Data Assimilation

Douglas W. Nychka and Jeffrey A. Anderson
National Center for Atmospheric Research
P.O. Box 3000, Boulder, CO 80307

June 20, 2008

Data assimilation refers to the statistical techniques used to combine numerical models with observations to give an improved understanding of the state of a system or process. Typically a data assimilation problem has a sequential aspect where data coming in over time is being used to continually update the state or parameters of a dynamical system. It is usually distinguished from more traditional statistical time series applications because the system may have complicated nonlinear dynamical behavior and the state vector and observations may be large. One of its primary roles is in estimating the state of a physical process when applied to geophysical models and physical measurements. Data assimilation has its roots in Bayesian inference and the restriction to linear dynamics and Gaussian distributions fits within the methods associated with the Kalman filter. Because data assimilation also involves estimating an unknown state based on possibly irregular, noisy or indirect observations it it also has an interpretation as solving an *inverse* problem (e.g. [24]). One goal of this article is to tie these concepts back to a general Bayesian framework.

One of the most successful applications of data assimilation is in numerical weather prediction where a large and heterogeneous set of observations are combined with a sophisticated physical model for the evolution of the atmosphere to produce detailed and high resolution forecasts of weather (see e.g. [17] for an introduction). The application to weather forecasting and in general to assimilation of atmospheric and oceanographic observations has a distinctly spatial aspect as the processes of interest are typically two and three dimensional fields. For this reason it is important to included this topic in this handbook. Although there are other applications of assimilation, such as target tracking or process control, and a more general class of Bayesian filtering methods (see [27], [6]) such as particle filters, these topics tend not to emphasize spatial processes and so are not as relevant to this handbook.

Spatial methods in data assimilation typically involve non-Gaussian fields and infer the spatial structure dynamically from a physical model. In this way the dynamical model and a statistical model are connected more closely than in a standard application of spatial statistics. In addition the sheer size of data assimilation problems requires approximate solutions that are not typical for smaller spatial data sets. In this article these differences will be highlighted by reviewing the principles behind current methods. We also point out some new areas where more standard space-time statistical models might be helpful in handling model error. A large scale example for the global atmosphere is included at the end this article to illustrate some of the details in practical data assimilation for atmospheric prediction.

1 Bayesian formulation of data assimilation

The basic ingredients of a data assimilation problem are the state vector giving a complete description of the system or process and a vector of observations made on the system. In addition one requires conditional distributions for propagating the system forward in time and for relating the observations to the system state. This overall organization is similar in concept to a Bayesian Hierarchical model (BHM) [26] and is known as a state space formulation in the context of Kalman filtering. In either interpretation we have a observation level that relates the observed data to the (unobserved) state of the system and a supporting, process level describing the evolution of the state over time. Throughout this discussion we also point out some parallel terminology from the geosciences. In particular the data assimilation process is separated into an *update* or *analysis* step and a *forecast* step. This distinction is helpful because most of the spatial statistical content is in the update step. In geoscience applications the state vector is usually a multivariate spatial field on a fine regular grid with observations being irregular and noisy measurements of some components. Typically the forecast step is completed by a deterministic, physically based model.

1.1 Bayes theorem

To introduce the Bayesian statistical model we will use a bracket notation for distributions where square braces denote a probability density function (pdf). Accordingly, $[Y]$ is the pdf for the random variable Y and $[Y|X]$ the conditional pdf for the random variable Y given the random variable X. Let $\boldsymbol{x}_t$ denote the state vector at time t and $\boldsymbol{y}_t$ be a vector of observations made at time t. To streamline the exposition assume that the times are indexed by integers, $t = 1, 2, \ldots$ although handling unequally spaced times usually does not add much difficulty. We will assume the likelihood: $[\boldsymbol{y}_t|\boldsymbol{x}_t]$ and a reference or prior distribution for the state $[\boldsymbol{x}_t]$. In the geosciences $[\boldsymbol{x}_t]$ is also know as the *forecast* distribution because it has been derived from forecasting the state based on data at prior times. The joint distribution of observations and the state is the product $[\boldsymbol{y}_t|\boldsymbol{x}_t][\boldsymbol{x}_t]$ and by Bayes theorem:

$$[\boldsymbol{x}_t|\boldsymbol{y}_t] = \frac{[\boldsymbol{y}_t|\boldsymbol{x}_t][\boldsymbol{x}_t]}{[\boldsymbol{y}_t]} \tag{1}$$

1.2 The update step

The first part of the assimilation process is to apply Bayes theorem (1) to obtain the posterior, $[\boldsymbol{x}_t|\boldsymbol{y}_t]$. In words, this is the conditional distribution of the system state *given* the observations at time t. The prior distribution for the state has been updated in light of information provided by new observations. This result is the *analysis* pdf in a geoscience context. Here the term analysis originates from the analyzed fields used to assess the current state of the atmosphere for weather forecasting. Although (1) is strictly an Bayesian formulation it should be noted that if the prior, $[\boldsymbol{x}_t]$, has a frequency interpretation then the analysis will also have frequentist content. Typically practitioners are concerned about the skill of the assimilation and so make a direct comparison between the center and spread of the analysis density and the true state of the system. We will illustrate this point of view in the atmospheric example at the end of this article.

1.3 The forecast step

The second part of the assimilation process is to make a forecast at a future time, e.g. $t+1$. In general, we assume that the dynamics of the state process are known and can be abstracted as: $[\boldsymbol{x}_{t+1}|\boldsymbol{x}_t]$. This Markov property implies that future states of the system only depend on the current state and is appropriate for many physical processes. The forecast distribution is then

$$[\boldsymbol{x}_{t+1}|\boldsymbol{y}_t] = \int [\boldsymbol{x}_{t+1}|\boldsymbol{x}_t][\boldsymbol{x}_t|\boldsymbol{y}_t]d\boldsymbol{x}_t \tag{2}$$

The mean of this distribution could be used as a point forecast of the state and the pdf quantifies the uncertainty in the forecast.

