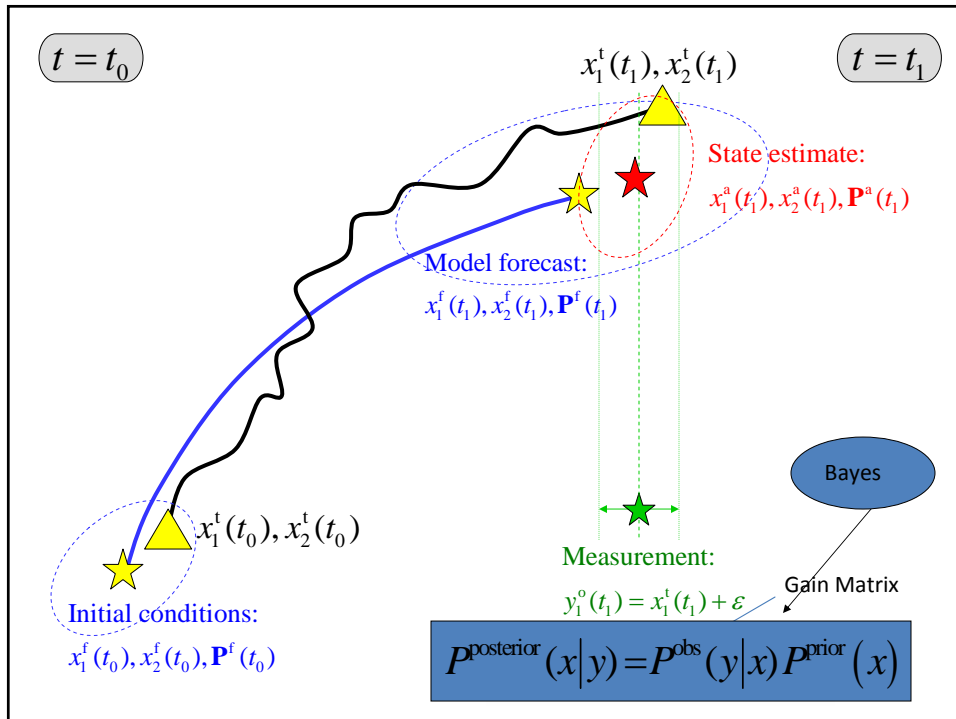


# A Bayesian Approach to Lagrangian Data Assimilation

Chris Jones, UNC-CH and University of Warwick  
 Andrew Stuart, Jochen Voss, University of Warwick  
 Amit Apte, TIFR Bangalore



Supported by the Office of Naval Research and the National Science Foundation



## Framework for DA Approach

$$\frac{dx}{dt} = f(x) \quad x \in \mathbb{R}^n$$

$$x(0) = x_0 \square \zeta$$

$p_\zeta(x_0)$  is the pdf of initial conditions

$$\text{obs: } t = t_1, \dots, t_K$$

$$y_k = h(x(t_k)) + \eta_k$$

## Bayesian formulation

Prior distribution:  $P^{\text{prior}}(x)$  from initial "initial condition"

Observational likelihood:  $P^{\text{obs}}(y|x)$  from (Lagrangian) data

$$P^{\text{obs}}(y|x) \propto \exp\left[-\frac{1}{2\sigma^2}(y - h(x))^2\right]$$

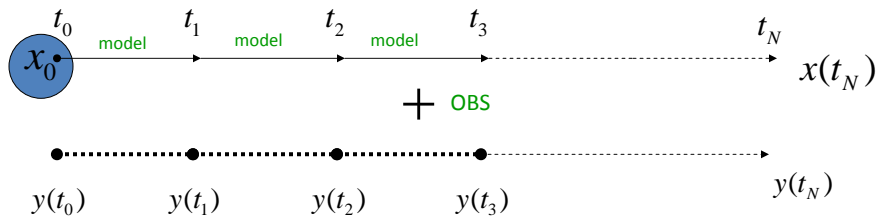
Bayes rule:  $P^{\text{posterior}}(x|y) \propto P^{\text{obs}}(y|x)P^{\text{prior}}(x)$

Ultimate Goal: Obtain the posterior distribution of initial conditions

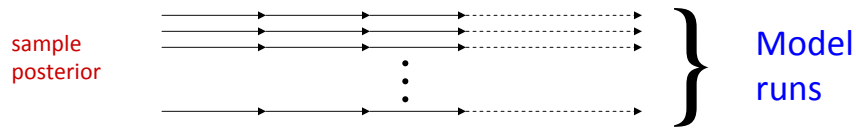
Assumption: Perfect model, but in principle unnecessary

