QUALITY CONTROL AND SELECTION ALGORITHMS IN OPERATIONAL DATA ASSIMILATION

1. PRE-PROCESSING AND PRE-SELECTION OF DATA

New quality control procedures require a much more comprehensive interface between the Reports Data Base (RDB) and the Analysis System. Some years ago it was sufficient for the analysis to be supplied with the basic observed variables and the position of the observation. As the quality control and pre-processing becomes more sophisticated the analysis needs to know much more about the observation to produce a good judgement.

1.1 General considerations

Problems in the use of specific data-types are discussed below. Here it is worthwhile to mention some general considerations. To improve the resolution of an analysis has the inevitable corollary that the whole world becomes more data sparse, and so it is necessary to rely more heavily on other aids to quality control besides the accuracy of the first guess. Prior information about the quality and reliability of individual components of the observing system is particularly important.

There is a need for time continuity checks on all data sources for which this is possible or meaningful, particularly SYNOPS , SHIPS and Aircraft. If possible, aids to this work such as flight-plans for aircraft, or route-plans for ships should be made available. The recently installed hardware at ECMWF, such as the Solid State Disc and the Mass-Storage System, are ideal for these applications.

1.2 Pre-processing and pre-selection of TEMP data

There are several aspects of TEMP data which require further investigation - bias correction, cross-checking of standard and special level data, quality control of special level data, and pre-selection decisions in mountainous terrain.

In the light of the Japanese effort it would be worthwhile to re-evaluate the checks on TEMP data. In particular the checking of special level data needs re-examination. It is desirable that information be exchanged between the major centres on the techniques in current use, and that an improved set of algorithms should be developed.

If a good set of algorithms were available as a standard software package, many of the developing NWP centres would find them useful.

The current quality control of TEMPS (and other types) make no reference to the first guess estimate. Given the high accuracy of this first guess, this accuracy should be fully exploited in quality control algorithms at the RDB level.

The ECMWF is close to implementing a system designed for bias corrections for radiosonde data. The remaining questions concern the stability in time of the corrections, the value or otherwise of correcting stations on the basis of sonde type or on an individual basis, the value of identifying sub-groups of stations which will not be modified, the necessity for an accurate specification of the atmospheric tides and the necessity of separating model and instrumental error. The documentation and exchange of information on the bias corrections between centres would be a valuable cross-check.

1.3 Pre-selection decisions for SATEM and SATOB

The pre-selection of SATEM data over land seems to vary from organisation to organisation. Some centres use the data extensively over land, and others not at all. An objective basis for clarifying the quality of the data over land is needed. Study of the question at ECMWF seems to suggest that the data is good above 500 mb, but further studies are necessary.

Similar studies of the behaviour of upper level winds over land suggest that earlier problems with this data have been circumvented.

1.4 Use of SATOB data

Comparison of the SATOB observations obtained from the four different geostationary satellites (METEOSAT, GOES-E , GOES-W and HIMAWARI) indicates that the quality of the observations has much improved during the past years and differences in the quality of the data between the different satellites have decreased significantly. However, collocation studies [Morgan] and the study of mean difference from the guess-field [Delsol] indicate that there remain large bias errors of high level winds in the subtropical jets.

Attempts should be made to correct such bias error in the observation by either correcting the observations themselves or incorporating bias error in the optimum interpolation scheme itself. Methods to make such bias corrections require further study. Also contact should be kept with the data producers so that the cause of such bias error can be investigated and ways found to reduce such errors by the use of additional information on tracer clouds (e.g. cloud depth) or by other means.

1.5 Use of SATEM data

The first priority is to investigate the information content of the SATEM reports currently available at 14 levels. It is noted that the information content can be fairly easily estimated by the use of empirical orthogonal function analysis.

Observational error characteristics of SATEM data needs to be further examined. For example, the error characteristics are known to vary with orbit; this is also related to the limb correction.