1.4 The assimilation cycle

At this point we have come full circle in the assimilation cycle. Confronted with new observations at time $t+1$, say $\boldsymbol{y}_{t+1}$ one just identifies $[\boldsymbol{x}_{t+1}]$ with the forecast pdf and applies Bayes theorem. Update and forecast steps are repeated as time advances and new observations arrive. An important concept to draw from this process is that spatial information about the distribution of $\boldsymbol{x}_t$ can be generated in (2) from the dynamics of the process. This inheritance is explicit in considering the special case of the Kalman filter linear equations. Although one needs to prescribe a spatial prior explicitly for $\boldsymbol{x}_1$ in the first update often this initial information is discounted by subsequent update/forecast cycles with more observations.

One subtly in this process is the assumption that the prior contains all information about past observations before time t. Equivalently we are assuming that if the observations at time t are conditioned on the state at time t then they are independent. In bracket notation,

$$[\boldsymbol{y}_t, \boldsymbol{y}_s|\boldsymbol{x}_t, \boldsymbol{x}_s] = [\boldsymbol{y}_t|\boldsymbol{x}_t][\boldsymbol{y}_s|\boldsymbol{x}_s] \tag{3}$$

. Given this conditional independence and updates found by Bayes theorem, the sequential assimilation process outlined above will result in a posterior that is the conditional distribution of the current state based on *all* past observations.

Throughout this discussion we have take the perspective of numerical weather forecasting where one is interested in future predictions of the state and does not have observations at later times to update $\boldsymbol{x}_t$. However, in a retrospective analysis, one would use past, present and future observations with respect to time t for updating the state $\boldsymbol{x}_t$. This process is termed *smoothing* as opposed to forecasting which is termed $filtering$. Given the Markov assumptions for propagating the state and conditional independence for the observations, the Bayesian computation for smoothing simplifies to a forward pass through the observation sequence followed by backward pass through the sequence updating the $[\boldsymbol{x}_t|\boldsymbol{y}_t]$ with $\boldsymbol{y}_{t+1}$. For large geophysical problems, the computational costs for exact smoothing are often prohibitive and approximations are needed. We suggest one possible approximate smoother as an extension of the ensemble Kalman filter in Section 3.

1.5 Sequential updating

Up to now we have assumed the full vector of observations is assimilated in a single application of Bayes Theorem. A important feature of this problem is that the update can be performed sequentially on the components of $\boldsymbol{y}_t$ provided that the components are conditionally independent

given $\boldsymbol{x}_t$ and that the posterior is computed exactly. To make this explicit, generically split the observation vector into two parts $\boldsymbol{y}_t = (Y^{(1)}, Y^{(2)})$ and we assume conditional independence:

$$[\boldsymbol{y}_t|\boldsymbol{x}_t] = [Y^{(1)}, Y^{(2)}|\boldsymbol{x}_t] = [Y^{(1)}|\boldsymbol{x}_t][Y^{(2)}|\boldsymbol{x}_t].$$

The full posterior for the update can be rewritten as

$$[\boldsymbol{x}_t|\boldsymbol{y}_t] \propto [\boldsymbol{y}_t|\boldsymbol{x}_t][\boldsymbol{x}_t] \propto [Y^{(2)}|\boldsymbol{x}_t][Y^{(1)}|\boldsymbol{x}_t][\boldsymbol{x}_t] \propto [Y^{(2)}|\boldsymbol{x}_t]\left([Y^{(1)}|\boldsymbol{x}_t][\boldsymbol{x}_t]\right) \propto [Y^{(2)}|\boldsymbol{x}_t][\boldsymbol{x}_t|Y^{(1)}]$$

In words this string of proportions indicates the full posterior can be obtained by first finding the posterior by updating with respect to $Y^{(1)}$ and then using this intermediate result as a subsequent prior to update with respect to $Y^{(2)}$. Since $Y^{(1)}$ and $Y^{(2)}$ are an arbitrary split of $\boldsymbol{y}_t$ by induction one can show rigorously that the full posterior can be found by updating with each component of $\boldsymbol{y}_t$ sequentially. Each update can involve a scalar observation and the posterior computation can simplify greatly when this is done. It is important to emphasize that a sequential update is only valid under conditional independence among the components of the observation vector. But if this holds the order of the sequential updating does not matter. An intriguing connection with spatial statistics is that this same sequential result can be used for Bayesian spatial prediction and there is the potential to transfer the efficient parallel algorithms for this described in the next section to find approximate solutions for large spatial problems.

2 The Kalman Filter and assimilation

The Kalman filter (KF) was first developed by Kalman and Bucy [15], [16] in an engineering context and as a linear filter. Although the KF can be interpreted as a optimal linear estimator, to streamline this discussion, we will add the assumption of joint Gaussian distributions to fit into the Bayesian paradigm given above.(See [14] for more background.)

2.1 The KF update step

Assume that $[\boldsymbol{y}_t|\boldsymbol{x}_t]$ is multivariate normal with mean vector $H\boldsymbol{x}_t$ and covariance Σ_o. H is a known matrix hat maps the state into the expected value of the observations and Σ_o is the observation error covariance matrix. (Both H and Σ_o can depend on t although we will not add this additional index.) This conditional distribution can also be represented as

$$\boldsymbol{y}_t = H\boldsymbol{x}_t + \boldsymbol{e}_t \tag{4}$$

where now $\boldsymbol{e}_t \sim N(0, \Sigma_o)$. As mentioned above, a subtle point assumed here is that $\boldsymbol{e}_t$ are uncorrelated over time and independent of the state. Without loss of generality Σ_o can be assumed to be diagonal by redefining $\boldsymbol{y}$ and H through a linear transformation. We will further assume that $[\boldsymbol{x}_t] \sim N(\boldsymbol{\mu}_f, \Sigma_f)$ where "f" indicates this is the forecast distribution. The update step yields a distribution that is again multivariate normal $[\boldsymbol{x}_t|\boldsymbol{y}_t] \sim N(\boldsymbol{\mu}_a, \Sigma_a)$ Here the "a" indicates the analysis distribution and the mean and covariance are

$$\boldsymbol{\mu}_a = \boldsymbol{\mu}_f + \left[\Sigma_f H^T (H\Sigma_f H^T + \Sigma_o)^{-1}\right](\boldsymbol{y}_t - H\boldsymbol{\mu}_f) \tag{5}$$

and

$$\Sigma_a = \Sigma_f - \Sigma_f H^T (H\Sigma_f H^T + \Sigma_o)^{-1} H\Sigma_f \tag{6}$$

The matrix expression in square brackets in (5) is the Kalman gain and transforms an difference between the observation vector and its expectation with respect to the forecast distribution into an adjustment to the state. To derive these expressions note that $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are jointly distributed multivariate normal and (5) and (6) can derived from the properties of the conditional multivariate normal.

