

An adaptive buddy check for observational quality control

By DICK P. DEE^{1,2*}, LEONID RUKHOVETS^{1,2}, RICARDO TODLING^{1,2}, ARLINDO M. DA SILVA² and JAY W. LARSON³

¹*General Sciences Corporation, USA*

²*NASA/Goddard Space Flight Center, USA*

³*Argonne National Laboratory, USA*

(Received 22 September 2000; revised 31 March 2001)

SUMMARY

An adaptive buddy-check algorithm is presented that adjusts tolerances for suspect observations, based on the variability of surrounding data. The algorithm derives from a statistical hypothesis test combined with maximum-likelihood covariance estimation. Its stability is shown to depend on the initial identification of outliers by a simple background check. The adaptive feature ensures that the final quality-control decisions are not very sensitive to prescribed statistics of first-guess and observation errors, nor on other approximations introduced into the algorithm.

The implementation of the algorithm in a global atmospheric data assimilation is described. Its performance is contrasted with that of a non-adaptive buddy check, for the surface analysis of an extreme storm that took place over Europe on 27 December 1999. The adaptive algorithm allowed the inclusion of many important observations that differed greatly from the first guess and that would have been excluded on the basis of prescribed statistics. The analysis of the storm development was much improved as a result of these additional observations.

KEYWORDS: Adaptive tolerances Data assimilation Quality control

1. INTRODUCTION

This article describes an adaptive algorithm for statistical quality control of observations for use in atmospheric data assimilation. The method is based on the so-called buddy check, which assumes that the observables are spatially coherent, so that nearby measurements (buddies) should tend to confirm each other. If a suspect observation cannot be supported by its buddies, then it may well be corrupt, in which case it must be discarded. On the other hand, that observation may contain genuine information about an unexpected event, in which case it should be used. Choosing one or the other of these two opposites can have a large impact on the assimilated product. The most important feature of our new buddy-check algorithm is that the rejection limits, which are traditionally based on a priori statistics, are made to depend as well on the current observed variability of the local flow. This results in relatively tolerant acceptance criteria in synoptically active situations, and in more stringent criteria when conditions are calm.

Variants of the buddy check have been implemented in sequential statistical analysis schemes (Lorenc 1981; Woollen 1991) and more recently in the framework of variational assimilation (Andersson and Järvinen 1999). The buddy check represents only a single component of observational quality control; it is usually preceded by various sanity checks and other preliminary quality-control procedures that are tailored to specific types of observations (e.g. Gandin 1988; Collins 1998). For a general statement and detailed discussion of the problem of quality control of meteorological observations see Lorenc and Hammon (1988) and Collins and Gandin (1990). Some of the early papers on numerical weather analysis and prediction (Bergthórsson and Döös 1955; Staff Members, Joint Numerical Weather Prediction Unit 1957; Bedient and Cressman 1957) contain many interesting comments on quality control that are still pertinent today.

It is intuitively clear that some notion of reasonable differences among nearby observations is required in order to distinguish contaminated data from genuine information

* Corresponding author: NASA/GSFC, Data Assimilation Office, Mail Code 910.3, Greenbelt, MD 20771, USA.
e-mail: dee@dao.gsfc.nasa.gov

about the atmosphere. Such differences are due to the natural small-scale variability of the observables and to the inherent uncertainties of the measurement process. It is also clear that quality control is not strictly a deterministic problem, in the sense that it is not always possible to know with certainty whether an observation has been corrupted or not. The performance of a quality test therefore must be measured in terms of probabilities. For the ideal test the probability of failing a genuine observation is bounded by some small fraction (the significance level of the test), while the probability of failing a contaminated observation (the power of the test) is maximal. This is a classic problem in the theory of statistical hypothesis testing (e.g. Lehmann 1997), which, however, can be solved only if the probability distributions of both types of observations (genuine as well as contaminated) are known.

Using Bayesian methods rather than the formalism of hypothesis testing, Lorenc and Hammon (1988) analysed the case of uniformly distributed errors (representing *gross errors*, i.e. data contamination) superimposed upon Gaussian background and observation noise (representing normal errors). They derived an expression for the optimal buddy-check tolerances for this case, which is a function of the background- and observation-error variances and of the *a priori* probability of gross error for the observation being checked. They also discussed in detail the use of an observation-processing database for estimating and updating the gross error statistics for different observing systems, which is required information for their method. The implementation of Bayesian quality-control methods in a variational framework was developed by Dharssi *et al.* (1992), Ingleby and Lorenc (1993), and by Andersson and Järvinen (1999).

A general problem with the application of statistical analysis methods to atmospheric observations is the difficulty in modelling the errors. Flow-dependent aspects are not well represented by the covariance models used in operational data-assimilation systems. Quantitative information on the reliability of different observing systems is hard to come by; statistics on gross errors are based on past performance and are likely to depend on the methodology that has been used to detect them. The danger with a quality-control algorithm that primarily relies on *a priori* statistics, typically based on time- and space-averaging, is that it may tend to enforce those very statistics, for example, by excessively rejecting observations whenever the local variability is larger than usual. This can happen during the onset of an extreme weather event, if the forecast is poor and the error variances are underestimated. Any available observations are particularly valuable under those circumstances. The challenge is then to design an effective and robust quality-control algorithm that performs well in continuously changing conditions.

For these reasons we develop our approach to statistical quality control with the primary goal of reducing the dependence on prescribed error statistics as much as possible. In well-observed regions this can be achieved, as described in this paper, by extracting additional information about local errors from the observations themselves. Our method initially uses the same statistical information as the global analysis system in which it is embedded, including (possibly correlated) error covariances for different observations, and (possibly dynamic) background-error covariances. However, the implied tolerances are locally adjusted for each buddy check, based on the observed variability of nearby data that have already passed the quality control. In relatively data-dense areas this results in quality-control decisions that are not very sensitive to prescribed statistics. As the data density decreases, the dependency on prescribed statistics gets larger, and in very poorly observed situations the method essentially reverts to an iterated conventional buddy check.

The outline of this paper is as follows. In section 2 we present the derivation of an adaptive and iterative buddy-check algorithm. Given an initial classification of the

observations as either suspect or not, we formulate the quality-control problem as a statistical test of the hypothesis that the observed discrepancies among the data are reasonable in view of their presumed probability distributions. The test is made adaptive by locally readjusting the prescribed error estimates during each iteration. We prove that each iteration of the algorithm is stable, in the sense that the adaptive tolerances are bounded if the initial identification of suspects is based on a simple background check. We also prove the convergence of the algorithm in a finite number of iterations. We then illustrate the performance of the method, and the effect of the adaptive feature in particular, by means of some simple contrived examples.

In section 3 we describe the implementation of statistical quality control in an early version of the Goddard Earth Observing System Data Assimilation System (GEOS DAS, version 3). A background check, based on prescribed error statistics for the global analysis system, is used to define an initial set of suspect observations, but does not itself reject any observations. Each suspect observation is then subject to a buddy check with adaptive tolerances. The procedure is repeated until no additional observations pass the buddy check. We briefly discuss the practical use of the background check in monitoring the validity of the prescribed error statistics. We show that the quality control tends to favour the most reliable observing systems, as expected. We then examine in detail the performance of the GEOS DAS quality control for the analysis of a severe storm that took place over Europe on 27 December 1999. The development of this storm was well observed but poorly analysed by several operational centres. We show that the GEOS DAS quality control allowed many observations into the analysis that would have been excluded by a non-adaptive statistical algorithm. We also show that the final quality-control decisions were in fact rather insensitive to the prescribed error statistics.

2. THE ADAPTIVE BUDDY CHECK

Let the vector \mathbf{w}^o denote a subset of the observations which are subject to quality control. We will be flexible with regard to the specific composition of this subset, although we have in mind a mix of observations of different types, located within a limited spatial region. In any case, the starting point for the buddy check is a preliminary classification of all elements of \mathbf{w}^o as either *suspect* or not. The observations that are not suspect are *buddies*. Observations may be flagged as suspect simply because they are outliers, or for any other reason. The buddy check tests the extent to which suspect observations are supported by their buddies. This is done by first predicting the values of the suspect data from the buddies, and then to test whether the discrepancy between the predictions and the actual observed values is reasonable or not.

The test can be formulated conveniently in terms of differences between the observations and some background estimate, which, in a cycling data-assimilation system, is usually a short-term model forecast \mathbf{w}^f . The observed-minus-forecast residual vector \mathbf{v} is then defined by

$$\mathbf{v} = \mathbf{w}^o - \mathbf{h}(\mathbf{w}^f), \quad (1)$$

where \mathbf{h} is the observation operator associated with \mathbf{w}^o . In general, this operator involves nonlinear forward models relating observables (e.g. radiances) to model variables (e.g. temperatures). For direct observations of forecast model state variables, \mathbf{h} is simply an interpolation from model grid points to observation locations. The residual \mathbf{v} is

often referred to as the *innovation*, because it represents that part of the observational information which is not contained in the forecast*.

