A Two-Stage Ensemble Kalman Filter for Smooth Data Assimilation

Craig J. Johns Jan Mandel *

March 31, 2005

Abstract

The ensemble Kalman Filter (EnKF) and variants derived therefrom have become important techniques in data assimilation problems. One breakdown of the EnKF method is that in the case of sparsely observed, accurate data, least squares properties of the EnKF create posterior ensemble members that are not compatible with the dynamic model. We propose a modification of Kalman and EnKF filters by imposing a constraint either by projection or by a penalty. A two-step ensemble Kalman Filter is proposed that imposes smoothness as a penalized constraint. The smoothing step consists of another application of the same EnKF code with the smoothness constraint as an independent observation. The utility of the method is demonstrated on a nonlinear dynamic model of wildfire.

KEYWORDS: Dynamic Data Assimilation, Ensemble Kalman filter, State-space Model, Constraints, Tikhonov regularization

1 Introduction

The discrete time state-space model in its most general form is an application of the Bayesian update problem: the modeled system is advanced in time until an *analysis* time, when the distribution of the system state before the update, called the *prior* or the *forecast* distribution, and the *data likelihood* are combined to give the new system state distribution, called the *posterior* or the *analysis* distribution. The system is then advanced until the next analysis time. Kalman [20] and Kalman and Bucy [21] provided simple recursive formulas for the system mean and covariance under the assumptions that the probability distributions are normal and the system is linear, thus an advancement in time preserves normality. The simplicity of Kalman updates and its least squares properties [29] have made the method popular in areas as diverse as medicine [19], economics [30] and geosciences [11]. The method is of such utility that it is standard fare in time series texts such as [31] and [4], as well as the primary emphasis of complete texts [14, 16].

1.1 The Kalman Filter

In the Bayesian update, the probability density $p(\mathbf{x})$ of the system state $\mathbf{x}$ before the update (the prior) and the probability density $p(\mathbf{y}|\mathbf{x})$ of the the data $\mathbf{y}$ given an assumed value of

*Department of Mathematics, University of Colorado, Denver, CO, 80217-3364

the system state $\mathbf{x}$ (the data likelihood) are combined to give the new probability density of the system state $p(\mathbf{x}|\mathbf{y})$ (the posterior) by

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}). \tag{1}$$

where $\propto$ means proportionality. Equation (1) determines the posterior density $p(\mathbf{x}|\mathbf{y})$ completely because $\int p(\mathbf{x}|\mathbf{y})\, d\omega(\mathbf{x}) = 1$. Consider the case of linear observation operator H: given system state $\mathbf{x}$, the data value, $\mathbf{y}$, would be $H\mathbf{x}$ if the model and the data were perfect with no errors. Of course, in general, the given data $\mathbf{y} \neq H\mathbf{x}$, so discrepancies are modeled with the likelihood $p(\mathbf{y}|\mathbf{x})$. Assume that the prior has normal distribution with mean μ and covariance Q, and the data likelihood is normal with mean $H\mathbf{x}$ and covariance R,

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}} Q^{-1}(\mathbf{x}-\mu)\right), \quad p(\mathbf{y}|\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{y}-H\mathbf{x})^{\mathrm{T}} R^{-1}(\mathbf{y}-H\mathbf{x})\right).$$

Denote the posterior system state by $\hat{\mathbf{x}}$ instead of $\mathbf{x}|\mathbf{y}$. It can be shown by algebraic manipulations [1] that the posterior is also normal,

$$p(\hat{\mathbf{x}}) \propto \exp\left(-\frac{1}{2}(\hat{\mathbf{x}}-\hat{\mu})^{\mathrm{T}} P^{-1}(\hat{\mathbf{x}}-\hat{\mu})\right),$$

where the posterior mean $\hat{\mu}$ and covariance P are given by the update formulas

$$\hat{\mu} = \mu + K(\mathbf{y} - H\mu), \quad P = (I - KH)\, Q, \tag{2}$$

and

$$K = QH^{\mathrm{T}}\left(HQH^{\mathrm{T}} + R\right)^{-1} \tag{3}$$

is the Kalman gain matrix. The posterior mean $\hat{\mu}$ also equals the solution $\mathbf{x}$ of the least-squares problem

$$S(\mathbf{x}) = (\mathbf{x}-\mu)^{\mathrm{T}} Q^{-1}(\mathbf{x}-\mu) + (\mathbf{y}-H\mathbf{x})^{\mathrm{T}} R^{-1}(\mathbf{y}-H\mathbf{x}) \to \min. \tag{4}$$

cf., [27] or Appendix A.

A wide variety of researchers using the dynamic linear models have developed numerous modifications devised to make a Kalman filter like method suitable for problems of missing observations [30], irregular observation times [19], nonlinear updates in state or observations equations [14] and work to remove the Gaussian assumptions [24, 28, 6]. Johns and Shumway [18] modified the model in [6] to account for censored observations. Kalman filters combined with Tikhonov regularization to enforce smoothness where studied in [17, 23]. In spite of their immense utility and proven optimality in many cases, Kalman filters suffer from the drawback that they need to compute, store, and advance in time the covariance matrix of the state, and they are thus unsuitable for systems with a large number of degrees of freedom.

