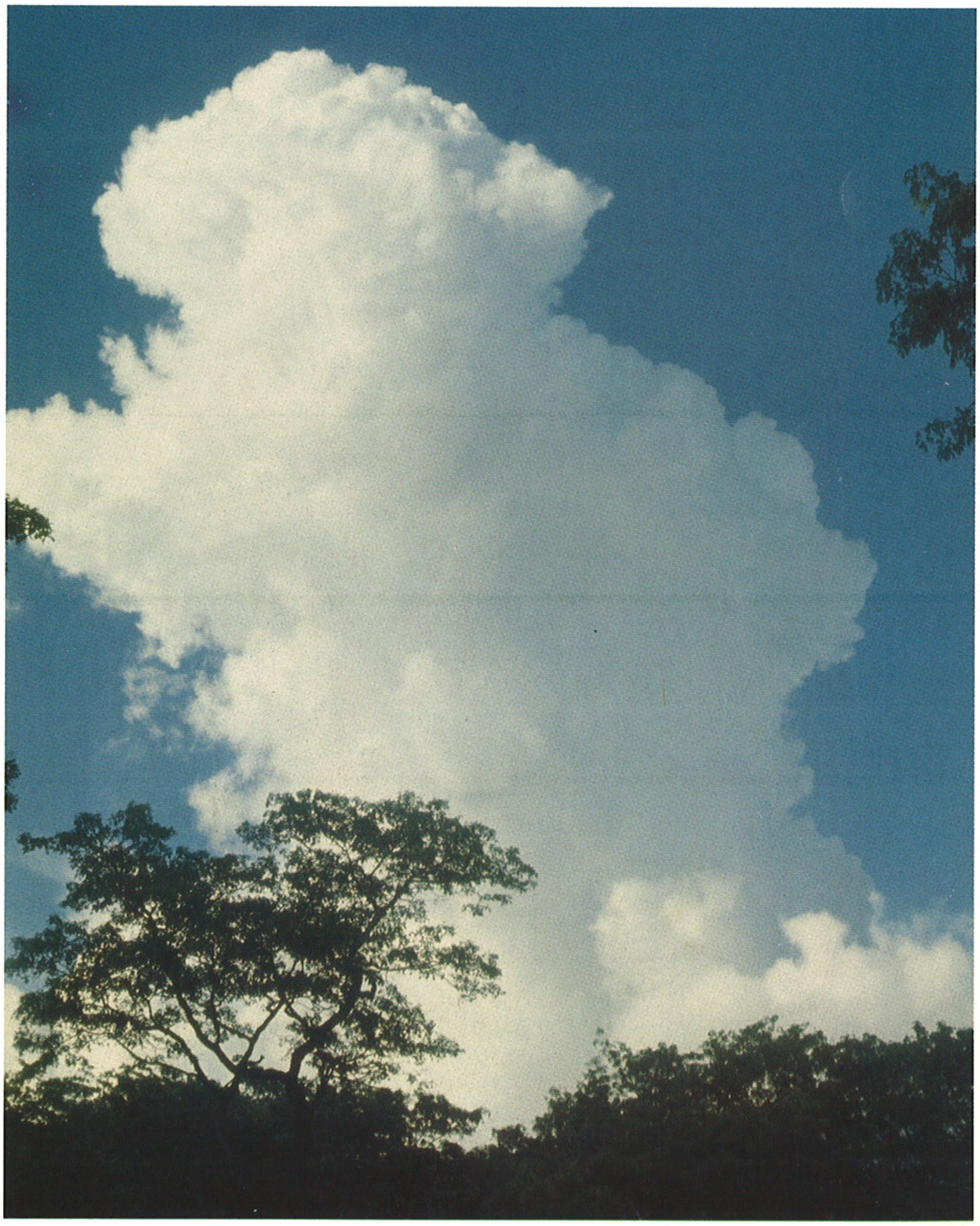
DNMI
Det norske meteorologiske institutt

NORDKLIM: Nordic Co-Operation Within Climate Activities

Quality Control of Climate Data: An Automatic Approach

Report no. 12/99

Petter Øgland



KLIMA

DNMI - REPORT

ISSN 0805-9918

REPORT NO.

12/99 KLIMA

DATE

Mar 18 1999

NORWEGIAN METEOROLOGICAL INSTITUTE
BOX 43 BLINDERN N-0313 OSLO

PHONE: +47 22 96 30 00

TITLE

Quality Control of Climate Data: An Automatic Approach

AUTHOR

Petter Øgland

PROJECT CONTRACTOR

DNMI - Climatology Division / NORDKLIM

SUMMARY

This report is a part of the NORMET project NORDKLIM «Nordic Co-Operation Within Climate Activities» and is intended as a documentation of parts of the quality control documentation for project task no. 1.2 «Quality Control».

The KLIMA_KONTR and PRECIP_KONTR programs are used at DNMI for mimicking the manual processing of climate data. The programs consist of running quality control subprograms, making output analysis, updating test data tables with duplicate climate data and collecting evaluation statistics.

The programs make use of concepts and methods of Artificial Intelligence (AI) and Statistical Process Control (SPC).

