METEOROLOGICAL OFFICE

Short-range Forecasting Research

Short Range Forecasting Division

Scientific Paper No. 10

Bayesian quality control using multivariate normal distributions

by

N.B. Ingleby and A. Lorenc

July 1992

Meteorological Office
London Road
Bracknell
Berkshire
RG12 2SZ
United Kingdom

Short Range Forecasting Research

Scientific Paper No.10

Bayesian quality control using multivariate normal distributions

by

N.B. Ingleby and A. Lorenc

July 1992

Short Range Forecasting Research
Meteorological Office
London Road
Bracknell
Berkshire RG12 2SZ
ENGLAND

Bayesian quality control using multivariate normal distributions

N Bruce Ingleby and Andrew C Lorenc

Abstract

An expression for the probability density of any distribution of observed values (given background values of known accuracy) is derived from the properties of multivariate normal distributions. This is used in the quality control of observations - 'good' and 'bad' observations are assumed to have errors from a normal distribution or a distribution giving no useful information respectively.

Three methods of quality control are presented and compared; two of these are based on the probability density derived above, and the third is based on a related maximum probability analysis. They differ in the optimality principal used: Individual Quality Control finds the most likely quality (i.e. good or bad) for each observation, given information from all the others, Simultaneous Quality Control finds the most likely combination of qualities, while Variational Quality Control is based on a Variational Analysis which finds the most likely true values. The multi-observation framework used includes the 'background' check as a special case, and it is extended to deal with observations with common sources of gross error. Applications to multilevel checks, bias checks and checks for known error patterns are sketched.

As a by-product the standard statistical interpolation formulae are derived from the properties of normal distributions thus demonstrating the implicit dependence of statistical interpolation on the normal distribution.

1. Introduction

Objective analysis schemes are used to provide initial conditions for numerical weather prediction. Current operational methods are sensitive to large errors in observations and so a preliminary quality control (qc) step is included to remove 'bad' observations. For instance Lorenc (1981) adapted the full statistical interpolation (OI) method to check the data being used, and Lorenc and Hammon (1988) used Bayesian methods to perform objective quality control (these methods are used operationally at ECMWF and the Meteorological Office respectively). Alternatively it is possible to design analysis systems that are insensitive to 'bad' observations, and so to combine the quality control and analysis functions (eg Dharssi et al, 1992). This paper arose from an attempt to understand the relationship between the different methods, and to provide the theoretical framework essential to enable informed decisions about the implementation of practical approximations to optimal objective[1] quality control methods.

We present a probabilistic framework for checking observations using information from nearby observations ('buddies') and from a background field (eg from a forecast). Checks using the background alone are included as a special case. The framework is quite general and applies to an arbitrary mix of observation types. The theory can be used directly, or used to guide the development of approximate methods - some of which are presented. The main assumption made is that we know the error distributions of good and bad observations, and of the background. Furthermore, to make the problem manageable, we assume that the background and good observations have normally distributed errors.

Lorenc and Hammon (1988) presented a method for quality control of observations based on Bayesian probability theory and assumptions about the distribution of observations both with gross errors, and without (=bad/good observations respectively). They explored the single observation case, and gave an extension to two observations, and by an approximate iterative method to n observations. In this paper the equations for the multiple observation case are derived from first principles (Section 2) - first for the case where there are no bad observations, then including gross errors with properties as assumed by Lorenc and Hammon (1988). In section 3 the theory is expanded to include gross errors affecting several observed values, and a brief comparison with previous work is made. The application to special tests and correction of observations is also sketched. Simple examples and a discussion of the different methods are given in section 4, and we summarise in section 5.

[1] By 'objective' we imply more than the automatic application of ad hoc rules, rather that the rules themselves have some theoretical foundation.

2. Theory

Notation

y, **x**, $\boldsymbol{\mu}$ etc (lower case bold) are column vectors, with elements y_i.
0, **B**, $\boldsymbol{\Sigma}$ etc (upper case bold) are matrices.
Superscript T (eg $\mathbf{y}^T$, $\mathbf{K}^T$) denotes matrix transpose.

P(A) is the probability of event A and P(**y**) the probability that the vector takes a value in a volume **dy** surrounding **y**. P(**y**) = *p*(**y**)**dy** where *p* is the probability density function (pdf). (For convenience **y** is being used to represent two different things: a vector, and the event that the vector takes a value in **dy**).

A∩B denotes 'A and B'. P(A|B)=P(A∩B)/P(B) is the conditional probability of A, given B. P(A∩B) can be written in two ways to give Bayes' Theorem:

$$P(A\cap B) = P(A|B)P(B) = P(B|A)P(A) \qquad \Rightarrow \qquad P(A|B) = P(B|A)P(A)/P(B)$$

2.1 Some properties of multivariate normal distributions

A vector **y**, of length n, is normally distributed if it has a probability density function

$$p(\mathbf{y}) = \{(2\pi)^n|\boldsymbol{\Sigma}|\}^{-0.5} \exp[-0.5(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})] \qquad (1)$$

where $\boldsymbol{\Sigma}$ is an n×n symmetric positive definite matrix with determinant $|\boldsymbol{\Sigma}|$, and $\boldsymbol{\mu}$ is a vector of length n. **y** is said to be distributed as $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, for the pdf we use the notation $p(\mathbf{y}) = n(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. $p(\mathbf{y})$ is normalised so that

$$\int p(\mathbf{y})\ \mathbf{dy} = 1 \qquad (2)$$

y has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, i.e.

$$\int \mathbf{y}\ p(\mathbf{y})\ \mathbf{dy} = \boldsymbol{\mu} \qquad (3)$$

and

$$\int (\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T\ p(\mathbf{y})\ \mathbf{dy} = \boldsymbol{\Sigma} \qquad (4)$$

Any linear combination of the y_i is also normally distributed. For any m×n matrix **W** of rank m ≤ n then **z** = **Wy** has pdf

$$p(\mathbf{z}) = n(\mathbf{z}|\ \mathbf{W}\boldsymbol{\mu},\ \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^T) \qquad (5)$$

Consider **y** split into two vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ of length k and n-k respectively and μ and Σ partitioned in the same way

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \; \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \; \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{6}$$

Σ_{11} and Σ_{22} are the covariance matrices of $\mathbf{y}_1$ and $\mathbf{y}_2$ of order k and n-k respectively. $\Sigma_{12}=\Sigma_{21}^T$ has dimensions k×(n-k) and is the covariance of $\mathbf{y}_1$ with $\mathbf{y}_2$.

The marginal pdf of $\mathbf{y}_2$ is defined as

$$p(\mathbf{y}_2) = \int p(\mathbf{y}) \, \mathbf{dy}_1 = \int p(\mathbf{y}_1 \cap \mathbf{y}_2) \, \mathbf{dy}_1 \tag{7}$$

The conditional probability density of $\mathbf{y}_1$, given the value of $\mathbf{y}_2$ is

$$p(\mathbf{y}_1|\mathbf{y}_2) = p(\mathbf{y}_1 \cap \mathbf{y}_2)/p(\mathbf{y}_2) = p(\mathbf{y})/p(\mathbf{y}_2) \tag{8}$$

and so the density of $\mathbf{y}$ is given by the product of (7) and (8)

$$p(\mathbf{y}) = p(\mathbf{y}_1|\mathbf{y}_2)p(\mathbf{y}_2) \tag{9}$$

For the multivariate normal distribution the marginal pdf is particularly simple, being multivariate normal with mean and covariance given by just picking the relevant elements of μ and Σ.

$$p(\mathbf{y}_2) = \mathrm{n}(\mathbf{y}_2|\mu_2, \Sigma_{22}) \tag{10}$$

The conditional distribution $p(\mathbf{y}_1|\mathbf{y}_2)$ is also multivariate normal

$$p(\mathbf{y}_1|\mathbf{y}_2) = \mathrm{n}(\mathbf{y}_1| \; \mu_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2-\mu_2), \; \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{11}$$

The mean $\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2-\mu_2)$ is known as the regression function of $\mathbf{y}_1$ on $\mathbf{y}_2$.

The convolution of two multivariate normal distributions of the same order (see example in next section) is

$$\int \mathrm{n}(\mathbf{y}_1|\mathbf{y}, \Sigma_{11})\mathrm{n}(\mathbf{y}|\mathbf{y}_2, \Sigma_{22})\mathbf{dy} = \mathrm{n}(\mathbf{y}_1|\mathbf{y}_2, \Sigma_{11}+\Sigma_{22}) \tag{12}$$

We shall also quote the result that the quadratic form $(\mathbf{y}-\mu)\Sigma^{-1}(\mathbf{y}-\mu)$ has a χ^2 distribution with n degrees of freedom.

These properties of multivariate normal distributions, except (12), can be found in Chatfield and Collins (1980) (chapter 6) or the extensive description by Anderson (1984) (section 2.3 for results (1) to (4), section 2.4 for (5) and (10), section 2.5 for (11) and section 3.3 for the quadratic form). A proof of (12) is given in Tarantola (1987) problem 1.20. Some of these properties are given in the Appendices of Thiébaux and Pedder (1987), and their Appendix III 'The statistical basis of statistical objective analysis' relates to some of the analysis equations in this paper.

When applied to meteorological analysis systems 'multivariate' usually means a simultaneous analysis of mass and wind variables. In this paper 'multivariate' simply means many variables in the statistical sense, and the vector $\mathbf{y}$ can contain pressure, wind, temperature or humidity variables or some mixture of them. The observations $\mathbf{y}_o$ can come from one observing platform or from many.

2.2 Analysis using multivariate normal distributions

We start with background and observed values assumed to have multivariate normal probability density functions. For simplicity the interpolation to observation points is ignored, i.e. the background is taken to be available at observation locations. All probabilities will be taken to be conditional on knowing the background values (which we always have available).

$$\mathrm{P}(\mathbf{y}) \;=\; p(\mathbf{y})\mathbf{dy} \;=\; \{(2\pi)^n|\mathbf{B}|\}^{-0.5}\,\exp[-0.5(\mathbf{y}-\mathbf{y}_b)^T\mathbf{B}^{-1}(\mathbf{y}-\mathbf{y}_b)]\mathbf{dy} \qquad (13)$$

$$\mathrm{P}(\mathbf{y}_o|\mathbf{y}) = p(\mathbf{y}_o|\mathbf{y})\mathbf{dy}_o = \{(2\pi)^n|\mathbf{O}|\}^{-0.5}\,\exp[-0.5(\mathbf{y}-\mathbf{y}_o)^T\mathbf{O}^{-1}(\mathbf{y}-\mathbf{y}_o)]\mathbf{dy}_o \qquad (14)$$

$\mathbf{B}$ and $\mathbf{O}$ are the background and observation error covariance matrices. $\mathbf{y}$, $\mathbf{y}_b$ and $\mathbf{y}_o$ are the vectors of true, background and observed values respectively ($\mathbf{y}_a$ will denote analysed values). n is the number of observations.

The joint probability of $\mathbf{y}$ and $\mathbf{y}_o$ (denoted $p(\mathbf{y}\cap\mathbf{y}_o)$) is

$$\begin{aligned}p(\mathbf{y}\cap\mathbf{y}_o) &= p(\mathbf{y}_o|\mathbf{y})p(\mathbf{y})\\ &= \{(2\pi)^n|\mathbf{B}+\mathbf{O}|\}^{-0.5}\,\exp[-0.5(\mathbf{y}_o-\mathbf{y}_b)^T(\mathbf{B}+\mathbf{O})^{-1}(\mathbf{y}_o-\mathbf{y}_b)]\\ &\quad\{(2\pi)^n|(\mathbf{B}^{-1}+\mathbf{O}^{-1})^{-1}|\}^{-0.5}\,\exp[-0.5(\mathbf{y}-\mathbf{y}_m)^T(\mathbf{B}^{-1}+\mathbf{O}^{-1})(\mathbf{y}-\mathbf{y}_m)]\end{aligned}$$

where $\quad \mathbf{y}_m = \mathbf{y}_b + \mathbf{B}(\mathbf{O}+\mathbf{B})^{-1}(\mathbf{y}_o-\mathbf{y}_b) \qquad (15)$

The probability of $\mathbf{y}_o$ occurring is

$$\begin{aligned}p(\mathbf{y}_o) &= \int p(\mathbf{y}\cap\mathbf{y}_o)\mathbf{dy}\\ &= \{(2\pi)^n|\mathbf{B}+\mathbf{O}|\}^{-0.5}\,\exp[-0.5(\mathbf{y}_b-\mathbf{y}_o)^T(\mathbf{B}+\mathbf{O})^{-1}(\mathbf{y}_b-\mathbf{y}_o)] \qquad (16)\end{aligned}$$

(15) and (16) (the product and convolution of two multivariate normal distributions) follow from the properties of normal distributions eg Tarantola

(1987) problem 1.20 ($\mathbf{y}_m$ can be evaluated from his $\mathbf{A}^{-1}\mathbf{b}$). The normalisation coefficients can be derived directly ($\mathbf{B}(\mathbf{B}^{-1}+\mathbf{O}^{-1})\mathbf{O} = \mathbf{O}+\mathbf{B} \Rightarrow |\mathbf{O}||\mathbf{B}| = |\mathbf{O}+\mathbf{B}||(\mathbf{B}^{-1}+\mathbf{O}^{-1})^{-1}|$) or simply deduced from the normalisation condition.

Thus the probability density of the observations is a multivariate normal function of the differences from the background values. This result will be used extensively in the following sections.[2]

From (15) and (16) we obtain the distribution of $\mathbf{y}$ given $\mathbf{y}_o$

$$p(\mathbf{y}|\mathbf{y}_o) = p(\mathbf{y}\cap\mathbf{y}_o)/p(\mathbf{y}_o)$$

$$p(\mathbf{y}|\mathbf{y}_o) = \{(2\pi)^n|(\mathbf{B}^{-1}+\mathbf{O}^{-1})^{-1}|\}^{-0.5} \exp[-0.5(\mathbf{y}-\mathbf{y}_m)^T(\mathbf{B}^{-1}+\mathbf{O}^{-1})(\mathbf{y}-\mathbf{y}_m)] \qquad (17)$$

The posterior distribution of $\mathbf{y}$ is the conditional probability density given by (17) (cf equation 1.37 of Tarantola, 1987). It has a multivariate normal distribution with mean $\mathbf{y}_m$ and covariance $(\mathbf{B}^{-1}+\mathbf{O}^{-1})^{-1} = \mathbf{B} - \mathbf{B}(\mathbf{B}+\mathbf{O})^{-1}\mathbf{B}$. These equations can be generalised to account for interpolation of the background to observation locations (see Appendix A). In both cases $p(\mathbf{y}|\mathbf{y}_o)$ has a multivariate normal distribution with mean and covariances given by the statistical interpolation equations.

The analysis at observation points, $\mathbf{y}_a$, has to be deduced from $p(\mathbf{y}|\mathbf{y}_o)$. The maximum probability analysis is $\mathbf{y}_{max}$ which maximises $p(\mathbf{y}|\mathbf{y}_o)$ (the mode of the distribution). The mean analysis is $\bar{\mathbf{y}} = \int \mathbf{y}p(\mathbf{y}|\mathbf{y}_o)d\mathbf{y}$. $\bar{\mathbf{y}}$ is also the minimum variance analysis, i.e. $\mathbf{y}_1=\bar{\mathbf{y}}$ minimises $\int (\mathbf{y}_1-\mathbf{y})^T(\mathbf{y}_1-\mathbf{y})p(\mathbf{y}|\mathbf{y}_o)d\mathbf{y}$. For normal distributions $\mathbf{y}_{max}$ and $\bar{\mathbf{y}}$ are identical, and equal to $\mathbf{y}_m$.