1.2 Ensemble Kalman Filter

Computational models of geophysical phenomenon typically rely on sophisticated dynamic, large, nonlinear systems of partial differential equations, which are solved numerically on a spatial mesh. Modeling considerations and demands for numerical accuracy typically result in state vectors of the order $10^5 - 10^7$ or larger; geoscientists always need as large model

as the current supercomputers can handle. The size of the data vector can be of the order $10^3 - 10^5$ or much more, e.g. for pixel data from airborne photographs. In particular, note that storage and computation of covariance matrices is not feasible for state vectors of this size. To accommodate this problem, ensemble Kalman filters were developed [10, 15] that represent the distribution of the system state using a random sample, called an ensemble. Each ensemble member represents a possible solution and is propagated in time by a model running independently up to analysis time. At analysis time, ensemble members are assumed to be a sample from the prior distribution, which is combined with the data likelihood, also assumed to be a normal, to create a sample from a distribution that approximates the posterior distribution. The benefit of the ensemble comes in situations where the ensemble members come from a distribution with high correlations such that a relatively small sample can reproduce the large-scale behavior of the covariance. This situation is typical of models of geophysical systems. Since the updated ensemble is a random sample from a distribution, moments can be estimated from the sample if needed. Furthermore, covariance matrices of conditional distributions need not be explicitly stored for further computations. Thus, the method is computationally feasible for nonlinear dynamic models with many parameters.

Particle filters (PF) [9] are related to EnKF but do not make the Gaussian assumption. They require much larger ensembles or shorter analysis cycles so that the prior and the posterior are not much different, and usually are applied to systems with few degrees of freedom. However, PF were used successfully in geophysical applications as well [33]. Other filters use combinations of Gaussian distributions to approximate more general distributions [2, 3] in attempt to avoid the assumption of normality and yet preserve the efficiency of the EnKF. Ensemble square root filters [32] were introduced to avoid random perturbations that are otherwise needed in the analysis step [5]. For more comprehensive surveys of related methods, see [11, 12].

The results in this paper were motivated by the observation that straightforward application of ensemble filters to a simple wildfire model [25] often fails within few steps: due to statistical variability of the ensemble members, the ignition temperature is exceeded in many members in many places, which starts secondary fires and the whole ensemble quickly ignites. Another motivation is that realistic models often have deterministic constraints on the state vectors, caused e.g. by redundant representation of the physical variables on multiple nested grids connected by interpolation operators [7, 22]. In this paper, we recognize that these are two sides of the same phenomenon: there are additional constraints on the system state, imposed either strongly or weakly, which are not accounted for in the classical state space model. For this reason, we investigate modifications of the Kalman filter paradigm by adding constraints by projection and penalty techniques.

The paper is organized as follows. In Sec. 2, we modify the constrained Kalman filter by loosening the constraint by penalty methods leading to a two stage Kalman filter. There we also observe that the filter can be realized as assimilation of the observation and then of the constraint, viewed as another independent observation. An ensemble form of the constrained filters is considered in Sec. 3, and results a simple fire model problem are presented in Sec. 4. Appendices contain the derivation of formulas for the constrained filters.

2 Two-Stage Kalman Filter

In applications, sometimes the state is constrained deterministically to a subspace. Such constraints may originate in the numerical methods used for the discrete state equation, such as interpolation relations between nested levels in complicated models [22]. One can then run the Kalman filter in a subspace, and, in particular, find the posterior mean by solving the least-squares problem (4) subject to the constraint. So, suppose that the state $\mathbf{x}$ is deterministically constrained so that $A(\mathbf{x} - \mu) = 0$ for a given matrix A. By enforcing the constraint $A(\mathbf{x} - \mu) = 0$ using Lagrange multipliers, the posterior mean is given by

$$\hat{\mu}_c = \hat{\mu} - PA^{\mathrm{T}}(APA^{\mathrm{T}})^{-1}A(\hat{\mu} - \mu), \tag{5}$$

where $\hat{\mu}$ and P are given by (2), cf., Appendix B. Geometrically, $\hat{\mu}_c$ is the projection of $\hat{\mu}$ onto the subspace given by $A(\mathbf{x} - \mu) = 0$, with respect to the the inner product given by P^{-1}. The covariance of the constrained conditional mean is given by

$$\begin{aligned} E\left[(\mathbf{x} - \hat{\mu}_c)(\mathbf{x} - \hat{\mu}_c)^T\right] &= Q - QH^TK^T(I - C)^T - (I - C)KHQ \\ &+ (I - C)QH^T(R + HQH^T)^{-1}HQ(I - C) \\ &= P + CQH^T(R + HQH^T)^{-1}HQC^T \end{aligned} \tag{6}$$

where $C = PA^T(APA^T)^{-1}A$. Clearly, the variance has increased from the Kalman filter solution as the constrained solution is not as flexible as the unconstrained solution.

The filter in (5), (6) strictly requires that the deviations of the state from the prior mean be in the nullspace of A. To relax the constraint, consider minimizing the penalized least squares

$$(\mathbf{x} - \mu)^{\mathrm{T}}Q^{-1}(\mathbf{x} - \mu) + (\mathbf{y} - H\mathbf{x})^{\mathrm{T}}R^{-1}(\mathbf{y} - H\mathbf{x}) + (\mathbf{x} - \mu)^{\mathrm{T}}A^{\mathrm{T}}D^{-1}A(\mathbf{x} - \mu) \to \min \tag{7}$$

over $\mathbf{x}$, where D is a given symmetric positive definite (covariance) matrix. The added term can be thought of as a variant of Tikhonov regularization for ill-conditioned systems [13]. Repeating twice the algebra used to derive (2), it is simple to show that the solution of (7) is

$$\hat{\mu}_p = \hat{\mu} + K^\star A(\mu - \hat{\mu})$$

where

$$K^\star = PA^{\mathrm{T}}\left(APA^{\mathrm{T}} + D\right)^{-1}$$

and $\hat{\mu}_1$ and P are as in (2), cf., Appendix C. Note that $\hat{\mu}_p \to \hat{\mu}_c$ as $D \to 0$.

