# DATA SELECTION AND QUALITY CONTROL IN THE ECMWF ANALYSIS SYSTEM

Peter Lönnberg and Dave Shaw

European Centre for Medium Range Weather Forecasts

Reading, U.K.

## 1. INTRODUCTION

This paper describes the quality control procedures and the data selection algorithms in the ECMWF analysis system. A unique feature of the ECMWF statistical interpolation scheme is the use of large analysis volumes with horizontal dimensions of order 1000 km and vertical dimension of at least a third of the atmosphere. Physical constraints can be imposed on the analysis changes (increments) on scales enclosed by an analysis volume. However, large analysis boxes may lead to discontinuities in the increment field at the boundaries between the volumes. A necessary requirement for a smooth increment field is an extensive and homogeneous data selection. By spatially overlapping the calculation of the analysis changes, and by averaging the contributions from different boxes, discontinuities are spread out over a large area. Naturally this means that the number of data which can be considered sufficient for the analysis of a volume must be quite large, typically of the order of 100 or more.

Hollingsworth et al. (1985) demonstrated that the first-guess and the analysis of the ECMWF assimilation system have an error level which is comparable to the observation errors in data rich regions. Even in data sparse regions the six hour forecast is generally quite close to the actual atmospheric state. This property of the system makes is possible to identify both random and systematic errors in the observing systems.

Data selection and checking algorithms are required for the processing of observations and in the main analysis. The aim of the pre-analysis is to provide a representative set of data for the analysis with gross errors identified. The pre-analysis step also compresses redundant information. For the main analysis the statistical interpolation technique (OI) provides a means to check the spatial consistency of the observations.

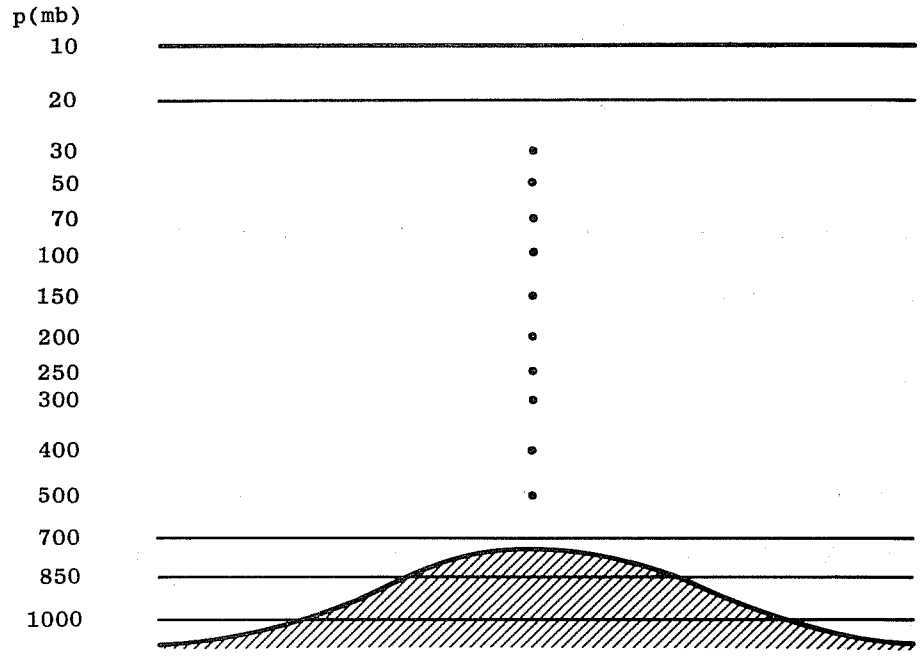
The observation processing and the pre-analysis data checking are described in Section 2. The selection algorithms for the OI check and for the calculation of the analysis increments are discussed in Section 3. The final data checking is done using the statistical interpolation scheme (Section 4).

The main features of the analysis system are summarized in Fig. 1. Lorenc (1981) describes the mathematical formulation of the ECMWF analysis system and further details can be found in the ECMWF Data Assimilation Documentation (hereafter called DAD) edited by Lönnberg and Shaw (1983).

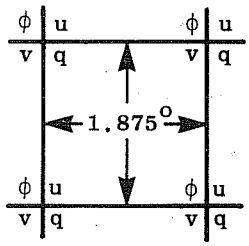
In summary, the quality of the data is checked by several procedures:-

(i) The Reports Data Base checks the formats of the reports, performs an internal consistency check, and compares the meteorological information against climatological extremes.

(ii) The pre-analysis step compares the observations against the six-hour forecast (Section 2.3).




Disposition of analysis levels in the vertical




Analysis variables and their disposition in the horizontal

| | | |
|---|---|---|
| Analysis method | Mass and wind: | 3 dimensional multi-variate (15 analysis levels, see above) |
| | Humidity | : 2 dimensional for the 5 layers up to 300 mb |
| | Surface | : Sea surface temperature from NMC analysis<br>Soil water content using rainfall observations, estimated evaporation<br>Snow depth using snow depth and snowfall observations |
| Independent variables | $\lambda, \theta, p, T$ | |
| Dependent variables | $\phi, u, v, q$ | |
| Grid | Non staggered, standard pressure levels for atmospheric variables<br>Gaussian for surface variables | |
| First guess | 6-hr forecast | |

Data assimilation frequency 6-hr ($\pm$ 3-hr time window)

Initialisation method Non-linear normal mode, 5 vertical modes, non-adiabatic

Figure 1: Schematic summary of the analysis method

(iii) The pre-analysis also looks for spatial consistency to identify data redundancy for super-observation formation (Section 2.4).

(iv) A test against an independent analysis at the observation position is done using the OI formulation (Section 4.2).

## 2. PRE-ANALYSIS

The main functions of the pre-analysis are to select a representative set of observations from all bulletins received from the Global Telecommunication System (GTS) within the six hour analysis time window, and to identify data with gross errors. The data are expressed as departures from the first-guess for a set of pressure levels or layers. For some observations an interpolation of the data to the closest analysis level is necessary. In the ECMWF assimilation system, the pre-analysis is run twice; to extract data for the mass and wind analysis, and for the humidity analysis. The mass and wind analysis makes use of height and wind data, in component form, at 15 analysis levels, and of 14 intermediate thicknesses. Information on moisture in TEMPs and SYNOPs is converted to precipitable water content for the humidity analysis, which is done for 5 layers.

The observation extraction can be regarded as a form of data selection and these aspects are described in Section 2.1. The implications of some practical limits in the analysis code on the amount of data used in the analysis are also discussed in that section. The use of asynoptic information is briefly discussed in Section 2.2. Two checks are applied in the pre-analysis phase to the data; these are described in Section 2.3. The procedure to compress information which the analysis scheme is unable to resolve is described in Section 2.4.