KEYWORDS

1. Quality control
2. Statistical process control
3. Climate data
4. NORDKLIM

SIGNATURE

Petter Øgland

Research Scientist

Bjørn Aune

Head of Climatology Division

Table of Contents

1. INTRODUCTION

In the NORDMET-project NORDKLIM: *Nordic Co-Operation Within Climate Activities*, there are two main tasks. Task 1 is dealing with **Climate Data** and has four subtasks (1.1 Network design; 1.2 Quality control; 1.3 Operational precpitation correction; 1.4 Long-term datasets). At a meeting in Oslo in January this year, the NORDKLIM Advisory Commitee decided that highest priority within Task 1 should be given to subtask *1.2 Quality Control (QC)*. All Nordic countries have an urgent need for improved systems for controlling climate data and correcting suspect values. The main aim in 1999 for Task 1.2 is according to the NORDKLIM project plans to work out a *joint report on Nordic algorithms for QC of climate data, and suggestions for QC routines on real-time data (incl. data from Automatic Weather Stations)*.

This report is intended as a documentation of the KLIMA_KONTR and PRECIP_KONTR simulation programs and aid for data processing of climate data, developed at the DNMI. The systems are designed in order to simulate how climate data are manually quality controlled and processed at the DNMI by use of running quality control programs, analysis programs, updates in test data tables and producing evaluation statistics.

We begin with a discussion of technical details concerning automatical data processing of climate data. The first section defines concepts and methods of artificial intelligence applied here and is followed by a section on quality control of climate data at DNMI. This is then followed by sections on the KLIMA_KONTR and PRECIP_KONTR simulation systems, describing how these particular systems have developed and how they perform. The main results are summerised in the last section.

All experiments and results relate to the KLIBAS climatological database system at DNMI, which is built over an Oracle database system and running in a client/server architecture with personal computers linking up to a Silicon Graphics mainframe.

2. ARTIFICIAL INTELLIGENCE

This chapter is, unless otherwise stated, based on a book on Artificial Intelligence (AI) for students of psychology (Chaudet and Pellegrin, 1998). The authors are research scientists at the Laboratory of Biomathematics and Computer Science at the Medical Faculty of Marseilles.

The term *intelligence* in our setting is purely technical, meaning that if we are unable to distinguish the process operative by the computer from an identical process operated by an intelligent human being, the computer is artificially intelligent. The test or distinguishing process is often refered to as the *Turing test* and is used as a basic criterion in applied research.

Although Chaudet and Pellegrin write interestingly on various aspects of AI, only the first five chapters of their book are commented on here, leaving natural languages, neural networks and robotics for later studies as these areas do not seem all that relevant at the moment, although neural networks, in particular, may become an important issue when dealing with algorithms for interpolation and quality control at a later stage.

Chaudet and Pellegrin define AI as *une disipline dont l'objectif est l'étude et la construction de systèmes artificiel de traitement des connaissances* (a disipline for objectively studying and constructing artificial systems for handling knowledge/consciousness). The reason for applying AI in the field of quality control of climate data at DNMI is that the *process* of quality control of climate data is something that seems hard to computerise by the traditional formal approach of system design. Firstly, it is complex and even growing more complex as more data need to be handled, and, secondly, in its present state much of the quality control relies on subjective decision making, in other words, consciousness.

By keeping the focus on consciousness, as suggested by Chaudet and Pellegrin, the specific techniques of AI may be related to

1. *Representation of consciousness.* In the context of quality control of climate data this is how to represent the internal states of the quality control system so that it is aware of what it is doing.
2. *Handling consciousness.* This is how to alter between states, how to make decisions, the reasoning and emotions of the system.
3. *Architctual organisation of the system.* Different tasks need to be done by the system, such as running quality control programs, analysing the results, comparing with other results, making updates and producing evaluation statistics. All these things have to be organised in a practical manner.

4. *The interaction between the system and its environment, investigating, for example, the systems ability to evolve.* As the type, regularity and causes for errors in meteorological observations may create a very complicated environment, the system need to adapt and take advantage of how the environment actually is.

2.1. Representing consciousness

The reason why we want to represent consciousness is that when the system reads information, we want it to act on this information. In other words, consciousness in terms of a quality control system for climate data is what happens between each time the system reads and acts.

Important processes that need to be implemented in this gap between sensoring and acting is that the system has some kind of model or representation of its enviroment in order to anticipate the consequences of its actions. It must have some notion of what it has been recently doing so that it can learn from its interaction with the environment.

It seems particularily important that the system is able to detect changes in the environment, which it may read partly as consequences of its own interaction and partly due to uncontrolled action. In the case of quality control of climate data it is important for the system to have a representation of various aspects of modifications on the data set, both due to updates done by the system and updates done outside the system.

If the system is expected to interact with its environment in a conscious manner, it is necessary that it may be able to reflect on its actions on the enviroment. More so, in order to have the system understand which actions are successful and which are not, it need to have emotions that will enforce successful behaviour and prevent it from repeating unsuccessful behaviour.

2.1.1. Models for representing consciousness

In the field of AI there are four main types of models for representing consciousness.

1. *Mathematical logic.* The main idea here is to have a language that the system may use in order to model its environment and make statements about, trying to clearify what is going on. In terms of quality control of climate data a propriate language could be the language of probability theory. The system could then make assumptions and draw conclusions and actions based on these assumptions.

2. *Semantic networks.* By having the system think of its enviroment in terms of "things" or entities, it may be able to draw connections and see how entities are related. For quality control applications it may be important to compare output from different quality control programs, checking, for instance, both for time consistence, space consistency and internal parameter consistency.

3. *Frames and scripts describing stereotypical action.* As the purpose of automatic data processing is not only understanding why observations seem to be wrong, but also to devise means for correcting these observations, the concept of frames could be useful for modeling behaviour which intially would have to be reflex behaviour but, as it grows due to understanding of the enviroment, could become more sophisticated.

4. *Rules for production describing successive stages and means for solving a problem.* These are special methods used in expert systems.