Note that the repetition of the algebra to find $\hat{\mu}_p$ implies that the same analysis code can be simply called twice. The first call assimilates the observation difference $\mathbf{y} - H\mathbf{x}$ into the system state, while the second call will modify the state towards satisfying the constraint. An alternative interesting interpretation of the penalized filter (7) is assimilation of the two independent observations $H\mathbf{x} - \mathbf{y} \sim N(0, R)$ and $A\mathbf{x} - \mathbf{z} \sim N(0, D)$. It is a well known consequence of Bayes theorem (1), that two independent observations $\mathbf{y}$ and $\mathbf{z}$ can be assimilated independently without changing the result:

$$p(\mathbf{x}|\mathbf{y}, \mathbf{z}) \sim p(\mathbf{y}, \mathbf{z}|\mathbf{x})\, p(\mathbf{x}) = p(\mathbf{z}|\mathbf{x})\, p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}) = p(\mathbf{z}|\mathbf{x})\left[p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x})\right]. \tag{8}$$

So, the Kalman filter that corresponds to the penalized least squares problem (7) is obtained by applying the Kalman formulas (2) twice for the two different observations.

Since the assimilation of the two observations $\mathbf{y}$ and $\mathbf{z}$ is independent, the same posterior distribution is obtained by assimilating $\mathbf{z}$ first. Consequently, the proposed penalized Kalman filter is equivalent to a modification of the prior to impose weakly the constraint, and then applying the observation $\mathbf{y}$ in the usual manner.

3 Two-Stage Ensemble Kalman Filter

For the base ensemble Kalman filter, we follow the algorithm from [5] and note on efficient implementation using contemporary numerical software. Suppose that $\mathbf{X}$ is a p by n matrix whose columns are a random sample from the prior distribution for the state vector $\mathbf{x}$, which is assumed to be normal with covariance Q. Further replicate the observations $\mathbf{y}$ into matrix $\mathbf{Y}$ with d rows and n columns so that each column consists of the observed vector $\mathbf{y}$ plus a random vector from $N(0, R)$. Then it follows from (2) and (3) that the columns of

$$\hat{\mathbf{X}} = \mathbf{X} + K(\mathbf{Y} - H\mathbf{X})$$

form a random sample from the posterior distribution.

The ensemble filter involves two approximations. First, in the Kalman gain matrix $K = QH^{\mathrm{T}}\left(HQH^{\mathrm{T}} + R\right)^{-1}$, the state covariance Q is unknown, so it is replaced by the sample covariance computed from the ensemble members,

$$\hat{Q} = \frac{\mathbf{E}\mathbf{E}^{\mathrm{T}}}{n-1}, \quad \mathbf{E} = \mathbf{X} - \mathbf{X}\frac{\mathbf{e}_n\mathbf{e}_d^{\mathrm{T}}}{n},$$

where $\mathbf{e}_k$ is column vector of all ones of size k. This gives the matrix form of the analysis ensemble

$$\hat{\mathbf{X}} \approx \hat{\mathbf{X}}_{ens} = \mathbf{X} + \frac{\mathbf{E}(H\mathbf{E})^{\mathrm{T}}}{n-1}\left(\underbrace{\frac{(H\mathbf{E})(H\mathbf{E})^{\mathrm{T}}}{n-1} + R}_{B}\right)^{-1}(\mathbf{Y} - H\mathbf{X}). \tag{9}$$

The second approximation results from the fact that if the state evolution is nonlinear, then the prior distribution is not necessarily normal. Nevertheless, the distribution from which analysis ensemble members in $\hat{\mathbf{X}}_{ens}$ are drawn from is usually a good approximation of the posterior. Note that the dimension of $H\mathbf{E}$ is reasonably sized if the data dimension d is not too large and if the ensemble size n is reasonable, and that efficient Choleski decomposition of the $d \times d$ matrix B in (9) is possible because of the positive-definiteness of R. Also, the matrix B is well conditioned if the data covariance matrix R is well conditioned and relatively large compared to the other term in B; this will be often the case because the matrix R is diagonal when the observation errors are independent, and because the variability of the ensemble should be comparable to that of the data. In this case, iterative solution of the system with B may be often more suitable than decomposition. Also, the matrix-matrix product $H\mathbf{E}$ does not need to be computed directly, because

$$H\mathbf{E} = H\mathbf{X} - \frac{(\mathbf{H}\mu_{ens})\,\mathbf{e}_d^{\mathrm{T}}}{n}, \quad \mu_{ens} = \frac{\mathbf{X}\mathbf{e}_n}{n}, \tag{10}$$

which consists of the matrix-matrix product $H\mathbf{X}$, multiplication of the matrices $\mathbf{X}$ and $\mathbf{H}$ by a vector, and rank one update. The dominant operations are full matrix operations efficiently implemented in the Level 3 BLAS routines [8], and also readily parallelizable. In many cases, computation is further reduced as H may be sparse since a component of $\mathbf{y}$ represents a characteristic of the state vector at a single point or the integrated value over some small region. In addition, if the observation errors are independent, the computation may be split into assimilating a part of the observations at a time sequentially as in (8), thus reducing the data size d that needs to be considered in any one computation.

