Chapter 5

Introduction to Estimation Theory

The general problem of data assimilation is this: given a set of observations and a model of some physical parameters, what does knowledge of the observations tell us about the model state? Let us represent the observations by an m-vector, $\mathbf{z}$, and the model state by a n-vector, $\mathbf{x}$. Then, the information we want to know is given by the conditional p.d.f. (probability density function):

$$p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z}) = \frac{p_{\mathbf{xz}}(\mathbf{x},\mathbf{z})}{p_{\mathbf{z}}(\mathbf{z})} \tag{5.1}$$

where $p_{\mathbf{z}}(\mathbf{z}) \neq 0$. But this is very difficult to obtain, in practice, especially if we are considering large, complex models (as we do in environmental applications). Are there then, important attributes of $p_{\mathbf{x}|\mathbf{z}}$ which can help us estimate $\mathbf{x}$? Let us call this estimate $\hat{\mathbf{x}}$ to distinguish it from the random variable, $\mathbf{x}$. This estimate will depend on what has actually been observed. Thus, when

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{z})$$

is viewed as a function of $\mathbf{z}$, $\hat{\mathbf{x}}$ is termed an *estimator*. An estimator tells us what the best estimate of $\hat{x}$ is, for a given $\boldsymbol{z}$, based on some principle.

Estimators can have various desirable properties. One such property is *unbiasedness*, i.e.

$$E(\mathbf{x}^t - \hat{\mathbf{x}}) = 0, \tag{5.2}$$

where $\mathbf{x}^t$ is the true value of $\mathbf{x}$ projected onto the model basis. If $\mathbf{x}$ is biased, the bias is defined by:

$$b(\hat{\mathbf{x}}) = E(\mathbf{x}^t - \hat{\mathbf{x}}) = E(\mathbf{x}^t) - \hat{\mathbf{x}}. \tag{5.3}$$

Since the expectation is with respect to $\mathbf{x}$, and our estimator is a function of $\mathbf{z}$ only, the mean of the estimator is the estimator.

If we have two estimators, the one with lower variance is preferred. However, if the two estimators are biased, then the one with the least variance is not necessarily preferable, since it may also have a larger bias. In this case, the one with the lowest *mean square error* may be preferable. The mean square error or MSE is defined by

$$\begin{aligned}
\text{MSE} &= E[(\mathbf{x}^t - \hat{\mathbf{x}})^2] = E[(\mathbf{x}^t - E(\mathbf{x}^t) + E(\mathbf{x}^t) - \hat{\mathbf{x}})^2] \\
&= E[(\mathbf{x}^t - E(\mathbf{x}^t) + b(\mathbf{x}^t))^2] \\
&= E[(\mathbf{x}^t - E(\mathbf{x}^t))^2] + E[b(\hat{\mathbf{x}})^2] + 2E[(\mathbf{x}^t - E(\mathbf{x}^t))b(\hat{\mathbf{x}})] \\
&= E[(\mathbf{x}^t - E(\mathbf{x}^t))^2] + b(\hat{\mathbf{x}})^2 + 2E[\mathbf{x}^t - E(\mathbf{x}^t)]b(\hat{\mathbf{x}}) \\
&= var(\mathbf{x}^t) + b(\hat{\mathbf{x}})^2.
\end{aligned} \tag{5.4}$$

For unbiased estimators, the MSE is equal to the variance, but for biased estimators, the MSE equals the variance plus the square of the bias. The RMS or Root Mean Square Error is defined by:

$$\text{RMS} = \sqrt{\text{MSE}}.$$

5.1 Example

Consider the measurement equation,

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{v}, \tag{5.5}$$

where $\mathbf{z}$ is the m-dimensional observation vector, $\mathbf{x}$ is the n-dimensional prior or background estimate of the state, and $\mathbf{v}$ is the m-dimensional vector of observation errors. $\mathbf{H}$ is an $m \times n$ matrix representing the mapping from model to observed variables and locations. $\mathbf{H}$ is called the "observation operator" or "forward model", and includes the transformation to observed variables (here assumed to be linear) and a spatial interpolation from model space to observation locations. $\mathbf{x}$ and $\mathbf{v}$ are both random variables, and we shall, for this example, assume them to be Gaussian, and independent of each other. Thus $\mathbf{x}$ is N($\boldsymbol{\mu}$,$\mathbf{P}$) and $\mathbf{v}$ is N($\mathbf{0}$,$\mathbf{R}$) where $\mathbf{P}$ is $n \times n$ and $\mathbf{R}$ is $m \times m$. Thus the observation error is assumed to be unbiased, and the background estimate is $\boldsymbol{\mu}$. For a given set of observations, $\mathbf{z}$, we want to know

$$p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z}) = \frac{p_{\mathbf{z}\mathbf{x}}(\mathbf{z},\mathbf{x})}{p_{\mathbf{z}}(\mathbf{z})} = \frac{p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x})p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{z}}(\mathbf{z})}. \tag{5.6}$$

Now,

$$p_{\mathbf{z}\mathbf{x}}(\mathbf{z},\mathbf{x}) = p_{\mathbf{v}\mathbf{x}}(\mathbf{v},\mathbf{x})\text{Jac}\left(\frac{\mathbf{v},\mathbf{x}}{\mathbf{z},\mathbf{x}}\right)$$

where

$$\text{Jac}\left(\frac{\mathbf{v},\mathbf{x}}{\mathbf{z},\mathbf{x}}\right) = \left|\begin{array}{cc} \frac{\partial \mathbf{v}}{\partial \mathbf{z}} & \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \\ \frac{\partial \mathbf{v}}{\partial \mathbf{x}} & \frac{\partial \mathbf{x}}{\partial \mathbf{x}} \end{array}\right| = \left|\begin{array}{cc} \mathbf{I} & (\mathbf{H}^{\mathrm{T}})^{-1} \\ \mathbf{0} & \mathbf{I} \end{array}\right| = |\mathbf{I}| = 1. \tag{5.7}$$

Thus, the joint p.d.f. can be written as:

$$\begin{aligned} p_{\mathbf{z}\mathbf{x}}(\mathbf{z},\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}}\frac{1}{(2\pi)^{m/2}|\mathbf{R}|^{1/2}} \\ &\times \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \end{aligned} \tag{5.8}$$

Since $\mathbf{z}$ is a linear combination of Gaussian r.v.s, $\mathbf{z}$ is Gaussian with mean

$$E(\mathbf{z}) = E(\mathbf{H}\mathbf{x}+\mathbf{v}) = \mathbf{H}\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathbf{z}}$$

and covariance matrix

$$\begin{aligned} \mathbf{P}_{\mathbf{z}} &= E[(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})^{\mathrm{T}}] = E[(\mathbf{H}\mathbf{x}+\mathbf{v}-\mathbf{H}\boldsymbol{\mu})(\mathbf{H}\mathbf{x}+\mathbf{v}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}] \\ &= E[\mathbf{H}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}] + E[\mathbf{v}\mathbf{v}^{\mathrm{T}}] \\ &= \mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}} + \mathbf{R}. \end{aligned} \tag{5.9}$$

To obtain the second line, we used the fact that $\mathbf{z}$ and $\mathbf{x}$ are independent, i.e.

$$E[(\mathbf{x}-\boldsymbol{\mu})\mathbf{v}^{\mathrm{T}}] = 0.$$

Thus, we can immediately write down the p.d.f of $\mathbf{z}$:

$$p_{\mathbf{z}}(\mathbf{z}) = \frac{1}{(2\pi)^{m/2}|\mathbf{HPH}^{\mathrm{T}}+\mathbf{R}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{HPH}^{\mathrm{T}}+\mathbf{R})^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})\right\}. \tag{5.10}$$

Now we can simply substitute (5.8) and (5.10) into (5.6) to find the conditional probability:

$$p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z}) = \frac{|\mathbf{HPH}^{\mathrm{T}}+\mathbf{R}|^{1/2}}{(2\pi)^{m/2}|\mathbf{P}|^{1/2}|\mathbf{R}|^{1/2}} \exp\left\{-\frac{1}{2}J\right\} \tag{5.11}$$

where

$$\begin{aligned} J &= (\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x}) + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{HPH}^{\mathrm{T}}+\mathbf{R})^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu}). \end{aligned} \tag{5.12}$$

Now we can write

$$J = (\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{P}_x^{-1}(\mathbf{x}-\hat{\mathbf{x}}) \tag{5.13}$$

where

$$\hat{\mathbf{x}} = \mathbf{P}_x(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} + \mathbf{P}^{-1}\boldsymbol{\mu}) \tag{5.14}$$

and

$$\mathbf{P}_x^{-1} = \mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}. \tag{5.15}$$

A proof of this is given in the appendix. Thus the conditional p.d.f. we are looking for is

$$p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z}) = \frac{|\mathbf{HPH}^{\mathrm{T}}+\mathbf{R}|^{1/2}}{(2\pi)^{m/2}|\mathbf{P}|^{1/2}|\mathbf{R}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{P}_x^{-1}(\mathbf{x}-\hat{\mathbf{x}})\right\} \tag{5.16}$$

There is also another way to get the numerator of (5.6). Since $\mathbf{z}$ and $\mathbf{x}$ are Gaussian, then $p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x})$ is also Gaussian. So, we need only determine its mean and covariance to completely determine its p.d.f. The conditional mean is

$$\begin{aligned} E(\mathbf{z}|\mathbf{x}) &= E((\mathbf{H}\mathbf{x}+\mathbf{v})|\mathbf{x}) \\ &= E(\mathbf{H}\mathbf{x}|\mathbf{x}) + E(\mathbf{v}|\mathbf{x}) \\ &= \mathbf{H}\mathbf{x} + E(\mathbf{v}) = \mathbf{H}\mathbf{x} \end{aligned}$$

since $\mathbf{x}$ and $\mathbf{v}$ are independent. Also, since $\mathbf{z} = \mathbf{H}\mathbf{x}+\mathbf{v}$,

$$\begin{aligned} E((\mathbf{z}-\mathbf{H}\mathbf{x})(\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathrm{T}}|\mathbf{x}) &= E(\mathbf{v}\mathbf{v}^{\mathrm{T}}|\mathbf{x}) \\ &= E(\mathbf{v}\mathbf{v}^{\mathrm{T}}) \\ &= \mathbf{R}. \end{aligned}$$

So the complete p.d.f. is

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}) = \frac{1}{(2\pi)^{m/2}|\mathbf{R}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x})\right\}.$$

Now multiply by $p_x(\mathbf{x})$ to get

$$\begin{aligned} p_{\mathbf{z},\mathbf{x}}(\mathbf{z},\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}}\frac{1}{(2\pi)^{m/2}|\mathbf{R}|^{1/2}} \\ &\times \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \end{aligned}$$

In summary, we have assumed that both the background state, $\mathbf{x}$, and the observation error, $\mathbf{v}$ are Gaussian r.v.s, and because of this assumption, we were able to write down the complete p.d.f. of $\mathbf{x}$ given $\mathbf{z}$. However, for non-Gaussian r.v.s, it will not be so easy to write down the complete p.d.f. The best that we could hope for is some parameter defining the p.d.f. such as its mean, mode or median. In what follows, we shall define some estimators of $\mathbf{x}$ by minimizing the expected value of a cost function. By choosing different cost functions, we obtain different estimators.

