QUALITY CONTROL

Andrew C Lorenc
Meteorological Office
Bracknell, England

Summary: Probabilities can be used to quantify the extent of our knowledge about past events. Bayes' Theorem tells us how to combine these probabilities. It enables us to calculate weights for the information sources, as a function of their error distributions. With the aid of very simple examples, I review the bayesian derivation of standard statistical methods used to combine observations, leading for instance to a quadratic variational penalty function. Introduction of a simple model for gross observational errors is shown to substantially modify the behaviour, with a non-Gaussian posterior probability distribution that is often multi-modal. The choice of the "best" analysis, maximizing the expected benefit, is discussed, and shown usually to correspond to rejecting those observation more likely than not to contain gross errors.

I discuss the theoretical basis for three different practical approaches (based on the Met Office bayesian sequential buddy checking scheme, the ECMWF "OI" scheme, and a variational scheme). Possible weaknesses are pointed out, again illustrated with a simple example. Finally I stress the importance of monitoring.

1 INTRODUCTION

1.1 Why Quality Control?

The quality control we do in data-assimilation has two reasons:

1. We have physical reasons for believing certain events may occur which affect the observed value. We wish to detect these events.
2. The distribution of errors associated with a datum is such that there is a non-negligible probability of errors that would be unacceptably large *for the use we are making of the datum.*

Note that 2 depends on our use of the observation. If we are using an analysis method based on a quadratic penalty function, it is linear in the observed values. A single large error can then be disastrous (figure 1a). However if instead we minirise the mean absolute deviation (We shall see that this is the correct norm if the observational error probability distribution function (pdf) is proportional to an exponential of the absolute deviation), the bad datum is ignored (figure 1b). Analysis methods designed to ignore such outliers are also considered to be quality control methods.

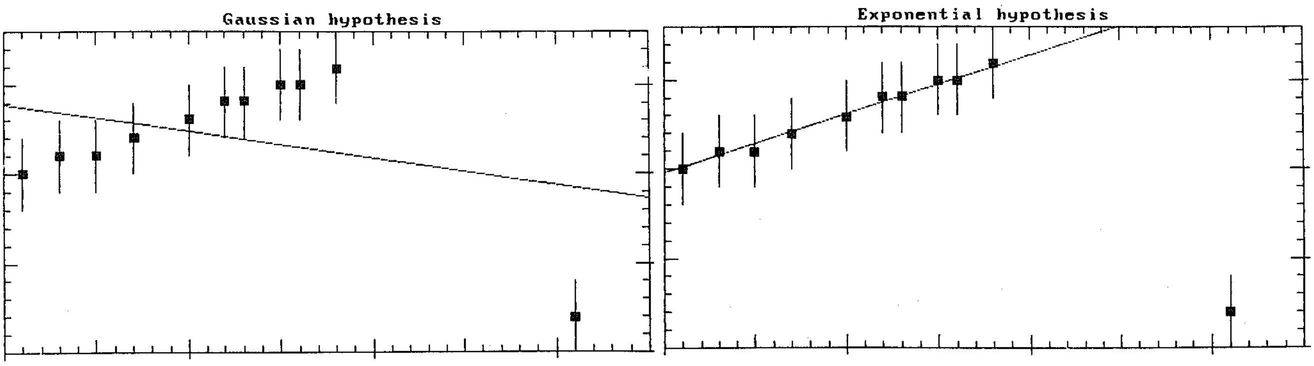

Fig.1. Best fit straight line to data including a gross error, (a) using a quadratic (L2) norm, (b) using a mean absolute (L1) norm (from Tarantola 1987).

1.2 What is the Best Analysis?

In section 2 we shall see how the Bayesian approach can, in theory, give us the posterior pdf, describing how likely it is that each atmospheric state is correct. However to start a forecast we need a single best analysis[1]. In our simple example of figure 5, is the best analysis in the tallest peak, or in the peak with the largest area? The theoretical approach to this is outlined in section 2.7. But in practice this is impracticable; more pragmatic judgements of what is best are implicit in the practical schemes described in section 3.

2 BAYES THEOREM

2.1 Bayesian Probabilities

Nothing is certain in life, especially in weather forecasting. Forecasters are used to using probabilities to express this. "There is a 25% chance of rain tomorrow" does not necessarily imply that the atmosphere is random, but rather that the forecaster is uncertain. The use of probabilities to quantify the extent of our knowledge, about future or past events, is the key to the Bayesian approach.

2.2 Discrete Events

Suppose we have discrete events A and B, then we can use $P(A)$ and $P(B)$ to describe the probability of them occurring in the future, or the extent of our knowledge about them, if they have occurred in the past. Similarly we use $P(A \cap B)$ to denote the probability that A and B both occur, and $P(A \mid B)$ to denote the conditional probability of A given that B has occurred.

[1] The choice of the best ensemble of analyses, to span the range of possible forecasts, is a major research area beyond the scope of this paper.

We then have two ways of expressing $P(A \cap B)$:

$$\begin{aligned} P(A \cap B) &= P(B)\ P(A|B) \\ &= P(A)\ P(B|A) \end{aligned} \tag{1}$$

These lead directly to Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)} \tag{2}$$

When we apply this, we often use relative probabilities (i.e. we do not bother with the factor $P(B)$), or else we calculate $P(B)$ from:

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \tag{3}$$

Example: Bayesian Dice

I have two dice. One is weighted towards throwing sixes. I have performed some experiments with them, and have the prior statistics that:

for the weighted (W) die, $P(6|W) = 58/60$
for the good (G) die, $P(6|G) = 10/60$

I choose one at random: $P(W) = P(G) = 1/2$

I throw this die, and it shows a six. Now:-

$$\begin{aligned} P(6) &= P(6|W)\ P(W) + P(6|G)\ P(G) \\ &= 58/60\ \ 1/2\ + 10/60\ \ 1/2 \\ &= 34/60 \end{aligned}$$

We can now apply Bayes' Theorem:

$$\begin{aligned} P(G|6) &= P(6|G)\ P(G)\ /\ P(6) \\ &= 10/60\ \ 1/2\ \ /\ 34/60\ = \ 5/34 \end{aligned}$$

$$\begin{aligned} P(W|6) &= P(6|W)\ P(W)\ /\ P(6) \\ &= 58/60\ \ 1/2\ \ /\ 34/60\ = 29/34 \end{aligned}$$

The information that I have thrown a six has added to my knowledge, so that the posterior probability that the chosen die is weighted has increased. If I were to throw again, and get another six, the probability would increase again.

2.3 Continuous Variables

We can go from discrete events to continuous variables by defining events such as:

X: the true value x_t is such that $x \leq x_t < x + \delta x$

Then $P(X) = p(x)\ \delta x$

Taking the limit $\delta x \rightarrow 0$, $p(x)$ is a probability density function (pdf).

Zero-dimensional Bayesian Analysis

A Gaussian pdf about a prior value x^b with background error variance V_b is:

$$p(x) = N(x|x^b,V_b) = (2\pi V_b)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(x-x^b)^2}{V_b}\right) \qquad \textbf{(4)}$$

where $N(x|m,V)$ denoted a normal distribution with mean m and variance V.

