DATA SELECTION AND QUALITY CONTROL IN THE FRENCH PRE-ANALYSIS AND ANALYSIS SYSTEMS

J. Pailleux
Direction de la Meteorologie, Paris, France

1. INTRODUCTION

For several years, a 3-dimensional optimum interpolation hemispheric analysis system has been used operationally in Paris. This system has been recoded recently for a Cray computer (1984) and at the same time a meso-scale analysis system was implemented on the same computer: see Pailleux, 1982.

The purpose of this paper is to explain in detail what messages are used in these two analysis schemes, what data are used from each type of observation and what processing is applied to the data in the pre-analysis, data selection and quality control steps.

The different types of messages which are currently used are: SYNOP, SHIP, DRIBU, TEMP, PILOT, AIREP, SATEM and SATOB; bogus data (created by a forecaster) are also used.

In section 2 we describe the procedures which are used in the large-scale hemispheric analysis recoded for the Cray 1, while in section 3 we mention what procedures in the meso-scale analysis are different from those in the large-scale scheme.

2. THE LARGE-SCALE HEMISPHERIC ANALYSIS

2.1 Main characteristics of the analysis system

The preliminary manipulation of the observations and the format used to store them are very dependent on the analysis system itself, and especially on the analysed parameters and the vertical coordinate. For example, the data which are actually used in a TEMP report are generally different for a system analysing the geopotential height on standard surfaces than for one analysing temperatures on sigma surfaces.

Therefore it is important to describe the main features of the analysis scheme itself. We can summarise them in the following way:

Analysed parameters

The parameters analysed are Φ, u, v and T at 16 pressure levels from 1000 to 50 hP, plus relative humidity Hu in 9 layers from 1000 to 300 hP and surface parameters (2m temperature and relative humidity). The pressure levels are the standard ones from 1000 to 50 hP, plus levels at 950, 900, 800 and 600 hP. The presence of these 4 levels explains why we use TEMP B in addition to TEMP A. Above 100 hP only the 70 and 50 hP levels are analysed, so TEMP C is used but not TEMP D. The same holds for PILOT reports. It was decided to analyse the temperature and relative humidity fields at 2 m because the dew-point and temperature SYNOP data can be used directly : these analysed fields are then used to initialise some surface parameters of the model through an increment interpolation technique.

Interpolation method

3-dimensional multivariate optimum interpolation is used for Φ, u and v, but univariate optimum interpolation for relative humidity and temperature (including the surface parameters).

Area and horizontal resolution

The northern hemisphere analysis uses a latitude/longitude grid (Δlat=1.5°, Δlon=2°).

Guess-field

In the usual operational context the guess-field is the 6-hour forecast produced by a 15-level spectral PE model which covers the northern hemisphere. The guess-field could also be a 12, 18 or 24 hour forecast, or a climatology (cold start).

To analyse one grid-point

We do not use all the data to analyse one grid-point - a limited number of data is selected around the grid-point to build the OI linear system (see section 2.3).

2.2 Preparation of the data

Surface observations

For the height and wind analysis, SYNOP and SHIP reports are discarded when they are very near to a radiosonde. The main reason is that in the preprocessing, Φ at 1000 mb is recalculated for most of the TEMP reports if it is not available and this information is redundant since the sea-level pressure from surface observations is available. Nevertheless the ocean weather ship surface observations are kept; moreover a very high quality is assumed for them in the quality control checks (see section 2.4). The SYNOP surface winds are discarded if the altitude of the station is greater than 150 m. With regard to the DRIBU messages, we sometimes get several reports from the same buoy on the GTS in a short period; only the report which is the nearest in time to the synoptic hour is kept for the analysis.

For relative humidity and surface analysis all the surface observations are kept. At the preprocessing step all the surface data are submitted to simple quality control checks (Td $\leqslant$ T or climatological checks for example).

TEMP and PILOT messages

Φ, u, v and T are generally provided by the TEMP A and C reports on all the standard levels. If TEMP A is absent and TEMP B present, the last one is used to recalculate the data, as far as possible. A vertical integration is performed from the observed T and Td to compute the relative humidity in the 9 isobaric layers 1000/950, 950/900, ... 400/300. Note also that a hydrostatic check and a vertical consistency check have been performed on the TEMP messages in the preprocessing program. The consistency between parts A, B, C and D has been checked at that stage too.

SATEM observations

SATEM observations are used everywhere, even over the continents. No special processing is applied to the thicknesses which are directly inserted in the height/wind analysis through the multivariate scheme. The precipitable water content (PWC), which is available in some SATEM

reports, is converted to an observed humidity profile (for the 9 layers) by using the temperature of some layers and also the guess-field humidity profile.