5.2 Bayesian approach

How can we choose an estimator? The best estimator will depend on what we call "best". So, let us devise some objective rules which define "best". To do this, let us define a risk function, J, and try to minimize the risk or expected cost function, J:

$$\begin{aligned} \mathcal{J}(\hat{\mathbf{x}}) &= E(J(\tilde{\mathbf{x}})) = \int_{-\infty}^{\infty} J(\tilde{\mathbf{x}}) p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(\tilde{\mathbf{x}}) p_{\mathbf{x}\mathbf{z}}(\mathbf{x},\mathbf{z}) d\mathbf{z} d\mathbf{x} \end{aligned} \tag{5.17}$$

$\tilde{\mathbf{x}}$ will soon be defined as the error of the estimator. Thus, we would like to choose cost functions that are large for large $|\tilde{\mathbf{x}}|$ but small when $\tilde{\mathbf{x}}$ is near 0. Two common cost functions are the quadratic cost function and the uniform cost function. The quadratic cost function is

$$J(\tilde{\mathbf{x}}) = |\tilde{\mathbf{x}}|_S^2 = \tilde{\mathbf{x}}^{\mathrm{T}}\mathbf{S}\tilde{\mathbf{x}} \tag{5.18}$$

where $\mathbf{S}$ is a non-negative definite, symmetric matrix (e.g. of scaling factors, or an error covariance matrix). The uniform cost function is

$$J(\tilde{\mathbf{x}}) = \begin{cases} 0 & |\tilde{\mathbf{x}}| < \epsilon \\ \frac{1}{2\epsilon} & |\tilde{\mathbf{x}}| \geq \epsilon \end{cases} \tag{5.19}$$

5.2.1 Minimum variance

The minimum variance estimate minimizes the risk function associated with the quadratic cost function, (5.18). The estimation error is

$$\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$$

so the variance of estimate is

$$J(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^{\mathrm{T}}\mathbf{S}\tilde{\mathbf{x}} = (\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{S}(\mathbf{x}-\hat{\mathbf{x}}).$$

Thus we want to minimize:

$$\begin{aligned}\mathcal{J}(\hat{\mathbf{x}}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(\tilde{\mathbf{x}})p_{\mathbf{xz}}(\mathbf{x},\mathbf{z})d\mathbf{x}d\mathbf{z} \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}(\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{S}(\mathbf{x}-\hat{\mathbf{x}})p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x}\right]p_{\mathbf{z}}(\mathbf{z})d\mathbf{z}. \qquad (5.20)\end{aligned}$$

Since the outer integral does not involve, $\hat{\mathbf{x}}$, we can minimize

$$\mathcal{J}_{\mathrm{MV}}(\hat{\mathbf{x}}|\mathbf{z}) = \int_{-\infty}^{\infty}(\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{S}(\mathbf{x}-\hat{\mathbf{x}})p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x}$$

with respect to $\hat{\mathbf{x}}$. Setting the first derivative to zero yields

$$0 = -2\mathbf{S}\int_{-\infty}^{\infty}(\mathbf{x}-\hat{\mathbf{x}}_{\mathrm{MV}})p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x}$$

or

$$\int_{-\infty}^{\infty}\hat{\mathbf{x}}_{\mathrm{MV}}p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x} = \int_{-\infty}^{\infty}\mathbf{x}p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x}. \qquad (5.21)$$

But, the left hand side can be written

$$\int_{-\infty}^{\infty}\hat{\mathbf{x}}_{\mathrm{MV}}p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x} = \hat{\mathbf{x}}_{\mathrm{MV}}\int_{-\infty}^{\infty}p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x} = \hat{\mathbf{x}}_{\mathrm{MV}} \qquad (5.22)$$

and the right side is

$$\int_{-\infty}^{\infty}\mathbf{x}p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x} = E(\mathbf{x}|\mathbf{z}) \qquad (5.23)$$

so that

$$\hat{\mathbf{x}}_{\mathrm{MV}} = E(\mathbf{x}|\mathbf{z}). \qquad (5.24)$$

The minimum variance estimate is the conditional mean of the $p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})$. This p.d.f. is called the *a posteriori* distribution, because the estimate is obtained after the observations are known. It is interesting to note that this estimator is independent of the scaling matrix, $\mathbf{S}$. However, when proving that the minimum variance estimate is really a minimum, (by computing the second derivative at $\hat{\mathbf{x}}_{\mathrm{MV}}$), we use the fact that $\mathbf{S}$ is non-negative definite, so $\mathbf{S}$ is not arbitrary.

It is easy to show that the minimum variance estimator is unbiased since

$$E(\mathbf{x}-\hat{\mathbf{x}}_{\mathrm{MV}}) = E(\mathbf{x}-E(\mathbf{x}|\mathbf{z})) = E(\mathbf{x}) - E(E(\mathbf{x}|\mathbf{z})) = 0 \qquad (5.25)$$

where we used (2.42), the "chain rule" for conditional expectation.

Optimal Interpolation or OI is an example of a data assimilation scheme that uses a minimum variance estimator. The "optimal" in the name refers to this fact. However, optimality occurs only

when the error statistics are known. If they are only approximately estimated, then the optimality will not hold, and the scheme is referred to as "Statistical Interpolation" or SI.

In our example, we chose Gaussian error statistics, so we know the complete *a posteriori* distribution. The minimum variance estimator is the the mean of (5.16):

$$\hat{\mathbf{x}}_{\mathrm{MV}} = E(\mathbf{x}|\mathbf{z}) = (\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} + \mathbf{P}^{-1}\boldsymbol{\mu}). \tag{5.26}$$

Using the Sherman-Morrison-Woodbury formula

$$(\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} = \mathbf{P}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1} = \mathbf{K} \tag{5.27}$$

and adding and subtracting $\boldsymbol{\mu}$ from the right side gives

$$\hat{\mathbf{x}}_{\mathrm{MV}} = \boldsymbol{\mu} + \mathbf{K}(\mathbf{z} - \mathbf{H}\boldsymbol{\mu}). \tag{5.28}$$

This is the familiar OI analysis equation, where $\boldsymbol{\mu}$ is our background state, $\mathbf{x}^b$ (i.e. a 6-h forecast). In OI we try to solve for $\mathbf{K}$ using (5.27), then solve for the analysis using (5.28). Note that the solution for (5.27) involves a matrix inversion. The size of the matrix being inverted is the size of $\mathbf{R}$, i.e. $m \times m$. For the global Numerical Weather Prediction (NWP) problem, m can be O(10^5), so this matrix inversion is not possible without approximations.

Although, in our example, we assumed Gaussian error statistics and arrived at the OI equations by minimizing the analysis error variance, if we had not made the Gaussian assumption, we could still obtain the same estimator by minimizing the error variance of a *linear estimator*.

5.2.2 Maximum a posteriori probability estimation

Another estimator can be defined by using the risk function for the uniform cost function.

$$\begin{aligned}
\mathcal{J}_U(\hat{\mathbf{x}}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(\tilde{\mathbf{x}}) p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z}) d\mathbf{x} p_{\mathbf{z}}(\mathbf{z}) d\mathbf{z} \\
&= \int_{-\infty}^{\infty}\left[\frac{1}{2\epsilon}\int_{-\infty}^{\hat{\mathbf{x}}-\epsilon} p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x} + \frac{1}{2\epsilon}\int_{\hat{\mathbf{x}}+\epsilon}^{\infty} p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x}\right] p_{\mathbf{z}}(\mathbf{z})d\mathbf{z} \\
&= \int_{-\infty}^{\infty}\frac{1}{2\epsilon}\left[1 - \int_{\hat{\mathbf{x}}-\epsilon}^{\infty} p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x} + \int_{\hat{\mathbf{x}}+\epsilon}^{\infty} p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x}\right] p_{\mathbf{z}}(\mathbf{z})d\mathbf{z} \\
&= \int_{-\infty}^{\infty}\frac{1}{2\epsilon}\left[1 - \int_{\hat{\mathbf{x}}-\epsilon}^{\hat{\mathbf{x}}+\epsilon} p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x}\right] p_{\mathbf{z}}(\mathbf{z})d\mathbf{z}.
\end{aligned} \tag{5.29}$$

Again, the outer integral does not involve $\mathbf{x}$ so we can minimize

$$-\frac{1}{2\epsilon}\int_{\hat{\mathbf{x}}-\epsilon}^{\hat{\mathbf{x}}+\epsilon} p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})d\mathbf{x}$$

with respect to $\mathbf{x}$. Using the mean value theorem for integrals:

$$\frac{1}{2\epsilon}\int_{-\epsilon}^{\epsilon} f(x)dx = \frac{1}{2\epsilon}2\epsilon f(\xi) = f(\xi) \tag{5.30}$$

where ξ is some value of x in the interval $[-\epsilon, \epsilon]$, we see that we want to minimize

$$\mathcal{J}_U(\hat{\mathbf{x}}|\mathbf{z}) = -p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z}) \tag{5.31}$$

or maximize $p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})$. In other words,

$$\frac{\partial}{\partial \mathbf{x}} p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})\Big|_{\mathbf{x}=\hat{\mathbf{x}}_{\mathrm{MAP}}} = 0. \tag{5.32}$$

In this case, we are looking for the maximum or "mode" of the *a posteriori* p.d.f. of the state given the observations.