## 2.1 Observation extraction

Most of the observations available on the GTS are used by the ECMWF analysis system. The processing of each observation type is described in detail in DAD. In the following we describe briefly what information is extracted from each observation type. Table 1 summarizes the use of data by the ECMWF analysis system for each observation type.

For each analysis box a maximum of 80 reports can be presented to the analysis. This number is seldom reached since observation redundancy is reduced by super-obbing and by removing asynoptic observations.

**Synops and Ships**

Firstly, stations which have extrapolated their pressure measurements by more than about 800 m are excluded from further processing. The reported pressure or height is then converted to a height value at the nearest analysis level assuming that the reported temperature is representative of the extrapolation layer. If no temperature report is available then the first-guess layer mean temperature is used. Finally the height departure is the difference between the extrapolated height report and the first-guess height.

The wind reports from land stations reflect to a high degree the local topographic conditions and may be inconsistent with the 10 m winds generated by the global model. Furthermore, the model 10 m winds have been found to be too strong over land and since May 1984 no land surface winds are used in the analysis (Shaw et al., 1984).

Each ship report, excluding those that have ambiguous call signs like SHIP, RIGG and so on, are subjected to a check against the analysis (Illari,

# DATA USAGE IN ECMWF ANALYSIS SCHEME

| | P/Z | DZ | V | T | TD | WW | CLOUDS |
|---|---|---|---|---|---|---|---|
| SYNOP | X | | | ● | ● | ● | ● |
| SHIP | X | | X | ● | ● | ● | ● |
| AIRCRAFT | | | X | | | | |
| SATOB | | | X | | | | |
| DRIBU | X | | X | | | | |
| TEMP | X | | X | ● | ● | | |
| PILOT | | | X | | | | |
| SATEM | | X | | | | | |
| PAOB | X | | | | | | |

Table 1: Use of data in the ECMWF analysis system; data marked by X are used in the mass and wind analysis, and those marked by ● in the humidity analysis. WW refers to the current weather information.

pers.comm.). The departure from the analysis is compared to the combined distribution of the analysis and observation errors. The ship is blacklisted if on the majority of the reporting occasions during the last 48 hours it had a departure exceeding 3 standard deviations of the combined error. It is required that the ship has reported at least 3 times during that period. The subsequent report from this ship enters the analysis in passive mode.

The 48-hour period was chosen instead of 24 hours because the analysis can incorrectly flag data in rapidly developing synoptic situations. No distinction is made between pressure or wind exceeding 3 standard deviations; an unreliable parameter frequently means that the whole report is dubious. On average about 10-15 ships are on the blacklist.

**Dribus**

PMSL and winds from drifting buoys are extracted and presented to the analysis.

**Aircraft reports**

The aircraft wind report is moved to the nearest analysis level using the first guess wind shear.

**Cloud winds**

Cloud winds are extrapolated by the same technique as aircraft winds. However, no cloud winds are used over land due to uncertainty of their quality, particularly over high terrain (Gustafsson and Pailleux, 1981).

**Temps and Pilots**

Winds and heights for the 15 analysis levels are extracted by a 2 step procedure. First, all acceptable values on analysis levels are extracted.

Acceptable means that it has not been flagged by the Reports Data Base as incorrect. In the second step, reports from adjacent levels are used to fill gaps by linear interpolation in ln p. The second step is applied only to wind data.

**Satellite thicknesses**

Over sea, the complete satellite temperature or thickness profile is used. However, over land only the stratospheric thicknesses, i.e. above 100 mb, are presented to the analysis. The SATEM reports are partitioned to give thicknesses for consecutive non-overlapping layers. Then if any observed thickness covers two or more analysis layers, it is split into observations for single analysis layers using the first-guess temperature profile.

**Australian pseudo-observations (PAOBs)**

The use of bogus data supplied by Melbourne WMC is restricted to ocean regions south of 19°S and to surface pressure information only.

## 2.2 Use of asynoptic observations

If more than one report from a SYNOP, SHIP or DRIBU, with unique station identifier, is available within the analysis time window (± 3 hours), then only the observation closest to the analysis time is taken, provided that the reports are within 0.5 degrees of latitude and longitude of each other.

The ascribed observation error variance of an asynoptic observation is the sum of measurement error variance and the estimated persistence error variance.

## 2.3 Pre-analysis data checks

The main quality control test in the pre-analysis is a check on the departure of the observation from the six hour forecast. If the departure exceeds n standard deviations of the combined distribution of the observation and forecast error, then that piece of information is rejected; n depends on the type of variable as follows:-

| variable | n |
|---|---|
| height, thickness | 8 |
| vector wind | 5.66 |

The basic assumption of this check is that the distribution of the forecast and observation errors are normal. This assumption is reasonable for the forecast error distribution, but does not consider any gross errors in the measurements.

Hollingsworth et al. (1985) showed that gross observation errors occur quite frequently. They arise from different sources like encoding and decoding of the report, transmission problems and so on. The current ECMWF analysis quality control algorithms do not consider the possibility of such errors. The decision to accept or reject a datum depends only on the magnitude of the departure and the long-term statistics. The weight given to the datum depends only on whether it has been accepted or not. Furthermore, it is assumed that both the first-guess and the observations are unbiased. Systematic deficiencies in observing systems have been demonstrated by several investigators, e.g. Delsol (1985), Lange (1985), Hollingsworth et al. (1985) and others.

Gross errors can be included in the formalism by a Baysian approach (Lorenc, 1985). This takes into account the likelihood of gross errors and provides a different basis for rejection. Also with the Baysian approach, the weights can be a function of the magnitude of the departure.

Data sparse regions pose a special problem for quality control. A statistical approach on a single time level is inefficient in identifying observation errors. This can be understood from the distributions of the forecast and observation errors; in data sparse regions the former can be significantly larger and thus hide large observation errors. Still, careful tuning gives a possibility to eliminate extreme observation errors (Hollingsworth et al., 1985).

An additional quality control check is applied, following the first-guess check, to multi-level observations. Each variable is examined in turn, taking account of the flags already set by either the first guess check or the preceding data-base checks. If four successive levels have flags $\geq 1$ then those data are rejected. Furthermore, for geopotential and thickness, all data at levels above the levels in question are also rejected. Data most commonly rejected by this test are TEMP geopotential data in the upper troposphere and stratosphere. The limits that are applied are 4 standard deviations for height and thickness and 2.83 standard deviations for vector wind. The usefulness of the multilevel test is demonstrated by an example in Shaw et al. (1984).

Statistics on data rejections are presented in Section 4 together with numbers for the OI check.