Figure 1 provides a simple one dimensional illustration of these equations for particular values of $\mathbf{y}_o$. The dashed line represents the background pdf $p(\mathbf{y}|\mathbf{y}_b)$, the dotted line the observation pdf $p(\mathbf{y}_o|\mathbf{y})$, and the solid line is $p(\mathbf{y}_o\cap\mathbf{y})$ - the product of the other two. The area under the

[2] Collocated observations:

If some of the observations measure the same quantity then the covariance matrix $\mathbf{B}$ is singular, and (13) would have to be replaced by a lower dimensional normal distribution. However this can be treated as the limit when the off-diagonal correlations tend to 1, and all the other equations remain valid. In particular in (16) $\mathbf{B}+\mathbf{O}$ is non-singular so the individual and simultaneous quality control methods (section 2.4) are not affected. In the variational analysis method (section 2.5) equation (13) has to be modified.

solid curve is the probability density of the observed value, which gets smaller the further the observation is from the background (compare a and b). If the solid curve is normalised so that its area is 1 then it is $p(\mathbf{y}|\mathbf{y}_o)$, the posterior distribution of $\mathbf{y}$. The shape of $p(\mathbf{y}|\mathbf{y}_o)$ is independent of $\mathbf{y}_o$.

2.3 Gross errors in observations

We follow Lorenc and Hammon (1988) in assuming that an observation is either good, in which case its error is from a normal distribution, or it has a gross error, in which case all plausible values are equally likely. We also assume that observational errors are independent. Thus we replace (14) with

$$\begin{aligned} p(\mathbf{y}_o|\mathbf{y}) &= \prod_{i=1}^{n} p(y_{oi}|y_i) \\ &= \prod_{i=1}^{n} \{p(y_{oi}|y_i \cap \bar{G}_i)P(\bar{G}_i) + p(y_{oi}|y_i \cap G_i)P(G_i)\} \\ &= \prod_{i=1}^{n} \{p(y_{oi}|y_i \cap \bar{G}_i)P(\bar{G}_i) + k_i P(G_i)\} \qquad (18) \end{aligned}$$

i is an index over the observations, G_i is the event that observation i contains a gross error and $\bar{G}_i$ is its converse. $p(y_{oi}|y_i \cap G_i)$, the pdf for observations with gross errors, is taken to be constant. Strictly speaking we take $p(y_{oi}|y_i \cap G_i) = k_i$ within an interval of width $1/k_i$ and zero outside. The interval should cover all non-negligible values of $p(y_i)$ as we will approximate $\int p(y_i)p(y_{oi}|y_i \cap G_i)dy_i$ by $k_i \int p(y_i)dy_i = k_i$ in (23).

The total probability density of $\mathbf{y}_o$ is

$$\begin{aligned} p(\mathbf{y}_o) &= \int p(\mathbf{y} \cap \mathbf{y}_o)\mathbf{dy} = \int p(\mathbf{y})p(\mathbf{y}_o|\mathbf{y})\mathbf{dy} \\ &= \int p(\mathbf{y}) \prod_{i=1}^{n} \{p(y_{oi}|y_i \cap \bar{G}_i)P(\bar{G}_i) + k_i P(G_i)\}\ \mathbf{dy} \qquad (19) \end{aligned}$$

If we multiply out the product there is a term for all subsets of $\{G_1,\ldots,G_n\}$ i.e. 2^n terms. We introduce a (binary) notation for all these combinations of good and bad data:

$$\begin{aligned} C_0 &= G_n \cap G_{n-1} \ldots G_2 \cap G_1 \\ C_1 &= G_n \cap G_{n-1} \ldots G_2 \cap \bar{G}_1 \\ C_2 &= G_n \cap G_{n-1} \ldots \bar{G}_2 \cap G_1 \\ &\vdots \\ C_{2^n-1} &= \bar{G}_n \cap \bar{G}_{n-1} \ldots \bar{G}_2 \cap \bar{G}_1 \qquad (20) \end{aligned}$$

For each C_α, α is written in binary and the bits numbered from 1 to n starting from the right. Bit i = 0 (1) corresponds to G_i ($\bar{G}_i$). C_0 denotes that all observations have gross errors, C_{2^n-1} that no observations have gross errors.

Thus (19) can be rewritten as

$$p(\mathbf{y}_o) = \int p(\mathbf{y}) \sum_{\alpha=0}^{2^n-1} p(\mathbf{y}_o|\mathbf{y}\cap C_\alpha)P(C_\alpha)\ \mathbf{dy} \tag{21}$$

or $$p(\mathbf{y}_o) = \sum_{\alpha=0}^{2^n-1} p(\mathbf{y}_o|C_\alpha)P(C_\alpha) \tag{22}$$

Let us treat a term corresponding to a particular combination C_α where the first k observations ($\mathbf{y}_{o1}$) do not have gross errors, but the remaining n-k ($\mathbf{y}_{o2}$) do. If the gross error mechanisms in each datum are independent (as we implicitly assumed in (18)), then $P(C_\alpha)$ is the product over i of $P(G_i)$ or $P(\bar{G}_i)$ as appropriate.

$$p(\mathbf{y}_o|C_\alpha)P(C_\alpha) = \int p(\mathbf{y})p(\mathbf{y}_o|\mathbf{y}\cap C_\alpha)P(C_\alpha)\ \mathbf{dy}$$

$$= \int p(\mathbf{y}) \prod_{i=1}^{k} p(y_{oi}|y_i\cap\bar{G}_i)P(\bar{G}_i) \prod_{j=k+1}^{n} k_j P(G_j)\ \mathbf{dy} \tag{23}$$

The components of $\mathbf{y}_2$ only appear within $p(\mathbf{y})$. Then

$$p(\mathbf{y}_o|C_\alpha)P(C_\alpha) = \prod_{j=k+1}^{n} k_j P(G_j) \int \prod_{i=1}^{k} p(y_{oi}|y_i\cap\bar{G}_i)P(\bar{G}_i) \left(\int p(\mathbf{y})\mathbf{dy}_2\right)\mathbf{dy}_1 \tag{24}$$

From (10) the marginal probability of $\mathbf{y}_1$ is just the multivariate normal distribution with correlation matrix $\mathbf{B}_{11}$

$$\int p(\mathbf{y})\mathbf{dy}_2 = \{(2\pi)^k|\mathbf{B}_{11}|\}^{-0.5} \exp[-0.5(\mathbf{y}_1-\mathbf{y}_{b1})^T\mathbf{B}_{11}^{-1}(\mathbf{y}_1-\mathbf{y}_{b1})]$$

$$= p(\mathbf{y}_1) \tag{25}$$

If the $p(y_{oi}|y_i\cap\bar{G}_i)$ are each normal then we can write

$$\prod_{i=1}^{k} p(y_{oi}|y_i\cap\bar{G}_i) = \{(2\pi)^k|O_{11}|\}^{-0.5} \exp[-0.5(\mathbf{y}_1-\mathbf{y}_{o1})^T\mathbf{O}_{11}^{-1}(\mathbf{y}_1-\mathbf{y}_{o1})]$$

$$= p(\mathbf{y}_{o1}|\mathbf{y}_1) \tag{26}$$

Substituting (25) and (26) in (24) we have

$$p(\mathbf{y}_o|C_\alpha)P(C_\alpha) = P(C_\alpha) \prod_{j=k+1}^{n} k_j \int p(\mathbf{y}_1)p(\mathbf{y}_{o1}|\mathbf{y}_1)\mathbf{dy}_1$$

$$= P(C_\alpha) \prod_{j=k+1}^{n} k_j \{(2\pi)^k|\mathbf{B}_{11}+\mathbf{O}_{11}|\}^{-0.5} \exp[-0.5(\mathbf{y}_{b1}-\mathbf{y}_{o1})^T(\mathbf{B}_{11}+\mathbf{O}_{11})^{-1}(\mathbf{y}_{b1}-\mathbf{y}_{o1})]$$

$$\text{where } P(C_\alpha) = \prod_{i=1}^{k} P(\bar{G}_i) \prod_{j=k+1}^{n} P(G_j) \tag{27}$$

as in (16). Some notes on the computation of (27) are given in Appendix B. The assumption that $\mathbf{O}$ is diagonal - used in writing (18) - is not necessary for equations (27) to (30). Any combination C_α can be reordered into the form used above; in general $\mathbf{y}_1$ is formed by selecting those observations assumed useful, and $\mathbf{B}_{11}$ and $\mathbf{O}_{11}$ are formed from the corresponding elements of **B** and **O**.

The probability of combination C_α is

$$P(C_\alpha|\mathbf{y}_o)=p(\mathbf{y}_o|C_\alpha)P(C_\alpha)/p(\mathbf{y}_o) \qquad (28)$$

which we can calculate from (27) and (22).

We can also calculate the probability of gross error in individual observations $P(G_i|\mathbf{y}_o)$ using

$$\begin{aligned}P(G_i|\mathbf{y}_o) &= p(\mathbf{y}_o|G_i)P(G_i)/p(\mathbf{y}_o)\\ &= p(\mathbf{y}_{o-i})k_iP(G_i)/p(\mathbf{y}_o) \qquad (29)\end{aligned}$$

where $\mathbf{y}_{o-i}$ is the vector of observations excluding i. Multiplying out as in (21).

$$P(G_i|\mathbf{y}_o) = \sum_{\alpha=0}^{2^n-1} \gamma_i(\alpha)p(\mathbf{y}_o|C_\alpha)P(C_\alpha) \Big/ \sum_{\alpha=0}^{2^n-1} p(\mathbf{y}_o|C_\alpha)P(C_\alpha)$$

where $\gamma_i(\alpha)=1$ if $G_i \in C_\alpha$, 0 otherwise (30)

There are 2^{n-1} non-zero terms in the numerator, these are identical to the terms in the denominator which contain G_i. The terms are given by (27).

The derivation given here makes it clear that $P(C_\alpha|\mathbf{y}_o)$ and $P(G_i|\mathbf{y}_o)$ are derived from integration over all possible states. Their use is discussed in the next section.

The observation values in vector $\mathbf{y}_o$ are being compared with each other to determine how likely each is to contain a gross error - a 'buddy check'. $\mathbf{y}_o$ could be a radiosonde temperature ascent - in which case the buddies are other values in the same ascent (with correlated observation errors - non-diagonal **O**). If we take $\mathbf{y}_o$ to be a single observation then (30) reduces to the 'background check' of Lorenc and Hammon (1988).

Lorenc and Hammon (1988), considering two observations, calculated $p(\mathbf{y}_o|\bar{G}_1 \cap \bar{G}_2)$ using statistical interpolation theory, this can be shown to give a bivariate normal distribution - a special case of (16). Their calculation of $P(G_i|\mathbf{y}_o)$ is thus equivalent to (28) for the two observation case. Their extension to more observations involves repeated application of this pairwise check giving an approximation to $P(G_i|\mathbf{y}_o)$. See appendix C for details.

2.4 Individual and Simultaneous Quality Control

Individual quality control (IQC)

One way of quality controlling the observations is to calculate $P(G_i|\mathbf{y}_o)$ for each observation using (30), and to choose C_α such that it contains:

G_i if $P(G_i|\mathbf{y}_o) > 0.5$

$\bar{G}_i$ if $P(G_i|\mathbf{y}_o) \leq 0.5$

This will be referred to as individual quality control (IQC) in the subsequent discussion, because although all observations are being considered the calculation (30) has to be repeated for each observation. This is essentially the method of Lorenc and Hammon (1988), although as mentioned above they calculated an approximation to $P(G_i|\mathbf{y}_o)$.

In IQC decisions about the quality of each datum are taken one by one, after the data have been compared with the background and each other. This could conceivably, in a borderline case, lead to a set of contradictory observations being accepted (see examples in section 4).

Simultaneous quality control (SQC)

The combined quality control and analysis problem consists of finding the best analysis given a multidimensional probability distribution which may well be multimodal, particularly if quality control is important. If gross observational errors following the simple model of this paper are the only source of nonlinearity, then the analysis probability distribution function will be a linear combination of multidimensional normal distributions. One reasonable strategy for choosing the best analysis is to choose the normal distribution with the largest integrated probability, and to set the analysis to its mean.

This strategy will be referred to as simultaneous quality control (SQC). It is equivalent to choosing the combination of accept/reject decisions for each datum that is most likely to be correct, and using it in a linear statistical interpolation analysis. SQC provides a final quality control decision, rather than a probability of gross error, and the denominator $p(\mathbf{y}_o)$ in (28) need not be evaluated. All that is required is to find the C_α with the maximum $p(\mathbf{y}_o|C_\alpha)P(C_\alpha)$, where $p(\mathbf{y}_o|C_\alpha)$ is a multivariate normal distribution (27).

The states C_α correspond to vertices of a hypercube. There is no simple way to guarantee finding the maximum $p(\mathbf{y}_o|C_\alpha)P(C_\alpha)$ without searching them all. One approximate method is to move from vertex to adjacent vertex - i.e. changing the decision on one observation at a time so that $p(\mathbf{y}_o|C_\alpha)P(C_\alpha)$ increases each time. This will converge to a local maximum, but not necessarily the global maximum. This method is similar to the simplex algorithm in linear programming, however in linear programming there is only one maximum, to which the simplex algorithm is guaranteed to converge. To increase the probability of finding the global maximum several randomly selected starting conditions could be used. Fischler and Bolles (1981) take this sort of approach starting from just a few observations each time and seeking to add consistent observations. However in the meteorological context most observations have small $P(G_i)$ (<0.1) and it seems to be a better strategy to start with all or most observations and seek to remove inconsistent ones.

This is similar to the ECMWF buddy check method of Lorenc (1981): 'all data ... are compared with an interpolated value obtained not using the datum being checked or any flagged data. If the absolute value of this deviation is more than four times the estimated interpolation error then the datum is considered to have failed. If more than one fail, then the worst is flagged and the rest of the failures are rechecked not using it.' In this paper the accept/reject criterion is expressed in probabilistic terms, its relationship to the deviation from an interpolated (analysed) value excluding the datum being checked is clarified in Appendix D.

Using this kind of algorithm only a small subset of the total 2^n terms will be evaluated, and so the approximate SQC will be significantly cheaper than the full SQC or IQC.

2.5 **Maximum probability analysis** (VAN and VQC)

Yet another alternative is to combine the quality control and analysis steps, and to determine a maximum probability analysis directly (suggested by Purser, 1984). The basic equation is

$$p(\mathbf{y}|\mathbf{y}_o) = p(\mathbf{y}_o|\mathbf{y})p(\mathbf{y})/p(\mathbf{y}_o) \propto p(\mathbf{y}_o|\mathbf{y})p(\mathbf{y}) \qquad (31)$$

Nonlinear optimization methods which attempt to search the phase space are computationally prohibitively expensive for application to the large volumes of meteorological data. Variational methods, using descent algorithms, are not guaranteed to find the best solution in nonlinear cases although in some instances they do. Dharssi et al (1992) described such an iterative non-linear analysis scheme for finding **y** which maximises (31) when $p(\mathbf{y})$ is given by (13) and $p(\mathbf{y}_o|\mathbf{y})$ by (18). The weight given to each observation is multiplied by a term we denote here by $1-P_y(G_i)$. $P_y(G_i)$ can be thought of as the probability of G_i given that the current iteration's estimate of **y** is correct; thus the weight given to observations with a high probability of gross error is nearly zero. $P_y(G_i)$ is given by

$$P_y(G_i) = k_iP(G_i)\Big/\left\{k_iP(G_i) + P(\bar{G}_i)(\sigma_{oi}^2 2\pi)^{-0.5}\exp\left(-(o_i-y_i)^2/\ 2\sigma_{oi}^2\right)\right\} \qquad (32)$$

After the iteration has converged to $\mathbf{y}_{max}$, $P_y(G_i)$ is given by (32) with $y_i=(\mathbf{y}_{max})_i$. This is not identical to $P(G_i|\mathbf{y}_o)$. The latter is the expected probability of gross error, taking into account all possible true values and their probabilities. $P_y(G_i)$ is the probability of gross error for one (the most likely) true value.