In our example where we assumed Gaussian statistics, $\hat{\mathbf{x}}_{\mathrm{MAP}}$ is the mode of the conditional Gaussian (5.16). For Gaussian, p.d.f.s we know that the mean, mode and median are all equal, so in this case the minimum variance (conditional mean) and MAP estimators are identical. Is the estimator unbiased?

$$\begin{aligned} E(\hat{\mathbf{x}}_{\mathrm{MAP}} - \mathbf{x}) &= (\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}E(\mathbf{z}) + \mathbf{P}^{-1}\boldsymbol{\mu}) - E(\mathbf{x}) \\ &= (\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\boldsymbol{\mu} + \mathbf{P}^{-1}\boldsymbol{\mu}) - \boldsymbol{\mu} \\ &= \boldsymbol{\mu} - \boldsymbol{\mu} = \mathbf{0}. \end{aligned} \tag{5.33}$$

In this example, the MAP/MV estimator is unbiased. While this is true for MV estimators in general (from (5.25)), this is not necessarily true for $\hat{\mathbf{x}}_{\mathrm{MAP}}$.

We can also determine the estimation error of the MAP/MV estimator. First note that

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{MAP}} - \mathbf{x} &= (\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} + \mathbf{P}^{-1}\boldsymbol{\mu}) - \mathbf{x} \\ &= \mathbf{P}_x(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{x} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{v} + \mathbf{P}^{-1}\boldsymbol{\mu}) - \mathbf{P}_x(\mathbf{P}_x)^{-1}\mathbf{x} \end{aligned} \tag{5.34}$$

using the measurement equation, $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{v}$ and the definition of $\mathbf{P}_x$ from (5.15). This can be further simplified to:

$$\hat{\mathbf{x}}_{\mathrm{MAP}} - \mathbf{x} = -\mathbf{P}_x\mathbf{P}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \mathbf{P}_x\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{v}. \tag{5.35}$$

Now we can easily show that

$$E[(\hat{\mathbf{x}}_{\mathrm{MAP}} - \mathbf{x})(\hat{\mathbf{x}}_{\mathrm{MAP}} - \mathbf{x})^{\mathrm{T}}] = \mathbf{P}_x = (\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1} \tag{5.36}$$

on using the fact that $\mathbf{x}$ and $\mathbf{v}$ are independent.

Let us now consider some special cases. If there is only a background and no observations, then $\mathbf{R}^{-1} = \mathbf{0}$ and

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{MAP}} = \hat{\mathbf{x}}_{\mathrm{MV}} = \boldsymbol{\mu} \\ \mathbf{P}_{\mathrm{MAP}} = \mathbf{P}_{\mathrm{MV}} = \mathbf{P}. \end{aligned} \tag{5.37}$$

Thus, just as in our scalar example of the first lecture, when there are no observations, the analysis reverts to the background $\boldsymbol{\mu}$, and the analysis error is simply the background error, $\mathbf{P}$. If, on the other hand, there is no information on the background state, then $\mathbf{P}^{-1} = \mathbf{0}$ and

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{MAP}} = \hat{\mathbf{x}}_{\mathrm{MV}} = (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} \\ \mathbf{P}_{\mathrm{MAP}} = \mathbf{P}_{\mathrm{MV}} = (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}. \end{aligned} \tag{5.38}$$

As we shall see in the next subsection, this is exactly the maximum likelihood (ML) estimate. The ML estimator is a special case of the MAP estimator, when there is no information about

the background. Thus the ML estimator, like the MAP estimator, approximates the mode of the *a posteriori* distribution of the state, given the observations. Finally, consider the case when additionally, $\mathbf{H} = \mathbf{I}$, i.e. when observations are plentiful. In this case,

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{MAP}} &= \hat{\mathbf{x}}_{\mathrm{MV}} = \mathbf{z} \\ \mathbf{P}_{\mathrm{MAP}} &= \mathbf{P}_{\mathrm{MV}} = \mathbf{R}. \end{aligned} \tag{5.39}$$

Thus, if there are lots of observations, and no information about the background, then the analysis should equal the observed state and the analysis error is equal to the observation error.

5.2.3 Maximum Likelihood estimation

Now suppose we know nothing about the error statistics of the background state. For MAP estimation, we are looking for the mode of the *a posteriori* distribution of the state given the observations. Thus, we want to maximize

$$p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z}) = \frac{p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}) p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{z}}(\mathbf{z})}$$

with respect to $\mathbf{x}$ using Bayes' theorem. We could equally well maximize the log of the *a posteriori* distribution, or

$$\ln p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}) + \ln p_{\mathbf{x}}(\mathbf{x}).$$

We omitted the term involving $p_{\mathbf{z}}(\mathbf{z})$ because it is independent of $\mathbf{x}$. If we know nothing about the p.d.f. of the background state, then it sometimes turns out that

$$\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{x}}(\mathbf{x}) = 0.$$

For example, consider the case of a Gaussian p.d.f. If $\mathbf{x}$ is N($\boldsymbol{\mu}$,$\mathbf{P}$), then

$$\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{x}}(\mathbf{x}) = -\mathbf{P}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

The state of no information corresponds to infinite error variance, or $\mathbf{P}^{-1} = \mathbf{0}$ so that $\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{x}}(\mathbf{x}) = 0$.

Maximum likelihood (ML) estimation is similar to MAP estimation, but where no *a priori* information about $\mathbf{x}$ is assumed. In ML estimation we require that

$$\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}) = 0.$$

The maximum likelihood estimate is an estimate of the mode of the *a priori* p.d.f. of the observations given the background state.

Now, returning to our example, our *a priori* p.d.f. is (5.10):

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\mathbf{z}\mathbf{x}}(\mathbf{z}, \mathbf{x})}{p_{\mathbf{x}}(\mathbf{x})} \tag{5.40}$$

and since $p_{\mathbf{z}\mathbf{x}}$ is given by (5.8), we can divide through by $p_{\mathbf{x}}$ to obtain

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}) = \frac{1}{(2\pi)^{m/2}|\mathbf{R}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{H}\mathbf{x})^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{z} - \mathbf{H}\mathbf{x})\right\}. \tag{5.41}$$

To maximize this p.d.f., we can minimize the exponent with respect to $\mathbf{x}$. i.e.

$$\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x})\Big|_{\mathbf{x}=\hat{\mathbf{x}}_{\mathrm{ML}}}=0.$$

Thus

$$\hat{\mathbf{x}}_{\mathrm{ML}}=(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z}. \tag{5.42}$$

We have already seen that the ML estimator is the same as the MAP estimator when there is no background information $\mathbf{P}^{-1}=\mathbf{0}$. Is this estimator unbiased?

$$\begin{aligned} E(\hat{\mathbf{x}}_{\mathrm{ML}}-\mathbf{x}) &= (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}E(\mathbf{z})-E(\mathbf{x}) \\ &= (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\boldsymbol{\mu}-\boldsymbol{\mu}=0. \end{aligned} \tag{5.43}$$

The ML estimator is unbiased, in this example, because the *a priori* distribution is Gaussian. For Gaussian p.d.f.s, the mode and mean coincide and the mode estimate is unbiased. For general p.d.f.s, there is no reason to expect the mode estimate to be unbiased. The estimation error is

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{ML}}-\mathbf{x} &= (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z}-\mathbf{x} \\ &= (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\mathbf{x}+\mathbf{v})-\mathbf{x} \\ &= (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{v} \end{aligned} \tag{5.44}$$

so that the estimation error covariance matrix is easily seen to be

$$\begin{aligned} E[(\hat{\mathbf{x}}_{\mathrm{ML}}-\mathbf{x})(\hat{\mathbf{x}}_{\mathrm{ML}}-\mathbf{x})^{\mathrm{T}}] &= (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}E[\mathbf{v}\mathbf{v})^{\mathrm{T}}]\mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1} \\ &= (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1} \end{aligned} \tag{5.45}$$

where we used the symmetry property of error covariance matrices.

Now, because $\mathbf{P}$ the background error covariance is non-negative definite,

$$\mathbf{P}_{\mathrm{MV}}=(\mathbf{P}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\leq(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}=\mathbf{P}_{\mathrm{ML}}.$$

The estimation error for the ML estimate is always larger than that for the MV estimate. This makes sense because the MV estimate is that which corresponds to the minimum of the Bayes risk.

