The probability distribution of observed minus expected sea-level pressures; a quality control perspective

R.S. Seaman
Centre for Australian Weather and Climate Research, Melbourne, Australia

(Manuscript received August 2007; revised January 2008)

Observed minus expected (O-E) sea-level pressures from Australian synoptic weather observations (SYNOPs), together with a synoptic assessment of which large (O-E) values correspond to good or bad observations, indicate that the (O-E) probability distribution for good observations is heavier tailed than in a normal (Gaussian) distribution. This is particularly so when the expected value is obtained using cross-validated increments from a prediction model 6 h forecast, but it is also true if the expected value is simply a 6 h forecast. This heavy tailed characteristic implies that quality control of such observations may be more difficult than would be the case for a Gaussian distribution.

A mixture of two Gaussian distributions, corresponding to good observations, together with a flat distribution, corresponding to bad observations, appears to provide a reasonable and parsimonious representation of observed (O-E) values.

Stratifying the data by means of a predicted space-mean pressure gradient indicates that large (O-E) values are more likely when the gradient is strong, which in turn suggests that a rejection tolerance, or observational de-weighting strategy, should vary accordingly.

The principles illustrated in the paper appear applicable to other types of observations and data assimilation systems.

The quality control problem; overview

Many meteorologists will recall situations in which the use of an incorrect observation has degraded a subsequent forecast. Quality control of observations is therefore an essential ingredient, both of synoptic analysis and forecasting, and of numerical data assimilation and prediction.

Corresponding author address: R.S. Seaman, Centre for Australian Weather and Climate Research, GPO Box 1289, Melbourne, Vic. 3001, Australia.
Email: r.seaman@bom.gov.au

To decide whether an observation is good or bad, it needs to be compared with an expected value. If the observation is 'too far' from the expected value, it may either be discarded, or accorded a lesser weight. An expected value is problem dependent. Quality control of meteorological data often uses considerations of spatial and temporal consistency – for example a prediction from an earlier time, or consistency with neighbours. What is 'too far' depends upon the statistical properties of observed minus expected (O-E) values.

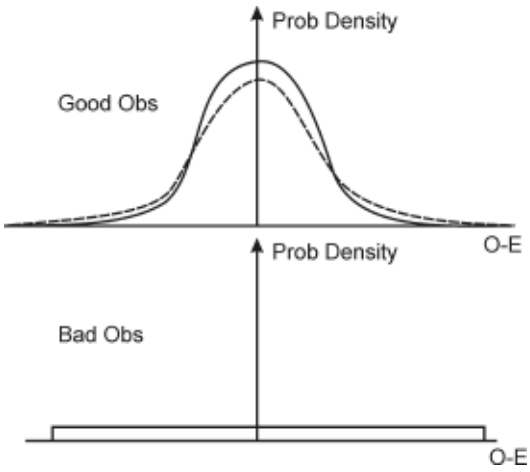
Quality control of SYNOP sea-level pressures is the focus of this paper. Observed Australian sea-level pressures have been compared with two different sets of expected values, namely (a) a 6 h first guess from the Bureau of Meteorology's global data assimilation system (GASP; National Meteorological and Oceanographic Centre, 2006), and (b) a spatial consistency ('cross-validation') estimate using (O-E) values at neighbouring stations.

Given that the purpose of quality control is to detect 'bad' observations, it seems reasonable to consider (O-E) values as coming from a mixed distribution of good and bad values, with many more good than bad. Under this assumption, a simple conceptual framework for the quality control problem, following the ideas of Lorenc and Hammon (1988), is shown by the two solid curves in Fig. 1. The (O-E)s for good observations have a probability distribution function (pdf) which is bell shaped (but not necessarily Gaussian). The (O-E)s for bad observations have a flat distribution. It is also assumed that completely absurd (as opposed to merely doubtful) observations have already been eliminated. The sum of the areas under the two solid curves in Fig. 1 is unity, and the plot of the total distribution of (O-E), for good and bad observations combined, may be obtained by graphically adding the two separate plots. From a quality control perspective, the critical (O-E) values are in the tails of the pdf for good observations, where the ordinates for good and bad observations are about equal. An observed (O-E) in this range is about equally likely to correspond to a good or bad observation, and it appears reasonable to locate a quality control tolerance here. See Lorenc and Hammon (1988) for a further discussion of this aspect, using Bayes' Theorem.

