IMPLEMENTATION OF VARIATIONAL QUALITY CONTROL

Erik Andersson
European Centre for Medium-Range Weather Forecasts,
Shinfield Park, Reading, U.K.

1. INTRODUCTION

The first ECMWF implementation of three dimensional variational assimilation (3D-Var) (Courtier *et al.* 1997; Rabier *et al.* 1997; Andersson *et al.* 1997) relied on the OI (Optimal Interpolation) analysis system for quality control of data. All data which passed the OI analysis quality controls were then used in the 3D-Var analysis. The OI-check (Lorenc 1981) compared each datum in turn against an analysis based on surrounding data and the first-guess, excluding the datum being checked. Data were flagged and subsequently rejected if the observation departure from the analysis exceeded a certain multiple of the estimated analysis error.

The OI-check has now been replaced, following ideas of Dharssi *et al.* (1992) and Ingleby and Lorenc (1993), by a quality control mechanism incorporated within the variational analysis itself. A modification of the observation objective function (or cost function) to take into account the non-Gaussian nature of gross errors, has the effect of reducing the analysis weight given to data with large departures from the current iterand (or preliminary analysis). Data are not irrevocably rejected, but can regain influence on the analysis during later iterations if supported by surrounding data.

The method is named variational quality control (VarQC), and is based on Bayesian formalism. First, an *a-priori* (PGE_I) estimate of the probability of gross error is assigned to each datum, based on study of historical data. Then, at each iteration of the variational scheme, an *a-posteriori* estimate of the probability of gross error (PGE_F) is calculated, given the current value of the iterand (the preliminary analysis). VarQC modifies the gradient (of the observation cost function with respect to the observed quantity) by the factor ($1-PGE_F$), which means that data which are almost certainly wrong ($PGE_F \approx 1$) are given near-zero weight in the analysis. For the purpose of this paper, data with a $PGE_F > 0.75$ are considered 'rejected'.

The implementation of VarQC is straight forward for data with un-correlated errors. The extension to correlated data involves computing probabilities for all possible combinations of events, rejected/not rejected, amongst the correlated set of data. For a set of n correlated data 2^n terms would need to be evaluted for the combined probability density function. In this paper we have used an approximate form which gives acceptable results for radiosonde height data; the only data which are currently assumed (at ECMWF) to have correlated observation errors. The approximate form can detect up to three incorrect levels in the radiosonde reports. If more than three levels are likely to be affected by gross errors the whole report is rejected.

The next section presents the VarQC method for un-correlated and correlated data, respectively. Results from three assimilations comparing VarQC with OIQC and no QC are presented in section 3, followed by conclusions in section 4.

2. METHOD

As in Dharssi *et al.* (1992) and Ingleby and Lorenc (1993) it is assumed that an observation belongs to either of two populations: one which follows the normal Gaussian distribution N, representing random errors, and one which is modelled by a flat (box-car) distribution F, representing gross errors. Other models to represent the probability density of incorrect data could also be used. The choice of a flat distribution is convenient since it corresponds to the assumption that those data provide no information to the analysis.

With a prior probability of gross error (PGE_I in the introduction)

$$P(G_i) = A_i \tag{1}$$

and a probability of not having a gross error

$$P(\overline{G_i}) = (1 - A_i) \tag{2}$$

we can write the probability density function p_i for observation i as a sum of two terms

$$p_i = N_i P(\overline{G_i}) + F_i P(G_i) \tag{3}$$

$$N_i = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{y_i - H(x)}{\sigma_i}\right)^2\right] \tag{4}$$

$$F_i = \frac{1}{L_i} = \frac{1}{2l_i\sigma_i} \tag{5}$$

where L_i is the range of possible values of y_i (i.e. the width of the box-car). y_i is the observed value, with observation error σ_i. x is the control variable and H is the observation operator. In (5) the interval L_i has been written in terms of multiples of the observation error.