5.2.4 Least squares

In the maximum likelihood method, no knowledge of background error statistics was required. What if we also have no knowledge of the observation error statistics? We simply look for a least square fit to the observations. Consider again the observation process:

$$\mathbf{z}=\mathbf{H}\mathbf{x}+\mathbf{v}, \tag{5.46}$$

where no knowledge of the error statistics of $\mathbf{v}$ is assumed. Here $\mathbf{x}$ is not a random variable, but simply a set of numbers. To minimize the error of the fit of $\mathbf{x}$ to the observations, we want to minimize

$$J(\mathbf{x})=\frac{1}{2}(\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{O}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathrm{T}}. \tag{5.47}$$

The matrix $\mathbf{O}$ is not an error covariance matrix, but just some symmetric, positive semi-definite scaling matrix that indicates the confidence level. Here J is a deterministic function and the minimum is easily found to be:

$$\hat{\mathbf{x}}_{\mathrm{LS}} = (\mathbf{H}^{\mathrm{T}}\mathbf{O}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{O}^{-1}\mathbf{z}. \tag{5.48}$$

If $\mathbf{H} = \mathbf{I}$ then $\hat{\mathbf{x}}_{\mathrm{LS}} = \mathbf{z}$. For plentiful observations, the least squares estimate closely fits the observations. Todling (1999) shows how a sequential least squares algorithm can be built. That is, suppose k observations are taken and the estimate is obtained. If a $k+1$st observation is taken, we don't want to redo the complete matrix inversion in (5.48). Todling shows that we can avoid dealing with all our past observations, if the estimate based on $k+1$ observations is obtained by linearly combining the estimate based on k observations and the $k+1$st observation.

It is evident from comparing (5.42) and (5.48) that the LS and ML estimates are equal when $\mathbf{O} = \mathbf{R}$. That is, when $\mathbf{v}$ is actually a random variable of observation errors and $\mathbf{O}$ is the error covariance matrix for $\mathbf{v}$. In general, however, Least Squares refers to a deterministic problem of fitting parameters to observations while Maximum Likelihood tries to maximize the *a priori* p.d.f. of the observations given the background.

5.3 3DVAR

Let us return to the simple scalar example of chapter 1. Given two pieces of information, x^o and x^b, each of which has unbiased errors, and variances of $(\sigma^o)^2$ and $(\sigma^b)^2$, respectively, we can solve for the weights which minimize the analysis error variance for the analysis

$$\mathrm{x}^a = \mathrm{x}^b + \mathrm{K}(\mathrm{x}^o - \mathrm{x}^b).$$

If the background and observation errors are uncorrelated, then the optimal weights are given by

$$\mathrm{K} = \frac{(\sigma^b)^2}{(\sigma^o)^2 + (\sigma^b)^2}.$$

With these weights, the analysis error variance is

$$(\sigma^a)^{-2} = (\sigma^o)^{-2} + (\sigma^b)^{-2}.$$

What if, instead of looking for a minimum variance estimator, we wanted a MAP estimator? In addition to the above assumptions, we also assume that the errors are Gaussian. Thus,

$$p(\epsilon) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

Then the joint p.d.f. of the observation and background errors (using the fact that they are independent because they are Gaussian and uncorrelated) is

$$p(\epsilon^o, \epsilon^b) = p(\epsilon^o)p(\epsilon^b) = \frac{1}{2\pi\sigma^b\sigma^o} \exp\left(-\frac{(\epsilon^o)^2}{2(\sigma^o)^2} - \frac{(\epsilon^b)^2}{2(\sigma^b)^2}\right)$$

Now, reintroduce the true state into this p.d.f.:

$$p(\mathrm{x}^t) = \frac{1}{2\pi\sigma^b\sigma^o} \exp\left(-\frac{(\mathrm{x}^o - \mathrm{x}^t)^2}{2(\sigma^o)^2} - \frac{(\mathrm{x}^b - \mathrm{x}^t)^2}{2(\sigma^b)^2}\right).$$

Which state is most likely? The one that maximizes $p(x^t)$ or minimizes

$$J(\mathrm{x}) = \left[\frac{(\mathrm{x}^o - \mathrm{x})^2}{2(\sigma^o)^2} + \frac{(\mathrm{x}^b - \mathrm{x})^2}{2(\sigma^b)^2} \right].$$

The solution is:

$$\mathrm{x}^a = \mathrm{x}^b + \frac{(\sigma^b)^2}{(\sigma^o)^2 + (\sigma^b)^2}(\mathrm{x}^o - \mathrm{x}^b)$$

which is exactly the same as that obtained in the minimum variance case. This is because for linear observation operators and Gaussian statistics, the MAP and MV estimators are identical. It may appear that additional assumptions were required in the MAP estimation (that the errors were Gaussian), compared to the MV estimation. However, the idea of minimizing variance really only makes sense for a unimodal p.d.f. so there is an implicit assumption being made when one chooses to minimize the variance of an estimator. For example, consider Fig. 5.1. A Gaussian with low variance is a tall and narrow curve. A uniform p.d.f. with low variance will also be tall and narrow. However, it is not clear that minimizing the variance of a multimodal p.d.f. is meaningful.

This simple scalar example can be made multidimensional. Given, an observation vector, $\mathbf{z}$, and a background vector $\mathbf{x}^b$ which are Normally distributed with zero means and covariance matrices $\mathbf{R}$ and $\mathbf{P}^b$, respectively, the joint p.d.f. of the observation and background errors is given by

$$p(\mathbf{e}^r, \mathbf{e}^b) = \frac{1}{(2\pi)^N \det(\mathbf{P}^b)^{1/2} \det(\mathbf{R})^{1/2}} \exp\left(-\frac{1}{2}(\mathrm{e}^b)^{\mathrm{T}}(\mathbf{P}^b)^{-1}\mathbf{e}^b - \frac{1}{2}(\mathbf{e}^r)^{\mathrm{T}}\mathbf{R}^{-1}\mathrm{e}^r \right).$$

More simply,

$$p(\mathbf{e}^r, \mathbf{r}^b) \propto \exp\left(-\frac{1}{2}(\mathbf{x}^t - \mathbf{x}^b)^{\mathrm{T}}(\mathbf{P}^b)^{-1}(\mathbf{x}^t - \mathbf{x}^b) - \frac{1}{2}(\mathbf{z} - H(\mathbf{x}^t))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z} - H(\mathbf{x}^t)) \right).$$

The maximum *a posteriori* estimate is obtained by minimizing

$$J(\mathbf{x}^t) \quad = \quad \frac{1}{2}(\mathbf{x}^t - \mathbf{x}^b)^{\mathrm{T}}(\mathbf{P}^b)^{-1}(\mathbf{x}^t - \mathbf{x}^b) + \frac{1}{2}(\mathbf{z} - H(\mathbf{x}^t))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z} - H(\mathbf{x}^t)).$$

Here J describes the 3DVAR cost function. The solution is exactly the same as the OI solution.

Let us now return to our discussion of estimation theory and put the 3DVAR algorithm in this context. As we have already seen in section 5.2.2, for MAP estimation, we want to maximimize

$$\frac{p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x})p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{z}}(\mathbf{z})}.$$

But, since $p_{\mathbf{z}}(\mathbf{z})$ is not a function of $\mathbf{x}$ it is sufficient to maximize just the numerator. Thus, in our example, assuming Gaussian statistics, the MAP estimator could have been obtained by minimizing the negative of the exponent of (5.8):

$$J_{\mathrm{MAP}}(\mathbf{x}) \quad = \quad \frac{1}{2}(\mathbf{z} - \mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\mathbf{x}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{5.49}$$

This is exactly the 3DVAR (3D-variational assimilation) cost function. The solution obtained by minimizing this cost function is the 3DVAR analysis. Thus, the 3DVAR problem solves for the maximum *a posteriori* estimator, assuming Gaussian p.d.f.s for the random error processes.

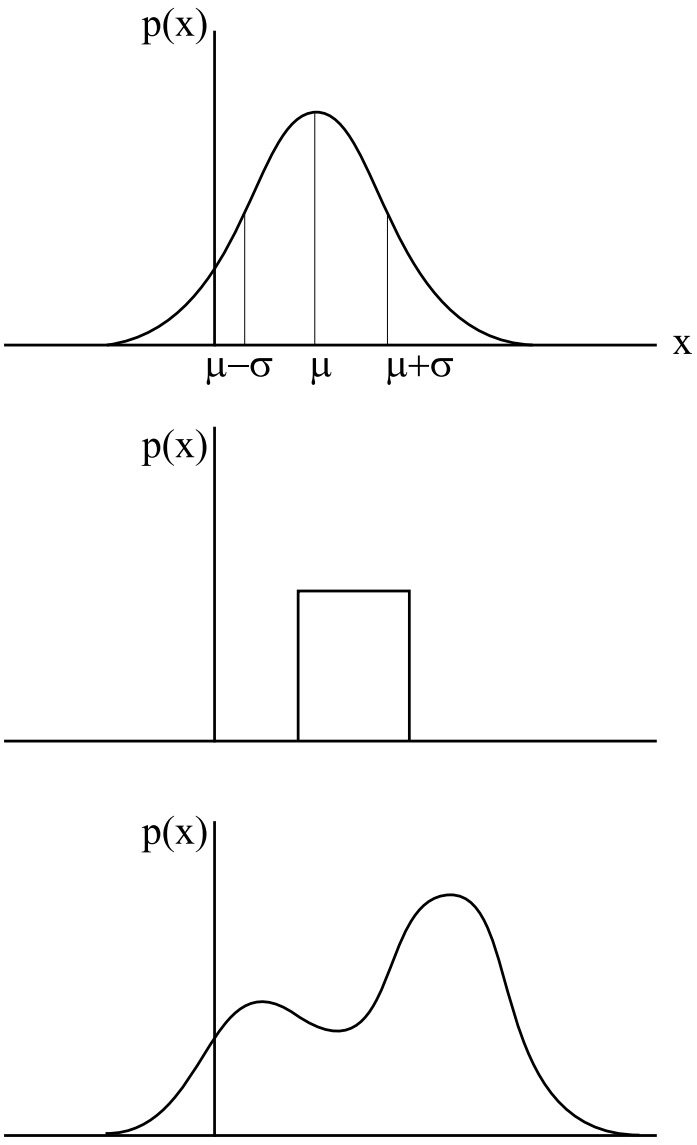

Figure 5.1: The variance of a Gaussian p.d.f. is clearly defined. For an arbitrary p.d.f., the variance may not be a useful measure. This is most evident for multimodal p.d.f.'s.

