Getting the Most Out of Your Models and Observations: A Short Introduction to My Work

Jon Poterjoy

National Research Council Research Associates Program
NOAA Atlantic Oceanographic and Meteorological Laboratory
NOAA National Severe Storms Laboratory

Thursday, 9^{th} March 2017

Definitions

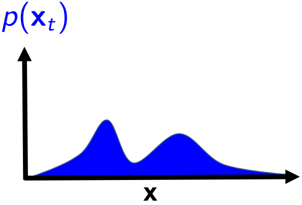
Let $\mathbf{x}_t$ be a vector of state variables, and $\mathbf{y}_t$ be a vector of observations that are valid at time t.

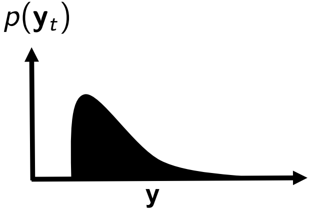
$\mathbf{x}_t$ and $\mathbf{y}_t$ are given by a nonlinear set of equations:

$$\begin{aligned} \mathbf{x}_t &= M(\mathbf{x}_{t-1}) + \eta_t, \\ \mathbf{y}_t &= H(\mathbf{x}_t) + \epsilon_t. \end{aligned}$$

Definitions

We don't know $\mathbf{x}_t$ and $\mathbf{y}_t$ exactly, so the best we can do is estimate their pdfs:





Our goal

What we're really after is the pdf of $\mathbf{x}_t$ conditioned on all information we have:

$$p(\mathbf{x}_t|\mathbf{x}_{0:t-1}, \mathbf{y}_{1:t}).$$

This pdf considers past and present observations, model states, and their errors. In practice, we only need most recent obs and model state:

$$p(\mathbf{x}_t|\mathbf{x}_{0:t-1}, \mathbf{y}_{1:t}) \approx p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t).$$

Our goal

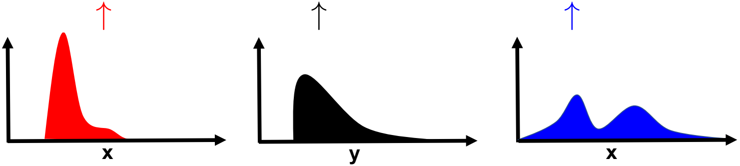
Bayes' theorem tells us where to begin:

$$\underbrace{p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{y}_t)}_{\uparrow \atop \textbf{Posterior (analysis)}} \propto \underbrace{p(\mathbf{y}_t|\mathbf{x}_t)}_{\uparrow \atop \textbf{Likelihood}} \times \underbrace{p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{y}_{t-1})}_{\uparrow \atop \textbf{Prior (background)}}.$$

Our goal

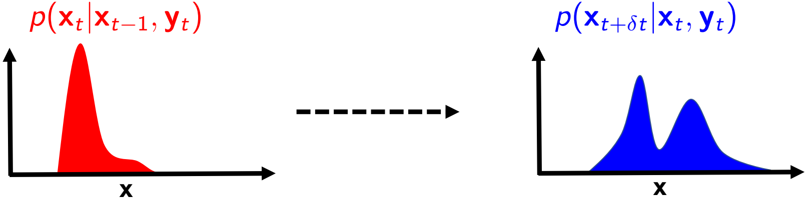
Bayes' theorem tells us where to begin:

$$\textcolor{red}{p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t)} \propto p(\mathbf{y}_t|\mathbf{x}_t) \times \textcolor{blue}{p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{t-1})}.$$



Our goal

For weather prediction, the ultimate objective is to estimate various parts of the forecast probability density $p(\mathbf{x}_{t+\delta t}|\mathbf{x}_t, \mathbf{y}_t)$.



Our goal

For weather prediction, the ultimate objective is to estimate various parts of the forecast probability density $p(\mathbf{x}_{t+\delta t}|\mathbf{x}_t, \mathbf{y}_t)$.