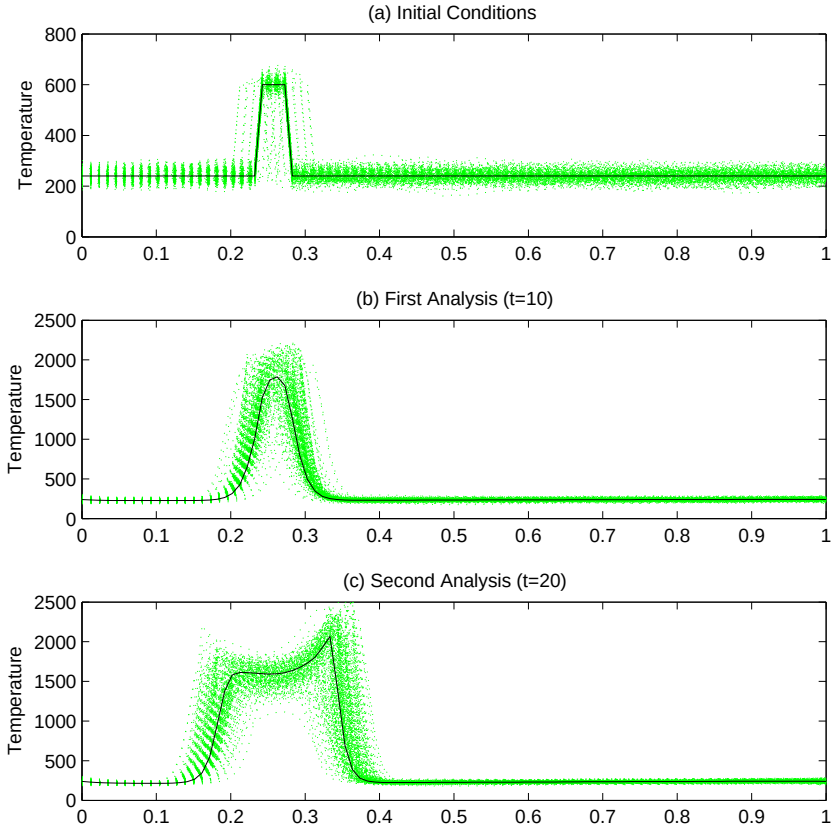

Figure 1: Panel (a) shows the temperature profiles of the reference solution (solid, dark line) as well as initial distribution of ensemble members (dotted, light lines) for the fire model. Panels (b) and (c) are similar and show the effect of forward propagation on the reference solution (solid dark line) as well as the ensemble members (dotted, light lines) to the first (t=10) and second (t=20) analysis times, respectively.

The two-stage Kalman filter is implemented in ensemble form simply by applying the ensemble filter formulas twice, assimilating first the observation $H\mathbf{x}-\mathbf{y} \sim N(0,R)$ and then the constraint as another, independent observation

$$A\mathbf{x} - \mathbf{z} \sim N(0, D), \tag{11}$$

where $\mathbf{z} = A\mu_{ens}$ with μ_{ens} being the forecast ensemble mean from (10).

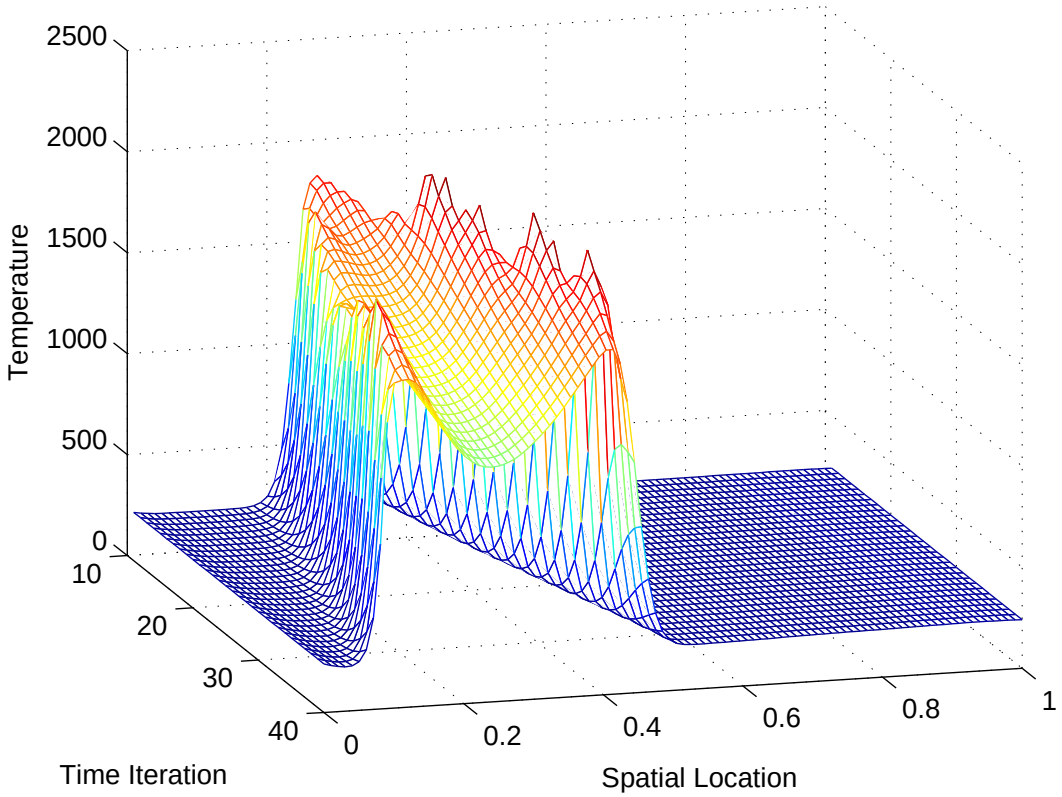

Figure 2: The reference solution of temperature for the fire model over 30 time iterations starting from the first analysis time period (t=10).