## 2.4 Super-observation formation

The OI scheme acts as a scale dependent filter on the observations (Hollingsworth and Lönnberg, 1985). Observation densities beyond the analysis resolution may safely be compressed by forming average observations. In the ECMWF system this is done by averaging pairs of observations which are within approximately 100 km of each other. The process is repeated several times so that previously averaged observations may be combined.

Furthermore, it is required that the two observations differ from each other and from the first-guess by less than some predescribed limits. Only compatible, i.e. certain combinations of observations, data can be super-obbed.

The first requirement is that the two data should be mutually compatible, expressed by

$$(\delta_i^o - \delta_j^o)^2 \leq 2 \left(\varepsilon_i^{o2} + \varepsilon_j^{o2}\right) \qquad (1)$$

Here $\delta_i^o$, $\delta_j^o$ are the normalised departures of the observations i,j, and $\varepsilon_i$, $\varepsilon_j$ are the ascribed rms normalised observations errors. For a SYNOP pressure, with an ascribed rms error of 1 mb, this requires the difference between the two data to be less than 2 mb.

The second compatibility requirement is that each datum should have a normalised departure which is less than

$$\delta_i^{o2} \leq 9 \ (1 + \varepsilon_i^{o2}) \qquad (2)$$

For a SYNOP pressure over Europe this means that the departure must be less than 4.4 mb from the first-guess.

Of all potential super-observation pairs 25% fail to be averaged as a result of exceeding one of these two checks.

Fig. 2 gives an example of the amount of super-obbing that is performed in the pre-analysis. The most likely combinations are between two SYNOPs or a SYNOP and TEMP.

## 3. DATA SELECTION FOR OI DATA CHECK AND ANALYSIS

The technique of collectively analysing all variables for several gridpoints puts severe constraints on the data selection. It should provide a representative set of data for all gridpoints and levels and should give spatial continuity between the analysis boxes. The selection of data for the OI data check and the main analysis is divided into three steps:-

(a) Selection of influencing boxes.

(b) Selection of observations from the influencing boxes.

(c) Selection of data from the chosen observations, separately for each slab of the atmosphere.

### 3.1 Box selection

The box selection starts with the central box itself. All immediate neighbours to the central box are then selected. The selection continues out

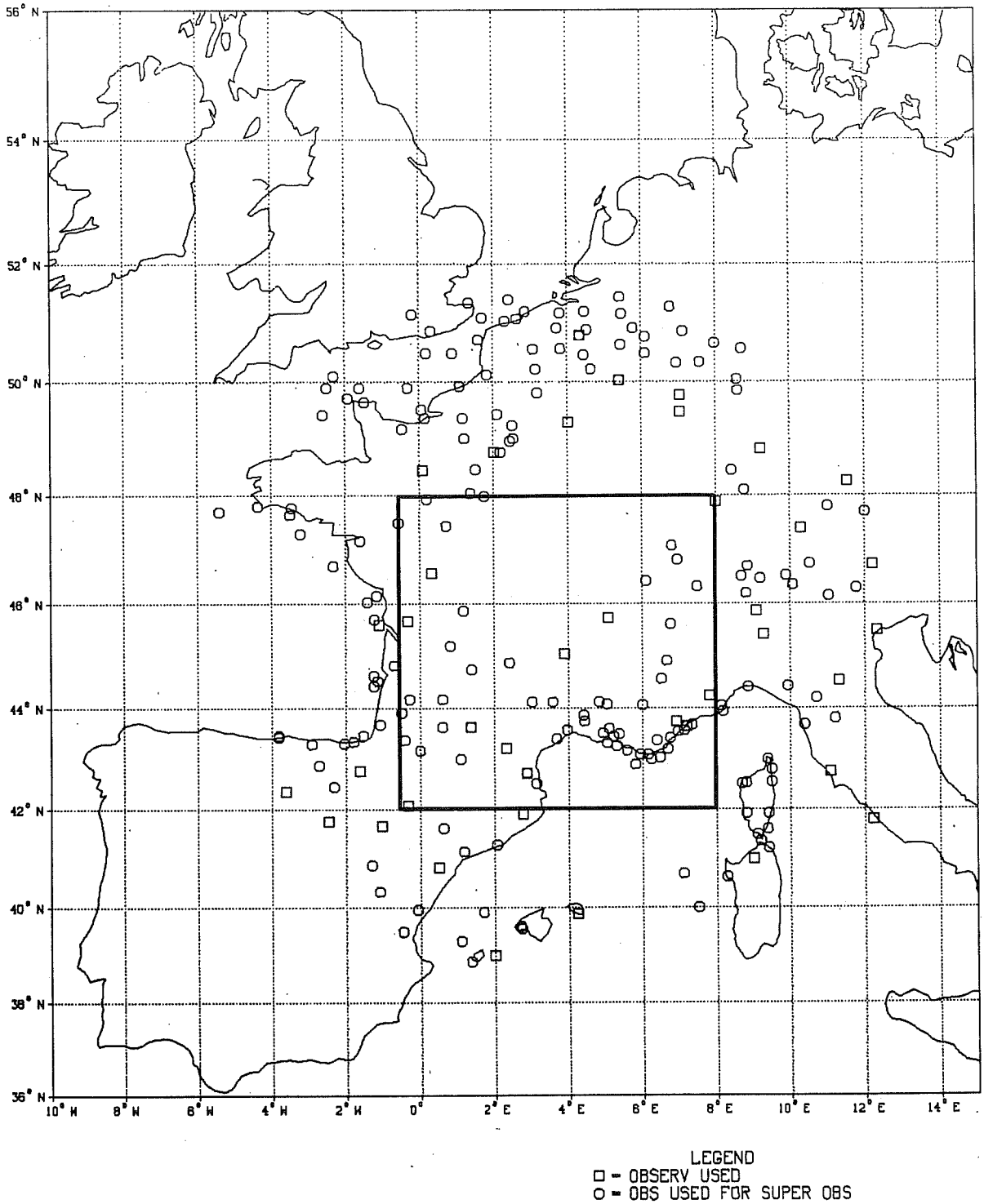



Figure 2: Observations selected for the analysis of the rectangular box shown. Observations used to form super-observations are indicated by circles. 12 GMT, 7 May 1983.