Now, because of the assumption of Gaussian statistics, the estimator is also a minimum variance estimator. In exercise 3, it is demonstrated that the solution obtained by minimizing (5.49), is exactly that given by (5.28) and (5.27). Thus, we have proved that for linear observation operators, and Gaussian background and observation errors, OI and 3DVAR are equivalent. Of course, this equivalence is only theoretical. In practice, the methodology of 3DVAR and OI are rather different. For practical applications with large state and observation vectors, approximations are made, such as data selection in OI (to limit the size of the matrix $\mathbf{HPH}^{\mathrm{T}} + \mathbf{R}$ which must be inverted), or a preset number of descent steps in the 3DVAR minimization of (5.49). Thus, it is the details of our implementation of a data assimilation algorithm that ultimately determine the performance of our scheme.

While OI solves for the gain matrix (weights to observations) directly, 3DVAR instead solves for the analysis through a minimization problem which does not involve the gain matrix. Thus 3DVAR avoids the the inversion of $\mathbf{HPH}^{\mathrm{T}} + \mathbf{R}$ which is a $m \times m$ matrix, where m is the dimension of the observation vector. 3DVAR therefore utilises all observations simultaneously, and solves for the analysis of the whole domain. Thus one of major advantages of 3DVAR over OI is that there is no more need for data selection– all data can be used. However, in reality, compromises must be made for computational reasons. Therefore, the global 3DVAR problem is only solved approximately, by limiting the number of iterations in the minimization algorithm. A maximum number of iterations can be predefined or a certain amount of reduction (say, two orders of magnitude) of $|\nabla J_{\mathrm{MAP}}(\mathbf{x})|$ can be required. Thus while OI exactly solves for many approximate local analyses, 3DVAR approximately solves the global problem. Which method is better? It will depend on the details of the particular problem, and how the assimilation scheme is implemented.

The minimization problem is solved using packaged solvers based on quasi-Newton or conjugate gradient methods, most typically. The software requires a method (subroutine) of calculating the cost function, and its gradient. The gradient of (5.49) is

$$\nabla J_{\mathrm{MAP}}(\mathbf{x}) = -\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\mathbf{x}) + \mathbf{P}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{5.50}$$

The operator, $\mathbf{H}^{\mathrm{T}}$ is the adjoint of $\mathbf{H}$ with respect to a given inner product. Adjoints will be discussed more fully later. Here, you should simply note that 3DVAR requires the adjoint of the linearized "forward model", $\mathbf{H}$. Note that $H(\mathbf{x})$ takes the model variables and converts them into observed variables. For observed variables which are linearly related to model variables, we can write

$$H(\mathbf{x}) = \mathbf{H}\mathbf{x}.$$

However, this operator can, in general, be nonlinear. In fact, the most important advantage of 3DVAR over OI, is the relative ease with which the extension for nonlinear observation operators, $H(\mathbf{x})$ can be made. In 3DVAR, simply minimize:

$$J_{\mathrm{MAP}}(\mathbf{x}) = \frac{1}{2}(\mathbf{z} - H(\mathbf{x}))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z} - H(\mathbf{x})) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{5.51}$$

To evaluate the observation term of (5.49), we use the fact that observation errors from different instrument types should be uncorrelated. However, if representativeness errors are a function of the true state, then these errors may be correlated for different observation types. Nevertheless, it is assumed that representativeness errors are not a function of the state, and that these errors are also uncorrelated for different observation types. In that case, the observation term of (5.49) may

be written as a sequence of sums:

$$J^o = \sum_{i=1}^{N_{type}} J_i^o \tag{5.52}$$

$$J_i^o = \frac{1}{2}(\mathbf{z}_i - H_i(\mathbf{x}^f))^{\mathrm{T}}\mathbf{R}_i^{-1}(\mathbf{z}_i - H_i(\mathbf{x}^f)) \tag{5.53}$$

where $\mathbf{z}_i$ is the observation vector for observation type i, and there are N_{type} different observation types. Thus $\mathbf{R}_i^{-1}$ is only needed for each observation type, and $\mathbf{R}$ is block diagonal.

In practice, the need for $\mathbf{P}^{-1}$ can avoided by redefining the control variable. If we can find $\mathbf{L}$ such that $\mathbf{P} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ then we can define a new variable $\mathbf{x} = \mathbf{L}\boldsymbol{\chi}$ and minimize (5.49) with respect to $\boldsymbol{\chi}$. In other words, minimize

$$J_{\mathrm{3dvar}}(\boldsymbol{\chi}) \;=\; \frac{1}{2}(\mathbf{z} - H(\mathbf{L}\boldsymbol{\chi}))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z} - H(\mathbf{L}\boldsymbol{\chi})) + \frac{1}{2}(\boldsymbol{\chi} - \boldsymbol{\mu}_\chi)^{\mathrm{T}}(\boldsymbol{\chi} - \boldsymbol{\mu}_\chi) \tag{5.54}$$

where $\boldsymbol{\mu}_\chi = \mathbf{L}\boldsymbol{\chi}$. The gradient of the second or background term is especially simply. Although the inversion of $\mathbf{P}^{-1}$ is avoided, access to both $\mathbf{L}$ and $\mathbf{L}^{\mathrm{T}}$ is necessary. $\mathbf{L}$ need not be a matrix since it could be very large and difficult to store. Instead, only a series of operations (subroutines) is needed where the input is $\mathbf{x}$ and the output is $\mathbf{L}\mathbf{x}$. More details about the practical implementation of 3DVAR can be found in Gauthier et al. (1999a,b), for example.

In the 1990's, 3DVAR began to replace OI at world meteorological centres. In 1991, 3DVAR was implemented operationally in the U.S., (Parrish and Derber, 1992). Other centres followed suite; first ECMWF in 1996 (Andersson et al. 1998, Courtier et al. 1998, Rabier et al. 1998), CMC in 1997 (Gauthier et al. 1999a) and the British Met Office in 2000 (Lorenc et al. 2000). 3DVAR was also implemented at Météo-France and in Australia (Steinle and Seaman 1996).

We have seen that for Gaussian statistics, OI and 3DVAR both solve the exact same analysis equations, and have the same analysis error covariance matrix. The only difference is in the method of solution. 3DVAR avoids the matrix inversion in (5.27) by solving instead, a global minimization problem (that is, over the whole Earth using all observations simultaneously). The "control variable" for the minimization is $\mathbf{x}$ which has dimension $\mathrm{O}(10^7)$ typically (for global NWP problems). Thus the minimization problem may not be solved exactly, but for only a certain number of descent steps. Of course, in OI, we could have solved for (5.28) using a variational method. To see this, write (5.28) and substitute for $\mathbf{K}$ using (5.27) to get:

$$\hat{\mathbf{x}}_{\mathrm{MV}} = \boldsymbol{\mu} + \mathbf{P}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}(\mathbf{z} - \mathbf{H}\boldsymbol{\mu}). \tag{5.55}$$

This equation can be solved in two stages:

$$\hat{\mathbf{x}}_{\mathrm{MV}} = \boldsymbol{\mu} + \mathbf{P}\mathbf{H}^{\mathrm{T}}\mathbf{y} \tag{5.56}$$

$$(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}} + \mathbf{R})\mathbf{y} = (\mathbf{z} - \mathbf{H}\boldsymbol{\mu}). \tag{5.57}$$

(5.57) is solved first, variationally, by minimizing:

$$J_{\mathrm{PSAS}}(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}} + \mathbf{R})\mathbf{y} - \mathbf{y}^{\mathrm{T}}(\mathbf{z} - \mathbf{H}\boldsymbol{\mu}).$$

Then, as in 3DVAR the global problem is solved simultaneously. That is, no data selection is required. For this minimization problem, the control variable would be an m-vector, which was only $\mathrm{O}(10^5)$ for the global NWP problem. This two-stage method involving (5.56) and (5.57) is

referred to as a global OI (or PSAS- Physical Space Analysis System in NASA's Data Assimilation Office), or 3DVAR in observation space (as in NRL-Naval Research Laboratory). Like OI, PSAS is solved in observation space of dimension m, but like 3DVAR, PSAS needs no data selection. When the observation vector is much smaller in size than the state vector ($m \ll n$), then it makes sense to consider a PSAS formulation. When the state vector is smaller than the observation vector ($n \ll m$), then 3DVAR may be preferable. At NASA's Data Assimilation Office (DAO), PSAS was implemented for their climate model, GEOS (see Cohn et al. 1998). PSAS was also implemented at NRL (Naval Research Laboratory) (see Daley and Barker 2001).

It should now be apparent what the J function is, when it is presented on slides during data assimilation seminars. For variational assimilation, J is the argument of the *a posteriori* p.d.f. Thus variational assimilation (whether 3DVAR or 4DVAR) provides a conditional mode estimate. We minimize J because this is equivalent to maximizing the *a posteriori* p.d.f. of the state given the observations, when this p.d.f. is Gaussian. Thus 3DVAR assumes Gaussian background and observation errors.

5.4 Global background covariances: The "NMC" method

In OI, to compute background error covariance matrices, statistics of innovations were computed for the radiosonde network over North America. In 3DVAR, with no data selection, truly global covariances are needed. Covariances derived over North America would define waves no longer than wavenumber 10 (because of the maximum continental extent of 4000 km). In addition, the horizontal resolution of the radiosonde network is limited by observation spacing to about a few hundred km. Thus scales finer than this cannot be resolved by the covariances.

With 3DVAR and a global forecast model, all observations can be treated simultaneously (i.e. no data selection) so the need for global background error covariances is apparent. For a global spectral model, the model state is represented in spectral (rather than grid point) space. Thus, a scalar variable $\mathbf{x}$ is represented as a sum of contributions to a set of orthogonal functions:

$$\mathbf{x} = \sum_{n=0}^{N} \sum_{m=-n}^{n} x_n^m Y_n^m.$$

Here Y_n^m are spherical harmonics (wavelike exponential functions in the zonal direction, associated Legendre polynomials in the meridional direction) and x_n^m are the coefficients. The model state is then the vector of x_n^m for all m, n. The mapping from model space to physical space is simply done by applying an inverse spectral transform. Thus, the background error covariance matrix in physical space, $\mathbf{P}^b$, can be written as

$$\mathbf{P}^b = \mathbf{S}^{-1}\tilde{\mathbf{B}}(\mathbf{S}^{-1})^{\mathrm{T}}.$$

Note that $\tilde{\mathbf{B}}$ is the covariance matrix between forecast errors in spectral, not physical, space.