2.2. Quality control of climate data viewed as a game

Playing games, chess especially, has been much used in the field of AI research, developing frameworks for consciousness with memory, strategy, emotions and so on. Viewing the process of quality control of climate data as a game there may be valuable ideas to be adapted from how AI is used in different games.

Not only taking advantage of stepping into a field that has already been explored with success, it seems likely that viewing quality control of climate data as a game may generate ideas and solutions by itself as is often the case when humans play games or try to understand games.

In order to explain the ideas of how AI may be applied in games, Chaudet and Pellegrin use tic-tac-toe as an example, introducing terminology for consciousness applied to game playing.

1. *A state and a space of states.* A state is a possible internal configuration of sensor data such as a specific matrix of crosses and naughts in tic-tac-toe or an internal representation of the quality of climate data for the control system, based on representation of knowledge as discussed above. A space of states is the set of all possible states with the possibiliy of adding mathematical structure such as order and metrics. Two states are of particular interest, the initial state and the final state which in the case of tic-tac-toe represents the start and finish of the game. In quality control the initial and final state would mean start and finish of quality control.

2. *The operators.* In order to get from one state to another, certain rules have to be followed, both in tic-tac-toe and quality control. The set of legal moves or

transitions are called the operators.

3. *The game tree.* The set of all operations on the set of states generates the game tree. From the initial state, the set of all operations generate the first of the possible transitions, which by a new set of operations generates a second set of states and so on. The terminal states represent the leaves in this tree. The sequence of operations from one state to another is called a path.

4. *Search strategies.* In order to get from a state where climate data has not yet been controlled to a state where the control is completed, there are different possible paths. The problem of selecting efficient paths is called applying for search strategies. For simple games such as tic-tac-toe the game tree is finit and may be tranversed in finite time. For a more complex game such as quality control, the structure of the game has to be investigated before it is possible to say anything about how to produce search strategies.

2.3. Quality control expert systems

Chaudet and Pellegrin (page 55) present expert systems as a development of game playing, stressing the use of knowledge trees and search strategies in early expert systems such as *Logic Theorist (LT)* and *General Problem Solver (GPS)* developed in the late 1950's. Later trends in expert systems has, however, been to restrict the system to process knowledge over a more limited domain, indicating that the use of expert systems in the field of quality control of climate data may be a practical application.

Expert systems have been particulary popular in medicine, but there are at the moment a vast array of interesting applications that border on the area of quality control. RBEST, an expert system for disk failure diagnoises during manufacturing (Bramer, 1990) seems like a good example of what an expert system for diagnosis for poor quality in climate data could look like.

2.4. Distributed intelligence

When dealing with problems having to do with monitoring and managing the KLIBAS database system by having programs monitoring themselves and communicating with each other, Kelley (1994) was a great inspiration. His focus on social intelligence in systems resulted in a change of view on how to solve problems, as can be seen in certain monthly statistics report (Øgland, 1996c)[1] and by

[1] The introduction to the monthly report was extensively rewritten for the July statistics, including charts explaining the KLIBAS system from a theory of systems point of view, focus on interaction between different modules, communication, abstraction and evolution.

use of a symptoms file and use of e-mail protocols for interaction between programs. An example of this is the version 3.0 of the SYNO_INN program (Øgland, 1996d), anticipating more complex program interaction and communication in 1997 and 1998.

Even though the concepts of intelligent agents or blackboard reasoning, the two subjects that Chaudet and Pellegrin bring up, was not a prime focus in Kelley's book, his reports on social behaviour among entities of different sorts with consequent group intelligence seemed like a very relevant approach indeed when addressing problems related to development of the KLIBAS database system (Øgland, 1997).

The latest additions to this line of thought are the KLIMA_KONTR and PRECIP_KONTR systems (Øgland, 1998c; 1999c) that are to be discussed in sections below. While these systems are designed as systems of multi-agents, main issues discussed by Chaudet and Pellegrin, such as methods of interaction and communication, have not yet been fully explored.

At present the CHECK_CONT2 analysis module (Øgland, 1999b) in the KLIMA_KONTR system is the centre of attention, a map of how the analysis module is to interact and communicate with other programs, such as the CONTSYN2 quality control program, being, in fact, the central issue in order to give the system direction for further development. In order to achieve a viable system, both theoretical and practical work has to be done in the area of applying the concepts and methods of AI, intellgent agents in particular, to quality control of climate data.

3. PROCESSING OF CLIMATE DATA

By the end of 1997 there were 855 operative meteorological stations run by the DNMI or jointly with the DNMI. Of the there were 234 synoptical stations (real-time data) and 621 climate stations.

The routine for processing climate data consists of subroutines for processing traditional weather stations, automatic weather stations, semi-automatic weather stations, precipitation stations, precipitation-intensity stations (Plumatic stations), totalisator stations and evaporation stations. In this document only traditional weather stations and precipitation stations are considered.

3.1. Weather stations

As the routine for data processing of weather stations is currently under revision (Aune, 1999), this section will only sketch some elements of the present routine relevant for the section on automatic data processing to be discussed in a later section.

The basic data processing of weather stations at DNMI is defined by Hesselberg (1945) with a guide by Johannesen (1957) in order to accomodate the use of digital computers. Revised guides have been written later (Andresen, 1984), although most of the KLIBAS software definitions have been based on the crude description given by Håland and Øgland (1994).

At present there are about 150 regular weather stations in the climate stations network. Of these, 124 are synoptic stations in terms of reporting regularily to DNMI at least three times a day, 27 of these being either semi-automatic weather stations or PIO stations (Moe, 1999; Øgland, 1998c).