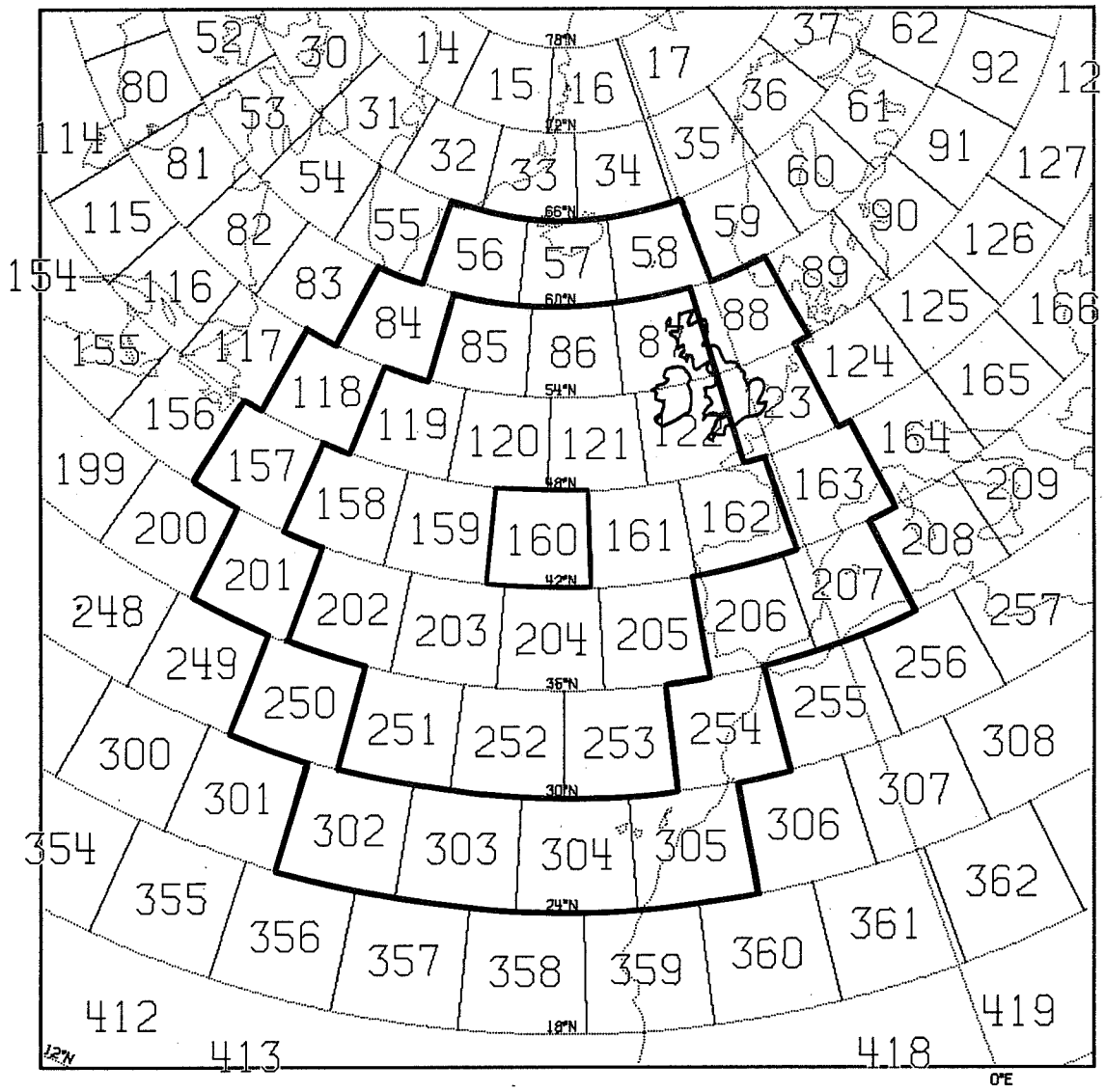



Figure 3: Selection of boxes for data checking and analysis of box 160. The outer boundary indicates the area from within the observations are taken. All observations within the inner boundary are used for the selection of representative observations.

to a distance of 3 boxes from the central box (see Fig. 3).

In the humidity analysis all boxes within a distance of two boxes are selected.

### 3.2 Selection of observations

The next step in the selection of information, for either the analysis or the OI data check, is to choose a representative set of observations from the influencing boxes. No more than 200 observations may be selected. Observations are taken in increasing box distance and all observations within a box distance of 2, i.e. neighbour's neighbour (see Fig. 3), are always taken provided that the number of observations does not exceed 200.

Beyond this limit of two neighbours, observations are selected only if they have an item of information for a level and a variable at which less than 5 + 2 data have been found. The additional 2 data are for possible later rejection. The search is done among the boxes chosen by the box selection.

The observations are then sorted according to distance from the midpoint of the central box. Next, the observation set is truncated to contain observations which are within a distance $D_{max}$ from the centre of the analysis box. $D_{max}$ is given by

$$D_{max} = D_{corner} + n \times b \qquad (3)$$

where $D_{corner}$ is the distance from the centre of the box to a corner and is 471 km; b is the forecast error scale length, and n is 1 in the OI data check and 2 in the analysis (see Fig. 4). The reason for having a smaller search radius for the OI data check is discussed in Section 4.