$P(y^o|x) = p(y^o|x)\ dy^o$, is the probability of getting an observation y^o, given that the true value is x.[2] For example a Gaussian pdf with observational error variance V_o is:

$$p(y^o|x) = N(y^o|x,V_o) = (2\pi V_o)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(y-x)^2}{V_o}\right) \qquad \textbf{(5)}$$

We can get $p(y^o)$ by integrating over all x:

$$p(y^o) = \int p(y^o|x)p(x)dx \qquad \textbf{(6)}$$

Since the convolution of two Gaussians is another Gaussian, for our examples this gives:

$$p(y^o) = N(y^o|x^b,V_o+V_b) \qquad \textbf{(7)}$$

Bayes' Theorem in continuous form is:

$$p(x|y^o) = \frac{p(y^o|x)p(x)}{p(y^o)} \qquad \textbf{(8)}$$

[2] Note that all pdfs are conditional on knowing the prior value x^b. To simplify, we do not show this explicitly in the notation.

$p(x|y^o)$ is called the posterior distribution, $p(x)$ the prior distribution, and $p(y^o|x)$ is the likelihood function for x.

Substituting the pdfs from our example gives:

$$p(x|y^o) = N(x|x^a,V_a) \tag{9}$$

where

$$\begin{aligned} \frac{x^a}{V_a} &= \frac{y^o}{V_o} + \frac{x^b}{V_b} \\ \frac{1}{V_a} &= \frac{1}{V_o} + \frac{1}{V_b} \end{aligned} \tag{10}$$

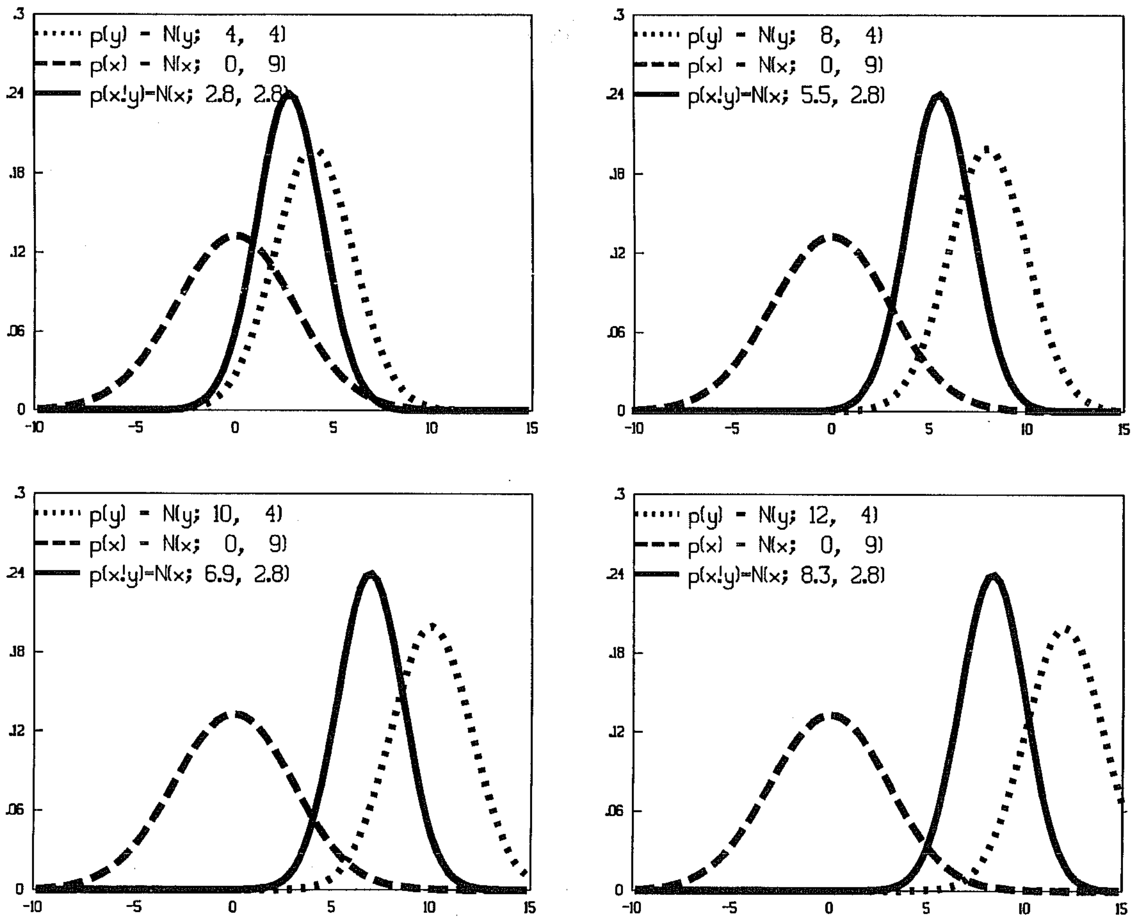

Fig.2. Prior pdf $p(x)$ *(dashed line)*, posterior pdf $p(x|y^o)$ *(solid line)*, and likelihood of observation $p(y^o|x)$ *(dotted line)*, plotted against x for various values of y^o. (Adapted from Lorenc and Hammon 1988).

This is the standard formula for the combination of observations, studied by Gauss (1823). It is illustrated in figure 2. Note that, due to the unique properties of Gaussian distributions, the posterior distribution has the same shape for all values of y^o and x^b.

2.4 Multi-dimensional Bayesian Analysis

I have only space here to sketch out how this approach can be extended to data assimilation, in notation consistent with Ide et al., (1996); a fuller derivation is in Lorenc (1986). The derivation is illustrated with the simplest possible example - a one-dimensional model consisting of two grid-point values, and a single observation at their midpoint. For this simple case the equivalent of figure 2 can be shown as a contour plot (figure 3). The values that we wish to analyse are combined in a vector $\boldsymbol{x}$: in our example the two grid point values, x_1 and x_2.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{11}$$