Dharssi et al (1992) also showed that it is possible, and computationally cheaper, to use the variational analysis scheme to quality control subsets of the observations using (32) and then use a conventional linear analysis. The distinction between using the variational analysis directly, and using it to quality control the observations used in a linear analysis (these will be referred to as VAN and VQC respectively) is illustrated in the examples (section 4).

The maximum probability analysis has to cope with the presence of multiple minima in the cost function: for large problems a full search of the state space is out of the question. The technique used by Dharssi et al (1992) was to set the observation errors to very large values initially and then gradually reduce them during the iterations to the 'true' values. This worked well on the simulated data studies of Dharssi et al (1992), but it may be that the data density was sufficient so that there were no really difficult decisions.

3. Generalisations and Applications

Section 2 presents the basic theory for the case when observations have independent errors - either normal or gross. Section 3.1 considers the extension of IQC and SQC to the case where there are several causes of gross error some of which may affect several observed values; this is compared to alternative approaches based on the quadratic form in section 3.2. Similar extensions to the maximum probability analysis are considered briefly in section 3.3.

Tests on linear functions of the observation increments, tests for errors with characteristic patterns, the possibility of using the theory to correct observations and a current Bayesian quality control system are mentioned in sections 3.4 to 3.7.

3.1 Several sources of gross error - Bayesian approach

What if a subset of observations $\mathbf{y}_o^S$ (coming from the same instrument or subject to the same processing) has a common source of error G_S in addition to the independent sources of gross error, G_i, in each observation i? Assuming that the events G_i and G_S are independent we replace (18) by

$$p(\mathbf{y}_o|\mathbf{y}) = P(\bar{G}_S)\prod_{i=1}^{n} \{p(y_{oi}|y_i \cap \bar{G}_i \cap \bar{G}_S)P(\bar{G}_i) + k_i P(G_i)\}$$

$$+ P(G_S)p(\mathbf{y}_o^S|\mathbf{y}^S \cap G_S) \prod_{i \notin S} \{p(y_{oi}|y_i \cap \bar{G}_i)P(\bar{G}_i) + k_i P(G_i)\} \qquad (33)$$

The second product is over the observations not in $\mathbf{y}_o^S$. Terms not affected by any error are normal as before. We have assumed that $p(\mathbf{y}_o^S|\mathbf{y}^S \cap G_S)$ is independent of the G_i, it will depend upon the type of error, for the present we assume

$$p(\mathbf{y}_o^S|\mathbf{y}^S \cap G_S) = \prod_{i \in S} k_i \qquad (34)$$

If the individual or simultaneous quality control methods are being applied (Sections 2.3 and 2.4) then if (34) holds the extension to an arbitrary number of causes of gross error each affecting different subsets of the observed data is straightforward. We redefine C_α as having the same 'good' and 'bad' observations as before, but where the gross errors can come from any cause. The prior probability $P(C_\alpha)$ can be estimated and used in (27), but it is no longer equal to $\prod_{i=1}^{k} P(\bar{G}_i) \prod_{j=k+1}^{n} P(G_j)$.

Position errors

As an example consider that G_S is the event that an observation has been reported with the wrong position, as can happen for ship-borne observing systems. Let 1 denote the correct position and 2 the reported position, we assume that these have the same climate, but that they are sufficiently separated to be independent. The calculated observation increment has covariance

$$\langle(\mathbf{y}_o^1-\mathbf{y}_b^2)(\mathbf{y}_o^1-\mathbf{y}_b^2)^T\rangle = \langle(\mathbf{y}^2-\mathbf{y}_b^2)(\mathbf{y}^2-\mathbf{y}_b^2)^T\rangle + \langle(\mathbf{y}^1-\mathbf{y}^2)(\mathbf{y}^1-\mathbf{y}^2)^T\rangle + \langle(\mathbf{y}^1-\mathbf{y}_o^1)(\mathbf{y}^1-\mathbf{y}_o^1)^T\rangle$$

$$= \mathbf{B} + 2\mathbf{C} + \mathbf{O} \qquad (35)$$

where <.> denotes an ensemble average and the observation errors and the background errors are assumed to be uncorrelated with each other and with the true values **y**. $\langle(\mathbf{y}^1-\mathbf{y}^2)(\mathbf{y}^1-\mathbf{y}^2)^T\rangle = 2\mathbf{C}$ where **C** is the covariance matrix of the distribution (climate) from which $\mathbf{y}^1$ and $\mathbf{y}^2$ were taken. 2**C** will usually be much larger than **B** and **O**. If **y** is normally distributed then

$$p(\mathbf{y}_o^S|\mathbf{y}^S\cap G_S) = n(\mathbf{y}_o^S|\ \mathbf{y}_b^S,\ \mathbf{B} + 2\mathbf{C} + \mathbf{O}) \qquad (36)$$

If we consider a surface report from a ship then **y** consists of pressure, temperature and wind. The temperature and pressure will have some correlation through the hydrostatic equation, but under the geostrophic assumption the wind is uncorrelated with the both of them, and thus **C** is almost diagonal.

However if a radiosonde profile from a ship is assigned the wrong position (as occasionally happens) then the $\mathbf{y}_o^1-\mathbf{y}_b^2$ differences will vary reasonably smoothly in the vertical, because there is substantial correlation in **C** between adjacent levels. Unfortunately a significant error in development in the background forecast will have much the same signature. If one level contains an individual gross error in addition then the generally large, but smooth differences will contain a spike (an example where $p(\mathbf{y}_o|\bar{G}_i\cap G_S) \neq p(\mathbf{y}_o|G_i\cap G_S)$).

Equation (36) is specific to a particular type of error, a malfunction of one of the radiosonde sensors would have a different signature. Equation (34) implies a more catch-all test for any differences from the assumed normal distribution. It can be considered an approximation to the diagonal elements of **C** (and is the form used by Lorenc and Hammon (1988)). It assumes that the different elements of $\mathbf{y}_o^1-\mathbf{y}_b^2$ are independent of each other.

3.2 The quadratic form

This section considers the special case where a set of observations has a single cause of gross error which corrupts all the observations. An approximate example is a satellite sounding in which an error in one radiance channel or the cloud clearing algorithm can corrupt the retrieved temperatures at almost all levels. In this context Barwell and Young (1991) calculate the quadratic form

$$r^2 = (\mathbf{y}_b-\mathbf{y}_o)^T(\mathbf{B}+\mathbf{O})^{-1}(\mathbf{y}_b-\mathbf{y}_o) \qquad (37)$$

which is a measure of the distance between $\mathbf{y}_b$ and $\mathbf{y}_o$.

If r exceeds a critical value then the whole sounding is rejected. Their derivation is based on the multivariate normal distribution, and $-r^2/2$ is the exponent in (16). If $(\mathbf{y}_b-\mathbf{y}_o)$ is unbiased and normally distributed with covariance matrix $\mathbf{B}+\mathbf{O}$ then r^2 is distributed according to the χ^2 distribution with n degrees of freedom (because of the normalisation by $(\mathbf{B}+\mathbf{O})^{-1}$ r^2 can be rewritten as the sum of squares of n independent random variables each distributed as N(0,1)). Note that even if **B** is almost singular the addition of **O** makes it nonsingular.

Using the Bayesian approach of the last section (and setting $P(G_i)=0$, for all i) we only need consider C_0 (all observations corrupted) and C_{2^n-1} (no observations corrupted).

$$p(\mathbf{y}_o|C_0)P(C_0) = P(G_S)p(\mathbf{y}_o|G_S) \qquad (38)$$

$$p(\mathbf{y}_o|C_{2^n-1})P(C_{2^n-1}) =$$
$$P(\bar{G}_S)\ \{(2\pi)^n|\mathbf{B}+\mathbf{O}|\}^{-0.5}\ \exp[-0.5(\mathbf{y}_b-\mathbf{y}_o)^T(\mathbf{B}+\mathbf{O})^{-1}(\mathbf{y}_b-\mathbf{y}_o)] \qquad (39)$$

In practice this is very similar to the test used by Barwell and Young in that both reject large values of $(\mathbf{y}_b-\mathbf{y}_o)^T(\mathbf{B}+\mathbf{O})^{-1}(\mathbf{y}_b-\mathbf{y}_o)$, the difference lies in the calculation of the rejection limit.

The r^2 statistic has also been proposed as a preliminary step in multi-observation tests by Purnell (1990). It would be used to test the hypothesis that the observations come from a multivariate normal distribution, and only if the hypothesis was rejected would further tests be carried out. The choice of significance level used is somewhat arbitrary, and the method cannot take into account variations in the prior probability of gross error in different observations. The test statistic is related to $P(C_{2^n-1})$, the probability that no observations are corrupted. As with any method which does not involve calculation of $P(C_\alpha)$ for all α this test could give incorrect results: large r^2 could still correspond to C_{2^n-1} (although unlikely) having larger probability than any other combination and the test could also fail to detect some cases where C_{2^n-1} is not the most likely combination. The Bayesian approach presented here has similarities but is more general.

3.3 Maximum probability analysis with several sources of gross error

We have looked briefly at the case where the observations have correlated normal errors (14), and at the case where the observations are independent but contain gross errors (18). The maximum probability analysis would in theory extend to observations with both gross errors and correlated normal errors. However the evaluation of the gradient of the observation penalty function would have to be split into separate cases. For a correlated set S of m observations terms for all 2^m subsets of S would have to be calculated. Bearing in mind that this is an iterative analysis method with these terms being evaluated each iteration this would be quite expensive. One compromise would be to have a preliminary quality control step to deal with multi-level data (the main source of correlated observation errors) before a variational analysis/VQC to detect errors in single-level data.

If the observations have dependent gross errors but uncorrelated normal errors then there is less extra work involved. In the case given by (33) there is one extra term in the penalty function (corresponding to $P(G_S)$). More generally for each subset of observations with a common source of error then there is one extra term in the penalty function provided that the subsets do not overlap.

3.4 Linear functions and bias checks

In testing for bias in a set of measurements, say a radiosonde temperature profile, the obvious test statistic to use is the mean of the observation minus background values.

$$\bar{d} = \mathbf{W}(\mathbf{y}_o-\mathbf{y}_b) \qquad (40)$$

where $\mathbf{W}$ is a 1×n weights matrix with elements w_i. The weights w_i should be chosen on a physical basis, for radiosonde temperatures which are usually reported at irregular intervals it may be appropriate to weight by the time interval that each measurement represents. If $\mathbf{y}_o-\mathbf{y}_b$ is normally distributed as in (16) then $\bar{d}$ is normally distributed with variance $\mathbf{W(B+O)W}^T$ by (5). A standard significance test can be used, or if prior information on the distribution of radiosonde temperature biases is available then a Bayesian test can be performed. In practice it would be necessary to omit elements of $\mathbf{y}_o$ that appear to contain gross errors before calculating and testing $\bar{d}$.

In a similar way any linear function of $\mathbf{y}_o-\mathbf{y}_b$ could be tested, including the tropospheric stability index developed by Kelly et al (1991) for checking satellite soundings.

3.5 Non-uniform distribution of gross errors (special tests)

Some observations are known to have particular preferred errors eg

a) satellite cloud track winds tend to have too low wind speeds in jet regions

b) some aircraft report a wind speed of zero instead of missing data

c) low level profiler winds are close to zero if corrupted by 'ground clutter'.

To take account of these situations the pdf for gross errors $p(y_{oj}|y_j\cap G_j)$ (usually = k_j) can be increased to take account of the increased frequency of errors. In theory we have to replace $k_jP(G_j)$ in (23) with $p(y_{oj}|y_j\cap G_j)P(G_j)$ and recalculate the convolution with $p(\mathbf{y})$. But if $p(y_{oj}|y_j\cap G_j)$ is independent of $\mathbf{y}$ (as we can usually assume) then it merely replaces k_j in (27).

3.6 Correcting observations

Other possible events such as 'the observation has a +10 mb error', denoted by E_{+10} can be added to the basic formalism. A single term from (18) would then become

$$p(y_o|y\cap\bar{G})(1-P(G)-P(E_{+10})) + p(y_o|y\cap E_{+10})P(E_{+10}) + kP(G) \qquad (41)$$

where $p(y_o|y\cap E_{+10})$ is the same normal distribution as $p(y_o|y\cap\bar{G})$ but shifted by 10. $P(E_{+10})$ has to be estimated from prior knowledge, $P(E_{+10}|y_o)$ can then be calculated, and if larger than the alternative probabilities the observation can be corrected with some confidence. Of course other events such as E_{-10},

E_{+20} and E_{-20} can be considered at the same time.

It should be stressed that any corrections should be based on a careful study of observation characteristics such as that by Collins and Gandin (1990). Errors in a single digit, such as 10 mb errors are associated with manual coding or transmission in character format, and it would not be appropriate to use such tests on data from an automatic weather station. Digit (or sign) tests need to be based on the quantities reported eg wind speed in knots or m/s rather than wind components.

There is also possible application to the dealiasing of scatterometer winds. Scatterometer aliases could be put in as separate observations a1 and a2 with $P(\bar{G}_{a1} \cap \bar{G}_{a2}) = 0$.

3.7 Implementation at the Meteorological Office

Methods based on the Bayesian quality control theory of Lorenc and Hammon (1988) have been used for operational quality control of surface marine observations since March 1988. In June 1991 the Bayesian quality control was extended to all observation types (Ingleby and Parrett, 1991). The buddy check is the pairwise method of Lorenc and Hammon (recast in symmetric form), modified by an *ad hoc* 'damping' method (see Appendix C). Only surface, aircraft, radiosonde and satellite cloud track wind observations are buddy checked, and only against other observations in the same category. A check for corruption of the whole observation (typically position error) is applied to single level data following section 5(b) of Lorenc and Hammon. Preliminary work has been done to extend this to multi-level data using the theory of section 3.1 of this paper.

VQC has been tested on aircraft data (approximately 10000 observations) with encouraging results. The implementation used a descent algorithm to find a maximum of the pdf; it was therefore an approximation to the ideal, as it was not guaranteed to find the global maximum. The operational Meteorological Office and ECMWF schemes can be regarded as approximations to IQC and SQC respectively. Because of the millions of data per day used in numerical weather prediction, it is not practical to implement any of the schemes in its theoretically ideal form. This paper clarifies the relationship between the different methods and the approximations made. In the development of operational quality control systems further work on simple test cases to elucidate the properties of the different methods will be a necessary supplement to large scale realistic tests.

4. Discussion and Examples

4.1 Summary of different methods

From section 2 we have three possible quality control methods which can be summarised as follows:

IQC. Individual quality control. All observations are compared with each other, but the decisions are taken independently. All 2^n combinations of possible gross errors should be considered. (In practice a sequential approximation can be used.)

For each observation $P(G_i|\mathbf{y}_o)$ is evaluated using (30) and (27).

If $P(G_i|\mathbf{y}_o)$ is over 0.5 the observation is rejected.

SQC. Simultaneous quality control. Choose the most likely from 2^n subsets of observations. (In practice only the most likely combinations of gross errors are examined.)

