Implicit Sampling for Data Assimilation in Geophysics

M. Morzfeld and A. Chorin (LBNL)
R. Miller, Y. Spitz, and **B. Weir** (OSU)

Data assimilation

Uncertain model

$$\mathbf{X}_m = \mathbf{X}_{m-1} + \tau f(\mathbf{X}_{m-1}, \theta, t_{m-1}) + \sqrt{\tau} G(\mathbf{X}_{m-1}, \theta, t_{m-1}) \mathbf{E}_m$$

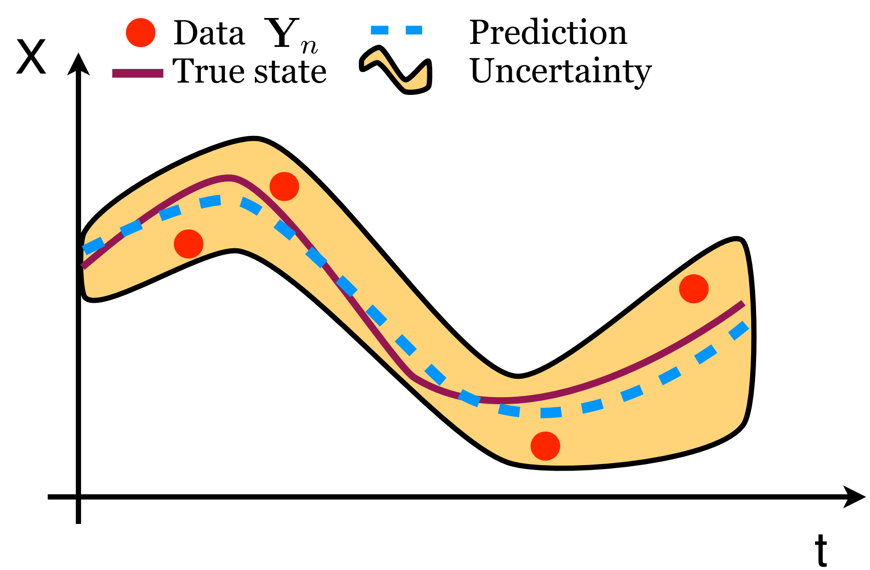
+

Incomplete, noisy observations

$$\mathbf{Y}_n = h(\mathbf{X}_{\mathfrak{m}(n)}, \theta, t_n) + \sqrt{R}\mathbf{D}_n$$

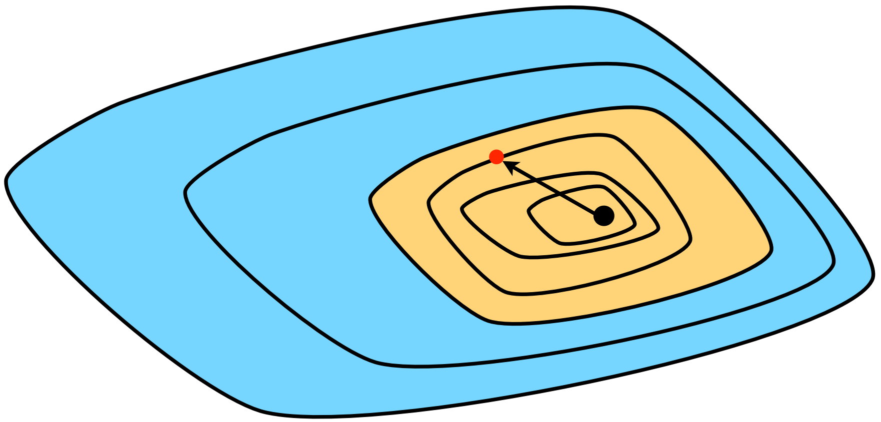
Prediction + uncertainty

$p(\mathbf{x}_{0:m(k)}, \theta | \mathbf{y}_{1:k})$, the target



Implicit sampling

- Monte Carlo method for importance sampling
- No forecast distribution, work directly with target
- Apply particle by particle
- Use numerical optimization to identify high probability regions based on model and observations
- Focus the sampling within these regions

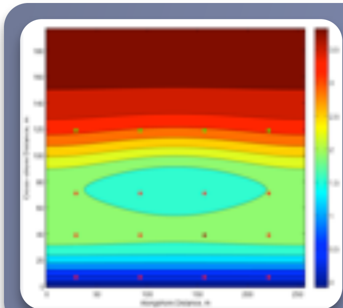


- Black dot: mode
- Red dot: sample
- Yellow area: < 3 std. devs.
- Blue area: > 3 std. devs.

