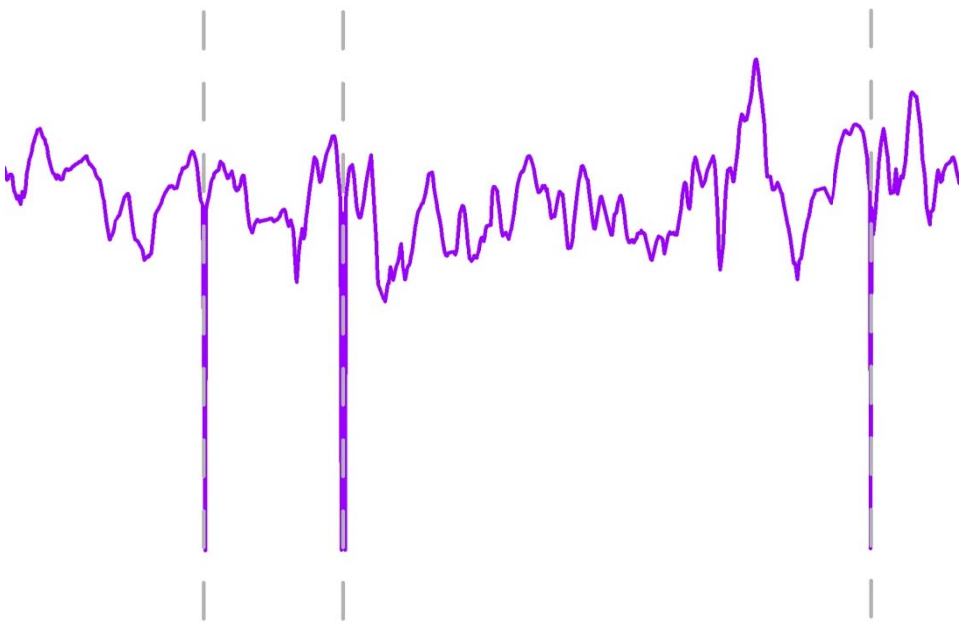
Technical Report 11-16

Quality control of Greenlandic weather and climate data series

1958-2010

Supplement to Technical Report 11-15

Laust Boas and Peter Riddersholm Wang



Copenhagen 2011

Colophon

Serial title:
Technical Report 11-16

Title:
Quality control of Greenlandic weather and climate data series 1958-2010

Subtitle:
Supplement to Technical Report 11-15

Author(s):
Laust Boas and Peter Riddersholm Wang

Other contributors:
Ellen Vaarby Laursen, Claus Kern-Hansen and John Cappelen

Responsible institution:
Danish Meteorological Institute

Language:
English

Keywords:
Greenland, climate data, wind, temperature, cloud cover, air pressure, humidity, precipitation, depth of snow, 1958-2010, Greenland dataset, quality control

URL:
www.dmi.dk/dmi/tr11-16.pdf

ISSN:
1399-1388 (online)

Version:
19.08.2011

Website:
www.dmi.dk

Front page:
Example of erroneous peak values

Content

Abstract

The purpose of this report is to describe the methods used to quality control the data series in DMI Technical Report 11-15. Each parameter is described for characteristics and trends.

Resumé

Formålet med denne rapport er at beskrive de metoder der er brugt til at kvalitetssikre dataserierne i DMI Tekniske Rapport 11-15. Den enkelte parameter er beskrevet for karakteristika og tendenser.

1. Introduction

This report presents the method used in the quality control of the data series published in DMI Technical Report 11-15.

2. Quality control

A description of the method used in the quality control is presented in this section followed by a more detailed description of the quality control for each parameter. The last section presents a short statistical description of the data series.

2.1 Method

All quality control is done visually and errors are corrected manually. Each time series has been plotted individually or together with other related parameters and have been scrutinized for apparent errors, which consequently are omitted from the data series.

The quality control is done subjectively and no permanent exclusion limits are set for observations. The Greenlandic weather stations are located in a wide range of climatologically different sites, and the data series between stations are therefore not immediately comparable. Permanent limits excluding high or low values automatically would therefore need to be specifically set for each station, which is beyond the scope of this report.

A definition of station specific limits would merely catch the extremely apparent errors, e.g. caused by a frozen instrument or an instrument out of order in other ways. Less significant errors e.g. caused by calibration issues would not as easy be located, because the values are not extreme, but merely offsets from the trend of the data series. However, such errors are relatively easy to locate through a visual quality control.

The quality control has been done in two steps. First step was the most detailed with comparison of time series and sporadic research in climate conditions for suspicious events. Second step comprised a visually look-through of the plotted data that resulted from the first step. Corrections have been noted manually as timestamps, and subsequently the observations have been removed from the data series.

2.2 Error types

The errors have not been categorized systematically and statistics of the distribution of errors are therefore not available. Only simple statistics for the length and number of errors are available, see section 2.4.

Nevertheless, five different types of errors have been found, and each type is individually described in table 1.