The initial partitioning of all observations as either suspect or not is important, because only non-suspect observations are used in the prediction step of a buddy check. If some, but not all, suspect observations pass the check, then the partitioning should be updated accordingly and the buddy check should be repeated. This leads to an iterative procedure that terminates when no additional observations pass the test. The remaining suspect observations are then rejected. We begin by describing a single iteration of the algorithm.

(a) *The buddy check as a hypothesis test*

The buddy check can be regarded as a statistical test of the assumption that none of the observations have been contaminated. We introduce the null hypothesis

$$\mathbf{v} \sim \mathcal{N}(0, \mathbf{S}), \tag{2}$$

or, in words, that all residuals are jointly normally distributed with zero mean and known covariance \mathbf{S} . As discussed in more detail below, the matrix \mathbf{S} can be obtained from the global analysis system and incorporates error statistics for different observations as well as information about background-error covariances. Nevertheless, we will later make allowance for the fact that, in practice, (2) always represents an idealization. For now we assume that rejection of the null hypothesis implies that some of the data are likely contaminated.

Given the information expressed in the null hypothesis, an optimal estimate of the suspect residuals can be obtained as follows. Partition the residual vector \mathbf{v} as

$$\mathbf{v} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \tag{3}$$

where \mathbf{x} contains the residuals associated with suspect observations, and \mathbf{y} those associated with buddies. The corresponding blocks of the residual covariance are

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_x & \mathbf{S}_{xy} \\ \mathbf{S}_{xy}^T & \mathbf{S}_y \end{bmatrix}, \tag{4}$$

where superscript T is the transpose. Under the null hypothesis, the conditional distribution of \mathbf{x} given \mathbf{y} is multivariate normal,

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mathbf{x}^*, \mathbf{S}^*), \tag{5}$$

with

$$\mathbf{x}^* = \mathbf{S}_{xy}\mathbf{S}_y^{-1}\mathbf{y}, \tag{6}$$

$$\mathbf{S}^* = \mathbf{S}_x - \mathbf{S}_{xy}\mathbf{S}_y^{-1}\mathbf{S}_{xy}^T, \tag{7}$$

(Jazwinski 1970, theorem 2.13). Each of these quantities is well defined when \mathbf{S} is positive definite. In particular, $\mathbf{x}^* = \mathbf{x}^*(\mathbf{y})$ is the optimal estimate of \mathbf{x} based on \mathbf{y} (Jazwinski 1970, theorem 5.3), and the matrix \mathbf{S}^* is the error covariance of this estimate. The main computation involved in (6) is the solution of the linear system $\mathbf{S}_y\mathbf{z} = \mathbf{y}$.

* This usage is imprecise and somewhat wishful in the context of data assimilation: observed-minus-forecast residuals are innovations only when the assimilation is optimal. See, for example, Anderson and Moore (1979, section 5.3) for a mathematical definition of the innovations process.

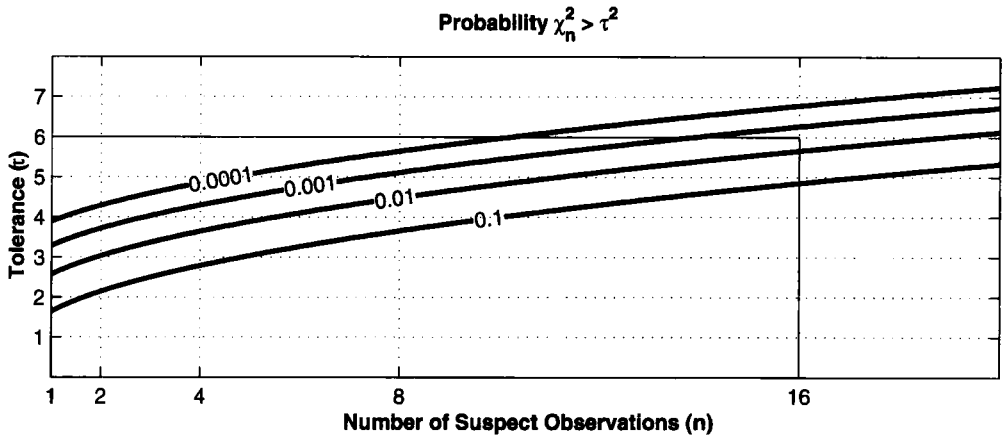


Figure 1. Chi-squared probabilities as a function of the number of suspect observations n , and of the tolerance parameter τ .

Computation of (7) requires n additional solves of this system, where $n = \dim \mathbf{x}$ is the number of suspect observations.

With (5) to (7) in hand, we can apply rigorous statistical tests to the null hypothesis. The general idea is to compute the probability p that the suspect observations are consistent with the null hypothesis. If p is smaller than some prescribed threshold δ , we reject the hypothesis and conclude that it is likely that at least some of the data are corrupt. The value of δ is called the *significance level* of the test*.

To be precise, (5) to (7) imply that the scalar quantity

$$\chi_n^2 = \frac{1}{n}(\mathbf{x} - \mathbf{x}^*)^T (\mathbf{S}^*)^{-1} (\mathbf{x} - \mathbf{x}^*) \tag{8}$$

has a chi-squared distribution with n degrees of freedom (e.g. Tarantola 1987, section 4.3.6). It follows that for any tolerance parameter $\tau > 0$,

$$p = p(\chi_n^2 > \tau^2) = 1 - P\left(\frac{n}{2}, \frac{\tau^2}{2}\right), \tag{9}$$

where $P(a, x)$ is the incomplete gamma function defined by

$$P(a, x) = \frac{1}{\Gamma(a)} \int_x^\infty e^{-t} t^{a-1} dt, \tag{10}$$

where Γ is the gamma function.

Figure 1 shows the probability $p(\chi_n^2 > \tau^2)$ as a function of the number of suspect observations n , and of τ . The latter must be specified by the user. For example, we read from the figure that if $n = 16$ and the value of χ_n^2 is $6^2 = 36$, then we may reject the null hypothesis at the 1% significance level.

Once we conclude that one or more of the suspect data is likely corrupt, we still need a procedure for marking individual observations for rejection. A simple approach

* The significance level bounds the probability that the null hypothesis is falsely rejected, i.e. the probability of failing a genuine observation. Its value does not imply the probability of contamination, and it is therefore misleading to state that the null hypothesis may be rejected with confidence $1 - \delta$. See von Storch and Zwiers (1998, chapter 4) for a lucid discussion of this and other subtleties associated with statistical hypothesis testing.

is to consider the marginal distribution of each suspect residual element x_i in \mathbf{x} . All marginal distributions of a normal distribution are themselves normal (Jazwinski 1970, theorem 2.12), which, together with (5) to (7) implies

$$x_i | \mathbf{y} \sim \mathcal{N}(x_i^*, S_{ii}^*), \tag{11}$$

with x_i^* , S_{ii}^* the appropriate elements in \mathbf{x}^* , \mathbf{S}^* , respectively. It follows that

$$p = p\left(\frac{|x_i - x_i^*|^2}{S_{ii}^*} > \tau^2\right) = \sqrt{\frac{2}{\pi}} \int_{\tau}^{\infty} e^{-t^2/2} dt. \tag{12}$$

For example, when $\tau = 3$ we have $p < 0.01$, which means that any observation for which $|x_i - x_i^*| > 3\sqrt{S_{ii}^*}$ may be rejected at the 1% significance level. Equation (12) corresponds to (9) with $n = 1$, so that these probabilities are also shown in Fig. 1.

In practice, testing procedures and criteria for rejection must be designed to depend on the nature of the observations. For example, if \mathbf{x} corresponds to a simultaneous rawinsonde temperature and moisture report, then we might choose to reject both measurements if together they fail the chi-squared test. Similarly, if a height profile obtained from satellite data fails the test, then that profile should be rejected in its entirety. The advantage of simultaneously applying a single test to multiple data is that error correlations among the suspect residuals themselves can be properly taken into account. These and other possibilities are a matter of strategy and of practical viability, depending, for example, on available information about error covariances.

(b) *Adaptive tolerances*

The theory so far presumes that rejection of the null hypothesis implies a strong likelihood of the presence of contaminated data. A test of the null hypothesis, however, may also fail due to inaccurate statistical information. In that case (2) will not hold even if all observations are genuine. For the buddy check to be effective, therefore, it must be robust with respect to any prescribed statistical parameters.

