Data selection and quality control algorithms used at SMHI.

Nils Gustafsson
Swedish Meteorological and Hydrological Institute.

1. Introduction

Quality control of observational data is a necessary component of any system for processing of meteorological data, not only for utilization of observations in the construction of initial fields for numerical models but also for e.g. climatological processing of meteorological data.

A basic objective of quality control is to detect errors in the observational data and, if possible, to correct these errors. An equally important objective of quality control is to monitor the performance of various observing systems in order to be able to inform the data producers about their possible random and/or systematic mistakes and errors. The former objective has, so far, been of main concern for the computerized quality control carried out at the Swedish Meteorological and Hydrological Institute (SMHI). Very little attention has been paid to the second objective. Recently, however, the near-operational application of a meso-scale objective analysis scheme has made it possible to carry out a limited monitoring of the performance of the Swedish surface data network. As an example, systematic errors in the reduced sea-level pressure observations have been noticed for a number of Swedish surface stations.

A number of quality control efforts are carried out at various stages of the data processing at SMHI. Unfortunately, most of these quality control efforts are performed with a significant amount of overlapping in the quality control tasks and independently of each other. As an example, various parameters in the national surface data reports are controlled independently of each other during manual analysis of the surface charts, during application of a meso-scale objective analysis scheme and during climatological processing of the data. An integrated system for quality control of observational data at SMHI has been suggested (Dahlström 1980) but the development and operational implementation of this system have been delayed. One of the basic concepts of this suggested system is communication of quality control information back to the data source.

2. Overview of data processing and quality control at SMHI

Some of the main data processing systems including quality control tasks at SHMI are briefly illustrated in Figure 1. Quality control is carried out in the following steps:

* Quality control of message bulletin headings is carried out in the Telecommunication System. Errors in bulletin headings are corrected manually.

* The evaluation and coding of upper-air message reports are computerized and include certain quality control tasks.

* Some of the manual surface stations are delivering observed information via a micro-processor at the station site. These micro-processors have been programmed to carry out certain simple quality control algorithms and the observer is immediately informed about detected mistakes.

* Observations from the automatic surface stations are quality controlled locally by station micro-processors. Quality control information is included in the messages from the automatic stations.

* All data entering the central computer system of SMHI are transmitted from the Telecommunication System and taken care of by a pre-processing system. Decoding, data format checks, limit checks, internal consistency checks and time consistency checks are carried out and observed data are stored in the synoptic data base together with quality control flags (section 3).

* Surface and upper-air data are displayed on charts which are presented to the forecasters for e.g. manual analysis of surface charts involving a certain amount of quality control (no feed-back to the data base).

* The Limited Area Model calculations are carried out four times a day up to 36 hours for an area covering Europe, the Northern Atlantic and parts of the Arctic and Northern America. Quality control of observed deviations from first guess forecasts and spatial consistency check are done during objective analysis for the Limited Area Model (section 4). The results of this quality control are not presently entered into the Synoptic Data Base.

- Meso-scale objective analysis over an area covering Southern Sweden and surrounding sea areas is performed in a near-operational mode (four times a day). A rather powerful check of some surface parameters is done within the meso-scale objective analysis (section 6). No results from this quality control are entered into the Synoptic Data Base.

- Almost all required climatological information from the synoptic network is extracted from the synoptic data base. Additional quality control is carried out manually (no feed-back to the synoptic data base).