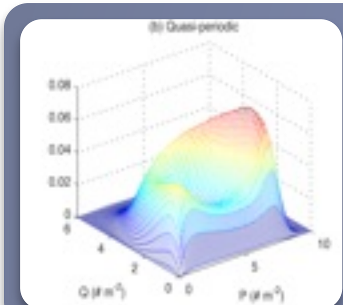
What makes this a good idea?

- Nonparametric
  - Strong theoretical basis for nonlinear/non-Gaussian problems
- Sequential/on-line
  - Can assimilate any number of observations with each application and discard them thereafter
- Optimized for observations
  - Computational resources are directed toward important regions of sample space
  - Avoids sampling “blindly” like many particle filters, i.e., produces samples with non-negligible information
- Many implementations, tuned for application …

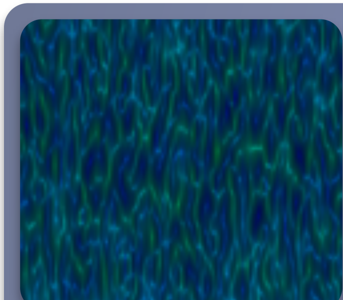
Twin experiments



Shallow water



Predator-prey



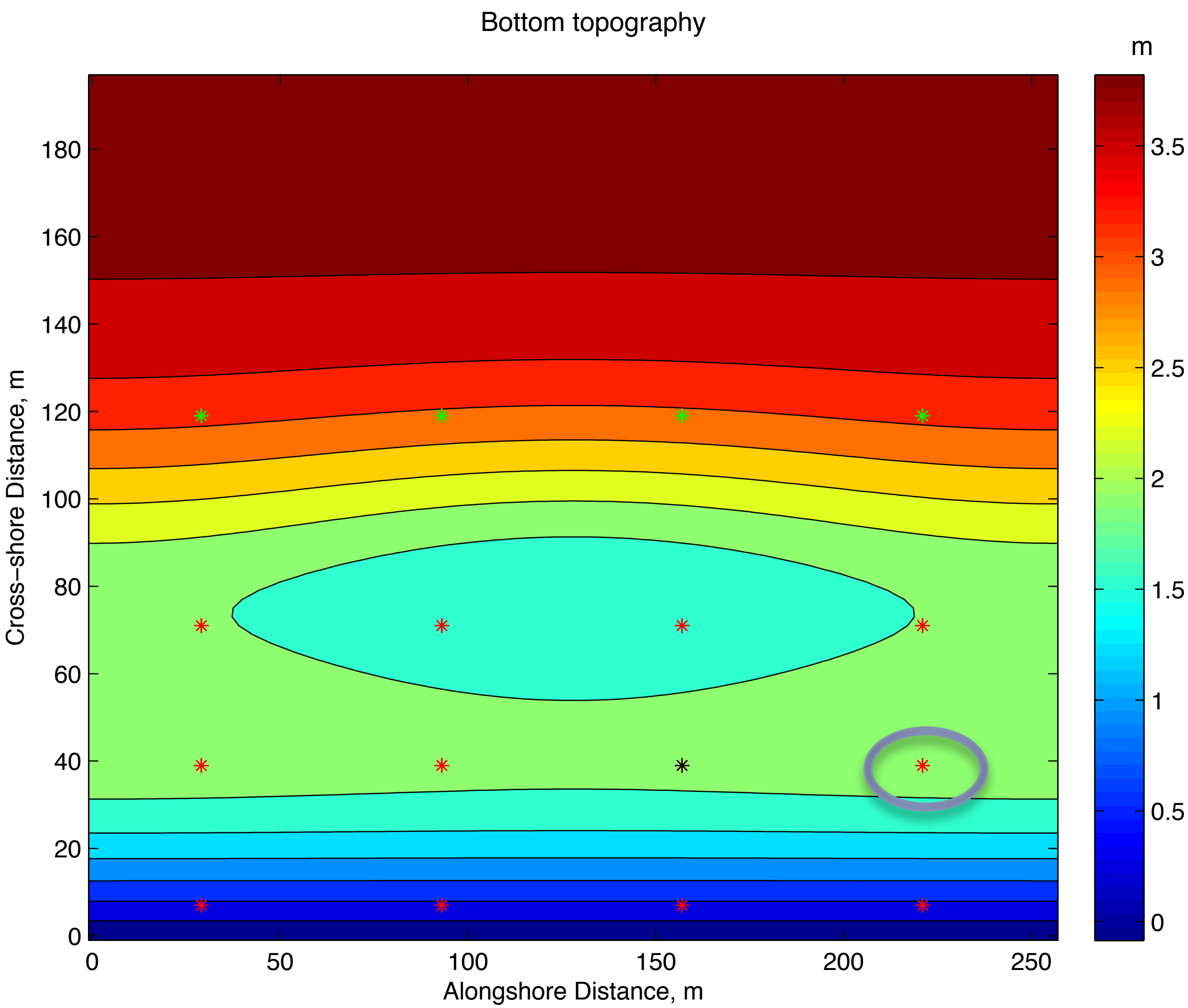
Geomagnetism

1. Shallow water

- State dimension O(30k): height and 2 velocity components on a 100x100 horizontal grid
- Weak dissipation + strong advection + forcing by wave breaking = strong nonlinearity
- Assimilate velocity data every time step at 16 points using 10 particles
- Cost fcn., $\mathcal{J}^{(i)} = -\log\left[p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k+1}, \mathbf{x}_k^{(i)})\right]$

$$= \frac{1}{2}(\mathbf{x}_{k+1} - \mathbf{x}^*)^T H (\mathbf{x}_{k+1} - \mathbf{x}^*)$$