4 The Fire Model

A simple example useful for demonstrating the utility of the constrained algorithm proposed above is the simplified model of wildfire by a reaction-convection-diffusion equation [26],

$$\frac{\partial T}{\partial t} = -\nabla \cdot (k\nabla T) - c_1 \cdot \nabla T - c_2(T - T_a) + c_3 \frac{\partial S}{\partial t}$$
$$\frac{\partial S}{\partial t} = -c_4 \max\{0, T - T_i\}^{\alpha}$$

The first equation is the heat balance, where T is the temperature, $-\nabla \cdot (k\nabla T)$ is the diffusion of heat, $-c_1 \cdot \nabla T$ is the heat transport by wind, $-c_2(T - T_a)$ is the heat escaping to the environment with the ambient temperature T_a, and $c_3 \frac{\partial S}{\partial t}$ is the heat generated by burning. The second equation models the fuel supply; the right-hand side is the intensity of burning. This is a very simplified model and we do not use any physical units, yet it appears to capture some essential qualitative fire behavior. The MATLAB code is available from `http://www-math.cudenver.edu/~cjohns/fire1d`.

In this paper, we consider the differential equation in one spatial dimension. The model has ignition point of 300 temperature units and includes a fuel break at spatial locations between 0.45 and 0.50. The reference solution profile of the initial conditions for temperature

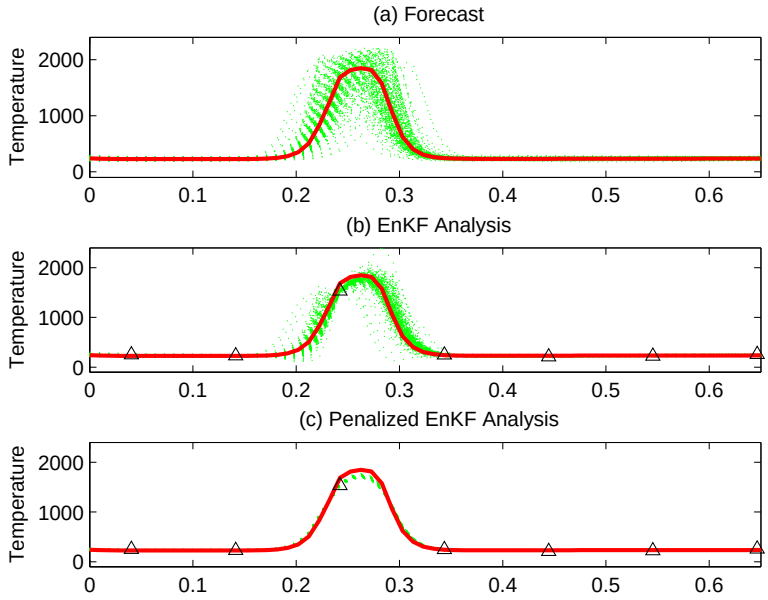

Figure 3: Panel (a) shows the temperature profiles of 250 ensemble members (dotted lines) forecast to the first analysis cycle (10 time iterations from initial conditions) and the corresponding reference solution (solid line). Panel (b) shows the temperature profiles (dotted lines) in the analysis ensemble corresponding to the EnKF update and includes the information in the data (△). Panel (c) compares the penalized analysis ensemble with the reference solution where the second analysis smooths the temperature via an approximate spatial derivative constraint.

is shown in the top panel of Figure 1. Panels (b) and (c) show how temperature profile of the reference solution propagates to time periods 0.05 and further to 0.10.

An initial ensemble of size 250 was generated by perturbing the reference solution temperature and fuel supply values with spatially correlated normal deviates. Six of the ensemble members were generated by shifting the reference solution right and left spatially, and then adding perturbations. Each ensemble member is propagated forward ten time steps via the fire model to reach the time period shown in Figure 2 to generate the forecast or prior ensemble. Panel (a) of Figure 3 shows the 250 members of the forecast ensemble (dotted lines) and the associated true temperature profile (solid). Data (△) collected every 10 spatial units are shown in Panel (b) along with the ensemble (dotted lines) updated to the EnKF analysis stage. Note that the analysis step does a good job of cinching the temperature profiles to the few observations; however, the data information does not carry over to nearby locations (e.g. the analysis ensemble near 0.20 or 0.30 on the x-axis in panel (b).) This cinching phenomenon is a byproduct of the least-squares approach in the filter and can produce temperature profiles not compatible with common sense, theory, and numerical approximation schemes.

To force a measure of smoothness upon the analysis, we impose an approximate spatial first derivative constraint on the temperature analysis field using a second application of the EnKF code. The penalized constraint analysis field is shown in panel (c) of Figure 3 where the "data" for the second application of the the EnKF code penalizes ensemble members

that deviate from $\mathbf{z} = A\mu_{ens}$ and $A = [A^\star, 0]$

$$A^\star = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 \\ \vdots & & & & \vdots & \\ 0 & \dots & & 0 & 1 & -1 \\ & & & & & \end{bmatrix} \tag{12}$$

and 0 is a matrix of zeros of the correct dimension so that A only works on the temperature values in $\hat{X}$. The intent of the constraint is to force temperature values that are neighboring spatially to be nearly equivalent, subject to the covariance in D. The effect of pre-multiplication by A can be loosely interpreted as a spatial derivative for small mesh sizes in the spatial domain. Specifically, for $i, i+1$ indexing spatially adjacent temperature values and t indexing time period, the (i,t) element of $A(\hat{X} - \bar{x})$ is $(\hat{X}_{i,t} - \hat{\mu}_i) - (\hat{X}_{i+1,t} - \hat{\mu}_{i+1})$. The penalty for deviation from the approximate first order smoothness is given by the diagonal variance matrix D as in (7). In this example, D is diagonal with non-zero elements $d_{ii} = |z_i|/2$. That is, the variance expected in the smoothness constraint for temperature is proportional to the magnitude of the average prior ensemble smoothness in temperature. Panel (c) of Figure 3 demonstrates the effect of the smoothing step carried out by running EnKF code again with the "data" being the smoothness constraints.