The contrast between the solid and dashed curves in the top half of Fig. 1 shows the importance of whether the distribution of good observations is close to Gaussian (solid), or heavy tailed (dashed). In the case of a heavy tailed distribution, the critical point where the ordinates for good and bad distributions are equal is less well defined than it would be for a Gaussian distribution.

The focus in what follows is therefore upon the tails, rather than upon the centre, of the total (O-E) distribution, that is, upon the unusual rather than the usual observations, which are either potentially valuable, or wrong. An implication is that any non-Gaussian characteristic in the (O-E) pdf for good observations may be quite important. Specifically, the usual criterion of two or three standard deviations characterising an outlier and perhaps a reject, should be viewed with scepticism. Both Devenyi and Schlattler (1994) and Ingleby and Lorenc (1992) suggest that pdfs for quantities analogous to the (O-E)s for good observations here may be heavier tailed than in a Gaussian distribution.

Fig. 1 **A schematic representation of probability distribution functions (pdfs) of (O-E) for good and bad observations. An (O-E) value is equally likely to be good or bad where the ordinates of the two pdfs are equal. After Lorenc and Hammon (1988). See text for further explanation.**



How does one verify a quality control decision ?

In operational practice, borderline quality control decisions are made all the time, but they may be difficult to verify individually, as there is no objective ground truth. One can assess whether, using alternative quality control methodologies, the predictions from one set of analyses are better or worse than predictions from an alternative set. But such an assessment doesn't answer the question of whether a particular observation was in error.

The latter question may be addressed by asking whether, in each individual case, a skilful manual analyst would have drawn to, or 'paid', the observation. This approach recognises that a manual analyst can provide a somewhat independent view, via synoptic analysis skills. Therefore at least it provides a basis, however imperfect, for assessing individual quality control decisions. One particular feature of manual synoptic analysis that distinguishes it from the GASP data assimilation is that a manual analyst would usually consider observations of full fields (pressures in hPa etc.), while the GASP data assimilation and quality control is based upon normalised increments from a background field. Moreover, a manual analyst can take into account known local

effects in the orography, and phenomena such as land and sea-breezes. The shortcomings of this approach are, firstly, its subjectivity and, secondly, the likelihood of many doubtful cases. Nevertheless, it is the approach adopted in this paper.

The following sections will (a) describe the basic data and its processing, (b) present the pdf of the (O-E) data, (c) fit an analytic functional representation, (d) discuss an important covariate (pressure gradient), and, finally, relate the preceding sections to the conceptual framework for quality control.

Data and methods

Australian six-hourly SYNOP data were extracted for a period of 12 months (October 2005 to September 2006). In order to calculate (O-E), these data were compared with (a) 6 h forecasts (first guesses) from the Bureau's GASP system, run at spectral truncation T239, and (b) a cross-validation estimate using increments from the first guess at neighbouring stations. Previous work (Seaman 1999) has shown that (b) usually provides a better estimate than (a), as would be expected because (b) takes into account spatial consistency with neighbours. The version of GASP used also differed slightly from that used by Bureau operations, the most important difference being that here the assimilation component was constrained to use all Australian SYNOPs (the operational version combines closely located SYNOPs). In addition, the system was run with 29 rather than 33 vertical levels.

Because (O-E) is characteristically larger in magnitude in mid-latitudes than in the tropics, and because systematic biases may be present, normalised values of (O-E) are used throughout. In a current half-month, the observed minus expected pressures $(O\text{-}E)_i$ at station i are normalised according to

$$(O\text{-}E)_{in} = ((O\text{-}E)_i - (O\text{-}E)_{im}) / s_{im} \qquad \ldots 1$$

where $(O\text{-}E)_{in}$ is the normalised observed minus expected pressure, $(O\text{-}E)_{im}$ is the station-specific mean over the preceding two half-months, and s_{im} is the station-specific standard deviation over the preceding two half-months.