Equation (5) and (6) are the same equations for a spatial conditional inference. Interpreting the state vector as being values of a spatial field on a regular grid, $\boldsymbol{\mu}_a$ is the conditional mean for $\boldsymbol{x}$ given the observations and Σ_a the conditional covariance matrix. Here one interprets the "prior" $N(\boldsymbol{\mu}_f, \Sigma_f)$ as a Gaussian process model with Σ_f being constructed from a spatial covariance function. The important distinction for data assimilation is that Σ_f is generated by the process not from an external spatial model. This becomes clear in examining the forecast step for the linear KF.

2.2 The KF forecast step

For the forecast step for the KF assume that the process evolves in a linear way from t to $t+1$, possibly with an additive Gaussian random component

$$\boldsymbol{x}_{t+1} = L\boldsymbol{x}_t + \boldsymbol{u}_t \tag{7}$$

Here $\boldsymbol{u}_t \sim N(0, \Sigma_m)$ independent of $\boldsymbol{x}_t$ and L is a matrix. Both L and Σ_m can depend on t. Based on all these assumptions it is straightforward to conclude that the forecast pdf is $N(L\boldsymbol{\mu}_a, L\Sigma_a L^T + \Sigma_m)$. Scrutinizing the forecast covariance matrix one sees that $L\Sigma_a L^T$ will be based on the previous forecast covariance matrix appearing in (6) and will also inherent the dynamical relationship from the previous time. Thus, in the situation of assimilation for a space-time process the spatial covariance for inference is built up sequentially based on past updates with observations and propagating the posterior forward in time as a forecast distribution. It is important to realize that this spatial information is the *error* between the conditional mean and the true field and is not the covariance of the process itself. For example if the observations are both dense and accurate and Σ_m is small the forecast covariance can be much smaller and have less structure than the covariance for $\boldsymbol{x}_t$ itself.

2.3 Sequential updates

In the previous section it was stated that the update can be done sequentially under conditional independence of the observations. It is helpful to describe how the KF update simplifies when the components of $\boldsymbol{y}_t$ are considered sequentially. Conditional independence holds among the components of $\boldsymbol{y}_t$ given our choice of Σ_o being diagonal and Gaussian distributions. Let i index the components of $\boldsymbol{y}_t$ and $\{\sigma_i^2\}$ be the diagonal elements of Σ_o. To notate the sequential aspect of the update set $\boldsymbol{\mu}_a^0 = \boldsymbol{\mu}_f$ and $\Sigma_a^0 = \Sigma_f$ and let $\boldsymbol{\mu}_a^{i-1}$ and Σ_a^{i-1} be the prior mean and covariance used in the update with the i^{th} component of the observation vector.

With this notation, the sequential update for the i^{th} component respect to the (5) simplifies to

$$\boldsymbol{\mu}_a^i = \boldsymbol{\mu}_a^{i-1} + \Sigma_a^{i-1}\boldsymbol{h}_i \left[\frac{\boldsymbol{y}_{t,i} - \boldsymbol{h}_i^T \boldsymbol{\mu}_a^{i-1}}{\boldsymbol{h}_i^T \Sigma_a^{i-1} \boldsymbol{h}_i + \sigma_i^2} \right]$$

where $\boldsymbol{h}_i$ is the i^{th} row of H and the expression in brackets is a scalar. The update for the covariance matrix is also simple and is a rank one correction to Σ_a^{i-1}. This form is important for the adjustment ensemble Kalman filter used in the example.

2.4 Problems in implementation of the KF

There are two major difficulties in implementing the standard KF in geophysical data assimilation problems: handling large matrices and accounting for nonlinear dynamics. Typically global or high resolution problems have large state vectors. For example, just considering an atmospheric model at medium resolution (grid cells of about 150×150 km at the equator) the atmosphere is divided up into a three dimensional grid of $128 \times 256 \times 26$ each having at least 5 variables (temperature, pressure, horizontal wind components and water vapor). Thus, $\boldsymbol{x}$ has nominally approximately five million elements. In addition, the observation vector typically has on the order of 10^5 elements. Dimensions of this size prohibit computing or storing the elements of the covariance matrices from the update and forecast steps and it is not possible to implement the KF exactly for large problems. One alternative is to fix Σ_f with a convenient form and this is known as three dimension variational data assimilation and described below.

Besides direct problems with linear algebra the KF poses difficulties with nonlinear models for the dynamics. For atmospheric models evolution over time is represented as a complicated nonlinear transformation $[\boldsymbol{x}_{t+1}|\boldsymbol{x}_t] = L(\boldsymbol{x}_t, t)$ based on the nonlinear equations of fluid dynamics and thermodynamics. L is explicitly defined by a computer code that implements a discretised model and usually has no closed form. Thus, even if L is a deterministic function calculating a closed form for the forecast distribution, essentially $[L(\boldsymbol{x}_t)|\boldsymbol{x}_t]$, is not possible. This problem is compounded by the fact that $\boldsymbol{x}$ may have high dimension as well. Finally, because of the nonlinear action of L, one may expect that the resulting forecast distribution will not be normal and so the assumptions of multivariate normality within the update step will not hold.

In summary although providing closed forms for the posterior and forecast distributions in a linear situation, the KF is not practical for the kinds of assimilation problems encountered in some geophysical settings. Some strategy for an approximate solution is needed. Ensemble Kalman filters approximate the posterior distribution with a discrete sample. Another approach is to avoid large covariance matrices by not updating of the forecast covariance and this leads to variational methods of assimilation. We present both of these practical alternatives in the next two sections.