Examples:

- Most probable TC track, intensity, structure, etc.
- Range of likely outcomes (variance in estimates).
- Range of possible outcomes (e.g., solutions that are improbable, but still plausible).

Our goal

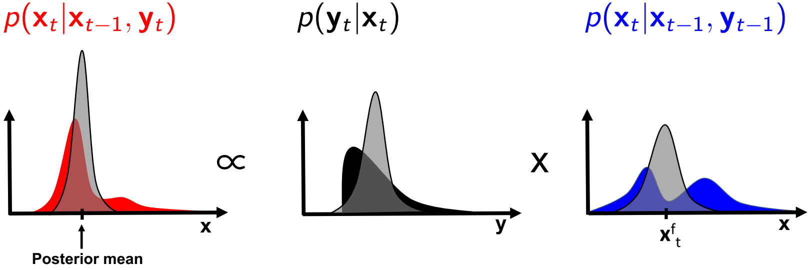
For weather prediction, the ultimate objective is to estimate various parts of the forecast probability density $p(\mathbf{x}_{t+\delta t}|\mathbf{x}_t, \mathbf{y}_t)$.

Improving this estimate requires advancements in

- representation of model processes and their uncertainty: $\mathbf{x}_t = M(\mathbf{x}_{t-1}) + \eta_t$
- observations and their uncertainty: data collection, OSSEs/OSEs, etc.
- data assimilation theory: $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t)$

How is this done on the DA side?

Optimal Interpolation (OI):

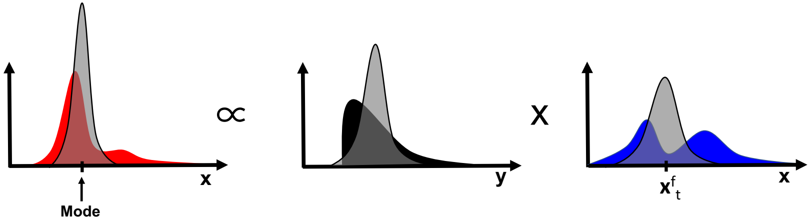


- Assume Gaussian obs errors and Gaussian prior with mean $\mathbf{x}_t^f$ and fixed prior error covariance from climatology.
- Solve explicitly for posterior mean.

How is this done on the DA side?

3D-Variational (3DVar):

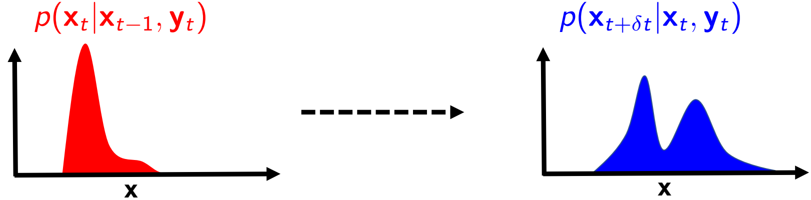
$$p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{y}_t) \quad \propto \quad p(\mathbf{y}_t|\mathbf{x}_t) \quad \times \quad p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{y}_{t-1})$$



- Assume Gaussian obs errors and Gaussian prior with mean $\mathbf{x}_t^f$ and fixed prior error covariance from climatology.
- Solve for posterior mode by minimizing a cost function (differs from OI under certain conditions).

Note on forecast pdf

First two methods do not allow for estimate of $p(\mathbf{x}_{t+\delta t}|\mathbf{x}_t, \mathbf{y}_t)$.
For high-dimensional systems, only computationally feasible estimate is through sequential Monte Carl methods.