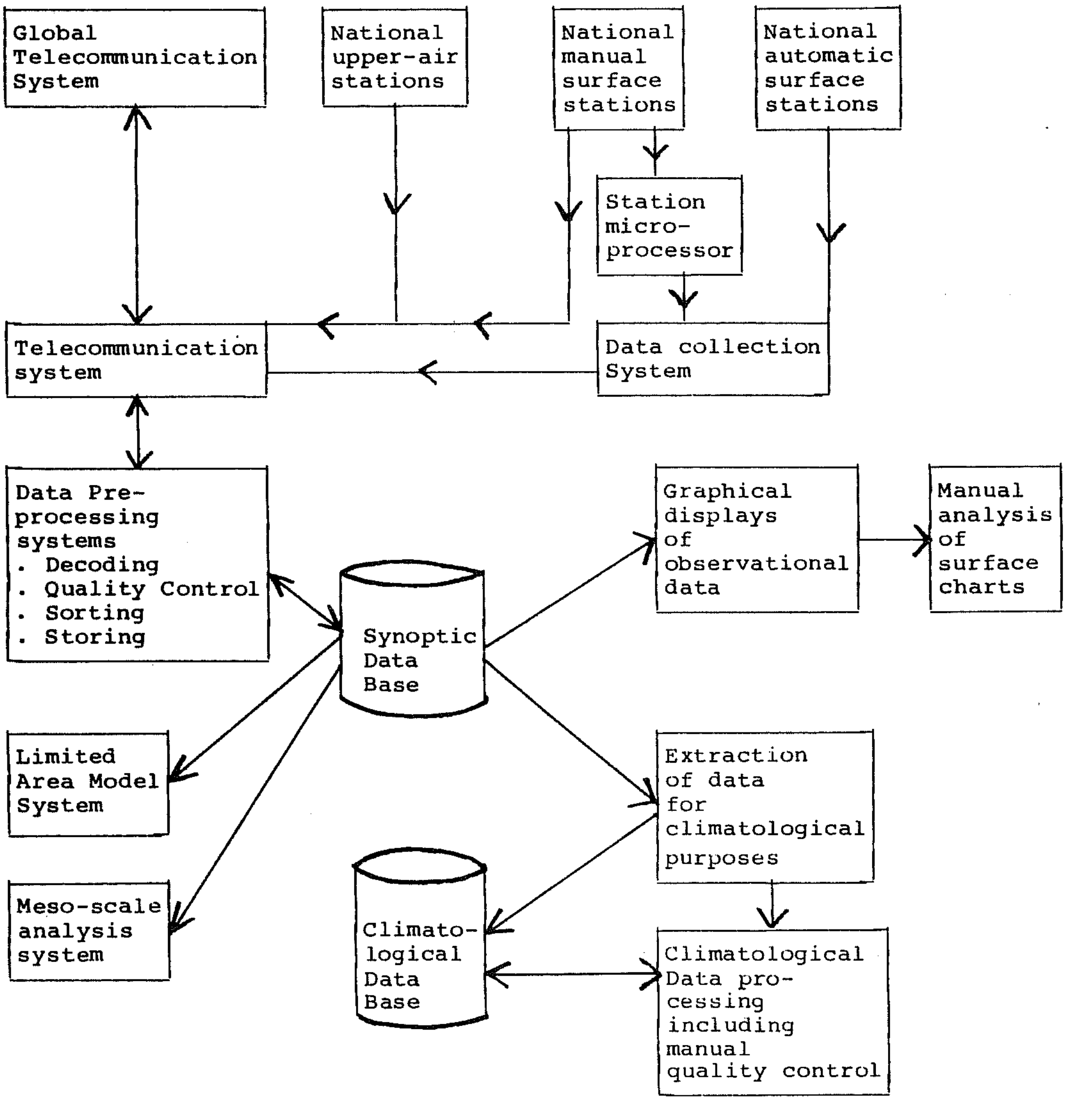
Global Telecommunication System
National upper-air stations
National manual surface stations
National automatic surface stations
Station micro-processor
Telecommunication system
Data collection System
Data Pre-processing systems
. Decoding
. Quality Control
. Sorting
. Storing
Synoptic Data Base
Graphical displays of observational data
Manual analysis of surface charts
Limited Area Model System
Meso-scale analysis system
Extraction of data for climatological purposes
Climato-logical Data Base
Climatological Data pro-cessing including manual quality control

Figure 1 **Overview of data processing and quality control at SMHI**

3. Quality control algorithms applied during pre-analysis processing

A number of quality control algorithms are applied during the pre-analysis processing of observational data at SMHI. Most of these algorithms are quite simple and similar to those applied by other centres (a description of the algorithms can be obtained from SMHI). Besides message format checks and other checks carried out during the decoding, the following types of algorithms are used for the various types of observational reports:

(A) Radiosonde and PILOT wind reports

- Check against climatological limits
- Comparison between significant and standard level data
- Check of vertical profiles for unreasonable lapse-rates and vertical wind shears.
- Hydrostatic check of standard pressure level geopotentials and temperatures.

(B) Surface data (SYNOP, SHIP and DRIBU)

- Check against climatological limits
- Time consistency check against reported values in the previously available reports from the same stations.
- Internal consistency check
- Check of position in surface ship reports.

(C) Satellite data (SATEM and SATOB) and aircraft data

- Check against climatological limits

The quality control algorithms listed above are partly based on previous quality control systems utilized operationally at SMHI since the middle of the 1960's and partly developed in connection with Sweden's participation as a main data management centre during the Global Weather Experiment (FGGE). Unfortunately, the operational implementation of these quality control algorithms has not been as successful as the quality control carried out in previous systems or during the FGGE. The results of the operational application of these algorithms with their output in the form of quality control flags have been found not too reliable. Therefore, the quality control flags produced by the pre-analysis quality control are presently not utilized by the analysis system!

An effort to improve the pre-analysis quality control algorithms is presently carried out at SMHI. From a general point of view, the quality control flags produced by the pre-analysis quality control should be useful in cases when the spatial quality control during the analysis is not able to classify the observed data as 'correct' or 'wrong' due to lack of reference information.