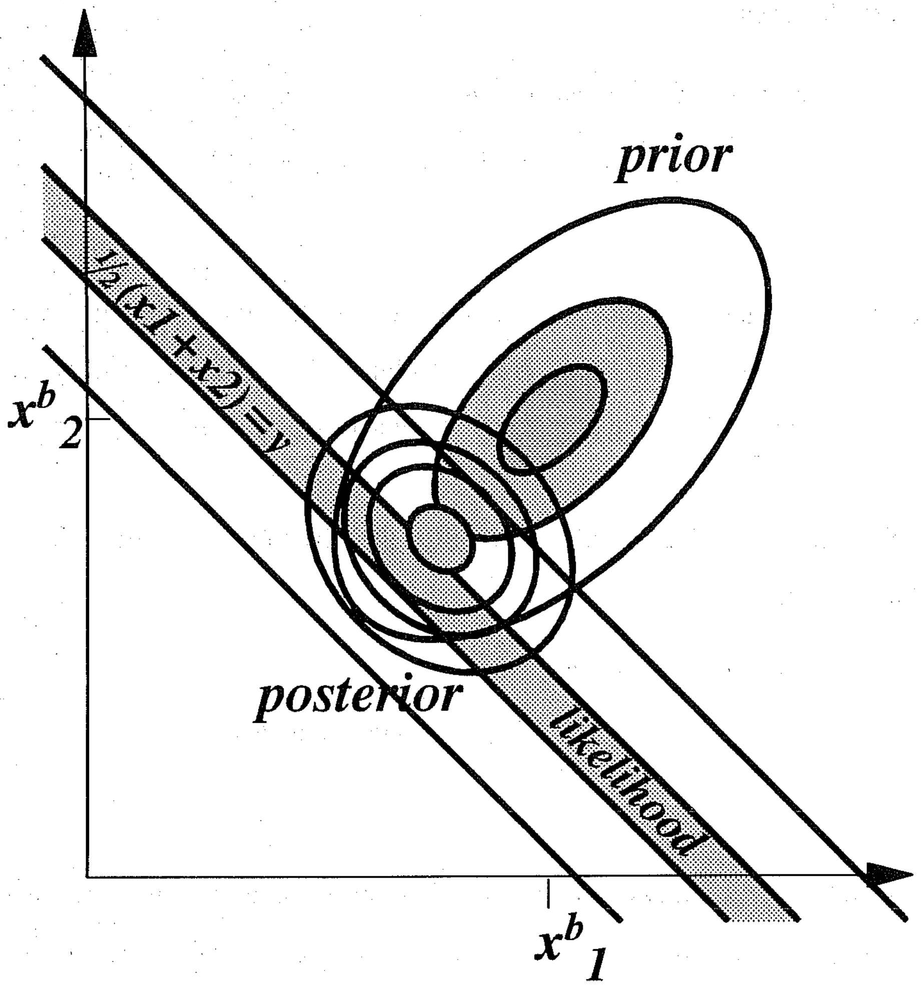

Fig.3. Contour plot of the prior pdf $p(x_1,x_2)$, for the simple example with a positive correlation between the background errors of $x^b{}_1$ and $x^b{}_2$.
Likelihood function: the probability ($p(y^o|x)$) of getting the observation, plotted as a function of x.

The observations are combined into a vector $\boldsymbol{y}^o$. We can make an estimate $\boldsymbol{y}$ of this from $\boldsymbol{x}$; in our example this is done by interpolation to the single observation:

$$\begin{aligned} y &= H(\boldsymbol{x}) = \tfrac{1}{2}x_1+\tfrac{1}{2}x_2 \\ &= \mathbf{H}\,\boldsymbol{x} = \begin{pmatrix}\frac{1}{2} & \frac{1}{2}\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} \end{aligned} \tag{12}$$

We have a prior estimate $\boldsymbol{x}^b$ for $\boldsymbol{x}$. Because the errors in the components of $\boldsymbol{x}^b$ are in general correlated, we cannot get the joint probability by multiplying $p(x_1)\times p(x_2)$. Instead we must use a multi-dimensional Gaussian:

$$\begin{aligned} p(x_1 \cap x_2) &= p(x) = N(\boldsymbol{x}|\boldsymbol{x}^b,\mathbf{B}) \\ &= \left((2\pi)^2|\mathbf{B}|\right)^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^b)^T\mathbf{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}^b)\right) \end{aligned} \tag{13}$$

where **B** is a covariance matrix. Since our example $\boldsymbol{x}$ only has two elements, we can plot $p(\boldsymbol{x})$, shown as the prior pdf in figure 3. The error correlation between the grid points leads to the elliptical shape - in a real model the ellipsoid has aspect ratios reaching 10^6 or more.

The instrumental error is described by the probability that the observed value lies between y^o and $y^o+\delta y^o$, given the true value[3] y^t :

$$\begin{aligned} p(y^o|y^t) &= N(y^o|y^t,\mathbf{E}) \\ &= (2\pi|\mathbf{E}|)^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(y^o-y^t)^T\mathbf{E}^{-1}(y^o-y^t)\right) \end{aligned} \tag{14}$$

Because we have only a finite representation of reality, in our example only two values, knowing $\boldsymbol{x}^t$ does not give us precise knowledge of y^t. This error of representativeness has the pdf:

$$\begin{aligned} p^t(y^o|\boldsymbol{x}^t) &= N(y^o|H(\boldsymbol{x}^t),\mathbf{F}) \\ &= (2\pi|\mathbf{F}|)^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(y^o-H(\boldsymbol{x}^t))^T\mathbf{F}^{-1}(y^o-H(\boldsymbol{x}^t))\right) \end{aligned} \tag{15}$$

The observational error is the sum of these two effects. Its pdf is obtained by integrating over all y which might be y^t:

$$\begin{aligned} p(y^o|\boldsymbol{x}^t) &= \int p(y^o|y)p^t(y|\boldsymbol{x}^t)dy \\ &= N(y^o|H(\boldsymbol{x}^t),\mathbf{E}+\mathbf{F}) \end{aligned} \tag{16}$$

It is common to replace **E+F** by a single observational error matrix **R**.[4] For our simple example,

[3] We define y^t to be what would be observed by a perfect instrument, with the same observing footprint as the real one.

[4] The above derivation shows that this "observational" error depends on the model resolution, and on the interpolation H. It is important to remember this when modelling observational errors for quality control algorithms; e.g. if y^o is a sounding of satellite radiance data, then poor cloud

we can plot $p(y^o|x)$ as a function of x_1 and x_2 - the likelihood [5] (figure 3).

Substituting into the Bayesian analysis equation, we get

$$\begin{aligned} p(x|y^o) &= \frac{p(y^o|x)p(x)}{p(y^o)} \\ &= \frac{N(y^o|H(x),\mathbf{R})\ N(x|x^b,\mathbf{B})}{\int [\ N(y^o|H(x),\mathbf{R})\ N(x|x^b,\mathbf{B})\ ]dx} \end{aligned} \qquad \textbf{(17)}$$

The product of two Gaussians can (as long as H is linearisable) be reorganised to collect the $\boldsymbol{x}$ terms into a single Gaussian:

$$N(y^o|H(x),\mathbf{R})\ N(x|x^b,\mathbf{B}) = N(y^o|H(x^b),\mathbf{R}+\mathbf{HBH}^T)\ N(x|x^a,\mathbf{A}) \qquad \textbf{(18)}$$

where $\boldsymbol{x}^a$ and $\mathbf{A}$ are defined by:

$$\begin{aligned} \mathbf{A} &= \mathbf{B}-\mathbf{BH}^T(\mathbf{HBH}^T+\mathbf{R})^{-1}\mathbf{HB} \\ x^a &= x^b+\mathbf{BH}^T(\mathbf{HBH}^T+\mathbf{R})^{-1}(y^o-H(x^b)) \end{aligned} \qquad \textbf{(19)}$$

Substituting (18) in the denominator of (17), the Gaussian in $\boldsymbol{x}$ integrates to one. Cancelling the other Gaussian top and bottom gives:

$$p(x|y^o) = N(x|x^a,\mathbf{A}) \qquad \textbf{(20)}$$

This is shown as the posterior pdf in figure 3.

2.5 Log(probabilities) - Penalty Functions

The Bayesian analysis equation, with its product of probabilities, is hard to handle. For variational algorithms, it is more convenient to take minus the logarithm. The equation then becomes:

assumptions can lead to gross errors in the radiative transfer model H; these are manifested as gross "observational" errors.