We compared the forecasts from the EnKF analysis (A) and the constrained analysis (C). The ensembles shown in panels (b) and (c) of Figure 3 were propagated forward 30 time steps and the pointwise squared errors (MSE) from the reference solution were averaged over each spatial location to simulate a prediction. The log of the ratio of MSE's for the two sets of predictions, MSE_A/MSE_C, is shown in Figure 4. In every case, the constrained predictions had a smaller MSE value than those based on the unconstrained EnKF analysis ensemble. Because the reference solution is the basis for this MSE analysis, the effects of bias are included in the calculations via the usual $bias^2 + variance$ formula. A graphical comparison of the bias showed that the bias for both methods were of the same magnitudes and generally smaller for the constrained method.

5 Conclusion

As mentioned in Sec. 2, including the penalty step with the EnKF can be considered changing the prior distribution. When housed in the setting of imposing further restraints on the model, the second step is easy to justify as adding information in a similar vein to the information added by the data likelihood. Furthermore, in many modeling situations, the interest is on the forward integration of the states for the purposes of prediction. The updated states under the usual Kalman filter are not required to be smooth and the lack of smoothness may make forward time computation numerically impossible. In particular, some least squares solutions where data with small variance is measured in a spatio-temporal region where the variance of the forecast is large (cf. Figure 3), the potential exists and indeed occurred in some of our simulations. This scenario is particularly likely when the basin of attraction for the model is narrow compared with the deviations in the prior distribution.

For the matrices A and D in the second step, several possibilities show promise. Where possible, known constraints imposed by a physical system or numerical approximation scheme will be beneficial to driving the behavior of the model and should be used. When such constraints are not available, using a form of A that represents spatial derivatives

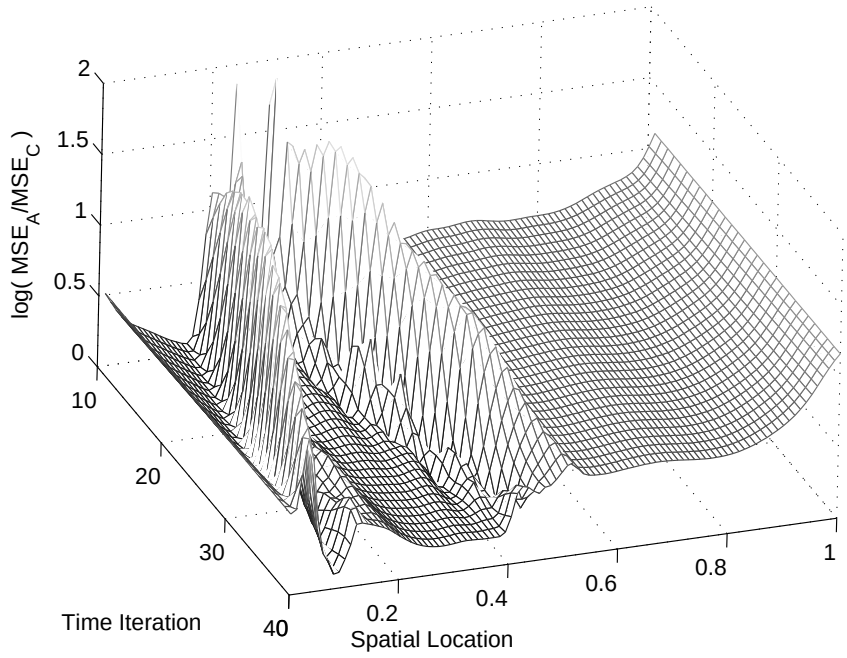

Figure 4: Log of the ratio of Mean-squared error (with respect to reference solution) of the analysis ensemble (MSE_A) and constrained analysis ensemble (MSE_C). Values greater than zero support the use of the two-stage shrinkage method.

provides sparse forms for A when the states represent variables of interest on a grid. These derivative forms also give a physical meaning to the linear constraints as well as some direction for building the magnitudes of variances in D for models with physical dimensions. In practice, it will be simplest to force D to be a multiple of the identity matrix. In small systems with fast updates, the effect of multiple forms of D can be considered similar to choosing a smoothing parameter in a nonparametric regression setting.

Acknowledgements

This research was supported by the National Science Foundation under grants 0325314 and 0417909.

References

[1] B. Anderson and J. Moore, *Optimal Filtering*, Prentice Hall, Englewood Cliffs, NJ, 1979.

[2] J. L. Anderson and S. L. Anderson, *A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts*, Monthly Weather Review, 127 (1999), pp. 2741–2758.

[3] T. Bengtsson, C. Snyder, and D. Nychka, *Toward a nonlinear ensemble filter for high dimensional systems*, J. of Geophysical Research - Atmospheres, 108(D24) (2003), pp. STS 2–1–10.

[4] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, Springer-Verlag, New York, second ed., 1996.

[5] G. Burgers, P. J. van Leeuwen, and G. Evensen, *Analysis scheme in the ensemble Kalman filter*, Monthly Weather Review, 126 (1998), pp. 1719–1724.

[6] J. B. Carlin, N. Polson, and D. S. Stoffer, *A Monte Carlo approach to nonnormal and nonlinear state-space modeling*, Journal of the American Statistical Association, (1992), pp. 494–500.

[7] T. L. Clark and R. D. Farley, *Severe downslope windstorm calculations in two and three spatial dimensions using anelastic interactive grid nesting: A possible mechanism for gustiness*, J. of the Atmospheric Sciences, 41 (1984), pp. 329–350.

[8] J. J. Dongarra, J. D. Croz, S. Hammarling, and I. Duff, *A set of level 3 Basic Linear Algebra Subprograms*, ACM Transactions on Mathematical Software, 16 (1990), pp. 1–17.