# State Estimation

Model runs + observations  $\longrightarrow$  state estimate



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



# Langevin Sampling

Langevin dynamics :  $\frac{dZ}{ds} = L(Z) + \sqrt{2} \frac{dW}{ds}$

in discretized version  $Z_{k+1} = Z_k + L(Z_k) \Delta_s + \sqrt{2\Delta_s} W$

with  $L(Z) = -\nabla_z \log \rho(Z)$

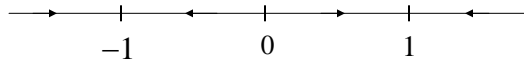
generates  $N$  samples  $\{Z_k\}_{k=1}^N$  from the distribution  $\rho(Z)$

Basis: Invariant distribution of the Langevin dynamics is  $\rho(Z)$

It works provided the Langevin equation is ergodic.

## Test case

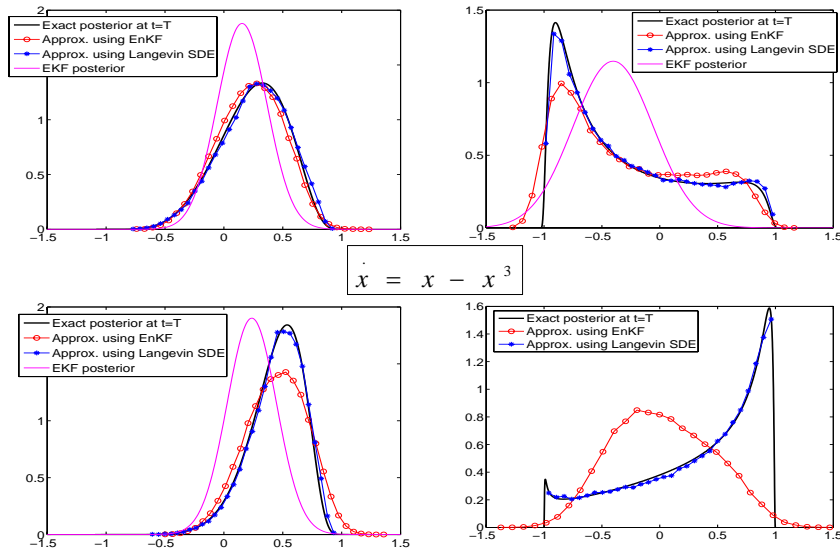
Deterministic model:  $\dot{x} = x - x^3$



Estimation problem: Initial condition  $x_0$

Observations:  $y_i = x(i\Delta) + \eta_i \quad \eta_i \sim N(0, R^2)$   
 $i = 1, \dots, N; \quad N\Delta = T$

## Comparing EnKF and Langevin sampling



## Lagrangian DA

- In ocean, subsurface info is often (quasi-) Lagrangian
- State estimation (as opposed to forecasting) is of interest in ocean
- Particularly appropriate for float data
- Natural to use an augmented state-space approach
- Obs are in a clearly defined low-dimensional subspace but encode key aspects of full dynamics
- Can potentially capture large-scale coherent features

## Lagrangian DA and State Estimation

Augmented model :  $\frac{dx_F}{dt} = M_F(x_F, t) ; \quad \frac{dx_D}{dt} = M_D(x_D, x_F, t)$   
with initial conditions  $x_0 \equiv (x_F(0), x_D(0))$

Observations :  $y_i = x_D(t_i) + \xi_i ; \quad \xi_i \sim N(0, R^2)$   
at times  $(t_1, t_2, \dots, t_m)$

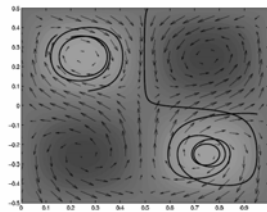
**Goal :** Estimate initial conditions  $x(0)$  using the observations

**Idea :** Use Bayesian formulation and Langevin sampling

# Model Problem for Lagrangian DA

Linearized shallow water model:

$$\begin{aligned}\frac{\partial u}{\partial t} &= v - \frac{\partial h}{\partial x}, \\ \frac{\partial v}{\partial t} &= -u - \frac{\partial h}{\partial y}, \\ \frac{\partial h}{\partial t} &= -\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y},\end{aligned}$$



2-mode approximation:

Geostrophic mode with amplitude  $u_0$   
Inertial-gravity mode that is time periodic