3 The ensemble Kalman filter

The term *ensemble* is used in the geosciences to refer to a sample either randomly drawn from a population or deliberated constructed. In data assimilation an ensemble of system states is used as a discrete approximation to the continuous, and often high dimensional distribution for $\boldsymbol{x}$. The basic idea of an ensemble Kalman filter is to use a sample of states to approximate the mean vectors and covariance matrices. Each ensemble member is updated by an approximation to Bayes theorem and is propagated forward in time using L giving a new ensemble for approximating the forecast distribution. This idea was proposed by Evensen [7] but has been developed by many subsequent researchers. It is a form of particle filter [6] with the ensemble members being "particles". One departure from standard particle filtering is that the ensemble members are modified at every update step rather than just being reweighted. It should be noted at the outset that this is a rich area of research and application within the geosciences and the overview in this section can not review many of the innovations and developments for specific problems. The details of implementing the update step with ensembles are important, especially when the ensemble size is small and one is concerned about the stability of the KF

over longer periods of time. The example at the end of this article gives some idea of practical issues and performance.

3.1 The ensemble update

The ensemble update step holds the main statistical details of the ensemble KF. By contrast, the forecast step for ensembles is both simple and explicit. To simplify notation, we will drop the time subscript because the computations are all at time t. Let $\{\boldsymbol{x}_f^j\}$ for $1 \leq M$ be an M member ensemble providing a discrete representation of the forecast pdf and $\{\boldsymbol{x}_a^j\}$ the corresponding ensemble for the analysis pdf (or posterior). The ensemble KF provides an algorithm using the observations and the update equations to transform the forecast ensemble into the analysis ensemble and so finesses the problem of working directly with high dimensional and continuous pdfs. If the dimension of the state vector is N then the storage for an ensemble is on order of $M \times N$ and can be much smaller than storing dense $N \times N$ covariance matrices. Examining the Kalman filter equations, the update equations (5) and (6) depend on the forecast mean $\boldsymbol{\mu}_f$ and the forecast covariance Σ_f. Given an ensemble, one replaces $\boldsymbol{\mu}_f$ by the *sample* mean and Σ_f by an estimate based on the *sample* covariance. Solving equation (5) results in a approximate posterior mean that we will denote $\hat{\boldsymbol{\mu}}_a$ and this will be taken as the sample mean of the updated ensemble. To understand how this is different from the the exact KF computations we give some details of this solution. Let $\bar{\boldsymbol{x}}_f$ be the ensemble mean forecast vector and

$$U_f = (\boldsymbol{x}_f^1 - \bar{\boldsymbol{x}}_f, \boldsymbol{x}_f^2 - \bar{\boldsymbol{x}}_f, ...\boldsymbol{x}_f^M - \bar{\boldsymbol{x}}_f) \tag{8}$$

be a matrix of the centered ensemble members. The sample forecast covariance has the form $\hat{\Sigma}_f = \frac{1}{M-1} U_f U_f^T$. Note that this estimate has effective rank M and when used in the update equations the linear algebra can exploit this reduced rank. Specifically, the full forecast covariance matrix need never be explicitly computed. Moreover in using iterative methods to solve the linear system in (5) the multiplication of $(H\hat{\Sigma}_f H^T + \Sigma_o)$ by an arbitrary vector can be done efficiency because of the reduced rank.

The other half of the update step involves the analysis covariance. The concept is to examine the form in (6) and modify the ensemble members to have a sample covariance close to this expression. There are two main strategies for doing this: a Monte Carlo approach, known as perturbed observations and a deterministic approach, know as a square root filter. For perturbed observations one generates M random vectors, $\boldsymbol{\epsilon}^j \sim N(0, \Sigma_o)$ that are further constrained so that the mean across j is zero. Now form M versions of "perturbed" observation vectors by adding these random deviates to the actual observation: $\boldsymbol{y}_t^j = \boldsymbol{y}_t + \boldsymbol{\epsilon}^j$. To update each ensemble member, apply the right side of equation 5 with the substitutions $\boldsymbol{x}_f^j$ for $\boldsymbol{\mu}_f$, $\hat{\Sigma}_f$ for Σ_f and $\boldsymbol{y}_t^j$ for $\boldsymbol{y}_t$ obtaining an analysis ensemble. Because the perturbed observations have zero sample mean the ensemble mean from this method will reproduce the posterior mean $\hat{\boldsymbol{\mu}}_a$. Moreover, as M goes to infinity the mean and covariance of the ensemble will match that of the posterior (see [8]).

Deterministic updates of the ensemble fall under the general ideas of square root Kalman filters ([25]). Given a matrix, A, the updated ensemble is generated through a linear transformation: $\boldsymbol{x}_a^j = U_f A + \hat{\boldsymbol{\mu}}_a$ Note that this ensemble will have mean vector $\hat{\boldsymbol{\mu}}_a$ and the key ideas is to choose A so that the sample covariance approximate the expression in (6). In other terms $UAA^TU^T = \hat{\Sigma}_a$. Note that A is only determined up to an orthogonal transformation and the choice is important in preserving physical structure of the state vectors as realizable states of

the system. Arbitrary transformations while preserving the mean and covariance may not yield ensemble members that are suitable for applying L in the forecast step. For large observation vectors the computations for an ensemble square filter may still be extensive and one approximation is to update the components of the state vector in a moving window of observations and is termed the Local ensemble Kalman filter (LETKF) [19].

3.2 The ensemble forecast step

Given an ensemble approximating the analysis distribution, we now describe the forecast step. An elegant property of ensemble methods is that the forecast step is exact to the extent that the discrete ensemble approximates the continuous distribution of the analysis pdf. Suppose that $\{\boldsymbol{x}_{a,t}^j\}$ are a random sample from $[\boldsymbol{x}_t|\boldsymbol{y}_t]$. Let $\boldsymbol{x}_{f,t+1}^j = L(\boldsymbol{x}_t^j, t)$ be the states obtained by propagating each member forward to $t+1$. By elementary probability this forecast ensemble will be a random sample from $[\boldsymbol{x}_{t+1}|\boldsymbol{y}_t]$ without requiring any additional assumptions on the distribution of the posterior.

