Overview of Approaches to Data Assimilation

Chris Jones

University of North Carolina at Chapel Hill

TIFR CAM, Mathematical and Statistical Foundations of Data Assimilation, Bangalore, India, July 4, 2011

- DATA: ever-improving experimental technology has led to vast amounts of accumulated data
- MODEL: ever-increasing computational capacity has led to greater model capability and output
- Scientific imperative: bring data and computations together to work in harmony to enhance prediction, state estimation and improve models

Coming to India

Urban legends
Solifugae are the subject of many urban legends and exaggerations about their size, speed, behaviour, appetite, and lethality. Members of this order of Arachnida apparently have no venom, with the possible exception of one species in India (*Rhagodes nigrocinctus*) as suggested in one study,[7] and do not spin webs.

Due to their bizarre appearance many people are startled by or even afraid of them. The greatest threat they pose to humans, however, is their defensive bite when handled. There is essentially no chance of death directly caused by the bite, but, due to the strong muscles of their chelicerae, they can produce a large, ragged wound that is prone to infection.

wind scorpions—camel spiders

Conversation with Amit Apte

CJ: Amit, I have been reading about wind scorpions in India and am pretty scared now about coming over to Bangalore.

AA: I have looked them up on Wikipedia and there is nothing to fear.

CJ: But it says there is one in India that may be venomous! Let's do this: could you try to estimate the number of wind scorpions in Bangalore in July? If it is less than 100 per sq km then I'll come.

Simple system of 2 variables:

$$x_1 = \text{ \# male wind scorpions in Bangalore}$$

$$x_2 = \text{ \# female wind scorpions in Bangalore}$$

MODEL: $$\left(x_1(t_{i+1}), x_2(t_{i+1})\right) = M\left(x_1(t_i), x_2(t_i), q\right)$$

$t = t_0$ $\quad$ $t = t_1$ $\quad$ $\bullet\bullet\bullet$ $\quad$ $t = t_N$

OBS: $$y(t_i) = H\left(x_1(t_i), x_2(t_i)\right)$$

MODEL: $$\left(x_1(t_{i+1}), x_2(t_{i+1})\right) = M\left(x_1(t_i), x_2(t_i), q\right)$$

- Reproduction of solifugae
- Life cycle
- Environmental factors
- Interaction with other species
- Availability of food
- Form model and set parameters

Note: all kinds of uncertainty…

MODEL ERROR

OBS: $$y(t_i) = H\left(x_1(t_i), x_2(t_i)\right)$$

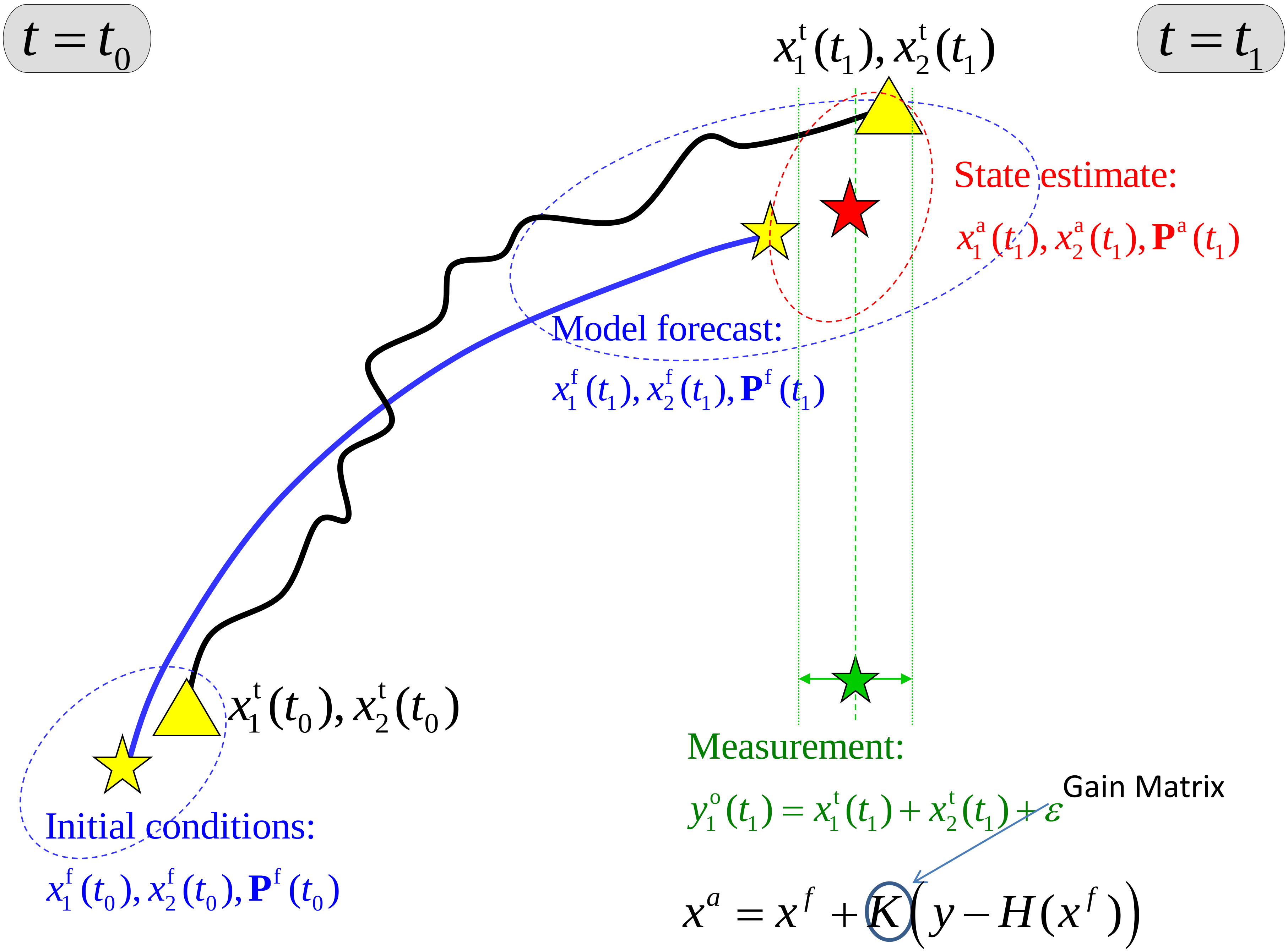
1. Observe only total # of wind scorpions