Name	**Type of error** (typical length of error)	**Description**
Peak values (*or* outliers)	*Single or few consecutive values (single observations)*	Obvious errors, that appears in most parameters as values that clearly are not consistent with the trend of the data series. The values are either implausible high or low. Difficult to detect for *precipitation*, as precipitation is naturally characterized by peak values.
Offset values	*Period of consecutive values (days to months)*	Typically instrument calibration issues cause values to change suddenly to and from a different level of values. Most frequently found in *humidity* data series.
Frozen values	*Period of consecutive values (days)*	When a mechanical instrument (typically anemometer) freezes or is in other ways stalled, values will be the same over a period of time. Frozen values are related to periods with temperatures below 0 °C. Frozen values also covers constant values independent from temperature conditions.
Scattered values	*Period of scattered values (months to years)*	Often appears at the beginning or the end of data series, when instruments are maintained and/or malfunctioning.
Missing values	*Period with no or very limited data (usually years)*	Frequently single or a few values appear in the midst of a longer period of missing values. In this case, the whole period has been omitted, assuming that the values are erroneous.

Table 1 Definitions of five different types of errors

2.3 Quality control of individual parameters

Each parameter will here be described individually to clarify the guidelines used to quality control the data series.

Wind direction - dd
The most frequent error in *wind direction* is caused by frozen instruments. This phenomenon is visible in data series when compared to *temperature* data series, which simultaneously show very low temperatures and always below 0 ˚C. These periods are omitted as illustrated in figure 1.

The *wind direction* data series frequently show peak values higher than 360 (except for 999) and these observations are omitted as well.

Wind speed - ff
There are often correlation between the type of errors in *wind direction* and *wind speed*. The anemometer measuring *wind speed* can freeze to a halt, and either shows 0-values (even though the wind is blowing) or the same value over a period of time.

Figure 1 shows this phenomenon visualized by raw data series for *temperature*, *wind direction* and *wind speed*.

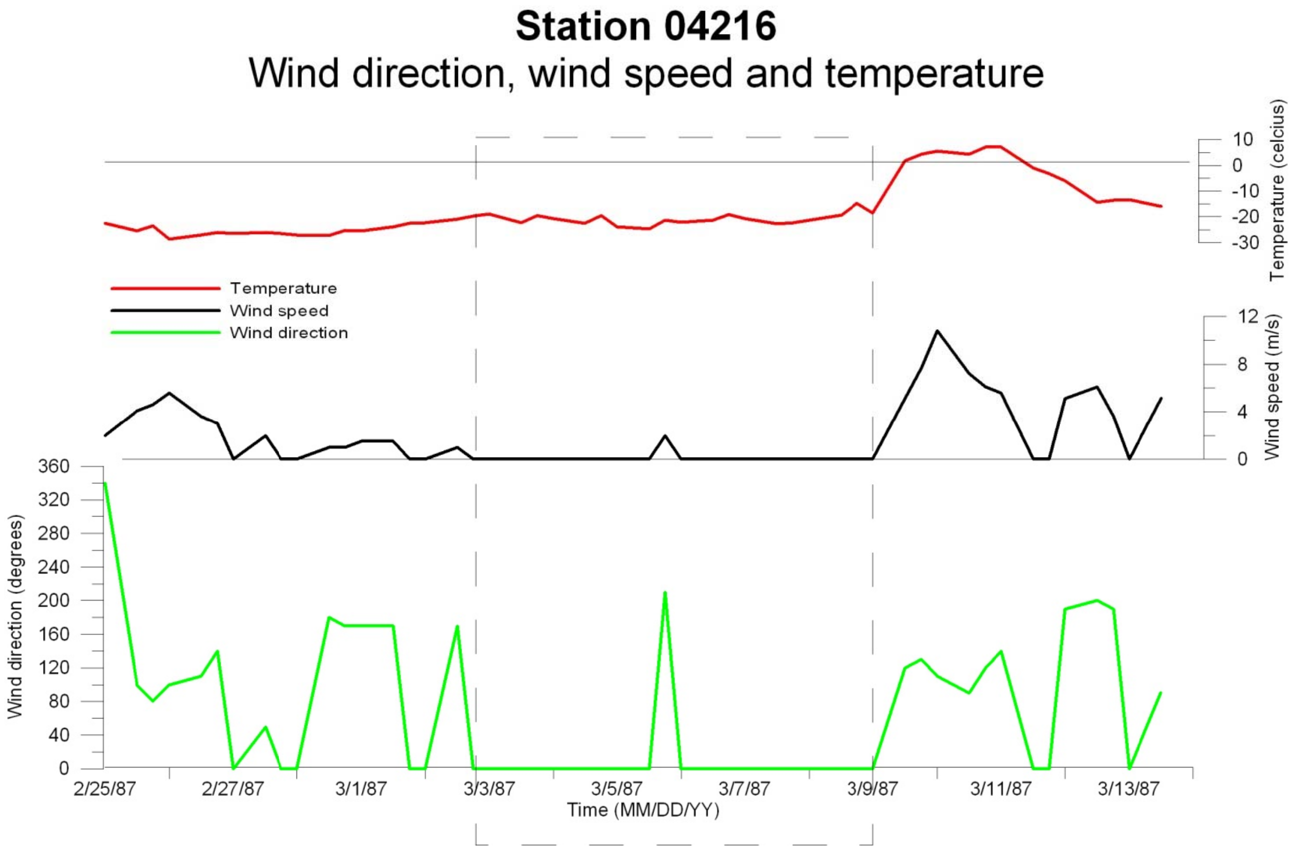


Figure 1 Example of frozen wind instruments. Notice the dramatic *temperature* rise around March 9th causing the instruments to thaw and function again. The dotted box indicates the omitted period in the *wind speed* and *wind direction* data series. A single observation in *wind direction* and *speed* in the middle of the omitted period is registered. This observation is assumed to be erronous too.