The pdfs of observed minus expected pressures

Table 1 shows the numbers and percentages of occurrences of normalised (O-E) in various standard deviation ranges, when the expected value is a cross-validation estimate (column 2), or a 6 h first guess (column 4). Also shown (column 6) are percentages of occurrences in each range that would be expected in a Gaussian distribution. Remember that the percentages in columns 2 and 4 correspond to observations that can be either good or bad. Focussing upon the tails of the total (O-E) distribution, it can be seen that nearly all of the observations over, say, three or four standard deviations would need to be 'bad' in order for the distribution of 'good' observations to be Gaussian.

For either choice of expected value there are, in absolute terms, non-negligible counts of (O-E) values in the over six standard deviations range, namely 307 in the cross-validation case and 43 in the first guess case. The larger number in the cross-validation case is explained by the fact that normalisation is performed on observed minus expected values. The expected values, and hence the normalisation, will be different for the cross-validation and first guess cases. The percentages of the total cases in the over six standard deviation range are admittedly very small (order .01 to .05%), but nevertheless it is relevant to inquire whether all these 'outliers' are bad observations. Despite the acknowledged shortcomings of such an approach (see earlier), an attempt was made to synoptically assess these outliers as probably correct, probably wrong, or simply

Table 1. Numbers (N) and percentages of (O-E) occurrences in standard deviation ranges, when expected values were cross-validation and 6 h first guess estimates. Percentages expected in a Gaussian distribution are also shown.

	Cross-validation		*First guess*		*Gaussian*
(O E) Sdev range	*N*	*%*	*N*	*%*	*%*
< 0.5	235656	40.12	226078	38.48	38.29
0.5 – 1.0	172931	29.44	174704	29.73	29.97
1.0 – 2.0	142559	24.27	154825	26.35	27.18
2.0 – 3.0	28243	4.80	27224	4.63	4.28
3.0 – 4.0	5751	0.98	3916	0.67	0.26
4.0 – 5.0	1487	0.25	672	0.11	0.01
5.0 – 6.0	484	0.08	123	0.02	0.00
> 6.0	307	0.05	43	0.01	0.00

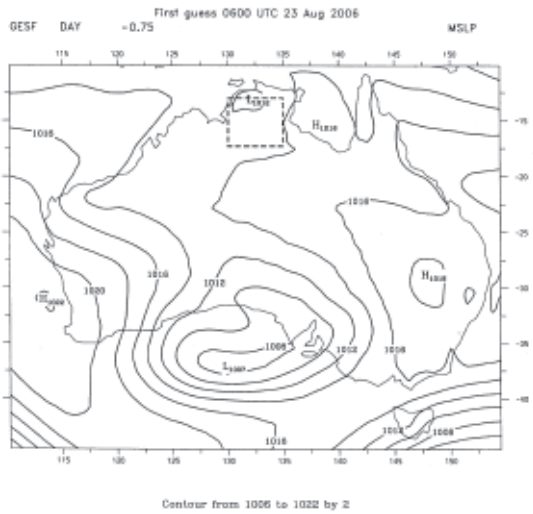
doubtful, on the basis of whether a manual analyst (in this paper, the author) would have paid the observation. As mentioned earlier, the synoptic assessments were based on observations of the full fields (not increments). This is in contrast to the outliers which, as previously discussed, were based upon normalised increments from an expected value.