3.3 Practical issues for small ensemble sizes

It is well known that for small ensemble sizes the sampling variability in $\hat{\Sigma}_f$ over the course of several update/forecast cycles can induce substantial error. A common effect is the collapse of the ensemble to the mean value. This occurs because errors in the covariance tend produce biases that favor less spread among the ensemble members. An artificially small forecast covariance results in the Kalman gain matrix decreasing the contribution of the observations to the update. The net result is a filter that ignores information from the observations and just propagates the ensemble forward in time. This behavior is known as filter divergence. Without incorporating observations and an L that has sensitive dependence to initial conditions the posterior distribution for $\boldsymbol{x}_t$ can diverge dramatically from the true state.

There are two important principles to counter this behavior in practice.

- *Localization* of the ensemble covariance estimate to improve the accuracy and reduce the effects of sampling
- *Inflation* to increase the spread of the ensemble.

Like other aspects of ensemble methods these two principles are implemented in many ways and we will just review some approaches that have a statistical or spatial thread.

3.4 Localization of the ensemble covariance

The simplest form of localization is to taper the sample covariance matrix based on distances. The rationale is that beyond a certain distance scale the assimilation errors for a spatial field should not be dependent and so the corresponding elements of the forecast covariance matrix should be set to zero and values at intermediate distances should be attenuated. (See [11] [12] for some background on localization.) Assume that each component of the state vector, $\boldsymbol{x}_k$ is associated with a location $\boldsymbol{u}_k$. and $d_{k,k'}$ is the distance between $\boldsymbol{u}_k$ and $\boldsymbol{u}_{k'}$. A tapered estimate is the direct (or Shure) matrix product of $\hat{\Sigma}_f$ with a correlation matrix.

$$[\tilde{\Sigma}_f]_{k,k'} = [\hat{\Sigma}_f]_{k,k'}\ \phi(d_{k,k'}) \tag{9}$$

where ϕ is a correlation function giving a positive definite correlation matrix. Based on the properties of the direct product, $\tilde{\Sigma}_f$ will remain nonnegative definite. To improve the computational efficiency the tapering is usually done with a compactly supported kernel ([9]). That is, $\phi(d)$ is identically zero for d sufficiently large, to introduce sparsity in the product covariance matrix. The result of this tapering is a covariance estimate that is biased but has less variance. Also introducing sparsity facilitates the matrix computations.

3.5 Inflation of the ensemble spread

Inflation is the operation of adjusting the ensemble spread beyond what is prescribed by the KF update formula to give a more appropriate measure of the uncertainty in the state. This adjustment can compensate not only for sampling variation of the ensemble but also in some cases for model error when the stochastic component, $\boldsymbol{u}_t$ from (7) has not be explicitly included. A useful assumption is that the forecast ensemble correlation structure is correct but that estimates of the variance of individual state vector components may be too small. After the model advance, but before the update step, the prior ensemble is inflated so that

$$\boldsymbol{x}_f^{j,inflated} = \sqrt{\boldsymbol{\lambda}_j}(\boldsymbol{x}_f^j - \boldsymbol{\mu}_f) + \boldsymbol{\mu}_f^j \tag{10}$$

There are fewer strategies for this than localization and while being effective are often *ad hoc* or global ([12], [11]) For example a standard approach is to multiply $\tilde{\Sigma}_f$ by a scalar that is greater than one , i.e. $\boldsymbol{\lambda}_j \equiv \lambda$. In the example in this article we describe a method of inflation that is based on sequentially inflating the state vector by a comparison of the forecast mean and variance to new observations [2]. In general there is a need for more work on statistical models and algorithms for handling inflation. Closely related to this issue is the need for statistical models to represent model error.

4 Variational methods of assimilation

Variational methods are more established than variants of the Kalman filter or a Bayesian framework and have been successful in operational settings for making rapid and reliable weather forecasts using large numerical models and massive data streams (for example see [23]). Essentially, variational methods estimate the state of the system by minimizing a cost function. As a starting point we identify the cost function problem that is equivalent to the update step in the KF and in terms of spatial statistics this will be equivalent to Kriging. However, despite this connection variational approaches often focus on crafting a cost function without relying on a Bayesian interpretation for motivation.

4.1 Three dimensional variational assimilation

Under the assumptions of linearity and multivariate normality the posterior of $[\boldsymbol{x}_t|y_t]$ is also multivariate normal. Moreover the mean of a Gaussian density is also the mode and so the posterior mode will also be the posterior mean in this case. Finally note that maximizing $[\boldsymbol{x}_t|y_t]$ is the same as minimizing minus the log of the joint distribution, $-(log([\boldsymbol{y}_t|\boldsymbol{x}_t]) + log([\boldsymbol{x}_t]))$, and where terms that do not depend on $\boldsymbol{x}_t$ or $\boldsymbol{y}_t$ can be omitted. Putting this remarks together we have motivated the variational problem:

$$\min_{\boldsymbol{x}} \quad (1/2)(\boldsymbol{y}_t - H\boldsymbol{x})^T {\Sigma_o}^{-1}(\boldsymbol{y}_t - H\boldsymbol{x}) + (1/2)(\boldsymbol{x} - \boldsymbol{\mu}_f)^T {\Sigma_f}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_f) \tag{11}$$

The minimizer of this cost function is the variational estimate of the state. It is a standard exercise to show that the minimum is

$$\hat{\boldsymbol{x}}_t = (H^T {\Sigma_o}^{-1} H + {\Sigma_f}^{-1})^{-1} H^T (\boldsymbol{y}_t - H\boldsymbol{x}) + \boldsymbol{\mu}_f \tag{12}$$

Based on the Sherwood-Morrison-Woodbury formula [10] this can be shown to be the same as the KF update in (5). Thus we have an alternative way of characterizing the mode of the analysis distribution. Since the Kriging equations are the same as 5 we have also outlined how Kriging can be interpreted as the solution to a variational problem.