[5] It does not integrate to one over x, so it is not a probability distribution function.

$$p(x|y^o) = p(x|y^o \cap G)\ P(G|y^o) + p(x|y^o \cap \bar{G})\ P(\bar{G}|y^o)$$
$$= N(x|x^b, \mathbf{B})\ P(G|y^o) + N(x|x^a, \mathbf{A})\ P(\bar{G}|y^o) \qquad (27)$$

The posterior pdf is the weighted sum of two Gaussian, corresponding to accepting or rejecting the observation. The weights given to each are the posterior probabilities of G and $\bar{G}$. When the peaks are distinct, these correspond to the areas under each. The results of allowing for gross errors in this way can be quite dramatic, even if $P(G)$ is small. Figure 5 shows the equivalent of figure 2, with errors appropriate for pressure observations from ships, which have about 5% gross errors. When the observation and the background agree, there is little difference from figure 2. But when they disagree, the posterior distribution becomes bi-modal.

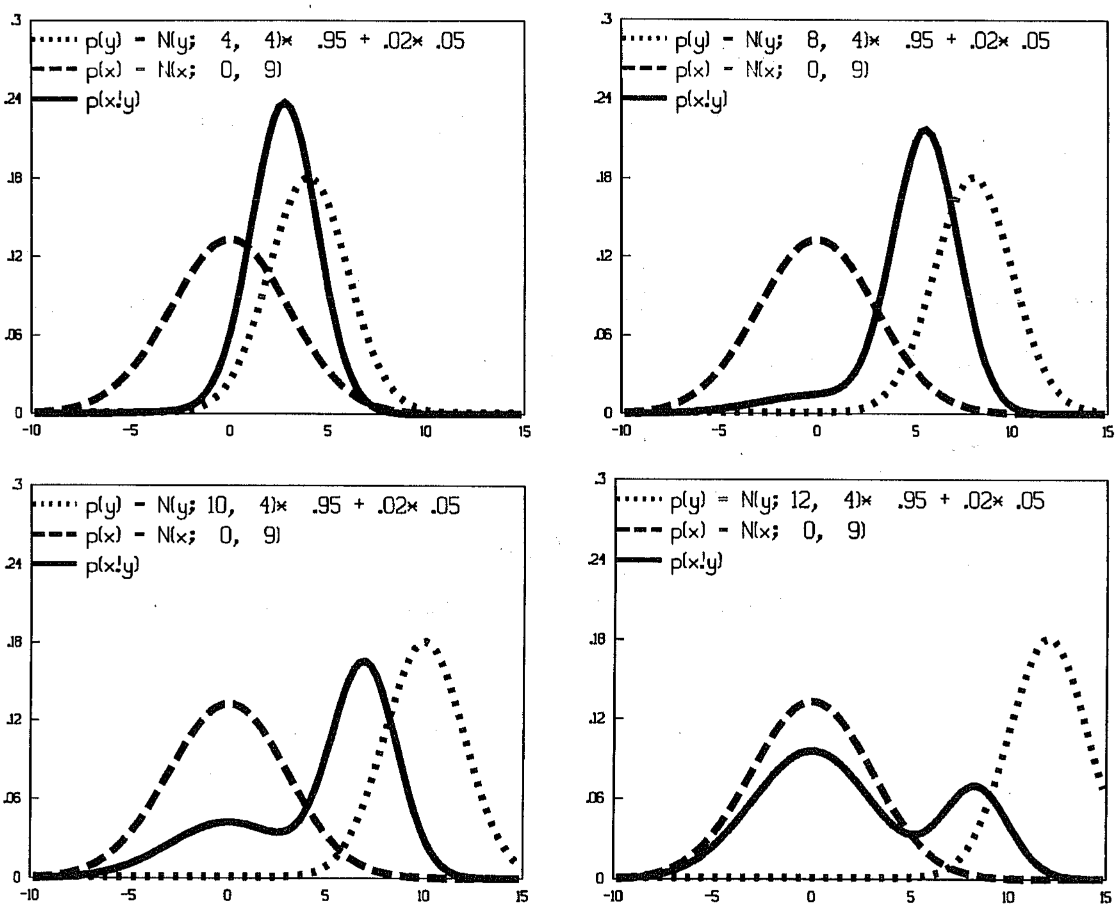

Fig.5. As figure 2 for an observation with a 5% chance of gross error.

Figure 6 shows the same examples in the log(probability) form of figure 4. The observational penalty is not quadratic; it has plateaus away from the observed value. Adding this to a quadratic background penalty can give multiple minima.

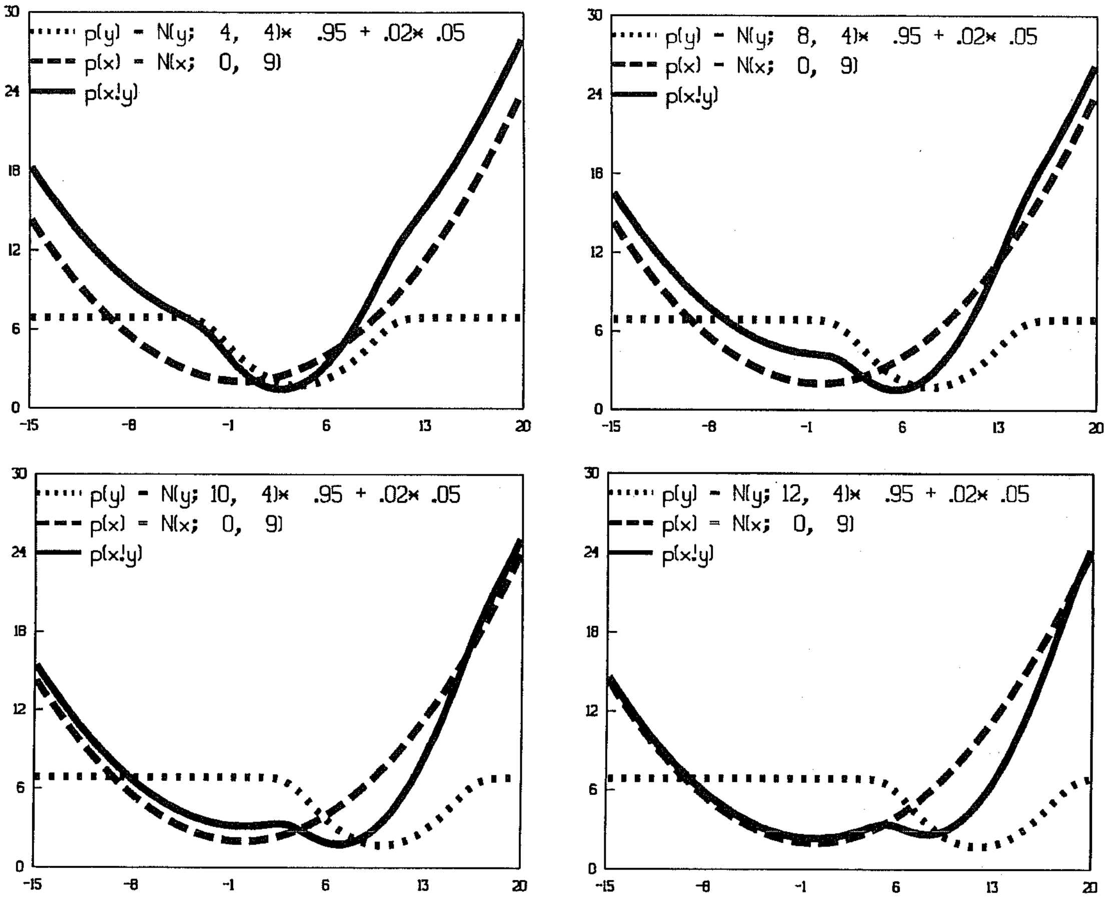

Fig.6. As figure 5 for -log(probabilities).

