Data Assimilation based on estimation theory: Some fundamentals

Lecture notes

These lecture notes are still under development. They come in support of oral lectures on data assimilation in geophysics (restricted to methods based on estimation theory) given annually in Grenoble, France, since 2006 as part of the "Formations doctorales" program.

The numbering of sections and subsections follows the outline of the oral lectures. The empty parts are those treated during the lectures, on the black board or with numerical illustrations. Comments are welcome.

Emmanuel Cosme
Université Grenoble Alpes, Institut des Géosciences de l'Environnement, Grenoble, France
Emmanuel.Cosme@univ-grenoble-alpes.fr

Last revision: January 18, 2021

Contents

1 Introduction

1.1 What is data assimilation?

The basic purpose of data assimilation is to combine different sources of information to estimate at best the state of a system. These sources generally are observations and a numerical model. Why not simply use observations? First, because observations are sparse or partial in geophysics. Some information is necessary to interpolate the information from observations to unobserved regions or quantities. A numerical model naturally does that. Second, because observations can be noised. Combining several noised data is an efficient way to filter out noise and provide a more accurate estimate.

A specificity (though not the only one) of data assimilation in geophysics is the time dimension. The dynamical aspect of geophysical fluids makes the problem difficult but interesting. Meteorologists and oceanographers talking about data assimilation always imply dynamics.

The data assimilation problem may be tackled with different mathematical approaches: signal processing, control theory, estimation theory for example. This course takes the point of view of estimation theory, from which the well known Kalman filter derives. Variational methods (3D-Var, 4D-Var...) comes from control theory.

The historical development of data assimilation for geophysical fluids can hardly be disconnected from meteorology. It is indeed a necessary step to provide a good initialization for a prediction, and until the 90's data assimilation has been developed and used in that only purpose. Today, its application generalizes to many other fields (atmospheric chemistry, oceanic biochemistry, glaciology, etc) and applications, for example:

- the estimation of the trajectory of a system to study its variability (reanalyses);
- the identification of systematic errors in numerical models;
- the optimization of observation network;
- the estimation of unobserved variables;
- the estimation of parameters.

1.2 A short history of data assimilation in meteorology

1.2.1 Subjective analysis (19th century)

Subjective analysis consists in extrapolating "by hand" a set of local observations (of pressure, historically) to provide a pressure map. Though subjective, this is a kind of data assimilation, where local observations are combined with the "good sense" and the experience of the meteorologist to provide a map.

1.2.2 Richardson's numerical weather prediction (1922)

Lewis Fry Richardson was the first scientist to try a numerical weather prediction. He made it by hand in 1917, while he was serving in a military unit in the north of France. Unfortunately, his attempt dramatically failed, due to a 145 mbar rise in pressure over 6 hours. The cause has been identified later: the prediction was initialized with in situ pressure observations, an unbalanced input for his numerical model. Lynch (1993) showed that with an appropriate smoothing of the initial condition, Richardson's prediction would have turned fairly accurate. His failure can thus be viewed as due to a deficiency with data assimilation.

1.2.3 Cressman's objective analysis (1950's)

The most relevant idea of Cressman was to acknowledge the poorness of the observation network, and to introduce a background: an *a priori* knowledge of the atmospheric state, to be modified when observations are available. At the grid point j, the correction writes:

$$\mathbf{x}_j^a = \mathbf{x}_j^b + \frac{\sum_{i=1}^{s} w(i,j)(\mathbf{y}_i - \mathbf{x}_i^b)}{\sum_{i=1}^{s} w(i,j)} \tag{1}$$

where $\mathbf{y}_i$ is the observation at the grid point i and $w(i,j)$ is the weight of $\mathbf{y}_i$ at the point j. To prescribe the weights, Cressman proposes:

$$\begin{cases} w(i,j) = \frac{R^2 - r(i,j)^2}{R^2 + r(i,j)^2} & \text{if} \quad r(i,j) \leq R \\ w(i,j) = 0 & \text{if} \quad r(i,j) > R \end{cases}$$

$r(i,j)$ is the distance between the points i and j. R is an influence radius to be prescribed. The main difficulty of this method is to determine objectively the weights. The method has other serious drawbacks: all the observations are processed identically, whatever their quality is; the physical balance is not controlled.

1.2.4 Nudging (1970's)

The idea is to force the numerical model toward the observations with an extra term for elastic relaxation. If the model writes:

$$\frac{d\mathbf{x}}{dt} = \mathcal{M}(\mathbf{x}) \tag{2}$$

then the nudging equation is:

$$\frac{d\mathbf{x}}{dt} = \mathcal{M}(\mathbf{x}) + \alpha(\mathbf{y} - \mathbf{x}) \tag{3}$$

where $\mathbf{y}$ is a direct observation of $\mathbf{x}$. This method also displays several drawbacks: the relaxation coefficient α must be determined. The method is no more applicable with undirect observations. Nudging, sometimes referred to as "the poor man's data assimation method", is also very simple to implement. For that reason it is still used for specific applications, mainly when the observations are not real observations but grided data from a reanalysis for example.

1.2.5 3Dvar and Optimal Interpolation (1980's)

This section appears for the sake of consistency, but is not meant to be detailed here. 3D-Var is a variational method, based on control theory. It is not detailed in this course, although it is briefly mentioned in passing later. Optimal interpolation is presented later as a approximation to the Kalman filter.

1.2.6 4Dvar and the Kalman filter (1990's)

This section also appears for consistency. 4D-Var is not adressed in this course. The Kalman filter is one of the core of the course and is detailed in section 6.

1.3 An illustration with Lorenz' model

See oral lecture.

The conclusion is:

- For various reasons, and as illustrated in this example, it is and it will always be impossible to perfectly know the states of the atmosphere or the ocean. Thus we simply decide to consider these states as random variables, and invoke estimation theory to catch them as precisely as possible;

- Over-simplistic data assimilation methods can be useful sometimes, but that is generally not the case.

2 Basic elements in probability and statistics

2.1 Probability

2.1.1 Random experiment

A random experiment is mathematically described by:

- the set Ω of all possible outcomes of an experiment, the result of which cannot be perfectly anticipated;
- the subsets of Ω, called events;
- a probability function, P: a numerical expression of a state of knowledge. P is such as, for any disjoint events A and B:

$$0 \leq P(A) \leq 1,$$
$$P(\Omega) = 1,$$
$$P(A \cup B) = P(A) + P(B)$$

2.1.2 Conditional probability

When the two events A and B are not independent, knowing that B has occurred changes our state of knowledge on A. This reads:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

2.2 Real random variables

The outcome of a random experiment is called a random variable. A random variable can be either an integer number (e.g., a die cast) or a real number (e.g., the lifetime of a electric light bulb).