This kind of cost function appears in many different areas and in general is characterized as a regularized solution to an inverse problem. $\boldsymbol{x}_t$ is estimated from data $\boldsymbol{y}_t$ by "inverting" H. A regularization term involving Σ_f is added to make this a well conditioned problem and insure a unique minimum to the cost function and regularization is a common strategy for creating a well posed cost function. Inverse problems cover a large range of applications in many different areas of science and engineering such as tomography and remote sensing, and variational methods of assimilation are just one special case. See [24] for some background. From the perspective of this article we can trace the regularization to the prior distribution for the state vector. Alternatively this term can be motivated by building in prior information about the state vector as an extra penalty in the cost function. For example if Σ_f is a covariance function for a smooth spatial field then as a regularization it will constrain the solution to also be smooth.

In the atmospheric sciences when $\boldsymbol{x}$ includes three dimensional fields this is known as 3DVAR. The important difference between 3DVAR and the update from the KF is that typically Σ_f is *fixed* and termed a background error covariance. Thus covariance information is not propagated based on the dynamics and the forecast step only involves $L(\hat{\boldsymbol{x}}_t)$. In the simplest form of 3DVAR there are no companion measures of uncertainty. This method has the advantage that it can be easily tuned by modifying or estimating parameters in the background covariance (e.g. [5]) and by incorporating physical constraints with additional cost terms.

An extension of 3DVAR is to add different time periods to the cost function. This is known as 4DVAR in the atmospheric sciences ([21], [20]) and seeks to find a single trajectory of the model that is consistent with observations at multiple times. An example of the cost function starting at $t = 1$ and going through $t = T$ is

$$\min_{\boldsymbol{x}_1} \quad (1/2) \sum_{t=1}^{T} \Big[(\boldsymbol{y}_t - H_t \boldsymbol{x}_t)^T {\Sigma_o}^{-1} (\boldsymbol{y}_t - H_t \boldsymbol{x}_t)] \Big] + (1/2)(\boldsymbol{x}_1 - \boldsymbol{\mu}_b)^T {\Sigma_b}^{-1} (\boldsymbol{x}1 - \boldsymbol{\mu}_b) \tag{13}$$

Here $\boldsymbol{\mu}_b$ and Σ_b refer to a background mean and covariance at the start of the period and provide some form of prior information on the initial state of the system. Implicit in this cost is that subsequent state vectors are found using a dynamical model, e.g. $\boldsymbol{x}_{t+1} = L(\boldsymbol{x}_t, t)$ that is deterministic. Despite the possible limitations from a high dimensional state vector and a complex dynamical model, this solution has an appealing interpretation. Based on initial conditions, $\boldsymbol{x}_1$, the solution is a *single* trajectory through time that best fits the observations. The only regularization is done on the initial conditions and the remaining times are constrained by the dynamical model. In general form a statistician would recognize 4DVAR as a large nonlinear ridge regression problem (where the parameters are $\boldsymbol{x}_1$). Note that as in 3DVAR only a point estimate is produced and measures of uncertainty need to generated by other methods. For problems with the complexity of an atmospheric model 4DVAR analysis is a

difficult computational problem partly because finding the minimum is a nonlinear problem and also the gradient of L may be hard to formulate or compute. Despite these hurdles, 4DVAR systems have been implemented as the primary assimilation algorithm in large operational models for weather forecasting (e.g. [23]).

5 An ensemble filter for an atmospheric model.

This section describes a specific ensemble based method and its application to a large numerical model. The sketch of this method known as the ensemble adjusted Kalman filter includes both localization and inflation and will illustrate some of the principles common to many other practical approaches. The atmospheric model and observations used in this example are at the scale of an operational system used to create global and complete fields of the atmosphere. The size of this problem is several orders of magnitude larger than typical applications of spatial statistics. The Data Assimilation Research Testbed (DART) (`http://www.image.ucar.edu/DAReS/DART`) is a open software environment that can be used to reproduce this analysis.

5.1 The adjusted ensemble Kalman Filter

The adjusted ensemble Kalman filter (EAKF) is based on a sequential update algorithm where an observation vector is assimilated as a sequence of scalar problems [1]. It may be surprising that this is efficient. In many geophysical applications one can take advantage of the observations errors being uncorrelated and a (great circle) distance based tapering to induce sparsity in the elements of Σ_f. Also, sequential updating and an ensemble approximation to the posterior are amenable to parallel computation, a necessary requirement for large problems [2]. The EAKF is a variant of the square root filter and so the modifications to the ensemble are deterministic. In the case of scalar updates the adjustment to each ensemble member is done to minimize the difference between its prior and updated values. This is in contrast to an approach such as perturbed observations where independent random components are added to each ensemble member and can produce more random shuffling among the ensemble members. Other square root filters may also induce significant differences between the forecast and analysis ensemble members.

5.2 Covariance localization and inflation

The main technique for localization will be recognized by a statistician as a shrinkage or decimation of the sample correlations found from the ensemble. The elements of $\hat{\Sigma}_f$ are tapered based on distance and further attenuated based on the size of the correlations. The amount of shrinkage is determined from an approximate resampling strategy where the ensemble is divided up in a small number of subsets and a shrinkage parameter is estimated by cross-validation [3]. In large problems with dense observations one anticipates that the data will provide substantial information of the atmospheric state over time. Moreover this state information can be reinforced by a physical model. From this perspective it is more important that localization be conservative in not updating components of the state vector due to spurious correlations in $\hat{\Sigma}_f$. Over multiple assimilation cycles these errors accumulate to cause filter divergence.