4. Data selection algorithms and quality control algorithms applied during the analysis for the Limited Area Model

The analysis for the Limited Area Model is performed on standard pressure levels. The basic analysis method is a 3-dimensional uni-variate statistical interpolation of observed deviations from 6 hour numerical forecast fields followed by a variational adjustment of the wind-field and the mass-field analysis increments to geostrophic balance. The main reasons for the uni-variate approach have been limitations in computer resources and the difficulty, in the multi-variate case, to make a proper local data selection using only 6-12 pieces of information. A main drawback of the uni-variate approach is considered to be the poor utilization of certain powerful multi-variate combinations, e.g. satellite sounding data (thermal winds) and single-level wind data.

The following aspects of quality control and data selection for the Limited Area Model system are discussed below:

(a) Pre-selection of observed parameters to enter the analysis computations.

(b) Check of observed values against numerical forecast values.

(c) Selection of data for analysis to grid points as well as to observational points for data checking purposes.

(d) Space consistency quality control during the analysis.

The quality control carried out during the analysis for the Limited Area Model appears to perform satisfactory but, it must be mentioned, very little effort has been devoted to monitor its performance.

4.1 Summary of observational data utilized in the analysis for the Limited Area Model

A subset of the complete observational information, available via the GTS, is utilized during the analysis for the Limited Area Model. Data parameters are extracted from the complete observational data records to fit any of the following types of analysis input data records:

Type 1:	Multi-level geopotential data
(Type 2:	Single-level -"- -"-)
Type 3:	Multi-level wind data
Type 4:	Single-level -"- -"-
Type 5:	Multi-level temperature data
(Type 6:	Single-level -"- -"-)
Type 7:	Multi-level humidity data
(Type 8:	Single-level -"- -"-)
Type 9:	Multi-layer thickness data
Type 10:	Sea-level pressure data

Data record types within parenthesis are not presently utilized in the operational analysis for the Limited Area Model.

Multi-level data are extracted for those standard pressure levels where the analysis is carried out. The multi-layer thickness information is given as thicknesses between the analysis standard pressure levels and 1000 mb. This particular representation of thickness data is motivated by the need to utilize the combined effect of sea-level pressure data and thickness data in the analysis of upper-level geopotentials.

The extraction of data to be utilized for the analysis is described below for the various types of observations.

(a) Radiosonde data (TEMP, TEMP SHIP)
Data on the standard pressure levels of the analysis are extracted for geopotential, wind, temperature and relative humidity. If necessary, standard pressure level data are obtained from significant level data by vertical interpolation.

(b) PILOT wind data (PILOT, PILOT SHIP)
If PILOT wind data are given on pressure levels, wind data on the standard pressure levels of the analysis are extracted, if necessary by the aid of vertical interpolation. If the PILOT wind data are given on height surfaces, the 6 hour geopotential forecasts are utilized to convert these wind data to standard level pressure surfaces.

(c) Surface land report (SYNOP)
Only sea-level pressure information is utilized. Data from stations reducing to other levels than the sea-level are not used.

(d) Surface ship reports (SHIP)
Sea-level pressure and 10 meter winds are extracted. Since the main purpose of the surface wind analysis is to improve the gradients of the sea-level pressure analysis, and since the first guess wind field is obtained by a gradient wind relation from the first guess pressure field, the observed 10 meter winds are corrected for frictional effects.

(e) Aircraft data (AIREP, ASDAR)
Wind data is extracted and vertical pressure level is assigned in accordance with the flight level information. Temperature is not utilized.

(f) Satellite sounding data (SATEM)
Thicknesses between the standard pressure analysis levels and 1000mb are extracted. SATEM reports which do not contain thicknesses all the way down to 1000mb (e.g. over mountainous areas) are not utilized. SATEM reports based on micro-wave data only in the troposphere are assigned larger assumed standard-deviations of observational errors than those based also on infrared information in the troposphere.

(g) Satellite wind data (SATOB)
Wind vectors are extracted and assigned to pressure levels as reported in the messages.

(h) Drifting buoy data (DRIBU)
Sea-level pressure is extracted.

The Limited Area Model is presently run with a 6 hour analysis cycle. For the following types of observations, data are extracted from a time period ± 3 hours around the analysis hour: TEMP, TEMP SHIP, PILOT, PILOT SHIP, AIREP (ASDAR), SATEM, SATOB and DRIBU. Surface reports (SYNOP and SHIP) are used from a time period of ± 0.5 hour around the analysis hour only.

4.2 Check of observed values against numerical forecast values

In order to eliminate the most obvious errors in the extracted observational data set, a crude check of each observed value against a numerical forecast value is carried out. All observed values with absolute deviations from the forecasted values above certain tolerances are rejected from further use during the analysis processing. The main objectives of this very crude checking algorithm are to minimize the computer time and to optimize the efficiency of the spatial consistency checking which is carried out in a subsequent phase of the analysis processing. The tolerances are given as functions of month, parameter and vertical level. Examples of tolerances for January and July are given in the table below.