The specification of the covariance matrix \mathbf{S} in particular will strongly influence the outcome. Equation (12) shows that the tolerance for a univariate test is proportional to the error variance S_{ii}^* , which, by (7), depends on \mathbf{S} . This means that the rejection limits of the buddy check depend on the statistical variability of the observed-minus-forecast residuals: outlier observations will be less readily rejected, for example, if the deviations are expected to be large. This behaviour is clearly desirable, but it crucially depends on the ability to specify locally accurate covariances for the data residuals.

We can show from (1) (e.g. Dee 1995) that the covariance matrix \mathbf{S} of the residual vector \mathbf{v} is

$$\mathbf{S} \approx \mathbf{R} + \mathbf{H}\mathbf{P}^f\mathbf{H}^T, \tag{13}$$

where \mathbf{R} is the observation-error covariance, \mathbf{H} is the linearized observation operator defined by

$$\mathbf{H} = \left. \frac{\partial \mathbf{h}}{\partial \mathbf{w}} \right|_{\mathbf{w}=\mathbf{w}^f}, \tag{14}$$

and \mathbf{P}^f is the forecast-error covariance. Equation (13) would be exact if forecast and observation errors were statistically independent and if the observation operator were linear. In practice, error covariances in operational data-assimilation systems are difficult to estimate, and many other assumptions leading to (13) may be violated as well.

The prescription of \mathbf{S} inherited from the global analysis system therefore represents, at best, a reasonable first guess of the residual covariance matrix, typically based on time- and space-averaged statistics. The buddy check, on the other hand, is a local test for which these statistics are not necessarily appropriate. We can attempt to make adjustments to the tolerances by introducing a parameter α to rescale the prescribed covariances. Replacing the null hypothesis (2) by

$$\mathbf{v} \sim \mathcal{N}(0, \alpha^2 \mathbf{S}), \tag{15}$$

the maximum-likelihood estimate of α based on the (non-suspect) data residuals \mathbf{y} is given by

$$\alpha^2 = \frac{1}{m} \mathbf{y}^T \mathbf{S}_y^{-1} \mathbf{y}, \tag{16}$$

where $m = \dim \mathbf{y}$ is the number of buddies (Dee 1995, section 4). This computation is practically cost-free, by virtue of (6). Equation (7) shows that to rescale the residual covariance \mathbf{S} is equivalent to rescaling the conditional covariance \mathbf{S}^* . This in turn has the effect of modifying the tolerance for the buddy check in (12), replacing τ by $\alpha \tau$.

A practical implementation must include a strategy for subdividing the observations into suitable subsets, because a uniform rescaling of the residual covariances as in (15) is not reasonable unless the observations are confined to a limited region in space. Even then, one could argue that perhaps the background- and observation-error variances should be rescaled independently, or that other parameters of the covariance models should be adjusted as well. It is possible to generalize (15) by introducing additional parameters (Dee 1995), but at the cost of complicating the algorithm. This would be difficult to justify in the absence of an accurate understanding of the local errors. Since the buddy-check tolerances are primarily sensitive to the total (background plus observation) error variances, a uniform rescaling appears to be both simple and effective.

Adapting tolerances based on actual observations requires some care, to ensure that the algorithm remains stable. Equation (16) implies

$$\frac{(\mathbf{y}^T \mathbf{y})/m}{\lambda_{\max}(\mathbf{S}_y)} < \alpha^2 < \frac{(\mathbf{y}^T \mathbf{y})/m}{\lambda_{\min}(\mathbf{S}_y)}, \tag{17}$$

where $\lambda_{\min}(\mathbf{S}_y)$ and $\lambda_{\max}(\mathbf{S}_y)$ denote the smallest and largest eigenvalues of \mathbf{S}_y , respectively. Thus, the inclusion in \mathbf{y} of a few extreme outliers could cause excessive amplification of the residual covariances, resulting in a buddy check which is unable to detect gross errors. This can be prevented by making sure that all non-suspect residuals are initially bounded, as in

$$y_i^2 < \tau_b^2 S_{y_{ii}} \quad \text{for } i = 1, \dots, m, \tag{18}$$

where τ_b is a prescribed tolerance parameter. This represents a simple *background check*, strictly based on prescribed statistics. We then have

$$\alpha^2 < \tau_b^2 \frac{(\text{trace } \mathbf{S}_y)/m}{\lambda_{\min}(\mathbf{S}_y)}, \tag{19}$$

which gives an upper bound for the rescaling parameter that depends on the prescribed covariances only. Note that λ_{\min} can be estimated based on (13); for example, in the case of independent observations with prescribed error standard deviation σ^0 , we have $0 < (\sigma^0)^2 < \lambda_{\min}$.

The variance of the maximum-likelihood estimate α given by (16) is proportional to m^{-1} for sufficiently large m (Dee 1995, section 4). However, the variance of the estimate may be large when m is very small and \mathbf{S}_y is poorly conditioned (see (17)). To alleviate this, one can introduce a smoothing parameter $m^* \geq 0$ and use instead

$$\alpha^2 = \frac{\mathbf{y}^T \mathbf{S}_y^{-1} \mathbf{y} + m^*}{m + m^*}. \quad (20)$$

This has the effect of reverting to prescribed error statistics when $m \ll m^*$.

(c) Use of prior information on gross errors

Our theory uses information about the distribution of observed-minus-forecast residuals in the absence of data contamination only. Assuming that the final adapted covariance estimates are accurate, this leads to a hypothesis test such that the probability of falsely rejecting a genuine observation is bounded by the significance level of the test, which is a known function of the parameter τ (see Fig. 1). The theory does not imply the probability of inadvertently accepting a contaminated observation, since this would require knowledge of the probability distribution of the data errors responsible for the contamination. If an accurate estimate of this distribution is available, then it is possible, in principle, to derive the optimal hypothesis test at any given significance level, such that the probability of accepting a corrupt observation is minimal.

Lorenc and Hammon (1988) presented a Bayesian analysis of this problem, in which it is assumed that the probability distribution of gross errors is available a priori. They derived, for the case of uniformly distributed gross errors, an expression relating τ to the prior probability $P(G)$ of gross error for the observation being checked (their Eq. (21)). This provides an objective means for specifying a value for the buddy-check rejection tolerance, consistent with the intuitive notion that this should depend on the reliability of the observing instrument. The optimal value of the parameter τ in the case of uniformly distributed gross errors turns out to be approximately proportional to $\sqrt{\log(1/P(G))}$ when $P(G)$ is small (see their Fig. 5).

Since the buddy-check rejection limits are most sensitive to the specified variances of the observed-minus-forecast residuals, it is important to improve the variance estimates when possible. A prior estimate of the probability of gross error, if it is available for the observation being tested, could be used to guide the specification of τ . This might be especially useful in very sparsely observed situations, when the adaptation of rejection limits based on surrounding data is not effective and the buddy check must rely on prescribed statistics.

(d) Summary of the algorithm

The following algorithm implements a simple background check, followed by an iterative, adaptive buddy check, applied to a subset \mathbf{v} of observed-minus-forecast residuals with prescribed covariance \mathbf{S} . The background check serves to define the initial set $\mathbf{x}^{(0)}$ of suspects, as well as to bound the non-suspect residuals used for the buddy check.

$$\mathbf{x}^{(0)} = \{v_i \in \mathbf{v} \text{ such that } |v_i| > \tau_b \sqrt{S_{ii}}\} \quad (\mathcal{A}.0)$$

for $k = 1, 2, \dots$

$$\mathbf{y} = \{v_i \in \mathbf{v} \text{ such that } v_i \notin \mathbf{x}^{(k-1)}\} \tag{A.1}$$

$$\mathbf{x}^* = \mathbf{S}_{xy} \mathbf{S}_y^{-1} \mathbf{y} \tag{A.2}$$

$$\mathbf{S}^* = \mathbf{S}_x - \mathbf{S}_{xy} \mathbf{S}_y^{-1} \mathbf{S}_{xy}^T \tag{A.3}$$

$$\alpha^2 = \frac{\mathbf{y}^T \mathbf{S}_y^{-1} \mathbf{y} + m^*}{\dim \mathbf{y} + m^*} \tag{A.4}$$

$$\mathbf{x}^{(k)} = \{x_i \in \mathbf{x}^{(k-1)} \text{ such that } |x_i - x_i^*| > \alpha \tau \sqrt{S_{ii}^*}\} \tag{A.5}$$

$$\text{if } \dim \mathbf{x}^{(k)} = \dim \mathbf{x}^{(k-1)} \text{ or } \dim \mathbf{x}^{(k)} = 0 \text{ then stop} \tag{A.6}$$

end.

The algorithm usually terminates after a small number of iterations when the set of suspects remains unchanged ($\dim \mathbf{x}^{(k)} = \dim \mathbf{x}^{(k-1)}$). Otherwise each iteration results in a decrease in the number of suspects by at least one, so that the convergence of the algorithm is guaranteed in at most $n = \dim \mathbf{x}^{(0)}$ iterations.