2.2.1 Probability density function

For real random variables, equality to a given number is not an event. Only the inclusion into an interval is an event. This defines the **probability density function**, commonly referred to as pdf:

$$P(a < X \leq b) = \int_a^b p_X(x)dx.$$

2.2.2 Joint and conditional pdf

If X and Y are two real random variables, $p_{X,Y}(x, y)$ is the joint pdf of X and Y. Conditioning applies as with the discrete random variables, so that

$$p_{X|Y=y^o}(x) = \frac{p_{X,Y}(x, y^o)}{p_Y(y^o)}.$$

2.2.3 Expectation and variance

A pdf is rarely known completely. Generally, only some properties are determined and handled. The two main properties are the expectation and the variance. The expectation of a random variable X is

$$\mathcal{E}(X) =< X >= \int_{-\infty}^{+\infty} x p_X(x) dx.$$

The variance is

$$\mathrm{Var}(X) = \mathcal{E}\left([X - \mathcal{E}(X)]^2\right) = \int_{-\infty}^{+\infty} [x - \mathcal{E}(x)]^2 p_X(x) dx.$$

The standard deviation is the square root of the variance.

2.2.4 The Gaussian distribution

The random variable X has a Gaussian (or normal) distribution with parameters μ and σ^2, which is noted $X \sim \mathcal{N}(\mu, \sigma^2)$, when

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

The Gaussian distribution possesses some very nice properties, in particular:

- It is a natural distribution for signal noises (a consequence of the central limit theorem);
- the parameters μ and σ^2 of the distribution are the expectation and the variance, respectively;
- If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are two independent variables, then $X_1 + X_2$ is also Gaussian and $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$;
- If a and b are real numbers and $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

2.3 Real random vectors

Real random vectors are vectors which components are real random variables. The pdf of a vector is the joint pdf of its real components.

2.3.1 Expectation and variance

The expectation vector is the vector formed with the expected values of the real components. The second moment of the distribution is the covariance matrix. If $\mathbf{X}$ denotes the random vector, the covariance matrix is defined by

$$\mathcal{E}\left[(\mathbf{X}-\mathcal{E}(\mathbf{X}))\,(\mathbf{X}-\mathcal{E}(\mathbf{X}))^T\right].$$

A covariance matrix is symmetric, positive. The terms on the diagonal are the variances of the vector components. The non diagonal terms are covariances. If X_i and X_j denotes two different components of $\mathbf{X}$, their covariance is

$$\mathrm{Cov}(X_i,X_j)=\mathcal{E}\left[(X_i-\mathcal{E}(X_i))\,(X_j-\mathcal{E}(X_j))\right]$$

and their correlation is

$$\rho(X_i,X_j)=\frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}}.$$

2.3.2 The multivariate Gaussian distribution

The random vector $\mathbf{X}$ of size n has a Gaussian (or normal) distribution with parameters μ and $\mathbf{P}$, which is noted $\mathbf{X}\sim\mathcal{N}(\mu,\mathbf{P})$, when

$$p_{\mathbf{X}}(\mathbf{x})=\frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{P}^{-1}(\mathbf{x}-\mu)\right].$$

μ and $\mathbf{P}$ are the expectation and the covariance matrix of $\mathbf{X}$, respectively. $|\mathbf{P}|$ denotes the determinant of $\mathbf{P}$. The component of $\mathbf{X}$ are said to be jointly Gaussian.

Any linear combination of Gaussian vectors is Gaussian. In particular, if $\mathbf{X}\sim\mathcal{N}(\mu,\mathbf{P})$, then $\mathbf{LX}\sim\mathcal{N}(\mathbf{L}\mu,\mathbf{LPL^T})$. By choosing the linear transformation $\mathbf{L}$ that selects a subset of $\mathbf{X}$ components, it can be easily shown that the marginal distribution for this subset is also Gaussian. These important properties will be used later.

2.4 Two fundamental rules of estimation theory

Two fundamental rules of estimation theory are:

Bayes' rule,

$$p_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^o}(\mathbf{x}) = \frac{p_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}^o)p_{\mathbf{X}}(\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y}^o)} \tag{4}$$

and the marginalization rule:

$$p_{\mathbf{Z}}(\mathbf{z}) = \int p_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z})d\mathbf{x} = \int p_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}(\mathbf{Z})p_X(\mathbf{x})d\mathbf{x} \tag{5}$$

3 Ingredients of data assimilation

Data assimilation generally consists in estimating a time sequence $(\mathbf{x}_k)$ of the state of a dynamical system based on some prior information if it exists, a numerical model, and a sequence $(\mathbf{y}_k^o)$ of observations. Index k is a time index here.

3.1 Discretization and *true* state

Most of the time, we aim to estimate with the best possible accuracy a geophysical field that vary continuously in space and time. This real, continuous (and possibly multivariate) field is noted $\mathbf{x}^c$.

Numerical models are often used for that purpose. But numerical models handle only discrete representations of the physical field. Proceeding this way, one implicitly drops the idea of estimating the real state $\mathbf{x}^c$, and tries to estimate a projection of $\mathbf{x}^c$ in a discrete space. Let Π be this projector, and $\mathbf{x}^t$ the projection of $\mathbf{x}^c$:

$$\mathbf{x}^t = \Pi(\mathbf{x}^c) \tag{6}$$

$\mathbf{x}^t$ is called the *true* state, and is the state to estimate. Because this true state is not known, it is represented by a random vector denoted $\mathbf{X}$.

3.2 *Prior* information

Often, particularly in geophysics, we have a *prior* knowledge of the state $\mathbf{X}$, under the form of the prior pdf $p_{\mathbf{X}}(\mathbf{x})$. The expectation $E(X)$ is often denoted $\mathbf{x}^b$. This is the *background* state. The background error is then defined as:

$$\epsilon^b = \mathbf{x}^b - \mathbf{X}. \tag{7}$$

The covariance matrix is then noted $\mathbf{P}^b$ (or simply $\mathbf{B}$, mainly in the framework of variational methods). Later in the text, we will see that the background state often comes from a model simulation. In this case the background is a *forecast* and is noted $\mathbf{x}^f$ instead. The forecast error is noted ϵ^f and the covariance matrix $\mathbf{P}^f$. When the state $\mathbf{X}$ is Gaussian, the background state and the covariance matrix are the parameters of the pdf.

3.3 Observations

- The **observation**Êis the measurement operation;
- The **observable**Êis the variable we want to measure ($\mathbf{y}^c$ below) ;
- The **observation result** $\mathbf{Y}$ represents the ensemble of values possibility attributed to the observable, given the observation setup and its uncertainties. It is a random vector;
- The **observation vector** (or value, for a scalar) is the numerical result of the observation. It is a realization of the observation result $\mathbf{Y}$.