- Target is Gaussian, find $\mathbf{x}^*$ and H from Kalman filter algebra



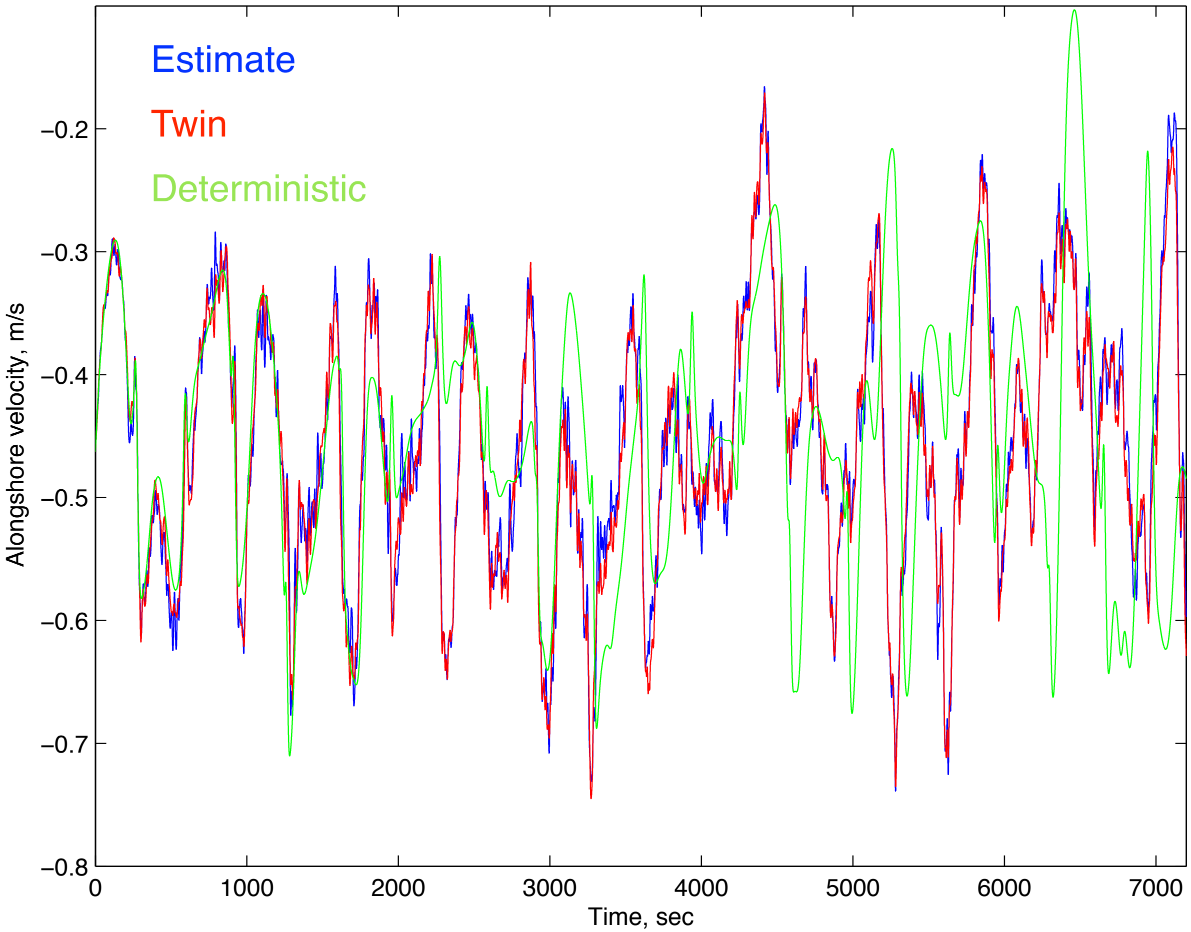
Shallow water

Bottom topography contours

Very shallow = strong forcing by breaking

Observation points indicated with asterisk

Comparison to follow indicated at the circled point



Shallow water

Good agreement between estimate and twin

Assimilation necessary: noticeable phase shift from deterministic and twin solutions after 5000 secs

2. Predator-prey

- Estimate 2 state variables P (prey) and Q (predator) and 7 unknown parameters $\theta = (\theta_1, ..., \theta_7)$.