SATOB observations

The wind data are used directly, without any special computation; the pressure given in the report is kept (no special technique is used to reassign the level of the wind data). The PWC and T observations given in some SATOB messages are not used in the analysis.

AIREP observations

The wind data are used directly, they are assigned to a pressure level which is generally computed from an altitude given in the report (in feet) through the standard atmosphere formula.

2.3 Data selection performed to analyse a given point

There are different techniques currently used to perform an optimum interpolation analysis. The first one consists of building a large correlation matrix to perform the analysis at all the grid-points of that area, while the second one selects a limited number of data to analyse one given parameter. The first technique is used at ECMWF, in Paris we use the second one. Let us give the different steps of the data selection algorithm used in Paris to analyse Φ, u and v.

The algorithm is divided in 2 main parts:

- the "geographical" selection which retains a limited number of observations for all the grid points in one vertical column of the analysis grid.
- the "statistical" selection which is performed for each analysis level and retains a limited number of data (up to 13) chosen from the observations retained by the geographical selection.

Different steps of the geographical selection

In a pre-analysis program, the observations have been sorted into latitude/longitude boxes (box size = 12^{o} x 12^{o}), and when we want to analyse Φ, u and v at all the levels of a given grid-point, we start by

selecting all the observation boxes which are near to the grid-point. Then we distinguish 5 different "types" of observations (SURFACE, AIREP, SATOB, TEMP/PILOT, SATEM) and the following steps are applied to each type:

a. We keep all the observations for which the distance to the grid-point is lower than a (a is a horizontal distance, depending on the scale of the structure functions).
b. We keep the N nearest observations. Presently N = 12 for SYNOP/SHIP/DRIBU, N = 6 for AIREP, N = 6 for SATOB, N = 12 for TEMP/PILOT and N = 6 for SATEM.
c. If 2 observations are very near to each other, one of them is discarded. Usually the nearest one is kept, but a TEMP is preferred to a PILOT; also a SYNOP is preferred to a SHIP which in turn is preferred to a DRIBU. Note that no "inter-distance" check is performed between observations of different types.
d. The distribution of the observations around the grid-point is controlled by a "quadrant check" which allows only n observations in each quadrant. n = 4, 2, 2, 4 and 3 for the different types of observations.

Statistical selection

For a given grid-point G and a given analysis level p, the correlations are calculated for each analysed parameter Φ, u, v with all the observational data retained by the geographical selection. Then the 6 data which have the highest correlations with Φ are retained (they could be Φ, u, v or $\Delta\Phi$) along with the 4 data (not among the 6 previous ones) which have the highest correlations with u and the 3 data (not already selected) which have the highest correlations with v. A linear OI system (maximum 13 x 13) is then constructed and solved to analyse simultaneously Φ, u and v at the grid-point G (this guarantees a certain consistency in the height and wind analysis).

General remarks

The main advantage of this data selection procedure is its flexibility: it is easy to tune it and, for example, to avoid the bad effects of systematic errors in some observing systems (see Gustafsson and

Pailleux, 1981). One of the drawbacks is that the horizontal and vertical consistency of the analysis is not as good as it could be in a system using the "box technique" and big matrices.

In the temperature and humidity analysis performed in Paris, the data selection is simpler because the interpolation is univariate and the maximum number of data kept is 8 instead of 13.

2.4 Quality control

The general principle of the analysis quality control is to analyse each observed parameter using all the available information except the observed parameter itself. Then, if the analysed value Ψ_a is sufficiently different from the observed value Ψ_o, Ψ_o is rejected.

We can describe the different steps of the algorithm more precisely as follows:

- In the data base, a quality code is associated with each data. It is:
  0 for a bad data: rejected by the preprocessing or the analysis.
  1 if we do not know anything (the usual case).
  2 for a data already accepted by a previous analysis (presumably correct).
  3 for a "forced" data; "forced" means that a forecaster has indicated by a manual monitoring that he wants the data to be accepted. The OWS surface observations receive also the code 3.

- A first check against the guess-field Ψ_p is applied to each data with a code equal to 1 or 3. If $|\Psi_o - \Psi_p| < \alpha\sqrt{\sigma_o^2 - \sigma_p^2}$, then the data Ψ passes successfully that test, if not the quality code is put to 0.

  $0 \longrightarrow 0$ bad data remain bad.
  $1 \longrightarrow 0$ or 1 : depending on the result of the test.
  $2 \longrightarrow 1$: data with 2 are not submitted to that test.
  $3 \longrightarrow 0$ or 3 : but α is considerably larger for "forced" data than for the usual data (so 3 generally remains 3, with the exception of a typing mistake by the forecaster.