The real field $\mathbf{x}^c$ results in a real signal, the observable $\mathbf{y}^c$ in the observation space. The causality relation involves a function h:

$$\mathbf{y}^c = h(\mathbf{x}^c) \tag{8}$$

Equation 8 is extremely simple but unusable in this form. First, $\mathbf{y}^c$ is not accessible. The observation result is a random vector $\mathbf{Y}$ affected by instrumental errors. Let ϵ^μ denote this measurement error, then

$$\mathbf{Y} = h(\mathbf{x}^c) + \epsilon^\mu \tag{9}$$

As seen earlier, the real state $\mathbf{x}^c$ is also not accessible and only $\mathbf{X}$ is searched for. Also, h represents the physics of the measure. Although the physical processes can be known, we cannot know and handle h numerically. In practice, the physics is represented by a numerical model $\mathcal{H}$, which applied to the discrete state $\mathbf{X}$, and is called the *observation operator*. Involving $\mathcal{H}$ and Π, equation 9 can be rewritten:

$$\mathbf{Y} = \mathcal{H}(\mathbf{X}) + \underbrace{h(x^c) - \mathcal{H}(\Pi(\mathbf{x}^c))}_{\epsilon^r} + \epsilon^\mu \tag{10}$$

ϵ^r is the *representativity error*. It includes the errors related to the representation of the physics in $\mathcal{H}$, and those due to the projection Π of the real state $\mathbf{x}^c$ on the discrete state space.

The sum of the measurement error and the representativity error,

$$\epsilon^o = \epsilon^r + \epsilon^\mu, \tag{11}$$

is the *observation error*, and the final equation that links the true state to the observation result is the *observation equation*:

$$\mathbf{Y} = \mathcal{H}(\mathbf{X}) + \epsilon^o \tag{12}$$

From the point of view of the observation system, the actual realization of the state $\mathbf{X}$ is "seen", therefore "known". The measurement pdf writes $p_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y})$ and is simply obtained with a translation of p_{ϵ^o} by $h(\mathbf{X})$. In practice, the observation provides an observation result $\mathbf{y}^o$. The measurement pdf evaluated at $\mathbf{y}^o$ is seen as a function of $\mathbf{x}$ and is named "likelihood".

For data assimilation purposes, the statistics of the observation error ϵ^o are often supposed known and Gaussian distributed with mean 0 and covariance matrix noted $\mathbf{R}$. Then, $p_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}^o) \sim \mathcal{N}(\mathbf{y}^o, \mathbf{R})$.

3.4 Numerical models

Here are considered the dynamical models, i.e. the models that compute the time evolution of the simulated state. Let $\mathbf{x}_k^c$ and $\mathbf{x}_{k+1}^c$ be the real (continuous) state at two consecutive observation times, k being a time index. They are related by a causality link:

$$\mathbf{x}_{k+1}^c = g(\mathbf{x}_k^c). \tag{13}$$

For the reasons already detailed previously, we only pretend to estimate the true state:

$$\mathbf{x}_{k+1}^t = \Pi\left(g(\mathbf{x}_k^c)\right). \tag{14}$$

g is like h: not known strictly, although we known (hopefully...) most of the physics involved in it. This physics is represented by our numerical model $\mathcal{M}$, which works with discrete states such as $\mathbf{x}^t$. Let us introduce this piece of knowledge into equation 14, and turn to a random vector formulation:

$$\mathbf{X}_{k+1} = \mathcal{M}_{k,k+1}(\mathbf{X}_k) + \eta_{k,k+1} \tag{15}$$

where

$$\eta_{k,k+1} = \Pi\left(g(\mathbf{x}_k^c)\right) - \mathcal{M}_{k,k+1}(\mathbf{X}_k). \tag{16}$$

The *model error* $\eta_{k,k+1}$ accounts for the errors in the numerical model (e.g., misrepresentation of physical processes) and for the errors due to

the discretization. Again, the actual value of this error is not known, it is thus considered as a random variable. This is why Equation 15 is written with random vectors. For data assimilation, a *transition* or *evolution* pdf $p_{\mathbf{X}_{k+1}|\mathbf{X}_k=\mathbf{x}_k}(\mathbf{x})$ is generally used. This pdf is drawn from $p(\eta_{k,k+1})$ by a translation of $\mathcal{M}_{k,k+1}(\mathbf{X}_k)$.

4 Bayesian formulation of the sequential data assimilation problem

The previous section focused on solving the analysis step using estimators. Here, we present the more general Bayesian formulation of the whole sequential data assimilation process.

4.1 Reminder: two fundamental rules of estimation theory

This is a reminder of section 2.4. Two fundamental rules of estimation theory are:

Bayes' rule,

$$p_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^o}(\mathbf{x}) = \frac{p_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}^o)p_{\mathbf{X}}(\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y}^o)} \tag{17}$$

and the marginalization rule:

$$p_{\mathbf{Z}}(\mathbf{z}) = \int p_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z})d\mathbf{x} = \int p_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}(\mathbf{z})p_X(\mathbf{x})d\mathbf{x} \tag{18}$$

4.2 Bayesian formulation of the sequential data assimilation problem

Let us note $\mathbf{X}_{0:k}$ the sequence of states $\mathbf{X}_0, \ldots, \mathbf{X}_k$, and $\mathbf{y}^o_{1:k}$ the sequence of observation vectors $\mathbf{y}^o_1, \ldots, \mathbf{y}^o_k$. The purpose of data assimilation is to estimate states based on observation vectors. This problem can take many forms, so we will focus on the *filtering* problem only, which consists in finding the pdf of $\mathbf{X}_k$ given past and present observation vectors, $\mathbf{y}^o_{1:k}$. This conditional pdf is $p_{\mathbf{X}_k|\mathbf{Y}_{1:k}=\mathbf{y}^o_{1:k}}(\mathbf{x})$. Smoothing problems, involving future observations, will be considered later in the course.

Let us recapitulate the available information first:

- A prior at time 0: $p_{\mathbf{X}_0}(\mathbf{x})$;
- The likelihoods $p_{\mathbf{Y}_i|\mathbf{X}_i=\mathbf{x}_i}(\mathbf{y}^o_i)$ for $i = 1, \ldots, k$;

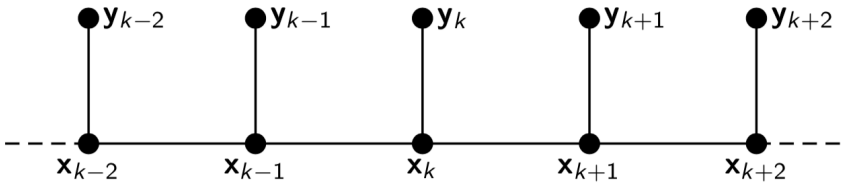

Figure 1: Hidden Markov chain. $\mathbf{x}$'s are states. $\mathbf{y}$'s are observations.

- The transition densities $p_{\mathbf{X}_{i+1}|\mathbf{X}_i=\mathbf{x}_i}(\mathbf{x})$.