Every month, observations reports are collected by the DNMI, stations that have not yet been digitalised now being punched into the ALV data table and quality control is ready to commence.

Quality control then, consists of two phases (Andresen, 1984), first the monthly reports are checked, then data are digitalised and checked once more by use of the CONTSYN2 quality control program (Øgland, 1996b; 1999a) by each of the persons responsible for digitalising the data.

As all observations have been checked, a more advanced control is executed by use of the quality control programs CONTSYN1, CONTSYN3, LISTER1, LISTER2, LISTER3, RELFUKT, KONTHUM, TGTN (Øgland, 1996b) and different means of manual interaction and control.

When all data has been thoroughly checked, and each corrected observation has been marked with a flag in the ALV data table, the data set of the month is inserted from ALV into the HLV main storage data tables. Corresponding observations in the real-time TELE data table are overwritten.

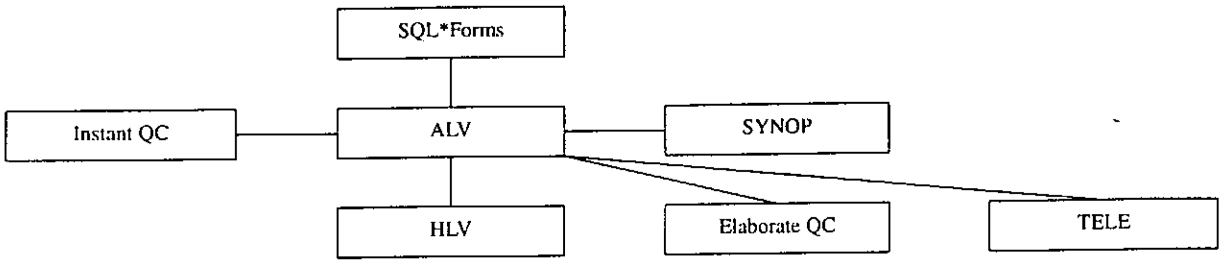


Figure 1. A rough sketch of the ALV data processing routine

Digitalising of reports is done by the SQL*Forms application. Instant quality control is done by use of the CONTSYN2 program. The data set in ALV also contains quality flags for each meteorological element where a flag '1' means that a value has been corrected. Though corrections are made at this stage, the flags are collectively updated at the end of the routine.

The more elaborate quality control is operated by quality control specialists aided by different kind of statistical and quality control software. When all checks have been done, and all observations controlled, the observations in the ALV data table are inserted into the HLN data tables and also into the TELE data table.

3.2. Precipitation stations

The description of the ALN data processing routine is derived from Kjensli (1996) with some updated statistics.

At present there are about 560 precipitation stations in the climate stations network where precipitation is measured on a daily basis, and reported to DNMI every week. As the reports are collected, two persons at DNMI are responsible for digitalising the information to fit into the KLIBAS database system via the data table ALN.

Additional 185 stations record precipitation at least twice a day and are part of the synoptic or automatic station network. Each daily amount of precipitation from these stations are updated from the TELE data table to the ALN data table.

Whenever observations are inserted into the ALN data table, dublicate values are inserted into the main storage tables (HLN).

By the end of the week, the observations are quality controlled, missing values or faulty values interpolated by estimates. The last day of the week, precipitation statistics are sent to groups of regular customers.

When all observations have been collected for a certain month, a new quality control will be performed on this data set. At the end of this final control, the HLN data tables are updated with the improved ALN data set, and the ALN series are deleted.

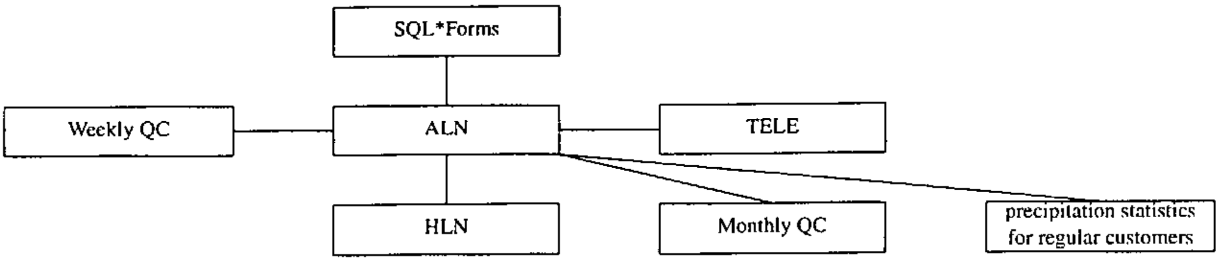


Figure 2. A rough sketch of the ALN data processing routine

Digitalising of reports is done by the SQL*Forms application. Weekly quality control consists presently of the programs RRSSU and GEOK. Both these programs aid the detection of errors which are then updated manually by another SQL*Forms application.

Monthly quality control makes use of a three time iteration of the programs RRUTM and GEOK, updates done in similar manner as with the weekly updates. In both the weekly and the monthly quality control routines, quality flags in the database is updated manually.

4. AUTOMATIC DATA PROCESSING SYSTEMS: KLIMA_KONTR

For data processing of weather stations, a program KLIMA_KONTR (Øgland, 1998c) is being developed. The program simulates the normal procedure of data processing for weather stations in terms of running computer programs for this routine in the same sequence as is done manually and makes an attempt of analysing the output in similar manner to what is done manually and will hopefully also produce suggestions on how the data set should be manipulated in similar fashion to how it is done manually.