2.7 **What is the best analysis**

To approach this objectively, we have to define how much it costs us to be wrong, or alternatively how much we benefit from being nearly right. If we have a quadratic cost function, then the best analysis is the mean of the posterior pdf. If we have a spike (delta function) benefit function, the best analysis is at the maximum of the posterior pdf. Probably the best simple model is between the two: a Gaussian shaped benefit function such that analyses close to the truth are useful, while those a long way from correct are equally worthless. A Gaussian shaped benefit function, with width specified by pseudo-covariance **C**, has the advantage of facilitating algebraic calculation of the benefit; convolving it with the posterior pdf from (27) we get:

$$\mathit{benefit}(x) = N(x|x^b,\mathbf{B}+\mathbf{C})\ P(G|y^o) + N(x|x^a,\mathbf{A}+\mathbf{C})\ P(\bar{G}|y^o) \qquad \textbf{(28)}$$

Figure 7 shows the expected benefit calculated using (28), for different values of C. C=0 corresponds to a delta function benefit; the curve is identical to the posterior pdf (similar to those shown in figure 5; this example is for a more accurate but equally unreliable observation). The analysis value which give the greatest expected benefit is shown by ×, and corresponds to the maximum of the posterior pdf. In practice, an analysis which is as accurate as the background x^b is still of some use. So it is more plausible that the width of the benefit function should be similar to that of the background pdf (which has variance 9 in our example). The expected benefit curve for this case has its maximum at +. It has chosen the peak from the posterior pdf which has the largest area (0.61 compared to 0.39). Finally, if we expand the width of the benefit function so that it is large compared to the separation between the peaks, then we get the curve with maximum at □. For very large C the gaussian benefit function becomes a quadratic, and the maximum is always at the mean value of the posterior pdf.

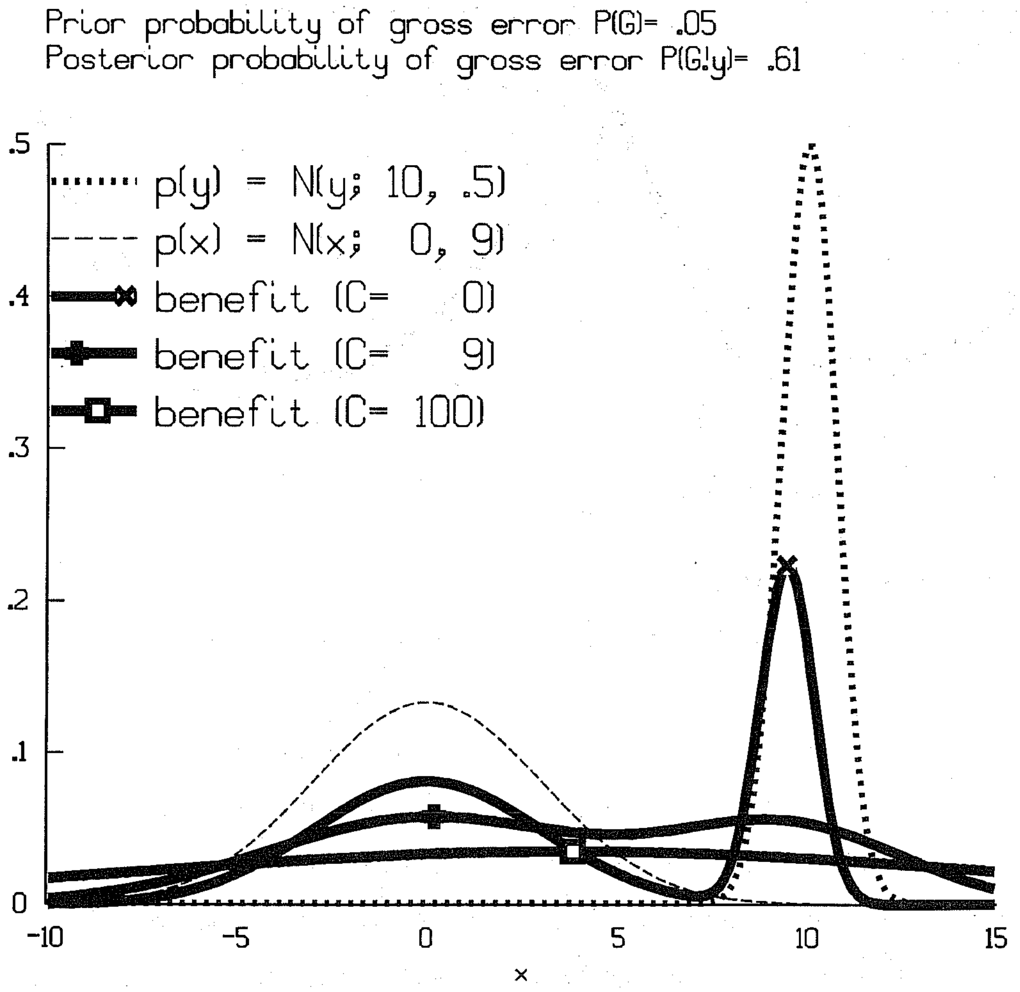

Fig.7. Expected benefit as a function of analysed value, from a case similar to those in figure 5. Curves are plotted for three different benefit functions, with widths C=0 (maximum at X), C=9 (maximum at +), and C=100 (maximum at □).
Shown for reference are the background pdf (with x^b=0, B=9), and the observational pdf (with y^o=10, R=0.5).
The prior probability of gross error was assumed to be $P(G)$=0.05, and the posterior probability was calculated to be $P(G|y^o)$=0.61.

3 PRACTICAL METHODS OF QUALITY CONTROL

3.1 Individual Quality Control

Lorenc and Hammon (1988) extended (26) to two observations with independent gross errors, and background values. This can be treated exactly by first calculating $P(G_i|y^o)$ for each, using (26), and then modifying them by a buddy check factor:

$$P(G_1|y_1^o \cap y_2^o) = P(G_1|y_1^o)\,\frac{P(y_1^o)P(y_2^o)}{P(y_1^o \cap y_2^o)} \qquad \textbf{(29)}$$

$$P(G_2|y_1^o \cap y_2^o) = P(G_2|y_2^o)\,\frac{P(y_1^o)P(y_2^o)}{P(y_1^o \cap y_2^o)} \qquad \textbf{(30)}$$

where

$$\frac{P(y_1^o)P(y_2^o)}{P(y_1^o \cap y_2^o)} = \left[1 - P(\bar{G}_1|y_1^o)P(\bar{G}_2|y_2^o)\left\{1 - \frac{P(y_1^o \cap y_2^o|\bar{G}_1 \cap \bar{G}_2)}{P(y_1^o|\bar{G}_1)P(y_2^o|\bar{G}_2)}\right\}\right]^{-1} \qquad \textbf{(31)}$$

The algebra, and computation, to extend this exact calculation to n observation goes as 2^n. However it has been found in practice that sequentially applying the two observation buddy check is a reasonable approximation. (See Lorenc and Hammon 1993 appendix C for more details of the pairwise buddy check). Finally, each observation i is used if $P(\bar{G}_i|y^o) > P(G_i|y^o)$. This is a generalisation of the decision made by the C=9 curve in figure 7.