$$y(t_i) = x_1(t_i) + x_2(t_i)$$

2. Observe in restricted region
3. Extrapolate to Bangalore

Note: all kinds of uncertainty…

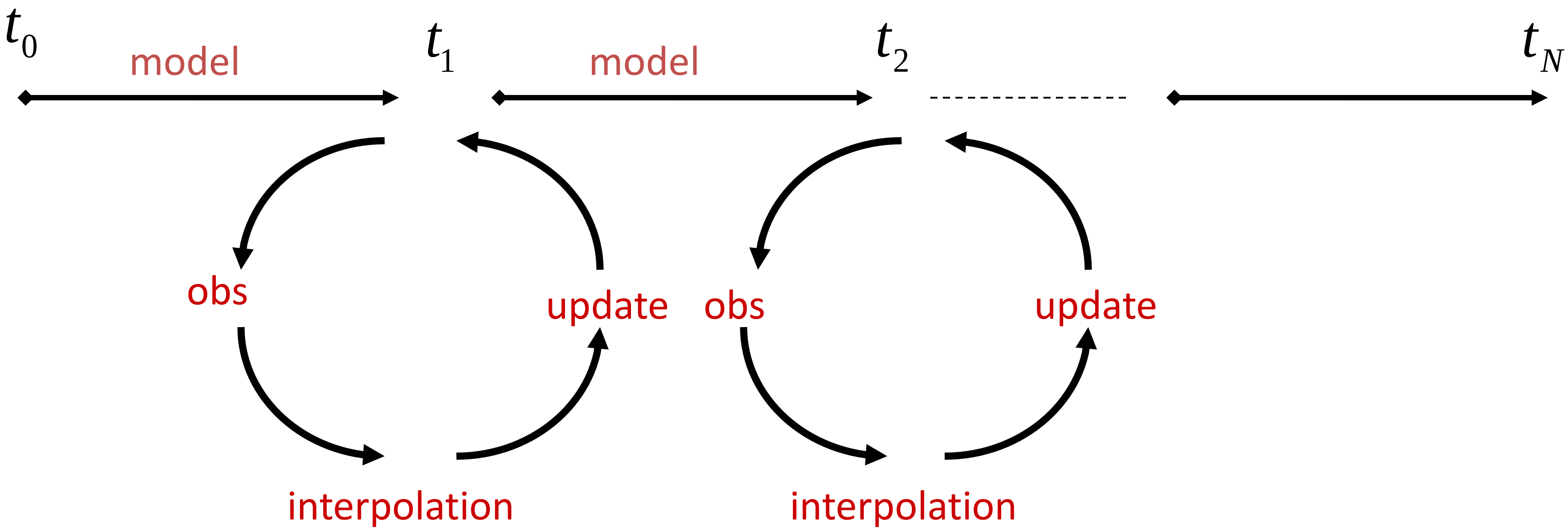
OBSERVATIONAL ERROR



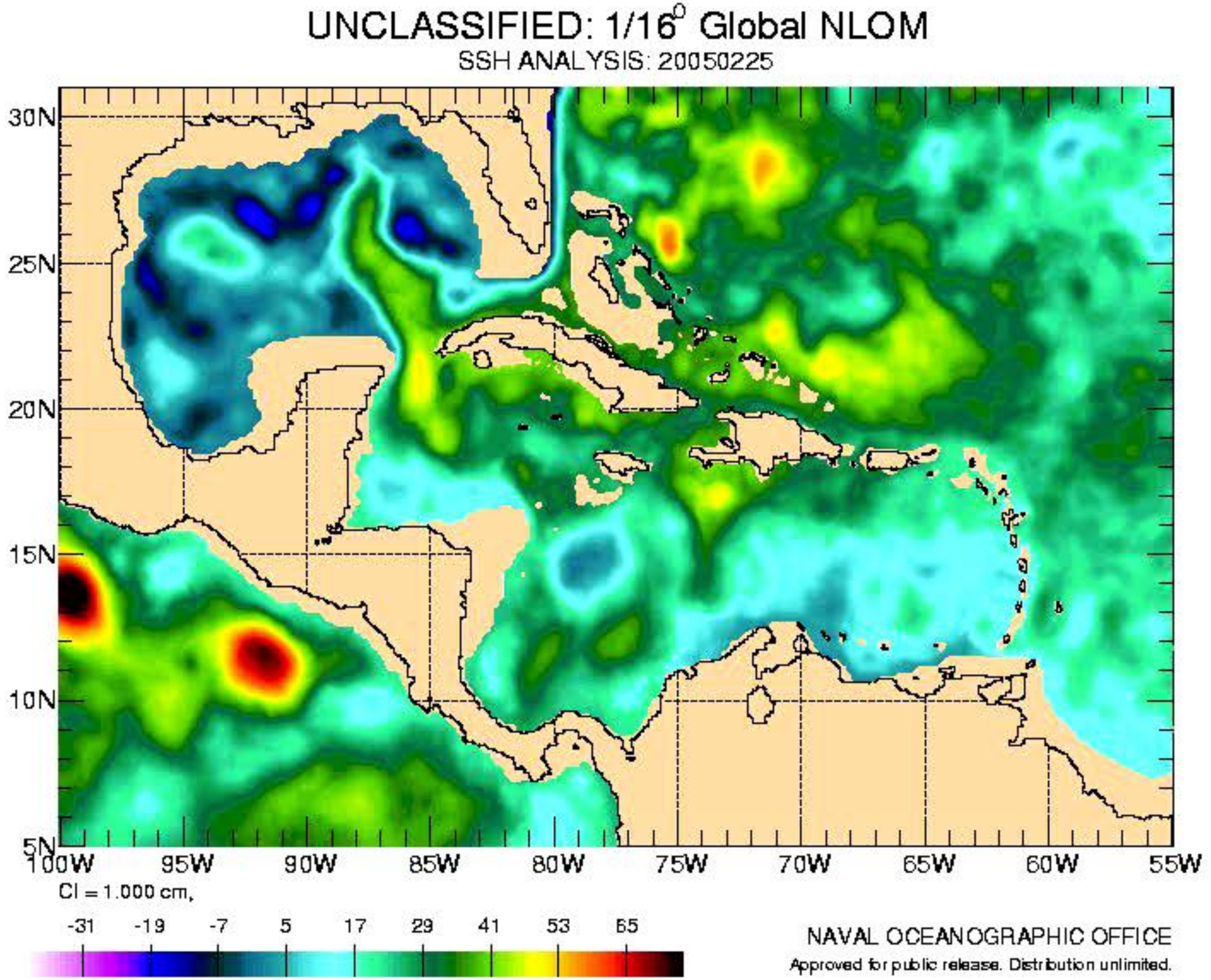
$t = t_0$
$t = t_1$
$x_1^t(t_1), x_2^t(t_1)$
State estimate:
$x_1^a(t_1), x_2^a(t_1), \mathbf{P}^a(t_1)$
Model forecast:
$x_1^f(t_1), x_2^f(t_1), \mathbf{P}^f(t_1)$
$x_1^t(t_0), x_2^t(t_0)$
Measurement:
$y_1^o(t_1) = x_1^t(t_1) + x_2^t(t_1) + \varepsilon$
Gain Matrix
Initial conditions:
$x_1^f(t_0), x_2^f(t_0), \mathbf{P}^f(t_0)$
$x^a = x^f + K\left(y - H(x^f)\right)$