The observation cost-function is defined as

$$J_i = -\ln p_i \tag{6}$$

Substiting Eq. (1) to (5) into (6) we obtain an expression for the VarQC observation cost function

$$J_i^{QC} = -\ln\left(\frac{\gamma_i + \exp[-J_i^N]}{\gamma_i + 1}\right) \tag{7}$$

and its gradient

$$\nabla_x J_i^{QC} = \nabla_x J_i^N \left(1 - \frac{\gamma_i}{\gamma_i + \exp[-J_i^N]}\right) \tag{8}$$

expressed in terms of the normal cost function J_i^N (based on Gaussian statistics), and γ_i:

$$\gamma_i = \frac{A_i/2l_i}{(1-A_i)/\sqrt{2\pi}}$$

The negative term modifying the gradient in (8) can be shown to be equal to the *a-posteriori* probability of gross error (PGE_F), given x, and assuming that $H(x)$ is correct (see Ingleby and Lorenc, 1993). Eq (8) shows that data which are likely to be incorrect are given reduced weight in the analysis. The probability depends on x which means that it is important to have an as good as possible preliminary analysis at the start of the variational quality control.

Equations (7) and (8) are illustrated in Fig. 1a and 1b. The figure shows the normal J^N and the VarQC J^{QC} cost functions (top), the VarQC gradient $\nabla_x J^{QC}$ (middle) and the factor within brackets in (8), i.e. 1-PGE_F (lower panel), plotted against observation departure. The parameters are $A = 1\%, l = 5$ and $\sigma = 2$. Figure 1a shows VarQC with a flat distribution for gross errors. The cost function flattens out for large values of the departure, and its gradient goes towards zero, as the probability of the observation being correct also drops towards zero. The interval within which the observation is partly used / partly rejected is relatively narrow for this set of parameters.

The second example (Fig. 1b) shows a similar diagram resulting from using a different statistical model for F (the probability density for incorrect data). Here a Gaussian with three times the standard deviation (3σ) has been used rather than a flat distribution. We can see that the resulting analysis does not reject the data with a high PGE_F but uses them with the larger observation error - thereby remaining robust to outlier data (Huber, 1977).

2.1 Rejection limits

VarQC does not require the specification of threshold values at which rejections occur - so called rejection limits. Rejections occur gradually and are a function of the estimate of PGE_F. We can, however, derive an expression for the normalized departure at which rejection occurs:

$$\left[\frac{H(x) - y}{\sigma}\right]^2 > 2\ln\left[\frac{PGE_F}{(1 - PGE_F)\,\gamma_i}\right] \tag{9}$$

If we define that a datum has been rejected if its gradient has been reduced by a factor 0.25, i.e. PGE_F=0.75 we obtain a relationship which is a function of A and l only. Note that in (9) the effective rejection limit is in terms of normalized departures from x, which in turn depend on specified background error statistics as well as on observational data in the vicinity of each observation.

2.2 Un-correlated observations

The combined probability for n un-correlated events is given by the product of the individual probabilities. For observations $i = 1, 2 \ldots$ to n we have

$$p = p_1 p_2 \cdots p_n = \prod_{i=1}^{n} \left[N_i P(\overline{G_i}) + F_i P(G_i)\right] \tag{10}$$

Expanding the product, in the special case of only two observations ($n = 2$), we obtain four terms, each representing a combination of events: $N_1 N_2$ - both correct, $N_1 F_2$ and $N_2 F_1$ - one observation correct the other not, and $F_1 F_2$ - both incorrect. In the general case there will be 2^n such terms, each representing a combination of quality control decisions. All of these combinations will be taken into account in the variational quality control simply by summing up the individual J_i contributions:

$$J = -\ln \prod_{i=1}^{n} p_i = \sum_{i=1}^{n} -\ln p_i \tag{11}$$

2.3 Correlated observations

In the case of n correlated observations (10) is not valid. Instead, the probability of each of the 2^n combinations of events has to be evaluated separately.

$$
\begin{aligned}
p \;=\; & P(\overline{G_1})P(\overline{G_2})\ldots P(\overline{G_n})N \;+ \\
& P(G_1)P(\overline{G_2})\ldots P(\overline{G_n})F_1N_{-1} \;+ \\
& P(\overline{G_1})P(G_2)\ldots P(\overline{G_n})F_2N_{-2} \;+ \\
& P(\overline{G_1})P(\overline{G_2})\ldots P(G_n)F_nN_{-n} \;+ \\
& P(G_1)P(G_2)\ldots P(\overline{G_n})F_1F_2N_{-(1,2)} \;+ \\
& \ldots + \ldots + \\
& P(G_1)P(G_2)\ldots P(G_n)F_1F_2\ldots F_n
\end{aligned}
\tag{12}
$$

where N_{-i} is the Gaussian probability density excluding observation number i.