Cloud cover - n
Unlike most other data types, no apparent correlation exists between subsequent values in *cloud cover* data series, hence no obvious tendency can be found in the series. In the same way, there is no apparent correlation with other parameters.

Erroneous periods in cloud cover series are often characterized by missing, “frozen”, or scattered single values over a long time, and the periods are often at the end or beginning of the data series. All observations during these periods are omitted, although not all time steps actually include *cloud cover* observations. See figure 2 for an example.

Erroneous values in the coherent part of data series are practically not noticeable, unless the value is outside the interval of 0-9. Some data series have only one or very few values during the entire period, and these values are consequently omitted.

Furthermore, a known specific error has been corrected at some stations. Due to technical difficulties the *cloud cover* instrument has been unable to measure 0 or "cloud free" since approx. 2004, and all observations in these periods have accordingly been omitted.

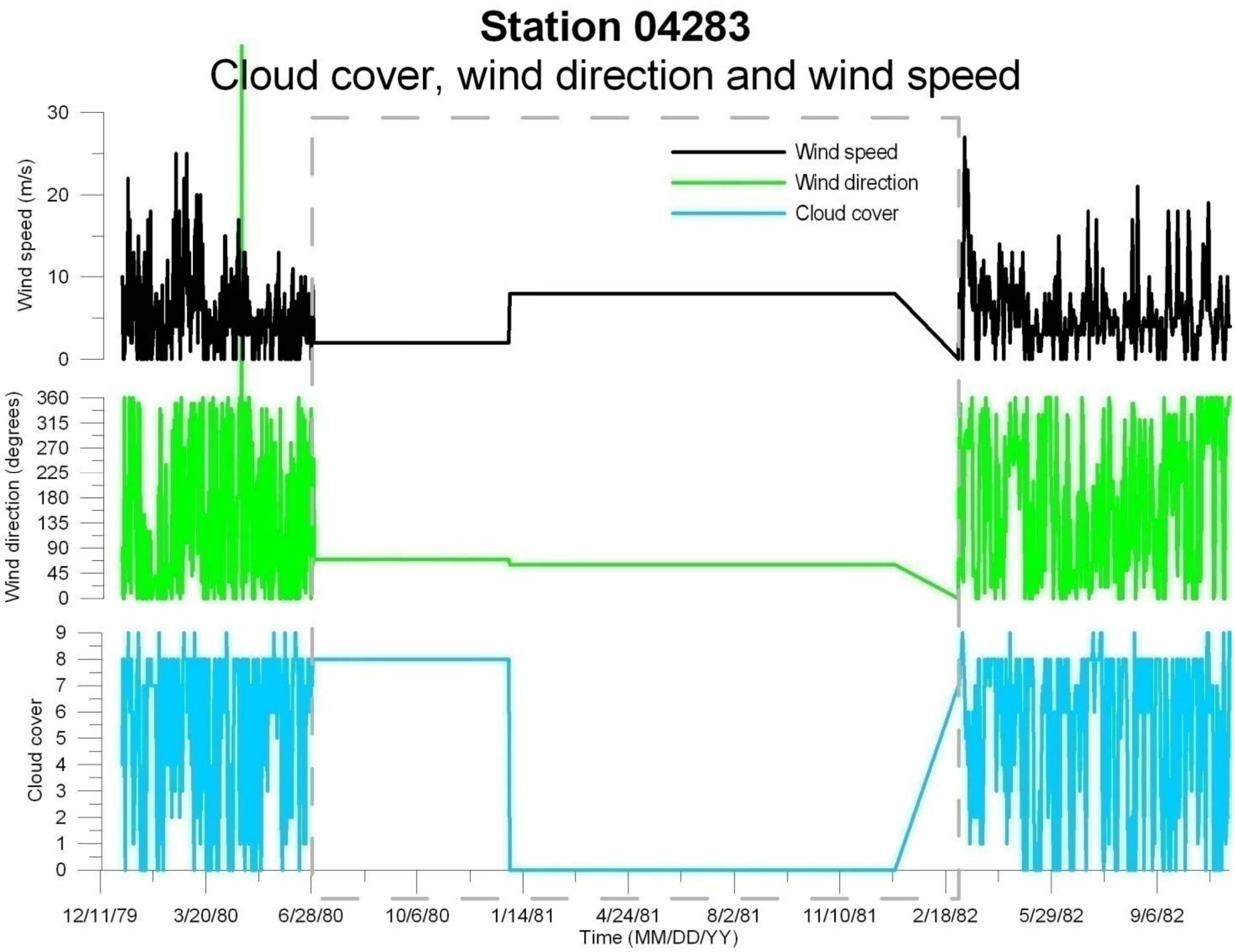


Figure 2 Example of omitted data due to technical issues. The dotted box indicates omitted period in the data series. It is often seen, that more than one parameter shows erroneous values in the same period, as shown here. The high *wind direction* observation in April 1980 is "999", variable directions.