Sequential Data Assimilation

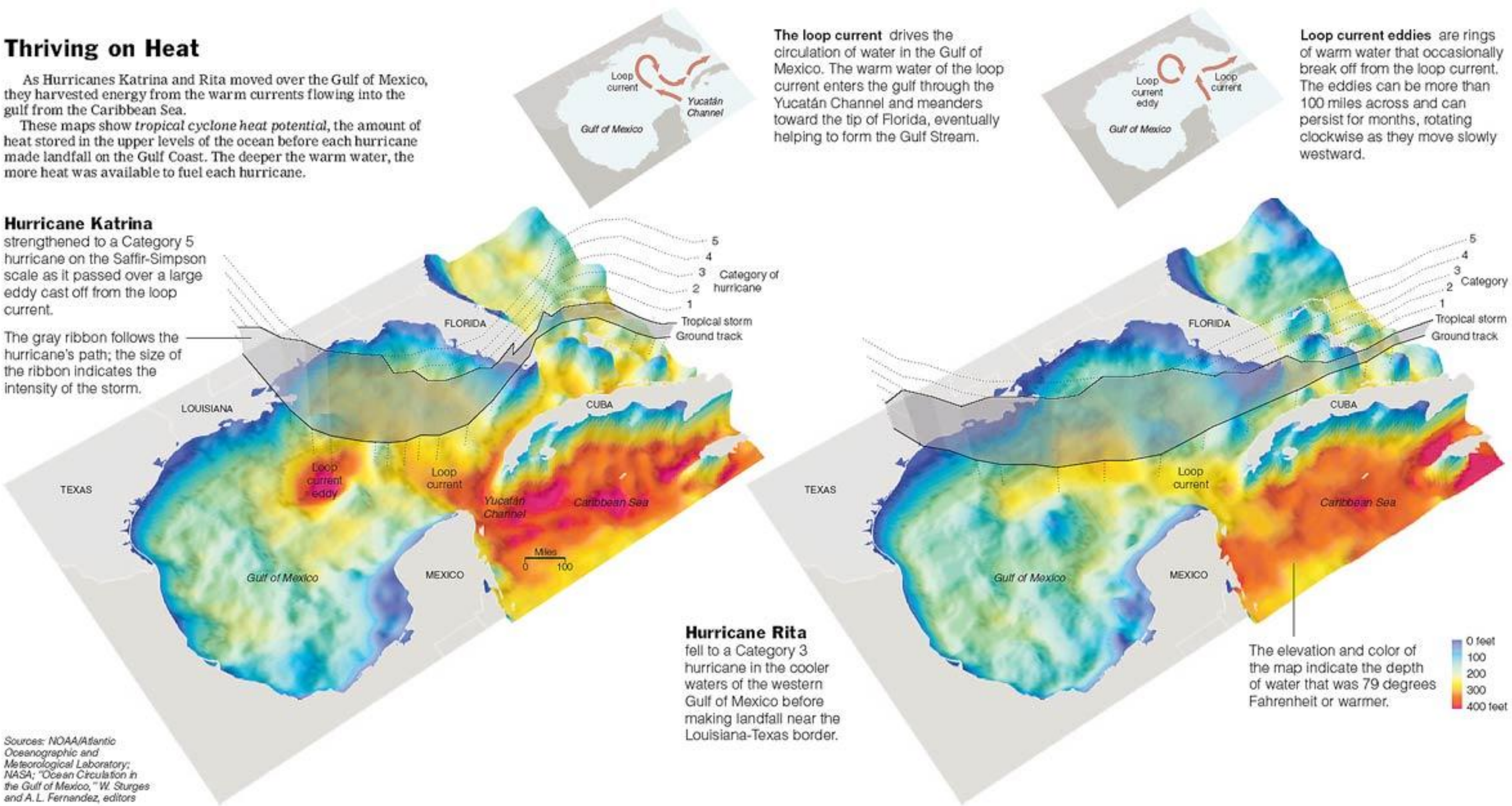
Model + observations $\longrightarrow$ prediction



Gulf of Mexico/Carribean



Hurricanes in the Gulf



Ocean model:

$\mathbf{x} \in \mathbf{R}^N$ – state vector comprising all relevant dynamical variables

(e.g. flow velocity, temperature, salinity, etc. at each grid point)

$d\mathbf{x}^{\mathrm{f}} = M(\mathbf{x}^{\mathrm{f}}, t)dt$ prognostic model

$d\mathbf{x}^{\mathrm{t}} = M(\mathbf{x}^{\mathrm{t}}, t)dt + \boldsymbol{\eta}(t)dt$ actual evolution

$E[\boldsymbol{\eta}(t)\boldsymbol{\eta}^T(t')] = \delta(t-t')\mathbf{Q}(t)$ covariance of the model residual

Observations:

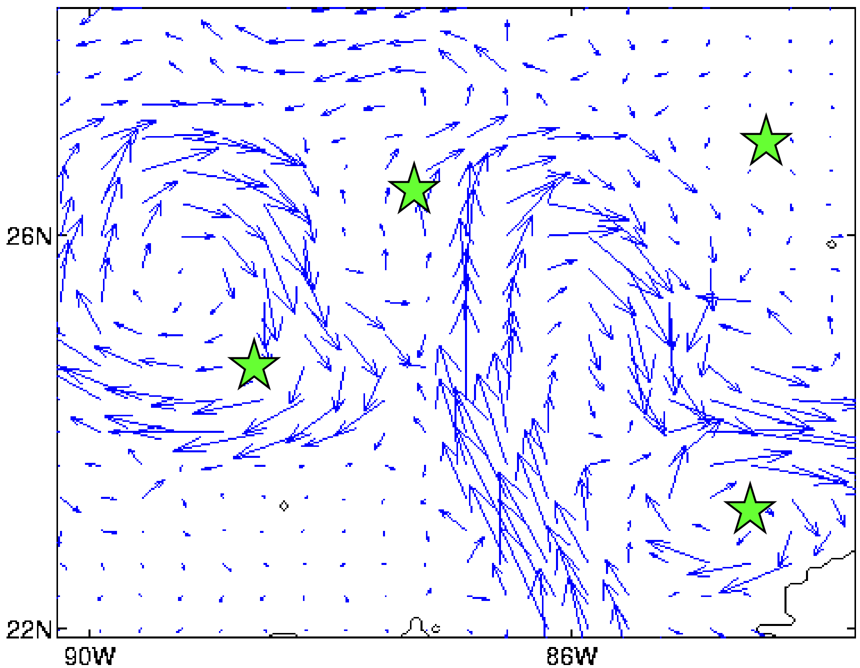
$\mathrm{y}_i^{\mathrm{o}} = H_i[\mathbf{x}_i^{\mathrm{t}}] + \boldsymbol{\varepsilon}_i$

H_i – observation operator

$\boldsymbol{\varepsilon}_i$ – observation error

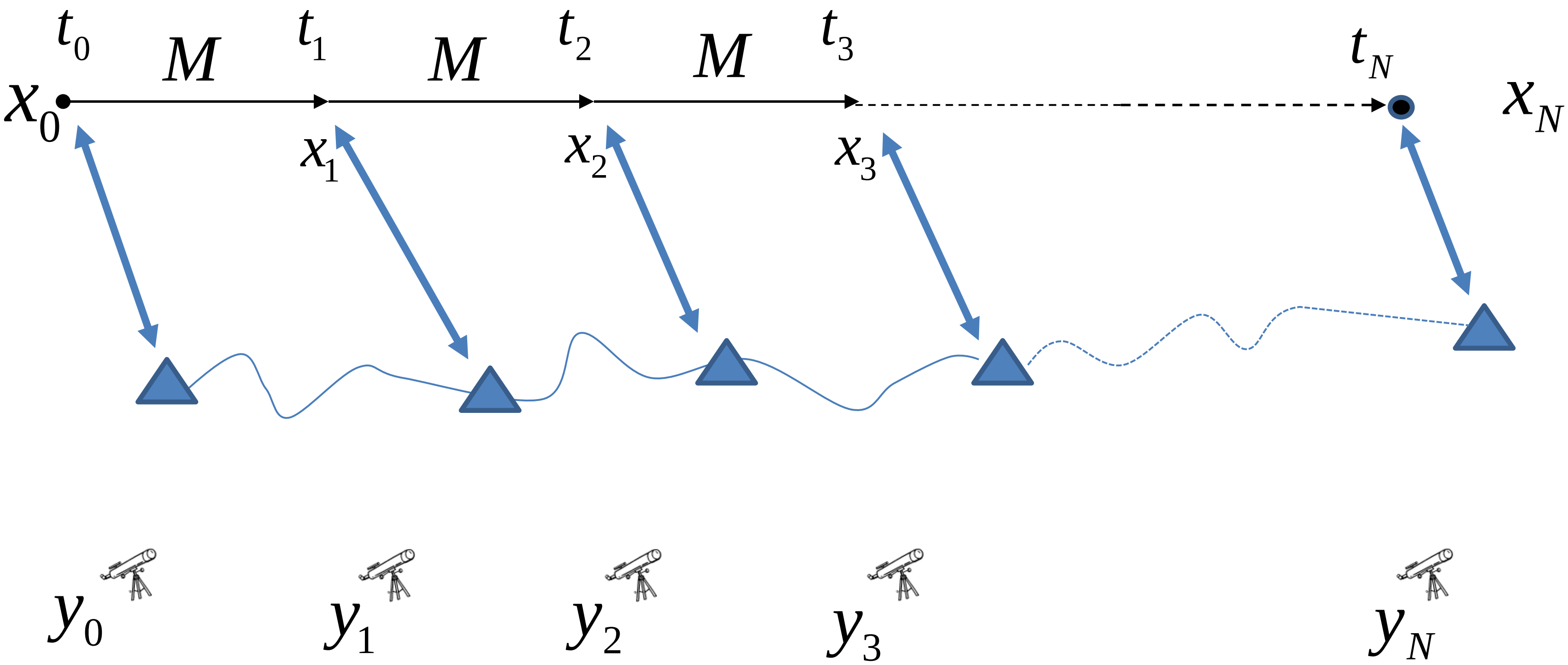
$\mathrm{y}_i^{\mathrm{o}} \in \mathbf{R}^L$, typically $L \ll N$

$E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_m^T] = \delta_{im}\mathbf{R}_i$ – observation error covariance



$$x_{i+1} = M(x_i, q)$$

$x_i \in \mathbb{R}^n \quad n \sim 10^6$ MODEL



$y_i \in \mathbb{R}^m$

$$H : \mathbb{R}^n \to \mathbb{R}^m$$

$(m \ll n)$ OBS

Typical problems:

1. Estimate state at current time
2. Estimate initial condition
3. Estimate parameters

1. Sequential DA (predictive mode-filtering)

Minimize the cost-function:

$$J(x) = \left\langle x - x_i, (P_i)^{-1} (x - x_i) \right\rangle + \left\langle y_i - H(x), R^{-1}(y_i - H(x)) \right\rangle$$

P_t = background error covariance

R = observational error covariance

2. Variational DA (reanalysis mode-smoothing)

Minimize the cost-function:

$$J(x) = \left\langle x - x_0^*, (B)^{-1} (x - x_0^*) \right\rangle + \sum_{j=1}^{N} \left\langle y_j - H(x), R_j^{-1}(y_j - H(x)) \right\rangle$$