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

$$v(x, y, t) = 2\pi k \cos(2\pi kx) \sin(2\pi ly) u_0 + \cos(2\pi my) v_1(t),$$

$$h(x, y, t) = \sin(2\pi kx) \sin(2\pi ly) u_0 + \sin(2\pi my) h_1(t),$$

$$k = l = m = 1$$

# Augmented System

$$\dot{u}_0 = 0,$$

$$\dot{u}_1 = v_1,$$

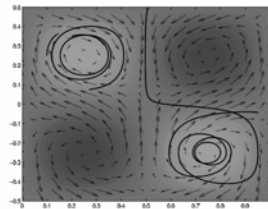
$$\dot{v}_1 = -u_1 - 2\pi m h_1,$$

$$\dot{h}_1 = 2\pi m v_1,$$

$$\dot{x}_i(t) = u(x_i(t), y_i(t), t),$$

$$\dot{y}_i(t) = v(x_i(t), y_i(t), t),$$

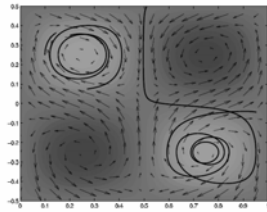
$$i = 1, \dots, M,$$



obs at:  $t_k = k\delta \quad k = 1, \dots, N$   
with Gaussian errors uncorrelated and independent of each other

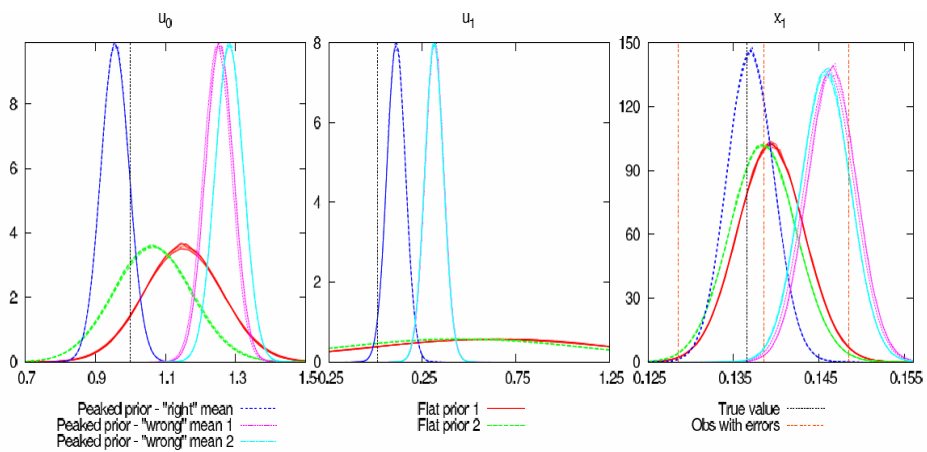
# Experiments and Methods

1. Short trajectory
2. Long trajectory staying in cell
3. Trajectory crossing cell boundaries

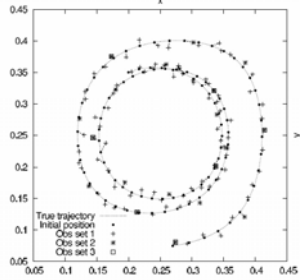


- A. Langevin Stochastic DE
- B. Metropolis Adjusted Langevin Algorithm
- C. Random Walk Metropolis Hastings
- D. EnKF

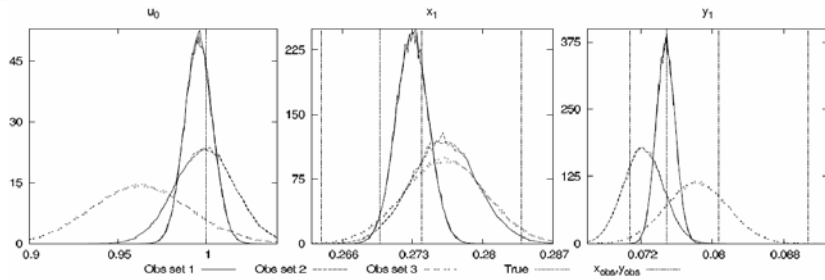
## Short Trajectory



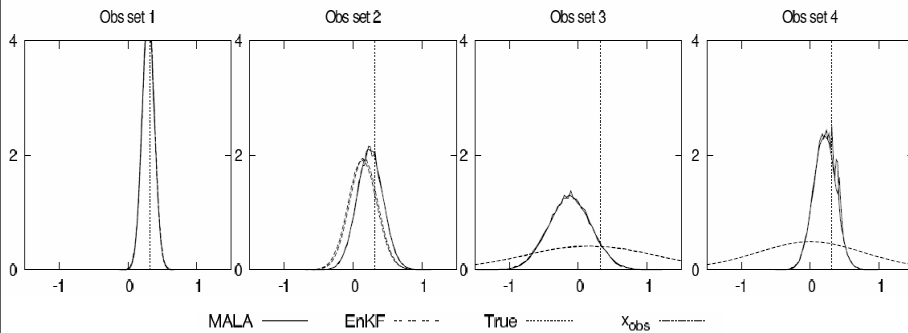
# Long Trajectory in Cell



3 observations sets, # of obs:  
 Obs set 1: 100  
 Obs set 2: 20  
 Obs set 3: 6