4.1. Design for the KLIMA_KONTR system

The KLIMA_KONTR program starts by making a list of stations that have not yet reported to the DNMI. This list is used for explaining missing values and choice of reference stations in areal analysis. The program then run the complete set of quality control programs designed for the routine, performing output analysis with the aid of the programs CHECK_CONT2, CHECK_RELFUKT and CHECK_KONTHUM, finally producing quality statistics in terms of counting quality flags for each station in ALV and ranging the stations according to number of flags.

4.1.1. Output analysis by CHECK_CONT2

So far only output from three of the nine quality control programs that make up the system are analysed, and the type of analysis is very simple in all cases. CHECK_CONT2 is, however, the program using the most advanced methods for output analysis when investigating the CONTSYN2 files.

In CHECK_CONT2, output from the CONTSYN2 quality control program is analysed, but so far only problems relating to air pressure is focused upon and even here only a tentative analysis is performed. The program is developing by analysing real cases of what is done in actual quality control of air pressure when the CONTSYN2 program is being used, describing these actions algoritmically, and have the program perform similar analysis whenever a similar problem arises again.

As the complexity of the program increases alarmingly with each problem being solved in this way, it is imperative, however, that methods of AI are being investigated in order to find out if there are methods, algorithms and data structures that could be used in order to reduce complexity and make the system more sustainable to growth and alterations.

Each time the program is run, the total number of problems on the CONTSYN2 file is counted, and daily measurements of this kind are used for producing *Shewhart diagrams* in a similar fashion to how quality control of real-time data is analysed (Øgland, 1999d). Below is a sample case.

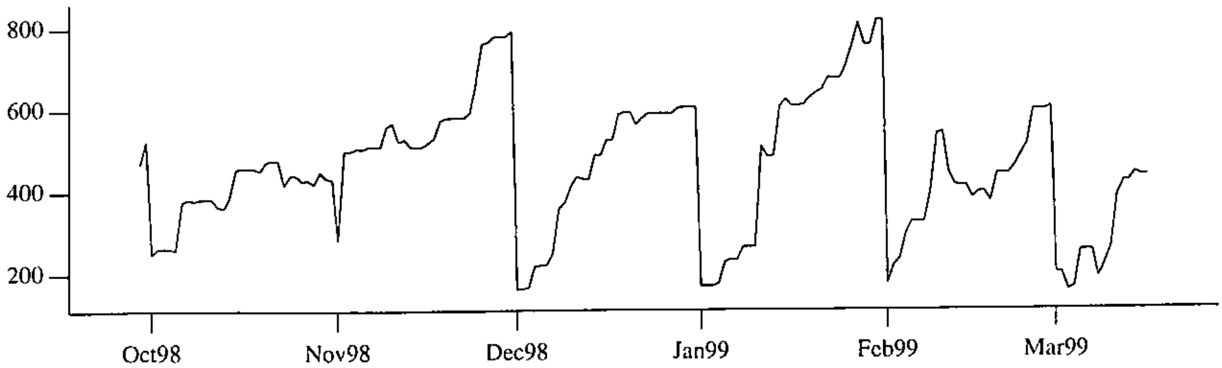


Figure 3. Day by day sum of problems recorded by CONTSYN2

The curve in figure 3 shows a gradual growth during each month as more and more uncontrolled data are inserted into the database system. The curve peaks in the end of November and the end of January, meaning that the total number of problems recorded during these two time intervals is the greatest total so far.

In order to emphasize on problems for particular "worst case" stations, the number of problems assosiated with one particular station is also recorded for each run. A list is then made up of the five most problematic stations on a daily basis.

In the most recent statistics we see, below, that the five most problematic daily instances related to one particular station corresponded with the first peak in figure 3 where also the total number of problems reached a local maximum. All instances in this case relate to the station no. 80101, SOLVÆR - SLENESET.

```
1. 1998.11.21: Number of problems recorded by CONT2 for station no. 80101 = 117
2. 1998.11.20: Number of problems recorded by CONT2 for station no. 80101 = 117
3. 1998.11.19: Number of problems recorded by CONT2 for station no. 80101 = 117
4. 1998.11.18: Number of problems recorded by CONT2 for station no. 80101 = 117
5. 1998.11.17: Number of problems recorded by CONT2 for station no. 80101 = 117
```

Presently, however, station no. 10400, RØROS, actually represents the worst case with 60 problems. With each run of the program, the worst station of the day is presented with relevant extract from the CONTSYN2 file.

Even though this present use of statistical process control is useful for a minimum monitoring of how the stations perform according to the CONTSYN2 quality control, as quality control of the previous month is not completed by the end of the next month, there are very few instances of the curve in figure 3 declining. The process is restarted at the beginning of each month, no matter if past data have been fully processed or not.

At the present stage, however, the statistical process control (SPC) is not defined adequately for the CHECK_CONT2 program. The purpose of the program is to analyse the CONTSYN2 file, so the SPC should really have been related to this task and not on data quality. The present SPC should, in other words, be totally redefined, exporting the current SPC as a byproduct of the CONTSYN2 program itself.

4.1.2. Output analysis by CHECK_RELFUKT and CHECK_KONTHUM

Output analysis for the quality control programs RELFUKT and KONTHUM is done in a similar manner to how CHECK_CONT2 works. The programs CHECK_RELFUKT and CHECK_KONTHUM programs are used (Øgland, 1998a; 1998b), although the techniques are even less sophisticated than those used by CHECK_CONT2.

CHECK_RELFUKT is the most advanced of the two programs. It identifies the station with the greatest deviation from the average relative humidity profiles given by the RELFUKT program, and present the numerical values for this station on the RELFUKT file in a graphical manner.