Air pressure - pppp

The characteristic of *air pressure* data series is that they are somewhat homogenous following a trend, meaning that erroneous values are often apparent.

Most errors found are peaks outside the valid interval (defined by the known extreme values) or outside the trend, as the example shows in figure 3.

On the other hand, *air pressure* can fluctuate significantly over a short period of time (hours), and a peak in a data series cannot conclusively be marked as erroneous. In some situations *air pressure* is compared to *wind speed* trend, but also parameters as *wind direction*, *temperature* and *humidity* can be evaluated. If these parameters show corresponding fluctuations, the peak in data series is most likely caused by a sudden change in weather conditions, and consequently is not omitted.

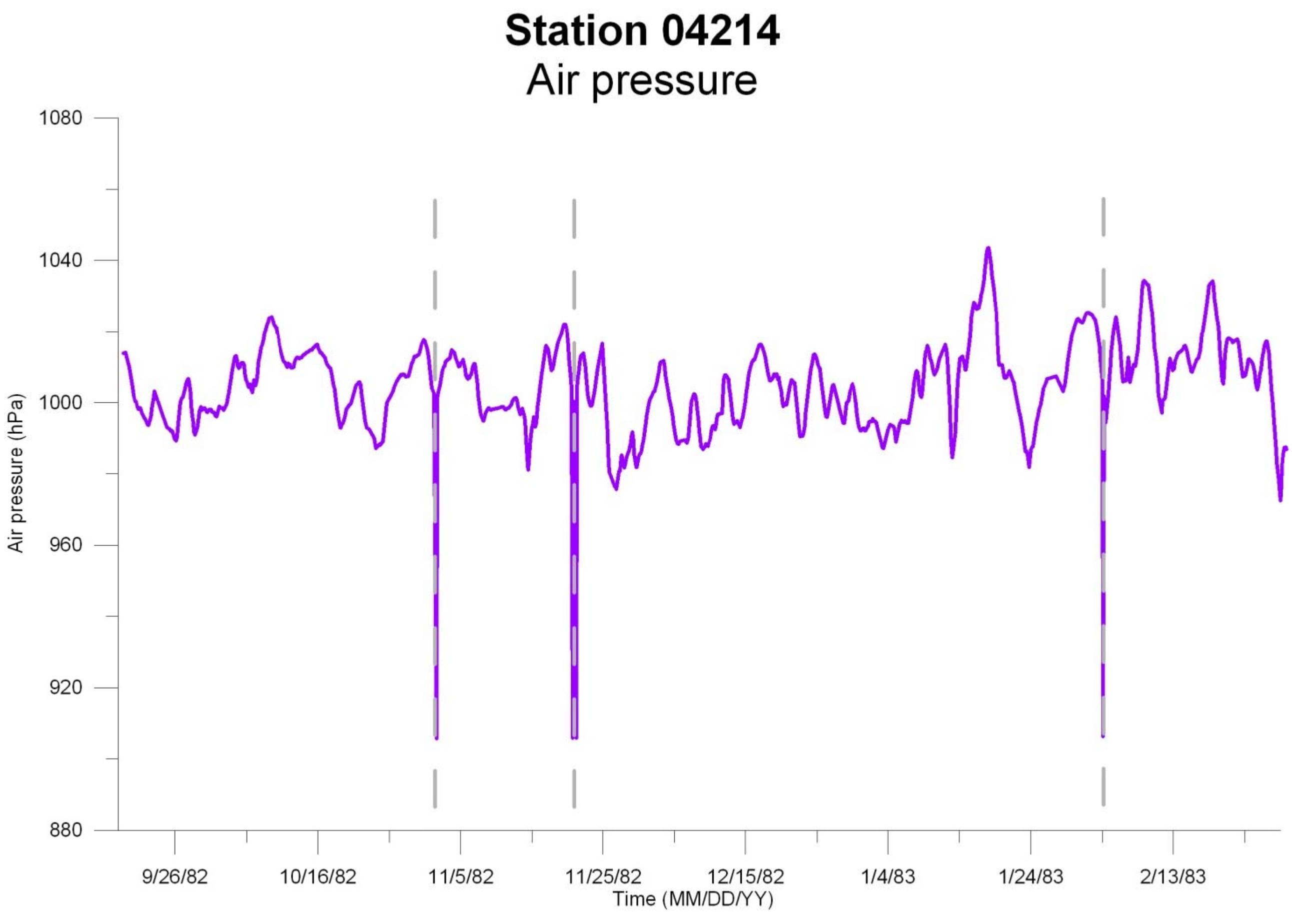


Figure 3 Example of peak values in *air pressure* data series. Three peak values marked by the dotted vertical lines are observations omitted from the data series. The middle peak actually includes 5 observations, which all are omitted, while the other two peaks consist of one observation each.

Temperature - ttt

Temperature data series have a characteristic annual period with cold temperatures during winter and warmer temperatures during summer, as well as a diurnal period caused by the dynamics of the sun, see figure 8. However, there are large variations in the absolute values between stations, due to the climatologically different localities of stations.

The typical errors found in *temperature* data series are peaks of single or few values of either implausible cold or warm temperatures that do not follow the trend. See figure 4 for an example.