Like Ingleby and Lorenc we define C_α as any combination of events, with C_1 referring to all data correct and C_{2^n} meaning all data in error.

Take for example the particular event C_{α_k} that the first k observations are correct and the remaining $n-k$ observations are in error. Its probability density function is

$$
p(C_{\alpha_k}) \;=\; \prod_{i=1}^{k} P(\overline{G_i}) \prod_{i=k+1}^{n} P(G_i)\; N_{-(k+1\to n)} \prod_{i=k+1}^{n} F_i \tag{13}
$$

This is a valid form for any of the terms in (11) after a suitable reordering of the observations.

$N_{-(k+1\to n)}$ is now a multi-dimensional Gaussian excluding observations $k+1$ to n:

$$
N_{-(k+1\to n)} \;=\; \left[(2\pi)^k |O_{1\to k}| \right]^{-0.5} \exp\left[-\frac{1}{2} (y - H(x))^T_{1\to k}\, O^{-1}_{1\to k} (y - H(x))_{1\to k} \right] \tag{14}
$$

The matrix $O^{-1}_{1\to k}$ is the inverse of the upper left kxk elements of O, the observation error covariance matrix, and $|O_{1\to k}|$ is its determinant.

Substituting (1), (2), (5) and (14) into (13), writing o for the correlation matrix with elements $o_{ij} = O_{ij}/\sigma_i\sigma_j$, noting that the determinant $|O| = |o| \prod_{i=1}^{n} \sigma_i^2$ we have

$$p(C_{\alpha_k}) = \prod_{i=1}^{n} \sigma_i^{-1} \prod_{i=1}^{k} (1-A_i) \prod_{i=k+1}^{n} \left(\frac{A_i}{2\, l_i} \right) \left[(2\pi)^k \, |o_{1\to k}| \right]^{-0.5} \exp\left[-\frac{1}{2} \left(\frac{y - H(x)}{\sigma} \right)^T_{1\to k} o^{-1}_{1\to k} \left(\frac{y - H(x)}{\sigma} \right)_{1\to k} \right] \tag{15}$$

The only data that are currently at ECMWF assumed to have correlated errors are the radiosonde height data. These are used at up to 15 pressure levels (standard levels chosen preferentially), with data from different stations assumed un-correlated. Thus $n = 15$ and $2^n = 32{,}768$ is the number of terms on the form given by (15) required for an exact computation.

2.3.1 *Evaluating all terms*

The full flexibility to allow the rejection of any combination of levels within a radiosonde ascent can only be achieved by calculating all the 2^n terms as above. This is not difficult but may be too expensive. It needs to be done once per iteration of the variational minimisation. The main cost is in the inversion of the different O-matrices involved. Some considerable savings could be done by pre-calculating all the matrix inverses for the fix set of standard pressure levels. However, non-standard level data may be used in cases when standard level data are missing.

2.3.2 *Omitting some terms*

It is feasible to simply neglect some of the terms, thereby limiting the procedure to reject only certain combinations of data. In particular it is possible to merely retain the possibility of rejecting the entire report - not individual levels within the report - by keeping only the first and the last terms of (12). This would be justifiable if one source of error is likely to affect the whole report (ie wrong station height or biased temperature sensor). The probability A would then refer to the chance of having a gross error affecting the report at one or more levels.

Incidently, this form of VarQC is the best solution for the quality control of wind data. It allows the probability of error in u- and v-components of wind observations to be computed jointly, neglecting the cross-terms $N_u F_v$ and $N_v F_u$. The two wind-components would then either be both rejected or both accepted (or both something-in-between).

2.3.3 *Diagonalising the problem*

The third solution assumes that the gross errors themselves are correlated in the same way as the random errors, ie according to the matrix O. We would then diagonalise the O-matrix: $O = E\Lambda E^T$, where E is the matrix of eigen-vectors and Λ is a diagonal matrix of eigen-values. We can define 'rotated' departures δ as $\delta = E(H(x) - y)$. We are now back to the un-correlated case where we can apply the quality control independently to the individual 'rotated' departures. The adjoint would include a multiplication by E^T to come back to the observed quantities, and to distribute the QC-weights to the observed data.