The CHECK_RELFUKT program is run systematically as an appendix to the RELFUKT program in the weather stations data processing routine, producing visual information for the routine that may help processing relative humidity. In its current state the program is not important in the KLIMA_KONTR system, as there is no way that KLIMA_KONTR can process the visual information. It is run as a part of this system, nevertheless, giving aid to manual inspection and monitoring of the system.

For the same reasons, the CHECK_KONTHUM program is not yet central to the KLIMA_KONTR system, although its statistical analysis of output from the KONTHUM files should prove useful as soon as an efficient communication between the CHECK_KONTHUM, CHECK_RELFUKT and CHECK_CONT2 programs is established. The CHECK_KONTHUM program is of importance for statistical process control where the general state of affairs may be described on a day to day basis and extreme cases more easily picked out for investigation.

4.2. Results from the KLIMA_KONTR system

The statistical output from the KLIMA_KONTR system consists of a ranking list of stations according to data quality measured from counting quality flags. At present quality flags are updated both manually and automatica.ly, automatical update done for PIO weather stations and Semi-Automatic Weather Stations (SAWS) when automatic alterations in the ALV table is done by the ALA2ALV program (Øgland, 1998e).

At the beginning of each month, quality flag statistics are presented for the previous month in the KLIBAS statistics report (Øgland, 1999b). Below is an extract of the statistics for February 1999, listing the 20 stations with the most quality flag updates during the past 12 months.

	Stnr	Name	FLTT	FLTN	FLTX	FLUU	FLRR	FLFF	FLDD	FLFX	FLFG	FLSS	FLSD	FLVV	FLN	FLP0
99910	NY-ÅLESUND	30	14	14	24	216	34	34	132	132	1474	0	5	0	30	30
50540	BERGEN - FLORIDA	825	417	416	2	1	27	27	198	18	0	0	1	1	1	2
10400	RØROS	1	2	1	123	41	220	221	668	0	7	4	221	221	1	33
39040	KJEVIK	11	8	8	2	409	6	6	487	487	0	1	0	0	10	10
16610	FOKSTUA II	37	37	37	80	2	39	41	468	468	0	3	2	2	35	36
43010	EIK - HOVE	69	94	94	267	5	75	75	193	184	4	5	11	11	69	69
80610	MYKEN	9	13	14	20	1	25	25	389	377	0	0	3	4	22	22
80700	GLOMFJORD	0	3	3	1	185	0	0	15	343	346	0	0	0	0	0
90800	TORSVÅG FYR	16	17	17	29	1	20	20	278	278	0	0	0	0	17	16
99950	JAN MAYEN	7	62	62	5	3	20	20	249	248	0	0	0	0	7	7
48330	SLÅTTERØY FYR	22	15	15	22	8	22	22	273	153	0	0	17	16	11	11
02950	MAGNOR	75	39	40	75	33	75	75	0	0	3	3	75	76	2	2
49580	EIDFJORD - BU	72	33	30	72	60	36	37	38	33	4	1	75	75	1	1
18700	OSLO - BLINDERN	14	7	7	11	408	9	9	14	14	2	0	8	7	9	9
96400	SLETTNES FYR	9	10	10	8	7	10	10	259	156	0	0	10	5	9	9
88690	HEKKINGEN FYR	24	11	8	19	0	24	24	26	269	0	0	2	0	0	0
32920	ØYFJELL - TROVATN	52	5	9	52	5	52	52	53	0	2	2	53	52	0	0
51670	REIMEGREND	39	27	27	41	25	42	42	45	0	0	0	42	43	0	0
86760	BØ I VESTERÅLEN II	41	24	24	42	24	42	44	41	0	0	0	42	45	0	0
65940	SULA	10	7	7	5	10	10	10	117	116	0	0	3	5	10	10

The major reason for collecting statistics of this kind is to put stations in quality categories which would help make priorities for the annual inspections and also input for the Instrument Division in case of low quality for automatic weather stations (Aune, 1992).

Presently more important, however, the material should be used for monitoring the development of the KLIMA_KONTR system. Statistical process control curves should be updated using flag statistics in order to measure and visualise the daily quality level. Presently there are no statistics showing how the routine is, in general, performing over time, and statistics based on quality flags would show how many weather elements are being updated at regular intervals.

Analysis of the SPC curves should in similar way to how quality control of real-time data is monitored and analysed, give vital information needed for developing and improving the KLIMA_KONTR system (Øgland, 1999d)

5. AUTOMATIC DATA PROCESSING SYSTEMS: PRECIP_KONTR

Similarily to the simulation of the weather stations data processing routine, a computer program, PRECIP_KONTR (Øgland, 1999c), has been developed for running the quality control software for the precipitation stations data processing routine (ALN routine) in sequence, make the necessary analysis, make suggestions on how to alter the data set in order better to conform with the quality control and make updates.

5.1. Design for the PRECIP_KONTR system

The idea of PRECIP_KONTR is to be a computerised version of the ALN data processing routine described in the previous section, simulating all aspects of the manual routine from the point when observations have entered the ALN data table.

The basic structure of PRECIP_KONTR consists of copying observations from ALN to ALN2, performing quality control analysis on ALN2 by using the same type of software that is being used for the ALN data table plus the computer programs ROMRR and FREYR, analyse the output and update ALN2, and, finally, produce statistics in order to compare the manual (ALN) and the automatic (ALN2) routines.

5.1.1. Areal quality control: ROMRR

When implementing the data processing routine for daily precipitation on the KLIBAS database system, areal quality control was one of the major concerns. A theoretical study of the method previously used on the ND-100 computer was done (Øgland, 1993; 1994a).