As expected, there were many doubtful cases (about 20%), but approximately equal numbers of cases (about 40%) were assessed as probably correct and probably wrong. While it is impractical to show all the relevant data used in the synoptic assessments, a few examples that give the general flavour are shown in Figs 2(b), 2(d), 2(f) and 2(h). The outliers, all corresponding to 'over six standard deviation' cases in Table 1, are underlined. The corresponding 6 h first guesses in Figs 2(a), 2(c), 2(e) and 2(g), over large areas, indicate the general synoptic situations, while the same first guesses, with a smaller contour interval, over smaller areas surrounding the outlier are also shown in Figs 2(b), 2(d), 2(f) and 2(h). The deviations, in hPa, from the first guesses (rounded standard deviations in brackets) for the four outliers were respectively 2.5(10), 2.8(7), 2.4(6) and 2.5(9). Cases (b), (d), (f) and (h) were assessed as wrong, doubtful, correct and wrong respectively. Case (d) may be one of representativeness error (correct but small scale), while in case (f) the guess field appears to be poor in the vicinity of the trough over southeastern Australia. Typically, the observed pressures in all the figures show considerable spatial structure in their deviations from the 6 h first guesses; this underlines the fact that (O-E) values are not only influenced by observational errors, but also by spatially correlated errors in the first guesses. Of course, systematic errors (that is, (O-E) values that tend to be consistently positive, or negative, on most days) are taken into account by the normalisation described earlier.

While the preceding assessments may be debatable, nevertheless if one accepts the author's overall judgement that some of the over six standard deviation cases are actually good observations, it follows that the pdf for good observations must be heavy tailed, as data outside six standard deviations essentially never occur in a truly Gaussian distribution. This conclusion, that the (O-E) pdf for good observations of sea-level pressure is heavy tailed, appears consistent with the work of Ingleby and Lorenc (1992), and Seaman (2002), both of whom suggest a mixture of Gaussian distributions may be more appropriate than a single Gaussian distribution.

Fig. 2 GASP 6 h first guesses for times (a) 0600 UTC 23/8/06, (c) 0000 UTC 7/9/06, (e) 1200 UTC 19/9/06 and (g) 0000 UTC 22/9/06, with corresponding windows (b), (d), (f) and (h) showing the outlier (underlined) and other observations (hPa). Latitudes and longitudes are shown along the window edges, and dashed boxes in Figs (a), (c), (e) and (g) show locations of the windows.

(a)



(b)

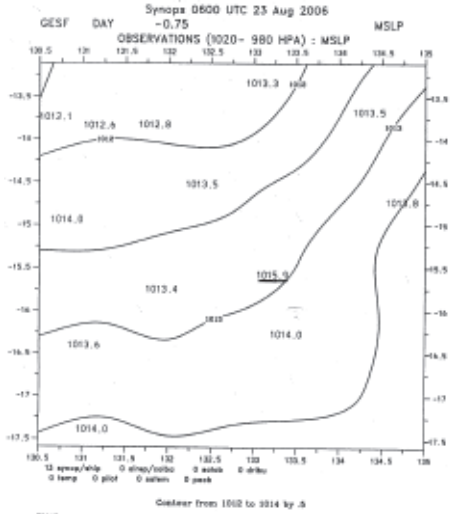
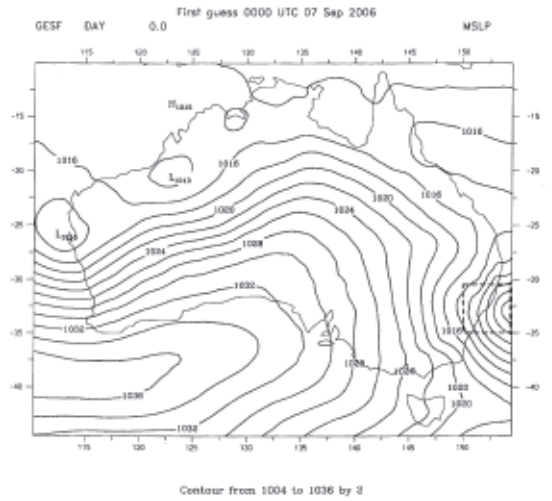
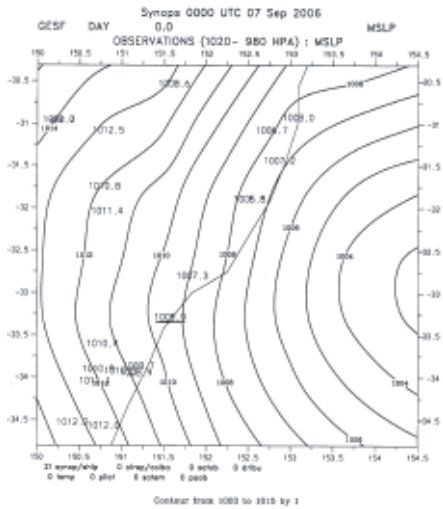


Fig. 2 Continued.