$$\frac{dP}{dt} = (\theta_1 + \theta_2 P)P + \theta_3 \frac{PQ}{1 + \theta_7 P}$$
$$\frac{dQ}{dt} = (\theta_4 + \theta_5 Q)Q + \theta_6 \frac{PQ}{1 + \theta_7 P}$$

- State and parameters have only one sign
- Transform (anamorphosis) to variables that are more nearly Gaussian, e.g., $\zeta = (\log P, \log Q, \log \theta)$

Predator-prey

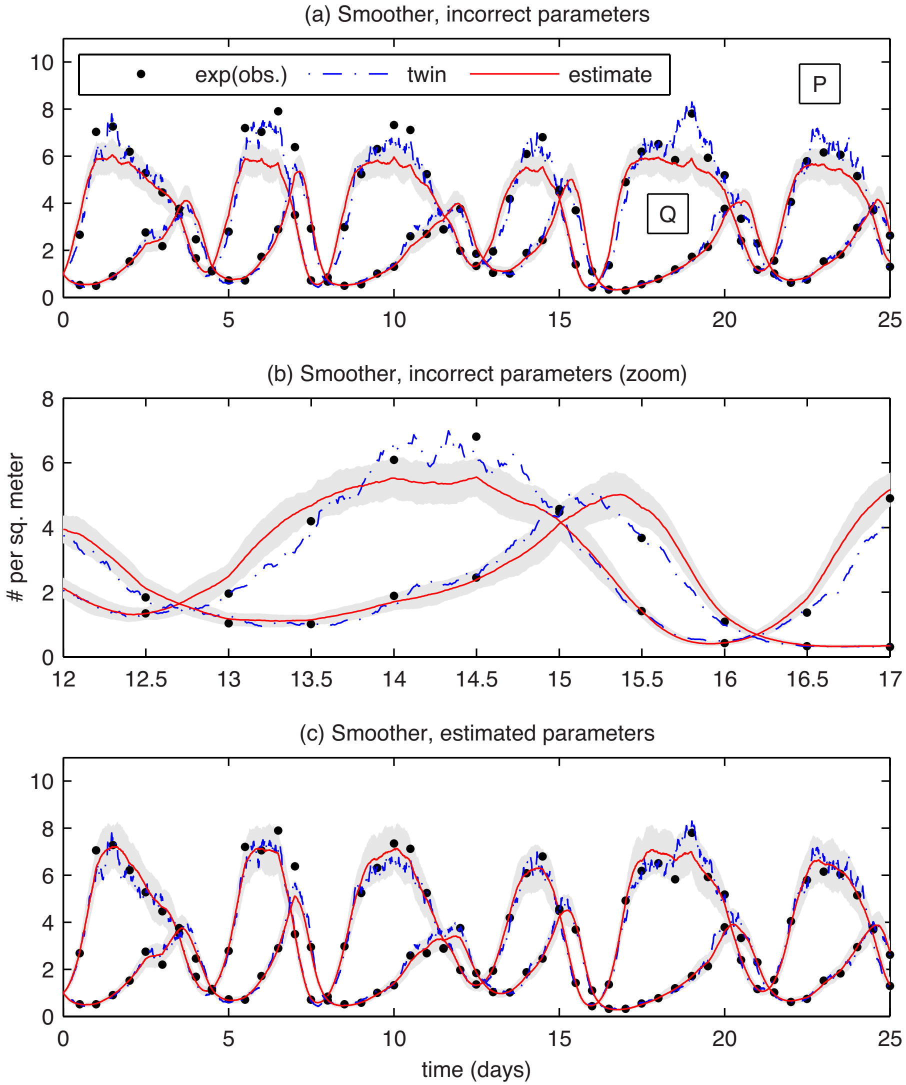
- Assimilate 50 observations every 50 times steps:
  - all at once (smoother), cost function is

$$\mathcal{J} = -\log\left[p(\mathbf{x}_{0:m(k)}, \theta|\mathbf{y}_{1:k})\right]$$

