The Local Ensemble Transform Kalman Filter

Eric J. Kostelich

Arizona State University
School of Mathematical & Statistical Sciences

ICTS Data Assimilation Research Program
July 7, 2011

Co-workers and Web resources

Co-workers:

Istvan Szunyogh, Brian Hunt, Edward Ott,

Eugenia Kalnay, Jim Yorke

and many others!

Thanks to: NSF, NASA, ASU

Papers, preprints, and codes:

http://www.weatherchaos.umd.edu
http://math.la.asu.edu/∼eric

Principal papers

Preprints: www.weatherchaos.umd.edu

Initial papers:
- E. Ott et al., Tellus A **56** (2004), 415–428.
- I. Szunyogh et al., Tellus A **57** (2005), 528–545.

Refined mathematical implementation: B. R. Hunt, E. K., I. Szunyogh, Physica D **230** (2007) 112–126.

Results with real data: I. Szunyogh, E.K. et al., Tellus A **60** (2008) 113–130.

Recap from last time

- In a chaotic process, every point is sensitive
- Uncertainties in initial conditions grow exponentially (at least for awhile)
- The weather (or at least a numerical model thereof) is chaotic
- The uncertainty in the global weather vector roughly doubles every 2 days
- Forecast horizon: about 2 weeks
- Frequent updates of the initial conditions with observations are required

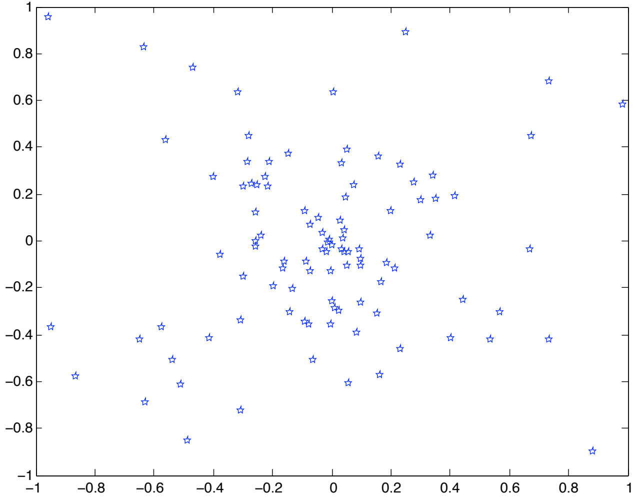
The ensemble dimension

- The ensemble dimension (E-dimension) of an $n \times k$ matrix is

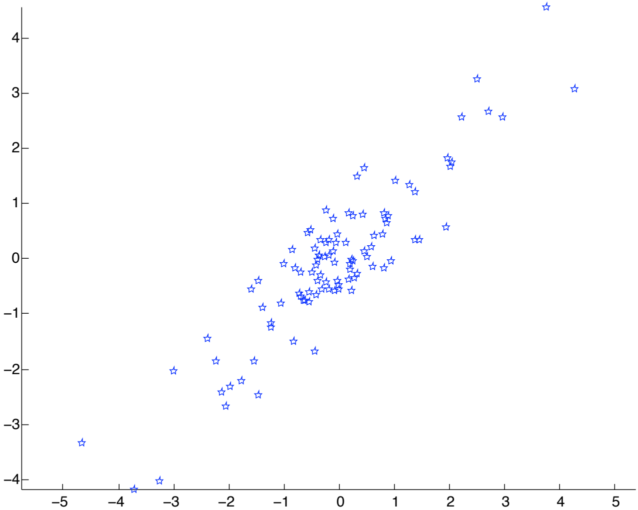
$$E \equiv \frac{(s_1 + s_2 + \cdots + s_k)^2}{s_1^2 + s_2^2 + \cdots + s_k^2}$$

- Measures the eccentricity of the “ellipse” of forecast uncertainty

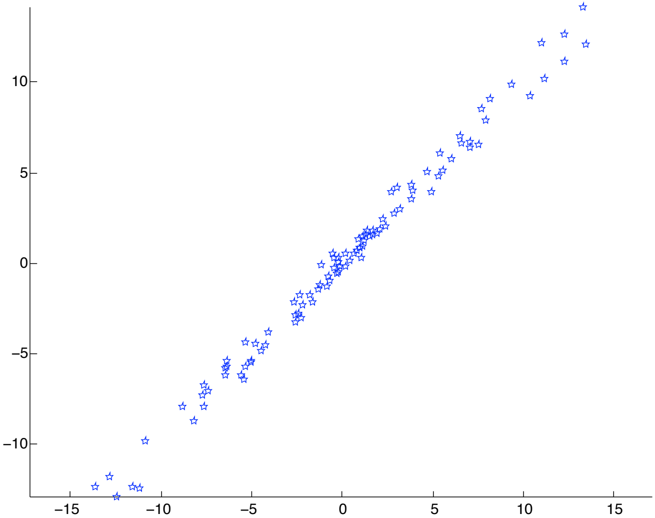
Example: $s_1 = 3.78$, $s_2 = 3.60$, $E_{\text{dim}} = 1.99$