The main weakness of this approach as implemented is its sequential nature. Nearby observations are not combined before checking an observation, but rather they are used one-by-one. So the way they support, or contradict, each other is only approximately allowed for. Note that for each observation the posterior pdf is split into two different peaks[8], leading to independent decisions for each observation which may not be consistent, as we shall see in 3.4.

[8] actually, they may not be distinct peaks.

3.2 Simultaneous Quality Control

Lorenc (1981) introduced the Optimal Interpolation (OI) analysis method used operationally at ECMWF (until replaced by 3DVAR). This performs an explicit solution of a quadratic variational problem. The solution is calculated in boxes for as many observations as we can afford to handle at once.

A key feature of the ECMWF system is the use of the same methodology for quality control. Lorenc (1981) shows how, once the inverse of the OI matrix **M** (=$\mathbf{HBH}^T$+**R** in our current notation) has been calculated, then it is possible with relatively few operations to solve the system of equations involving a smaller matrix omitting one (or a few) observations. He used this to check each observation in turn against a value analysed using all the other observations. An observation fails if:

$$(y^o - y^a)^2 > T^2(V_o + V_a) \tag{32}$$

where y^a, with error variance V_a, is the analysis obtained using the OI equations, at the position of the observation being checked, omitting the observation being checked and other rejected observations.

In the Lorenc (1981) paper the tolerance (T) was set in a somewhat empirical manner to 4.0. When, as the scheme developed, we tried to account for the better quality of weather ship observations by reducing their observational error V_o, we found that this resulted in more being rejected:- not what we wanted. (It was this behaviour, and the subjective tolerance in what was otherwise an objective analysis, that induced me to study the Bayesian approach.) It is shown in Lorenc and Hammon that, to match the criterion $P(\bar{G}_i|y^o)>P(G_i|y^o)$, the tolerance T should be given by:

$$T^2 = 2\ln\left[\frac{P(\bar{G})}{P(G)}\right] + \ln\left[\frac{k^{-2}}{2\pi(V_o+V_a)}\right] \tag{33}$$

where k is the probability density of observations with gross error, as defined in (27). T is shown in figure 8.

In applying this method, observations have to be either included in, or excluded from, the analysis. While an observation is checked, the decisions on other observations are frozen. The ECMWF scheme follows a pragmatic approach of rejecting the worst, then rechecking the others, iteratively until no more fail.

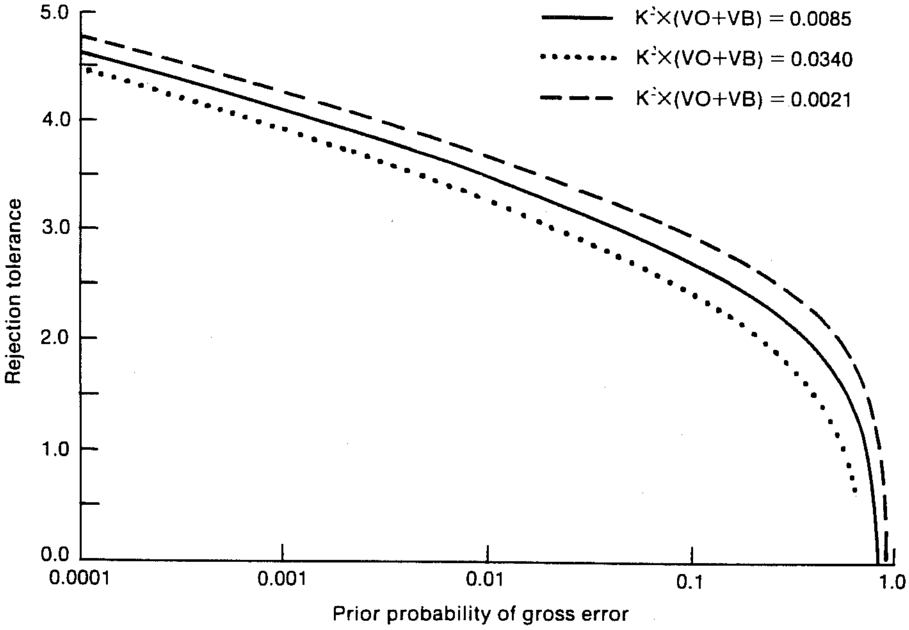

Fig.8. Rejection tolerance T, plotted against prior probability of gross error (from Lorenc and Hammon 1988).

Ingleby and Lorenc (1993) present equations for extending the Bayesian approach of Lorenc and Hammon. From the n gross error events G_i, they define 2^n new combined events C_α each corresponding to a particular set of rejections:

$$\begin{aligned} C_0 &= G_n \cap G_{n-1} \dots \cap G_2 \cap G_1 \\ C_1 &= G_n \cap G_{n-1} \dots \cap G_2 \cap \bar{G}_1 \\ C_2 &= G_n \cap G_{n-1} \dots \cap \bar{G}_2 \cap G_1 \\ &\vdots \\ C_{2^n-1} &= \bar{G}_n \cap \bar{G}_{n-1} \dots \cap \bar{G}_2 \cap \bar{G}_1 \end{aligned} \tag{34}$$

Bayes theorem can be applied to each of these combined events:

$$P(C_\alpha|y^o) = \frac{P(y^o|C_\alpha)P(C_\alpha)}{\sum_{\alpha'=0}^{2^n-1} P(y^o|C_{\alpha'})P(C_{\alpha'})} \tag{35}$$

Note that the denominator is the same in all the expressions; if we only want to find the most likely C_α it need not be evaluated. Evaluation of one $P(y^o|C_\alpha)$ involves evaluating only a single multivariate Gaussian. In fact the ECMWF OI method (with Bayesian tolerance T) is deciding which is more likely out of two C_α which differ just by the observation being tested. Because it is judging between sets of quality control decisions, we call this approach *Simultaneous Quality Control.*

The C_α can be regarded as being the vertices of an n-dimensional hyper-cube. The method starts from a first-guess set of rejections, and tests each observation in turn. This is equivalent to searching for the most probable of the adjacent vertices. It then iterates, re-testing some vertices adjacent to the new C_α. This strategy is like the SIMPLEX algorithm of linear programming, but applied to a non-linear problem.

The main weakness of this method as implemented is that its cost prohibits evaluation of more than a few of the possible combinations. The SIMPLEX-like algorithm is not guaranteed to find the absolute maximum, but depends on a good initial estimate as to which observations are correct. This is illustrated in section 3.4.

3.3 Variational Analysis with non-Gaussian Errors

It is possible to use our model of observational errors directly in a variational algorithm. Dharssi *et al.* (1992) did this for simulated windlidar observations. Instead of the quadratic $\frac{1}{2}(y^o-H(x))^T(\mathbf{E}+\mathbf{F})^{-1}(y^o-H(x))$, the observational penalty becomes (for diagonal **E+F**):

$$J_o = \sum_i -\ln\left[N(y_i^o|H(x),V_{r_i})P(\bar{G}_i)+k_iP(G_i)\right] \tag{36}$$

where V_{ri} $(=\mathrm{E}_{ii}+\mathrm{F}_{ii})$ is the observational error of i if it has not a gross error.