  - or sequentially (filter), cost function is

$$\mathcal{J}^{(i)} = -\log\left[p(\mathbf{x}_{m(k)+1:m(k+1)}, \theta|\mathbf{y}_{1:k+1}, \mathbf{x}^{(i)}_{m(k)}, \theta^{(i)})\right]$$

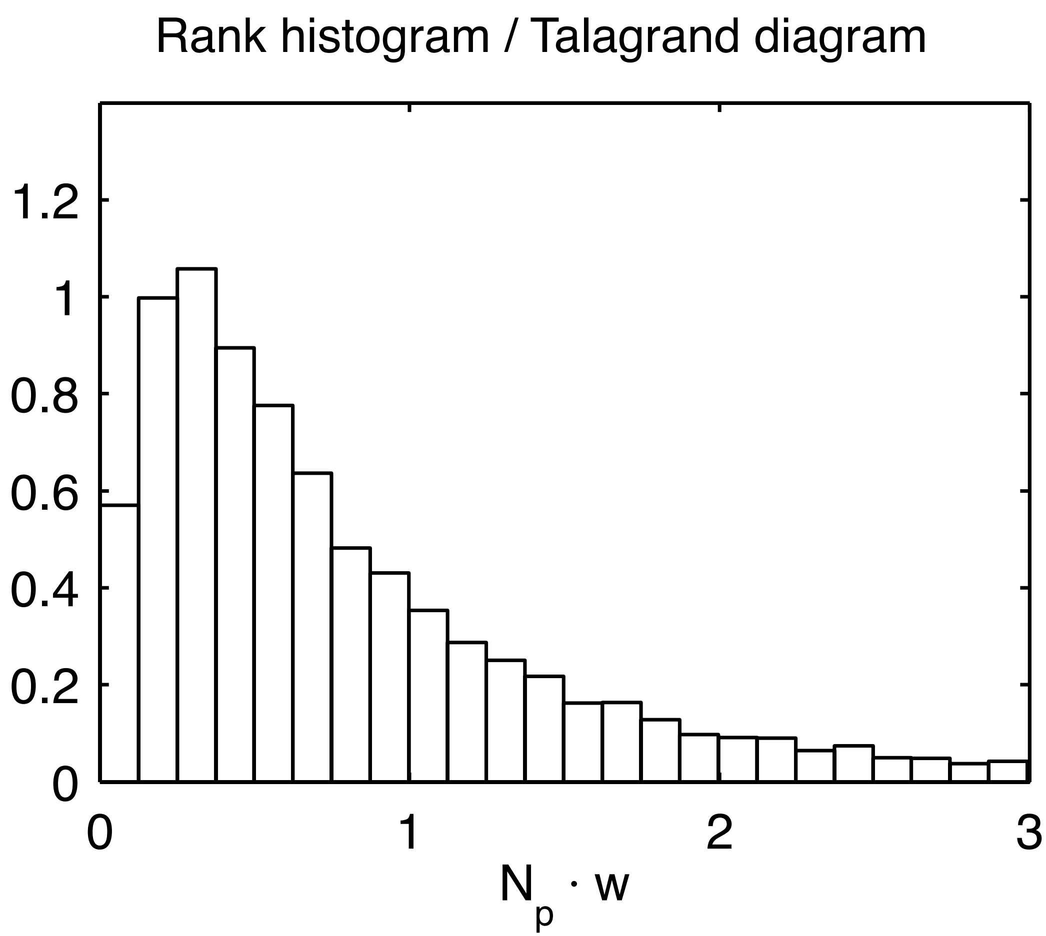
- In transformed variable ζ, cost is nearly, but not exactly quadratic:
  - optimize to find its min ζ*, and Hessian H at ζ*
  - use Gaussian importance sampling



Predator-prey (smoother)

Comparison of state estimate in two cases: (a,b) parameters fixed at incorrect values, (c) estimated parameters