For each subset $p(\mathbf{y}_o|C_\alpha)P(C_\alpha)$ is evaluated using (27).

VQC. Compare with a variational (maximum probability) analysis which incorporates the possibility of gross errors in observations.

The analysis is given by the maximum of (31) and $P_y(G_i)$ by (32).

If $P_y(G_i)$ is over 0.5 the observation is rejected.

Notes

1) IQC, SQC and VQC are all derived from the same background and observation pdfs (eg (13) and (18)), they differ in the optimality principle being used.

2) IQC and SQC are calculating probabilities integrated over all possible analysis states, they can be thought of as multidimensional background checks.

3) The optimality principles do not depend on the probability distributions, but normal distributions have many special properties which reduce the computational expense, particularly for the integrations in IQC and SQC.

4) The most fundamental problem is that we are trying to reduce a complicated, often multimodal, pdf into a single vector - the 'analysis'.

5) For normal distributions the analysis based method VQC uses the separate covariance matrices **B** and **O**, whereas the integrated methods IQC and SQC only use **B+O**.

6) None of the methods require a preliminary (background) check considering observations individually, although it might be useful, particular in SQC to guide the search for the most likely combination of observations.

7) IQC and VQC provide (different measures of) probabilities of gross error in individual observations whereas SQC just gives a pass/fail marking for each observation.

4.2 Examples

Much of the time one of the normal distributions has much larger probability than any of the others, and IQC, SQC and VQC will tend to make the same decisions. It is only in rather borderline cases that they disagree. For illustration we use cases involving two or three collocated observations, so that the pdfs are one dimensional and can be displayed as line graphs. The cases were chosen for disagreement between IQC and SQC and are given in figure 2 and table 1.

The dotted curves in figure 2 are the pdfs for each individual combination. They are each normal and are displaced towards zero relative to the observations as they also include background information (as for the solid lines in figure 1). The background error pdf, corresponding to all observations having gross errors, is centred on zero. The total pdf (the sum of all the individual contributions) is shown as a solid line. VAN picks the highest point of the total pdf (the mode) and VQC uses this to check the observations. SQC chooses the individual distribution with the largest area; its mean does not necessarily correspond with the mode, since it is based on the integral of the pdf whereas the height of the peak is a local quantity. The mean analysis can fall in the trough in pdf between the background and the observations, eg in figure 2a it is -3.2, making it an unsatisfactory choice for most purposes.

Table 1 gives the posterior probabilities of gross error and resulting analyses given by the different methods, for comparison the background check (only using observation i in the calculation) and the mean analysis are included. For the background check (BKC) the probability of gross error (PGE) is $P(G_i|y_{oi})$, for IQC $P(G_i|\mathbf{y}_o)$ is given, for VQC $P(G_i|\mathbf{y}=\mathbf{y}_{max})$. The observation is taken to have passed the quality control for values less than 0.5. For SQC a pass/fail indicator is given. For these quality control methods the resulting linear analysis is given in the analysis column, for VAN (MAN) the maximum (mean) probability analysis obtained directly is given. Case a) will be examined in some detail to explain the calculations. $p(\mathbf{y}_o|C_\alpha)P(C_\alpha)$ for different combinations of gross errors is given in table 2, along with the marginal sums. SQC chooses the largest individual value, which is 3.11E-6 for $\bar{G}_1\cap\bar{G}_2$. IQC calculates $P(G_1|\mathbf{y}_o)$ as the partial sum for G_1 over the total sum = 4.40/7.53 = 0.58. Similarly $P(G_2|\mathbf{y}_o)$ = 2.98/7.53 = 0.40. These are the values in the IQC line of table 1a.

A more graphical representation of IQC is given in figure 3. The solid line is $p(\mathbf{y}|\mathbf{y}_o)$ (= the total pdf from figure 2a normalised so that its area is 1), the dashed lines are $P(G_i|\mathbf{y})$ the probability of gross error in observation i given that $\mathbf{y}$ is the 'truth' (equation (32)). VQC simply reads off the values of $P(G_i|\mathbf{y})$ corresponding to $\mathbf{y}_{max}$. IQC calculates the integral of the product of $P(G_i|\mathbf{y})$ and $p(\mathbf{y}|\mathbf{y}_o)$ over the whole range of values of $\mathbf{y}$.

The difference between performing a maximum probability analysis directly and using it to quality control observations for a linear analysis (VAN and VQC) can be illustrated using the examples. VAN chooses $\mathbf{y}_{max}$ at the maximum of the solid line, whereas VQC chooses the peak of the individual pdf which is (in some sense) closest to $\mathbf{y}_{max}$. The cases here suggest that the differences between the resulting analyses are minimal. Any choice between VAN and VQC would be largely on ease of practical implementation.

In these simple examples all the terms are being evaluated, and a complete line search is made (with a grid interval of 0.1). In general search algorithms would be used, these are not guaranteed to find the global maximum. The total pdf can be multimodal (figure 2) causing difficulties for VQC. In table 2 the two largest values of $p(\mathbf{y}_o|C_\alpha)P(C_\alpha)$ are not adjacent, illustrating that a simplex-like algorithm for SQC could get stuck in a secondary maximum when two observations should either be accepted or rejected together.

	⟵	PGE	⟶	Analysis
a) $\mathbf{y}_o-\mathbf{y}_b=$	(-8 ,	-6)^T		y_a-y_b
BKC	0.99	0.67		0.0
IQC	0.58	0.40		-4.2
SQC	pass	pass		-5.7
VQC	0.06	0.00		-5.7
VAN				-5.7
MAN				-3.2
b) $\mathbf{y}_o-\mathbf{y}_b=$	(-9 ,	-9 ,	-6)^T	y_a-y_b
BKC	1.00	1.00	0.67	0.0
IQC	0.63	0.63	0.40	-4.2
SQC	fail	fail	fail	0.0
VQC	0.04	0.04	0.01	-7.0
VAN				-6.9
MAN				-3.6
c) $\mathbf{y}_o-\mathbf{y}_b=$	(-4 ,	4)^T		y_a-y_b
BKC	0.09	0.09		0.0
IQC	0.52	0.52		0.0
SQC	fail	pass		-2.8
VQC	0.01	1.00		-2.8
VAN				-2.8
MAN				0.0
d) $\mathbf{y}_o-\mathbf{y}_b=$	(-3 ,	3)^T		y_a-y_b
BKC	0.03	0.03		0.0
IQC	0.49	0.49		0.0
SQC	fail	pass		2.1
VQC	0.01	1.00		-2.1
VAN				-2.1
MAN				0.0

Table 1. Comparison of different quality control and analysis methods. The examples are based on collocated surface pressure observations with σ_o = 1.0 mb, σ_b = 1.5 mb, k = 0.043 mb^{-1} and P(G) = 0.04. See text for further description.

	G_2	$\bar{G}_2$	
G_1	2.96	1.44	4.40
$\bar{G}_1$	0.02	3.11	3.13
	2.98	4.55	7.53

Table 2. $p(\mathbf{y}_o|C_\alpha)P(C_\alpha)$ $(\times 10^6)$ of different combinations of gross errors for case a).

4.3 Choice of methods

To take a subjective view of the examples in table 1. In a) and b) the observations seem fairly consistent, so in a) SQC and VQC are better than IQC, in b) VQC is better than IQC which is better than SQC. c) and d) are both (somewhat contrived) symmetrical situations: IQC gives the same probability to both observations - a marginal fail in c), a marginal pass in d) - whereas SQC and VQC consider the two observations inconsistent so have to pick one of them (or not return an answer). This illustrates that the three methods are providing different types of consistency or optimality.

IQC can be characterised as 'even handed', giving 'compromise' solutions. SQC and particularly VQC are more 'decisive'. IQC and SQC tend to favour broad distributions whereas VQC chooses higher (often narrower) distributions. As the distributions get narrower the more observations are involved this suggests that VQC tends to 'draw to' the observations more than the other methods. The simple examples given here support this view, and would tend to favour VQC. However the dilemma becomes more acute for observations which are usually much more accurate than the background, but subject to gross errors. For example in figure 4 should we choose the very narrow peak and accept the observation (VQC) or the broader peak which has larger area and reject the observation (the 'safer' option followed by IQC and SQC). Correlated observation errors are dealt with more naturally by the IQC and SQC methods than by VQC. The question of which method is 'best' is discussed further in the next section.

With several different multi-observation checks available the options include using one method, using two or three methods and combining the results somehow, or using a cheap method (the pairwise check?) in an initial scan and then referring difficult cases to a better (less approximate) check. It seems likely that any of these methods would be applied to subsets of observations at a time - implicitly assuming that quality control and analysis will continue to be treated as two separate processes, for the time being at least.

An important practical aspect is that our knowledge of the observation and background error distributions is imperfect. The decisions made depend on these distributions, and ideally we would like a method that is relatively insensitive to the parameters that we are least sure of. Figures 5 and 6 both show the total distribution from figure 2a (solid line) and the effect of perturbations to the case. In figure 5 the observation at -8 is moved to -8.5 (dotted line) and then to -9 (dashed line), the size of the hump corresponding to both observations being correct decreases very rapidly, and in both cases both observations are rejected by all three qc methods.

In figure 6 the background error variance is modified: to $1.61 \approx 1.27^2$ (dotted line), to $2.89 =1.7^2$ (dashed line), the average of these two distributions is also used (dash-dotted line) this has the same variance $2.25 = 1.5^2$ as used in the solid line. When the background error distribution is narrowed (dotted line) the observations are rejected, whereas the observations easily pass the qc when the wide or composite pdfs are used. At any particular point the estimate of background error standard deviation could easily be in error by 0.2 mb, so this sensitivity is uncomfortable.

4.4 Wrong decisions

Any quality control system is subject to errors of two types: Type 1 is rejecting 'good' observations and Type 2 is accepting 'bad' observations. What levels of Type 1 and 2 errors are acceptable? The answer depends on at least two factors: how bad a particular 'bad' observation is, and how sensitive to initial conditions the atmospheric/model flow is in the region under consideration. The decisions are most important in areas sensitive to initial conditions, unfortunately model errors will on average be larger in these areas making the decisions more difficult.

Is it more important to avoid one type of error than the other? This question is not usually addressed explicitly, the implicit assumption seems to be that both types of error should be given about the same weight. If avoiding 'bad' observations was more important then we would tighten the tolerances allowed and accept the rejection of more 'good' data. With IQC or maximum probability analysis QC it would be possible to adjust the accept/reject threshold (usually 0.5) for $P(G_i|\mathbf{y}_o)$ or $P(G_i|v_i)$ to give different weights to Type 1 and Type 2 errors. Modifying SQC would be more difficult.

If we specify a cost function for analysis errors we can, at least in theory, take account of the two factors mentioned above. Let us denote the cost function by $f(\mathbf{y}_a,\mathbf{y})$, in the linear case it is a function of $\mathbf{y}_a-\mathbf{y}$. Usually f will increase as the difference between $\mathbf{y}_a$ and $\mathbf{y}$ increases (this measures the 'badness' of the analysis, rather than the observations). If we wish to take account of variations in perturbation sensitivity then a non-linear cost function has to be used.

If we know $p(\mathbf{y}|\mathbf{y}_o)$ and $f(\mathbf{y}_a,\mathbf{y})$ then the 'optimum' analysis is given by the minimum of their convolution i.e. the $\mathbf{y}_a$ that minimises $\int p(\mathbf{y}|\mathbf{y}_o)f(\mathbf{y}_a,\mathbf{y})d\mathbf{y}$. Very narrow cost functions have a delta function as a limit - i.e. 'anything other than a perfect analysis is useless'. This will give the maximum probability methods VQC and VAN. At the other extreme as f becomes very wide the optimum analysis will tend towards the mean of $p(\mathbf{y}|\mathbf{y}_o)$, this is $\sum_\alpha \mathbf{y}_a(C_\alpha)P(C_\alpha|\mathbf{y}_o)$ where $\mathbf{y}_a(C_\alpha)$ is the linear analysis using combination C_α and $P(C_\alpha|\mathbf{y}_o)$ is given by (28). This is related to SQC and IQC, but is **not** the linear analysis from either of them. In practice the cost function lies somewhere between these extremes.

The cost function and hence the 'best' quality control/analysis does depend on the use of the analysis; eg an analysis designed specifically to monitor frost/no-frost at a location has a step function cost. Hence the 'best' method of quality control for numerical weather prediction is still an open question.

4.5 Robustness

Gutowski and Hoffman (1985) suggested that it is desirable for the analysis $\mathbf{y}_a$ to be a continuous function of (insensitive to small changes in) the observations $\mathbf{y}_o$. This is related to robust estimation in the statistical literature. Robustness implies using all observations in the analysis, but to take account of 'bad' observations some sort of quality factor is included in the weightings. The robustness of quantities used in this paper are as follows.

a) the PGEs from IQC are a continuous function of $\mathbf{y}_o$

b) the PGEs from VQC are a piecewise continuous function of $\mathbf{y}_o$ containing jumps from one maximum to another

c) any rejection method (IQC, SQC or VQC) will introduce discontinuities

d) VAN is a piecewise continuous analysis

e) the mean analysis $\int \mathbf{y}\ p(\mathbf{y}|\mathbf{y}_o)\ d\mathbf{y}$ is fully continuous

If required the PGEs from VQC, equation (32), could be used to reconstruct VAN (a robust analysis) just by dividing the observation error variances by $P_y(\bar{G}_i)$ (Dharssi et al, 1992, equations 15-17).

We have found that the robust quantities such as the PGEs from IQC, while continuous, have fairly flat regions with quite sharp gradients in between - reducing the distinction between robust and non-robust methods. This sharp transition is related to the assumption that observation and background errors are normally distributed - longer tailed distributions give smoother transitions/more robustness (see example in Tarantola (1987) problem 1.9).

4.6 Use of Bayesian methods

The Bayesian approach explicitly depends on prior estimates of the observation distributions for all possible distributions (for 'good' observations and gross errors). In the meteorological context there is ample opportunity to build up such estimates (or models) from months or years of previous observations. If we have reasonable models of the distribution of gross errors then the Bayesian approach is much more powerful than the significance testing approaches (section 3.2). The significance tests do not use any information about the distribution of gross errors - they just test for a certain degree of consistency with the normal distribution. A statistical viewpoint is provided by Barnett and Lewis (1978), however they are mainly considering single sets of data without reliable prior information.

4.7 Normal distributions, linear analyses and statistical interpolation

If the observation and background errors are normally distributed then the posterior distribution of the truth is also normally distributed with mean (and mode) given by the statistical interpolation analysis, the covariance matrix is also given by the statistical interpolation equations (see (17) and also Appendix A). This is because least-squares/minimum variance/linear regression methods (including statistical interpolation) are implicitly based on the normal distribution. If the errors are not normal then these methods are sub-optimal (eg Tarantola (1987) section 1.7.2 and problem 1.11, also chapter 4 for least-squares methods in general).

The analysis equation $p(\mathbf{y}|\mathbf{y}_o) = p(\mathbf{y}\cap\mathbf{y}_o)/p(\mathbf{y}) = p(\mathbf{y}\cap\mathbf{y}_o)/\int p(\mathbf{y}\cap\mathbf{y}_o)d\mathbf{y}$ (see (16) and (17)) is quite general requiring only that $p(\mathbf{y}\cap\mathbf{y}_o)$ can be normalised. If the errors are normal then the location of the distribution (and hence the analysis) turns out to be a linear function of $\mathbf{y}_b$ and $\mathbf{y}_o$. However for non-normal errors the analysis derived from $p(\mathbf{y}|\mathbf{y}_o)$ will be non-linear in general.