In 3DVAR, the background error covariances are needed in the inverse form, for the computation of the background term of the cost function:

$$\mathrm{J}^b = \frac{1}{2}(\mathbf{Sx})^{\mathrm{T}}\tilde{\mathbf{B}}^{-1}(\mathbf{Sx})$$

where $\mathbf{x} = \mathbf{x} - \mathbf{x}^b$. Note that with a model state defined in spectral space, variances are not a function of hor. space. The sum of the diagonal elements of the covariance matrix in spectral space

is

$$\sigma^2 = \sum_{n=0}^{N} (2n+1) b_n$$

where

$$\sum_{m=-n}^{n} < x_n^m, x_n^m >= (2n+1) b_n.$$

This is not very good news because we know that forecast error variances depend on physical location. However, as we shall show later, we can define the model state slightly differently (first dividing by variances in physical space before taking a spectral transform) to allow for full spatial variation of forecast error variances.

One advantage of the spectral representation is that vertical and horizontal covariances are nonseparable:

$$B^{kl} = \sum_{n=0}^{N} \sum_{m=-n}^{n} < (x_{kn}^m)(x_{ln}^m)^* >= \sum_{n=0}^{N} B_n^{kl}.$$

Since there is an intrinsic dependence of vertical covariances on horizontal (total) wavenumber, covariances are no longer assumed separable in physical space. This means that the horizontal correlation length scale can change with height. This is good news if you want to allow for the increase in correlation length scales with altitude.

In order to obtain global background error covariances, a new method of estimating background error covariance model parameters was devised at the U.S. National Meteorological Center (NMC), the so-called NMC-method. (NMC was since renamed NCEP, National Center for Environmental Prediction but the method's name remains unchanged.) The NMC-method is empirically based and motivated by the need for global, multivariate correlations with full vertical and spectral resolution. Instead of computing statistics of innovations, statistics of 24-h and 48-h forecast differences are compiled. By choosing a 24-h forecast instead of an analysis, the "spin-up" problem is avoided. (The "spin-up" problem is due to the imbalance of the analysis with repect to the model dynamics and results in an under- or over-prediction of precipitation, divergence, etc.) A 24-h difference in the forecast lengths was chosen because it is short enough to claim some similarity with 0-6 hr forecast error (which is really what we want) but long enough that forecasts are dissimilar despite the lack of data to update the starting analysis.

Although the connection between 48-h and 24-h forecast differences and the background error (6-h forecast error) is tenuous at best, the method compares relatively well to previous results based on innovations. Thus the method is justified by the fact that "it seems to work". The method is or was used for background error covariance estimates at many centres: NCEP in USA (Parrish and Derber 1992); ECMWF in Europe (Rabier et al. 1998); CMC in Canada (Gauthier et al. 1999a,b); the British Met Office (Ingleby 2001); BMRC in Australia (Steinle et al. 1995); and Météo-France in France (Desroziers et al. 1995). The method is also being tried for limited area models (Berre 2000).

Properties of the NMC-Method have been investigated by Bouttier (1994). He compares what the NMC-Method calculates to what a Kalman filter would obtain. In the analysis, he assumed for simplicity that H is linear, variations evolve linearly (tangent linear hypothesis is valid), no model error, a 6-hr forecast difference (instead of 24-h). The analysis error at hour 0 for NMC method and KF are approximately given by

$$\mathbf{P}_0^{\mathrm{NMC}} = \mathbf{K}\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}} + \mathbf{K}\mathbf{R}\mathbf{K}^{\mathrm{T}} \tag{5.58}$$

$$\mathbf{P}_0^{\mathrm{KF}} \quad = \quad (\mathbf{I}-\mathbf{KH})\mathbf{P}^b(\mathbf{I}-\mathbf{KH})^{\mathrm{T}} + \mathbf{KRK}^{\mathrm{T}} \tag{5.59}$$

The method breaks down if there is no data between the launch of the 2 forecasts. In reality, with no data ($\mathbf{K}=\mathbf{0}$), the forecast error is due to the background error. The KF predicts this in (5.59) but (5.58) predicts a zero covariance. Thus the NMC-method tends to under estimate background error covariances in data sparse regions (S. Hemisphere, oceans, etc). For dense, good quality, horizontally uncorrelated observations, $\mathbf{KH} = \mathbf{I}$ and NMC method over estimates covariances. The first term of (5.58) becomes $\mathbf{P}^b$ in this case while the corresponding KF term is $\mathbf{0}$. If observations are available at every grid point, and the background and observation error variances are equal, then $\mathbf{KH} = 0.5\ \mathbf{I}$, and the NMC method is exactly equivalent to the KF. In this special case, both first terms of (5.59) and (5.58) are identical. This may seem a very special case, but it can occur. For example, over North America with its dense observation network this assumption is approximately satisfied. Because of the known drawbacks of the NMC-method, the derived statistics are only used to define the background error *correlation* matrix. The variances are obtained using innovations (statistics of background minus observation differences) as they were with OI.

Despite the difference in the method used to obtain background error correlations for OI and 3DVAR, the results were surprisingly similar. Of course the 3DVAR background error variances were scaled to match the OI variances which were based on innovation statistics. Both correlation models has advantages and disadvantages. The NMC-based correlation model was nonseparable (in the horizontal and vertical), while the OI model assumed separability. Both models assumed isotropic correlations. The OI model allowed for locally homogeneous correlations (within latitude bands of 60 degrees) while the NMC-based model produced globally homogeneous correlations. Since the main inhomogeneity in correlations is in the meridional direction, it would be very useful to have this flexibility. With another change of variables, it is possible to introduce anisotropy and inhomogeneity to background error correlation models (e.g. Stajner et al. 2002).

5.5 Practical implementations of 3DVAR

5.5.1 Incremental Formulation of 3DVAR

3DVAR solves a global minimization problem. The size of the minimization problem is related to the size of the state vector, $\mathbf{x}$. Since this can be very large, $\mathrm{O}(10^7)$, the problem is computationally expensive. Since we know that assimilation includes filtering, the analysis increments will likely only be defined in terms of large scales. In this case, why bother solving for the small scales? Why not solve directly for the analysis increments on a reduced resolution grid, which will reduce the cost of the 3DVAR minimization? This is the reasoning behind the incremental approach to 3DVAR (or 4DVAR). The goal is to solve for analysis increments on a reduced resolution grid. Therefore, a projection operator from high to low resolution is required: $\mathbf{x}_L = \pi\mathbf{x}$. Here the subscript L refers to the low resolution model state. This operator could correspond to a grid interpolation or to spectral truncation. Now define a reference state around which the observation operator will be linearized:

$$\mathbf{x}_n = \mathbf{x}_{n-1} + \pi\delta\mathbf{x}_n. \tag{5.60}$$

At the first iteration ($n = 0$), the reference state will be the background state, i.e. $\mathbf{x}_0 = \mathbf{x}^b$. The analysis increments are then defined as:

$$\Delta\mathbf{x}_n = \mathbf{x}_n - \mathbf{x}^b$$

where $\Delta\mathbf{x}_0 = 0$. We want to rewrite the general 3DVAR cost function given by:

$$J(\mathbf{x}) = \frac{1}{2}[\mathbf{z} - H(\mathbf{x})]^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{z} - H(\mathbf{x})] + \frac{1}{2}[\mathbf{x} - \mathbf{x}^b]^{\mathrm{T}}(\mathbf{P}^b)^{-1}[\mathbf{x} - \mathbf{x}^b]$$

in terms of the increments. Note that we can expand

$$\mathbf{x}_n = \Delta\mathbf{x}_{n-1} + \mathbf{x}^b + \pi\delta\mathbf{x}_n.$$

If H is linearized using

$$H(\mathbf{x}^b + \delta\mathbf{x}) \approx H(\mathbf{x}^b) + \mathbf{H}\delta\mathbf{x}$$

then

$$\begin{aligned}
\mathbf{z} - H(\mathbf{x}_n) &= \mathbf{z} - H(\Delta\mathbf{x}_{n-1} + \mathbf{x}^b + \pi\delta\mathbf{x}_n) \\
&\approx \mathbf{z} - H(\mathbf{x}^b) - \mathbf{H}(\Delta\mathbf{x}_{n-1} + \pi\delta\mathbf{x}_n)) \\
&= \mathbf{z}' - \mathbf{H}(\Delta\mathbf{x}_{n-1} + \pi\delta\mathbf{x}_n)).
\end{aligned}$$

where $\mathbf{z}' = \mathbf{z} - H(\mathbf{x}^b)$. Also,

$$\begin{aligned}
\mathbf{x}_n - \mathbf{x}^b &= \Delta\mathbf{x}_{n-1} + \mathbf{x}^b + \pi\delta\mathbf{x}_n - \mathbf{x}^b \\
&= \Delta\mathbf{x}_{n-1} + \pi\delta\mathbf{x}_n
\end{aligned}$$

Thus the 3DVAR cost function can be rewritten as:

$$\begin{aligned}
J(\delta\mathbf{x}_n) &\approx \frac{1}{2}[\mathbf{z}' - \mathbf{H}(\Delta\mathbf{x}_{n-1} + \pi\delta\mathbf{x}_n)]^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{z}' - \mathbf{H}(\Delta\mathbf{x}_{n-1} + \pi\delta\mathbf{x}_n)] \\
&+ \frac{1}{2}[\Delta\mathbf{x}_{n-1} + \pi\delta\mathbf{x}_n](\mathbf{P}^b)^{-1}[\Delta\mathbf{x}_{n-1} + \pi\delta\mathbf{x}_n]. \qquad (5.61)
\end{aligned}$$

While the original 3DVAR cost function involved a nonlinear observation operator, this cost function only involves a linearized observation operator. Thus the cost function is purely quadratic and is guaranteed to have a unique minimium. On the other hand, some assumptions were made to derive this cost function, namely, that (1) $H(\mathbf{x})$ is weakly nonlinear and (2) if $\pi \neq \mathbf{I}$, the analysis increment is well represented on the low resolution grid.