[9] A. Doucet, N. de Freitas, and N. Gordon, eds., *Sequential Monte Carlo in Practice*, Springer, 2001.

[10] G. Evensen, *Sequential data assimilation with nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, J. Geophys. Res., 99 (C5) (1994), pp. 143–162.

[11] G. Evensen, *The ensemble Kalman filter: Theoretical formulation and practical implementation*, Ocean Dynamics, 53 (2003), pp. 343–367.

[12] ——, *Sampling strategies and square root analysis schemes for the EnKF*, Ocean Dynamics, (2004), pp. 539–560.

[13] P. C. Hansen, *Rank-deficient and discrete ill-posed problems*, SIAM Monographs on Mathematical Modeling and Computation, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.

[14] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge, 1989.

[15] P. Houtekamer and H. L. Mitchell, *Data assimilation using an ensemble Kalman filter technique*, Monthly Weather Review, 126 (1998), pp. 796–811.

[16] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, 1970.

[17] T. A. Johansen, *On Tikhonov regularization, bias and variance in nonlinear system identification*, Automatica, 33 (1997), pp. 441–446.

[18] C. J. Johns and R. H. Shumway, *A state-space model for censored data*, Environmetrics, (2005), pp. 167–180.

[19] R. H. Jones, *Fitting multivariate models to unequally spaced data*, in Time Series Analysis of Irregularly Spaced Data, E. Parzen, ed., no. 25 in Lecture Notes in Statistics, Springer-Verlag, 1984, pp. 158–188.

[20] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. ASME J. Basic Eng., 82 (1960), pp. 35–45.

[21] R. E. Kalman and R. S. Bucy, *New results in filtering and prediction theory*, Trans. ASME J. Basic Eng., 83 (1961), pp. 95–108.

[22] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, 2003.

[23] K. Y. Kim, S. I. Kang, M. C. Kim, S. Kim, Y. J. Lee, and M. Vauhkonen, *Dynamic image reconstruction in electrical impedance tomography with known internal structures*, IEEE Transactions on Magnetics, 38 (2002), pp. 1301–1304.

[24] G. Kitagawa, *Non-Gaussian state-space modeling of nonstationary time series, (with discussion)*, Journal of the American Statistical Association, 400 (1987), pp. 1032–1044.

[25] J. Mandel, M. Chen, L. P. Franca, C. Johns, A. Puhalskii, J. L. Coen, C. C. Douglas, R. Kremens, A. Vodacek, and W. Zhao, *A note on dynamic data driven wildfire modeling*, in Computational Science - ICCS 2004, M. Bubak, G. D. van Albada, P. M. A. Sloot, and J. J. Dongarra, eds., vol. 3038 of Lecture Notes in Computer Science, Springer, 2004, pp. 725–731.

[26] ——, *Dynamic data driven wildfire modeling*, in Dynamic Data Driver Application System, F. Darema, ed., Kluwer, 2005, p. in print. UCD CCM Report 208, March 2004, http://www-math.cudenver.edu/ccm/reports/rep208.pdf.

[27] R. Meinhold and N. Singpurwalla, *Understanding the Kalman filter*, American Statistician, 37 (1983), pp. 123–127.

[28] ——, *Robustification of Kalman filter models*, Journal of the American Statistical Society, 84 (1989), pp. 479–486.

[29] C. C. Paige and M. A. Saunders, *Least squares estimation of discrete linear dynamic systems using orthogonal transformations*, SIAM J. Numer. Anal., 14 (1977), pp. 180–193.

[30] R. H. Shumway and D. S. Stoffer, *An approach to time series smoothing and forecasting using the EM algorithm*, Journal of Time Series Analsysis, 3 (1982), pp. 253–264.

[31] ——, *Time Series Analysis*, Springer-Verlag, New York, NY., 2000.

[32] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, *Ensemble square root filters*, Monthly Weather Review, 131 (2003), pp. 1485–1490.

[33] P. van Leeuwen, *A variance-minimizing filter for large-scale applications*, Monthly Weather Review, 131 (2003), pp. 2071–2084.

A The Least-squares View of a Kalman Filter

This Section is standard and we include it only for completeness as the basis of developments further on. From an optimization standpoint, the Kalman recursion for the state can be formulated as a least-squares solution of the inconsistent linear system

$$\mathbf{x} = \mu, \quad \mathbf{y} = H\mathbf{x}$$

for the unknown $\mathbf{x}$, where the least squares norms are given by the inverses of the covariance matrices. Hence, $\mathbf{x}$ is determined as the solution of the unconstrained minimization problem

$$S_u(\mathbf{x}) = (\mathbf{x}-\mu)^{\mathrm{T}} Q^{-1}(\mathbf{x}-\mu) + (\mathbf{y}-H\mathbf{x})^{\mathrm{T}} R^{-1}(\mathbf{y}-H\mathbf{x}) \to \min,$$

which is equivalent to

$$\nabla S_u(\mathbf{x}) = 2Q^{-1}(\mathbf{x}-\mu) - 2H^{\mathrm{T}} R^{-1}(\mathbf{y}-H\mathbf{x}) = 0,$$

which is in turn gives

$$\mathbf{x} = \hat{\mathbf{x}}_u = P(Q^{-1}\mu + H^{\mathrm{T}} R^{-1}\mathbf{y})$$

where

$$\begin{aligned} P &= (Q^{-1} + H^{\mathrm{T}} R^{-1} H)^{-1} = \left[Q - QH^{\mathrm{T}}(HQH^{\mathrm{T}} + R)^{-1} HQ\right] \\ &= (I - KH)Q \end{aligned}$$

with K equal to the Kalman gain as in equation (3). Consequently,

$$\begin{aligned} \hat{\mathbf{x}}_u &= P(Q^{-1}\mu + H^{\mathrm{T}} R^{-1}\mathbf{y}) \\ &= \mu + K(\mathbf{y} - H\mathbf{x}). \end{aligned} \tag{13}$$

Note that P is the posterior covariance from (2).