Differentiating this gives:

$$\frac{\partial J_o}{\partial y_i} = \frac{(y_i^o - y_i)}{V_{r_i}} \left\{ \frac{N\left(y_i^o \mid y_i, V_{r_i}\right) P(\bar{G}_i)}{N\left(y_i^o \mid y_i, V_{r_i}\right) P(\bar{G}_i) + k_i P(G_i)} \right\} \tag{37}$$

where y_i is the element of $H(\boldsymbol{x})$ corresponding to $\boldsymbol{x}$ interpolated to the ith observation position. The first term is just what we get if the observation error is Gaussian, as in variational methods with a quadratic penalty function. The term in braces is equal to the probability that observation i has not a gross error, given that $\boldsymbol{x}$ is exactly correct. So for each iteration u of the descent algorithm, all one has to do to allow for gross errors is to replace the observational error for each observation in the formulae for the standard gradient of a quadratic penalty by:

$$E_{o\,[u]} = \frac{V_r}{P(\bar{G} \mid y^o \cap H(x_{[u]}))} \tag{38}$$

i.e. the observational error variance should be inflated by one over the probability (given the current best estimate $\boldsymbol{x}_{[u]}$) that the observation has not a gross error.

Note that this only gives the correct gradient of J_o; it does not give the correct penalty (for which we need (36)) nor the correct second derivative. In general, analysis error estimates based on the second derivative, valid for Gaussian distributions (e.g. in (19)) will be over-optimistic for long-tailed distributions.

We saw in figure 5 that if errors are non-Gaussian, the penalty function is non-quadratic and can have multiple minima. So the end point of a descent algorithm iteration will depend on the first-guess. In a set of variational assimilation experiments with a one-dimensional shallow water model and its adjoint, I found that (for the example studied) the first-guess had to be very good to get convergence to the best solution (Lorenc 1988, figure 13).

Dharssi *et al.* (1992) studied various approaches for overcoming this problem for a simple example with two observations, so that the penalty function can be plotted as contours. Ordinary descent algorithms did not always find the lowest minimum (figure 9). Better results could be obtained by artificially increasing the assumed observational error for early iterations, slowly reducing it to its true value. Alternatively, one can artificially decrease the prior $P(G)$ to zero for early iterations. Neither approach always worked. In fact we saw in 2.7 that the extremum of the posterior pdf is not necessarily the best analysis, particularly for observations with low observational error but some gross errors. Increasing the observational error makes the modified posterior pdf more like the expected benefit curve for C=9 shown in figure 7. In simulations using rather dense observations with large probabilities of gross error (up to 50%), the method with increased observational error worked satisfactorily.

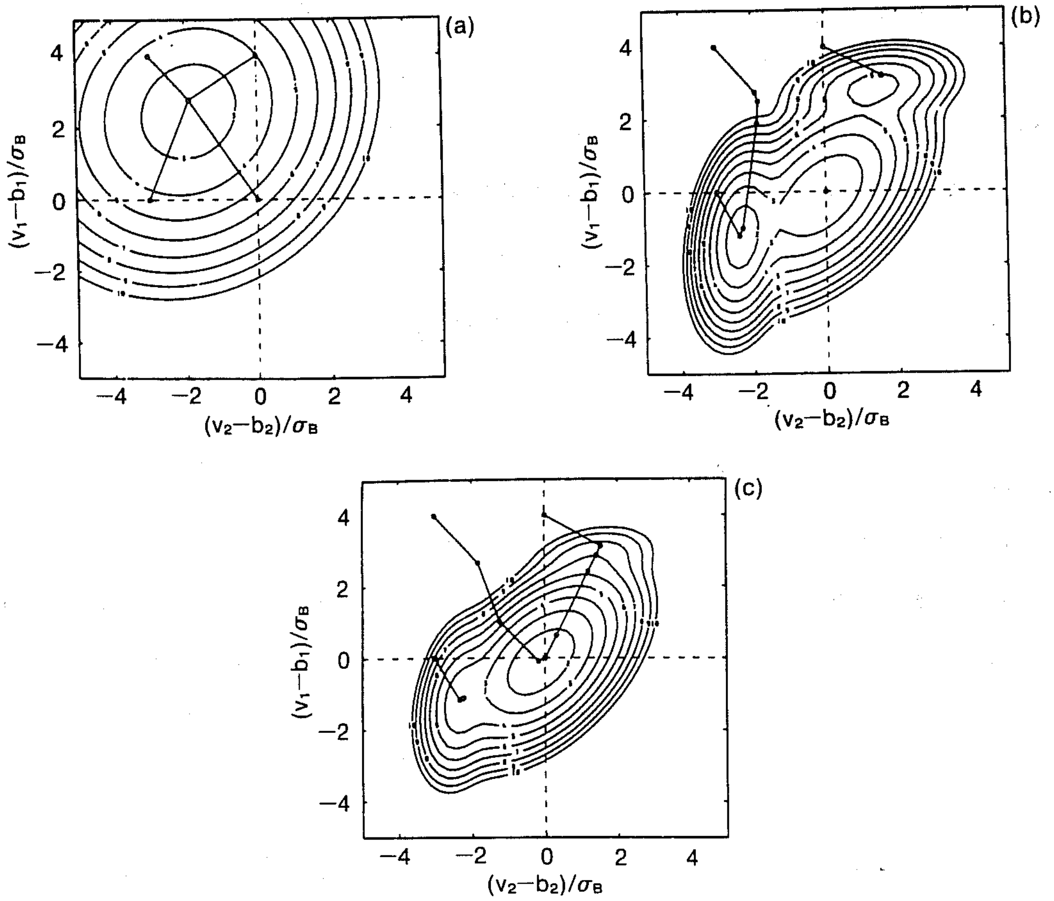

Figure 7. Contours of the penalty function (11) when only two observations are present. The observation values are given by $o_1 - b_1 = 4\,\sigma_B$, $o_2 - b_2 = -3\,\sigma_B$ and the background error correlation between the two observation points is set to 0.5. The observation error standard deviation $\sigma_0 = 0.5\,\sigma_B$ and $a_i = b_i$. (a) displays the contours when the initial probability of gross error $P_g = 0.0$. (b) and (c) display the contour maps when P_g is 0.1 and 0.5 respectively. The tracks superimposed (dots connected by solid line) represent the path taken by the iterative scheme (12) through this space, for four different starting points which are at (0, 0), (−3, 0), (−3, 4) and (0, 4).

Fig.9. Fig.7 from Dharssi *et al.* (1992).

The main weakness of this method is its use of a variational descent algorithm in situations where a discrete decision between distinct possibilities is needed. It is dependent on a good starting point initial guess. Another theoretical weakness is its search for the maximum of the posterior pdf, rather than the estimated benefit. We saw in 2.7 how an accurate but unreliable observation can give a tall but narrow peak, which is not actually the best analysis. This effect can be partially alleviated by artificially increasing the observational error.