Thus the incremental 3DVAR algorithm breaks the original 3DVAR minimization problem into a series of minimization problems. There is an outer loop in which the reference state is updated based on the current estimate of the analysis increment (5.60), and an inner loop in which the quadratic cost function (5.61) is solved for the increment.

The incremental 3DVAR algorithm can be summarized as follows:

1. Calculate innovations $\mathbf{z}'$ at full resolution.
2. Solve (5.61) to get analysis increments $\delta\mathbf{x}_n$ at low resolution. Define $\mathbf{x}_0 = \mathbf{x}^b$ so $\Delta\mathbf{x}_0 = 0$. Start from $n = 1$.
3. Interpolate $\delta\mathbf{x}_n$ from the low resolution grid to the high resolution grid.
4. Add $\mathbf{x}_n = \mathbf{x}_{n-1} + \pi\delta\mathbf{x}_n$ to get the analysis on the high resolution grid.
5. Go to 2.

5.5.2 CMC's 3DVAR

In this section we describe the change of variable from model state to a variable which avoids the need for inverting the background error covariance matrix. Further details on CMC's implementation of 3DVAR are found in Gauthier et al. (1999a,b).

At CMC, the 3DVAR implementation is based on the incremental formulation described above. The model state is $\mathbf{x} = [\boldsymbol{\psi}, \boldsymbol{\chi}, \mathbf{T}, \ln \mathbf{q}, \mathbf{p}_s)]^{\mathrm{T}}$. We want to define a change of variables that results in a simplification of the background term of the 3DVAR cost function. Therefore, let us define $\delta\mathbf{x} = \mathbf{L}\delta\mathbf{X}$, where $\mathbf{P}^b = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. The cost function is then

$$J_b = \delta\mathbf{x}^{\mathrm{T}}(\mathbf{P}^b)^{-1}\mathbf{x} = \delta\mathbf{X}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}(\mathbf{P}^b)^{-1}\mathbf{L}\delta\mathbf{X}.$$

With our choice of $\mathbf{L}$, we have that $\mathbf{L}^{\mathrm{T}}(\mathbf{P}^b)^{-1}\mathbf{L} = \mathbf{I}$ and $J_b = \delta\mathbf{X}^{\mathrm{T}}\delta\mathbf{X}$. How do you define $\mathbf{L}$? First of all, $\mathbf{L}$, may be viewed as a matrix, but it is never stored as such. Instead, $\mathbf{L}$ is simply a sequence of operations (or subroutines).

1) First transform to unbalanced variables,

$$\delta\mathbf{x} = \mathbf{K}\delta\mathbf{x}_u.$$

More explicitly,

$$\delta\mathbf{x} = \begin{pmatrix} \delta\boldsymbol{\psi} \\ \delta\boldsymbol{\chi} \\ \delta(\mathbf{T}, \mathbf{p}_s) \\ \delta \ln \mathbf{q} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 & 0 & 0 \\ \mathbf{E} & \mathbf{I} & 0 & 0 \\ \mathbf{N} & 0 & \mathbf{I} & 0 \\ 0 & 0 & 0 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \delta\boldsymbol{\psi}_u \\ \delta\boldsymbol{\chi}_u \\ \delta(\mathbf{T}, \mathbf{p}_s)_u \\ \delta \ln \mathbf{q}_u \end{pmatrix} = \mathbf{K}\delta\mathbf{x}_u. \tag{5.62}$$

Thus

$$\delta\boldsymbol{\chi} = \mathbf{E}\delta\boldsymbol{\psi} + \delta\boldsymbol{\chi}_u \tag{5.63}$$

$$\delta(\mathbf{T}, \mathbf{p}_s) = \mathbf{N}\delta\boldsymbol{\psi} + \delta(\mathbf{T}, \mathbf{p}_s)_u. \tag{5.64}$$

What we are doing is changing from physical variables to "unbalanced" variables because these are uncorrelated with each other. In fact, they are uncorrelated with each other, *by construction*. The matrices $\mathbf{E}$ and $\mathbf{N}$ are defined by statistical regression by removing the correlation between $\delta\boldsymbol{\psi}$ and $\delta\boldsymbol{\chi}$ and between $\delta\boldsymbol{\psi}$ and $\delta\mathbf{T}, \mathbf{p}_s$. To define $\mathbf{E}$, multiply (5.63) by $\delta\boldsymbol{\psi}^{\mathrm{T}}$ and take the ensemble average:

$$< \delta\boldsymbol{\chi}\delta\boldsymbol{\psi}^{\mathrm{T}} >= \mathbf{E} < \delta\boldsymbol{\psi}\delta\boldsymbol{\psi}^{\mathrm{T}} > + < \delta\boldsymbol{\chi}_u\delta\boldsymbol{\psi}^{\mathrm{T}} >= \mathbf{E} < \delta\boldsymbol{\psi}\delta\boldsymbol{\psi}^{\mathrm{T}} >$$

where we have defined $< \delta\boldsymbol{\chi}_u\delta\boldsymbol{\psi}^{\mathrm{T}} >= 0$. Thus

$$\mathbf{E} =< \delta\boldsymbol{\chi}\delta\boldsymbol{\psi}^{\mathrm{T}} >< \delta\boldsymbol{\psi}\delta\boldsymbol{\psi}^{\mathrm{T}} >^{-1} .$$

$\mathbf{N}$ can be defined similarly. Thus "unbalanced" here does not refer to ageostrophic or any other dynamical term, but to simply to the uncorrelated part of $\delta\boldsymbol{\chi}$ or $\delta\mathbf{T}$ or $\delta\mathbf{p})_s$. In practice, however, when you look at the correlated part of $\delta\psi$ and δT, for example, it mainly reflects a geostrophic and hydrostatic balance. Similarly, the correlated part of $\delta\psi$ and $\delta\chi$ is largest at the surface reflecting the action of Ekman pumping in turning the geostrophic wind toward low pressures. (In implementation, further assumptions are made. The ensemble average is replaced by a time and longitudinal average. $\mathbf{E}$ is solved for diagonal elements only, and $\mathbf{N}$ is solved for diagonal elements, but within latitude bands.)

Note that the inverse of $\mathbf{K}$ is easily defined as:

$$\mathbf{K}^{-1} = \begin{pmatrix} \mathbf{I} & 0 & 0 & 0 \\ -\mathbf{E} & \mathbf{I} & 0 & 0 \\ -\mathbf{N} & 0 & \mathbf{I} & 0 \\ 0 & 0 & 0 & \mathbf{I} \end{pmatrix}.$$

Thus the background error covariance matrix in physical space is related to the covariance matrix of unbalanced variables as follows.

$$\mathbf{P}^b = < \delta\mathbf{x}\delta\mathbf{x}^{\mathrm{T}} > = \mathbf{K} < \delta\mathbf{x}_u\delta\mathbf{x}_u^{\mathrm{T}} > \mathbf{K}^{\mathrm{T}} = \mathbf{K}\mathbf{P}_u^b\mathbf{K}^{\mathrm{T}}.$$

2) Normalize by variances in physical space, then spectrally transform:

$$\delta\tilde{\mathbf{x}}_u = \mathbf{S}\boldsymbol{\Sigma}_u^{-1}\mathbf{K}^{-1}\delta\mathbf{x} = \mathbf{S}\boldsymbol{\Sigma}_u^{-1}\delta\mathbf{x}_u$$

Note that $\boldsymbol{\Sigma}_u$ is a diagonal matrix of background error variances and $\mathbf{S}^{\mathrm{T}} = \mathbf{S}^{-1}$ for a spectral transform matrix.

$$\delta\mathbf{x}_u = \boldsymbol{\Sigma}_u\mathbf{S}^{-1}\delta\tilde{\mathbf{x}}_u$$

so that

$$\mathbf{P}_u^b = < \delta\mathbf{x}_u(\delta\mathbf{x}_u)^{\mathrm{T}} > = \boldsymbol{\Sigma}_u\mathbf{S}^{-1}\tilde{\mathbf{P}}_u^b\mathbf{S}\boldsymbol{\Sigma}_u$$

3) Eigenvalue decomposition:

$$\tilde{\mathbf{P}}_u^b = \mathbf{E}_u\Lambda_u\mathbf{E}_u^{\mathrm{T}}.$$

4) In total:

$$\delta\mathbf{X} = \Lambda_u^{-1/2}\mathbf{E}_u^{\mathrm{T}}\mathbf{S}\boldsymbol{\Sigma}_u^{-1}\mathbf{K}^{-1}\delta\mathbf{x} = \mathbf{L}^{-1}\delta\mathbf{x}$$
$$\delta\mathbf{x} = \mathbf{L}\delta\mathbf{X} = \mathbf{K}\boldsymbol{\Sigma}_u\mathbf{S}^{-1}\mathbf{E}_u\Lambda_u^{1/2}\delta\mathbf{X}$$

so that

$$\begin{aligned} \mathbf{P}^b &= \mathbf{K}\mathbf{P}_u^b\mathbf{K}^{\mathrm{T}} \\ &= \mathbf{K}\boldsymbol{\Sigma}_u\mathbf{S}^{-1}\tilde{\mathbf{P}}_u^b\mathbf{S}\boldsymbol{\Sigma}_u\mathbf{K}^{\mathrm{T}} \\ &= \mathbf{K}\boldsymbol{\Sigma}_u\mathbf{S}^{-1}\mathbf{E}_u\Lambda_u\mathbf{E}_u^{\mathrm{T}}\mathbf{S}\boldsymbol{\Sigma}_u\mathbf{K}^{\mathrm{T}} \\ &= \mathbf{L}\mathbf{L}^{\mathrm{T}} \end{aligned}$$

This means we can write

$$J_b(\delta\mathbf{X}) = \delta\mathbf{x}^{\mathrm{T}}(\mathbf{P}^b)^{-1}\mathbf{x} = \delta\mathbf{X}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}(\mathbf{P}^b)^{-1}\mathbf{L}\delta\mathbf{X} = \delta\mathbf{X}^{\mathrm{T}}\delta\mathbf{X}.$$

Now we can minimize with respect to $\delta\mathbf{X}$, then convert the solution to $\delta\mathbf{x}$ using $\delta\mathbf{X} = \mathbf{L}\delta\mathbf{x}$. Thus we've solved the 3DVAR problem with needing to invert the background error covariance matrix.