After convergence, any residuals that remain in $\mathbf{x}^{(k)}$ are rejected. The probability that a rejected residual is consistent with (15) is then bounded by the significance level of the test. For example, with $\tau = 3$ the significance level is about 0.3%, which means that the probability of false rejection is less than 0.003.

The user must also specify the tolerance parameter τ_b for the background check and the relaxation parameter m^* . For $\tau_b = 2$, (12) predicts that roughly 4.5% of all residuals should be initially marked as suspect. With $m^* = 0$ the algorithm fully adjusts the tolerances for the buddy check during each iteration, based on the current set of non-suspects. Setting $m^* \gg m$ effectively turns off the adaptive feature in the algorithm, which then completely relies on prescribed covariances.

(e) *A simple illustration*

We simulate a one-dimensional domain with 32 equally spaced observation locations, labelled $i = 1, 2, \dots, 32$. The analysis system operates under the null hypothesis

$$\mathbf{v} \sim \mathcal{N}(0, \mathbf{S}), \tag{22}$$

with

$$S_{ij} = \begin{cases} 1, & \text{for } i = j, \\ \frac{1}{2} e^{-0.2(i-j)^2}, & \text{otherwise.} \end{cases} \tag{23}$$

This model represents residuals with a spatially uncorrelated observation-error component and correlated forecast errors (see Dee and da Silva (1999)). In a more realistic application the prescribed covariance matrix \mathbf{S} would be more complex, e.g. due to the presence of multiple observing systems with different accuracies, and possibly flow-dependent background-error covariances. However, this simple example will suffice for the purpose of illustrating the basic behaviour of the adaptive buddy check.

In the first example we simulate actual residuals from the distribution

$$\mathbf{v} \sim \mathcal{N}(\mathbf{b}, \sigma^2 \mathbf{S}), \tag{24}$$

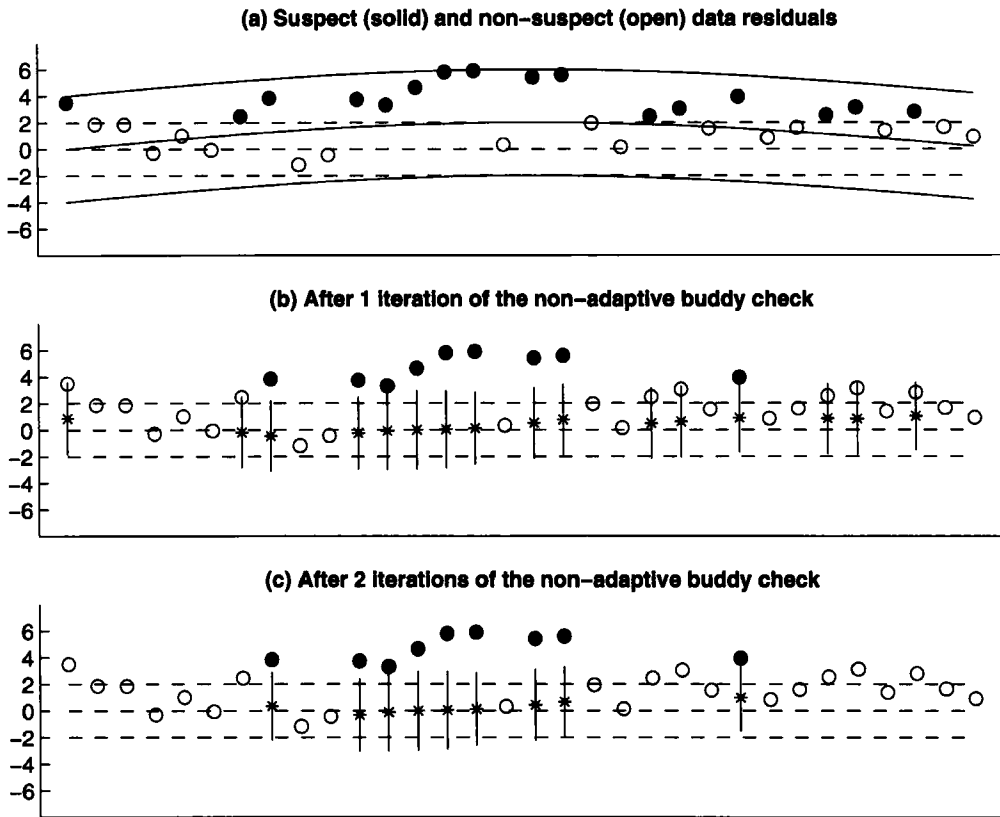


Figure 2. Illustration of the performance of the non-adaptive buddy check with prescribed tolerances, for (a) a contrived example in which the prescribed error statistics underestimate the actual errors (see text), (b) after one iteration and (c) after two iterations. The solid curves in (a) show the actual expected range for the residuals. The dashed curves indicate the tolerances used by the algorithm, which are based on the prescribed statistics.

with

$$b_j = 2 \sin \frac{\pi j}{32}, \tag{25}$$

$$\sigma = 2. \tag{26}$$

Here \mathbf{b} represents a bias in the residual, and σ is a noise amplification factor. Both \mathbf{b} and σ are unknown to the algorithm. This type of situation can easily arise at a time when forecast errors are spatially coherent and larger than usual in some region. The challenge for a buddy check is then to recognize that all observations are genuine, even though many of the residuals may be much larger than expected.

Figure 2 illustrates what happens when the non-adaptive algorithm ($m^* \gg m$) is applied, using $\tau_b = 2$ and $\tau = 3$. The circles in Fig. 2(a) mark a single realization of (24) produced with a random-number generator. The solid curves indicate the actual unconditional expected range $b_j \pm 2\sigma$ for the majority (about 95.5%) of residuals. The dashed horizontal lines show the unconditional range 0 ± 2 anticipated by the analysis system; all residuals that are outside this range (indicated by solid circles in the figure) are initially considered suspect.

Figure 2(b) summarizes the situation at the end of the first iteration of the algorithm. For each suspect residual x_i , an asterisk marks the conditional expectation x_i^* given

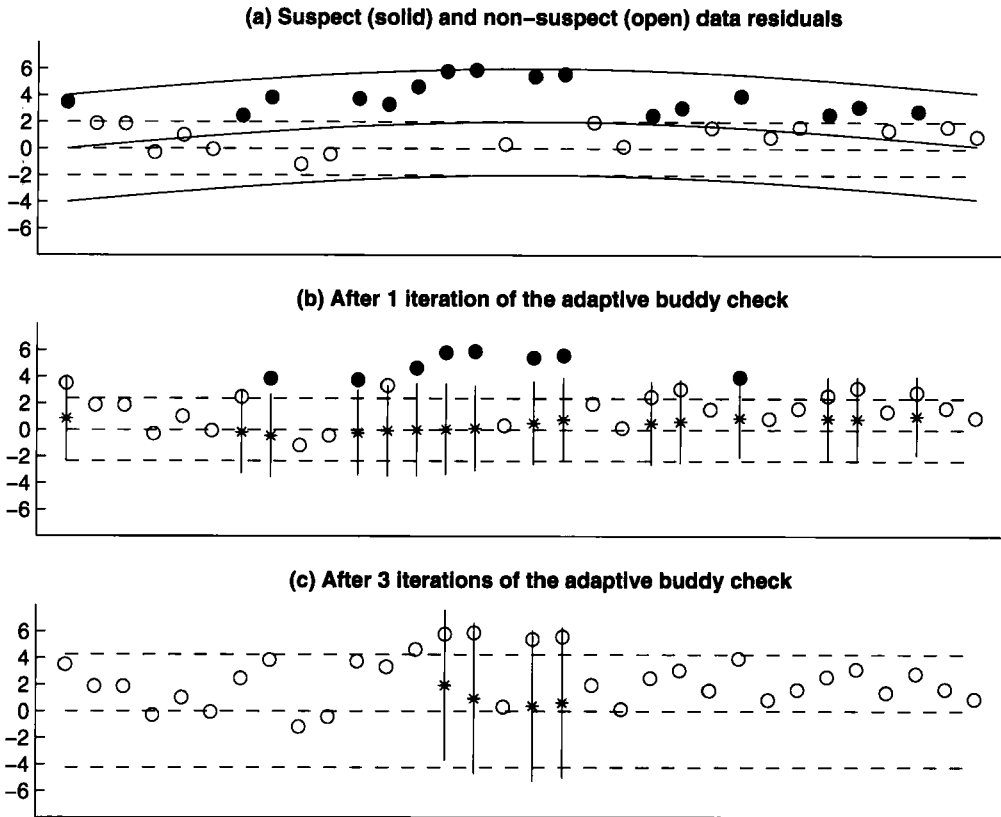


Figure 3. As Fig. 2, but for the adaptive buddy check and with (c) after three iterations. The tolerances (indicated by the dashed curves) expand during the iterations to accommodate the actual variability of the data.

all non-suspect data, and the vertical bars show the conditional range $x_i^* \pm 3\sqrt{S_{ii}^*}$. Three residuals are found to lie within this range, and are therefore no longer considered suspect. After updating the set of buddies, a second iteration of the algorithm does not change the status of any of the remaining suspects. The algorithm then terminates. All observations marked by solid circles in Fig. 2(c) end up failing the buddy check.