Level	Maximum permitted deviations between observed and first guess forecast values			
	Geopotential gpm		Wind vector m/s	
	January	July	January	July
1000 mb	240	160	25	15
500 mb	240	160	45	30
200 mb	385	300	55	40

4.3 Data selection for analysis of various parameters

A critical problem for all objective analysis schemes is the selection of influencing observations for each grid point value (and equivalently each observational point value during the spatial consistency checking) to make this value as accurate as possible. The computer time for solving the system of linear equations for the interpolation weights increases cubically as a function of the number of influencing observations. This means that only a rather limited number of observed values can be selected to influence each grid point value. In the analysis system for the Swedish limited area model the maximum number of influencing observations is presently set to ten (10).

Since it is not realistic to check all combinations of this limited number of influencing observations to get the best combination, it has been necessary to construct empirical rules for selection of influencing observations. These rules are described below.

No 'super-observations', in form of mean-values of closely situated observations, are formed.

4.3.1 Data selection for the mass-field analysis

For the analysis of the mass-field above the surface, radiosonde observations are considered to be the main observational data source. Therefore, radiosonde observations are first selected in the vicinity of the grid-point to be analyzed. The following selection algorithm is utilized (see also Figure 2):

(I) Select the closest radiosonde observation in each of the four quadrants around the gridpoint provided the two-dimensional correlation between the selected observations and the corresponding gridpoint value is larger than a specified lower limit (= 0.75).

(II) Divide the remaining available radiosonde observations into subsets according to a network of squares centered around the gridpoints. Select radiosonde observations, one from each square, in order of increasing distance from the gridpoint until the correlations between the observations and the gridpoint values are below the specified limit (= 0.75). Avoid those squares where observations were selected by rule (I).

(III) Select further remaining radiosonde observations from the squares enumerated during selection rule (II)!

(IV) Select observations from squares situated outside those enumerated by rule (II), thus from squares where the correlation between the observations and the gridpoint value is less than the specified limit (= 0.75).

During the process of selecting radiosonde reports, a data density index and a data distribution index are computed. The data density index is the number of selected radiosonde observations with a correlation to the gridpoint value being larger than the specified limit (= 0.75). The data distribution index is the number of quadrants with selected radiosondes having correlations to the gridpoint value which are above the specified limit. If the data density index is at least at or above a certain value (= 6) the analysis is performed two-dimensionally with radiosonde-observations from the actual analysis level influencing the analysis only (if, in addition, the data distribution index is above a certain value (= 3), the analysis is performed with a simple distance-weighting method). If the data density index is below the required value (= 6), analysis is performed 3-dimensionally. Four (4) radiosonde geopotential observations are selected, according to the selection rules above, and three (3) sea-level pressure values together with three (3) satellite thickness values are furthermore selected. The selection of these additional data is based on distance between the gridpoint and the observational positions. If data from the analysis level are missing in the radiosonde reports, data from a level in the vicinity of the analysis level are selected.

For the analysis of the lowest geopotential level (1000 mb), radiosonde data and sea-level pressure data are selected on the basis of distance between the gridpoint and the positions of the observations.

4.3.2 Data selection for the wind-field analysis

Data selection for the wind-field analysis is simplified by the fact that only wind observations are utilized.

Presently, observed wind vectors are selected three-dimensionally on the basis of stream-function correlation between the gridpoints and the positions of the observations. Only one level of data is selected from the multi-level observations for the analysis of each level.