B = background error covariance

R_j = jth observational error covariance

x_0^* = initial (initial condition) estimate

Interpolation and Gain Matrix

If H is linear (linearized) then cost-function is quadratic

$$J(x) = \left\langle x - x_i, \left(P_i\right)^{-1}\left(x - x_i\right)\right\rangle + \left\langle y_i - H(x), R^{-1}(y_i - H(x))\right\rangle$$

Solution is found by interpolation:

$$x_i^{\,a} = x_i + K\left(y_i - H(x_i)\right)$$

Gain Matrix:

$$K = P_i H^T \left(H P_i H^T + R\right)^{-1}$$

To Linearize or not to Linearize?

$$x_i^{\,a} = x_i^{\,f} + K\left(y_i - H(x_i^{\,f})\right) \qquad K = P_i H^T \left(H P_i H^T + R\right)^{-1}$$

- P_i constant background error covariance matrix (Optimal Interpolation=OI)
- If model is linear, P_i is evolved under model (Kalman Filter=KF)
- If model is nonlinear, P_i is evolved under tangent linear model (Extended Kalman Filter=EKF)
- P_i is built out of ensembles evolved under full nonlinear model (Ensemble Kalman Filter=EnKF)

For large models, linearization is enacted at some level

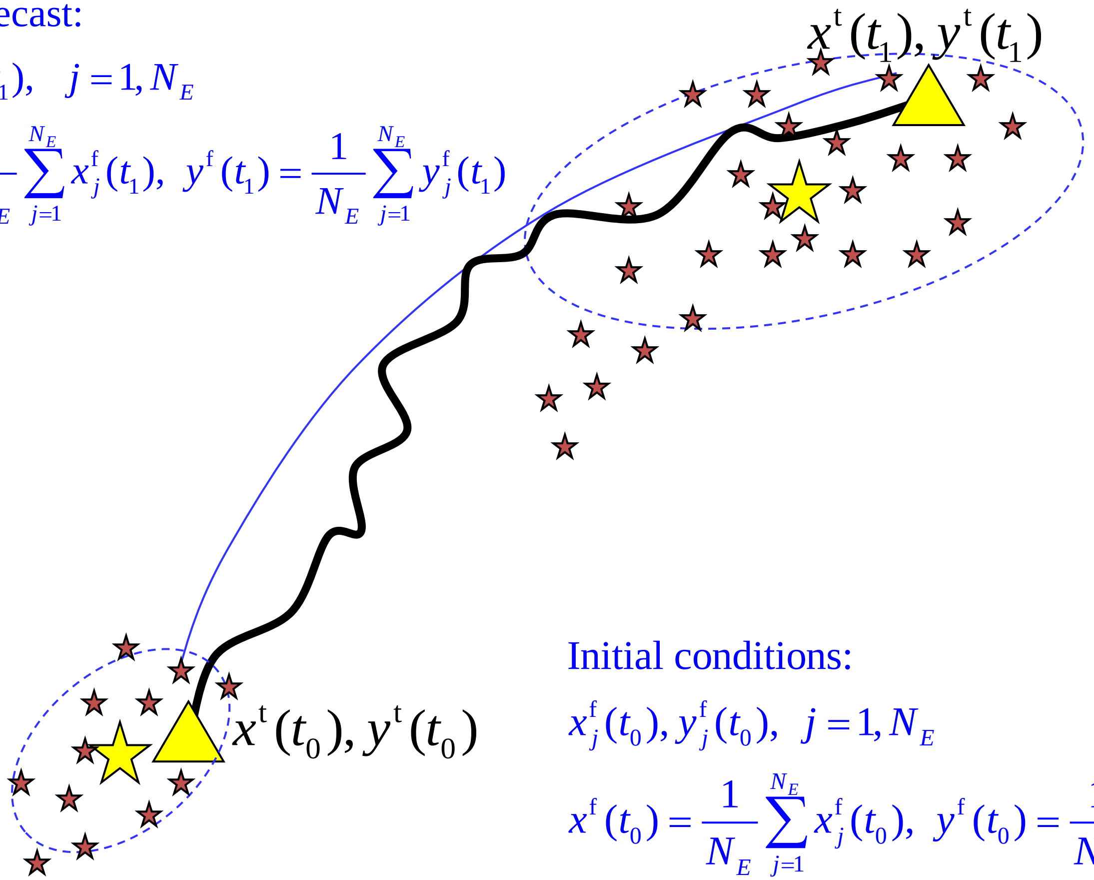
Ensemble Kalman Filter (EnKF)

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

Model forecast:

$x_j^{\mathrm{f}}(t_1), y_j^{\mathrm{f}}(t_1), \quad j = 1, N_E$

$$x^{\mathrm{f}}(t_1) = \frac{1}{N_E}\sum_{j=1}^{N_E} x_j^{\mathrm{f}}(t_1), \;\; y^{\mathrm{f}}(t_1) = \frac{1}{N_E}\sum_{j=1}^{N_E} y_j^{\mathrm{f}}(t_1)$$



Initial conditions:

$x_j^{\mathrm{f}}(t_0), y_j^{\mathrm{f}}(t_0), \;\; j = 1, N_E$

$$x^{\mathrm{f}}(t_0) = \frac{1}{N_E}\sum_{j=1}^{N_E} x_j^{\mathrm{f}}(t_0), \;\; y^{\mathrm{f}}(t_0) = \frac{1}{N_E}\sum_{j=1}^{N_E} y_j^{\mathrm{f}}(t_0)$$

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:

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

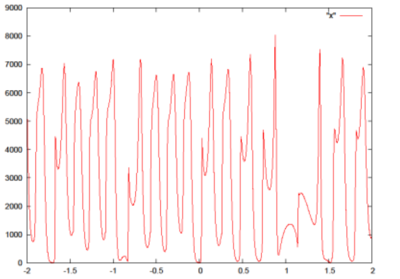
What if linearizing throws away too much information?