In summary, 3DVAR and OI analyses are equivalent in theory, but different in practice. 3DVAR allows easy extension for nonlinearly related observed and model variables. This also permits more flexibility in the choice of analysis variables. 3DVAR does not require data selection so analyses are in better balance. The improvement of 3DVAR over OI is not statistically significant when the observation sets are the same. However, systematic improvements of analyses are seen in the stratosphere and southern hemisphere for 3DVAR. With the addition of much more satellite data, improvements in analyses are evident even in the northern hemisphere.

Appendix

Show that

$$\begin{aligned} J &= (\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x}) + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu}). \end{aligned} \tag{5.65}$$

can be written as:

$$J = (\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{P}_x^{-1}(\mathbf{x}-\hat{\mathbf{x}}) \tag{5.66}$$

where

$$\hat{\mathbf{x}} = \mathbf{P}_x(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} + \mathbf{P}^{-1}\boldsymbol{\mu}) \tag{5.67}$$

and

$$\mathbf{P}_x^{-1} = \mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}. \tag{5.68}$$

First recall the Sherman-Morrison-Woodbury formula from the appendix of chapter 3:

$$(\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} = \mathbf{P}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1}. \tag{5.69}$$

Multiply both sides on the left by $\mathbf{R}^{-1}\mathbf{H}$:

$$\begin{aligned} \mathbf{R}^{-1}\mathbf{H}(\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} &= \mathbf{R}^{-1}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}})(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1} \\ &= \mathbf{R}^{-1}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}}+\mathbf{R}-\mathbf{R})(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1} \\ &= \mathbf{R}^{-1}(\mathbf{I}-\mathbf{R}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1}) \\ &= \mathbf{R}^{-1} - (\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1}. \end{aligned} \tag{5.70}$$

Second, note that (5.67) can be rearranged to:

$$\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} = \mathbf{P}_x{}^{-1}\hat{\mathbf{x}} - \mathbf{P}^{-1}\boldsymbol{\mu}. \tag{5.71}$$

Subtract $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\boldsymbol{\mu}$ from both sides to get:

$$\begin{aligned} \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu}) &= \mathbf{P}_x{}^{-1}\hat{\mathbf{x}} - \mathbf{P}^{-1}\boldsymbol{\mu} - \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\boldsymbol{\mu} \\ &= \mathbf{P}_x{}^{-1}\hat{\mathbf{x}} - (\mathbf{P}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})\boldsymbol{\mu} \\ &= \mathbf{P}_x{}^{-1}\hat{\mathbf{x}} - \mathbf{P}_x{}^{-1}\boldsymbol{\mu} \\ &= \mathbf{P}_x{}^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu}). \end{aligned} \tag{5.72}$$

Finally, note that

$$\begin{aligned} (\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x}) &= (\mathbf{z}-\mathbf{H}(\mathbf{x}-\boldsymbol{\mu})-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}(\mathbf{x}-\boldsymbol{\mu})-\mathbf{H}\boldsymbol{\mu}) \\ &= (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}} + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}(\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}} - (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}(\mathbf{x}-\boldsymbol{\mu}). \end{aligned} \tag{5.73}$$

Now we can expand (5.65) using (5.73):

$$\begin{aligned} J &= (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}} + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}(\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}} - (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}(\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu}) + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &= (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}[\mathbf{R}^{-1}-(\mathbf{H}\mathbf{P}\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1}](\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}} + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}[\mathbf{P}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}](\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}} - (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}(\mathbf{x}-\boldsymbol{\mu}) \\ &= (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}[\mathbf{R}^{-1}\mathbf{H}(\mathbf{P}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}](\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}} + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu}) - (\hat{\mathbf{x}}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\mathbf{x}-\boldsymbol{\mu}). \end{aligned} \tag{5.74}$$

To get the last line, we substituted the expressions (5.68) and (5.70). Finally, we can again recognize the expression, (5.68) in the last line to further simplify this to

$$\begin{aligned} J &= (\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{P}_x\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^{\mathrm{T}} + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu}) - (\hat{\mathbf{x}}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\mathbf{x}-\boldsymbol{\mu}). \end{aligned} \tag{5.75}$$

Substitute (5.72) into the first term on the right to get

$$\begin{aligned} J &= (\hat{\mathbf{x}}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\mathbf{P}_x)(\mathbf{P}_x)^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu}) + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu}) - (\hat{\mathbf{x}}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\mathbf{x}-\boldsymbol{\mu}). \end{aligned} \tag{5.76}$$

Simplification of the above yields:

$$\begin{aligned} J &= (\hat{\mathbf{x}}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu}) + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &- (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu}) - (\hat{\mathbf{x}}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{P}_x)^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &= [(\mathbf{x}-\boldsymbol{\mu}) - (\hat{\mathbf{x}}-\boldsymbol{\mu})]^{\mathrm{T}}(\mathbf{P}_x)^{-1}[(\mathbf{x}-\boldsymbol{\mu}) - (\hat{\mathbf{x}}-\boldsymbol{\mu})] \\ &= (\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}}{\mathbf{P}_x}^{-1}(\mathbf{x}-\hat{\mathbf{x}}) \end{aligned} \tag{5.77}$$

This is exactly the expression we were looking for, namely, (5.66).

REFERENCES

1. Andersson and co-authors, 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part III: Experimental results. *Q. J. Roy. Meteor. Soc.*, **124**, 1831-1860.

2. Berre, L., 2000: Estimation of synoptic and mesoscale forecast error covariances in a limited-area model. *Mon. Wea. Rev.*, **128**(3), 644-667.

3. Bouttier, F., 1994: Sur la prévision de la qualité des prévisions météorologiques. Thèse de doctorat de l'Université Paul Sabatier à Toulouse, 240 pp.

4. Cohn, S., A. da Silva, J. Guo, M. Sienkiewicz and D. Lamich, 1998: Assessing the effects of data selection in the DAO Physical- Space Statistical Analysis System. *Mon. Wea. Rev.*, **126**, 2913-2926.

5. Courtier, P. and co-authors, 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part I: Formulation. *Q. J. Roy. Meteor. Soc.*, **124**, 1783-1807.

6. Desroziers, G., V. Mathiot and F. Orain, 1995: A study of ARPEGE forecast error covariances. Proceedings of the WMO second international symposium on assimilation of observations in meteorology and oceanography, Tokyo, Japan, Vol. 1, 263-268.

7. Daley, R. and E. Barker, 2001: NAVDAS: formulation and diagnostics. *Mon. Wea. Rev.*, **129**(4), 869-883.

8. Gauthier, P., C. Charette, L. Fillion, P. Koclas and S. Laroche, 1999a: implementation of a 3D variational data assimilation system at the Canadian Meteorological Centre. Part I: The global analysis. *Atmosphere-Ocean*, **37**, 103-156.

9. Gauthier, P., M. Buehner and L. Fillion, 1999b: Background-error statistics modelling in a 3D variational data assimilation scheme: estimation and impact on the analyses. In *Diagnosis of data assimilation system*, Proceedings of a Workshop held at ECMWF on 2-4 November 1998, p. 131-145.

10. Ingleby, N. B., 2001: The statistical structure of forecast errors and its representation in The Met. Office Global 3-D Variational Data Assimilation Scheme. *Q. J. Roy. Meteor. Soc.*, **127** (571 Part A), 209-231.

11. Laroche, S., P. Gauthier, J. St-James and J. Morneau, 1999: Implementation of a 3D variational data assimilation system at the Canadian meteorological centre. Part II: The regional analysis. *Atmosphere-Ocean*, **37**, 281-307.

12. Lorenc, A., S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne, F. W. Saunders, 2000: The Met. Office global three-dimensional variational data assimilation scheme *Q. J. Roy. Meteor. Soc.*, **126**(570 Parb B), 2991-3012.

13. Parrish, D. F. and J. C. Derber, 1992: The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747-1763.

14. Rabier, F. and co-authors, 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part II: Structure functions. *Q. J. Roy. Meteor. Soc.*, **124**, 1809-1829.

15. Stajner, I., L. P. Riishojgaard, and R. B. Rood, 2001: The GEOS ozone data assimilation system: specification of error statistics. *Q. J. R. Meteorol. Soc.*, **127**:1069-1094.

16. Steinle, P., R. Seaman, W. Bourke and T. Hart, 1995: A generalized statistical interpolation scheme. Proceedings of the WMO second international symposium on assimilation of observations in meteorology and oceanography, Toyko, Japan, Vol. I, 205-208.

17. Steinle, P., R. Seaman, 1996: A new 3-D variational analysis scheme. (In: Manton, M. J.; Jasper, J. D. and Meighen, P. J. (eds.),, Bureau of Meteorology Research Centre. BMRC Research Reports, No. 54, Melbourne, Australia), 139-144.

18. Tarantola, A., 1987: *Inverse problem theory.* Elsevier, 613 pp.