If the background errors are not normally distributed then the integration over all analysis states may not be possible analytically, and the versions of IQC and SQC as presented here are not applicable. However if the background errors were represented by the sum of two normal distributions then $P(G_i|\mathbf{y}_o)$ can be calculated but the denominator of (29) would then contain 3^n terms. Alternatively $\mathbf{y}_o-\mathbf{y}_b$ could be transformed so that it was more nearly normally distributed, however this would distort any physical relationships (such as geostrophy or non-divergence) in the covariances. The maximum probability analysis formulation depends only on the distribution being differentiable, although the convergence would be affected by non-normal background errors (the cost function becomes less quadratic).

5. Summary

The Bayesian probability methods of observation quality control introduced by Lorenc and Hammon (1988) have been generalised. From the error distributions of observations and background (including non-normal observation errors) the posterior probability density function of the 'true state' can be found which encapsulates our knowledge. This can be used to determine which observations probably have 'gross' errors and/or the 'optimal' analysis. Several different optimality criteria are possible, we examine three in particular. They can be related to the Meteorological Office's operational Bayesian quality control scheme, the statistical interpolation qc scheme used at ECMWF and a non-linear variational scheme. They are illustrated with simple examples in section 4.

Much of the time the three methods perform similarly, mainly differing on the difficult cases where they show different levels of 'decisiveness' or 'compromise'. Which is best depends on the cost function of analysis errors. All the methods depend on our knowledge of the observation and background error distributions. It is important to improve this knowledge, and also to consider the sensitivity of qc methods to the various parameters.

Acknowledgement

We would like to thank D Carrington for useful comments on the manuscript.

References

Anderson, T.W. 1984 Introduction to multivariate statistical analysis. Second edition. *Wiley*

Barnett, V. and Lewis, T. 1978 Outliers in statistical data. *Wiley*

Barwell, B.R. and Young, J.C.L. 1991 Quality control of satellite temperature Soundings. Short-range forecasting research technical note no. 63. To appear in *The ISPRS J. of Photogrammetry and Remote Sensing*

Chatfield, C. and Collins, A.J. 1980 Introduction to multivariate analysis. *Chapman and Hall*

Collins, W.G. and Gandin, L.S. 1990 Comprehensive hydrostatic quality control at the National Meteorological Center. *Mon.Wea.Rev.*, **118,** 2752-2767

Dharssi, I. Lorenc, A.C. and Ingleby, N.B. 1992 Treatment of gross errors using maximum probability theory. to appear in *Q.J.R.Meteorol.Soc.*

Fischler, M. and Bolles, R 1981 Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Graphics and Image Processing*, **24,** 381-395

Gutowski, W.J. and Hoffman, R.N. 1985 Robust data quality control in a statistical interpolation analysis scheme. Pp. 428-433 in *Preprints, 9th Conf. on*

Probability and Statistics in Atmospheric Sciences. AMS

Ingleby,N.B. and Parrett,C.A. 1991 Quality control of atmospheric data. *Unified Model Documentation Paper 32.* (Available from Meteorological Office)

Kelly,G. Andersson,E. Hollingsworth,A. Lönnberg,P. Pailleux,J and Zhang,Z. 1991 Quality control of operational physical retrievals of satellite sounding data. *Mon.Wea.Rev.*,**119,**1866-1880

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. *Q.J.R.Meteorol.Soc.*,**112,**1177-1194

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

Purnell,D.K. 1990 A hierarchical test for corrupt data. WMO Int. Symp. on Assimilation of Observations in Meteorology and Oceanography.626-627

Purser,R.J. 1984 A new approach to the optimal assimilation of meteorological data by iterative Bayesian analysis. *AMS 10th Conf. on Weather Forecasting and Analysis.*102-105

Tarantola,A. 1987 Inverse Problem Theory. Methods for Data Fitting and Model Parameter Estimation. Elsevier, Amsterdam

Thiébaux,H.J. and Pedder,M.A. 1987 Spatial objective analysis: with applications in atmospheric science. Academic Press, London

Appendix A. General analysis equation

Here we extend the theory given in section 2.2 to the general case where to obtain an estimate of the observed quantities from a forecast requires a generalised interpolation K (see Lorenc, 1986). K may be non-linear, it will be linearised about the forecast state.

This appendix uses the same notation as the main paper with the following additions and modifications:

$\mathbf{x}$ true state of the atmosphere projected onto model representation

$\mathbf{x}_b$ prior estimate of $\mathbf{x}$ (e.g. from forecast) referred to as the background

$\mathbf{y}_o$ observations

$\mathbf{y}$ observations that would be given by error-free instruments

$K(\mathbf{x})$ forward operator for calculating $\mathbf{y}$ from $\mathbf{x}$

$\mathbf{K}$ tangent linear operator of K, such that $K(\mathbf{x}+\delta\mathbf{x})=K(\mathbf{x})+\mathbf{K}\delta\mathbf{x}+O(\delta\mathbf{x}^2)$.

We assume that

$$p(\mathbf{x}) = \mathrm{n}(\mathbf{x}|\mathbf{x}_b,\mathbf{B}) \tag{A1}$$

$$p(\mathbf{y}_o|\mathbf{y}) = \mathrm{n}(\mathbf{y}_o|\mathbf{y},\mathbf{O}) \tag{A2}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathrm{n}(\mathbf{y}|K(\mathbf{x}),\mathbf{F}) \tag{A3}$$

(cf Lorenc (1986) equations 14-16). Equation (A3) describes the error of K. If $K(\mathbf{x})=\mathbf{Kx}$ then $\mathbf{y}_b$ and $\mathbf{B}$ as used in the main paper correspond to $\mathbf{Kx}_b$ and $\mathbf{KBK}^T$ here. We take the convolution of (A2) and (A3)

$$\begin{aligned} p(\mathbf{y}_o|\mathbf{x}) &= \int p(\mathbf{y}_o\cap\mathbf{y}|\mathbf{x})d\mathbf{y} = \int p(\mathbf{y}_o|\mathbf{y}\cap\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{y} \\ &= \mathrm{n}(\mathbf{y}_o|K(\mathbf{x}),\mathbf{O}+\mathbf{F}) \end{aligned} \tag{A4}$$

assuming that $p(\mathbf{y}_o|\mathbf{y}\cap\mathbf{x})=p(\mathbf{y}_o|\mathbf{y})$ i.e. that knowledge of $\mathbf{x}$ does not add to our information about $\mathbf{y}_o$ if we already know $\mathbf{y}$. The convolution of (A4) and (A1) is

$$\begin{aligned} p(\mathbf{y}_o) &= \int p(\mathbf{x}\cap\mathbf{y}_o)d\mathbf{x} = \int p(\mathbf{y}_o|\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \mathrm{n}(\mathbf{y}_o|K(\mathbf{x}_b),\mathbf{O}+\mathbf{F}+\mathbf{KBK}^T) \end{aligned} \tag{A5}$$

This is the probability density of the observations given the background information (cf (16)). This result holds for linear K, and for nonlinear K providing it can be linearised over the range where the integrands are non-negligible.

We can write the joint distribution of $\mathbf{x}$ and $\mathbf{y}_o$ as

$$p(\mathbf{x}\cap\mathbf{y}_o) = \mathrm{n}(\mathbf{z}|\mathbf{z}_b,\mathbf{C}),$$

$$\text{where } \mathbf{z} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}_o\end{pmatrix}, \quad \mathbf{z}_b=\begin{pmatrix}\mathbf{x}_b\\ K(\mathbf{x}_b)\end{pmatrix} \text{ and } \mathbf{C} = \begin{pmatrix}\mathbf{B} & \mathbf{BK}^T\\ \mathbf{KB} & \mathbf{O+F+KBK}^T\end{pmatrix} \tag{A6}$$

The covariance matrix $\mathbf{C}$ is given by (A1) and (A5), and the ensemble average $<(\mathbf{x}-\mathbf{x}_b)(\mathbf{y}_o-K(\mathbf{x}_b))^T> = <(\mathbf{x}-\mathbf{x}_b)(\mathbf{y}_o-K(\mathbf{x}))^T> + <(\mathbf{x}-\mathbf{x}_b)(K(\mathbf{x})-K(\mathbf{x}_b))^T> = \mathbf{BK}^T$, since $<(\mathbf{x}-\mathbf{x}_b)(\mathbf{y}_o-K(\mathbf{x}))^T> = 0$ by the assumed independence of observation and background errors. Using standard results for conditional distributions (equations (9) to (11))

$$p(\mathbf{x}\cap\mathbf{y}_o) = \mathrm{n}(\mathbf{y}_o|K(\mathbf{x}_b),\mathbf{O+F+KBK}^T)\ \mathrm{n}(\mathbf{x}|\mathbf{x}_a,\mathbf{A}) \tag{A7}$$

where $\mathbf{x}_a$ and $\mathbf{A}$ are defined by

$$\mathbf{A} = \mathbf{B} - \mathbf{BK}^T(\mathbf{KBK}^T+\mathbf{O}+\mathbf{F})^{-1}\mathbf{KB} \tag{A8}$$

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{BK}^T(\mathbf{KBK}^T+\mathbf{O}+\mathbf{F})^{-1}(\mathbf{y}-K(\mathbf{x}_b)) \tag{A9}$$

If $K(\mathbf{x})=\mathbf{Kx}$, then equivalent definitions for $\mathbf{x}_a$ and $\mathbf{A}$ are:

$$\mathbf{A}^{-1} =\mathbf{B}^{-1} +\mathbf{K}^T(\mathbf{O}+\mathbf{F})^{-1}\mathbf{K} \tag{A10}$$

$$\mathbf{A}^{-1}\mathbf{x}_a=\mathbf{B}^{-1}\mathbf{x}_b+\mathbf{K}^T(\mathbf{O}+\mathbf{F})^{-1}\mathbf{y}_o \tag{A11}$$

The conditional probability

$$p(\mathbf{x}|\mathbf{y}_o) = p(\mathbf{x}\cap\mathbf{y}_o)/p(\mathbf{y}_o) = \mathrm{n}(\mathbf{x}|\mathbf{x}_a,\mathbf{A}) \tag{A12}$$

is the posterior probability density, i.e. the probability that the truth lies in a volume $\mathbf{dx}$ around $\mathbf{x}$, given the background $\mathbf{x}_b$ and the observations $\mathbf{y}_o$. The mean and its variance (A9) and (A8) agree with the maximum likelihood/minimum variance estimates derived in Lorenc (1986) (his equations 28 and 29).

In statistical terms $p(\mathbf{y}_o)$ is the marginal density at the point $\mathbf{y}_o$ and $p(\mathbf{x}|\mathbf{y}_o)$ is the conditional density of $\mathbf{x}$ given the value of $\mathbf{y}_o$.

Appendix B. Computational methods

The evaluation of (16) or (27) involves the calculation of $|\mathbf{B+O}|$ and $(\mathbf{y}_b-\mathbf{y}_o)^T(\mathbf{B+O})^{-1}(\mathbf{y}_b-\mathbf{y}_o)$. Because $\mathbf{B+O}$ is a covariance matrix it is symmetric and positive definite and it can be expressed as $\mathbf{B+O} = \mathbf{U}^T\mathbf{U}$, where $\mathbf{U}$ is upper triangular - the Cholesky decomposition. The determinant $|\mathbf{B+O}| = |\mathbf{U}|^2$, and $|\mathbf{U}|$ is just the product of the diagonal elements of $\mathbf{U}$ since it is triangular. Rather than calculate $(\mathbf{B+O})^{-1}$ it is cheaper and more accurate to calculate $\mathbf{x} = (\mathbf{B+O})^{-1}(\mathbf{y}_b-\mathbf{y}_o)$ as the solution of the linear equations $(\mathbf{B+O})\mathbf{x} = (\mathbf{y}_b-\mathbf{y}_o)$ using the Cholesky decomposition (eg the Cholesky decomposition can be performed by NAG routine F01BXF, and the solution of the linear equations by routine F04AZF.)

When calculating probabilities for several subsets of observations some savings can be made by noting that the Cholesky decomposition of **B+O** for observations 1 to m (in that order), contains within it the Cholesky decompositions for observations 1 to m-1, 1 to m-2 etc.

The special case where all off diagonal elements of **B+O** are equal corresponds to collocated observations, each with equal error variance. It is useful for test purposes as special forms of $(\mathbf{B+O})^{-1}$ and $|\mathbf{B+O}|$ are available:

The inverse can be calculated directly, i.e. if

$$\mathbf{B+O} = \sigma^{2n}\begin{pmatrix} 1 & \rho & \rho & \\ \rho & 1 & \rho & \\ \rho & \rho & 1 & \ldots \\ & & \vdots & \end{pmatrix} \quad \text{then } (\mathbf{B+O})^{-1} = \sigma^{-2n}\begin{pmatrix} \alpha & \beta & \beta & \\ \beta & \alpha & \beta & \\ \beta & \beta & \alpha & \ldots \\ & & \vdots & \end{pmatrix}$$

where $\alpha = \dfrac{1 + (n-2)\rho}{1 + (n-2)\rho - (n-1)\rho^2}$ and $\beta = \dfrac{-\rho}{1 + (n-2)\rho - (n-1)\rho^2}$.

To calculate the determinant we use the Cholesky decomposition (see eg Tarantola (1987) for details), in this case

$$\mathbf{U} = \sigma^n\begin{pmatrix} d_1 & u_1 & u_1 & \ldots u_1 \\ 0 & d_2 & u_2 & \ldots u_2 \\ 0 & 0 & d_3 & \ldots u_3 \\ & & \vdots & \end{pmatrix} \quad \text{where } d_i^2 = 1 - \sum_{k=1}^{i-1} u_k^2 \quad \text{and } u_i = \left(\rho - \sum_{k=1}^{i-1} u_k^2\right)\Big/d_i$$

If we set $S_{i-1} = \sum_{k=1}^{i-1} u_k^2$ then we can rewrite this as $d_i^2 = 1 - S_{i-1}$

and $S_i = S_{i-1} + u_i^2 = S_{i-1} + (\rho - S_{i-1})^2/d_i^2$ with $S_0 = 0$.

The determinant of **B+O** is $\sigma^{2n}\prod_{i=1}^{n} d_i^2$.

Appendix C. Pairwise buddy check of Lorenc and Hammon

To evaluate the joint probability of two observations y_{o1} and y_{o2} when neither of them have gross errors Lorenc and Hammon (1988) used

$$
\begin{aligned}
p(\mathbf{y}_o|\bar{G}_1\cap\bar{G}_2) &= p(y_{o1}\cap y_{o2}|\bar{G}_1\cap\bar{G}_2) \\
&= p(y_{o1}|y_{o2}\cap\bar{G}_1\cap\bar{G}_2)\ p(y_{o2}|\bar{G}_2) \qquad \text{(C1)}
\end{aligned}
$$

where

$$p(y_{o1}|y_{o2}\cap\bar{G}_1\cap\bar{G}_2) = n(y_{o1}|y_{a1},\sigma_{a1}^2+\sigma_{o1}^2) \qquad \text{(C2)}$$

and y_{a1} is the statistical interpolation estimate at position 1 using the background and observation 2. σ_{a1}^2 and σ_{o1}^2 are the error variances of y_{a1} and y_{o1} respectively. The resulting expression for $p(\mathbf{y}_o|\bar{G}_1\cap\bar{G}_2)$ was reduced to a symmetric form by B R Barwell (personal communication). It can be written as

$$p(\mathbf{y}_o|\bar{G}_1\cap\bar{G}_2) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{0.5}} \exp\left(\frac{-1}{2(1-\rho^2)}\left(\frac{x_1^2}{\sigma_1^2} - \frac{2\rho x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right)\right) \qquad \text{(C3)}$$

$$\text{where} \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_{o1}-y_{b1} \\ y_{o2}-y_{b2} \end{pmatrix} \quad \text{and} \quad \mathbf{O+B} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

C3 is a bivariate normal distribution with correlation ρ (a special case of (16)). Note that ρ is the total correlation between $y_{o1}-y_{b1}$ and $y_{o2}-y_{b2}$, not the background error correlation. The agreement with the different derivation in section 2 is gratifying, it is a special case of the general relationship demonstrated in Appendix D, equations D1-D3.