Figure 3 summarizes the performance of the adaptive algorithm ($m^* = 0$), for the same set of residuals and using the same prescribed covariance information and tolerance parameters. Figure 3(b) shows that, already after the first iteration, more of the suspect residuals lie within the conditional range. The reason is that the algorithm senses that the prescribed bounds are too conservative, based on the variability of the initial non-suspects. The dashed horizontal lines now indicate the adjusted range $0 \pm 2\alpha$, which has expanded slightly since $\alpha > 1$. Subsequent iterations allow the remaining suspect residuals to be confirmed by the growing set of buddies. Figure 3(c) shows that, in fact, all residuals end up passing the buddy check.

A second example illustrates the opposite situation when the actual variability of the data residuals is smaller than that implied by the prescribed error statistics. This can happen, for example, when the prevailing flow is calm, and background and observation errors are smaller than usual. In that case the adaptive buddy check will become stricter rather than more lenient, as shown in Fig. 4. In this example we set $b = 0$ and $\sigma = 0.5$ in (24). The solid lines in (a), as before, indicate the actual range $0 \pm 2\sigma$, which is

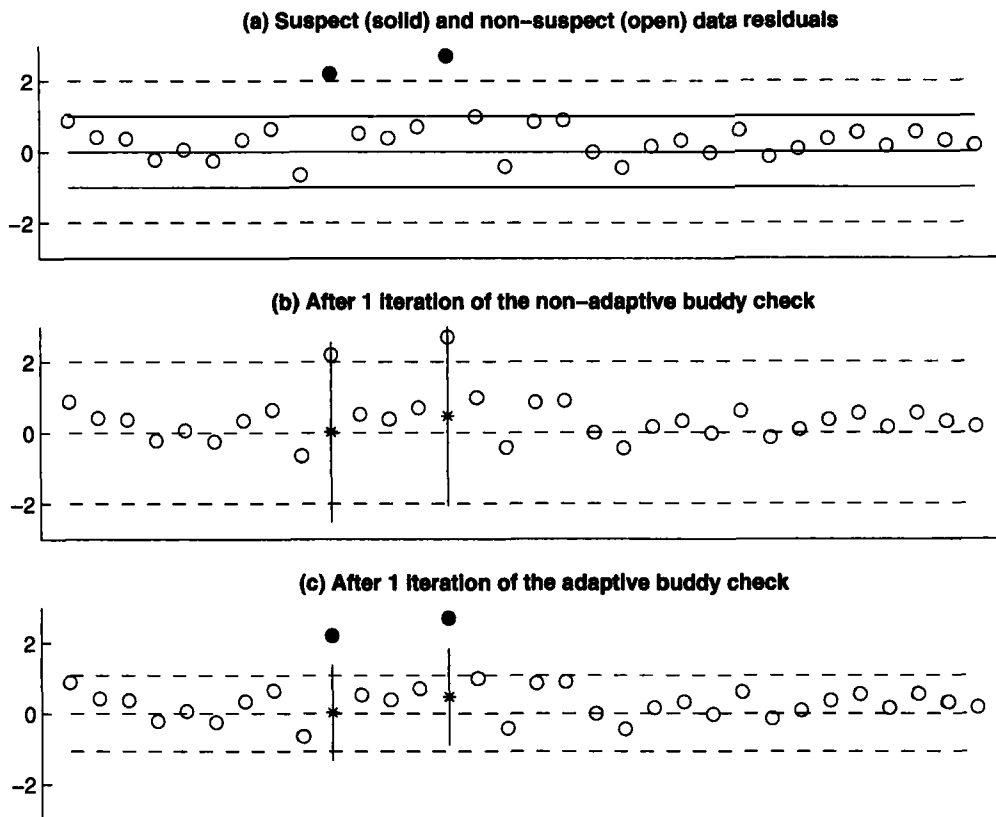


Figure 4. Illustration of the performance of (b) the non-adaptive and (c) the adaptive buddy checks after one iteration for an example in which the prescribed error statistics overestimate the actual errors. (See text for details of (a)).

now smaller than the range 0 ± 2 implied by the prescribed statistics, indicated by the dashed curves. The two residuals initially marked suspect (solid discs in (a)) correspond to observations that were purposely corrupted. They are both accepted by the non-adaptive buddy check, because the discrepancies are compatible with the prescribed statistics. The adaptive buddy check, on the other hand, estimates a reduced variance using the non-suspect residuals, and accordingly adjusts the tolerances. As a result the two erroneous observations are rejected.

3. IMPLEMENTATION IN GEOS DAS

The GEOS DAS version 3 produces global atmospheric datasets at 6-hourly intervals on a $1^\circ \times 1^\circ$ latitude-longitude grid and on 48 vertical levels. The core of the system consists of an atmospheric general-circulation model (Takacs and Suarez 1996), a physical-space statistical analysis system (Cohn *et al.* 1998), the statistical quality control (SQC) described here, and various interface functions. Apart from conventional observations, the system accepts geopotential heights retrieved from TIROS* Operational Vertical Sounder (TOVS) data, cloud-drift wind retrievals, and Special Sensor

* Television Infrared Observation Satellite.

Microwave/Imager (SSM/I) surface winds and total precipitable water. The final, assimilated data products are obtained from the analysed fields by means of the incremental analysis update procedure (Bloom *et al.* 1996). We will not dwell on the details of the DAS implementation, except as they pertain to the performance of the quality control, since many aspects of the system are undergoing rapid development.

The SQC is invoked just before computing the global analysis, and therefore represents the final line of defence against the inclusion of contaminated data. All observations have passed various sanity checks and other preliminary quality-control procedures by the time they are presented to the SQC, and some may have been flagged as suspect in the process. Input for the SQC consists of observed-minus-forecast residuals and a preliminary estimate of their variances, derived from prescribed error statistics for the global analysis system. Observation-error standard deviations for most GEOS DAS data types were estimated using maximum-likelihood techniques (Dee and da Silva 1999; Dee *et al.* 1999). Forecast-error standard deviations currently used in the GEOS DAS are global, spatially variable estimates based on monthly statistics of rawinsonde and TOVS observed-minus-forecast height residuals (DAO 1996).

(a) *The background check*

The SQC first performs a simple background check of all observations against the 6-hour forecast, as in (A.0). This test does not actually reject any observations, but marks as suspect all residuals v_i for which

$$v_i^2 > \tau_b^2 \{(\sigma_i^o)^2 + (\sigma_i^f)^2\}. \quad (27)$$

Here σ_i^o and σ_i^f are the prescribed observation- and forecast-error standard deviations, respectively, appropriate for the observation type and location.

The rate at which the background check flags data provides a useful consistency check on the prescribed statistics. Equation (12) predicts that normal, independently distributed residuals should fail the background check at an average rate of about 4.5% when $\tau_b = 2$, which is the value currently used in the GEOS DAS. Actual rates will vary even with correctly specified statistics, because residuals are neither normal nor independently distributed. Consistently large deviations from the predicted rate suggest a problem with either the prescribed error standard deviations or with other assumptions about errors incorporated into the analysis system. Intermittently large deviations may indicate transitional problems with the DAS, such as an exceptionally poor forecast or a breakdown of the observing system. Background-check results should be continuously monitored for each observing system and data type, broken down by region and vertical level.

Table 1 summarizes the SQC monitoring for the GEOS DAS during January 2000. The symbols z , u , v , q , and p_{sl} in the second column stand for the analysed quantities geopotential height, zonal wind, meridional wind, water vapour mixing ratio, and sea-level pressure, respectively. The first numeric column shows the number of scalar observations presented to the SQC for each data type (counting each wind vector report as two observations). The second numeric column shows the background check rates, in per cent. The final column contains the rejection rates for the buddy check, which we discuss in the next section.