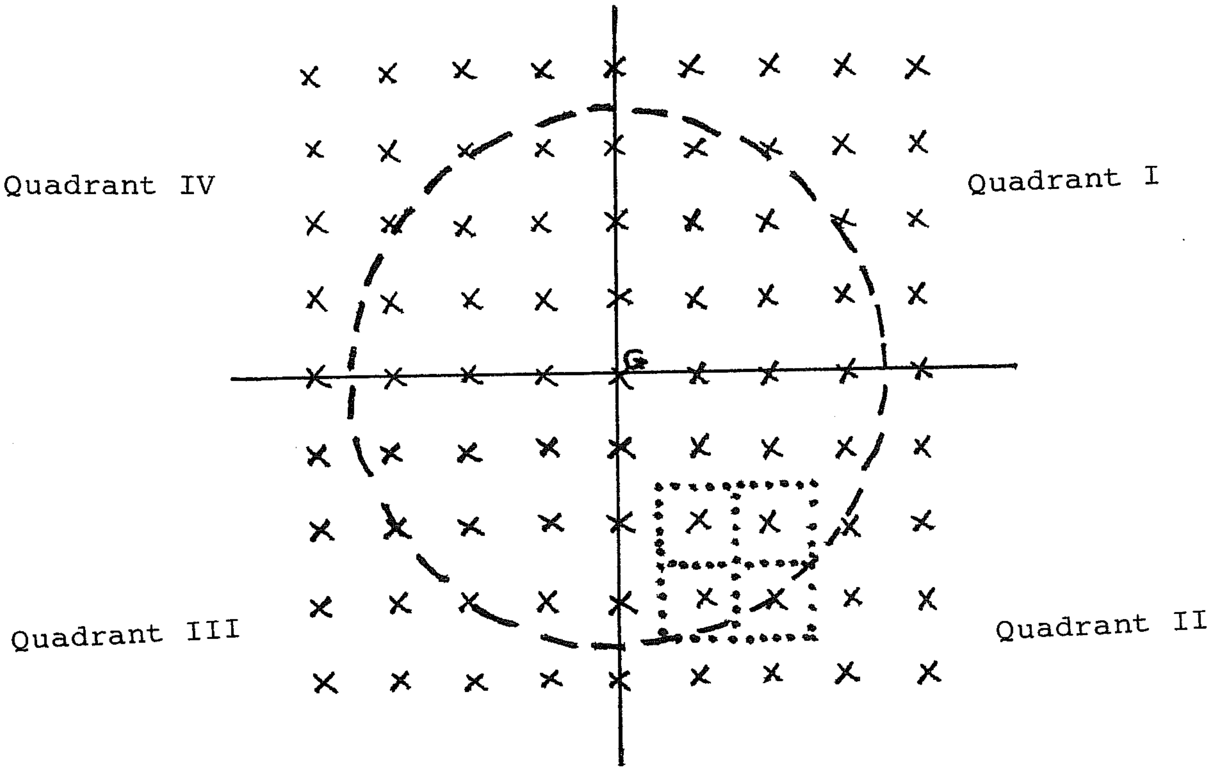

Figure 2 Geometry of data selection for the mass-field analysis (see text)

x Gridpoints

------ Distance at which correlation between observations and gridpoint G is 0.75

——— Division of the area into quadrants

...... Gridpoint-boxes for sorting of observed data (a few are indicated only)

4.4 Space consistency quality control

All observed values, which have passed the check against numerical forecasts, are checked against interpolated values obtained by the analysis procedure from observed values in the vicinity of the observed value to be checked. The observed value to be checked is excluded from influencing the interpolated value.

Suppose we are going to check the observed value f_i^{OBS}. By the statistical interpolation scheme we will obtain an interpolated value f_i^{INT} and an estimate of the mean square interpolation error E_i^{INT}. In addition, we will utilize the estimated variance σ_i^2 of the 'natural' observational errors associated with f_i^{OBS}. The following checking algorithms are applied:

(a) If $\left| f_i^{OBS} - f_i^{INT} \right| < K_f \cdot \sqrt{E_i^{INT} + \sigma^2{}_i}$, the observed value f_i^{OBS} is accepted to be used in the spatial interpolation for the gridpoints.

(b) If $\left| f_i^{OBS} - f_i^{INT} \right| \geq K_f \sqrt{E_i^{INT} + \sigma^2{}_i}$, it is considered that there exists a discrepency between the observed value f_i^{OBS} and those observed values influencing the interpolated value f_i^{INT}.

In case of (b) it is not clear from the first application of the checking algorithm whether it is f_i^{OBS} or any of the influencing observations, say g_j^{OBS} $j=1, \ldots,$ N, which is (are) the cause of the discrepency. Therefore, spatial interpolations and application of the checking algorithm are repeated, first with g_1^{OBS} excluded, and then with g_2^{OBS} excluded, and so on, until:

(c) All g_j^{OBS} $j=1, 2, \ldots,$ N have been excluded with the same interpolation and checking result (case (b) above). In this case f_i^{OBS} is rejected for further use in the analysis processing.

(d) For a certain observed value g_j^{OBS} excluded from influencing the interpolated value f_i^{INT}, the criterion of algorithm (a) is fulfilled and the observed value f_i^{OBS} is accepted for further use in the analysis processing.

5. Data collection delay problems in utilizing data from the GTS

The present operational collection of observational data via the GTS includes certain time delays which are crucial for the application of Limited Area Models:

(a) Surface and upper-air reports from some remote stations are collected by radio communication.

This procedure involves a manual treatment of the data which may cause delayed data distribution and increased probability for introduction of errors.

(b) Satellite sounding data produced from the TOVS-instrument on board the NOAA-satellites enter the GTS with a time-delay of several hours. This time delay is caused by the necessity to store the data on board the satellite before transmission to ground stations. The processing from raw radiance data to temperature and humidity profiles at NOAA-Washington is another time delaying factor.

(c) Retrieval and localization of drifting buoy data through the ARGOS system also require time delays of the order of hours.