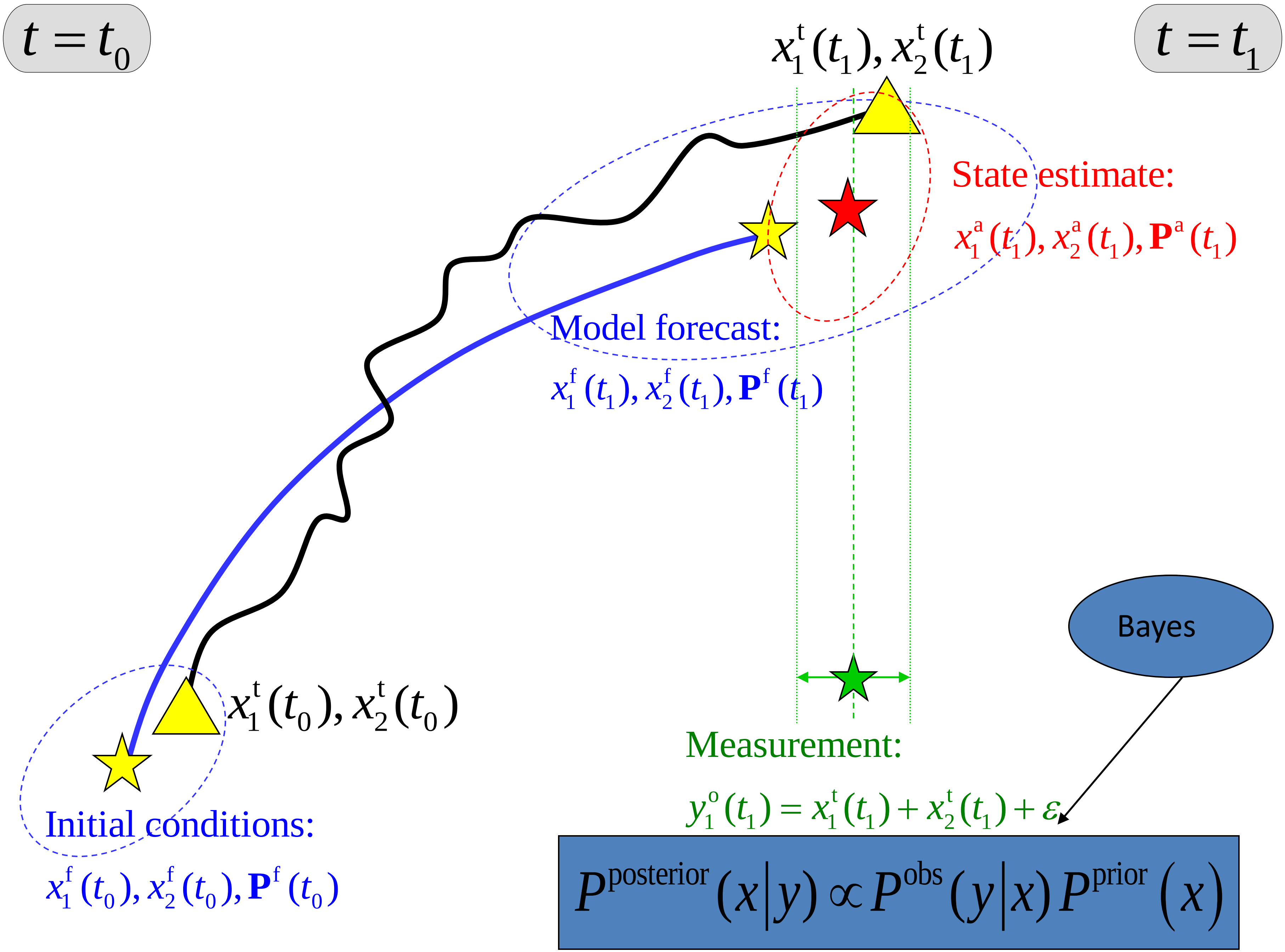
For instance, the obs operator *H* may be nonlinear

$$J(x) = \left\langle x - x_i, (P_i)^{-1}(x - x_i) \right\rangle + \left\langle y_i - H(x), R^{-1}(y_i - H(x)) \right\rangle$$

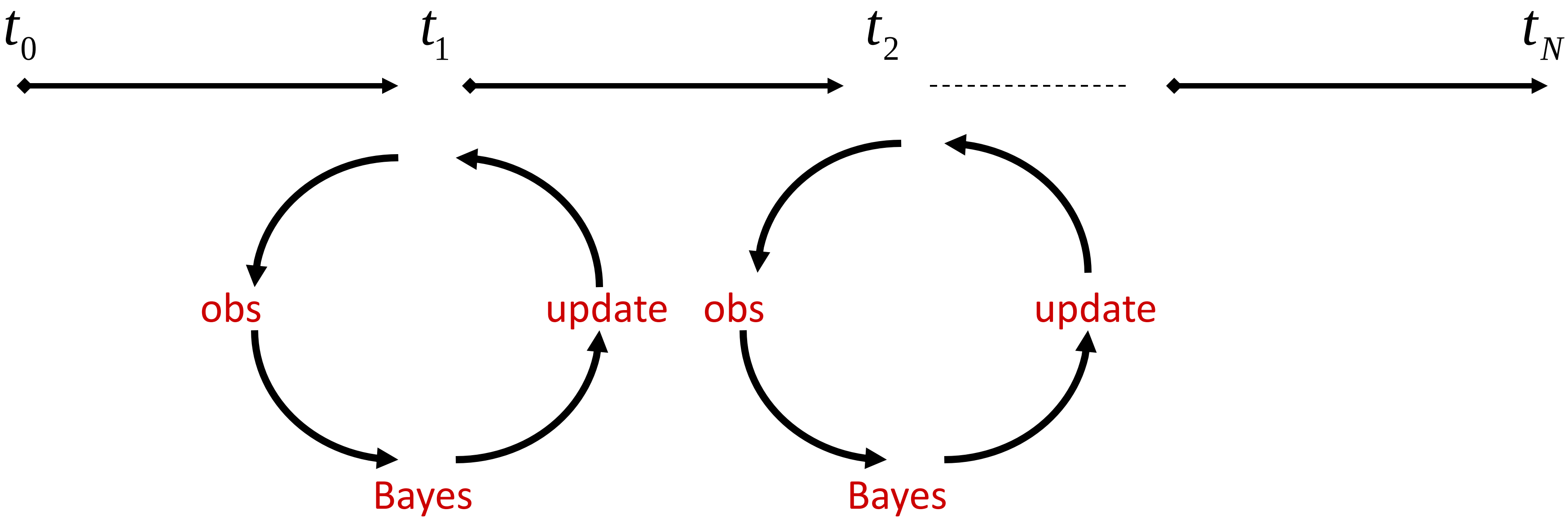
Then *J* is not necessarily quadratic

1. We now need to find a global minimizer in the presence of possibly many local minimizers
2. Even if we do find a global minimizer, is that what we want?