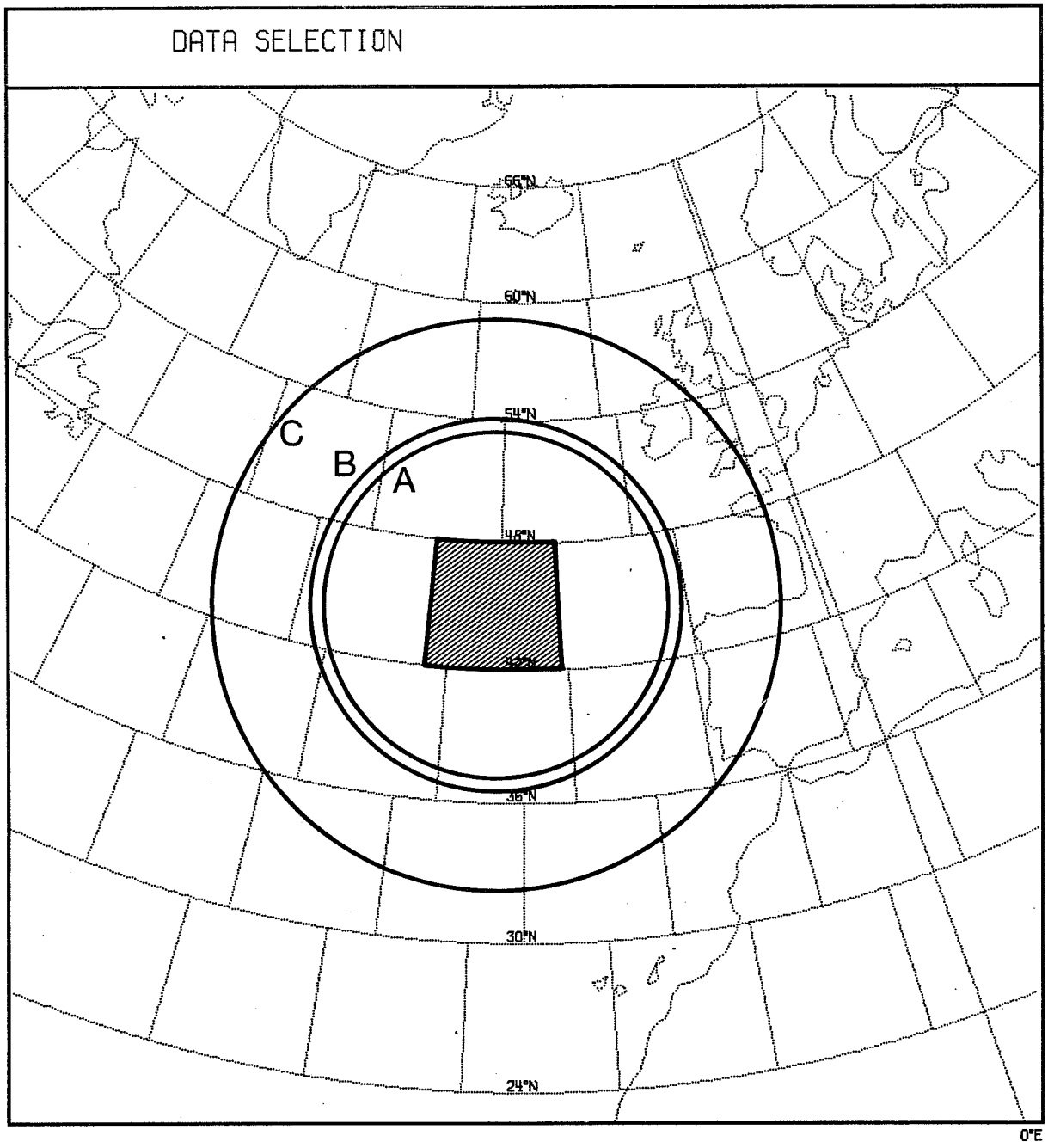



Figure 4: Data selection areas for indicated box:
(A) Minimum data area
(B) Maximum data area for data checking
(C) Maximum data area for analysis

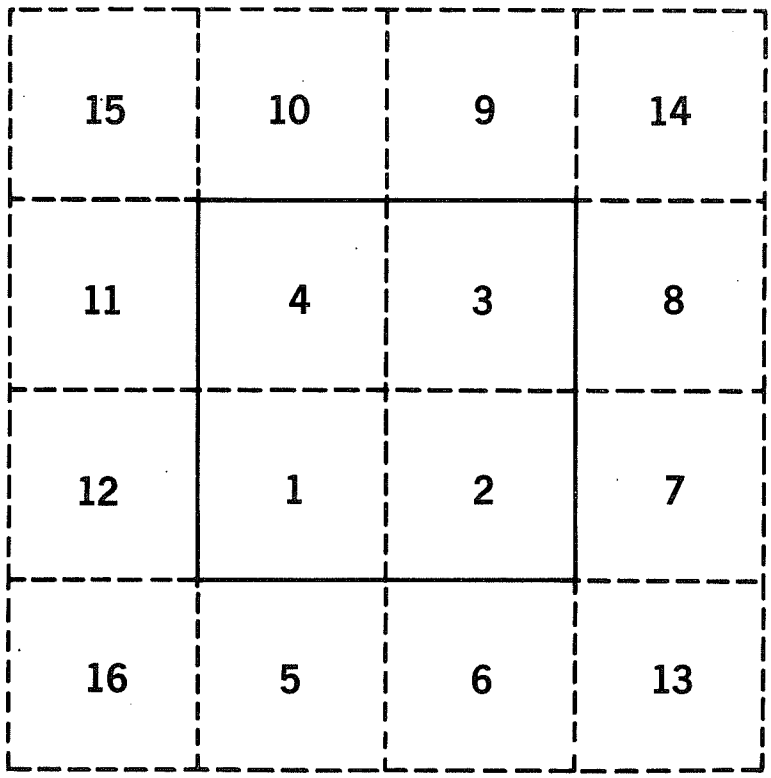



Figure 5: Subregion division of analysis boxes used in the new selection algorithm. The full lines mark subregion boundaries. The numbers indicate the sequence in which subregions are sampled.

At this point, the observations are ordered by distance from the centre of the box. If the number of data in the observations remaining from the previous step exceeds 191, which is the maximum matrix size, preference would be given to close observations (At the end of February 1985 the maximum matrix size was increased to 255). However there is an overriding need to make the selection representative. The central box and the outlying regions defined by a half box extension of the central box are divided into 16 subregions as shown in Fig. 5. These 16 subregions represent the minimum area from which data should be drawn to provide a reasonable analysis for the central box. For a matrix size of 191 the 16 subregions can be permitted to contribute approximately 12 data each. To this end the observations are reordered, drawing observations from the subregions in the sequence indicated in the numbering of Fig. 5. Preference is given to particular types of observations, the pecking order being TEMP, PILOT, SATEM, SYNOP/SHIP, AIREP, SATOB, DRIBU. Each observation is deemed to contribute a fixed number of data to the matrix. When the selection of data for a subregion exceeds 12, selection proceeds to the next subregion. When the subregions have been sampled sufficiently to exceed the matrix size, or when all subregions have been considered, the selection by subregion stops and all remaining observations are sequenced (retaining their distance-from-centre-order) to follow newly selected observations. This leaves the observations in an order which should ensure representativeness, giving preference to observations such as TEMPs.

3.3 Selection of data

If the number of data in the selected observations exceeds 191, the data checking or analysis is done separately for 3 layers. The data selection proceeds separately for each slab of the atmosphere by using the observations ordered as described in Section 3.2. The slab boundaries are 1000-700, 700-150 and 150-10 hPa. The boundary levels 700 and 150 hPa are analysed twice. All data within a radius of 943 km from the centre of the and within the boundaries of the slab are taken provided the 191 data limit is not hit. Beyond the 943 km limit, further data are selected in case the amount of selected data items so far does not exceed 5 at that level. The selection proceeds to the $D_{max}$ boundaries given by (3). If the matrix size has not yet been reached, the data set is supplemented by information outside the analysis volume. Preference is given to data closest to the slab.

By permitting an overlap from one slab to the next in terms of data selected, one mitigates some of the effects that can arise from totally distinct selections for adjacent slabs. However in practice this overlapping may still be insufficient to guarantee smooth transitions, in terms of analysis increments, as one proceeds from one slab to the next. A serious consequence is found in the vertical profile of geopotential increments, where a discontinuity across a slab boundary transforms to an erratic profile of temperature increments. An ultimate solution to this problem is to avoid the vertical partitioning of the atmosphere, and have a common data selection throughout the full depth of the atmosphere. This is not feasible within the framework of the current system, given the constraints on matrix size and the excessive size of the horizontal boxes in data rich regions.