1.6 Direct use of radiance data in optimum interpolation

The direct use of radiance data in the optimum interpolation scheme should be regarded as an important project. The basic approach is to use the differences of the first guess and the observed radiances. For this process, the method of diagnosing the radiance from the first guess field must be studied, although there is some evidence that this problem has been solved (provided the water content is known).

It is also important to investigate the number of channels to be used for the OI scheme. The information content of the 20 channels of radiance data needs to be investigated. Condensing the information by the use of empirical orthogonal functions is also worthwhile.

1.7 Use of other SATEM information

In some SATEM reports, we have information about precipitable water content and cloudiness. Effort should be put into studying how this information can be used for the analysis of humidity.

1.8 Use of PAOBS and other bogus data

The availability of moisture and typhoon bogus data from JMA and NESDIS, and global cloudiness data (cloud top brightness temperature data from Meteosat is available on the GTS) should be investigated. Cloud data of this type is of value for identifying areas of intense convection for the humidity analysis and for the initialisation of physical fields.

The error characteristics of such data and of the Australian PAOBS should be determined. Quality flags on the PAOBS would be useful. The possibility of horizontal error correlation needs particular attention.

2. DEVELOPMENTS IN QUALITY CONTROL

The theoretical and practical means are now available to put quality control algorithms on a sound statistical basis, rather than using empirically tuned rejection criteria. This should be of benefit in cases where there is only partial information redundancy, and checks give probabilistic rather than binary (right/wrong) results. In order to do this it is necessary to collect statistical information about:-

- the accuracy and error structure of the forecast background
- the accuracy of good observations
- the probability of each type of gross error in an observation, and the distribution of observed values for erroneous observations.

These quantities are all amenable to collection from databases associated with operational systems and so this work should continue with high priority. Quality control is more sensitive to these statistics even than statistical (optimum interpolation) grid-point analysis, and in empirically tuned quality control systems it has been found desirable to make criteria flow-dependent. Therefore the flow-dependence of the above statistics should be investigated, so that at least for quality control this can be taken into account.

Many different quality control checks are done at different stages of the data-base and analysis suite; the probabilistic approach provides a means for combining the results from these in a logical way. Consequently the probability of error should be carried, rather than a simple right/wrong flag for each check, to enable this to be done.

Data checking is highly sensitive to the data density. This in turn makes the selection of data checking algorithm important and so there should be an investigation of the definition of data density zones, e.g. dense, medium and sparse.

In a data dense region a "buddy" check is probably sufficient to identify erroneous observations. The use of the buddy check is, however, limited by the atmospheric variability.

For the data sparse regions there probably is no appropriate analysis checking procedure. In these regions an accurate time consistency check is imperative.

In medium density regions, the ECMWF optimum interpolation checking would result in most erroneous data being flagged. The sensitivity of this check to the statistics in general, and in particular to the assumptions about Gaussian statistics should be examined in real data cases. In addition, the importance of the volume of data included in the checking should be investigated. A balance needs to be found between the statistical formulation, the error statistics, the data density and the analysis resolution.

Except in extreme non-linear cases, such as a ship observation in a fjord, quality control should be independent of the resolution of the analysis for which the observation is to be used. Errors of representativeness should be accounted for in the weighting of observations, not in the quality control.

A more detailed effort to compare the quality control algorithms used in the major centres should be undertaken. The exchange of information on rejection rates may be of value. The Centre should undertake a thorough review of all aspects of the quality control algorithms used in the operational suite. It is recognised that the quality control of humidity data may require special techniques.

3. DATA SELECTION ALGORITHMS

The data selection algorithms depend on whether one is using a box or grid point analysis technique, so the relative merits of both methods need to be considered.

3.1 Box techniques and grid point analysis

Before a definite answer to the question whether to use a box technique or a grid point analysis can be formulated, experimentation on the box size should be done. The optimal box size, the real prediction error correlation decay length and the correlation length actually used are closely interconnected. Reasonable a priori statements on the box size can only be made under the assumption that the used prediction error correlation is appropriate for the meteorological situation.