None of the actual background check rates shown in Table 1 exactly match the ideal rate. Sea-level-pressure data from surface stations are flagged at a higher rate because the distribution of residuals for that data type is slightly peaked, with thick tails relative to the normal distribution. The tails primarily result from the use of (extrapolated) sea-level-pressure data over topography. The low rate for TOVS height retrievals appears

TABLE 1. SUMMARY OF GEOS DAS STATISTICAL QUALITY CONTROL MONITORING FOR JANUARY 2000

Source	Type	Data count	Background check	Buddy check
			(per cent)	
Rawinsondes	(z)	466 903	6.07	1.25
	(u, v)	822 632	6.34	1.90
	(q)	197 582	6.83	1.15
TOVS	(z)	6 271 375	2.93	0.61
Aircraft (sat. relay)	(u, v)	381 488	6.92	0.59
Aircraft (manual)	(u, v)	256 648	10.05	3.29
Pilot balloons	(u, v)	147 172	5.56	2.02
Cloud drift	(u, v)	1 061 248	4.03	1.07
Surface stations	(p _{sl})	790 857	9.41	1.22
Marine stations	(p _{sl})	42 851	8.65	0.04
	(u, v)	66 906	9.42	0.87
Ships	(p _{sl})	88 669	2.90	1.28
	(u, v)	171 280	4.45	3.51
Fixed buoys	(p _{sl})	79 073	5.62	0.32
	(u, v)	151 124	5.34	1.69
Drifting buoys	(p _{sl})	124 632	5.24	2.41
	(u, v)	35 154	4.07	2.46
SSM/I	(u, v)	4 638 572	0.27	0.04

For explanation of 'Type' and numeric columns, see text.

to be due to an earlier removal of outliers during the retrieval process. It may also result from a statistical dependence between retrieval errors and forecast errors, as well as a systematic overestimation by the analysis system of stratospheric forecast errors. The exceedingly low rate for SSM/I winds indicates that the prescribed observation-error standard deviations for this data type are too high and need to be adjusted.

(b) *The buddy check*

Following the background check, the SQC implements an iterative and adaptive buddy-check algorithm similar to that described in section 2(d). During each iteration, each suspect observation is tested separately against a restricted set of buddies. This set is constructed by first locating all observations of the same variable within a suitably large three-dimensional domain centred at the location of the suspect observation. These candidate buddies are then ranked according to the weights they would receive in an optimal statistical analysis at the location of the suspect observation (based on prescribed error statistics). This leads to the selection of a limited number of *best buddies*. These are then used to predict the value of the suspect observation as in (A.2) and to estimate the scaling factor for the buddy check limits as in (A.4). Depending on the result of the test (A.5), the suspect observation is marked for reacceptance. At the end of the iteration, after all suspect observations have been tested in this manner, the entire process is repeated, unless no additional observations are reaccepted.

This particular scheme for the selection of buddies does not depend on a predefined domain decomposition. It has the advantage that the sets of observations used to test two nearby suspect observations overlap. The horizontal and vertical extent of the search area for buddies is defined in terms of the de-correlation length-scales associated with the background errors, such that any observations outside the search area would receive negligible weight in predicting the value of the suspect observation. Each buddy check is local, and uses all available statistical information about observation and forecast errors. The inflation (or deflation) factor α for the tolerance is computed separately for

each suspect observation from its buddies, i.e. based on local data only, and recomputed during each iteration.

We introduced several approximations in the operational implementation of the algorithm. The most significant of these is that the buddy check is univariate, in the sense that only data of the same type (but, if available, from different instruments) are used to test each suspect observation. In a multivariate check, confirmation of an extremely low sea-level-pressure observation, for example, might be found in nearby cyclonic wind observations. A second simplification is that the local analysis performed in each iteration of the buddy check uses a single step of the successive corrections method (Daley 1991, chapter 3) rather than an optimal statistical interpolation. It would be more elegant to call the analysis component of the DAS to solve (A.2) in each instance, but that is currently not practical.

In the current configuration of the GEOS DAS the algorithm is allowed to be fully adaptive ($m^* = 0$) and uses $\tau = 3$ for the buddy-check tolerances for all observing systems (no information about prior probabilities of gross errors is prescribed). The maximum number of best buddies is set to 25. Wind vector data are excluded when either of the components fail the buddy check. Similarly, an entire TOVS height profile is excluded if any single height residual from that profile fails the buddy check.

The last column in Table 1 shows the average rejection rates for each of the data types assimilated in the GEOS DAS during January 2000. The rates vary significantly among the different data types. For example, the scheme is approximately 30 times less likely to reject sea-level-pressure observations from stationary marine platforms than reports from ships. Compare this with Lorenc and Hammon (1988), who used historical information on the reliabilities of different types of ship data to configure and test their Bayesian quality-control method. Similarly, our method rejects aircraft wind observations that are automatically relayed much less frequently than those reported verbally. In a statistical sense, therefore, the method is able to differentiate between observing systems with different reliabilities, without the benefit of a priori probabilities of gross error for these systems.

The computational cost of the SQC is typically less than 2% of the total cost of a global analysis. The main portion is expended during the first iteration of the algorithm. The majority (typically 85–90%) of observations that are initially flagged as suspect are reaccepted after the first iteration. The cost can be further reduced by increasing the tolerance parameter for the background check, which would result in a smaller pool of initial suspects. Our current choice of $\tau_b = 2$ is conservative and could probably be increased without significantly changing the final result of the buddy check after convergence. The total number of iterations of the algorithm needed for convergence is typically between three and six.

In spite of the approximations, the buddy check as implemented in the SQC retains the main features described in section 2: it is based on a local analysis of nearby data which is both adaptive and iterative. In analogy with the simple contrived example presented in section 2(e), we illustrate the performance of the SQC using the following case-study.

(c) Storm of 27 December 1999

Two severe storms hit Europe in succession at 06 UTC on 26 December 1999 and at 18 UTC on 27 December 1999, causing significant damage and a great deal of media attention (Bell *et al.* 2000). Both storms were poorly predicted by many weather services, in spite of the fact that they were well observed by a variety of observing systems. Preliminary indications are that the forecast problems were largely due to inadequate

data-assimilation procedures, since many models were able to predict the storms several days ahead but lost track of them in subsequent short-range forecasts (P. Undén, personal communication; see also the *Official SRNWP/EUCOS Report on the December 1999 Storms*, available on the internet at <http://srnwp.sma.ch/workshops/FinalReport.html>). Because of background-error covariance specifications that are inappropriate for extreme situations, the available observations were interpreted incorrectly and, in some cases, were excluded from the analysis altogether. At Météo France the medium-range forecasts of the storms were reasonably good, but a large number of crucial surface observations were not assimilated when the second storm hit the coast. These observations were excluded by a simple statistical background check because the model first-guess became very poor (J.-N. Thépaut, personal communication).

To illustrate this problem, we show in Fig. 5 the result of an experiment with the GEOS DAS in which the adaptive feature of the buddy check is switched off. The three panels in the figure show the development of the second storm in a sequence of sea-level-pressure analyses. Each panel contains the locations of all available pressure observations that took place within a 6-hour window centred at the analysis time (some locations correspond to multiple observations). The colour green indicates that the observations at that location passed the background check, yellow means that at least one observation failed the background check but subsequently passed the buddy check, and red means that at least one observation failed the buddy check and therefore did not enter the analysis. Not shown are the variety of near-surface wind observations that were analysed as well, originating from ships, buoys, and SSM/I retrievals. Both the background check and the buddy check in this case are strictly based on prescribed statistics.

At 12 UTC and at 18 UTC the background check flagged a large number of observations, at a rate much higher than usual because of the poor 6-hour forecast. A fair number of these ultimately did pass the buddy check, but the most crucial observations in the vicinity of the depression were rejected. The analysis at 12 UTC shows a weak low of 989.3 hPa located over the Celtic Sea. The actual low is further to the south-west, in the vicinity of 8°W, 48°N, where an entire cluster of observations was rejected by the buddy check. The lowest three in this cluster averaged 974.2 hPa, which is about 17 hPa below the analysed minimum. At 18 UTC the depression over France is too weak by about 10 hPa; the analysed low there is 978.8 hPa, while the three lowest observations at that time averaged only 968.8 hPa.

Figure 6 shows the GEOS DAS analyses obtained by using the adaptive buddy check. Almost all of the observations that were flagged by the background check were ultimately allowed into the analysis. The definition of the storm is now much better. The analysed low at the height of the storm is 973.6 hPa, which is 5.2 hPa deeper as a result of the additional observations. The first-guess low was 980.2 hPa, about 4.5 hPa deeper due to the improved initial conditions at 12 UTC. Although the model still has difficulty capturing the storm in its full strength, the inclusion of the available observations in this case clearly improves the assimilation.