Example: $s_1 = 19.24$, $s_2 = 4.35$, $E_{\text{dim}} = 1.43$



Example: $s_1 = 83.65$, $s_2 = 4.33$, $E_{\text{dim}} = 1.10$

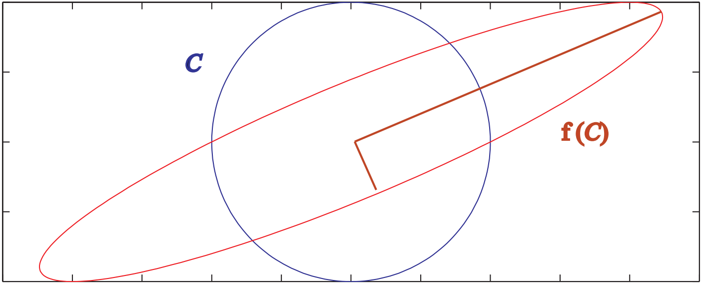


Local low dimensionality

- A medium-resolution weather model has about 3000 variables over a typical $1000 \text{ km} \times 1000 \text{ km}$ synoptic region
- So an ensemble of size k can be regarded as a $3000 \times k$ matrix
- The ensemble dimension is ~ 40 for $100 \leq k \leq 200$
- Key empirical finding of

  D. J. Patil et al. PRL **86** (2001), 5878–5881.

Cartoon for weather models



- The expanding subspace is low dimensional
- The contracting subspace is high dimensional
- Goal: worry only about the expanding subspace

The Global Forecast System (GFS)

- Dynamical variables in the GFS:
  - natural logarithm of surface pressure
  - virtual temperature
  - divergence and vorticity of the wind field
- Passive tracers:
  - Ozone
  - Water vapor
- Principal measurements:
  - barometric pressure
  - sensible temperature
  - relative humidity
  - wind speed and direction
  - satellite radiances (complicated!)

Mathematical assumptions

- Observations: $\mathbf{y} \in \mathbb{R}^p$, $\mathbf{y} = \mathbf{H}(\mathbf{x}_t) + \varepsilon$
- Observation errors: $\mathbf{E}(\varepsilon) = \mathbf{0}$, $\mathbf{E}(\varepsilon\varepsilon^{\mathrm{T}}) = \mathbf{R}$
- Model forecast ("background"): $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}_b = \mathbf{x}_t + \eta$
- Model errors: $\mathbf{E}(\eta) = \mathbf{0}$, $\mathbf{E}(\eta\eta^{\mathrm{T}}) = \mathbf{P}_b$
- Goal: minimize the objective function

$$J(\mathbf{x}) = [\mathbf{y} - \mathbf{H}(\mathbf{x})]^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{y} - \mathbf{H}(\mathbf{x})] + (\mathbf{x} - \mathbf{x}_b)^{\mathrm{T}}\mathbf{P}_b^{-1}(\mathbf{x} - \mathbf{x}_b)$$

- Minimization produces an analysis $\mathbf{x}_a$ with associated covariance $\mathbf{P}_a$

Simplest assumptions

- The observation errors ε are normally distributed with mean $\mathbf{0}$ and covariance $\mathbf{R}$
- Model errors similarly: $N(\mathbf{0}, \mathbf{P}_b)$
- When the underlying model is linear, it can be shown the the minimizer $\mathbf{x}_a$ of J is unique, unbiased and has minimum variance among all linear estimators
- Weather models are "linear enough" over 6-hour intervals, but there is no guarantee of optimality