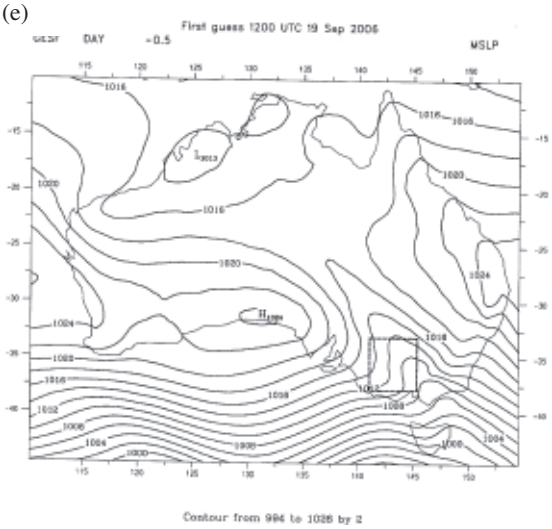
(c)



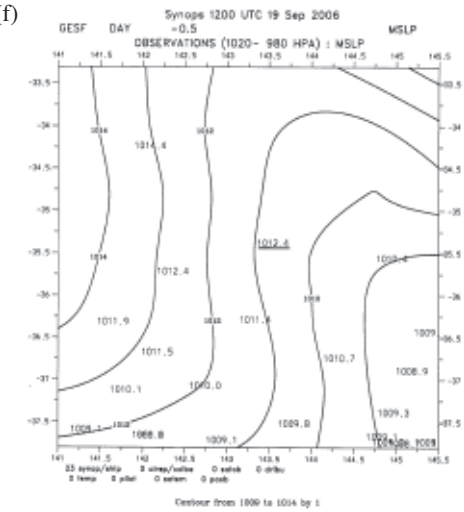
(d)



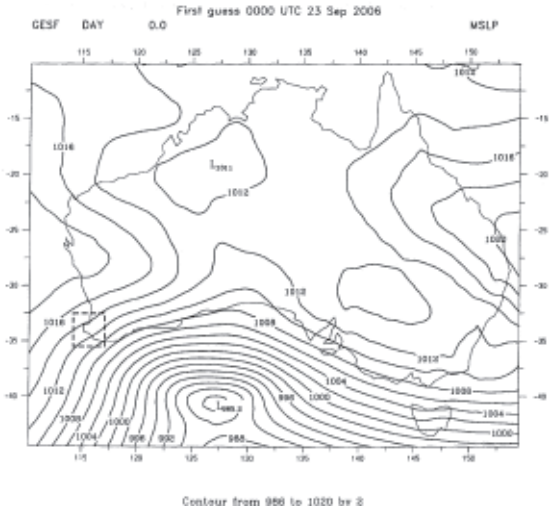
(e)



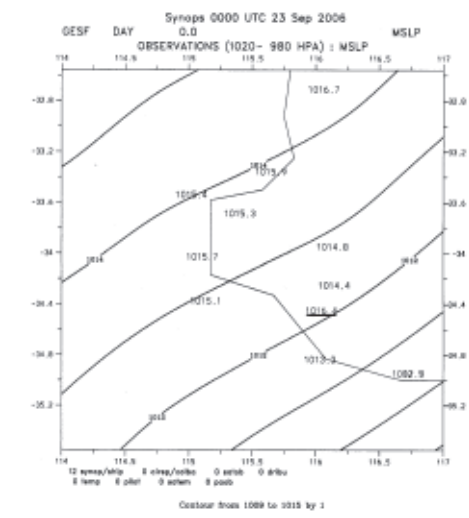
(f)



(g)



(h)