The current approach, which mitigates the effects of the discontinuity, is to replace the analysis of geopotentials in the two upper slabs by an analysis of thickness. By using the bottom slab to provide a reference level, one is

effectively permitting the plentiful near-surface data to have an impact on the geopotential structure in the upper slabs without the data being explicitly used in those slabs. Such an approach has been incorporated in the current scheme (Undén, 1984).

### 3.4 Data selection in humidity analysis

The selection algorithms for the humidity analysis differ from the mass and wind ones in several respects. All observations from boxes up to a distance of 2 neighbours are selected in order of distance. No more than 300 observations are accepted. As the moisture analysis is two-dimensional, only data from the analysis layer are used.

## 4. OI DATA CHECKING

Statistical interpolation is a powerful technique to combine information from various observing systems with different error characteristics. It also provides a method to check data by an independent analysis of surrounding observations. A basic problem of this technique is to determine proper error statistics for all types of situations. Detailed stratification of the error covariances according to flow type is a laborious and difficult process. Usually, the error statistics are stratified according to season and location and it is assumed that they are precise. This means that every observation is assumed to improve the analysis regardless of synoptic situation or distance between observation and analysis points. The impact of using of non-optimal statistics is discussed in Section 4.1.

A derivation of the OI data checking algorithm can be found in Lorenc (1981); here we only briefly describe the technique and our implementation (Section 4.2). In Section 4.3 we present rejection statistics and compare the OI technique with the first-guess and multi-level checks.

## 4.1 Impact of non-optimal statistics

In an operational environment we estimate the forecast error statistics from past performance of the assimilation system. Usually, the data is stratified according to season and area. We then use the population mean in the OI scheme, and ignore any departures from it. That is we apply the statistics for all synoptic situations and disregard that the population may consist of subpopulations with different characteristics (Seaman, 1977). In the following we will investigate the sensitivity of the OI scheme to situations when we use the ensemble statistics for subpopulations with significantly different properties than the mean. Our main concern is the impact of non-optimal statistics in the data checking.

Seaman's calculations have been repeated for a scenario that simulates the OI check. We have a 1-dimensional problem with 5 equidistant points and we estimate the analysis error at the central point using observations from the four surrounding locations. We assume that the observation errors are constant, the normalised error being 0.4, and independent. The forecast error correlation is modelled by a Gaussian type function. We calculate the analysis error as a function of the observation spacing for 5 different situations. These situations represent cases when the actual forecast error scale length departs from the long-term mean. All other assumptions are correct, i.e. we know the amplitude of the forecast and observation errors in this situation. In the two examples (Figs. 6 and 7) we show the analysis error normalised by the forecast error (ordinate) as function of observation spacing normalised by the assumed forecast error scale length (abscissa).

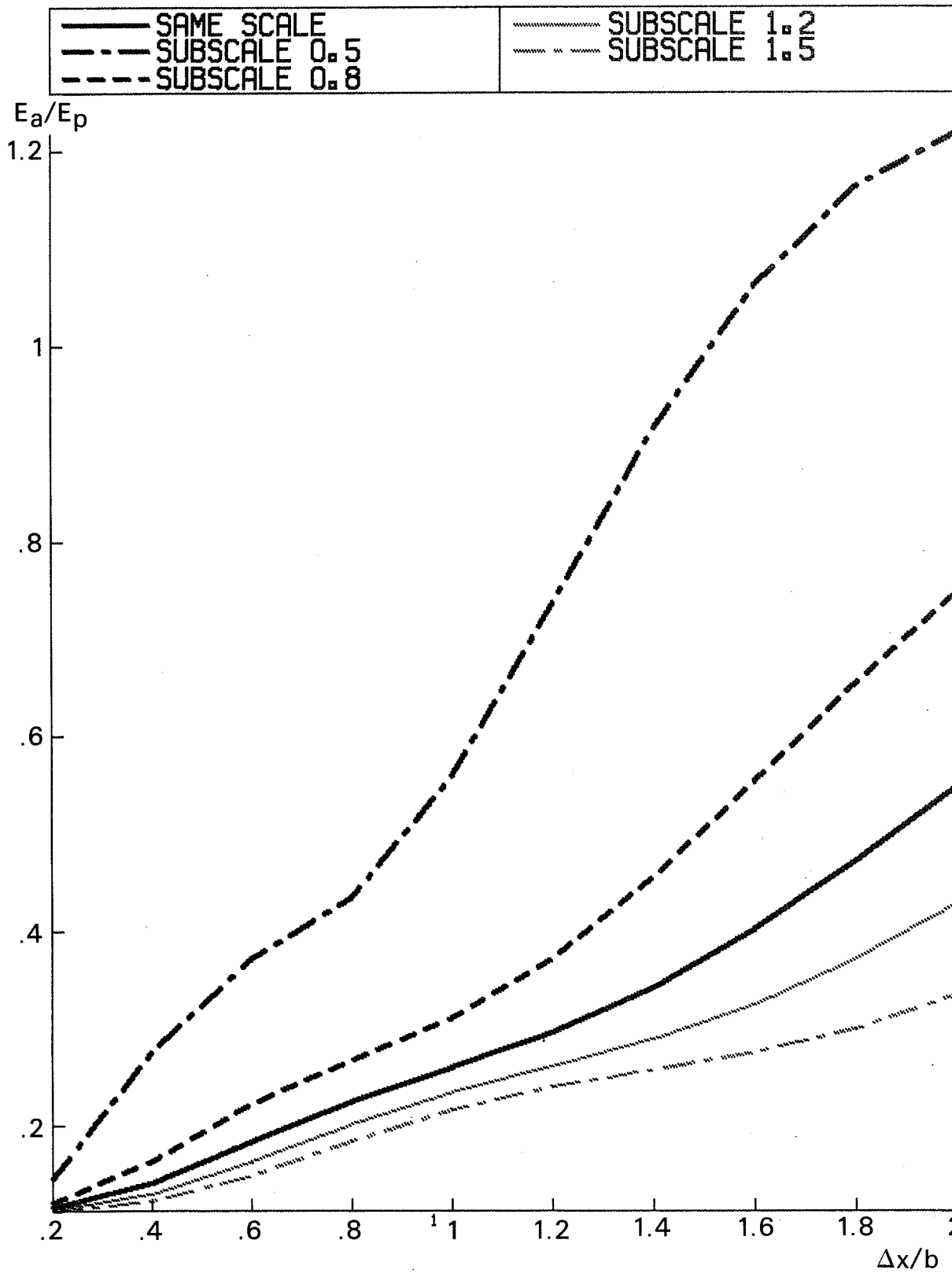



Figure 6: Normalised analysis error ($E_a/E_p$) as a function of the normalised observation spacing ($\Delta x/b$) for the analysis of height using height data. In the five cases shown the real forecast error scale-length is 0.5, 0.8, 1.0, 1.2 and 1.5 of the assumed one.