When performing the tests, a non-operative areal check program was run, and two new different systems were made when wanting to put the system in operation (Øgland, 1994b; 1994c). Neither of these are currently being used by the manual routine, but the latter system is an integrated part of the PRECIP_KONTR system.

5.1.2. Expert system for output analysis: FREYR

During the preparations for the first meeting of the FREYR project in Helsinki, January 23.-24. 1997 (Rissanen, 1997), a software framework for an expert system and decision aid on the quality control of daily preciptiation data was implemented at the DNMI.

The FREYR system consisting of an administrative module and two programs EPR (Expert System/Pattern Recognition) and STC (Space-Time-Control). The system is not documented, but theoretical aspects of the system are discussed in the specifications to the Nordic FREYR system (Vejen, Rissanen and Øgland, 1997).

5.2. Results from the PRECIP_KONTR system

The statistical output from the PRECIP_KONTR system, so far, consists of a list of stations where the preciptiation values for ALN are compared with those of ALN2. Only a limited range of stations are being tested at present, and there is no actual quality control control in the system for the moment, neither the ROMRR nor the FREYR program are working properly, so the output is fairly random.

Nevertheless, in order to give an impression on how the results are presented in the monthly statistics reports, a sample ranking is given below.

```
00600 GLØTVOLA:                   rmse=293.9, bias=170.7, count=12
00420 HEGGERISET - NORDSTRAND:    rmse=41.0, bias=43.5, count=12
00250 ØRSJØSETRA:                 rmse=39.7, bias=40.8, count=12
00100 PLASSEN:                    rmse=38.5, bias=35.2, count=12
00730 VALDALEN:                   rmse=36.7, bias=46.6, count=12
00290 TÅGMYRA:                    rmse=35.8, bias=37.2, count=12
00060 LINNES:                     rmse=32.2, bias=32.0, count=12
00700 DREVSJØ:                    rmse=31.4, bias=27.1, count=18
00610 GLØTVOLA - TRØAN:           rmse=28.8, bias=30.1, count=12
00770 ELLEFSPLASS:                rmse=28.5, bias=28.9, count=12
```

These statistics show that among the query stations, station no. 00600, GLØTVOLA, is the one with the greatest root mean square error or distance between ALN and ALN2, and, in this case, it is also the stations with the greatest bias, although only 12 pairs of observations have been used to produce these statistics.

Statistics calculated by the SQL interface code in select codes from the ALN and ALN2 data tables. In this early version of PRECIP_KONTR, the standard error,

$$stde^2 = \sum_{i=1}^{n}((RR^{ALN}{}_i - RR^{ALN2}{}_i) - bias)^2/(n-1)^2 \quad (5.1)$$

has been used as an estimate of *rmse*, using *bias* defined defined as

$$bias = \frac{1}{n}\sum_{i=1}^{n}(RR^{ALN}{}_i - RR^{ALN2}{}_i) \quad (5.2)$$

where n is the number of observations.

Daily comparison of this kind is important for finding difficult stations and interesting problems that may have vital impact on the system when being solved. In order to measure the general development and improvement of the system, however, methods of statistical process control similar to those used for managing and improving the system for quality control of real-time data (Øgland, 1999d) should be applied.

6. DISCUSSION AND CONCLUSION

Concepts and methods of Artificial Intelligence (AI) are presently under consideration for use in data processing of climate data at the DNMI. For the two main climate data processing routines, the data processing for weather stations and the data processing for preciptiation stations, software is being developed in order to simulate and validate the manual run of the routine. The goal is firstly to produce software that will make manual data processing more efficient and secondly automate as many trival parts of the routine as possible.

For data processing of weather stations, a program KLIMA_KONTR is being developed. The program simulates the normal procedure of data processing for weather stations in terms of running computer programs for this routine in the same sequence as is done manually, making an attempt of analysing the output in similar manner to what is done manually, and, hopefully, producing suggestions on how the data set should be manipulated in similar fashion to how it is done manually.

So far, only output from three of the nine quality control programs that make up the system are analysed, and the type of analysis is very simple in all cases. CHECK_CONT2 is the program with the most advanced methods for analysis, analysing air pressure quality warnings on the output files from the CONTSYN2 program. The analysis is very simple, however, but growing and developing.

The development of the CHECK_CONT2 program is based on running the program on CONTSYN2 output and find out what kind of manual reactions the output would generate in real cases. This study, then, result in constructing algorithms for performing or simulating the reactions.

In the case of the precipitation data processing routine, a computer program PRECIP_KONTR is under development. For this program the emphasis has so far been on producing test environment for running an fully automatic system in parallell with the current manual processing routine and to measure the difference in results for the two routines. The PRECIP_KONTR system is planned to follow suggestions from the FREYR project in addition to remodelling the current state of affairs.

Currently, the most important work for making progress is to make a study in the field of AI in order to coordinate the development so far with general experiences in this area of research. Methods of statistical process control (SPC) also need to be used as an integrated part of system monitoring and development.

REFERENCES

Andresen, L., 1984: *Kontorassistentenes arbeid med kontroll av data fra manuelle værstasjoner*, KLIMA-Arbeidsnotat no. 21, DNMI, Oslo, January 1984.

Aune, B., 1992: Note to the databasegroup 17.11.1992 (appendix I in:) *Kontroll av data - spesifikasjonsrapport*, KLIBAS-report no. 01/93, DNMI, Oslo, May 1993.

Aune, B., 1999: *Prosjekt Databehandlingsrutiner (DATRUT)*, KLIMA-note no. 14/99, DNMI, Oslo, March 1999.

Bramer, M. (ed.), 1990: *Practical Experience in BUILDING EXPERT SYSTEMS*, John Wiley and sons, Chichester, UK.