Fitting an analytic function

Having established that the distribution of (O-E) values is probably heavier-tailed than Gaussian, an obvious next step is to further quantify the shape of that distribution in terms of Fig. 1. As discussed in the opening section, a comparison of the upper and lower frames of Fig. 1, using the dashed curve in the upper frame, will give an indication of where to locate a quality control tolerance. A simple representation of the (O-E) distribution was therefore attempted using an equal mixture of two Gaussian distributions, with standard deviations in the ratio '1 : *sdratio*', together with a flat distribution of bad observations comprising a proportion '*pbad*' of the entire (O-E) population. This representation therefore has two free parameters *sdratio* and *pbad*. The abscissa bounds of the flat distribution for bad observations correspond to plus and minus ten standard deviations (the gross bounds check in GASP).

Optimal values of the free parameters are determined by direct search in the relevant two-dimensional parameter space. Using the 24 half-month batches of (O-E) data, referred to in the third section, the optimal value of the parameter *sdratio* had a median value of about 1.25, with a variability between batches from around 1.0 to 1.4. Similarly the parameter *pbad* had a median value of about .010, with a variability between batches from .005 to .050.

Clearly there is scope for a more complex parametrisation than that described above; these results should be regarded as indicative orders of magnitude. In particular, one could attempt to allow for asymmetry (skewness) in the (O-E) distribution.

An important covariate; the space-mean pressure gradient

In the course of synoptically assessing the over six standard deviation cases, it was noticed that such cases appeared to occur more often when the background field pressure gradients in the vicinity of the relevant observations (space-mean pressure gradients) were strong. See the Appendix for details of how the space-mean pressure gradient, and its cumulative distribution were calculated. This subjective conclusion was subsequently confirmed quantitatively, as indicated by Table 2. Those (O-E) values of more than about three standard deviations are clearly associated with space-mean pressure gradients above the median, as is evident from the columns headed 'ratio'. Such an effect is only weakly evident in the correlation coefficients between the magnitudes of (O-E) and the space-mean gradients which are only of order 0.1, as the vast majority of (O-E) values are less than three standard deviations. This result exemplifies the benefit of considering only 'large' deviations from the mean when testing apparent associations (e.g Shahani 1969).

It is fairly easy to explain synoptically why those large (O-E) values that are associated with good observations should occur preferentially in strong gradient situations; in such circumstances a small phase or location error in a synoptic feature of a background field can result in a large (O-E) value (in hPa terms). However, it is less obvious why, or indeed whether, large (O-E) values associated with bad observations should also occur preferentially in strong gradients, and further investigation of this aspect is needed.

Implications for quality control

As with all binary (yes/no) decisions, it is necessary to trade off the probability of detection (POD) of bad observations, against the false alarm rate (FAR - rejection of good observations). The heavy tailed distribution of (O-E) for good observations, discussed in the section before last, is a particularly relevant consideration. What it means is that (a) even if a rejection toler-

Table 2. Numbers (N) of (O-E) occurrences in standard deviation ranges, cross-tabulated against space-mean gradients above and below the median, when expected values were cross-validation and 6 h first guess estimates. Ratios above/below shown.

	Cross-validation			*First guess*		
(0-E)	*space-mean gradient*			*space-mean gradient*		
Sdev range	*> median N*	*< median N*	*Ratio*	*> median N*	*< median N*	*Ratio*
< 0.5	108269	127387	0.85	105440	120638	0.87
0.5–1.0	81951	90980	0.90	82734	91970	0.90
1.0–2.0	72727	69832	1.04	77454	77371	1.00
2.0–3.0	16252	11991	1.36	15981	11243	1.42
3.0–4.0	3812	1939	1.97	2544	1372	1.85
4.0–5.0	1075	412	2.61	470	202	2.33
5.0–6.0	365	119	3.08	100	23	4.35
> 6.0	242	65	3.72	35	8	4.38

ance was set to (say) six standard deviations, some good observations will still be wrongly rejected, and (b) the critical point in Fig. 1, at which the pdf curves for good and bad observations intersect (that is, have equal ordinates), is rather poorly defined. In other words, the heavy tailed characteristic of the (O-E) pdf for good observations makes the quality control task more difficult than it would be if the pdf were truly Gaussian, resulting in a lesser POD and/or higher FAR.