It is generally assumed that the model errors are independent of each other; observations errors are independent of each other; model and observation errors are independent. Given the Markov form of Eq. 12 and 15, this reduces to consider the transition from $\mathbf{X}_i$ to $\mathbf{X}_{i+1}$ and the transformation from $\mathbf{X}_i$ to $\mathbf{Y}_i$ as Markov processes. This Markovian property can be understood graphically, by representing a *Hidden Markov chain*, as on Figure 1. For a given variable, some node may hide dependencies to other, remote variables. A few examples are (in compact notations):

- $p_{\mathbf{Y}_k|\mathbf{X}_k=\mathbf{x}_k,\mathbf{X}_{k-1}=\mathbf{x}_{k-1}}(\mathbf{y}) = p_{\mathbf{Y}_k|\mathbf{X}_k=\mathbf{x}_k}(\mathbf{y})$;
- $p_{\mathbf{X}_{k+1}|\mathbf{X}_k=\mathbf{x}_k,\mathbf{Y}_k=\mathbf{y}_k^o}(\mathbf{x}) = p_{\mathbf{X}_{k+1}|\mathbf{X}_k=\mathbf{x}_k}(\mathbf{x})$.

The natural procedure alternates *propagation* steps, using the marginalization rule and the transition density:

$$p_{\mathbf{X}_k|\mathbf{Y}_{1:k-1}=\mathbf{y}^o_{1:k-1}}(\mathbf{x}) = \int p_{\mathbf{X}_k|\mathbf{X}_{k-1}=\mathbf{x}'}(\mathbf{x}) p_{\mathbf{X}_{k-1}|\mathbf{Y}_{1:k-1}=\mathbf{y}^o_{1:k-1}}(\mathbf{x}') d\mathbf{x}', \quad (19)$$

with analysis (or update) steps, using Bayes' rule and the likelihood:

$$p_{\mathbf{X}_k|\mathbf{Y}_{1:k}=\mathbf{y}^o_{1:k}}(\mathbf{x}) \propto p_{\mathbf{X}_k|\mathbf{Y}_{1:k-1}=\mathbf{y}^o_{1:k-1}}(\mathbf{x}) p_{\mathbf{Y}_k|\mathbf{X}_k=\mathbf{x}}(\mathbf{y}^o_k). \quad (20)$$

Other methods are possible, but rarely used in practice.

5 The particle filter

The two pillars of estimation theory was presented in section 2.4. But they were not explicitly used in the Kalman filter. Indeed the Kalman filter handles only the first two moments of the pdfs, not the pdfs themselves.

This is optimal only in the linear gaussian case, as it was shown in section 6. What happens if the models are nonlinear and the pdfs non gaussian? The Kalman filter is no more optimal and, more importantly, can easily fail the estimation process. Other approaches must be used, then comes the particle filter. The particle filter works sequentially in the spirit of the Kalman filter, but unlike the latter, it handles an ensemble of states (the particles) which distribution approximates the pdf of the true state. The two pillars are explicitly used in the estimation process. The linear and gaussian hypotheses can then be ruled out, in theory. In practice though, the particle filter cannot yet be applied with high dimensional systems (this is often referred to as "the curse of dimensionality").

5.1 Particle implementation of Bayes' rule (analysis step)

The time index k is dropped in this section for clarity.

Very often, the measurement model pdf $p(\mathbf{y}|\mathbf{x})$ can be accurately approximated with an analytical function (Gaussian for instance), $f(\mathbf{x}, \mathbf{y})$. The prior pdf is not fully known, but is represented by an ensemble of M states $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^M$. These states, called particles, form a discrete approximation of the pdf:

$$p(\mathbf{x}) \simeq \sum_{i=1}^{M} \omega^i_{\text{prior}} \delta(\mathbf{x} - \mathbf{x}^i), \qquad \text{with} \qquad \sum_{i=1}^{M} \omega^i_{\text{prior}} = 1 \tag{21}$$

Bayes' rule provides a particle approximation of the posterior pdf:

$$p(\mathbf{x}|\mathbf{y}) \simeq \sum_{i=1}^{M} \omega^i_{\text{posterior}} \delta(\mathbf{x} - \mathbf{x}^i) \tag{22}$$

where

$$\omega^i_{\text{posterior}} = \frac{\omega^i_{\text{prior}} f(\mathbf{x}^i, \mathbf{y})}{\sum_{i=1}^{M} \omega^i_{\text{prior}} f(\mathbf{x}^i, \mathbf{y})}. \tag{23}$$

In Equation 23, the denominator is simply a normalization factor, so that the sum of the $\omega^i_{\text{posterior}}$'s is 1.

However, a problem arises very quickly with this update step, especially when the vector dimension is large: a few particles tend to acquire relatively large weights, while the others get negligible weights, resulting in a negligible contribution to the posterior pdf. This problem is known and solutions have been proposed, which consist in resampling the posterior pdf. The

most common and easy-to-understand is the bootstrap filter (Gordon et al, 1993). It consists in selecting the particle with high weights and ruling out the others. Then, the surviving particles are cloned a number of time proportional to their respective weights. After the resampling, the ensemble gathers again M particles, some of them identical, with approximately equal weight.

Below is presented a very simple resampling procedure, only for illustration. Such simple procedures are rarely efficient enough for real applications.

For each particle n, do:

- *draw a random number u from the uniform pdf over* $[0, 1[$*;*
- *set* $j = 1$*;*
- *set* $S_w = w(1)$*;*
- *while* $S_w < u$ *do:*
  - $j = j + 1$*;*
  - $S_w = S_w + w(j)$*;*
- *Particle j is retained and replace particle n.*

5.2 Particle implementation of forecast step

The analysis step provides (after resampling) a particle approximation of the pdf $p(\mathbf{x}_k|\mathbf{y}_k)$:

$$p(\mathbf{x}_k|\mathbf{y}_k) \simeq \sum_{i=1}^{M} \omega^i_{\text{posterior},k} \delta(\mathbf{x}_k - \mathbf{x}^i_k) \tag{24}$$

The forecast step consists in calculating a particle approximation of the pdf $p(\mathbf{x}_{k+1}|\mathbf{y}_k)$. Assuming the existence of a (numerical) model given by

equation 15, the forecast step is performed using the marginalization rule:

$$
\begin{aligned}
p(\mathbf{x}_{k+1}|\mathbf{y}_k) &= \int p(\mathbf{x}_{k+1}|\mathbf{x}_k, \mathbf{y}_k) p(\mathbf{x}_k|\mathbf{y}_k) d\mathbf{x}_k && (25)\\
&= \int p(\mathbf{x}_{k+1}|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{y}_k) d\mathbf{x}_k && (26)\\
&\simeq \int p(\mathbf{x}_{k+1}|\mathbf{x}_k) \sum_{i=1}^{M} \omega^i_{\text{posterior},k} \delta(\mathbf{x}_k - \mathbf{x}^i_k) d\mathbf{x}_k && (27)\\
&\simeq \sum_{i=1}^{M} \omega^i_{\text{posterior},k} \int p(\mathbf{x}_{k+1}|\mathbf{x}_k) \delta(\mathbf{x}_k - \mathbf{x}^i_k) d\mathbf{x}_k && (28)\\
&\simeq \sum_{i=1}^{M} \omega^i_{\text{posterior},k} \delta\left(\mathbf{x}_{k+1} - \mathcal{M}_{k,k+1}(\mathbf{x}^i_k) - \eta_{k,k+1}\right) && (29)
\end{aligned}
$$

In other word, each particle is simply propagated with the dynamical model. It is important to introduce a model noise term to each particle. This way, particles identical after the resampling step separate.