These time delays in data collection, data distribution and data reception were found to be most crucial for application of the Limited Area Model at the Swedish Weather Service. Due to requirements on availability from the operational forecasters and due to the limited computer resources, it was decided to run the Swedish LAM with a data cut-off time of only 2 hours and 30 minutes. The operational experience with this early cut-off time has been quite negative. It was found that radiosonde reports fom the Arctic areas (Greenland, the Norwegian Sea and Northern Canada) and from the North Atlantic Weather Ships often were missing in the operational LAM runs. (A short test period indicated a data recovery of only 30%). In addition, satellite sounding data never seem to arrive in time for the operational LAM runs and also many drifting buoy reports arrived too late via the GTS. Although no thorough study on the impact of these effects of the early data cut-off time on the forecast quality has been carried out in Sweden, we are quite convinced that the resulting data loss is one of the main reasons for many poor forecasts produced by the Swedish LAM in comparison with e.g. the ECMWF forecasts. Figure 3, showing data distributions for one case comparing 2 1/2 hour and 5 hour cut-off times, clearly illustrates the problem of obtaining observational data for short-range forecasting with Limited Area Models.

If resources are made available, there are certainly measures which can be taken to improve the observational data sets for short-range forecasting with Limited Area Models:

(1) Delay the start of operational LAM runs until sufficient GTS data are available. Availability requirements on the forecast products may require more computer resources for the forecast model computations.

(2) Rerun the data assimilation cycles (analysis and 6 hour forecasts) to obtain improved first guess fields for the next forecast runs.

(3) Obtain certain observational data by direct read-out from satellites (e g drifting buoy data and satellite vertical sounding data).

(4) Improve critical data collection and data distribution functions of the WWW.

