Efficient assimilation of atmospheric data: a Local Ensemble Transform Kalman Filter

Brian Hunt

University of Maryland
Chaos/Weather Group

http://keck2.umd.edu/weather/
Goal of this Talk

- Describe a framework for the analysis step of an Ensemble Kalman filter that is:
  - Computationally efficient and robust
  - Mathematically simple and flexible

- This approach combines elements of the Ensemble Transform Kalman Filter [Bishop et al. 2001] and the Local Ensemble Kalman Filter [Ott et al. 2004]; I call it the LETKF.
Starting Point

• Data assimilation by statistical interpolation [Lorenc 1981].

• Given a background (first-guess) model state $x^b$ with (presumed) covariance $B$, observations $y$ with covariance $R$, and a function $H$ from model space to observation space, we seek to minimize the function

$$J(x) = (x - x^b)^T B^{-1} (x - x^b) + [y - H(x)]^T R^{-1} [y - H(x)].$$

• The minimizing state $x^a$ is called the analysis.
Ensemble Kalman Filtering

- Ensemble Kalman filters [Evensen 1994] evolve an ensemble of initial conditions with the nonlinear model and let $x^b$ and $B$ be the sample mean and covariance of the ensemble forecast states at the analysis time.

- **Bad news**: The background covariance $B$ reflects only uncertainties in the space $S$ spanned by the ensemble.

- **Good news**: The analysis takes place in the low-dimensional space $S$ (computationally efficient).
Guiding Question

- Which linear combination of the ensemble states best fits the data?

- Let $k$ be the number of ensemble members, and let $X$ be a matrix of normalized ensemble perturbations: each column of $X$ is the difference between an ensemble state $x^b_j$ and the ensemble mean $x^b$, divided by $\sqrt{k - 1}$ so that $B = XX^T$.

- Express a model state $x$ in $S$ as $x = x^b + Xw$, where $w$ is a $k$-dimensional weight vector. Which $w$ is best?
Linearization of $H(x)$ in Ensemble Space $S$

- For each ensemble state $x^b_j$, let $y^b_j = H(x^b_j)$. Let the mean of this background observation ensemble be $y^b$, and let $Y$ be the matrix with columns $(y^b_j - y^b)/\sqrt{k-1}$.

- Make the linear approximation $H(x^b + Xw) \approx y^b + Yw$.

- In terms of $w$, the function to be minimized is

$$J(w) \approx w^Tw + (y - y^b - Yw)^TR^{-1}(y - y^b - Yw).$$

(The background mean of $w$ is 0 and its background covariance is the identity matrix.)
Analysis Mean and Covariance

• In the \( w \) coordinate system, the analysis mean \( w^a \) and covariance \( A \) are given by the standard Kalman filter equations

\[
A = (I + Y^T R^{-1} Y)^{-1}
\]

\[
w^a = A Y^T R^{-1} (y - y^b)
\]

• The matrix that is inverted to find \( A \) is small (\( k \) by \( k \)) and has no small eigenvalues.

• So far this approach is essentially equivalent to the ETKF [Bishop et al. 2001].
Analysis Ensemble

- To form the analysis ensemble weight vectors $w^a_j$, add to $w^a$ the columns of the symmetric matrix $W = [(k - 1)A]^{1/2}$; this ensures the correct analysis covariance.

- Any matrix for which $WW^T = (k - 1)A$ would do; this is the choice available in a square root filter [Tippett et al. 2003].

- Our choice minimizes the distance between the background and analysis ensembles and ensures that, when done locally, the analysis ensemble varies continuously from one region to the next [Ott et al. 2004].
Localization

- For an ensemble of moderate size \((k < 100)\), the linear combination that best fits the data in one region may be significantly different from the best linear combination in another region.

- Localize by doing a separate analysis at each model grid point, ignoring data beyond a certain distance [Houtekamer & Mitchell 1998].

- One can choose which data to use, and do the resulting analysis, independently at each grid point.
Asynchronous Observations

• In an operational setting, data cannot be assimilated as frequently as it is taken; several hours worth of data is assimilated at one analysis time.

• Which linear combination of the ensemble trajectories best fits the data?

• For observations taken at time $t$, apply $H$ to the ensemble states at time $t$ when forming the background observation ensemble vectors $y^b_j$, and proceed exactly as before. This simplifies the 4D approach described in [Hunt et al. 2004].
Preliminary Results

- Based on tests by my colleagues J. Harlim on a toy model [Lorenz 1996] and E. Kostelich and I. Szunyogh on the NCEP GFS, the LETKF described here produces analyses of similar quality to our group’s LEKF [Ott et al. 2004, Szunyogh et al. 2004].

- The LETKF runs about three times as fast as the LEKF on the GFS; it is able to assimilate 1.5 million (simulated) observations on 0.5 million grid points in under 5 minutes using a 40 member ensemble on a cluster of 40 Pentium processors.
Conclusions

• The LETKF can efficiently assimilate a large amount of atmospheric data.

• It does not require linearizing the observation operator.

• The amount of localization is easily adjusted.

• Observations taken at different times can be assimilated simultaneously.

• More from our group:

http://keck2.umd.edu/weather/