6 The Kalman filter

6.1 Introduction

The system is now dynamical. Instead of a unique state, we aim at estimating a series of states $\mathbf{x}_k$, where the subscript k is a time index pointing observation dates. We assume to have the following *a priori* pieces of knowledge:

- the initial state $\mathbf{x}^t_0$ is Gaussian with mean $\mathbf{x}^b_0$ and covariance $\mathbf{P}^b_0$;
- a linear dynamical model $\mathbf{M}_k$ that describes the state evolution;
- the model errors η_k are Gaussian with mean 0 (unbiased error) and covariance $\mathbf{Q}_k$;
- the model errors are white in time: $< \eta_k \eta^T_j >= 0$ if $k \neq j$;
- linear observation operators that link the states to the observations;
- the observation errors ϵ^o_k are Gaussian with mean 0 (unbiased errors) and covariance $\mathbf{R}_k$;
- the observation errors are white in time: $< \epsilon^o_k \epsilon^{oT}_j >= 0$ if $k \neq j$;

- Errors of different types are independent: $< \eta_k \epsilon_j^{oT} >= 0$, $< \eta_k \epsilon_0^{bT} >= 0$, $< \epsilon_k^o \epsilon_0^{bT} >= 0$.

6.2 Analysis step

At time t_k, $p(\mathbf{x}_k|\mathbf{y}_{1:k-1})$ is known through the mean $\mathbf{x}_k^f$, the covariance matrix $\mathbf{P}_k^f$, and the assumption of a gaussian distribution. The analysis step consists in updating this pdf using the observation $\mathbf{y}_k$ available at time t_k, and find $p(\mathbf{x}_k|\mathbf{y}_{1:k})$. This is done using Bayes' rule. Notation f in superscript is used because this comes from a previous forecast, as it will be shown in the next section. If $k = 0$, f must be replaced by b. Since both state and observation are from time k here, the time index is dropped for conciseness in this section.

Starting with:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x}^f, \mathbf{P}^f), \quad p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}^f|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}^f)^T \mathbf{P}^{f\,-1}(\mathbf{x}-\mathbf{x}^f)\right],$$

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{H}\mathbf{x}, \mathbf{R}), \quad p(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{R}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right].$$

Then Bayes'rule provides the posterior pdf. Given the realization $\mathbf{y}^o$ of the observation $\mathbf{y}$,

$$p(\mathbf{x}|\mathbf{y}) \propto \exp(-J) \tag{30}$$

with

$$J(\mathbf{x}) = \frac{1}{2}\left[(\mathbf{x}-\mathbf{x}^b)^T \mathbf{P}^{b\,-1}(\mathbf{x}-\mathbf{x}^b) + (\mathbf{y}^o - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x})\right]. \tag{31}$$

It can be easily shown that equation 31 leads to

$$J(\mathbf{x}) = \frac{1}{2}\left[(\mathbf{x}-\mathbf{x}^a)^T \mathbf{P}^{a\,-1}(\mathbf{x}-\mathbf{x}^a)\right] + \beta, \tag{32}$$

with

$$\mathbf{P}^a = \left[\mathbf{P}^{b\,-1} + \mathbf{H}^T \mathbf{R}^{-1}\mathbf{H}\right]^{-1}, \tag{33a}$$

$$\mathbf{x}^a = \mathbf{P}^a\left[\mathbf{P}^{b\,-1}\mathbf{x}^b + \mathbf{H}^T \mathbf{R}^{-1}\mathbf{y}^o\right]. \tag{33b}$$

and β is independent of $\mathbf{x}$. With the help of the Sherman-Morrison-Woodbury (SMW) formula:

$$[\mathbf{A} + \mathbf{U}\mathbf{D}\mathbf{V}]^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}[\mathbf{D}^{-1} + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}]^{-1}\mathbf{V}\mathbf{A}^{-1}, \tag{34}$$

it can be shown that these are the BLUE equations. The *a posteriori* pdf 30 is thus gaussian, and its parameters are given by the BLUE equations. Hence with gaussian pdfs and linear observation operator, there is no need to use Bayes'rule: the BLUE equations can be used instead to compute the parameters of the resulting pdf. Since the BLUE provides the same result as Bayes'rule, it is the best estimator of all.

In passing, one can recognize the 3D-Var cost function. By minimizing this cost function, 3D-Var finds the MAP estimate of the gaussian pdf, what is equivalent to the MV estimate found by the BLUE.

6.3 Forecast step

After the analysis step, the gaussian pdf $p(\mathbf{x}_k|\mathbf{y}_{1:k})$ is known through the mean $\mathbf{x}_k^a$ and the covariance matrix $\mathbf{P}_k^a$. The forecast step provides $p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k})$ using the marginalization rule, Eq. 19.

6.4 Synthesis

The Kalman filter is initialized with an forecast state vector $\mathbf{x}_0^f$ and the associated error covariance matrix $\mathbf{P}_0^f$. The assimilation sequence is performed according to the Kalman filter equations:

Initialization: $\mathbf{x}_0^f$ and $\mathbf{P}_0^f$

Analysis step:

$$\mathbf{K}_k = (\mathbf{H}_k\mathbf{P}_k^f)^T[\mathbf{H}_k(\mathbf{H}_k\mathbf{P}_k^f)^T + \mathbf{R}_k]^{-1}, \tag{35a}$$

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k(\mathbf{y}_k^o - \mathbf{H}_k\mathbf{x}_k^f), \tag{35b}$$

$$\mathbf{P}_k^a = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^f. \tag{35c}$$

Forecast step:

$$\mathbf{x}_{k+1}^f = \mathbf{M}_{k,k+1}\mathbf{x}_k^a, \tag{36a}$$

$$\mathbf{P}_{k+1}^f = \mathbf{M}_{k,k+1}\mathbf{P}_k^a\mathbf{M}_{k,k+1}^T + \mathbf{Q}_k. \tag{36b}$$

Another formulation for the analysis starts with the computation of the inverse of the covariance matrix (sometimes called the *information* matrix):

$$\mathbf{P}_k^{a\,-1} = \mathbf{P}_k^{f\,-1} + \mathbf{H}_k{\mathbf{R}_k}^{-1}\mathbf{H}_k, \tag{37a}$$

$$\mathbf{K}_k = \mathbf{P}_k^a{\mathbf{H}_k}^T{\mathbf{R}_k}^{-1}, \tag{37b}$$

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k(\mathbf{y}_k^o - \mathbf{H}_k\mathbf{x}_k^f), \tag{37c}$$

but this is not usually used in geophysics.