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

Bayesian View of Sequential DA



$x =$ state

$y =$ obs

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

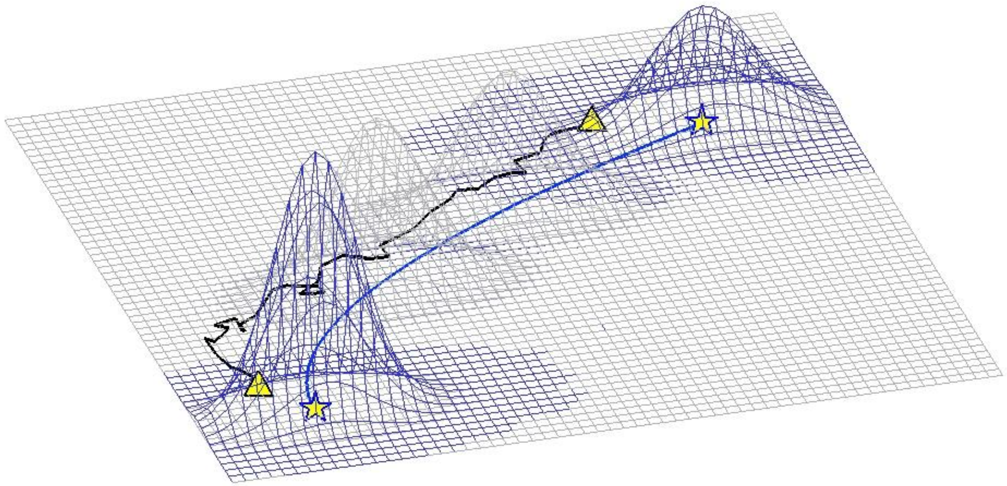
Key question: how do we obtain the distributions on RHS?

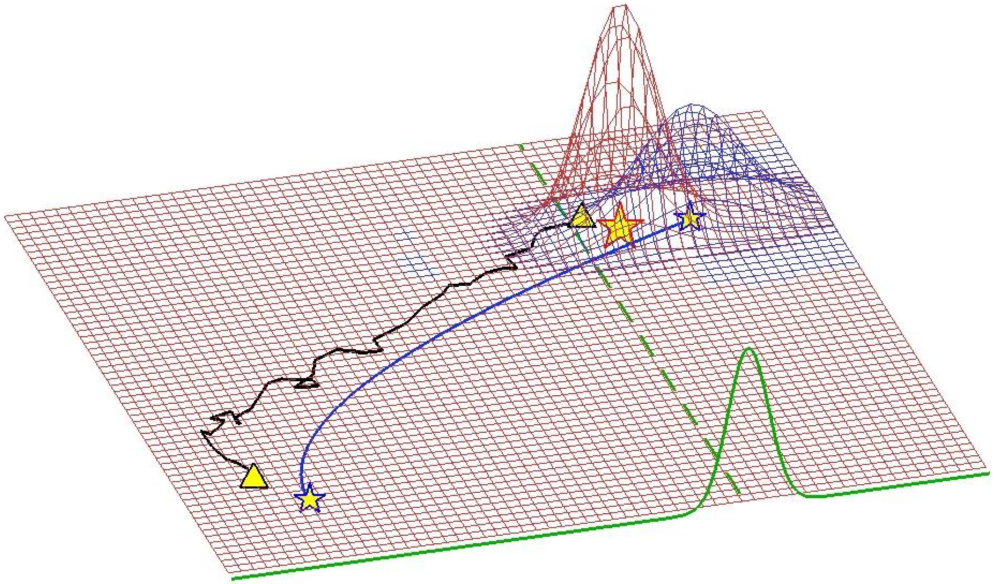
In principle: evolve pdf under Fokker-Planck eqn for model

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):

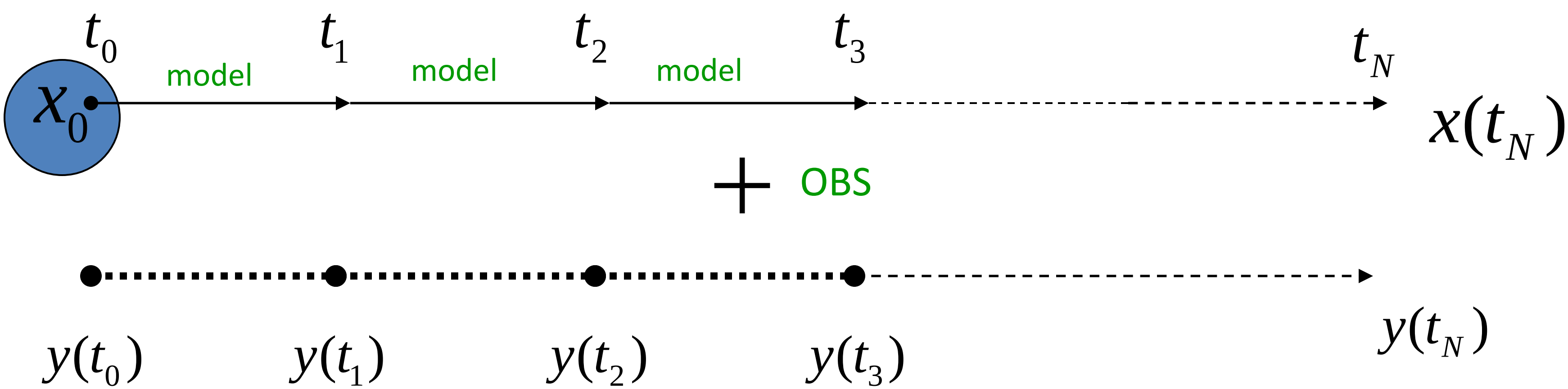
$$p(\mathbf{x}, t_1) \rightarrow p(\mathbf{x}, t_1 \mid \mathbf{y}^{\mathrm{o}})$$

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

But: computationally prohibitive ~ 10^6

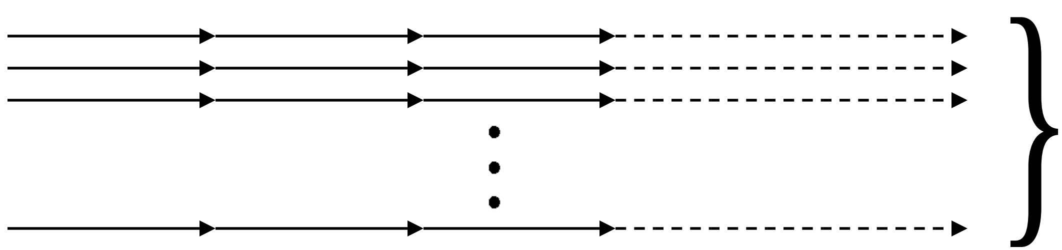
Variational DA

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

Bayes Theorem

If not linear(ized), there may be multiple minima for the cost-function! Is the global minimizer necessarily the desired answer?