The dimensionality problem

- Must evaluate

  $$J(\mathbf{x}) = [\mathbf{y} - \mathbf{H}(\mathbf{x})]^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{y} - \mathbf{H}(\mathbf{x})] + (\mathbf{x} - \mathbf{x}_b)^{\mathrm{T}}\mathbf{P}_b^{-1}(\mathbf{x} - \mathbf{x}_b)$$

  where $\mathbf{y} \in \mathbf{R}^p$, $\mathbf{x} \in \mathbf{R}^n$
- Current NCEP operations: $p \sim 1.75$ million and $n \sim 3$ billion
- We need $\mathbf{R}^{-1}$ $(p \times p)$ and $\mathbf{P}_b^{-1}$ $(n \times n)$

The computational complexity problem

- Inversion of a $k \times k$ matrix is an $O(k^3)$ algorithm
- If a 100×100 matrix takes ~ 1 sec to invert, then computing $10^9 \times 10^9$ matrix $\mathbf{P}_b^{-1}$ takes $\sim 10^{18}$ sec (Note: 1 year $\approx 3 \times 10^7$ sec)
- $\mathbf{R}$ is nearly diagonal if observation errors are mostly uncorrelated
- $\mathbf{P}_b$ is not diagonal
- Computing $\mathbf{P}_b(t+\Delta t)$ from $\mathbf{P}_a(t)$ requires integration of the tangent linear model

Complexity reduction strategies

- Localization: Try to do the minimzation over smaller regions of the globe
- Estimate and precompute $\mathbf{P}_b^{-1}$: Assume that the forecast uncertainty is approximately constant from one day to the next. (Used in all current operational DA systems)
- Process the observations sequentially and discard those regarded as redundant

Each strategy has drawbacks

- Assuming $\mathbf{P}_b \approx$ constant ignores the “errors of the day”
- Generally regarded as one of the key impediments to better forecasts
- The result of sequential assimilation of observations depends on the order of processing
- Must assure continuity at the boundaries of the smaller regions

The Local Ensemble Transform Kalman Filter (LETKF)

- Addresses many of these problems
- Exploits the local low dimensionality of numerical weather models to reduce the matrix sizes but still account for errors of the day
- Assimilates all the data at once
- Uses localization and sets of observations that vary slowly in space to help assure continuity
- Permits efficient implementation on massively parallel computers

LETKF ideas for reducing computational complexity

- Given the set of forecast vectors $\{\mathbf{x}_b^i\}_{i=1}^k$, form their mean $\bar{\mathbf{x}}_b$ and the $\ell \times k$ background perturbation matrix $\mathbf{X}_b$ whose ith column is $\mathbf{x}_b^i - \bar{\mathbf{x}}_b$ where each column of $\mathbf{X}_b$ is a forecast
- Local low dimensionality means $\widehat{\mathbf{P}}_b = (k-1)^{-1}\mathbf{X}_b\mathbf{X}_b^{\mathrm{T}}$ is a good low-rank approximation of the forecast covariance at each grid point
- Localize the observations to find the "best" linear combination of ensemble solutions at each grid point (in the sense of minimizing J)

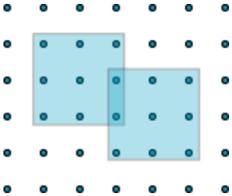
A technical detail

- The objective function J requires that we compute $\mathbf{P}_b^{-1}$
- The columns of $\mathbf{X}_b$ are $\mathbf{x}_b^i - \bar{\mathbf{x}}_b$, $i = 1, \ldots, k$
- But given k ensemble solutions, the rank of the $k \times k$ matrix $\widehat{\mathbf{P}}_b = (k-1)^{-1}\mathbf{X}_b\mathbf{X}_b^{\mathrm{T}}$ is at most $k-1$
- So think of $\mathbf{X}_b$ as a linear transformation from an abstract k dimensional space $\widetilde{S}$ to a $k-1$ dimensional subspace $S \subset \widetilde{S}$ (S is the column space of $\mathbf{X}_b$)
- We perform the minimization of J in $\widetilde{S}$

Technical detail, 2

- If $\mathbf{w} \in \widetilde{S}$ then $\mathbf{X}_b\mathbf{w} \in S$
- Also $\mathbf{x} = \bar{\mathbf{x}}_b + \mathbf{X}_b\mathbf{w}$ is the corresponding model state
- If $\mathbf{w}$ is Gaussian with mean $\mathbf{0}$ and covariance $(k-1)^{-1}\mathbf{I}$, then $\mathbf{x}$ is Gaussian with mean $\bar{x}_b$ and covariance $\widehat{\mathbf{P}}_b$
- Goal: find the linear combination of background forecast that minimizes $\widetilde{J}(\mathbf{w}) = (k-1)^{-1}\mathbf{w}^{\mathrm{T}}\mathbf{w} + [\mathbf{y} - \bar{\mathbf{y}}_b - \mathbf{Y}_b\mathbf{w}]^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{y} - \bar{\mathbf{y}}_b - \mathbf{Y}_b\mathbf{w}]$
- The minimizer $\mathbf{w}_a$ of $\widetilde{J}$ is perpendicular to the null space of $\mathbf{X}_b$, and $\bar{\mathbf{x}}_a = \bar{\mathbf{x}}_b + \mathbf{X}_b\mathbf{w}_a$ minimizes the original J