We now take a close look at the observations that failed the buddy check. The single rejected observation at 12 UTC, located near the centre of the depression, is a ship report of 879.8 hPa, which is clearly in error. On the other hand, for the 18 UTC analysis a report of 971.3 hPa from a surface station on the French coast was rejected, even though it was probably accurate. Three successive reports were issued from that station, valid at 17 UTC (971.3 hPa), 18 UTC (976.8 hPa), and 19 UTC (982.8 hPa). These values are consistent with a rapid inland movement of the storm. The first report was rejected by the buddy check because it differed greatly from the first-guess value of 983.6 hPa

and could not be confirmed by surrounding observations (including the two later reports from the same station).

For similar reasons, seven sea-level-pressure ship reports were excluded from the 06 UTC analysis, all of which were probably accurate. These observations, whose locations are marked by the four red discs in the top panel of Fig. 6, all took place toward the end of the 6-hour analysis window and showed a drop in sea-level pressure of about 10 hPa compared with reports just a few hours earlier. In addition, there were some late wind reports from nearby buoys (not shown) that indicated a change to easterly winds associated with a developing cyclonic circulation. A multivariate buddy check might have taken advantage of this information. If these data had been included in the analysis, the model perhaps would have detected the developing depression at an earlier stage.

The underlying problem here is that it is not possible to represent accurately rapidly moving storm systems when all observations within a 6-hour time window are treated as if they occurred simultaneously. This is clearly a shortcoming of the assimilation system. Work is well underway to reduce the length of the analysis window in the GEOS DAS.

The most important practical advantage of the adaptive buddy check is that the final quality-control decisions are fairly robust with respect to the prescribed forecast- and observation-error statistics, at least in well-observed situations. To demonstrate this we changed the tolerance parameter for the background check, which has the effect of modifying the initial pool of suspects, similar to what would be the case if different error statistics were prescribed. The case just discussed uses $\tau_b = 2$, which results in 57 suspects at 06 UTC, 153 at 12 UTC, and 186 at 18 UTC. We repeated the quality control with $\tau_b = 3$, which reduced the number of initial suspects to 10, 84, and 111, respectively. However, the final result of the buddy check was unchanged for all observations.

4. CONCLUSION

We have presented an adaptive statistical quality-control algorithm with flow-dependent tolerances for outlier observations. The algorithm is similar to a conventional buddy check in that it tests suspect observations against local statistical analyses based on nearby data, using information about background and observation errors provided by the global statistical analysis system. The adaptive aspect was designed in order that the final quality-control decisions depend more on the data and less on prescribed statistics. This is an important practical advantage, since the information about background and observation errors in atmospheric data-assimilation systems tends to be least accurate in situations that are difficult to forecast, which is precisely when quality-control decisions have their largest potential impact.

We illustrated the performance of the method with several examples. We first used contrived observations, drawn from error distributions that are very different from that initially presumed, to contrast the behaviour of the adaptive buddy check with that of a buddy check which relies on prescribed statistics only. These examples show that the adaptive mechanism can result in either increased or decreased leniency of the quality control, depending on the situation. We also presented a realistic case-study based on the GEOS DAS analysis of an extreme storm event. Here the application of an iterative buddy check with fixed tolerances led to the exclusion of many important observations. The adaptive algorithm, on the other hand, was able to adjust the tolerances and bring in the great majority of the observations, resulting in an improved storm analysis.

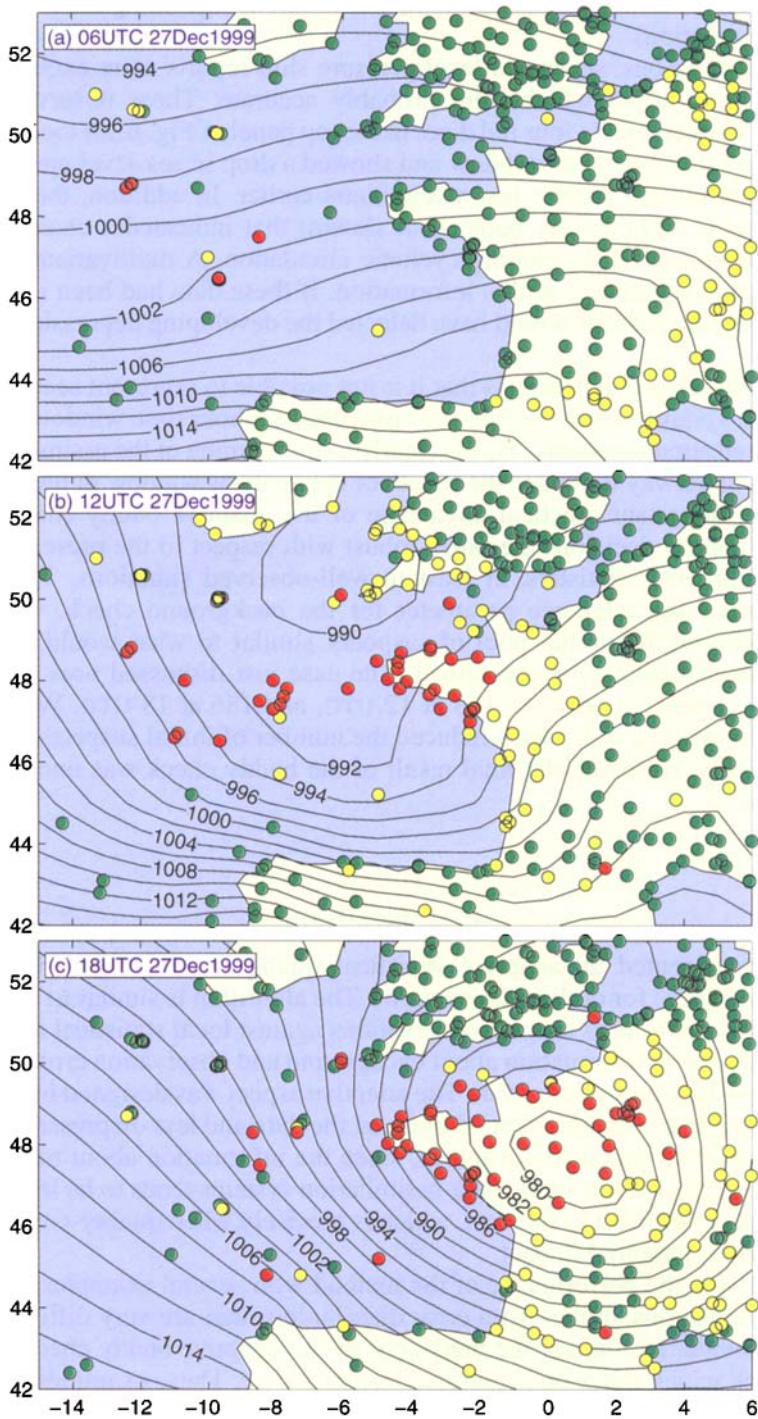


Figure 5. Successive GEOS surface analyses for (a) 06, (b) 12, and (c) 18 UTC showing the development of the 27 December 1999 storm, obtained when using a buddy check whose tolerances are based on prescribed statistics. Red discs mark the locations of sea-level-pressure observations that were rejected by the buddy check, yellow discs correspond to observations that failed the background check but subsequently passed the buddy check, and green discs correspond to those that passed the background check.