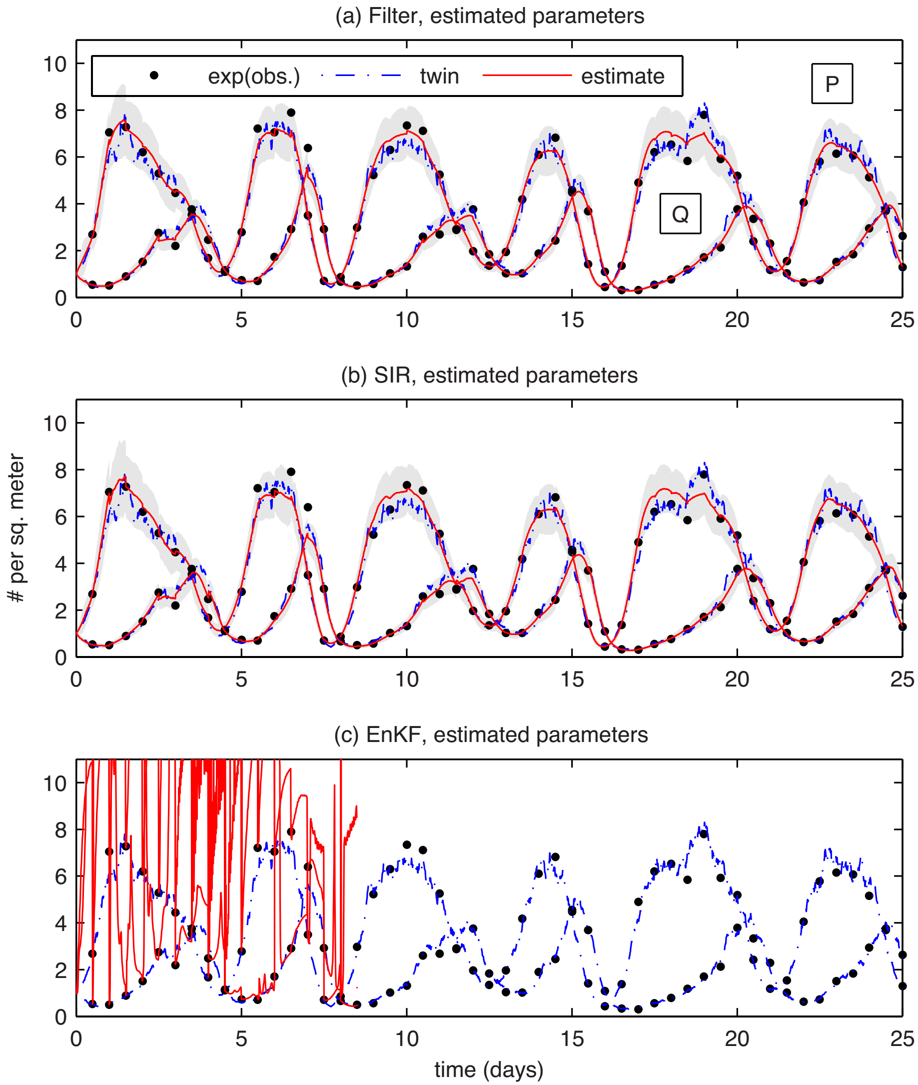
240 particles



Predator-prey (smoother)

Rank histogram computed with 240,000 particles

Noticeable drop-off in distribution before zero



Predator-prey (filter)