The LETKF algorithm



- Proceed grid point by grid point
- Use only observations within a prespecified region about each grid point
- Grid points are processed independently
- The set of observations used for each local analysis varies slowly in space (important for continuity)

Treatment of observations

- Compute the observation operator for each ensemble solution to form $\{\mathbf{H}(\mathbf{x}_b^i)\}$ and its mean $\bar{\mathbf{y}}_b$
- Suppose the local region contains s observations
- Linearize $\mathbf{H}$ about the ensemble mean as $\mathbf{H}(\bar{\mathbf{x}}_b + \mathbf{X}_b\mathbf{w}) \approx \bar{\mathbf{y}}_b + \mathbf{Y}_b\mathbf{w}$ where $\mathbf{Y}_b$ is the $s \times k$ matrix whose ith column is $\mathbf{H}(\mathbf{x}_b^i) - \bar{\mathbf{y}}_b$
- Only the s components of $\mathbf{H}(\mathbf{x}_b^i)$ belonging to the given local region are used in the computation of $\bar{\mathbf{y}}_b$ and $\overline{\mathbf{Y}}_b$

Net result

- The minimizer is the k-vector $\mathbf{w}_a = \mathbf{Q}\mathbf{Y}_b^T\mathbf{R}^{-1}(\mathbf{y} - \bar{\mathbf{y}}_b)$ where $\mathbf{Q} = [(k-1)\mathbf{I} + \mathbf{Y}_b^T\mathbf{R}^{-1}\mathbf{Y}_b]^{-1}$ is $k \times k$
- In model space, the analysis mean becomes $\bar{\mathbf{x}}_a = \bar{\mathbf{x}}_b + \mathbf{X}_b\mathbf{w}_a$
- The analysis perturbations are given by $\mathbf{X}_a = \mathbf{X}_b\mathbf{W}_a$ where $\mathbf{W}_a = [(k-1)\mathbf{Q}]^{1/2}$ and we take the symmetric square root
- This choice assures that $\mathbf{W}_a$ depends continuously on $\mathbf{Q}$ and that the columns of $\mathbf{X}_a$ sum to zero (for the correct sample mean)

A word about matrix square roots

- Suppose that $\mathbf{A}$ is a symmetric positive definite matrix
- If $\mathbf{A} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$, then $\mathbf{S}$ is a square root of $\mathbf{A}$
- $\mathbf{S}$ can be found from the Cholesky decomposition of $\mathbf{A}$
- $\mathbf{S}$ is not unique: If $\mathbf{U}$ is any orthonormal matrix, then

$$(\mathbf{SU})(\mathbf{SU})^{\mathrm{T}} = \mathbf{SUU}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}} = \mathbf{SS}^{\mathrm{T}}$$

- The Cholesky decomposition is not continuous as a function of the elements of $\mathbf{A}$

A word about matrix square roots, 2

- Instead, we consider the symmetric square root of $\mathbf{A}$
- Suppose $\mathbf{D}$ is the diagonal matrix containing the eigenvalues of $\mathbf{A}$
- Let $\mathbf{V}$ be the matrix whose corresponding columns are the (orthonormal) eigenvectors (so $\mathbf{AV} = \mathbf{VD}$)
- Then $\mathbf{S} = \mathbf{VD}^{1/2}$ is a square root of $\mathbf{A}$
- This choice assures that the columns of $\mathbf{X}_a$ sum to $\mathbf{0}$ and that $\mathbf{S}$ depends continuously on $\mathbf{Q}$
- Technical details in Hunt, E. K., and Szunyogh, Physica D **230** (2007) 112–126

Computational complexity

- In principle, the only free parameters are the ensemble size k and the size of the local regions
- Given a local region with s observations and an ensemble of size k
- The most expensive step ($> 90\%$ of cpu cycles): computing $\mathbf{Y}_b^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}_b$, which is $O(k^2 s)$
- Second most expensive step: computing the symmetric square root, which is $O(k^3)$
- Observation lookup is $O(\log L)$, where L is the size of the total observation set