Inflation of the ensemble follows by estimating a vector of inflation values $\boldsymbol{\lambda}$ that scale each component as in (10). The spatial adaptation is implemented in a manner to be computationally feasible and parallel in the state vector components and is a good illustration of an algorithm

that is effective but does not necessarily follow directly from first statistical principles. For a new observation $\boldsymbol{y}_{t,i}$, let $\hat{\sigma}^2 = \boldsymbol{h}_i^T \hat{\Sigma}_f \boldsymbol{h}_i$ be the forecast ensemble variance and $\hat{y} = \boldsymbol{h}_i^T \hat{\boldsymbol{x}}_f$ its ensemble forecast mean. With this notation we have the forecast error $(\boldsymbol{y}_{t,i} - \hat{y})$ with an expected variance of $\sigma^2 + \sigma_o^2$ without inflation. For the j^{th} component of the state vector a pseudo-likelihood is taken as

$$[\boldsymbol{y}_{t,i} | \boldsymbol{\lambda}_j] \sim N(\hat{y}, \theta^2) \tag{14}$$

with

$$\theta^2 = \left[1 + \gamma_j (\sqrt{\boldsymbol{\lambda}_j} - 1)\right]^2 \sigma^2 + \sigma_o^2 \tag{15}$$

and γ_j is the correlation between $\boldsymbol{x}_j$ and $y_{t,i}$ based on the ensemble forecast covariance. A normal prior is used for $\boldsymbol{\lambda}_j$ and the posterior is approximated by a normal. The intuition behind the choice of (15) is a psuedolikelihood criterion that only links the inflation and the forecast error if the correlation between the actual state and the observation is large. When $\gamma = 1$ and $\boldsymbol{\lambda} \equiv \lambda$ is a constant the variance of the forecast error is parameterized as $\lambda^2 \sigma^2 + \sigma_o^2$. Thus in this limit (15) will reduce to a more conventional likelihood and give an approximate Bayesian inference for a scalar inflation of the forecast variance.

5.3 Assimilation with the Community Atmospheric Model

The Community Atmospheric Model (CAM) 3.1 [4] is a mature global atmospheric model that forms the atmospheric component of the Community Climate System Model, a state-of-the-art climate model. For this example CAM is configured at a resolution of approximately 128×256 on a longitude/latitude grid and has 26 vertical layers. Observations consist of soundings from weather balloons, measurements made from commercial aircraft (ACARS) and satellite derived wind fields and the initial ensemble was initialized from a climatogical distribution from this season. An 80 member ensemble was used for the EAKF. Available observations were assimilated every 6 hours over the period 1-jan-2007 through 31-jan-2007. Overall the quality of the forecast fields are comparable to reanalysis data products produced by the US National Center for Environmental Prediction and the National Center for Atmospheric Research ([18]) so are close the best analysis available for the atmosphere. The following figures summarize some of the statistical results. Figure 1 illustrates the effect of adding inflation in terms of the global root mean squared for the height of the atmosphere at a pressure of 500 hPa, a standard field used for forecasting. Here adaptive inflation is increasing the accuracy in the ensemble mean and at the same time producing ensemble members with larger spread. It interesting that better filter performance can be obtained with a larger range in the ensemble and suggests that the inflation may also contribute to better characterizations of the forecast uncertainty. Figure 2 is a snapshot of the inflation field ($\boldsymbol{\lambda}$) estimated at the end of the assimilation period indicating how the inflation factor varies over space. One surprise is that a large inflation is required over North America where there is a high data density. At first it seems contradictory that a data rich area, partly due to intensive aircraft observations, should require inflating the spread, and hence the uncertainty in the state. However, this effect might be explained by the presence of model error. Without inflation model errors are ignored and given a dense data region the forecast variance will be small especially as sampling errors accumulate through the sequential updating of the EAKF. A large amount of data in this region also allows the method to identify the discrepancy between the forecast variances and the actual forecast errors and so makes it possible to estimate an inflation field different from one.

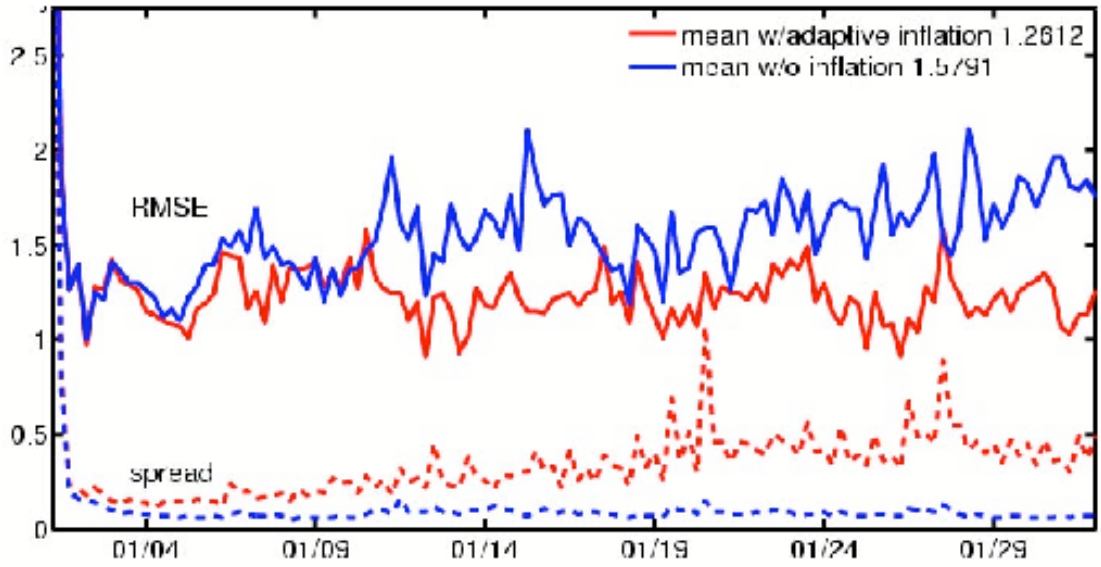

Figure 1: Root mean square errors (RMSE) for the 500 hPa heights from data assimilation derived using CAM 3.1 over the period 1/1/2007 through 1/31/2007. RMSE for the ensemble means based on assimilation with spatial inflation of the ensemble (red solid) and without adaptive inflation (blue solid). Dashed curves are the mean spread in the ensemble members.

It was noticed that adaptive inflation could be improved by building in some temporal correlation and the Bayesian estimation could be more formally integrated within the ensemble KF. Both of these are topics of future research. However, these results indicate that spatially varying parameters that control the assimilation process, even obtained by approximate or heuristic principles, are important and highlight an additional role for spatial statistics in data assimilation.