2.4 **Preferred method**

The implementation of VarQC for un-correlated observations is straight forward and computationally inexpensive. For correlated data it is too expensive and probably unnecessary to evaluate the probability for each combination of events (rejection/non-rejection), as in 2.3.1. On the other hand, it seems un-physical to assume that the gross errors have the same correlations as the random errors, as in 2.3.3 above. The gross errors are more likely to be un-correlated.

The best solution is one where a limited number of outcomes are accounted for. The thoughts in 2.3.2 can be elaborated to include not only the zero order term, but also the terms representing the rejection of one or two data per profile - the first and second order terms, *et cetera.* Let o be the highest order of terms kept. The first α_o outcomes are then treated fully as in (15) . The probabilities of the remaining $2^n - \alpha_o$ outcomes are not evaluated but replaced by the probability $P(G_{>o})$ of having more than o gross errors in the report. The number of terms of order o is $\frac{n!}{(n-o)!o!}$. In the case of 15 data $o = 2$ would leave $\alpha_o = 121$ terms to be evaluated per radiosonde report. The probability $P(G_{>o})$ is equal to the sum of all neglected probabilities:

$$P(G_{>o}) = \sum_{\alpha=\alpha_o+1}^{2^n} P(C_\alpha) \tag{16}$$

with

$$P(C_\alpha) = \prod_{i=1}^{k} P(\overline{G_i}) \prod_{i=k+1}^{n} P(G_i) \tag{17}$$

Noting that $\sum_{\alpha=1}^{2^n} P(C_\alpha) = 1$ we can alternatively write

$$P(G_{>o}) = 1 - \sum_{\alpha=1}^{\alpha_o} P(C_\alpha) \qquad (18)$$

which is easier to compute.

With $n = 15$ and $A = 1\%$ (for example) gives $P(G_{>2}) = 0.04\%$ indicating that neglected terms only represent a very small part of the total prior probability.

The operational algorithm is as follows: All radiosondes are checked with $o = 1$. If this check indicates that any level (or all levels) is likely to be in error, then the check is recomputed with $o = 3$. In this way up to 3 incorrect levels can be identified. If more than three levels are likely to be in error, PGE_F for all levels of the report becomes large and the weight of the whole report is reduced.

3. RESULTS

The first implementation of VarQC was tuned, as closely as possible, to reproduce the results of the OIQC. Histograms of OI rejections were studied to find out its rejection limits. The A and l parameters of VarQC were then set such that similar rejection limits were achieved, using (9). Figure 2 shows an example of such histograms. It shows all rejected SYNOP pressure data in a one-month period comparing OIQC and VarQC (filled). The diagram verifies that the two quality control methods perform similarly. The A and l parameters for all data types except TOVS and SCAT data, were set in a similar fashion. The OIQC did not apply to TOVS and SCAT data. The extension of VarQC to those data will be added later.

3.1 **Forecast impact**

The forecast impact of quality control was tested in a two week period: 950422 to 950506. Three 3D-Var assimilations were run at T106 resolution. All experiments had identical first-guess checks. One experiment was run without quality control against other data (NoQC), one used OIQC and one used VarQC. The impact of data quality control on the performance of the forecasts was very small in this period. Fig 3 shows a scatter diagram of 500 hPa geopotential rms error of NoQC against VarQC, for the two hemispheres at day 5. The scatter is small and the impact is neutral. The result with OIQC is similar (not shown).

4. CONCLUSION

Variational quality control has been implemented in the ECMWF 3D-Var data assimilation scheme and it has been shown to be an efficient and adequate replacement for OIQC.

Quality control is a very non-linear process which could slow down the convergence of the variational

minimisation. VarQC introduces multiple minima in the cost function, corresponding to different quality control decision. It is therefore important to have a good starting point for the VarQC minimisation. Moreover, the method assumes that the observation equivalent $H(x)$ is correct at each iteration, which makes the starting point doubly important. In our implementation a good starting point for VarQC is achieved by carrying out 30 iterations of the 3D-Var minimisation without quality control before switching on VarQC for the remaining 40 iterations.