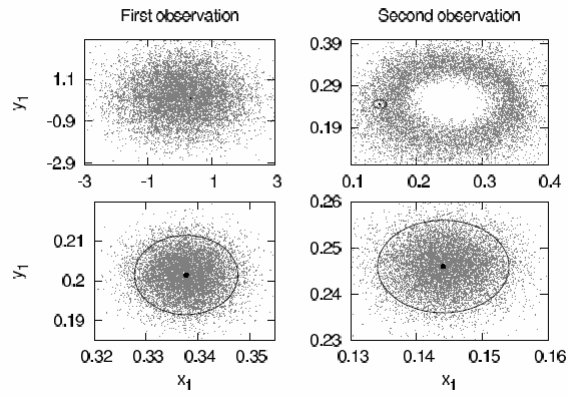
# Comparison



MALA improves with increased number of observations  
 (frequency kept same) but EnKF does not.

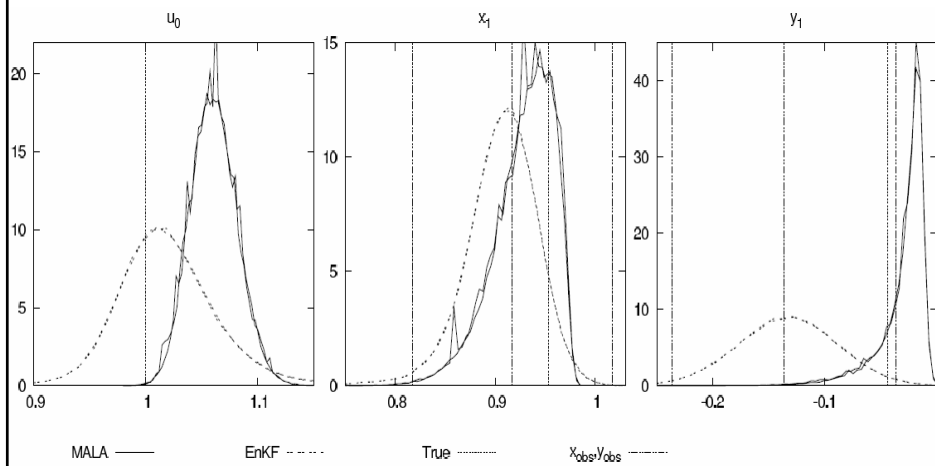
Obs set 4, has same  
 frequency as 3, but  
 extends trajectory  
 and makes 20 obs

# Scatter Plots

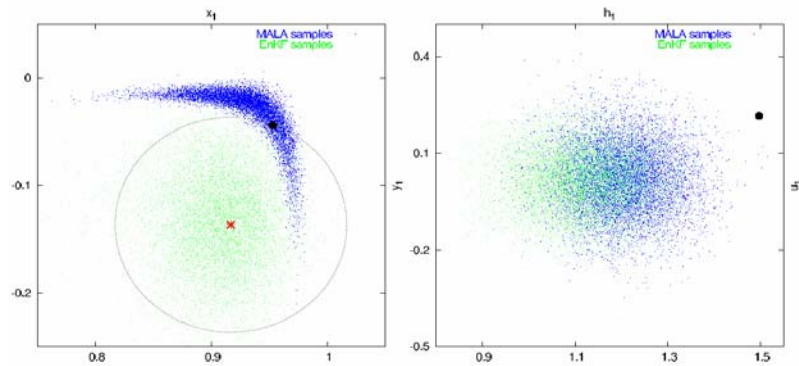


EnKF is handicapped by trying to effectively approximate by a Gaussian and thus not accounting for nonlinear effects

# Trajectory crossing cell boundary



## Scatter Plots



## Conclusions

- In model problem, a modified Langevin sampling does particularly well.
- Going to higher dimensions is obviously a challenge. Salman (2007) has a hybrid method that appears to work well
- Nonlinearity is well addressed, but saddle issue is not resolved
- Filtering vs. smoothing is a serious debate because of chaotic effects