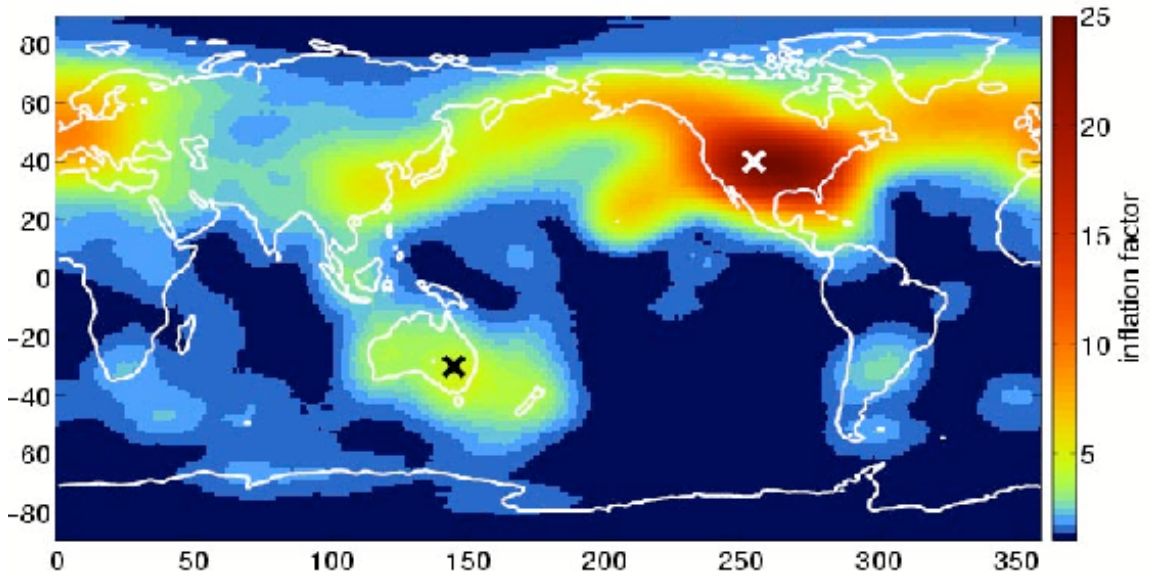

Figure 2: The field of spatial inflation values, $\boldsymbol{\lambda}$, at the end of the assimilation period.

References

[1] Anderson, J. L. (2001) An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review* 129, 2894-2903.

[2] Anderson, J., and N. Collins (2007). Scalable implementations of ensemble filter algorithms for data assimilation. *Journal of Atmospheric and Oceanography Technology A*, 24 1452-1463

[3] Anderson, J. L., (2007) Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D* 230, 99-111, doi:10.1016/j.physd.2006.02.011

[4] Collins, W. D., P. J. Rasch, and Others (2004). Description of the NCAR Community Atmosphere Model (CAM 3.0), *Technical Report NCAR/TN-464+STR*, National Center for Atmospheric Research, Boulder, Colorado, 210 pp.

[5] Dee, D. P. and A. M. da Silva (1999). Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: methodology. *Monthly Weather Reveiw* 127, 1822-1834

[6] A Doucet, N De Freitas (2001) Sequential Monte Carlo Methods in Practice Springer.

[7] Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi- geostrophic model using Monte Carlo methods to do forecast error statistics. *Journal Geophysical Research* 99(C5), 10143-10162.

[8] Furrer, R. and T. Bengtsson (2006). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal Multivariate Analysis* 98, 227-255.

[9] Gaspari, G. and S. E. Cohn (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal Royal Meteorological Society* 125, 723-757.

[10] Golub, G. H. , C. F. Van Loan (1996). *Matrix computations.* Johns Hopkins University Press, Baltimore.

[11] Houtekamer, P. L., and H. L. Mitchell (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review* 129, 123-137.

[12] Hamill, T. M., J. S. Whitaker, and C. Snyder (2001). Distance-dependent filtering of background-error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review* 129, 2776-2790.

[13] Harlim, J. and B. R. Hunt (2007) Four-dimensional local ensemble transform Kalman filter: numerical experiments with a global circulation model. *Tellus A*, 59(5),731-748, doi: 10.1111/j.1600-0870.2007.00255.x

[14] Jazwinski, A. H. (1970). *Stochastic Processes And Filtering Theory.* Academic Press, 376 pp.

[15] Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 82 (1): 35-45.

[16] Kalman, R. E., R. S. Bucy (1961). New Results in Linear Filtering and Prediction Theory. *Transactions of the ASME - Journal of Basic Engineering* 83, 95-107.

[17] Kalnay, E. (2002)*Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press.

[18] Kistler, R., E. Kalnay, W. Collins and others (2001). The NCEP-NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation. *Bulletin of the American Meteorological Society*, 82(2),247-268

[19] Szunyogh,I., E. J. Kostelich, G. Gyarmati, E. Kalnay, B. R. Hunt, E. Ott, E. Satterfield, J. A. Yorke (2008). A local ensemble transform Kalman filter data assimilation system for the NCEP global model. *Tellus A*, 60(1) , 113130 doi:10.1111/j.1600-0870.2007.00274.x

[20] Le Dimet F.-X., O. Talagrand (1968). Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus. A Dynamic meteorology and oceanography*, 38(2), 97-110

[21] Lewis, J. M. J. C. Derber (1985). The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus. Series A , Dynamic meteorology and oceanography* 37:44, 309-322.

[22] Mitchell, H. L. and P. L. Houtekamer (2000). An adaptive ensemble Kalman filter. *Monthly Weather Review* 128, 416-433.

[23] Rabier, F., H. Jarvinen, E. Klinker, J. F. Mahfouf, and A. Simmons (2000). The ECMWF operational implementation of four-dimensional variational assimilation. I: experimental results with simplified physics. *Quarterly Journal Royal Meteorological Society* 126, 1148-1170.

[24] Tarantola, A.(1987). *Inverse problem theory: methods for data fitting and model parameter estimation*. Elsevier Science Publishers.

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

[26] Wikle, C. (2008). Hierarchical Modeling with Spatial Data. *This volume*

[27] Wikle, C. K. and L. M. Berliner (2006). A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena* 230, 1-16

6 Acknowledgments

This work was supported by the National Science Foundation through a grant to the University Corporation for Atmospheric Research. Nychka also acknowledge the support of the Isaac Newton Institute for Mathematical Sciences.