B Constrained Kalman Filter

If the state $\mathbf{x}$ is constrained so that $A(\mathbf{x}-\mu) = 0$ for some matrix A, the updated state $\hat{\mathbf{x}}$ is the solution of the constrained problem

$$\begin{aligned} &S_c(\mathbf{x}) = (\mathbf{x}-\mu)^{\mathrm{T}} Q^{-1}(\mathbf{x}-\mu) + (\mathbf{y}-H\mathbf{x})^{\mathrm{T}} R^{-1}(\mathbf{y}-H\mathbf{x}) \to \min, \\ &\text{subject to } A(\mathbf{x}-\mu) = 0. \end{aligned}$$

Introducing Lagrange multipliers, this problem is equivalent to the linear algebraic system

$$2Q^{-1}(\mathbf{x}-\mu) - 2H^{\mathrm{T}} R^{-1}(\mathbf{y}-H\mathbf{x}) + A^{\mathrm{T}}\lambda = \mathbf{0} \tag{14}$$

$$A\mathbf{x} = A\mu \tag{15}$$

Solving for $\mathbf{x}$ from (14) gives

$$(Q^{-1} + H^{\mathrm{T}} R^{-1} H)\mathbf{x} = H^{\mathrm{T}} R^{-1}(\mathbf{y}-H\mathbf{x}) + \frac{1}{2} A^{\mathrm{T}}\lambda$$

so

$$\begin{aligned}\mathbf{x} &= P(Q^{-1}\mu + H^{\mathrm{T}}R^{-1}\mathbf{y}) - \frac{1}{2}PA^{\mathrm{T}}\lambda \\ &= \hat{\mathbf{x}}_u - \frac{1}{2}PA^{\mathrm{T}}\lambda\end{aligned}$$

where $\hat{\mathbf{x}}_u$ is the unconstrained solution from (13) and K is again from (3). Substituting to (15) gives

$$\lambda = \lambda_c = 2(APA^{\mathrm{T}})^{-1}A(\hat{\mathbf{x}}_1 - \mu),$$

hence

$$\begin{aligned}\mathbf{x} = \hat{\mathbf{x}}_c &= \hat{\mathbf{x}}_u - PA^{\mathrm{T}}(APA^{\mathrm{T}})^{-1}A(\hat{\mathbf{x}}_u - \mu) \\ &= \left(I - PA^{\mathrm{T}}(APA^{\mathrm{T}})^{-1}A\right)(\hat{\mathbf{x}}_u - \mu) + \mu.\end{aligned}$$

Note that $I - PA^{\mathrm{T}}(APA^{\mathrm{T}})^{-1}A$ is P^{-1}-orthogonal projection onto the nullspace of A.

C Penalized Kalman Filter

Changing the constraint to a penalty, one will minimize

$$S_p(\mathbf{x}) = (\mathbf{x}-\mu)^{\mathrm{T}}Q^{-1}(\mathbf{x}-\mu) + (\mathbf{y}-H\mathbf{x})^{\mathrm{T}}R^{-1}(\mathbf{y}-H\mathbf{x}) + (A\mathbf{x}-A\mu)^{\mathrm{T}}D^{-1}(A\mathbf{x}-A\mu) \to \min,$$

which gives

$$\nabla S_3(\mathbf{x}) = 2Q^{-1}(\mathbf{x}-\mu) - 2H^{\mathrm{T}}R^{-1}(\mathbf{y}-H\mathbf{x}) - 2A^{\mathrm{T}}D^{-1}A(\mathbf{x}-\mu) = 0,$$

which is equivalent to

$$\left(Q^{-1} + H^{\mathrm{T}}R^{-1}H + A^{\mathrm{T}}D^{-1}A\right)\mathbf{x} = Q^{-1}\mu + H^{\mathrm{T}}R^{-1}\mathbf{y} + A^{\mathrm{T}}D^{-1}A\mu$$

and so

$$\begin{aligned}\hat{\mathbf{x}}_p &= (R^{-1} + H^{\mathrm{T}}Q^{-1}H + A^{\mathrm{T}}D^{-1}A)^{-1}(R^{-1}\mu + H^{\mathrm{T}}Q^{-1}\mathbf{y} + A^{\mathrm{T}}D^{-1}A\mu) \\ &= (P^{-1} + A^{\mathrm{T}}D^{-1}A)^{-1}\left[R^{-1}\mu + H^{\mathrm{T}}Q^{-1}\mathbf{y} - P^{-1}\mu + \left(P^{-1} + A^{\mathrm{T}}D^{-1}A\right)\mu\right] \\ &= \left(P - \underbrace{PA^{\mathrm{T}}(APA^{\mathrm{T}} + D)^{-1}}_{K^*}AP\right)(R^{-1}\mu + H^{\mathrm{T}}Q^{-1}\mathbf{y} - P^{-1}\mu) + \mu \\ &= (I - K^*A)\,(\underbrace{P\left(R^{-1}\mu + H^{\mathrm{T}}Q^{-1}\mathbf{y}\right)}_{\hat{\mathbf{x}}_u} - \mu) + \mu \\ &= (I - K^*A)\,\mathbf{x}_u + K^*A\mu,\end{aligned}$$

that is

$$\hat{\mathbf{x}}_p = \mathbf{x}_u + K^*A\left(\mu - \mathbf{x}_u\right), \quad K^* = PA^{\mathrm{T}}\left(APA^{\mathrm{T}} + D\right)^{-1}.$$