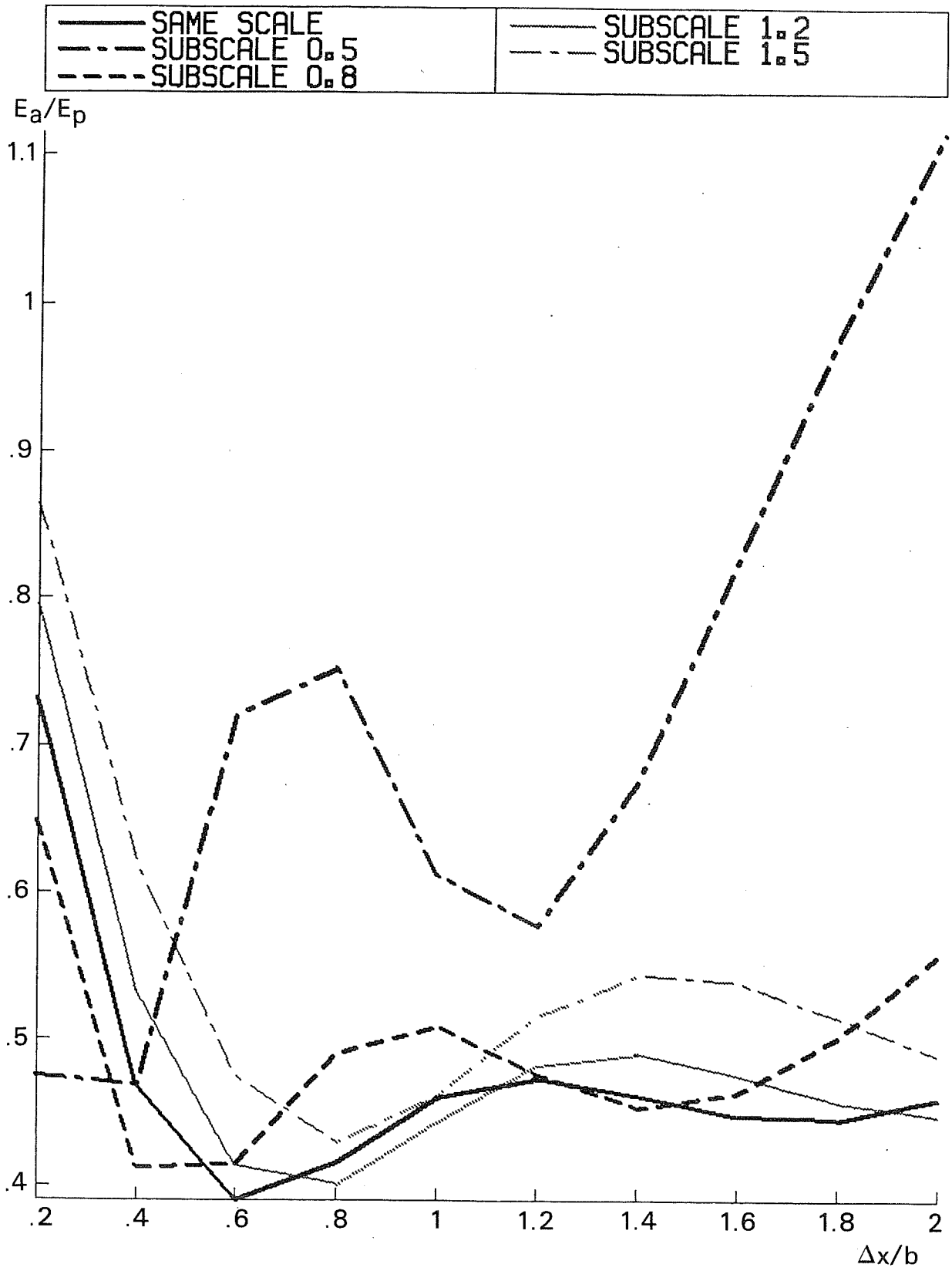



Figure 7: Same as in Fig. 6, but for the analysis of the wind normal to the line of points.

In the first example we calculate the analysis of height at the centre point using height data. We consider first the case when all our assumptions, including the scale length, are precise (indicated by 'same scale' in Fig. 6). We find that the normalised analysis error is about 0.25 for an observation spacing of one scale length and about 0.55 at two scale lengths. However, for cases when the actual scale length is larger the assumed one, we find a reduction in the analysis error. When the real scale length is shorter than the assumed one, an analysis which is worse than first-guess is produced for normalised observation spacing larger than 1.4 in case the real length scale is only half of the assumed one ('subscale 0.5' in Fig. 6).

It is quite clear that departures from long-term statistics affect the analysis accuracy and in particular the estimated analysis error. Fig. 6 shows also that, in a univariate analysis, the statistical assumptions become important for the observation spacings larger than 0.5 scale length. This means that we should not put much confidence in the OI data check in data sparse regions and we should use only close observations to decide on the quality of an observation.

Fig. 7 has the same setup as Fig. 6, but we calculate the error of the analysis of the transverse wind at the central point using height data at the four surrounding points. It is quite clear from Fig. 7 that the statistics are much more important in a multivariate environment than in a univariate case. Closely spaced height observations with random errors have little information on the height gradient. An optimal observation spacing is around 0.6 scale length when the statistical assumptions are reasonably good. Still the analysis error is always around 0.4 or more. With large deviations from the ensemble mean statistic, i.e. the case when the subpopulation has a scale

length which is half of the assumed one, only a small range of observation spacings produces an analysis with an error less than 0.5. This example demonstrates the difficulty in producing a reliable estimate of the wind from height data.

## 4.2 Description of method

The basic idea of the statistical interpolation check is to use the OI technique to provide an analysis at the point of the observation using neighbouring observations but not the observation to be checked itself. The expected squared difference between the observation and the independent analysis is the sum of the observation and analysis error variances. However, this is true only in the circumstance that the assumed statistics are appropriate for that particular situation. To accommodate non-optimal statistics we use an approach like the one proposed by Lorenc (1981) to add a constant to the OI estimate of the analysis error. Since May 1984 we have used the following formulation:

$$E_a = \left(E_{OI}^2 + E_{min}^2/E_p^2\right)^{\frac{1}{2}} \qquad (4)$$

where $E_a$ is the analysis error, $E_p$ the first-guess error, and $E_{OI}$ the optimum interpolation estimate of the analysis error. $E_{min}$ will be defined below. A datum is rejected if its departure exceeds 4 standard deviations of the distribution of the analysis and observation error.