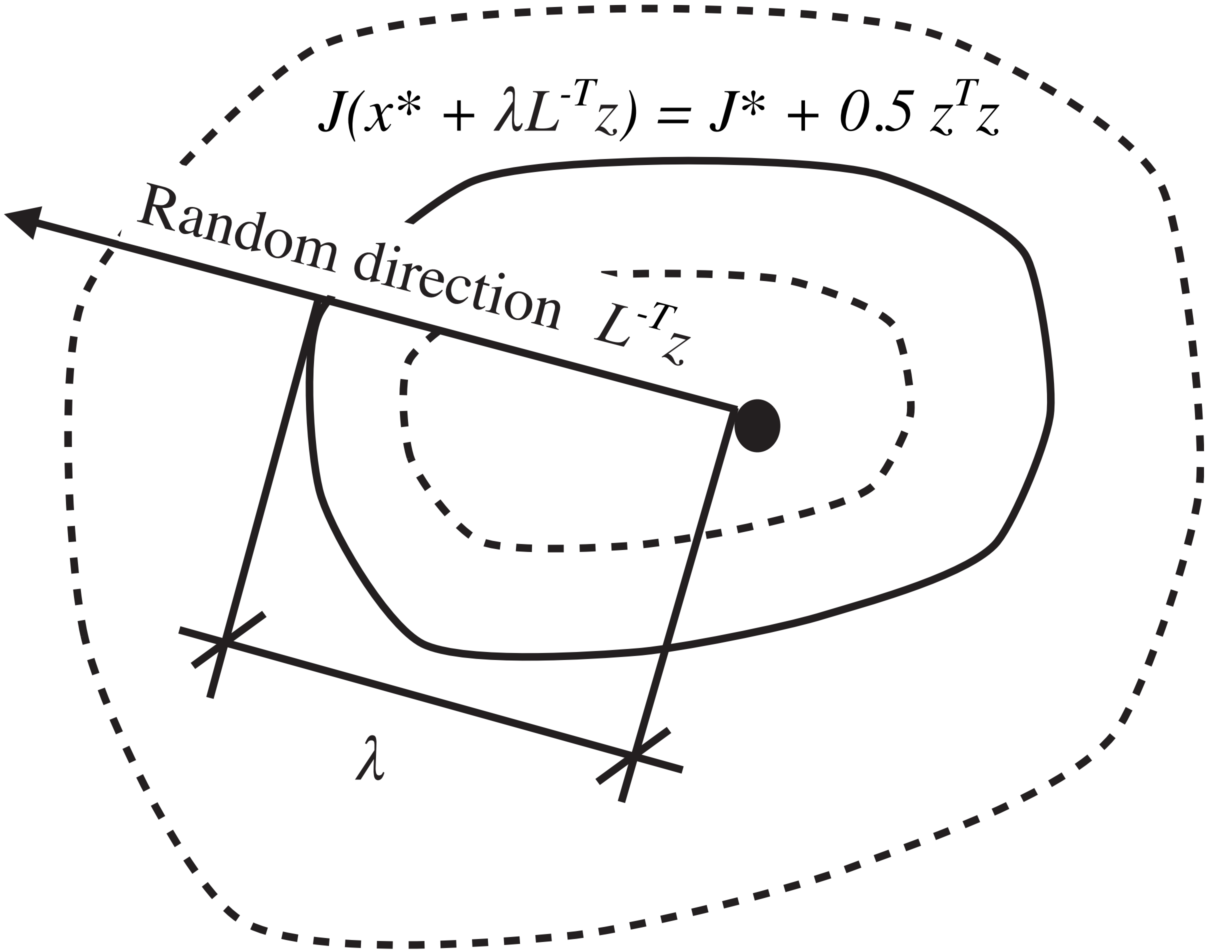
2400 particles for implicit and SIR filters and 240,000 for EnKF

EnKF covariance blows up, works with denser observations

3. Geomagnetism

- SPDEs for velocity and magnetic field
- Legendre spectral elements to transform to system of very stiff SDEs
- Observations of magnetic field only at 200 equally spaced locations
- Cost is far from quadratic: many steps between observations

$$\mathcal{J}^{(i)} = -\log\left[p(\mathbf{x}_{m(k)+1:m(k+1)}, |\mathbf{y}_{1:k+1}, \mathbf{x}^{(i)}_{m(k)})\right]$$



Geomagnetism

If target is far from Gaussian (skew, kurtosis, etc.), use random maps to sample:

1. Draw $z \sim N(0,I)$

2. Cholesky factor $H = LL^T$

3. Search along $L^{-T}z$ until proposal and target have the same level set

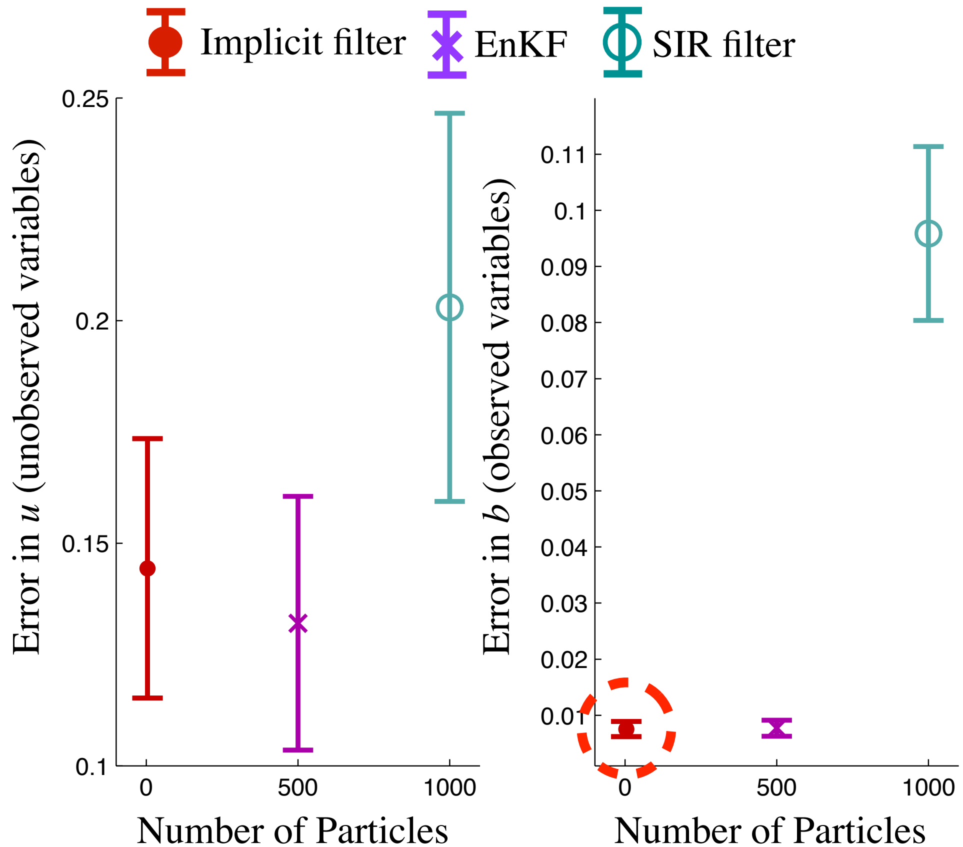


Fig. 5. Filtering results for data collected at a high spatial resolution (200 measurement locations). The errors at $t = 0.2$ of the implicit particle filter (red), EnKF (purple) and SIR filter (green) are plotted as a function of the number of particles. The error bars represent the mean of the errors and mean of the standard deviations of the errors.

Geomagnetism

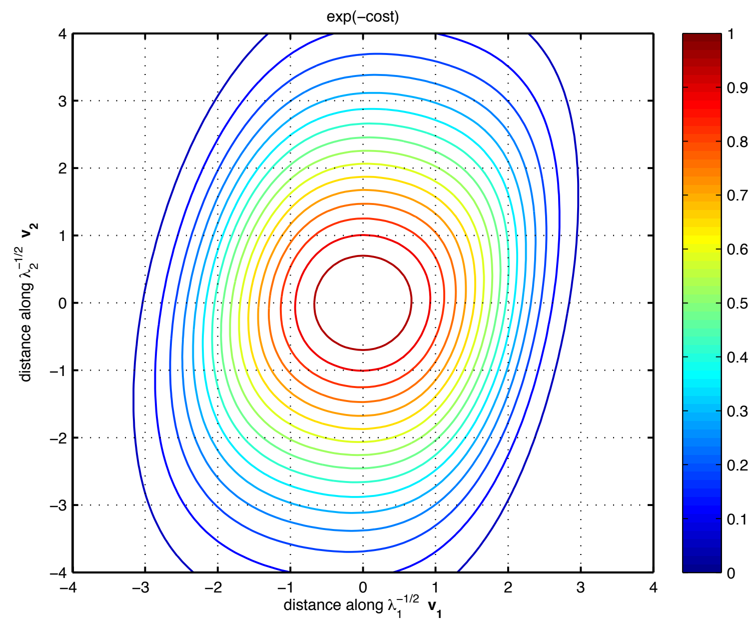
Comparable errors for implicit filter O(10) particles, EnKF O(100) particles, and SIR O(1000) particles

References

- Lorenz ’63, Geomagnetism, and Kuramoto-Sivashinsky (KS)
  - Morzfeld et al., JCP (2012)
  - Morzfeld and Chorin, NPG (2012)
  - Atkins et al., MWR (to appear)
- Predator-prey
  - Weir et al., BMB (to appear)
- KS and shallow water
  - Jardak et al., JNMF (2009)
  - Jardak et al., JGR (2010)
- High dimensional scaling
  - Bengtsson et al., IMS (2008)
  - Bickel et al., IMS (2008)
  - Snyder et al., MWR (2008)
- … many more
  - Miller, van Leeuwen (state estimation)
  - Kivman, Losa, Dowd, Fennel, Mattern (parameter estimation)

Conclusions

- Implicit sampling is successful in 3 applications to nonlinear models in O(10) to O(10k) dimensions
- Able to estimate state and parameters; handles constraints and multiplicative noise
- Can improve results with adaptive refinements of importance density*



*see poster: Estimation of Ecological Model Parameters by Implicit Sampling, Session 1 (today)

fin

merci