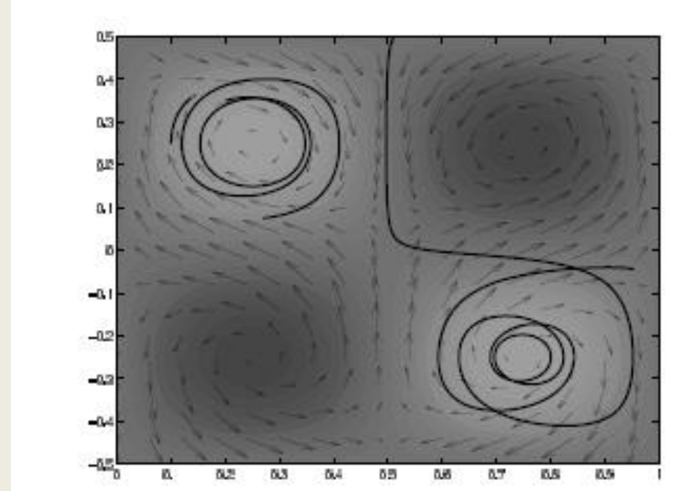
Assimilating from trajectory staying in one cell

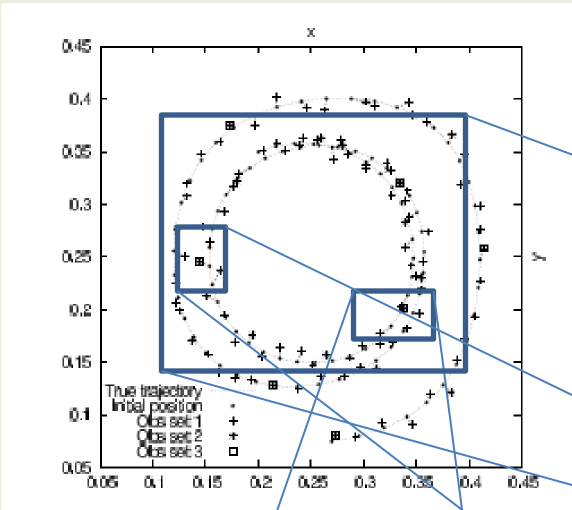
Compare:
- EnKF
- Metropolis-Hastings

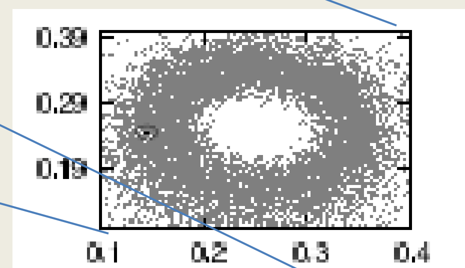




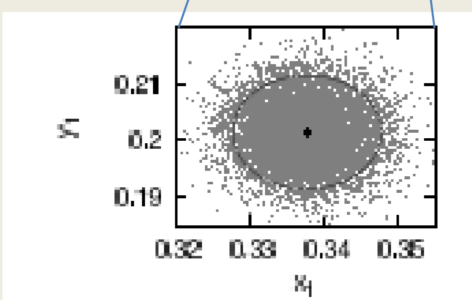
Problem with EnKF



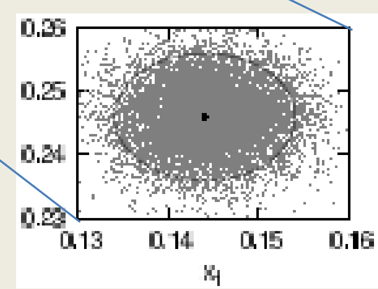




Before second observation



After first observation and assimilation



After SECOND observation and assimilation

Conclusions

Data and models

- Need balance in use of data and models
- Bayesian perspective provides framework
- Increasing amounts of data and model output should be exploited, but smartly!

Math and DA

- Can hope to filter effectively in low dimensions
- Do NOT avoid nonlinearity, use it as it is high in information content
- Seek data that has LOW dimension but HIGH information content