TOVS and SCAT data will also be subjected to VarQC shortly. The performance will be re-assessed on other periods, and the settings of the input parameters will be adjusted based on statistics of observation departures from the first guess and analyses.

References

Andersson, E., J. Haseler, P. Undén, P. Courtier, G. Kelly, D. Vasiljević,C. Branković, C. Cardinali, C. Gaffard, A. Hollingsworth, C. Jakob, P. Janssen, E. Klinker,A. Lanzinger, M. Miller, F. Rabier, A. Simmons, B. Strauss, J-N. Thépaut and P. Viterbo. The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part III: Formulation. Submitted to *Q. J. R. Meteorol. Soc.*

Courtier, P., E. Andersson, W. Heckley, J. Pailleux, D. Vasiljević, M. Hamrud, A. Hollingsworth, F. Rabier and M. Fisher, 1997: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part I: Formulation. Submitted to *Q. J. R. Meteorol. Soc.*

Dharssi, I., A. C. Lorenc and N. B. Ingleby, 1992: Treatment of gross errors using maximum probability theory. *Q. J. R. Meteorol. Soc.*, **118,** 1017-1036.

Huber, P. J., 1977: Robust statistical methods. *Soc. for Industrial and Applied Mathematics*, **27.**

Ingleby, N. B. and A. C. Lorenc, 1993: Bayesian quality control using multivariate normal distributions. *Q. J. R. Meteorol. Soc.*, **119,** 1195-1225.

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

Rabier, F., A. McNally, E. Andersson, P. Courtier, P. Undén, J. Eyre, A. Hollingsworth and F. Bouttier, 1997: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part II: Structure functions. Submitted to *Q. J. R. Meteorol. Soc.*