6.5 Particular cases

6.5.1 No observations

6.5.2 Complete and perfect observations

6.6 Two illustrations

6.7 Implementation issues

6.7.1 Definition of covariance matrices, filter divergence

If the input statistical information is mis-specified, the filtering system may come to underestimate the state error variances. Too much confidence is then given to the state estimation and the effects of the analyses are then minimized. In the extreme case, observations are simply rejected. This is a filter divergence.
Very often filter divergence is quite easy to diagnose: state error variances are small and the time sequence of innovations is biased. But it is not always simple to correct. The main rule to follow is not to underestimate model errors. If possible, it is better to use an adaptive scheme to tune them online.

6.7.2 Problem dimensions

The first limitation of the straightforward implementation of the Kalman filter is the problem dimension. In oceanography or meteorology, models generally involve several millions (very often tens of millions, even hundreds of millions sometimes) of variables. Let us call n the number of variables. A state covariance matrix is then $n \times n$. With the dimensions considered, the storage of such matrix is obviously impossible. The standard solution is *rank reduction*. The theoretical description holds in two steps:

Square-root decomposition of the covariance matrix: A covariance matrix is symmetric, positive definite. It can be square-root reduced as:

$$\mathbf{P}^f = \hat{\mathrm{S}}^f \hat{\mathrm{S}}^{f^T} \tag{38}$$

where $\hat{\mathrm{S}}^f$ is a $n \times n$ matrix. It is not unique: a Cholesky decomposition provides a lower triangular matrix. A singular value decomposition provides a unitary matrix multiplied by a diagonal matrix (holding the square roots

of the eigenvalues of $\mathbf{P}^f$. But anyway these methods can rarely be applied, because $\mathbf{P}^f$ cannot be explicited and stored.

Rank reduction: This consists in reducing by several orders of magnitude the number of columns of the square root matrix. A number m of typically a hundred of columns are considered and form a $n \times m$ matrix that we call $\mathbf{S}$ (f or a).

6.7.3 Evolution of the state error covariance matrix

The matrix propagation equation 36b theoretically provides a symmetric matrix. But its numerical implementation may not in practice. An example is: $\mathbf{W} = \mathbf{M}_{k,k+1}\mathbf{P}^a_k$, then $\mathbf{P}^f_{k+1} = \mathbf{M}_{k,k+1}\mathbf{W}^T + \mathbf{Q}_k$. In certain circumstances numerical truncation errors may lead to an asymmetric covariance matrix and to the collapse of the filter. A simple recipe is to add an extra step that forces symmetry, for instance: $\mathbf{P}^f_{k+1} = (\mathbf{P}^f_{k+1} + {\mathbf{P}^f_{k+1}}^T)/2$. Another way is to use the square root formulation of the covariance matrix, and compute

$$\mathbf{P}^f_{k+1} = (\mathbf{M}_{k,k+1}\mathbf{S}^a_k)(\mathbf{M}_{k,k+1}\mathbf{S}^a_k)^T + \mathbf{Q}_k. \tag{39}$$

6.7.4 Nonlinear dynamics

Nonlinear dynamics poses two problems to the Kalman filter. First, the transposed model is not defined. Then, nonlinearity destroys gaussianity of statistics. The way to proceed with nonlinearity is given by the Extended Kalman Filter (EKF), presented next. One must be aware that the EKF is no more optimal, even with initially gaussian statistics, and is valid only for weakly nonlinear dynamics.

6.8 The Extended Kalman Filter (EKF)

When the dynamical model $\mathcal{M}$ and the observation operator $\mathcal{H}$ are (weakly) nonlinear, the Kalman is said to be extended (to nonlinear models). $\mathbf{M}$ and $\mathbf{H}$ denote the tangent linear models of $\mathcal{M}$ and $\mathcal{H}$ respectively.

Initialization: $\mathbf{x}^f_0$ and $\mathbf{P}^f_0$
Analysis step:

$$\begin{aligned}
\mathbf{K}_k &= (\mathbf{H}_k\mathbf{P}^f_k)^T[\mathbf{H}_k(\mathbf{H}_k\mathbf{P}^f_k)^T + \mathbf{R}_k]^{-1}, && \text{(40a)}\\
\mathbf{x}^a_k &= \mathbf{x}^f_k + \mathbf{K}_k(\mathbf{y}^o_k - \mathcal{H}_k(\mathbf{x}^f_k)), && \text{(40b)}\\
\mathbf{P}^a_k &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}^f_k. && \text{(40c)}
\end{aligned}$$

Forecast step:

$$\mathbf{x}^f_{k+1} = \mathcal{M}_{k,k+1}(\mathbf{x}^a_k), \tag{41a}$$

$$\mathbf{P}^f_{k+1} = \mathbf{M}_{k,k+1}\mathbf{P}^a_k\mathbf{M}^T_{k,k+1} + \mathbf{Q}_k. \tag{41b}$$

7 Kalman filters for high dimensional problems

7.1 Optimal Interpolation (OI)

Optimal Interpolation (OI) can be seen as an approximation to the Kalman filter where the state error covariance matrix is not propagated using the dynamics of the physical model. But the matrix may change with time in other ways. When they were running OI, Numerical Weather Prediction centers had a different covariance matrix for each month typically. OI reaches its limits when the dynamics of the day significantly determine the covariance errors. OI is still in use in several oceanographic operational centers. There are two approaches to properly form the (static) covariance matrix.

7.1.1 Analytical formulation

The covariance matrix is formed from a vector of variances and a correlation matrix:

$$\mathbf{P} = \mathbf{D}^{1/2}\mathbf{C}\mathbf{D}^{1/2}$$

where $\mathbf{D}$ is a diagonal matrix holding the variances and $\mathbf{C}$ is a correlation matrix to be defined. One example is:

$$\mathbf{C}(i,j) = (1 + al + \frac{1}{3}a^2l^2)\exp(-al)$$

where a is a tunable parameter and l is the distance between the grid points i and j.

7.1.2 Asymptotic approximation

Gathering together the propagation and update equations of the Kalman filter (equations 35a, 35c, and 36b), the Ricatti equation is obtained:

$$\mathbf{P}^f_{k+1} = \mathbf{M}_{k,k+1}\left(\mathbf{P}^f_k - \mathbf{H}_k\mathbf{P}^f_k)^T[\mathbf{H}_k(\mathbf{H}_k\mathbf{P}^f_k)^T + \mathbf{R}_k]^{-1}\mathbf{H}_k\mathbf{P}^f_k\right)\mathbf{M}^T_{k,k+1} + \mathbf{Q}_k$$

what is a recurrence relation for the series $\left(\mathbf{P}^f_k\right)_k$. Under certain conditions not detailed here, this series converges towards a covariance matrix $\mathbf{B}$. The

recurrence relation can be iterated once before the assimilation sequence so as to find the limit $\mathbf{B}$. $\mathbf{B}$ is then used as the static covariance matrix in OI applications.