Basic experiments should concentrate on areas with a large gradient in data density; from an economic point of view it may be worthwhile to give attention to data dense areas. Most likely a single gridpoint analysis will generate more three-dimensional noise and less balance than the box technique. But a box analysis generates power in the wave with wave-length equal to the box size - especially if the real prediction error has large scales. It is not clear how this box size wave should be suppressed.

Overall it appears that the box technique is preferable to a single grid-point formulation: it has better overall noise and balance properties , and is more economical in data sparse areas(with respect to grid resolution). The reduction in noise and increase in balance with respect to the grid-point scheme, however, may go together with a less tight fit to the data. A sufficient density of data and good accuracy of the first-guess error statistics are essential for the box technique.

3.2 Data selection algorithms

It is realised that in many circumstances the data selection algorithm determines the final structure of the analysis more than the OI formulation itself. Therefore existing algorithms should be improved.

(a) During data selection, data redundancy should be removed. This applies to both data quality control and analysis runs. If two data agree according to the buddy check, they should be super-obbed. A third datum can now be checked safely against the superob. It is redundant to check this third datum against the constituents of the superob separately, provided that the reliability of the superob is estimated and used in a consistent way.

Therefore superobs should be formed for all data near an analysis gridpoint, to within the analysis resolution. Also data checking and analysis should be performed on the same set of normal observations plus superobs, where superobs have an appropriately reduced rejection probability and observation error.

(b) Data selection should (again) be dependent on the real prediction error covariance. If the prediction error has power in the large scale, data should be selected to represent this large scale. This may introduce the need for a separate large scale analysis.

(c) In the not too distant future, experiments should be carried out to verify the applied data selection algorithms. It has been suggested that an analysis should be carried out with many (say 1000) data per box, and then compare it with a usual analysis (191 data). Differences between the two analyses identify weaknesses in the data selection.

Finally, note that the means for monitoring the performance of the data thinning and data selection algorithms are desirable.

4. SUMMARY OF RECOMMENDATIONS

4.1 Preprocessing and pre-selection of data

(i) Information about the quality and reliability of individual components of the observing system is required.

(ii) There is a need for time continuity checks for data, particularly SYNOPS, SHIPS and Aircraft. To aid this, flight plans for aircraft or route-plans for ships should be made available.

(iii) Checks on TEMP data should be examined, especially the checking of special level data, and information exchanged between major centres about techniques in current use. Information about bias corrections should also be exchanged.

(iv) The quality of SATEM data over land should be investigated. The information content of SATEM reports should be estimated, and the error characteristics further examined.

(v) The bias in SATOB observations should be overcome by correcting the observations or incorporating bias error in the optimum interpolation scheme itself. Producers of this data should try to identify the cause of the bias.

(vi) The direct use of radiance data in the optimum interpolation scheme should be studied, and effort put into considering how information about precipitable water and cloudiness can be used for the analysis of humidity.

(vii) The use and error characteristics of PAOBS and other bogus data should be investigated.

4.2 Developments in quality control

(i) It is necessary to collect statistical information about the accuracy and error structure of the forecast background, the accuracy of good observations, the probability of each type of gross error in an observation and the distribution of observed values for erroneous observations.

(ii) The flow dependence of the above statistics should be investigated.

(iii) Data checking should be done in terms of the probability of error rather than right/wrong decisions.

(iv) The importance of the volume of data for data checking needs to be examined.

(v) Except in extreme non-linear cases, quality control should be independent of the resolution of the analyses for which the observations are used.

(vi) Efforts should be made to compare the quality control algorithms used in the major centres.

4.3 Data selection algorithms

(i) The relative merits of box techniques and gridpoint analysis require further investigation. The optimum size of a box needs to be determined and the formulation of a variable box size technique is recommended.

(ii) During data selection, data redundancy should be removed.

(iii) Data selection should be dependent on the real prediction error covariance.

(iv) Experiments should be carried out to verify the data selection algorithms.

(v) The data thinning and data selection algorithms should be monitored.