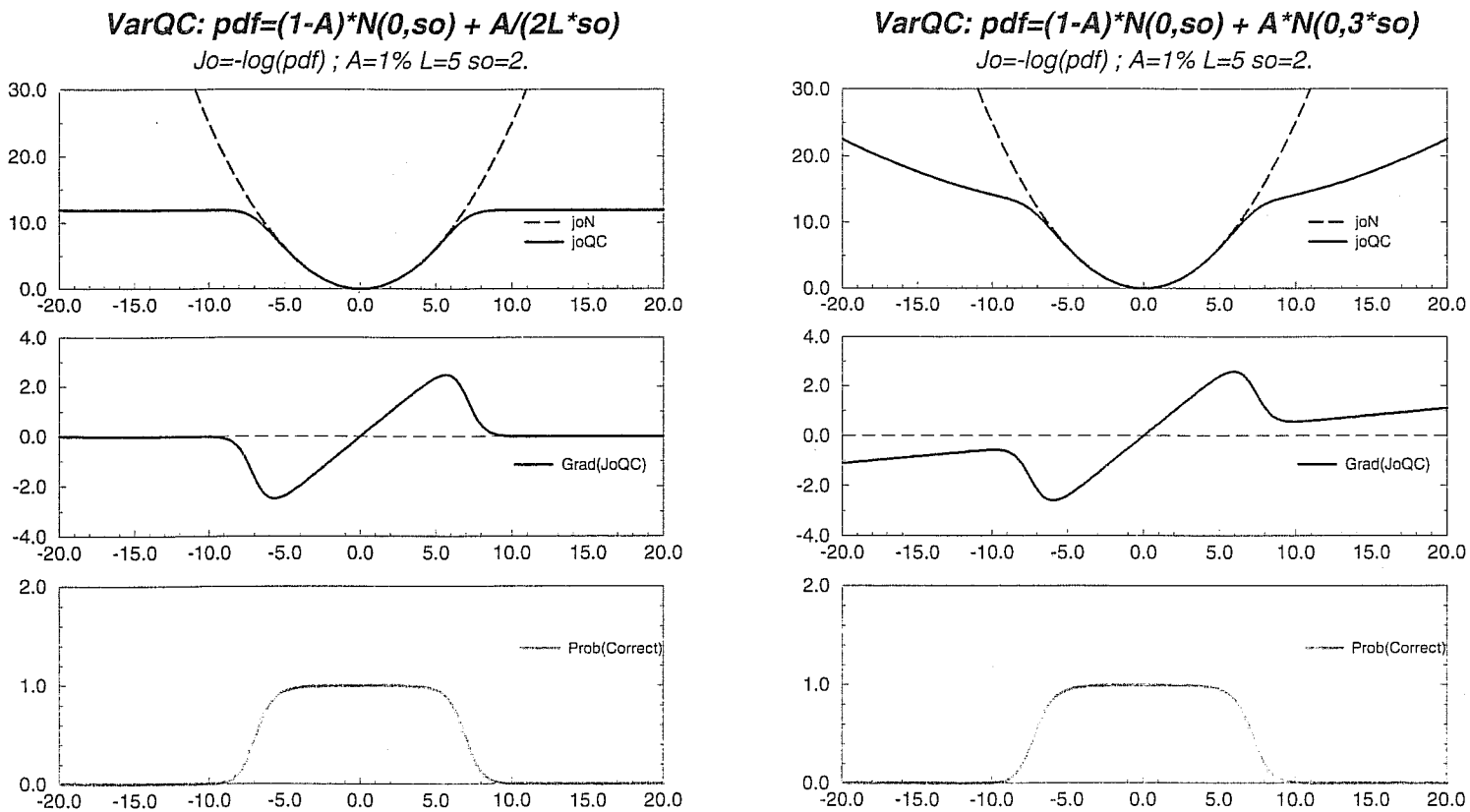

Fig. 1 Top panels show cost-function without QC (dashed) and with VarQC (full line). Middle panels show gradient of VarQC cost function and lower panels show the a-posteriori probability that the observation is correct. The distribution of gross errors is flat in a) and a Gaussian with three times the observation error in b).