We now evaluate the factor $p(\mathbf{y}_o|\bar{G}_1\cap\bar{G}_2)/p(y_{o1}|\bar{G}_1)p(y_{o2}|\bar{G}_2)$, which is central to the pairwise buddy check.

$$
\begin{aligned}
\frac{p(\mathbf{y}_o|\bar{G}_1\cap\bar{G}_2)}{p(y_{o1}|\bar{G}_1)\ p(y_{o2}|\bar{G}_2)} &= \frac{p_b(\mathbf{y}_o|\bar{G}_1\cap\bar{G}_2)}{(2\pi\sigma_1^2)^{-0.5}\exp(-x_1^2/2\sigma_1^2)dx_1\ \ (2\pi\sigma_2^2)^{-0.5}\exp(-x_2^2/2\sigma_2^2)dx_2} \\
&= \frac{2\pi\sigma_1\sigma_2}{2\pi\sigma_1\sigma_2(1-\rho^2)^{0.5}} \exp\left(\frac{-1}{2(1-\rho^2)}\left(\frac{x_1^2}{\sigma_1^2} - \frac{2\rho x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right) + \frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2}\right)
\end{aligned}
$$

$$= \frac{1}{(1-\rho^2)^{0.5}} \exp\left(\frac{-\rho^2}{2(1-\rho^2)} \left(\frac{x_1^2}{\sigma_1^2} - \frac{2x_1x_2}{\rho\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right)\right) \qquad \text{(C4)}$$

This factor is also an objective measure of the 'agreement' of two observations, taking values > 1 if they are consistent with each other, < 1 if the two observations disagree.

For two observations and i=1 (30) becomes

$$P(G_1|\mathbf{y}_o) = \left(p(\mathbf{y}_o|C_0)P(C_0) + p(\mathbf{y}_o|C_2)P(C_2)\right)/$$
$$\left(p(\mathbf{y}_o|C_0)P(C_0) + p(\mathbf{y}_o|C_1)P(C_1) + p(\mathbf{y}_o|C_2)P(C_2) + p(\mathbf{y}_o|C_3)P(C_3)\right) \qquad \text{(C5)}$$

where $C_0=G_1\cap G_2$, $C_1=\bar{G}_1\cap G_2$, $C_2=G_1\cap\bar{G}_2$, $C_3=\bar{G}_1\cap\bar{G}_2$ and $P(C_0)=P(G_1)P(G_2)$ etc

Following Lorenc and Hammon (1988) (equations 36, 38 and Appendix B) (C5) can be rewitten as

$$P(G_1|\mathbf{y}_o) = P(G_1|y_{o1})\left\{p(y_{o1})p(y_{o1})/p(\mathbf{y}_o)\right\} \qquad \text{(C6)}$$

$$p(\mathbf{y}_o)/p(y_{o1})p(y_{o1}) =$$
$$1 - P(\bar{G}_1|y_{o1})P(\bar{G}_2|y_{o2})\left\{1-p(\mathbf{y}_o|\bar{G}_1\cap\bar{G}_2)/\left(p(y_{o1}|\bar{G}_1)p(y_{o2}|\bar{G}_2)\right)\right\} \qquad \text{(C7)}$$

These equations are exact for the two observation case. Lorenc and Hammon extend them to more observations by sequential checking, in pairs, at each step multiplying the current estimate of the probability of gross error in each observation by the reciprocal of (C7). This involves an approximation and it has been found that, particularly in data dense areas, this buddy check becomes over-active - tending to reject some good observations close to background values (and preferentially passing observations further from the background if there are several that agree). To alleviate this behaviour an *ad hoc* 'damping' has been introduced. Two alternatives have been tried:

1) raise $p(\mathbf{y}_O|\bar{G}_1\cap\bar{G}_2)/p(y_{o1}|\bar{G}_1)p(y_{o2}|\bar{G}_2)$ (C4) to a power η_1 less than 1
2) raise $p(\mathbf{y}_o)/p(y_{o1})p(y_{o1})$ (C7) to a power η_2 less than 1

Method 2 damps 'agreement' less and 'disagreement' more than method 1 and it has been adopted as standard with η_2=0.5. It would be better to make η a decreasing function of observation density. However with this modification the pairwise buddy check works quite well, and it is not clear how much improvement can be gained without going to the more sophisticated multi-observation checks described in the main paper.

Appendix D. Adding and removing observations - conditional probabilities

In IQC or SQC (27) would be computed for many slightly different combinations of gross errors/observations accepted. It is desirable to find a method of adding or removing observations, so that the most expensive part, the matrix inverse, does not need to be recalculated from scratch.

It is convenient to consider the normal distribution $p(\mathbf{y}|\mu) = \mathrm{n}(\mathbf{y}|\mu,\Sigma)$ and to partition it as in (6) (after any necessary re-ordering). We are mainly considering cases where $\mathbf{y}_1$ contains only one or two observations, and $\mathbf{y}_2$ is a much longer vector. We will first calculate $p(\mathbf{y}|\mu)$ from $p(\mathbf{y}_2|\mu_2)$ (adding observations). Using (9), (10) and (11)

$$p(\mathbf{y}|\mu) = p(\mathbf{y}_1|\mathbf{y}_2\cap\mu)p(\mathbf{y}_2|\mu) \qquad \text{(D1)}$$

where the conditional density

$$p(\mathbf{y}_1|\mathbf{y}_2\cap\mu) = \mathrm{n}(\mathbf{y}_1|\ \mu_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2-\mu_2),\ \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \qquad \text{(D2)}$$

and the marginal density

$$p(\mathbf{y}_2|\mu) = \mathrm{n}(\mathbf{y}_2|\mu_2,\Sigma_2) = p(\mathbf{y}_2|\mu_2) \qquad \text{(D3)}$$

If we have already computed $p(\mathbf{y}_2|\mu_2)$ then we have Σ_{22}^{-1} (or a Cholesky decomposition of Σ_{22}) available and (D2) can be computed fairly cheaply. The required density (D1) is just the product of (D2) and (D3).

We now consider calculating $p(\mathbf{y}_2|\mu_2)$ from $p(\mathbf{y}|\mu)$ (removing observations). Let

$$\Sigma^{-1} = \mathbf{S} = \begin{pmatrix} \mathrm{S}_{11} & \mathrm{S}_{12} \\ \mathrm{S}_{21} & \mathrm{S}_{22} \end{pmatrix} \quad \text{so that} \quad \mathbf{S}\Sigma = \begin{pmatrix} \mathrm{S}_{11} & \mathrm{S}_{12} \\ \mathrm{S}_{21} & \mathrm{S}_{22} \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \mathbf{I} \qquad \text{(D4)}$$

in particular

$$\mathrm{S}_{11}\Sigma_{12} + \mathrm{S}_{12}\Sigma_{22} = 0$$

$$\Rightarrow \quad \Sigma_{12}\Sigma_{22}^{-1} = -\mathrm{S}_{11}^{-1}\mathrm{S}_{12} \qquad \text{(D5)}$$

We have already calculated and used $\mathbf{S}$ in the computation of $p(\mathbf{y}|\mu)$ so we obtain $\Sigma_{12}\Sigma_{22}^{-1}$ from (D5), substitute it twice in (D2) and calculate $p(\mathbf{y}_2|\mu_2)$ as $p(\mathbf{y}|\mu)/p(\mathbf{y}_1|\mathbf{y}_2\cap\mu)$. As before the matrix inversions/determinant calculations have the same order as $\mathbf{y}_1$.

The calculation of the analysis (using $\mathbf{y}_{o2}$ only) and its variance given by substituting (D5) in (D2) is the same as that in Lorenc (1981) section 3c, but in a rather different guise.

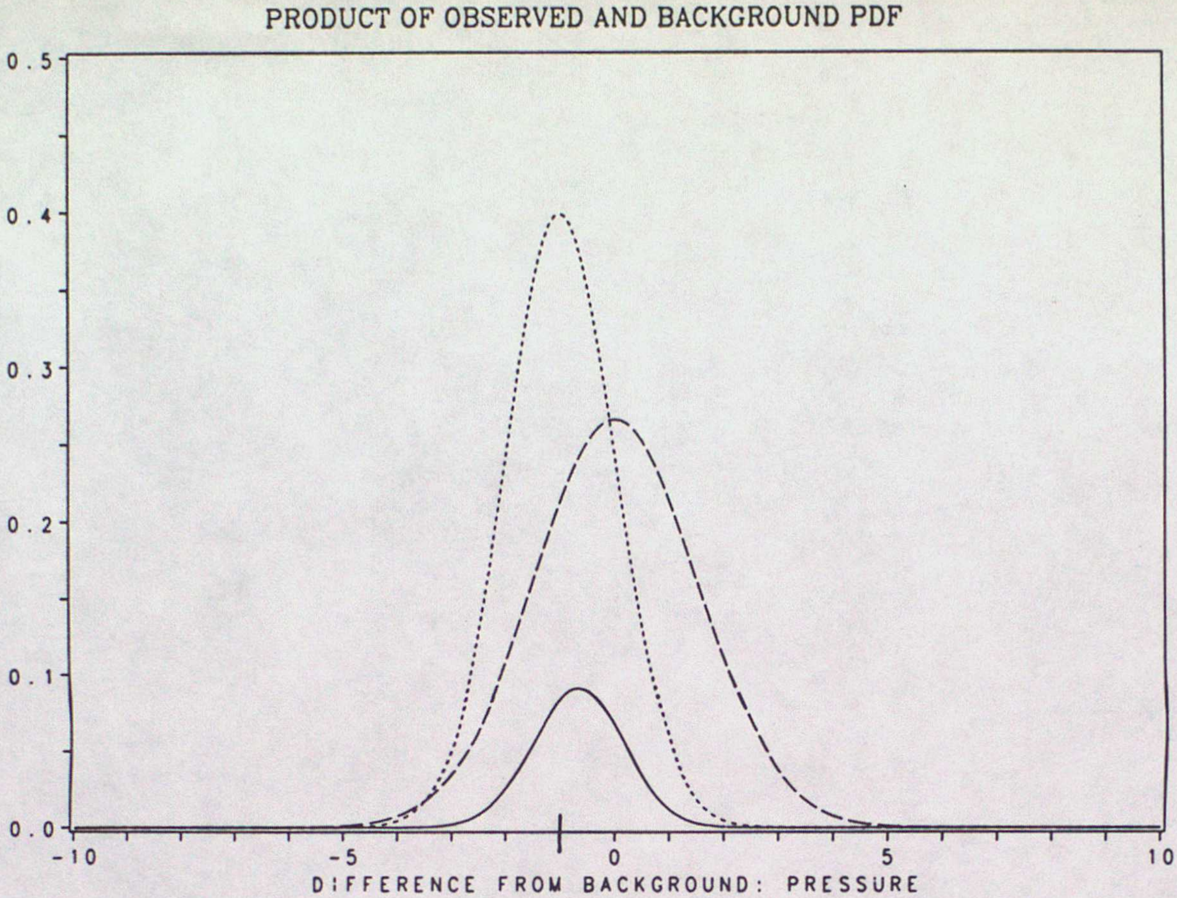

Figure 1a. $p(\mathbf{y}_o|\mathbf{y})$ (dotted, σ_o=1 mb), $p(\mathbf{y})$ (dashed, σ_b=1.5 mb) and their product $p(\mathbf{y}\cap\mathbf{y}_o)$ (solid) for one observation −1 mb from background. $p(\mathbf{y}_o)$ is area under solid curve, solid curve normalised gives $p(\mathbf{y}|\mathbf{y}_o)$.

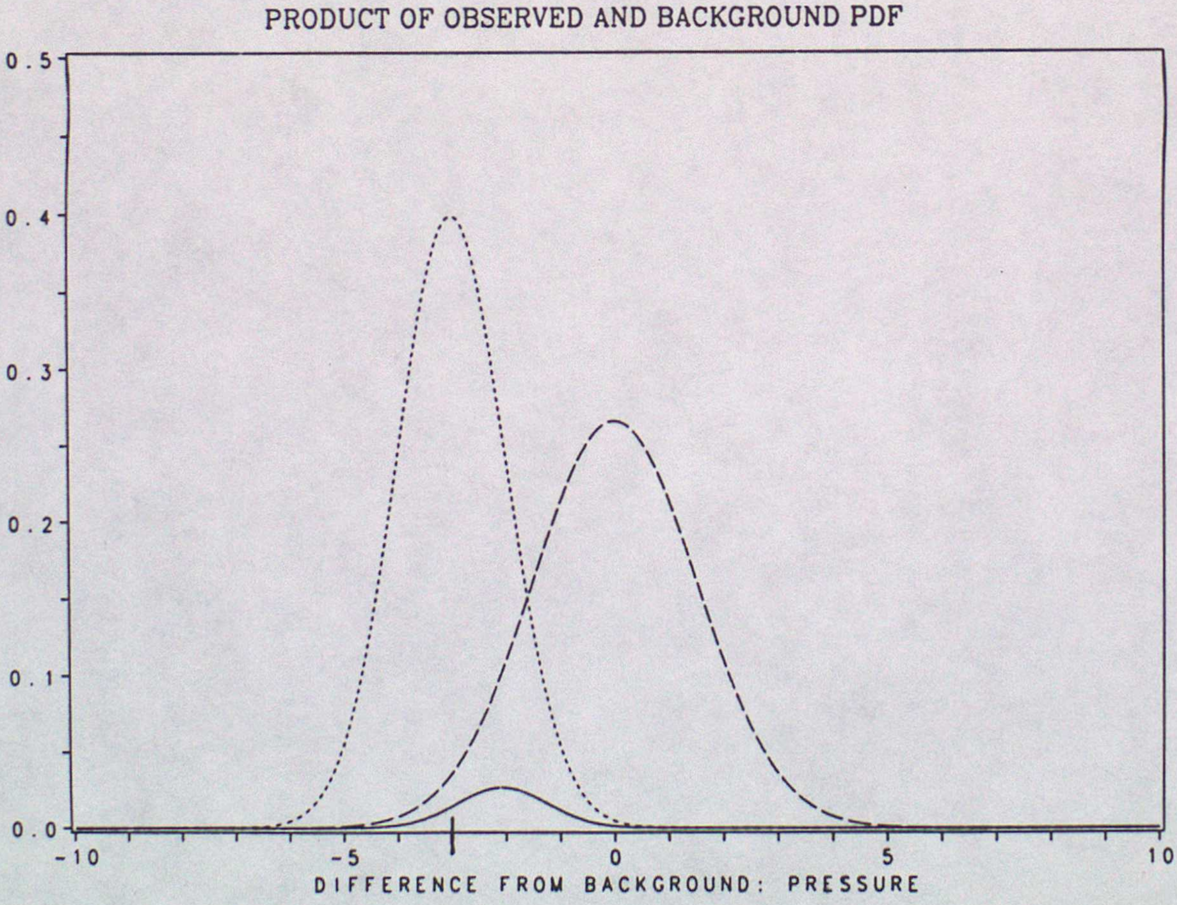

Figure 1b. As figure 1a except that observation is −3 mb from background.