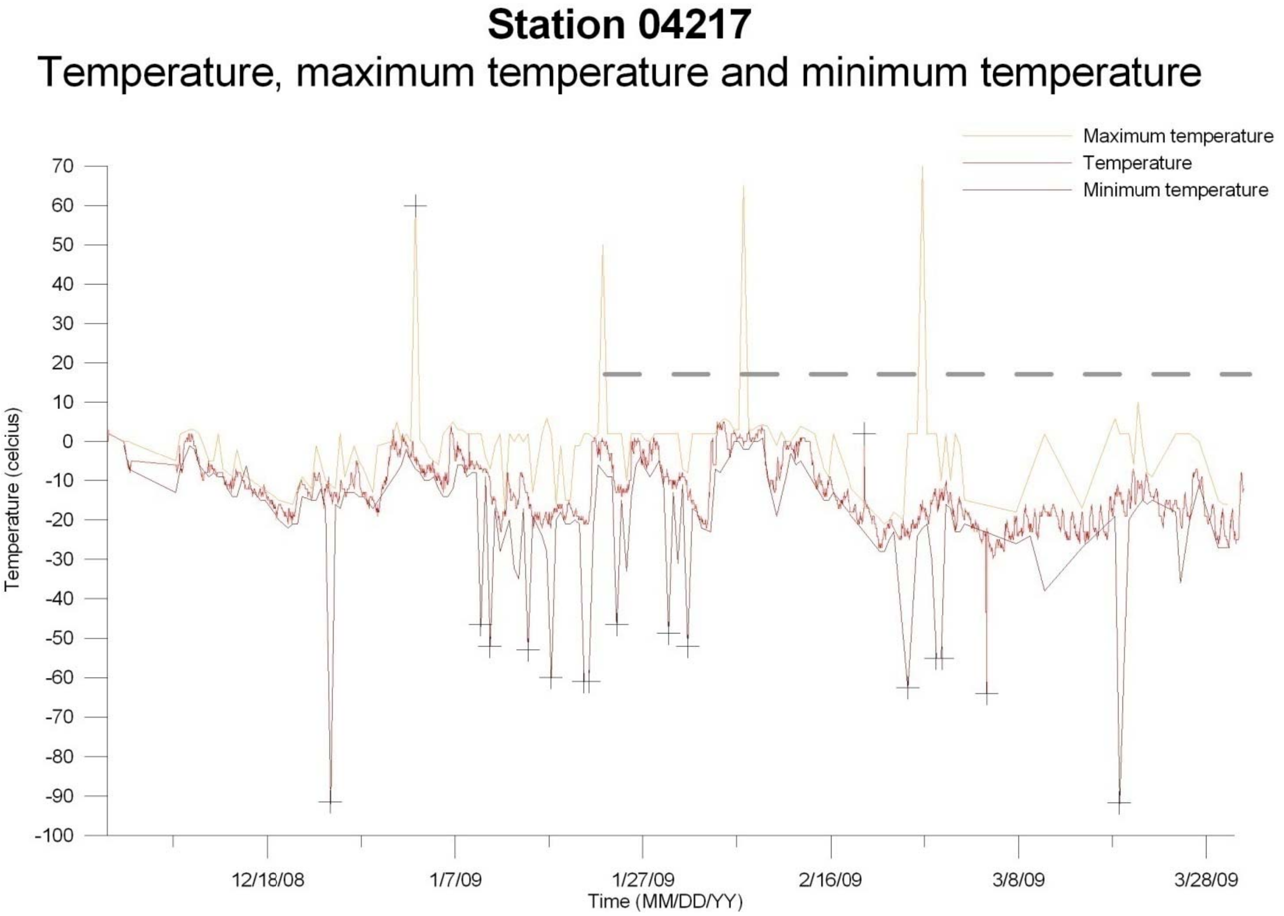


Figure 4 Example of peak values in *temperature* data series. All peak values marked with a crosshair in *temperature*, *maximum temperature* and *minimum temperature* are omitted from the final data series. The dotted horizontal line marks a period in *maximum temperature* which is omitted due to suspicious observations, e.g. an artificial limit of observations on 0 ˚C.

Absolute maximum temperature - txtxtx

The same characteristics as for *temperature* are valid for *absolute maximum temperatures*. However, another characteristic error is apparent in *maximum temperature* data series, namely peak values that clearly are erroneous when considering the general trend of the series. The peak values are not necessarily implausible if isolated as noted under temperature, but they are inconsistent with the cyclic behavior that characterizes *temperature* data series. See figure 4 for an example.

Absolute minimum temperature – tntntn

The same characteristics as for *absolute maximum temperature* are valid for *absolute minimum temperatures*. See figure 4 for an example.

Relative humidity - rh

Humidity is a challenging parameter to evaluate due to the large variations in values within short periods of time. At the same time, it is fairly easy to visually recognize an error in a data series, because most errors consist of downward peak values. The peak values are recognizable because they break the tendency in a series of data and/or because the peak value is implausible low. See figure 5 for an example.

Another frequent error is due to calibration issues. These errors are apparent as periods offset from the general tendency in the series (offset values – see table 1). Although these periods seem implausible, they are only omitted if the offset is especially pronounced (very low values).