Efficient parallel implementation

- All data needed to analyze a given model grid point is distributed only once
- Efficient load balancing algorithm provides nearly linear scaling up to 128 processors (at least)
- Assimilating $\sim$ 250,000 observations into the T62 GFS ($\sim$ 3,000,000 variables) takes 180 seconds on 16 single-core Xeon processors
- Greatest expense is computing the forecasts (also embarrassingly parallel)

Key properties of the LETKF

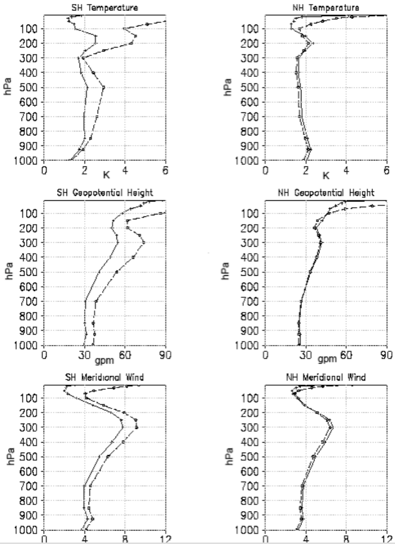
- Estimates of the most likely current state and also its uncertainty
- Assimilates all data at once
- Observations can be nonlinear functions of the state vector
- Can interpolate in time as well as space
- The only free parameters are the ensemble size and size of local regions
- Model independent (no adjoints!)

Validation experiments

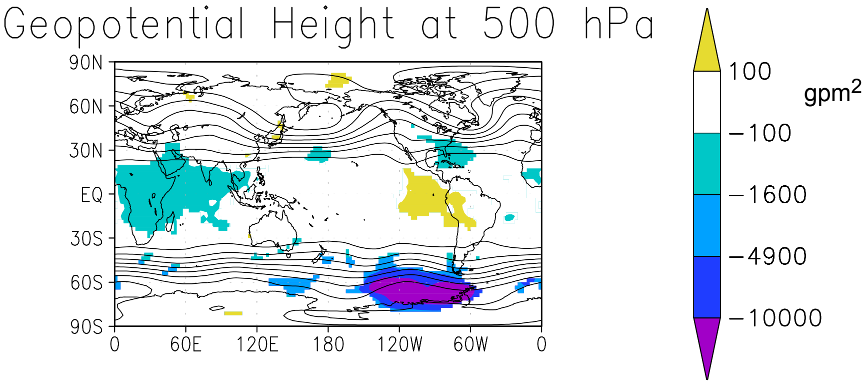
- Operational (T254, ~ 50 km) analyses with full observing network are taken as "truth"
- LETKF used with full observing network (minus satellite radiances but including satellite-derived wind estimates) and 60-member ensemble
- Analysis/forecast cycle run from Jan. 1 to March 1, 2004 using operational GFS at T62 resolution
- Benchmark analysis: NCEP operational system at T62 (~ 200 km) with reduced dataset
- LETKF analysis: LETKF using reduced dataset at T62
- Thanks to Zoltan Toth and NCEP

Sample results: 48-hour forecast error

- NCEP operational analyses taken as truth
- Compute $\langle \text{forecast} - \text{truth} \rangle$ using forecasts started from LETKF and NCEP T62 analyses from Jan. 11 to Feb. 27, 2004
- NCEP surface analyses form the boundary conditions



Locations of greatest 48-hour forecast improvement



Not all problems are solved!

- What if the model has systematic biases?
- Not all observations give sufficient information about the dynamics (the observability problem)
- If you could take more observations, where and when should you do so? (the targeted observation problem)
- Observation error has two components:
  - Instrumentation error
  - Representativeness error (from subgrid-scale dynamics)
- Nonlinearities cause $\widehat{\mathbf{P}}_b$ to underestimate the true uncertainty

Conclusions

- The LETKF merits serious consideration for research and operational applications
- LETKF's advantages are greatest where observations are sparsest
- Ideal for planetary atmosphere DA, if the model errors are sufficiently small
- Model independence makes it adaptable
- One of the most mature, flexible, computationally efficient, and well-tested ensemble-based Kalman filters

Our sponsors

Thanks to:

- National Science Foundation
- Keck Foundation
- McDonnell Foundation
- NASA
- Army Research Office
- National Centers for Environmental Research
- ASU College of Liberal Arts and Sciences