The figures and tables were produced using the SAS package.

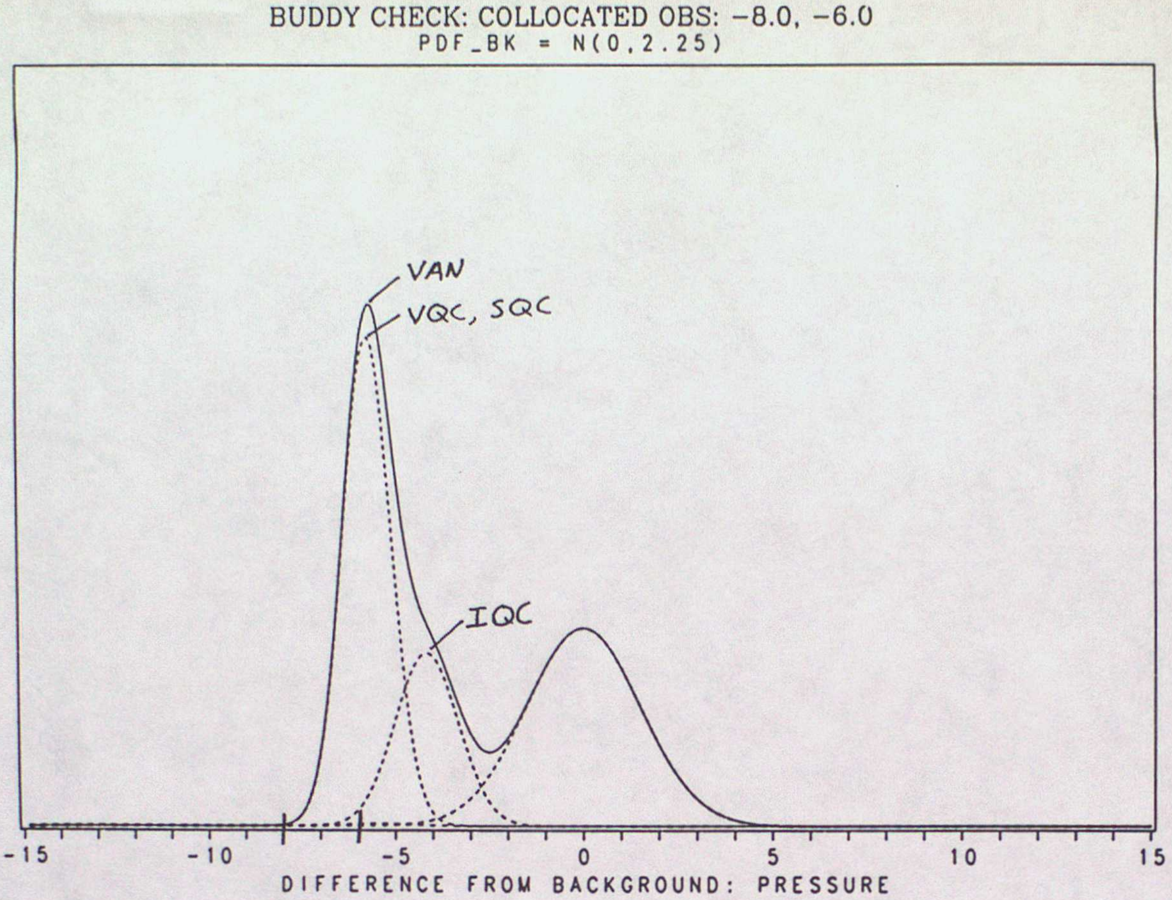

Figure 2a. Dotted curves give $p(\mathbf{y}\cap\mathbf{y}_o|C_\alpha)$ for all combinations of gross error C_α. Solid curve is sum of all dotted curves = $p(\mathbf{y}\cap\mathbf{y}_o)$. Observation increments are -8 and -6. Input pdfs and posterior PGEs as for table 1a.

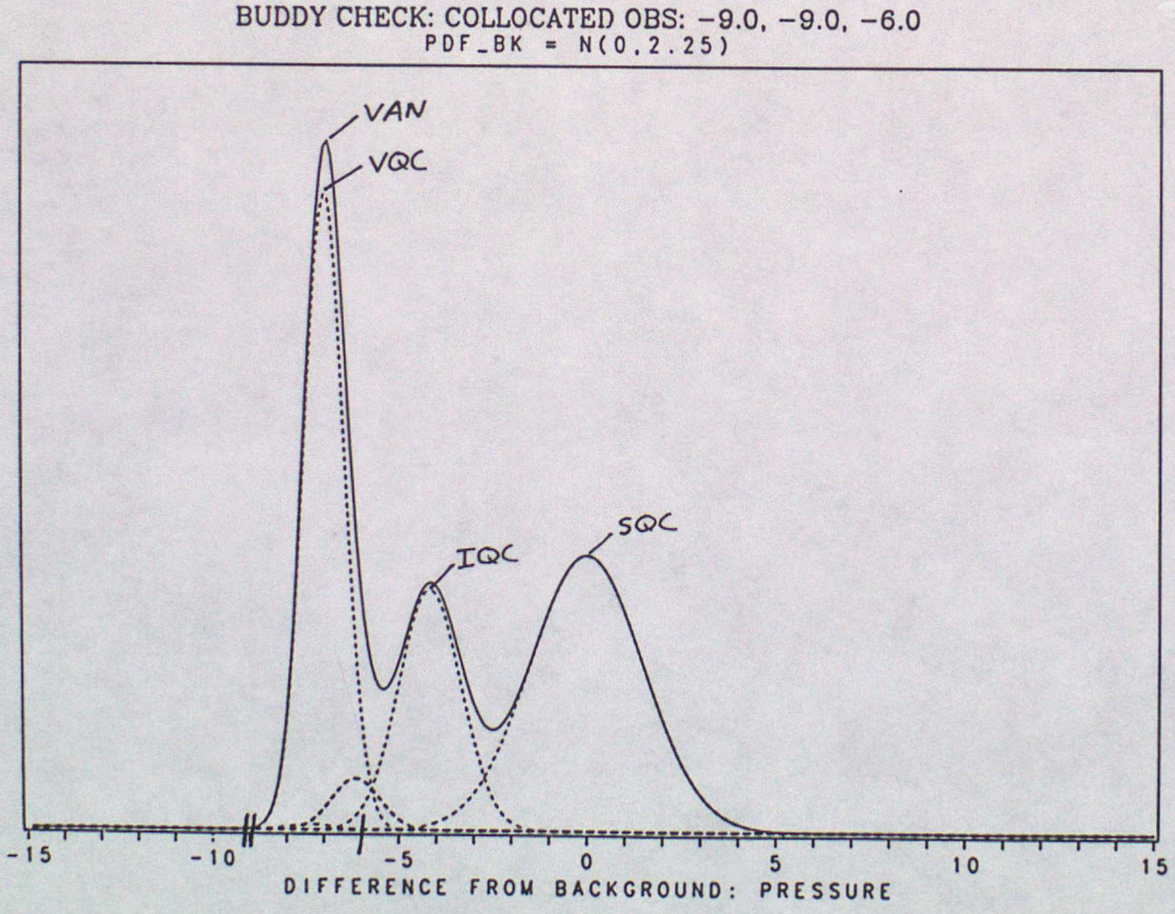

Figure 2b. As figure 2a except for case b, observation increments of -9, -9 and -6 (compare with table 1b).

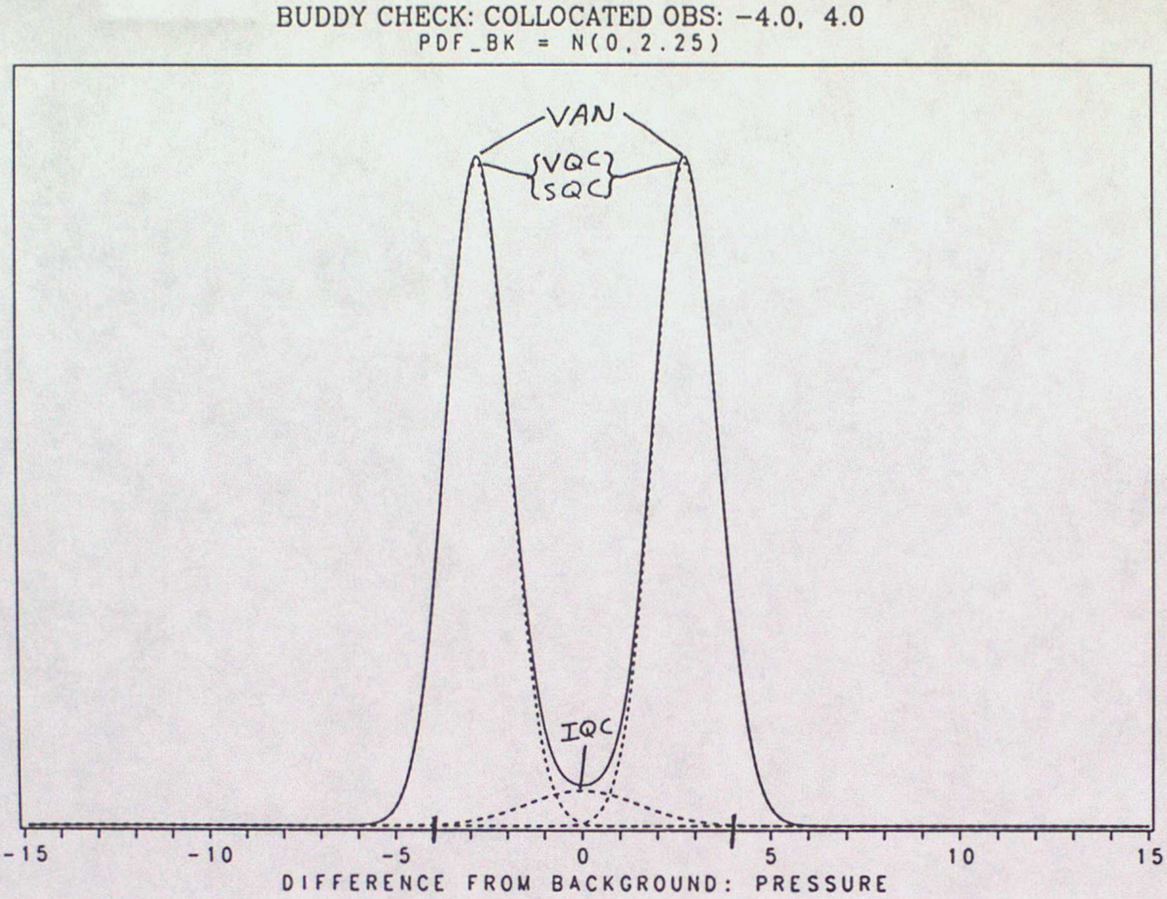

Figure 2c. As figure 2a except for case c, observation increments of -4 and 4 (compare with table 1c).

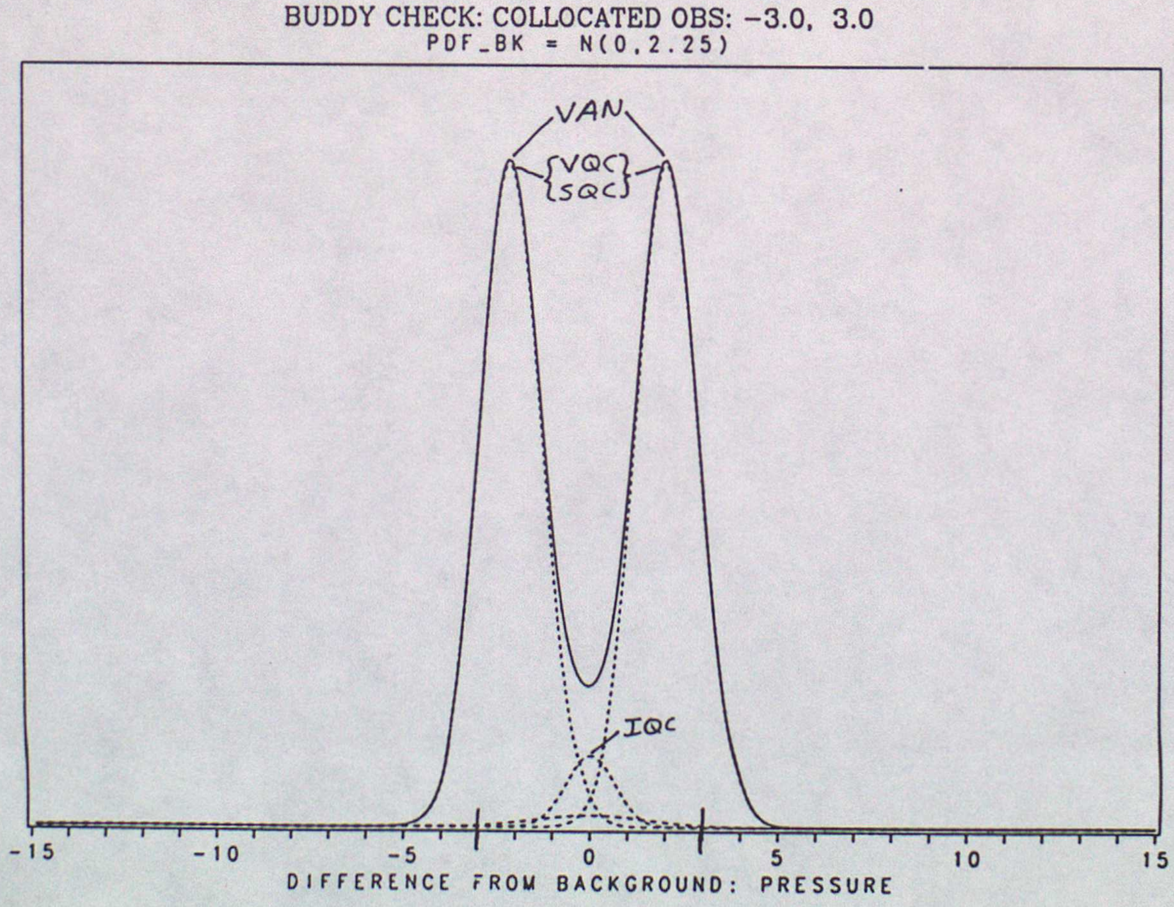

Figure 2d. As figure 2a except for case d, observation increments of -3 and 3 (compare with table 1d).