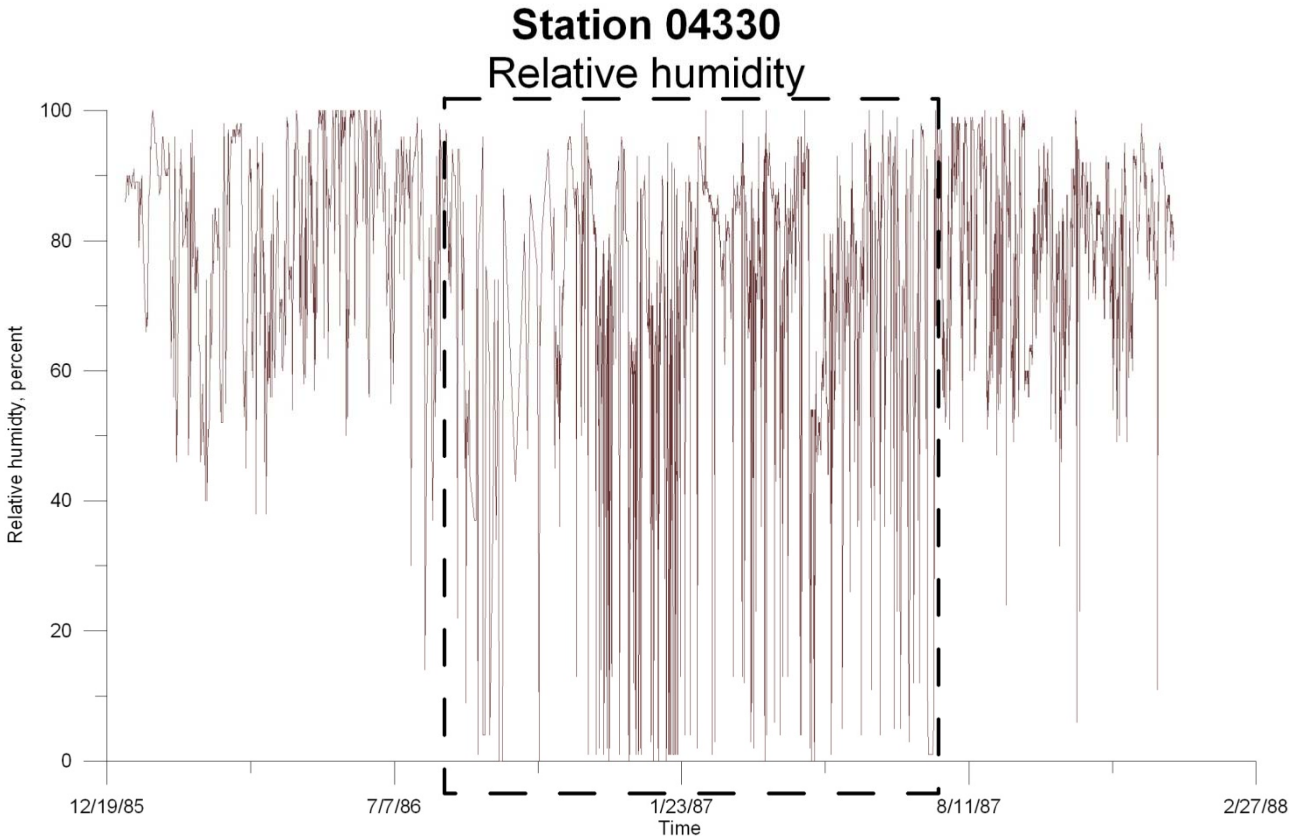


Figure 5 Example of erroneous *humidity* observations. The black square marks the omitted period. In this case it is a mix of offset observations and implausible low observations.

Accumulated precipitation – rrr6

Precipitation data series are by definition characterized by peak values. This is due to the nature of precipitation which generally either falls as showers with high values in a short period of time, or more steady with lower values over a longer period. This means that the only noticeable type of a peak error is an implausible high observation. This occurs rarely and that is why so few errors are logged in accumulated precipitation.

Furthermore, some data series only have one or very few values during the entire period, and such values are consequently omitted.

Snow depth - sss

Snow depth data series follow an annual period related to the dynamics of *temperature* data series. The snow accumulates at the beginning of winter, increases with precipitation and melts typically at the end of winter and beginning of spring, see figure 6.

Most typical errors in *snow depth* are peak values or scattered values, the latter often stretching over a long period of time. Depending on the number of observations in these long periods, the whole period has been omitted.