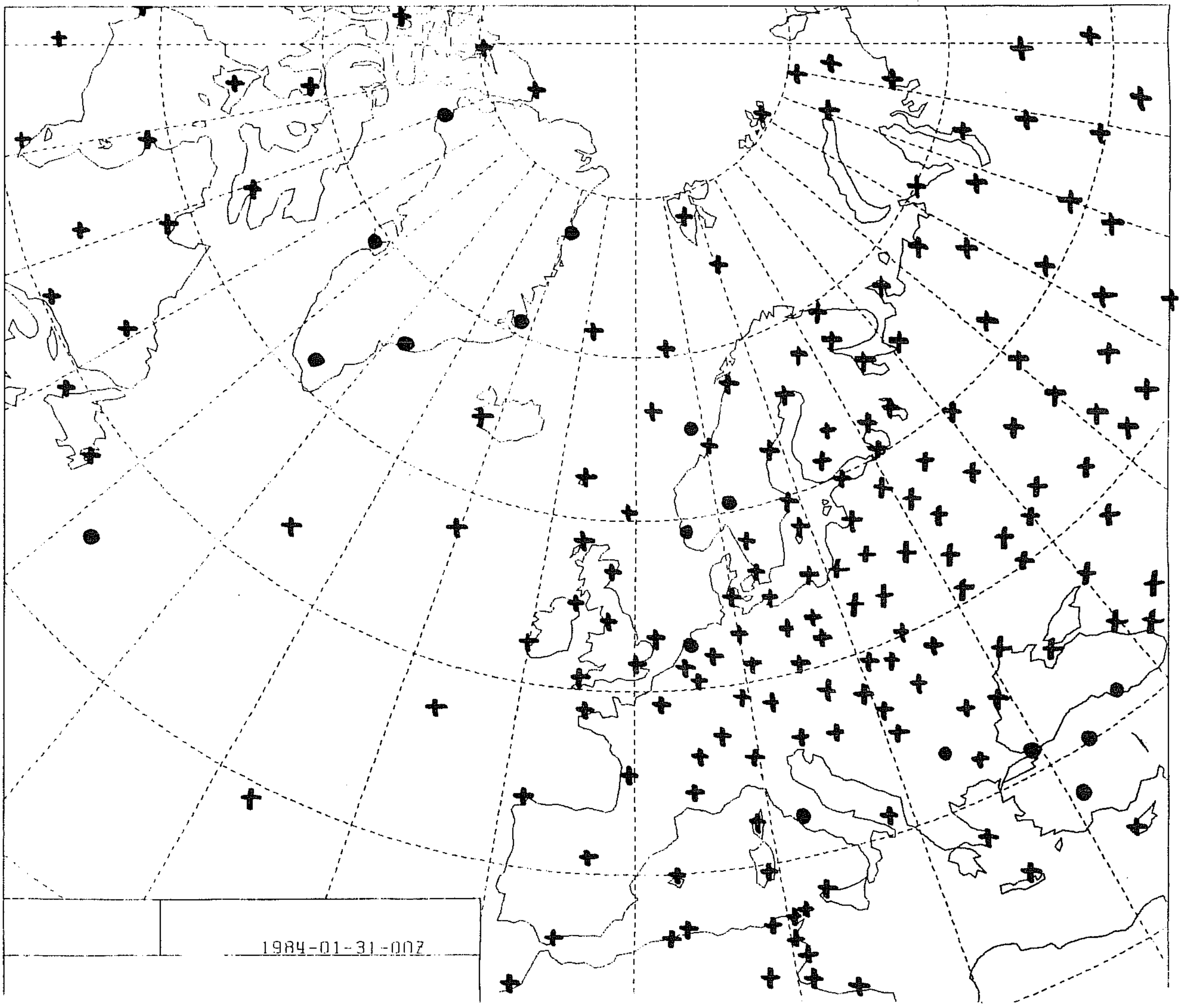

Figure 3 Observational data distributions for the Swedish LAM analysis runs (31 January, 1984, 00 GMT)

+ Cut-off time = 2 1/2 hours

● Additional data with cut-off time = 5 hours

6. Data selection and quality control for meso-scale analysis

Methods for meso-scale objective analysis are presently being developed at SMHI. The objectives of this development work are

* Determination of initial data fields for simple very short-range forecasting methods based on advection and/or extrapolation.

* Presentation of diagnostic maps to the forecaster as a forecasting tool for nowcasting and very short-range weather forecasting.

Initially, a rather pragmatic approach was taken for this development work. The LAM analysis system has been converted into a meso-scale analysis system applied on a grid with a horisontal resolution of 0.2^{o} x 0.2^{o}. The analysis area covers the Southern part of Sweden and surrounding sea and coastal areas. Main changes in the spatial interpolation algorithms are:

* Replacement of the Gaussian horizontal correlation functions by Bessel-function correlations in order to obtain a better representation of meso-scale structures.

* Introduction of coastline-dependent anisotropic correlation functions for analysis of 2 meter temperatures and relative humidities.

As regards data selection and data quality control the following changes and experienced problems are of interest:

(1) Large parts of the Baltic Sea are rather data-sparse when considering detailed analysis of meso-scale structures. There are only a few reports transmitted from ships in the Baltic Sea. A large percentage of these ship reports turned out to be rejected during the pre-analysis processing of the data (erroneously coded section 0 of the SHIP code, ship reports considered to be situated on land areas, a probable error in the software for elimination of duplicate reports, etc).

(2) Due to the relative sparseness of data over the Baltic, the local data selection during the analysis of sea-level pressure over the Baltic turned out to be a problem. When the set of selected sea-level pressure observations, mainly from coastal and island stations, was gradually altered in the middle of the Baltic (using the operational LAM data selection algorithms) discontinuities in the meso-scale sea-level pressure analysis were created.