The actual analysis error was estimated over Europe from data which had not been used by the analysis (Illari, pers. comm.). It was found that the analysis error for height is between 5 and 8 m in the lower troposphere; the corresponding OI estimate as produced by the ECMWF system is around 2 m. By

adding, in terms of variances, to the OI estimate a contribution from the unresolved scales and non-optimal statistics, we can obtain a reasonable estimate of the accuracy of the analysis. We call this additional contribution the "minimum analysis error", as it would be the real analysis error in case we had an infinite amount of observations. A value of 6 m for the minimum analysis error for height in the lower troposphere gives errors close to observed values. We then assumed that the ratio between minimum analysis error and six hour forecast error found for lower tropospheric heights is approximately preserved for other levels and variables. This assumption then gives the minimum analysis error for all variables and levels. The values thus obtained are used globally.

### 4.3 Rejection statistics

A fairly typical example of the rejection rates is presented in Table 2. The rejections are separated into two groups, those data rejected by the first-guess and the multi-level check and on the other hand the rejections by the OI check. The third column is the total number of data presented to the analysis. The first-guess check must be regarded as a gross check and should remove only obvious errors. For single level data the rejection rate against the first-guess is very low; SYNOP pressures and heights are most frequently eliminated but less than 0.5% are still rejected. The multi-level check traps internally consistent but incorrect radiosonde heights quite efficiently. More then 1% of them are found incorrect. The OI check is most efficient where data density is high and observation errors are uncorrelated, i.e. for SYNOPs and SHIPs.

Table 3 gives the vertical distribution of radiosonde rejections for height and wind. The wind rejections peak at jet level and are about 0.5% of total

## SUMMARY OF REJECTIONS (18.10.84 12 GMT)

| | FG/ML | OI | TOTAL |
|---|---|---|---|
| SYNOP/SHIP Z | 14 | 51 | 3,244 |
| SYNOP/SHIP V | 0 | 1 | 684 |
| AIRCRAFT | 1 | 6 | 612 |
| SATOB | 0 | 3 | 951 |
| DRIBU Z | 3 | 1 | 37 |
| DRIBU V | 0 | 0 | 3 |
| TEMP Z | 99 | 17 | 7,239 |
| TEMP V | 19 | 10 | 6,865 |
| PILOT | 3 | 2 | 603 |
| SATEM | 0 | 3 | 5,811 |
| PAOB | 0 | 0 | 155 |

Table 2: Data rejection summary for one analysis cycle (12 GMT, 18 October 1984). FG/ML is the amount of data rejected by the first-guess and multi-level check, and OI refers to the statistical interpolation check. TOTAL gives the number of data presented to the analysis.

## TEMP REJECTION RATES JUNE-AUGUST 1984

| | % | |
|---|---|---|
| MB | V | Z |
| 1000 | 0.16 | 0.86 |
| 500 | 0.28 | 1.07 |
| 300 | 0.42 | 1.84 |
| 200 | 0.55 | 2.57 |
| 100 | 0.48 | 3.48 |
| 50 | 0.23 | 3.5 |
| 20 | 0.16 | 3.84 |

Table 3: Global rejection rates of TEMP data, wind and height, at selected levels. 12 GMT observations, 1 June-31 August 1984.

amount of wind data. The rate of rejection for geopotential reports increase with height from about 1% in the lower troposphere to almost 4% in the stratosphere. An improper radiation correction applied to the radiosonde temperatures has the most serious consequences in the stratosphere. Further rejection statistics can be found in Shaw et al. (1984).

## 5. DISCUSSION

Sophisticated data selection algorithms are needed in an assimilation system that simultaneously analyses several variables in an extensive volume. The scheme must provide a homogeneous and comprehensive selection of data in data rich regions without heavy penalty in computer time.

In the ECMWF system, the final and most stringent quality control check is based on the statistical interpolation technique. The difficulty with this method is to find the right balance between noise suppression and dependence on past performance statistics.

In the ECMWF system two basic and quite severe assumptions are made. It is assumed that the observation and forecast error distributions are normal. This is clearly not true for some observing systems. Techniques to detect the non-Gaussian nature of the error distribution and methods to use it are clearly needed. The other assumption is that the forecast and the observations are unbiased. This problem may to some extent be remedied by correcting the observations for their systematic errors.

Isolated observations create special problems in any analysis system. A statistical check working on only one time level does not give a satisfactory means of identifying measurement errors. For this category of data a test on

temporal continuity would provide good guidance to detect erroneous observations. The other aspect of data checking is data assimilation frequency. Frequent data assimilation means a high likelihood for areas with sparse observations. In those areas the current methods of data checking will work unsatisfactorily.

REFERENCES

Delsol, F., 1985: Monitoring the availability and the quality of observations at ECMWF. ECMWF workshop on the Use and Quality Control of Meteorological Observations.

Gustafsson, N., and J. Pailleux, 1981: On the quality of FGGE data and some remarks on the ECMWF data assimilation system. ECMWF Tech.Memo.No.37, 29 pp.

Hollingsworth, A., D.B. Shaw, P. Lönnberg, L. Illari, K. Arpe and A.J. Simmons, 1985: Monitoring of observation quality by a data assimilation system. ECMWF Seminar/Workshop on Data Assimilation Systems and Observing System Experiments with Particular Emphasis on FGGE.

Hollingsworth, A., and P. Lönnberg, 1985: The statistical structure of short range forecast errors as determined from radiosonde data. Part I: The wind field. ECMWF Seminar/Workshop on Data Assimilation Systems and Observing System Experiments with Particular Emphasis on FGGE.

Lange, A., 1985: Quality of TEMP data. ECMWF Workshop on the Use and Quality Control of Meteorological Observations.

Lönnberg, P., and D. Shaw (Eds), 1983: ECMWF Data Assimilation Scientific Documentation. ECMWF Meteorological Bulletin M1.5/1. Research Manual 1.

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

Lorenc, A.C., 1985: Analysis methods for the quality control of observations. ECMWF Workshop on the Use and Quality Control of Meteorological Observations.

Seaman, R.S., 1977: Absolute and differential accuracy of analyses achievable with specified observational network characteristics. Mon.Wea.Rev., 105, 1211-1222.

Shaw, D., P. Lönnberg, and A. Hollingsworth, 1984: The 1984 revision of the ECMWF analysis system. ECMWF Tech.Memo.No.92, 69 pp.

Undén, P., 1984: Evaluation of analysis increments at model levels. ECMWF Tech.Memo.No.94, 25 pp.