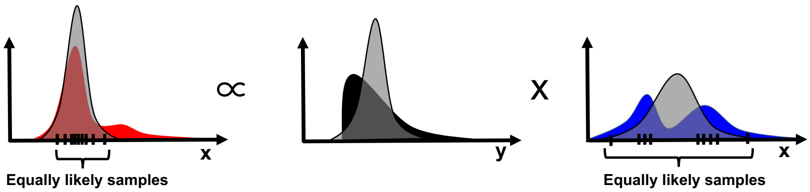


Draw $\mathbf{x}_t^n$ for $n = 1, 2, ..., N_e$, from $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t)$.

$\mathbf{x}_{t+\delta t}^n = M(\mathbf{x}_t^n) + \eta_t^n$, are then samples from $p(\mathbf{x}_{t+\delta t}|\mathbf{x}_t, \mathbf{y}_t)$.

Ensemble Kalman filter (EnKF):

$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t)$ $p(\mathbf{y}_t|\mathbf{x}_t)$ $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$

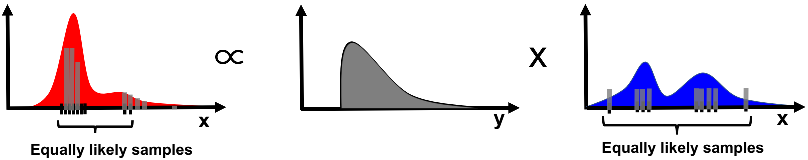


- Assume Gaussian obs errors and Gaussian prior estimated from equally likely ensemble realization of model integrations.
- Solve explicitly for sample (ensemble) that has correct posterior mean and covariance.

How is this done on the DA side?

Particle filter (PF):

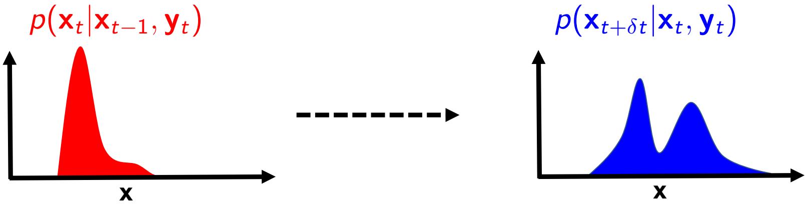
$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t) \qquad p(\mathbf{y}_t|\mathbf{x}_t) \qquad p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$



- Freedom to choose error distribution for obs; assume prior distribution is sum of delta functions centered on each member.
- Posterior given by weighted sum of prior members; samples drawn in neighborhood of members with high weights

Note on forecast pdf

Unlike EnKFs, PFs will converge to true $p(\mathbf{x}_{t+\delta t}|\mathbf{x}_t, \mathbf{y}_t)$ as $N_e \to \infty$, and as representation of $M(\mathbf{x}_t)$ and its errors improve.



Important points

Many operational centers use “hybrid” DA: combines strengths and weaknesses of Var and EnKF (ask about theoretical and practical strengths/weaknesses in weather models).

PFs provide a way of solving complex nonlinear problems that are difficult for EnKFs (see my talk later this month).

PF can be useful for high-dimensional applications by exploiting sparsity of data assimilation problem (Poterjoy 2016, Poterjoy and Anderson 2016)

Research objectives at AOML

Test PF for hurricane OSSEs and OSEs—builds off recent success applying the PF for convective-scale NWP at NSSL.

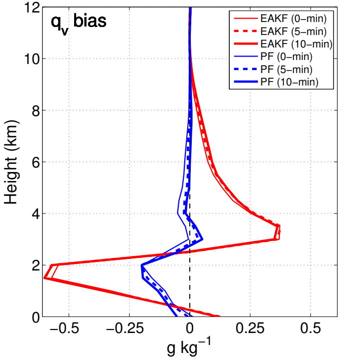
Use new DA framework to assimilate observing systems not used operationally, but may have large potential (e.g., cloudy radiances).

Improve estimates of observation errors used for DA in tropical cyclones.

Research objectives at AOML

Improve the model!

(Loading video...)



- Bias after 3 h of 5-min cycling DA using PF and EnKF with perfect model (Poterjoy et al. 2017).