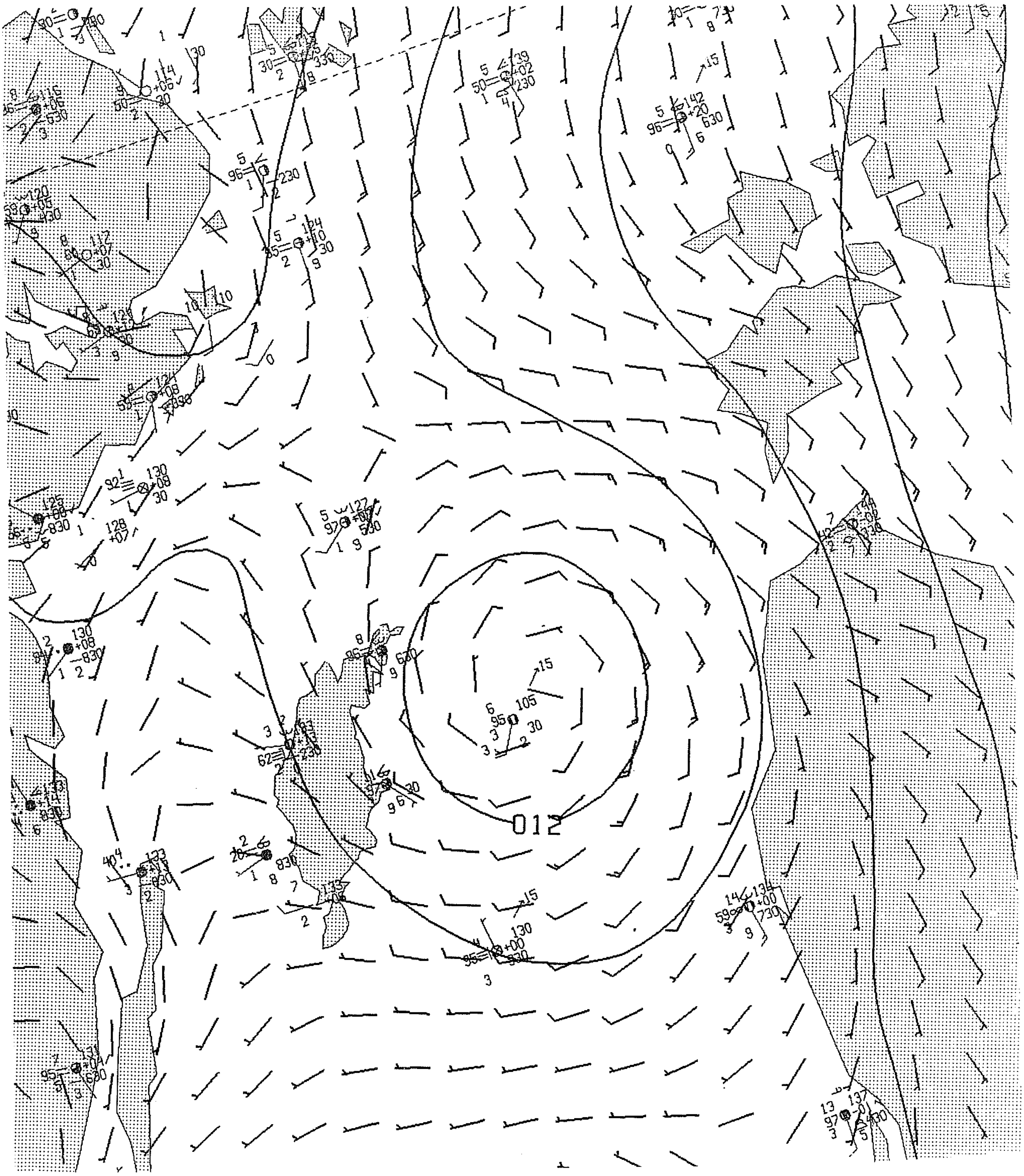

Figure 4 Example of SHIP report in the Baltic Sea causing problems for spatial consistency quality control of sea-level pressure. (12 April 1984 12 GMT). Isolines of sea-level pressure and wind derived by a simple dynamic simulation model are included in the figure.

This data selection problem is obviously present in any analysis scheme based on local data selection, but the problem appears very clearly only when high-resolution gridpoint fields are presented. A temporary solution was obtained by forcing the analysis scheme to select at least one sea-level pressure observation from each of the eight 45 degree sectors around the gridpoint. In a longer perspective, a data selection algorithm based on the 'box technique' should be preferable for meso-scale analysis. Also from a computational economics point of view, the box technique should be preferable since several gridpoints necessarily will use the same observational data in a meso-scale analysis grid.

(3) The spatial consistency quality control of observed data turned out to be crucial for the quality of the meso-scale analyses. As an example, an observational error of the order of 1 mb may be quite harmless to a synoptic scale objective analysis while the same error may cause significant problems for a meso-scale analysis for which observed differences of the order of 0.5 mb are of interest (see Figure 4). Especially it was found that the few available ship reports from the Baltic Sea were of poor quality.

(4) By making statistical evaluation of rejected data during the meso-scale analysis of sea-level pressure, it was found that data from some stations were rejected very often. A detailed manual analysis of the same data has confirmed that these stations have systematic deviations in sea-level pressures from surrounding stations. Absolute values of these deviations are of the order of 0.5-1.0 mb. The reasons for these systematic deviations are presently being examined. This simple monitoring of the performance of the surface station data will continue and also be generalized to other observed parameters than the sea-level pressure.

(5) It turned out to be difficult to utilize the reported 10 meter winds from the Swedish surface stations within the framework of the present meso-scale analysis system. There are several reasons for these difficulties, e.g.:

* Many wind observations represent micro-scale rather than meso-scale variations.

* Many surface stations are not equipped with wind-measuring instruments.

* The present analysis scheme uses non-divergent representation of the spatial correlations for the wind field.

At present, 10 meter wind fields are obtained by a simple dynamic model for simulation of low level winds (Danard, 1977).

7. Summary and conclusions

Data selection and quality control algorithms used at SMHI have been presented and discussed. Operational software for quality control of meteorological data suffers from methodological limitations (and probably also from trivial programming errors). Most operational quality control algorithms are based on 'ad-hoc' formulations. There is a need for development of a theoretical framework for quality control algorithms.

Limitations of man-power at SMHI do not allow for much development of data selection and quality control algorithms. As regards pre-analysis quality control algorithms, an international effort to create a standard software package should be encouraged.

All efforts for monitoring the performance of various observing systems are fully supported.

References

Dahlström, B., Ehlert, K., Gustafsson, N., 1980, A basic system for quality control of observations. Internal SMHI report, November, 1980.

Danard, M., 1977, A simple model for meso-scale effects of topography on surface winds. Monthly Weather Review, 1977.

Gustafsson, N., Törnevik, H., 1984, Development of an operational system for very-short-range forecasting at SMHI. Proceedings of the second International Symposium on Nowcasting, Norköping, Sweden, 3-7 September, 1984, pp. 473-477.