Chaudet, H. and Pellegrin, L., 1998: *INTELLIGENCE ARTIFICIELLE et psychologie cognitive*, Dunod, Paris, 1998.

DNMI, 1998: *Årsbetretnig 1997 (Annual report 1997)*, DNMI, 1998.

Førland, E.J., B. Dahlström, R. Heino, T. Jonsson, H. Madsen, 1998: *NORDKLIM. A Nordic project on Nordic Cooperation with Climate Activities*, DNMI/SMHI/FMI/VI/DMI, Oslo/Norrköping/Helsinki/Reykjavik/Copenhagen, 08.06.1998.

Førland, E.J., 1999: *Minutes from 1st NORDKLIM Steering Committee Meeting*, Oslo, Norway, 27.-28. January 1999.

Hesselberg, Th., 1945: *Regler for arbiedet i Klimaavdelingen 1945*, DNMI, Oslo, June 1945.

Håland, L. and Øgland, P., 1994: *Kvalitetskontroll av værstasjonsdata i Klimaavdelingen*, KLIBAS-report no. 23/94, DNMI, Oslo, October 1994.

Johannessen, Th. W., 1957: *Veiledning i kontrollen av dagbøkene før observasjonene punches*, DNMI, Oslo, April 1957.

Kelley, K., 1994: *Out of Control: the New Biology of Machines, Social Systems, and the Economic World*, Addison Wesley Publishing Company, 1994.

Kjensli, P.O., 1996: *Beskrivelse av nedbørrutinen ved Klimaavdelingen.*, KLIBAS-report no. 03/96, DNMI, Oslo, 1996.

Moe, M. 1999: *PIO STASJONER. Status februar 1999*, KLIMA-report no. 07/99, DNMI, Oslo, February 1999.

Rissanen, P., 1997: *Quality control of daily precipitation data: -definitions, enviroments and programs*, invitation to meeting, Helsinki, Finland, January 1997.

Vejen, F., Rissanen, P. and Øgland, P., 1997: *FREYR. Specification report on a system for quality control of meteorological data*, draft version 1.3, DMI/FMI/SHMI/DNMI, February 1997.

Øgland, P., 1993: *A Method of Weighted Linear Estimation Applied to Quality Control of Precipitation Values*. KLIMA-report no. 42/93, DNMI, Oslo, December 1993.

Øgland, P., 1994a: *Some recent developments regarding quality control of precipitation data in Norway*. Det 19nde nordiske meteorologmøte, Kristiansand, Norway, June 6th-10th, 1994, 179-183.

Øgland, P., 1994b: *Romkontroll for nedbørstasjoner over et statisk referansenettverk*, KLIBAS-report no. 07/94, DNMI, Oslo, April 1994.

Øgland, P., 1994c: *Romkontroll for nedbørstasjoner over et selvreviderende referansenettverk*, KLIBAS-report no. 39/94, DNMI, Oslo, December 1994.

Øgland, P., 1996a: *Automatisering av kvalitetskontroll av geofysiske data ved Klimaavdelingen*. KLIBAS-report no. 04/96, DNMI, Oslo, February 1996.

Øgland, P., 1996b: *Rutine for kvalitetskontroll av klimadata. Versjon 1.1*, KLIBAS-report no. 10/96, DNMI, Oslo, April 1996.

Øgland, P., 1996c: *Driftsrapport juli 1996*, KLIBAS-note no. 16/96, DNMI, Oslo, August 1996.

Øgland, P., 1996d: *Eksperimentell overføring av data fra syno-filer til Oracle-database. Versjon 3.0*, KLIBAS-report no. 15/96, DNMI, Oslo, October 1996.

Øgland, P., 1997: *Logging av driftsparametre for kvalitetssikring av driftsrutiner*, KLIBAS-report no. 45/97, DNMI, Oslo, May 1997.

Øgland, P., 1998a: *Computer program CHECK_RELFUKT*. KLIBAS-report no. 37/98, DNMI, Oslo, August 1998.

Øgland, P., 1998b: *PC i observasjonstjenesten: PIO_INN v.2.2*, KLIBAS-report no. 58/98, DNMI, Oslo, October 1998.

Øgland, P., 1998c: *Analysing Relative Humidity Statistics by CHECK_KONTHUM*. KLIBAS-report no. 38/98, DNMI, Oslo, August 1998.

Øgland, P., 1998d: *Computer program KLIMA_KONTR v.1.1*. KLIBAS-report no. 40/98, DNMI, Oslo, August 1998.

Øgland, P., 1998e: *Updating AWS in ALV and TELE by ALA2ALV v.1.2*. KLIBAS-report no. 66/98, DNMI, Oslo, December 1998.

Øgland, P., 1999a: *CONTSYN2 v.2.0: Some additional checks for the ALV routine*, KLIBAS-report no. 07/99, DNMI, Oslo, February 1999.

Øgland, P., 1999b: *Automatic air pressure quality control by CHECK_CONT2*, KLIBAS-report no. 10/99, DNMI, Oslo, February 1999.

Øgland, P., 1999c: *Interpolation and Quality Control status for February 1999*, KLIBAS-note no. 05/99, DNMI, Oslo, March 1999.

Øgland, P., 1999d: *Attempting to simulate the ALN routine by PRECIP_KONTR*, KLIBAS-report no. 17/99, DNMI, Oslo, March 1999.

Øgland, P., 1999e: *Quality Control of Meteorological Real-Time Data: Results from the S-T-F system*, KLIMA-report no. 11/99, DNMI, Oslo, March 1999.