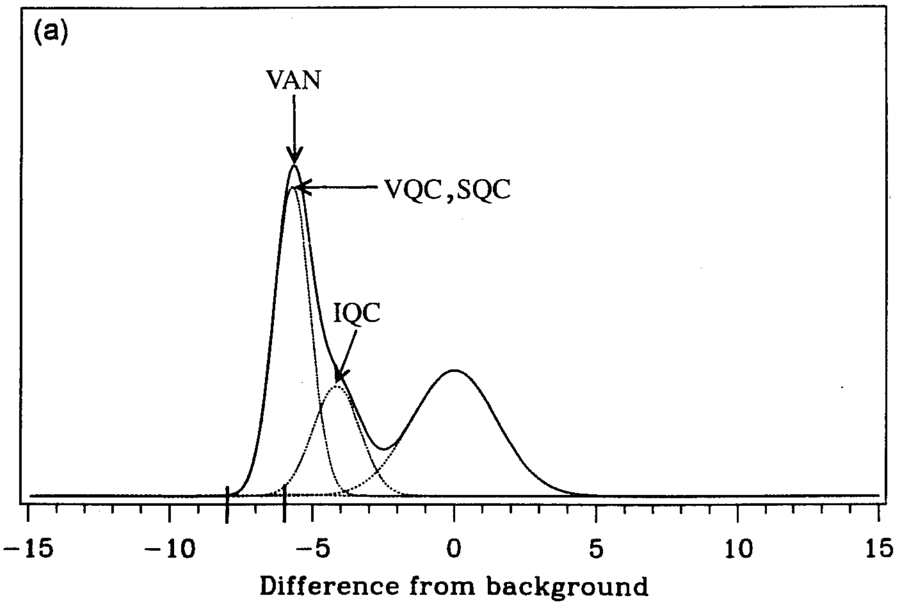

Fig.10 Posterior pdf for collocated observations differing by -8mb and -6mb from the background. The prior $P(G)$=.05 for each, the error variance of good observations is V_e=1.0mb^2, the background error variance is V_b=(1.5mb)^2, and the probability density of observations with gross errors is k=.043mb^{-1}. (from Ingleby and Lorenc 1993).

3.4 Comparison

Ingleby and Lorenc (1993) compared Individual Quality Control (IQC), Simultaneous Quality Control (SQC), and non-Gaussian Variational Analysis (VAN) for some simple examples. One is shown in figure 10. There are two observations and thus four combinations C_α, each corresponding to a dotted Gaussian curve in figure 10. The table below shows their posterior probabilities.

	G_2	$\bar{G}_2$	$G_2 \cup \bar{G}_2$
G_1	.393	.191	.584
$\bar{G}_1$	.003	.413	.416
$G_1 \cup \bar{G}_1$	.396	.604	1.0

The most likely combination is for them both to be correct: the table shows $P(\bar{G}_1 \cap \bar{G}_2)$=.413. This is the SQC result. Note that the simplex-like search algorithm would not work in this case: if we start at $G_1 \cap G_2$, both $\bar{G}_1 \cap G_2$ and $G_1 \cap \bar{G}_2$ are less likely, so we do not get to $\bar{G}_1 \cap \bar{G}_2$.

To get the probabilities for IQC we have to sum rows and columns: the table show $P(\bar{G}_1)$=.416,

$P(\bar{G}_2)$=.604. So observation 1 probably has a gross error, and observation 2 is probably correct. Note however that if we make these decisions individually, and then draw the analysis assuming $G_1 \cap \bar{G}_2$, then we are choosing a rather unlikely combination.

The variational analysis method chooses the highest point on the solid curve in figure 10. Note that a descent algorithm starting from the background would not have found this. If we convolve in a benefit function, as in 2.7, the best analysis will be near the mean of the larger peak in the posterior pdf, i.e. between the IQC and VAN/VQC/SQC results.

4 MONITORING

To the manager of a manufacturing company, Quality Control has a different meaning to that we have implied so far; he wants to know about and prevent errors. In NWP we call the equivalent process "monitoring". The purpose is to collect statistics on the performance of observing and processing systems, to detect systems that are not performing as expected, and to feed this information back so the deficiency is corrected at source. To do this we need:

- a comprehensive database of basic and processed observed values, independent estimates of the same quantities, and parameters affecting the processing
- software for categorising, sorting, and analysing the database
- effort to try categorisations and look for "unexpected" behaviour
- communications, willpower and persistence, to get errors from stages out of your direct control rectified.

Design of the monitoring system is as important as design of the data-assimilation scheme; it should not be added on as an afterthought. Monitoring by NWP centres of the operational World Weather Watch observations is probably the major cause of the significant increase in observational quality which has been seen in radiosondes (S. Uppala, personal communication), ships (C. Heasman, personal communication), and cloud track winds.

Another important product of monitoring is a good description of the observational error characteristics. If we are using the gross-error model, we need to know the prior probability of error, and the error distributions of gross-error and of "good" observations. Without these, the quality control is not objective [9]. Lorenc and Hammon (1988) showed how the statistics of observations processed by their quality-control scheme could be used (with some added "judgement") to "bootstrap" the assumed prior distributions.

For some observation types, more complicated error models are called for. For instance there are many different ways that a radiosonde temperature and geopotential report can be corrupted. Gandin *et al.* (1993) have devised a "Comprehensive Quality Control" scheme which looks for sixteen. Because of the redundancy of information in a radiosonde message, it is often possible to correct errors.

Many observations have bias errors, which monitoring statistics are useful in detecting and correcting. For instance many ships and buoys have mean surface pressure errors which persist until the instrument is recalibrated; the Met Office routinely updates a list of corrections for them. The "observational" errors in satellite radiance soundings are biassed by errors in the radiative transfer calculations used in *H*; all successful methods for using the radiances use empirical bias corrections obtained from a monitoring process.

[9] by "objective" I mean more than the automatic application of *ad hoc* rules, rather that the rules themselves have some statistical foundation.

References

Dharssi,I., Lorenc,A.C. and Ingleby,N.B. 1992
"Treatment of gross errors using maximum probability theory" *Quart.J.Roy.Met.Soc.*, **118**, 1017-1036

Gandin Lev S., L.L.Morone and W G Collins 1993
"Two years of operational comprehensive quality control at the National Meteorological Center" *Weather and Forecasting*, **8**, 57-72

Gauss 1823
"Theoria combinationis observationum erroribus minimis obnoxiae".

Ide, K., Courtier, P., Ghil, M., and Lorenc, A.C. 1996
"Unified notation for data assimilation: Operational, Sequential and Variational" *J.Met.Soc.Japan*, Special issue, *to appear*

Ingleby, N.B., and Lorenc, A.C. 1993
"Bayesian quality control using multivariate normal distributions". *Quart.J.Roy.Met.Soc.*, **119**, 1195-1225

Lorenc, A.C. 1981
"A global three-dimensional multivariate statistical analysis scheme." *Mon. Wea. Rev.*, **109**, 701-721.

Lorenc, A.C. 1986
"Analysis methods for numerical weather prediction." *Quart. J. Roy. Met. Soc.*, **112**, 1177-1194.

Lorenc, A.C. and Hammon, O., 1988
"Objective quality control of observations using Bayesian methods - Theory, and a practical implementation." *Quart. J. Roy. Met. Soc.,* **114**, 515-543

Lorenc, A.C. 1988
"Optimal nonlinear Objective Analysis." *Quart. J.,Roy. Met. Soc.,* **114**, 205-240.

Tarantola,A. 1987
"Inverse Problem Theory - Methods for data fitting and model parameter estimation". publ. Elsevier. 613pp. ISBN 0-444-42765-1