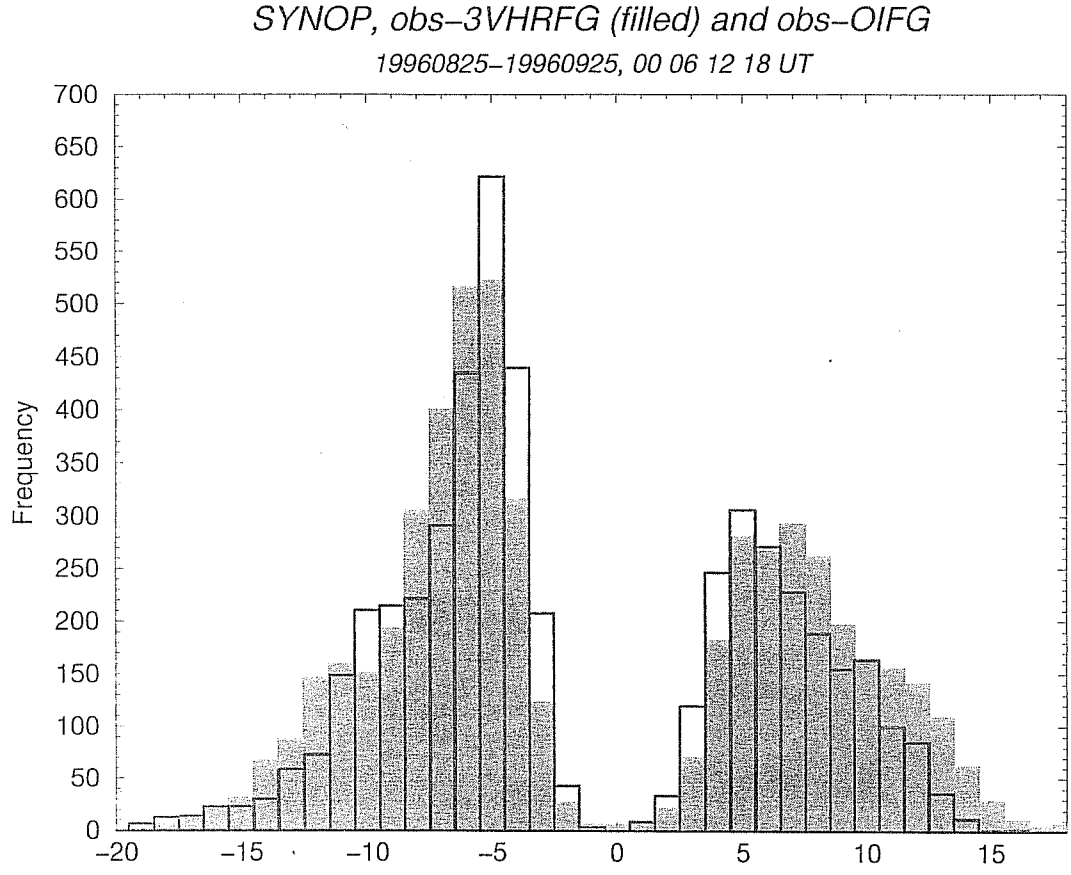

Fig.2 Histogram of obs-first guess departures for rejected SYNOP pressure data for OIQC (black outline) and VarQC (filled gray), in a 30 day period.

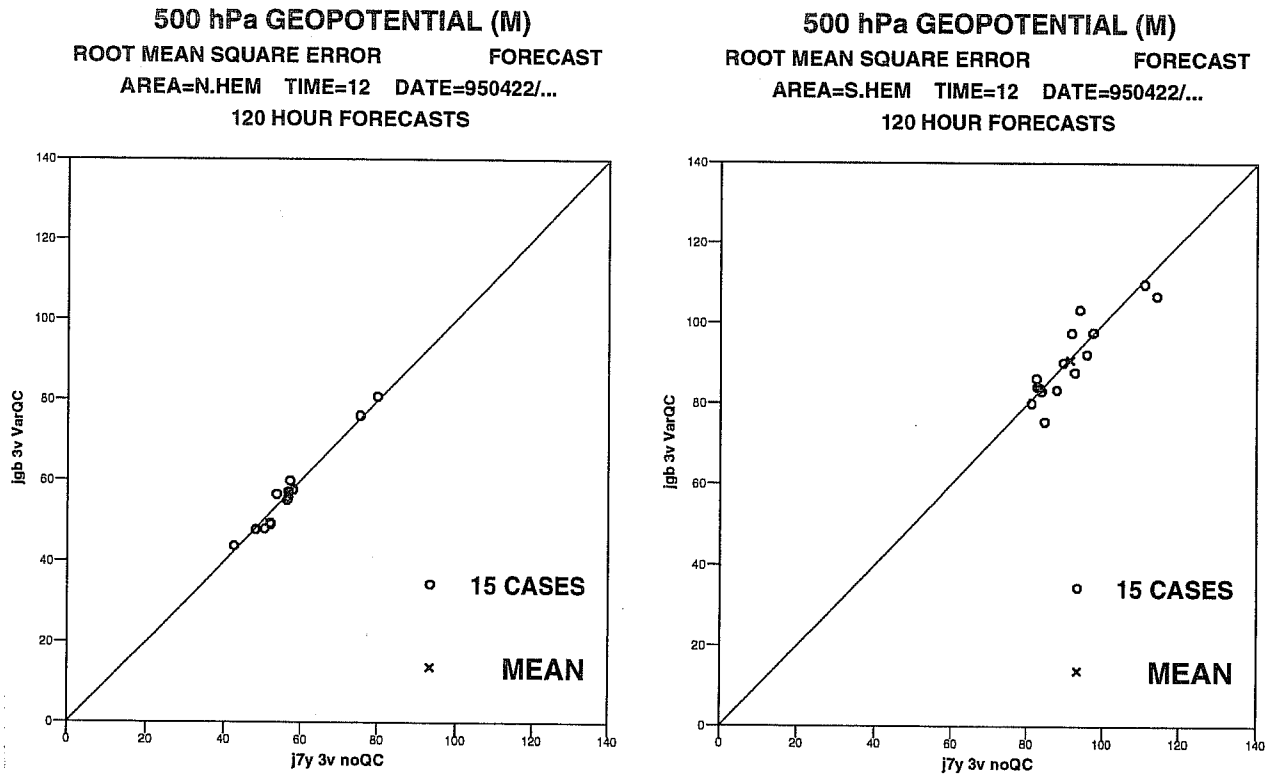

Fig. 3 Rms of 500 hPa geoptential forecast error (m) from two 15-day assimilation experiments: NoQC along x-axis and VarQC along y-axis, for the Northern (left) and Southern (right) Hemispheres.