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

$$P(x\,|\,y) \propto \exp(-J(x))$$

mode ←→ global min

$$J(x) = \left\langle x - x_0^*, (B)^{-1}\left(x - x_0^*\right)\right\rangle + \sum_{j=1}^{N} \left\langle y_j - H(x), R_j^{-1}(y_j - H(x))\right\rangle$$

See: Data Assimilation: Mathematical and Statistical Perspectives, Apte, J, Stuart and Voss, IJNMF 2008

Nonlinearity vs. Dimension

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

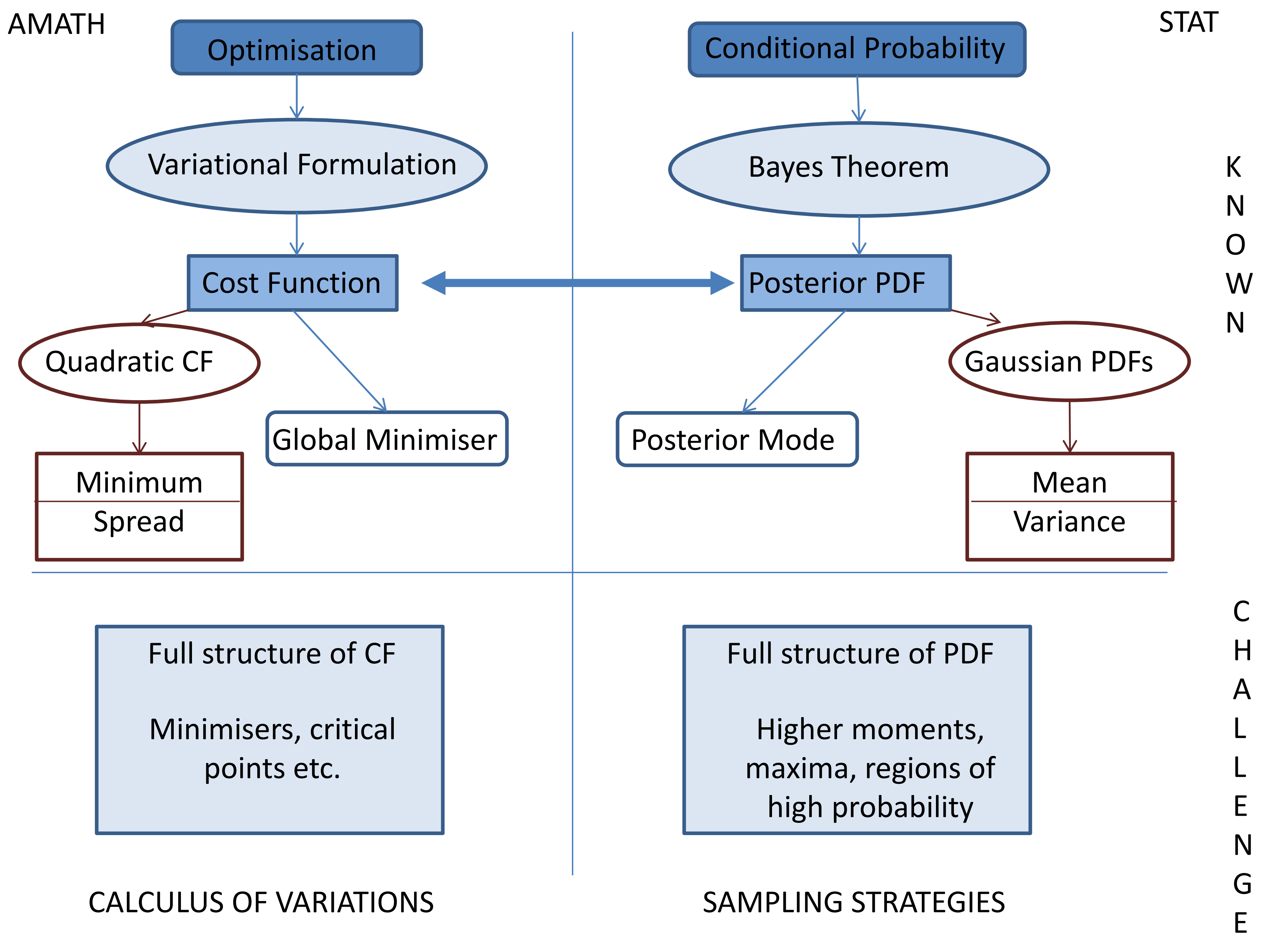
$$J(x) = \left\langle x - x_0^*, (B)^{-1}\left(x - x_0^*\right)\right\rangle + \sum_{j=1}^{N} \left\langle y_j - H(x^j), R_j^{-1}(y_j - H(x^j))\right\rangle$$

$$P(x \mid y) \propto \exp(-J(x))$$

Sampling strategies:

1. Particle filtering
2. Langevin sampling
3. Metropolis-Hastings
4. Importance sampling

But none are well developed for high-dimensional problems



AMATH
STAT
Optimisation
Variational Formulation
Cost Function
Quadratic CF
Global Minimiser
Minimum
Spread
Conditional Probability
Bayes Theorem
Posterior PDF
Gaussian PDFs
Posterior Mode
Mean
Variance
KNOWN
Full structure of CF
Minimisers, critical points etc.
Full structure of PDF
Higher moments, maxima, regions of high probability
CHALLENGE
CALCULUS OF VARIATIONS
SAMPLING STRATEGIES