7.2 The "stochastic" Ensemble Kalman Filter (EnKF)

The Ensemble Kalman Filter (EnKF) can be seen as an hybrid algorithm between the Kalman filter and the particle filter. In the EnKF, the pdf is represented by a sample (called the ensemble) of states (the members). The dynamical propagation is performed in the same way as in the particle filter. But the analysis step is computed in Kalman's fashion: the BLUE equations are applied to each member of the ensemble. For consistency with the observation error covariance matrix, the observations used need to be noised accordingly. In equation **??** below, ϵ_i^o are random drawings from the gaussian distribution $\mathcal{N}(0, \mathbf{R}_k)$. All the other statistical information necessary to the BLUE is calculated from the ensemble. Let m be the number of members in the ensemble, and i a member index running from 1 to m. The analysis equations are:

Initialization: $\mathbf{x}_{0,i}^f$

Analysis step:

$$\begin{aligned}
\bar{\mathbf{x}^f}_k &= \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_{k,i}^f \\
\mathbf{P}_k^f &= \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathbf{x}_{k,i}^f - \bar{\mathbf{x}^f}\right)\left(\mathbf{x}_{k,i}^f - \bar{\mathbf{x}^f}\right)^T \\
\mathbf{H}_k\mathbf{P}_k^f &= \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathcal{H}_k(\mathbf{x}_{k,i}^f) - \mathcal{H}_k(\bar{\mathbf{x}^f})\right)\left(\mathbf{x}_{k,i}^f - \bar{\mathbf{x}^f}\right)^T \\
\mathbf{H}_k\mathbf{P}_k^f{\mathbf{H}_k}^T &= \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathcal{H}_k(\mathbf{x}_{k,i}^f) - \mathcal{H}_k(\bar{\mathbf{x}^f})\right)\left(\mathcal{H}_k(\mathbf{x}_{k,i}^f) - \mathcal{H}_k(\bar{\mathbf{x}^f})\right)^T \\
\mathbf{K}_k &= (\mathbf{H}_k\mathbf{P}_k^f)^T[\mathbf{H}_k(\mathbf{H}_k\mathbf{P}_k^f)^T + \mathbf{R}_k]^{-1} \\
\mathbf{y}_{k,i}^o &= \mathbf{y}_k^o + \epsilon_i^o \\
\mathbf{x}_{k,i}^a &= \mathbf{x}_{k,i}^f + \mathbf{K}_k(\mathbf{y}_{k,i}^o - \mathcal{H}_k\mathbf{x}_{k,i}^f)
\end{aligned}$$

Forecast step:

$$\mathbf{x}_{k+1,i}^f = \mathcal{M}_{k,k+1}(\mathbf{x}_{k,i}^a) + \eta_{k,i} \tag{42}$$

The problem of storing the state covariance matrix, mentionned in section 6.7.2, is solved. "Only" m state vectors are stored. In the standard EnKF, the inversion of the innovation error covariance matrix is still required to compute the Kalman gain. The dimension of this matrix is $s \times s$, s being the number of observations. In real problems, s may easily become of the order of a few hundred, what makes the inversion prohibitive. The usual strategy to tackle this problem is to localize the analysis, i.e., to consider, for the correction at one grid point, only the observations present within a limited region in the close environment. Thus, the Kalman gain is different and must be recomputed for each grid point. But the local innovation error covariance matrix is of low dimension and its inversion is possible. Localization is a very important aspect in high dimensional Kalman filtering. As discussed later, it is not only useful for computational purpose.

7.3 The Ensemble Transform Kalman Filter (ETKF)

There are several reduced rank square root filters described in the literature. They generally have many common aspects and differ mostly in implementation details. Below is described the ETKF, also known as the Singular Evolutive Extended Kalman (SEEK) filter.

The analysis equations are re-written using the SMW formula (equation 34) with $\mathbf{U} = \mathbf{H}_k\mathbf{S}_k^f$ and $\mathbf{D} = \mathbf{I}$. The analysis equations become:

$$\begin{aligned}
\mathbf{\Gamma}_k &= (\mathbf{H}_k\mathbf{S}_k^f)^T\mathbf{R}_k^{-1}(\mathbf{H}_k\mathbf{S}_k^f), && \text{(43a)}\\
\mathbf{d}_k &= \mathbf{y}_k^o - \mathbf{H}_k\mathbf{x}_k^f, && \text{(43b)}\\
\mathbf{K}_k &= \mathbf{S}_k^f[\mathbf{I} + \mathbf{\Gamma}_k]^{-1}(\mathbf{H}_k\mathbf{S}_k^f)^T\mathbf{R}_k^{-1}, && \text{(43c)}\\
\mathbf{x}_k^a &= \mathbf{x}_k^f + \mathbf{K}_k\mathbf{d}_k, && \text{(43d)}\\
\mathbf{S}_k^a &= \mathbf{S}_k^f(\mathbf{I} + \mathbf{\Gamma}_k)^{-1/2}. && \text{(43e)}
\end{aligned}$$

This scheme is efficient when the observation error covariance matrix $\mathbf{R}$ is easily invertible. Most often it is considered diagonal. The scheme requires the inversion of a $m \times m$ matrix, m being the number of columns in $\mathbf{S}_k^f$. The algorithm is implemented so that the costly calculation ($U - D$ decomposition of $\mathbf{\Gamma}$) is done once, and the storage requirements are minimized (the subscripts are dropped for conciseness, although each element still depends on time):

1. Computation of $\mathbf{\Gamma}$:

$$\mathbf{\Gamma} = (\mathbf{H}\mathbf{S}^f)^T\mathbf{R}^{-1}(\mathbf{H}\mathbf{S}^f) \qquad (44)$$

2. $U - D$ decomposition of $\mathbf{\Gamma}$:

$$\mathbf{\Gamma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{45}$$

3. Computation of the innovation in the reduced space:

$$\delta = (\mathbf{H}\mathbf{S}^f)^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}^f) \tag{46}$$

4. Computation of the correction in the reduced space:

$$\gamma = \mathbf{U}[\mathbf{I} + \mathbf{\Lambda}]^{-1}\mathbf{U}^T\delta \tag{47}$$

5. Computation of $\mathbf{S}^f$ to $\mathbf{S}^a$ transformation matrix:

$$\mathbf{L} = \mathbf{U}[\mathbf{I} + \mathbf{\Lambda}]^{-1/2}\mathbf{U}^T \tag{48}$$

6. Computation of the analysis state:

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{S}^f\gamma \tag{49}$$

7. Computation of the analysis error covariance matrix:

$$\mathbf{S}^a = \mathbf{S}^f\mathbf{L} \tag{50}$$

The columns of $\mathbf{S}_k^f$ or $\mathbf{S}_k^a$ are called "error modes". The corrections brought to the forecast state are linear combinations of these modes.