The presence of a covarying space-mean pressure gradient has the obvious implication that the rejection tolerance should depend on the covariate, and a wider tolerance should be specified in strong gradient situations. A reasonable criterion might be to vary the tolerance in such a way that the rejection rate is independent of the gradient. Tables that are similar but more detailed than Table 2 could provide the information needed to specify such a criterion.

Not all assimilation systems use a binary strategy for quality control. Arguments corresponding to those in the preceding two paragraphs apply if, rather than an 'all or nothing' accept/reject strategy, one simply gives doubtful observations less weight than others.

Concluding remarks

The main conclusions of this study have been (a) confirmation of the heavy-tailed character of the (O-E) distribution of Australian sea-level pressures, and (b) the significant covariation of (O-E) with the pressure gradient. The preceding results should be regarded as illustrative of some things that can be learned, relevant to quality control, from the study of observed minus expected values. Obviously the specific quantitative results will be dependent upon the element under consideration, and upon the assimilation and prediction model configurations used to derive the expected values. But the principles illustrated here should still apply.

The use of covariates, like pressure gradient in the present paper, to modulate quality control tolerances, deserves particular attention. Indices of synoptic activity such as time tendencies of relevant variables, might also be considered. The use of covariates is conceptually similar to adaptively changing the background field error variance during the evolution of an ongoing assimilation (e.g. Parrett 1993). If the latter can be achieved realistically by four-dimensional variational methods, or by Kalman filtering, there may be no need to do any more by empirical statistical means, as in the present paper. This is clearly a question to be resolved by further research.

Acknowledgments

Thanks are extended to the reviewers and Associate Editor Graham Mills, whose presentational suggestions significantly improved the paper. Brett Harris, Peter Steinle and Blair Trewin provided useful comments on an earlier versions of the manuscript.

References

Devenyi, D. and Schlatter, T.W. 1994. Statistical properties of three hour prediction "errors" derived from a mesoscale analysis and prediction system. *Mon. Weath. Rev., 122,* 1263-80.

Ingleby, N.B. and Lorenc, A.C. 1992. Forecast errors and buddy checks. *Preprints, 12th Conference in Probability and Statistics in the Atmospheric Sciences*, Toronto, Am. Met. Soc., 207-13.

Lorenc, A.C. and Hammon, O. 1988. Objective quality control of observations using Bayesian methods. Theory and a practical implementation. *Q. Jl R. met. Soc., 114,* 515-43.

National Meteorological and Oceanographic Centre. 2006. Operational upgrade of GASP. *Analysis and Prediction Operations Bulletin No. 62.*

Parrett, C.A. 1993. An operational trial of synoptic-dependent background errors. *Forecasting Research Technical Report No. 67,* The Met Office.

Seaman R.S. 1999. Quality control of Australian sea level pressure observations. *Aust. Met. Mag., 48,* 123-31.

Seaman, R.S. 2002. The relevance of Benford's Law to background field errors in data assimilation. *Aust. Met. Mag., 51,* 25-33.

Shahani, A.K. 1969. A simple graphical test of association for large samples. *Appl. Stat., 18*, 185-90.

Appendix

Calculation of the space-mean pressure gradient

The spectral coefficients of daily GASP 6 h first guess fields are post-processed to a regular 2.5 by 2.5 degrees latitude-longitude grid of sea-level pressure, and values of pressure gradient are evaluated on this grid by standard finite differencing. The space-mean pressure gradient, either at a grid-point or at an observing point, is calculated by a spatially weighted average of the gridded pressure gradient values with a scan radius R of 500 km, where the weighting function of separation d (< 500 km) takes the form

$$(R*R-d*d)/(R*R+d*d).$$

The daily space-mean grid-point values are used to calculate the cumulative distribution of space-mean gradient at every grid-point. The space-mean value at an observing point at any particular time is located within the spatially interpolated cumulative distribution.