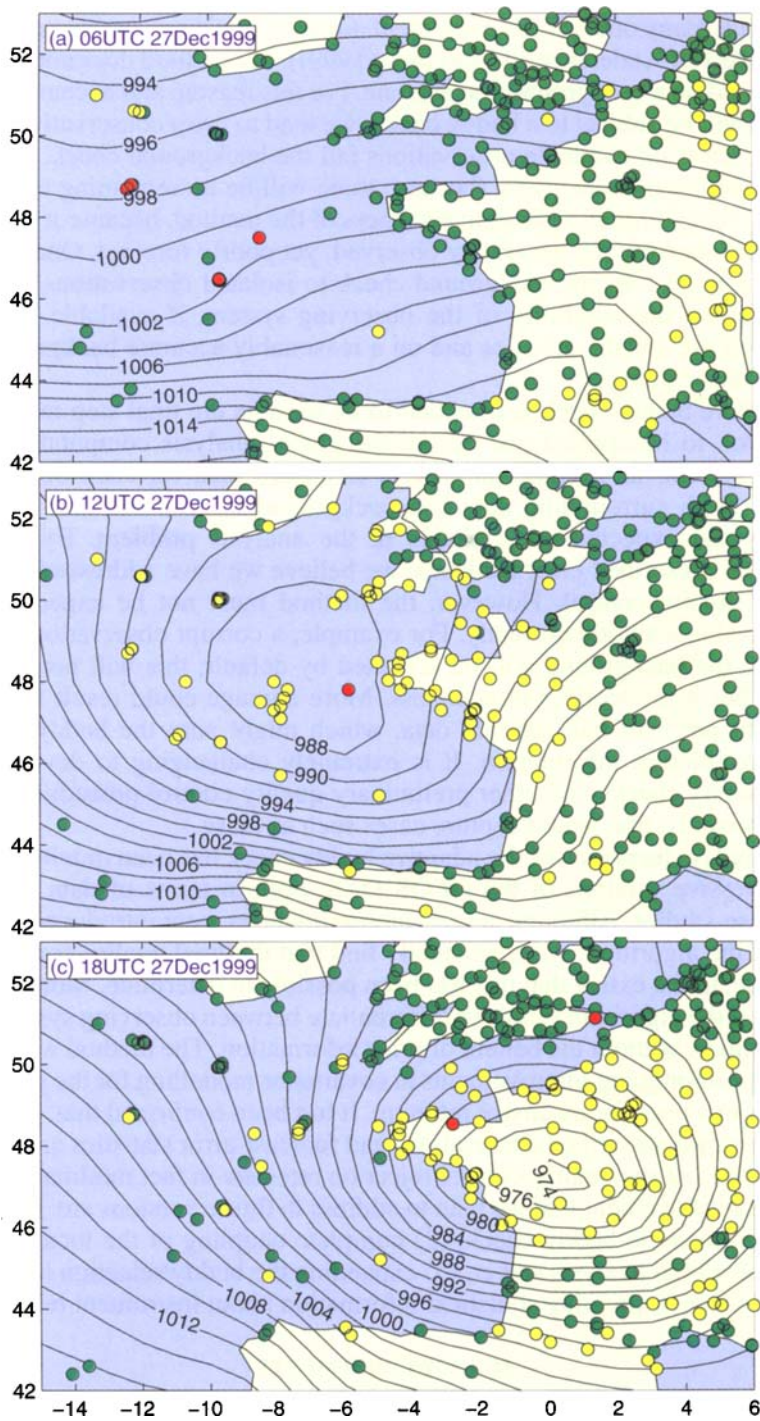


Figure 6. As Fig. 5, but using an adaptive buddy check.

The remaining dependence on prescribed statistics in our method is primarily through the background check, used for the initial identification of suspect observations. In contrast with many operational implementations of statistical quality control (see, for example, Table 1 of Andersson and Järvinen (1999)), our method does not reject any observations based on a background check alone. For this reason, and because only flagged observations will be subject to a buddy check, we tend to use a conservative background check. However, if *all* available observations fail the background check, then the algorithm will reject them collectively (because there will be no remaining observations to test against). This points to a possible weakness of the method, because it is conceivable that an event is sparsely but accurately observed, yet poorly forecast. One could fix this by applying a more tolerant background check to isolated observations, and by using information about the reliability of the observing system, if available. However, the dependence on prescribed statistics and on a reasonably accurate background estimate is inevitable in such cases.

The adaptive buddy check is intended to be used as the final step in observational quality control, to be applied just before the global analysis computation in a data-assimilation system. Its specific purpose is to ensure that, for each observation, any disagreement with surrounding data and background information can be reasonably explained by the expected uncertainties in the analysis problem. By reducing the dependence on prescribed error statistics, we believe we have addressed a major issue in statistical quality control. However, the method must not be expected to handle every conceivable situation correctly. For example, a corrupt observation that happens to agree with the background will be accepted by default; this will not greatly affect the analysis but it is incorrect nonetheless. More damage could result from a locally self-consistent patch of bad satellite data, which might pass the buddy check in the absence of conflicting information. It is extremely challenging to develop a system of effective sanity checks and other preliminary quality-control procedures, tailored to specific instruments, that would capture cases such as these.

Operational performance of the adaptive buddy check has been monitored since late 1998 in successive versions of the GEOS DAS, both in terms of data counts and in individual case-studies. Although several approximations were introduced in the implementation of the algorithm, we consistently find that the final quality-control decisions are reasonable, to the extent that this has been possible to determine. Monitoring results has shown that the algorithm is able to differentiate between observing systems with different reliabilities, without the benefit of prior information. The method was designed to take advantage of ongoing improvements in covariance modelling for the global analysis system, without a need for extensive retuning. It has been confirmed that, as the analysis system evolves and prescribed observation- and forecast-error statistics are adjusted, the resulting changes of the quality-control rejection rates are in fact minimal.

Areas for improvement that we plan to address in future versions are: (1) implementation of a multivariate buddy check, (2) complete coupling of the local buddy-check analysis to the global analysis solver, (3) expanding the buddy selection to a larger time window, and (4) incorporating statistical information about instrument reliability.

ACKNOWLEDGEMENTS

We thank the reviewers for their critical reading of the manuscript and for raising several important issues. Thanks also to Gerard Cats for first suggesting this application of adaptive error estimation, to Jim Stobie for constructive criticism, and to Jean-Noël Thépaut for pointing to the role of quality control in the analysis of the 27 December 1999 storm.

REFERENCES

- Andersson, B. D. O. and Moore, J. B. 1979 *Optimal filtering*. Prentice-Hall, Englewood Cliffs
- Andersson, E. and Järvinen, H. 1999 Variational quality control. *Q. J. R. Meteorol. Soc.*, **125**, 697–722
- Bediend, H. A. and Cressman, G. P. 1957 An experiment in automatic data processing. *Mon. Weather Rev.*, **85**, 333–340
- Bell, G. D., Halpert, M. S., Schnell, R. C., Higgins, R. W., Lawrimore, J., Kousky, V. E., Tinker, R., Thiaw, W., Chelliah, M. and Artusa, A. 2000 Climate assessment for 1999. *Bull. Am. Meteorol. Soc.*, **81**, S1–S50
- Berghórsson, P. and Döös, B. R. 1955 Numerical weather map analysis. *Tellus*, **7**, 329–340
- Bloom, S. C., Takacs, L. L., da Silva, A. M. and Ledvina, D. 1996 Data assimilation using incremental analysis updates. *Mon. Weather Rev.*, **124**, 1256–1271
- Cohn, S. E., da Silva, A., Guo, J., Sienkiewicz, M. and Lamich, D. 1998 Assessing the effects of data selection with the DAO physical-space statistical analysis system. *Mon. Weather Rev.*, **126**, 2913–2926
- Collins, W. G. 1998 Complex quality control of significant level rawinsonde temperatures. *J. Atmos. Oceanic Technol.*, **15**, 69–79
- Collins, W. G. and Gandin, L. S. 1990 Comprehensive hydrostatic quality control at the National Meteorological Center. *Mon. Weather Rev.*, **118**, 2752–2767
- Daley, R. 1991 *Atmospheric data analysis*. Cambridge University Press, Cambridge
- DAO 1996 'Algorithm theoretical basis document version 1.01'. Data Assimilation Office, NASA/Goddard Space Flight Center, Greenbelt, MD 20771, USA
- Dee, D. P. 1995 On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Weather Rev.*, **123**, 1128–1145
- Dee, D. P. and da Silva, A. M. 1999 Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology. *Mon. Weather Rev.*, **124**, 1822–1834
- Dee, D. P., Gaspari, G., Redder, C., Rukhovets, L. and da Silva, A. M. 1999 Maximum-likelihood estimation of forecast and observation error covariance parameters. Part II: Applications. *Mon. Weather Rev.*, **124**, 1835–1849
- Dharssi, I., Lorenc, A. C. and Ingleby, N. B. 1992 Treatment of gross errors using maximum probability theory. *Q. J. R. Meteorol. Soc.*, **118**, 1017–1036
- Gandin, L. S. 1988 Complex quality control of meteorological observations. *Mon. Weather Rev.*, **116**, 1137–1156
- Ingleby, N. B. and Lorenc, A. C. 1993 Bayesian quality control using multivariate normal distributions. *Q. J. R. Meteorol. Soc.*, **119**, 1195–1225
- Jazwinski, A. H. 1970 *Stochastic processes and filtering theory*. Academic Press, New York
- Lehmann, E. L. 1997 *Testing statistical hypotheses*. Springer, New York
- Lorenc, A. C. 1981 A global three-dimensional multivariate statistical interpolation scheme. *Mon. Weather Rev.*, **109**, 701–721
- 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
- Staff Members, Joint Numerical Weather Prediction Unit 1957 One year of operational numerical weather prediction. *Bull. Am. Meteorol. Soc.*, **38**, 263–268
- von Storch, H. and Zwiers, F. W. 1998 *Statistical analysis in climate research*. Cambridge University Press, Cambridge
- Takacs, L. L. and Suarez, M. J. 1996 'Dynamical aspects of climate simulations using the GEOS general circulation model'. NASA Technical Memorandum 104606, **10**
- Tarantola, A. 1987 *Inverse problem theory*. Elsevier, Amsterdam
- Woollen, J. S. 1991 'New NMC operational OI quality control'. Pp. 24–27 in Proceedings of the 9th conference on numerical weather prediction, Denver, USA. American Meteorological Society