In theory, the forecast step requires the linearized model (see equation 41b). In practice though, this is avoided using a finite difference approach: With $\mathbf{P}_{k-1}^a = \mathbf{S}_{k-1}^a{\mathbf{S}_{k-1}^a}^T$, we have

$$\mathbf{P}_k^f = \mathbf{S}_k^f{\mathbf{S}_k^f}^T$$

, if $\mathbf{S}_k^f = \mathbf{M}_k\mathbf{S}_{k-1}^a$, and each column j of $\mathbf{S}_k^f$ is computed as

$$\mathbf{S}_{k,j}^f = \mathcal{M}_{k-1,k}(\mathbf{x}_{k-1}^a + \mathbf{S}_{k-1,j}^a) - \mathcal{M}_{k-1,k}(\mathbf{x}_{k-1}^a)$$

A difficult step is to deal with model errors. Because of its size, the full matrix $\mathbf{M}\mathbf{P}^a\mathbf{M}^T$ cannot be form so that $\mathbf{Q}$ would be added. The simplest way to add model error is called *covariance inflation* of forgetting factor approach. It consists in multiplying $\mathbf{S}_k^f$ by a factor slightly higher than 1.

7.4 A key issue: Localization

Localization is not only useful to compute the Kalman gain in the EnKF. It has two other advantages:

- it prevents corrections due to distant observations. Such corrections are due to significant correlations between distant grid points. But these correlations are very often due more to the effect of subsampling rather than real physical and statistical reasons.
- it has the effect of enlarging the rank of the covariance matrix, which is basically of m at best. The correction at each grid point is therefore less dependent of those at the other grid points, the distant grid points in particular.

Two implementations are possible: B localization, and R localization, also called *Covariance Localization* and *Analysis Localization*, respectively. But the latter names may be confusing.

7.5 Covariance inflation

8 Linear smoothers

8.1 Smoothing problems and types of smoothers

8.2 Sequential smoother

8.3 RTS smoother

8.4 Forward-backward smoother

8.5 Ensemble smoother

A Optimal estimation and the BLUE

We here focus on the analysis step, where two pieces of information are merged together to improve our knowledge of the system state. Very often we do not have access to the full probability densities, but rather to an estimate with some uncertainty. Then the problem reduces to finding an optimal estimate of the system state. Optimality can be defined in different ways. Three optimal estimates are presented below, but we will then focus on the first one only.

A.1 Optimal estimation

The optimal estimate of the random variable $\mathbf{x}$ given the observation $\mathbf{y}$ is the value that best reflects what a realization of $\mathbf{x}$ can really be in regard to $\mathbf{y}$. This definition is subjective, so that several criteria can be proposed to define optimality. For illustration three optimal estimators are presented below, although in the rest of this course only the minimum variance estimator will be considered.

A.1.1 Minimum variance estimation

The estimate is defined such as the spread around it is minimal. The measure of the spread is the variance. If $p(\mathbf{x}|\mathbf{y})$ is the pdf of $\mathbf{x}$, the minimum variance estimate $\hat{\mathbf{x}}_{MV}$ is the solution to:

$$\frac{\partial \mathcal{J}(\hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}} = 0 \tag{51}$$

where

$$\mathcal{J}(\hat{\mathbf{x}}) = \int (\mathbf{x} - \hat{\mathbf{x}})^T (\mathbf{x} - \hat{\mathbf{x}}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x}.$$

It is easy to show that the solution is the expectation of the pdf, $\hat{\mathbf{x}}_{MV} = \mathcal{E}(\mathbf{x}|\mathbf{y})$.

A.1.2 Maximum *A Posteriori* estimation

The estimate is defined as the most probable value of $\mathbf{x}$ given $\mathbf{y}$, i.e., the value that maximizes the conditional pdf $p(\mathbf{x}|\mathbf{y})$. $\hat{\mathbf{x}}_{MAP}$ is such that

$$\frac{\partial p(\mathbf{x}|\mathbf{y})}{\partial \mathbf{x}} = 0 \tag{52}$$

With a Gaussian pdf, the minimum variance and the Maximum A Posteriori estimators are identical.

A.1.3 Maximum Likelihood estimation

The estimate is defined as the most probable value of $\mathbf{y}$ given $\mathbf{x}$, i.e., the value that maximizes the conditional pdf $p(\mathbf{y}|\mathbf{x})$. $\hat{\mathbf{x}}_{ML}$ is such that

$$\frac{\partial p(\mathbf{y}|\mathbf{x})}{\partial \mathbf{x}} = 0 \tag{53}$$

The ML estimator can be seen as the MAP estimator without any prior information $p(\mathbf{x})$.

A.2 The best linear unbiased estimate (BLUE)

Here we aim at estimating the true state $\mathbf{x}$ of a system, assuming a background estimate $\mathbf{x}^b$ and a partial observation $\mathbf{y}^o$ are given. These data are assumed unbiased and their uncertainties are also given, in the form of the covariance matrices $\mathbf{P}^b$ and $\mathbf{R}$, respectively. The observation operator is assumed linear. To summarize, we have the following pieces of information:

$$\begin{aligned} \mathbf{H} \quad , \quad \text{with} \qquad \mathbf{y}^o &= \mathbf{H}\mathbf{x} + \epsilon^o && (54)\\ \mathbf{x}^b &= <\mathbf{x}> && (55)\\ \mathbf{P}^b &= <\epsilon^b \epsilon^{bT}> && (56)\\ <\epsilon^o> &= = 0 && (57)\\ \mathbf{R} &= <\epsilon^o \epsilon^{oT}> && (58)\\ & && (59) \end{aligned}$$

The best estimate is searched for as a linear combination of the background estimate and the observation:

$$\mathbf{x}^a = \mathbf{A}\mathbf{x}^b + \mathbf{K}\mathbf{y}^o \tag{60}$$

where $\mathbf{A}$ and $\mathbf{K}$ are to be determined to make the estimation optimal. In that goal, some criteria must be formulated to define optimality. Given the information provided here, a wise choice is to search for an unbiased estimate, with minimum variance. Thus we try to find $\mathbf{A}$ and $\mathbf{K}$ that makes:

$$<\epsilon^a> = 0 \tag{61}$$

$$\mathrm{tr}(\mathbf{P}^a) \text{ minimum} \tag{62}$$

These requirements are reached when

$$\begin{aligned} \mathbf{A} &= \mathbf{I} - \mathbf{K}\mathbf{H} && (63)\\ \mathbf{K} &= \mathbf{P}^b\mathbf{H}^T(\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R})^{-1} && (64) \end{aligned}$$

and $\mathbf{K}$ is called the *Kalman gain*. The *a posteriori* covariance matrix can also be computed. The final form of the update equations is

$$\begin{aligned} \mathbf{K} &= \mathbf{P}^b\mathbf{H}^T(\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R})^{-1} && (65)\\ \mathbf{x}^a &= \mathbf{x}^b + \mathbf{K}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b) && (66)\\ \mathbf{P}^a &= (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^b && (67) \end{aligned}$$

and this constitutes the Best Linear Unbiased Estimate (BLUE) equations, under the constraint of the minimum variance.