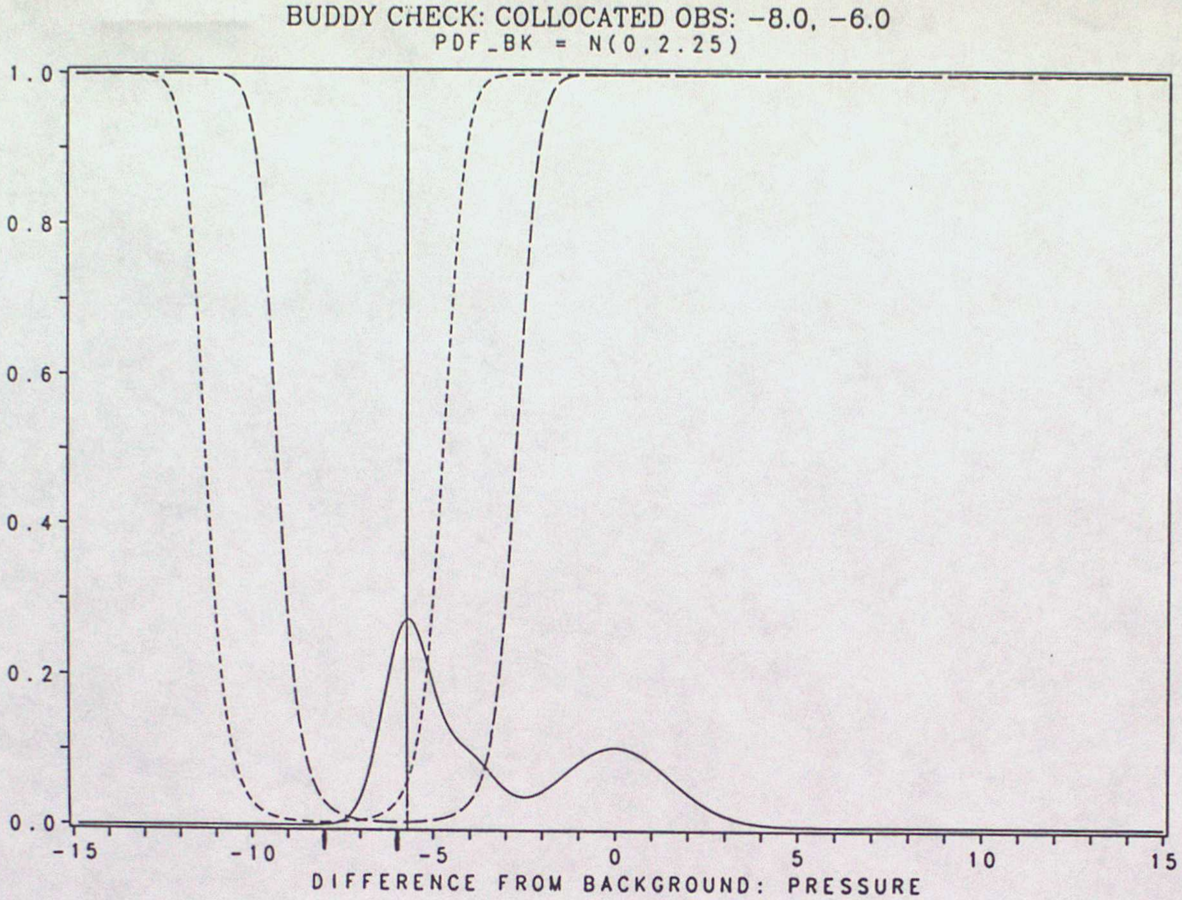

Figure 3. $p(\mathbf{y}|\mathbf{y}_o)$ (solid-normalised version of solid curve in figure 2a) and $P(G_i|\mathbf{y})$ for observations at -8 mb (dotted) and -6 mb (dashed).

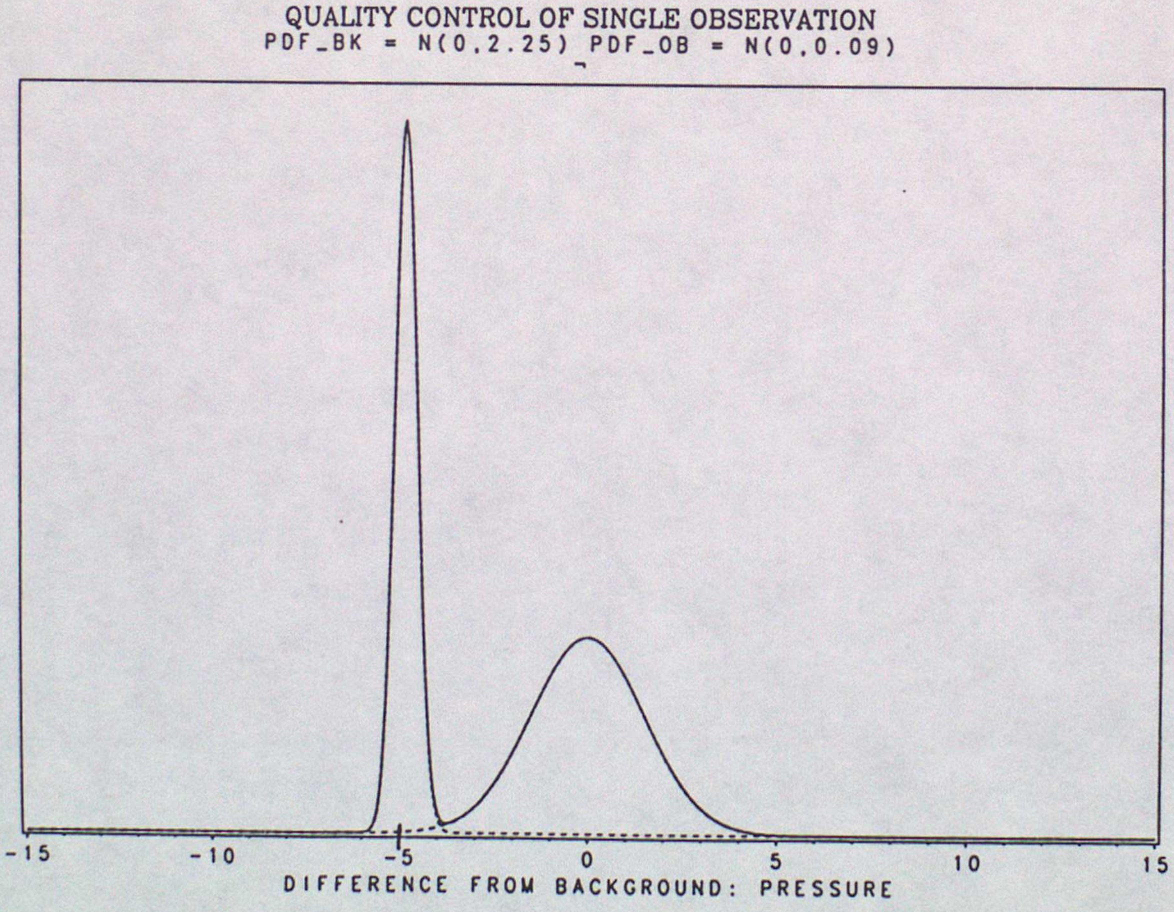

Figure 4. Posterior pdf for a single observation of high accuracy (σ_o=0.3 mb, σ_b=1.5 mb, k=0.043 mb^{-1}, P(G)=0.04) with an increment of -5 mb.

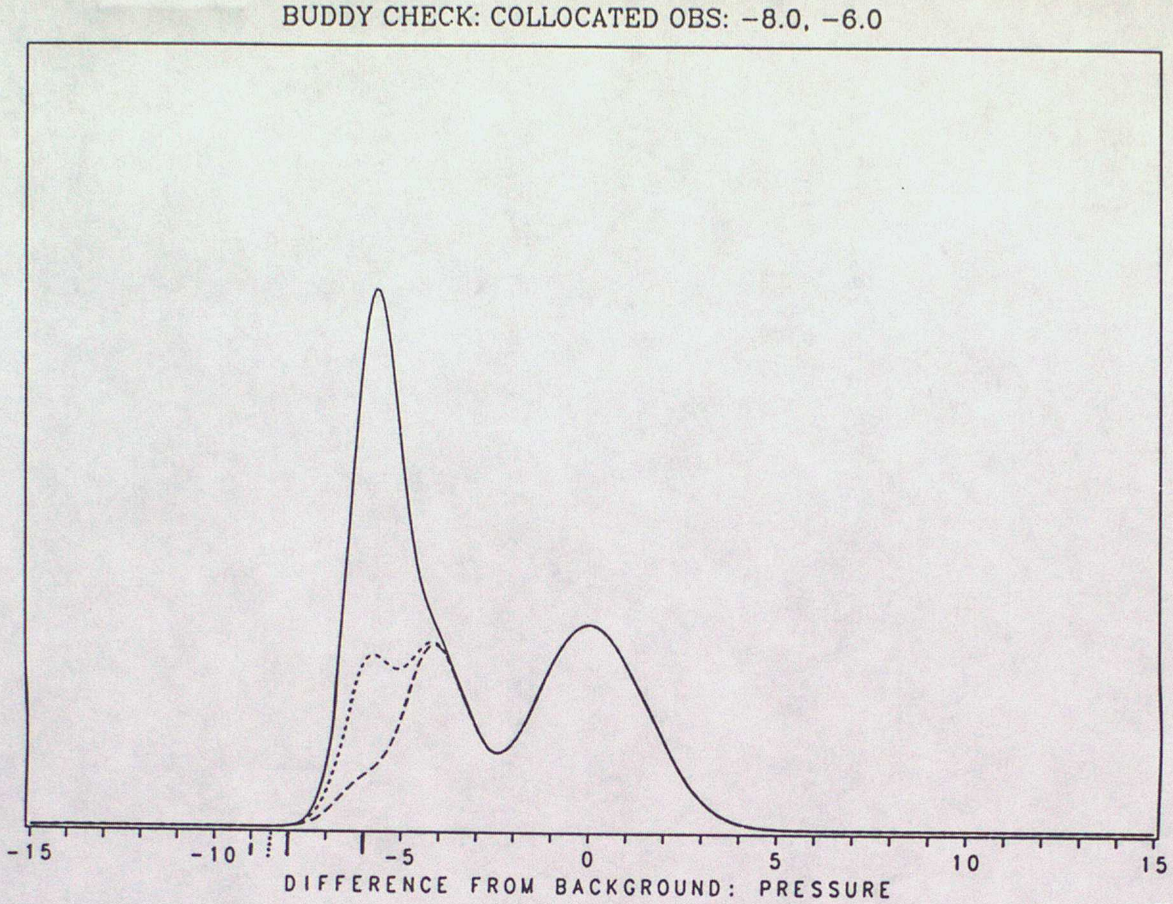

Figure 5. Solid line corresponds to the posterior pdf in figure 2a, with observations at -8 and -6. The dotted line is similar except for observations at -8.5 and -6, dashed line has observations at -9 and -6.

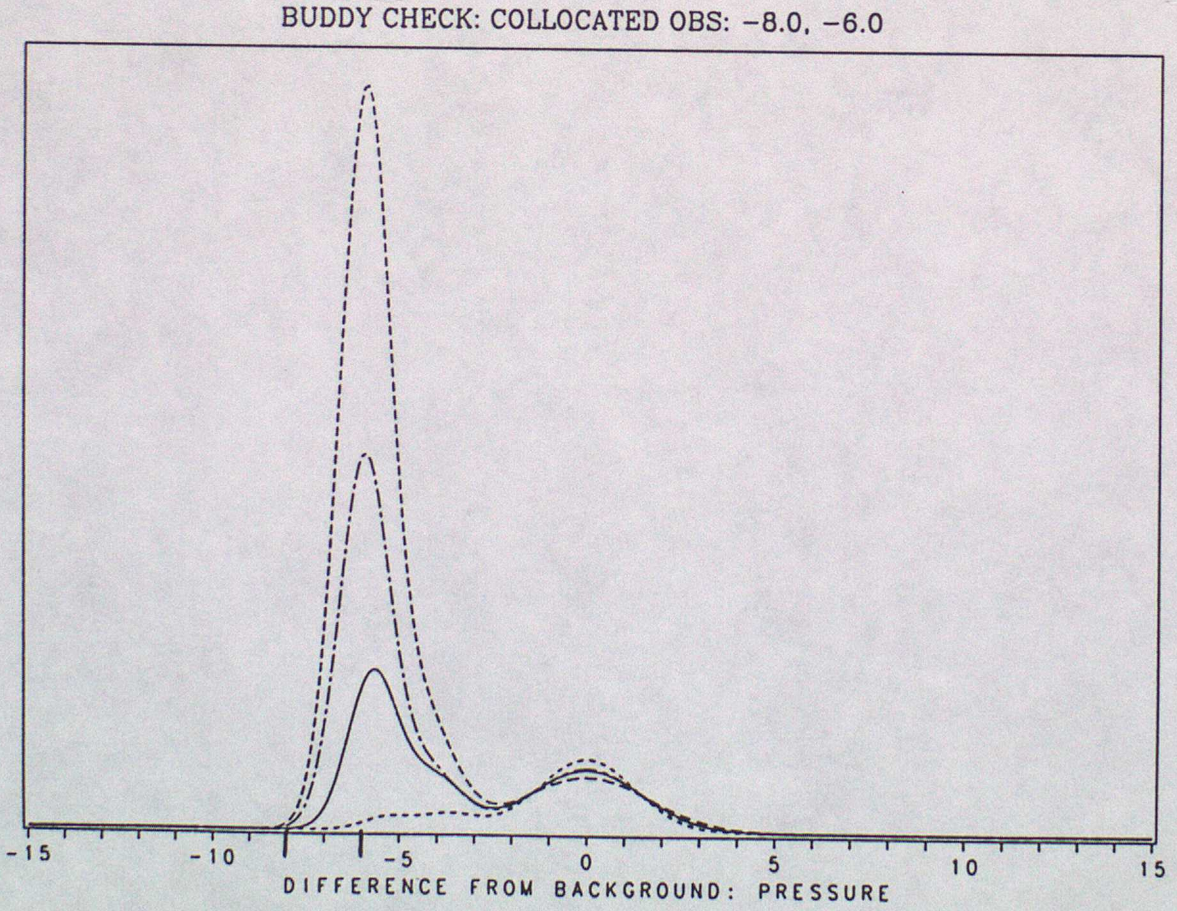

Figure 6. Solid line corresponds to the posterior pdf in figure 2a, with observations at -8 and -6 and σ_b=1.5 mb. The dotted and dashed lines are similar except that σ_b=1.27 and 1.7 mb respectively. Dash-dotted line uses background pdf $0.5(N(0,1.27^2)+N(0,1.7^2))$ which has σ=1.5.

SHORT RANGE FORECASTING DIVISION SCIENTIFIC PAPERS

This is a new series to be known as Short Range Forecasting Division Scientific Papers . These will be papers from all three sections of the Short Range Forecasting Research Division i.e. Data Assimilation Research (DA),Numerical Modelling Research (NM), and Observations and Satellite Applications (OB) the latter being formerly known as Nowcasting (NS). This series succeeds the series of Short Range Forecasting Research /Met O 11 Scientific Notes.

1. **THE UNIFIED FORECAST /CLIMATE MODEL .**
M.J.P. Cullen
September 1991

2. **Preparation for the use of Doppler wind lidar information in meteorological data assimilation systems**
A.C. Lorenc, R.J. Graham, I. Dharssi, B. Macpherson, N.B. Ingleby, R.W. Lunnon
February 1992

3. **Current developments in very short range weather forecasting.**
B.J. Conway
March 1992

4. **DIAGNOSIS OF VISIBILITY IN THE UK MET OFFICE MESOSCALE MODEL AND THE USE OF A VISIBILITY ANALYSIS TO CONSTRAIN INITIAL CONDITIONS**
S.P. Ballard, B.J. Wright, B.W. Golding
April 1992

5. **Radiative Properties of Water and Ice Clouds at Wavelengths Appropriate to the HIRS Instrument**
A.J. Baran and P.D. Watts
2nd June 1992

6. **Anatomy of the Canonical Transformation**
M.J. Sewell and I. Roulstone
27 June 1992

7. **Hamiltonian Structure of a Solution Strategy for the Semi-Geostrophic Equations**
I. Roulstone and J. Norbury
29 June 1992

8. **Assimilation of Satellite Data in models for energy and water cycle Research**
A.Lorenc
July 1992

9. **The use of ERS-1 data in operational meteorology**
A.Lorenc and others
July 1992

10. **Bayesian quality control using multivariate normal distributions**
N.B. Ingleby and A. Lorenc
July 1992