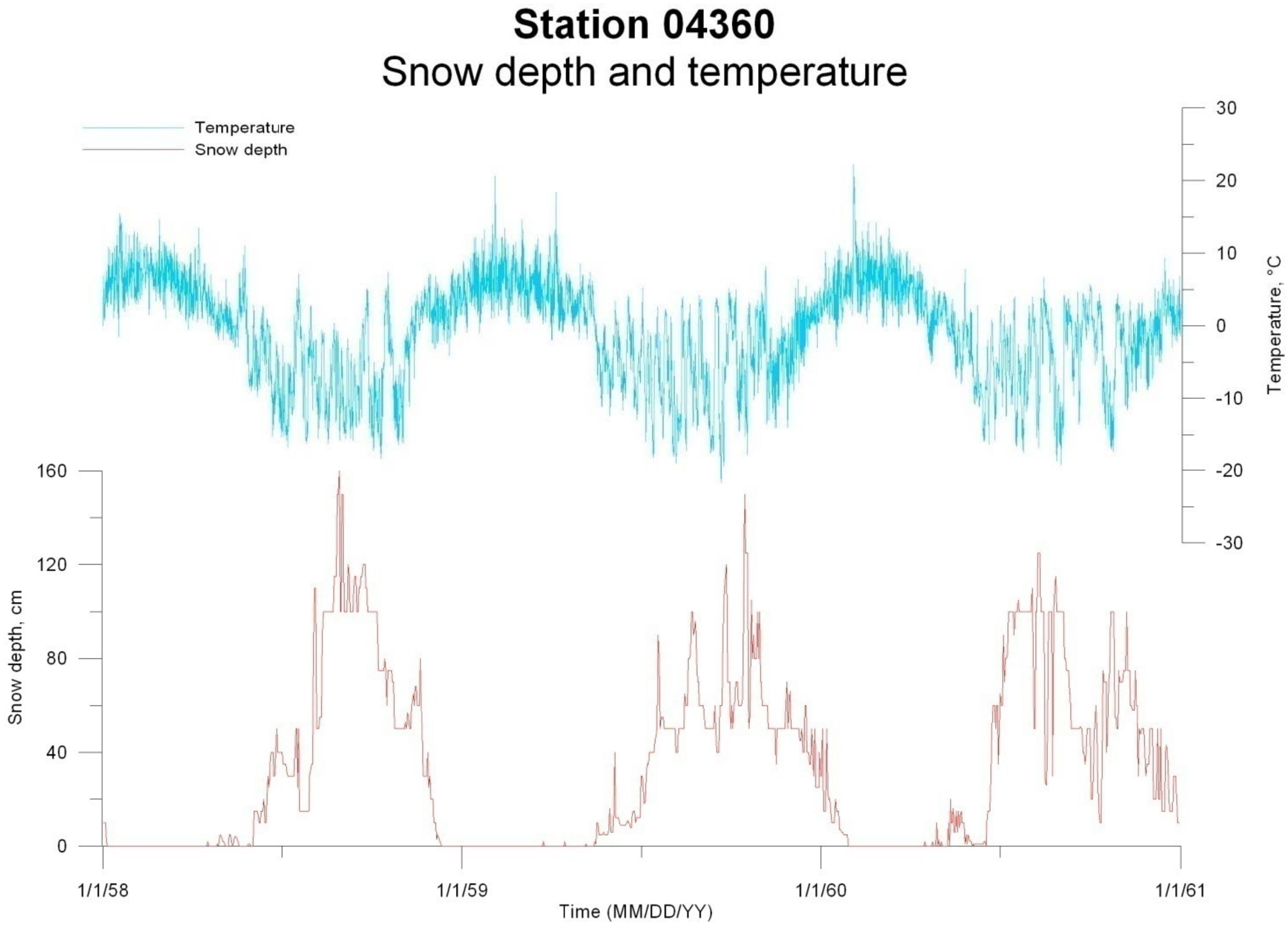


Figure 6 Example of the correlation between the annual periods of *snow depth* and *temperature*.

2.4 Statistics and trends

Table 2 shows statistics for the dataset and the errors that are omitted from it. Appendix 1 shows the distribution of errors on stations.

		Number of errors			Length of periods (days)	% of total errors (peaks)	% of total period	Data years
		Periods	Peaks	Total				
Wind direction	*dd*	389	390	779	96,485	11	16	1,787
Wind velocity	*ff*	206	684	890	80,367	12	13	1,802
Cloud cover	*n*	27	182	209	59,775	3	10	1,132
Air pressure	*pppp*	154	1379	1,533	61,932	21	10	1,724
Temperature	*ttt*	133	1238	1,371	48,374	19	8	1,808
Absolute maximum temperature	*txtxtx*	28	154	182	23,240	2	4	1,151
Absolute minimum temperature	*tntntn*	34	205	239	21,114	3	4	1,145
Relative humidity	*rh*	243	1648	1,891	109,571	26	18	1,736
Accumulated precipita-tion	*rrr6*	26	88	114	30,787	2	5	1,040
Snow depth	*sss*	55	135	190	67,044	3	11	716
	Sum	**1,295**	**6,103**	**7,398**	**598,688**	**100**	**100**	**14,041**

Table 2 Simple statistics for the complete dataset. See text below for further explanation and comments.

It is clear that the most common measured parameters, as seen by the number of data years, are *wind direction* and *speed*, *air pressure*, *temperature* and *humidity*. *Snow depth* is the least frequent parameter in the dataset, which follows from that *snow depth* are observed once a day and only at manned stations. *Humidity* and *wind direction* have the longest periods of errors. Peaks errors are most occurring in *humidity*, *air pressure* and *temperature* data series.

Omitted data years		**Data years omitted of total 14.041 years**
days	years	11.7 %
598,688	1,640	

Table 3 Length of error periods and the share of total periods. See text below for further explanation.