- An optimum interpolation analysis is performed for each observed parameter (result = Ψ_a) using all the data with a quality code equal to 1 or 3, and not using Ψ_o itself. If $|\Psi_o - \Psi_a| < \beta\sqrt{\sigma_o^2 + \sigma_p^2}$, then the data Ψo is kept for the analysis.

  0 ⟶ 0 or 1

  1 ⟶ 0 or 1

  3 ⟶ 0 or 3 but β is considerably larger for a "forced" data.

  Then all the observations with a code equal to 0 are rejected.

The last step could be repeated, even several times, to perform the quality control with only the data kept at the previous step, but this is not done at the present time.

When a thickness is rejected in a SATEM observation, the whole profile is rejected (to keep the consistency). For a TEMP report a "global decision" is taken to preserve the vertical consistency of the rejection algorithm.

For the temperature analysis, the quality control is reduced to a very simple test consisting of rejecting all the observations which are too far from the guess-field (the guess field is deduced from the analysed geopotential heights and is assumed to be very good).

2.5 Some remarks on quality control for humidity observations

The procedure described in section 2.4 seems to work well when we deal with analysed parameters which have a statistical distribution not very far from the Gaussian distribution. But the relative humidity has a special statistical distribution with most of the values near 0 or 1. Some experiments indicate that if we apply the tests described in section 2.4 to relative humidity observations with too stringent limits (α and β too low), then all the moist observations are rejected when the guess-field is dry and all the dry observations are rejected when the guess-field is moist. Then each grid-point value tends to become 0 or 1 in several assimilation cycles with spurious horizontal gradients. Possible ways of avoiding this problem are:

- to relax the quality control (to increase α and β).

- to take into account the special distribution of Hu.
- to analyse the specific humidity instead of Hu.

3. FINE-MESH ANALYSIS

3.1 Differences in the analysis scheme

- The analysed parameters are the prognostic variables of the model: T, u, v on the σ-layers (+ p_s).
- The scheme can use satellite clear radiances directly through the multivariate analysis technique.
- Horizontal mesh: 35 km on a polar stereographic projection.

3.2 Consequences on the preparation of the data

- For TEMP and PILOT messages, no special computation is needed to calculate data at levels 950, 900, 800 and 600 hP; instead TEMP B and D are used in a fully 3-dimensional way.
- The radiances are used instead of SATEM.

3.3 Data selection

The actual number of selected data differs from the examples given in section 2.3, but the general method is the same as for the large-scale analysis. The selection algorithm is a little more sophisticated because of two additions:

- The "interdistance check" is also performed between observations of different types.
- The "quadrant test" has been improved: in the statistical selection, if two adjacent quadrants are completely void of data, another attempt is made to select other data in these quadrants.

3.4 Quality control

The quality control procedure in the fine-mesh analysis is still being developed. At the present time the rejection algorithm is reduced to a check against the guess-field and a buddy check. Note also that the data which have been rejected by the large-scale analysis are considered as bad and are not taken into account: such a procedure is not necessarily to be recommended, and it may not be kept in the final suite.

3.5 Use of the radiance data in the fine-mesh analysis

The specific aspect of that analysis scheme is its ability to use the raw radiances directly in the optimum interpolation scheme. A radiance guess-field is needed: it is calculated from the predicted temperature and humidity fields. Then the radiance increments $\varepsilon R = R_o - R_p$ are inserted into the optimum interpolation scheme. Ad-hoc statistics have been derived to do this.

4. CONCLUSIONS

The main points of the data selection and quality control procedures in the DM operational analysis system are the following:

- Use of a manual monitoring of the observations (actually affecting a very limited amount of data).
- Use of a limited number of data to analyse a grid-point (limitation on the optimum interpolation matrix size).
- Use of the same data to analyse Z, u, v (or T, u and v in the fine-mesh analysis) at a given grid-point. (Attempt to improve the balance between mass and wind fields).
- In the data selection procedures, the distance of the data to the grid-point, the data quality as well as their distribution around the grid-point are taken into account.
- Direct use of the radiances in the fine-mesh analysis.

REFERENCES

Pailleux, J., 1982: Meso-scale analysis at DM. Proceedings of the ECMWF workshop on "Current problems in data assimilation". (November 1982).

Gustafsson, N. and Pailleux, J., 1981: On the quality of FGGE data and some remarks on the ECMWF data assimilation system. ECMWF Technical Memorandum no 37.