According to table 3, omitted data years correspond to 11.7 % of the total number of data years which is a significant. However, this is not the true share of errors. As seen in table 1 the error types "Scattered values" and "Missing values" refer to periods with limited data values meaning that the omitted periods contain only few observations. The periods with few observations still counts in the statistics as omitted data years. Therefore the number of omitted data years is higher than in reality, even though peak errors are not included in this calculation.

3. References

DMI Technical Report 00-18. John Cappelen, Bent Vraae Jørgensen, Ellen Vaarby Laursen, Lotte Sligting Stannius, Rikke Sjølin Thomsen: "The Observed Climate of Greenland, 1958-99 – with Climatological Standard Normals, 1961-90". Danish Meteorological Institute. Copenhagen 2001.

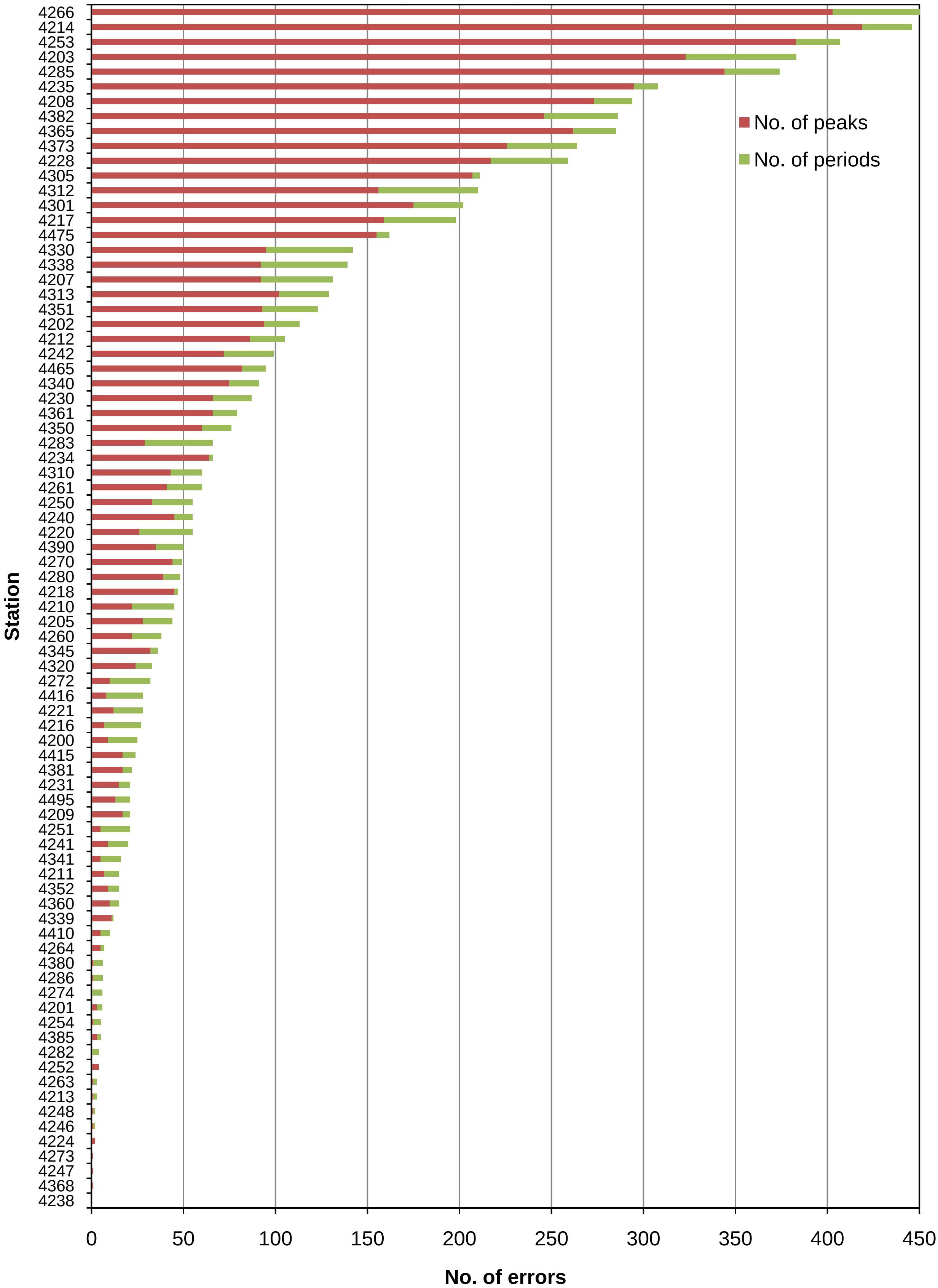
DMI Technical Report 11-10. Lone Seir Carstensen and Bent Vraae Jørgensen: "Weather and Climate Data from Greenland 1958-2010 – Dataset available for research and educational purposes". Danish Meteorological Institute. Copenhagen 2011.

DMI Technical Report 11-15. Laust Boas and Peter Riddersholm Wang: "Weather and Climate Data from Greenland 1958-2010. Observation data with description." Danish Meteorological Institute. Copenhagen 2011.

4. Previous reports

Previous reports from the Danish Meteorological Institute can be found on:
http://www.dmi.dk/dmi/dmi-publikationer.htm

Appendix 1 – Errors distributed on stations



Note that the scheme presents the *number of periods*, and not the actual *length of the periods*