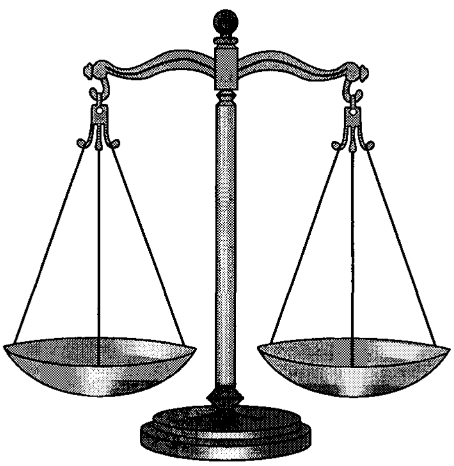
New Zealand's National Climate Database (CLIDB): audit report on the SURFACE_WIND table

John Sansom

Allan C. Penney



NIWA Technical Report 94

ISSN 1174-2631

2000

New Zealand's National Climate Database (CLIDB): audit report on the SURFACE_WIND table

John Sansom
Allan C. Penney

NIWA Technical Report 94
2000

Published by NIWA
Wellington
2000

Edited and produced by
Science Communication, NIWA,
PO Box 14-901, Wellington, New Zealand

ISSN 1174-2631
ISBN 0-478-23221-7

Citation:
Sansom, J. & Penney, A.C. 2000: New Zealand's National Climate Database (CLIDB): audit report on the SURFACE_WIND table.
NIWA Technical Report 94. 40 p.

The National Institute of Water and Atmospheric Research is New Zealand's leading provider of atmospheric, marine, and freshwater science

Visit NIWA's website at http://www.niwa.cri.nz

Contents

Abstract

Sansom, J. & Penney, A. C. 2000: New Zealand's National Climate Database (CLIDB): audit report on the SURFACE_WIND table. ***NIWA Technical Report 94.*** **40 p.**

The auditing of the dataset within New Zealand's National Climate Database which contains wind observations is described. Each row in the dataset consists of the place of observation, the date and time of observation, the observing interval, the mean and standard deviation of the wind speed and direction over that interval, and some minor attributes. All the attributes were checked individually and in groups so that any invalid values were found; consistency of the time sequence of wind observations at the same place was checked; extreme values were checked; contemporary values at neighbouring places were examined for large differences; and the temporal quality at a particular place was assessed through the number of years of observation and consistency of reporting during those years.

The grand total of changes made to **SURFACE_WIND** was 77 638, which is 0.3% of its total number of rows. About 90% of the changes came from just four sources: observation of calm conditions is stored with zeros for both speed and direction, but in 43 979 rows, where speed was zero, a non-zero direction was amended to zero; after examining the times series of observations for unlikely changes from one hour to the next, 12 233 rows were deleted; incorrect synoptic observation times were corrected by changing the time to the nearest correct synoptic observation time for 8927 rows, but in 2476 other rows the nearest synoptic hour already had an observation so these rows were deleted; the first 3053 rows from E95467/3632 up to September 1980 were deleted as it often reported calm although the manual station E95465/3630 less than 1 km away reported significant wind. The need for the changes to the most noticeable errors could have been found at any time and it is, perhaps, the other, more particular, changes which are the most valuable since the subtlety of many of the errors kept them so well hidden that only the auditing was likely to find them.

Apart from the changes to the data, some changes to programs were also made. In particular, a new procedure **WRITE_SURFACE_WIND** was written to follow all the rules which apply to the insertion and amendment of data into **SURFACE_WIND**. This new procedure was incorporated into the seven procedures that are regularly used to insert new data or amend existing data.

Introduction

This report is the fourth in a series which documents the auditing of particular data tables within New Zealand's National Climate Database (CLIDB). This is an ORACLE relational database consisting of a set of data tables; one for each type of climate data (e.g., rain, sunshine, wind) and other tables containing metadata such as station and instrument information. In this context, auditing simply means that the table concerned will be checked, usually without reference to other data tables, but its consistency with data in relevant metadata tables will be checked. Thus, these audits are expected to uncover data errors and provide some measures of quality. They have been motivated by the need to bring all the data within a table up to the current standard with which new data are entered into the table. Furthermore, any defects existing in current data entry procedures will be detected and fixed. This series of single table audits will provide the necessary experience to design better data entry procedures and raise the general level of quality so that it becomes viable to run audits on a more regular basis, perhaps annually.

A table is made up of rows and columns; the columns define what data are held in the table and the rows are separate observations. Each column can hold only one type of data such as number, date, character. However, for a column containing, for example, number data it may be that not all numbers are valid but that they should fall within a restricted range or be restricted to a set of values. Thus the values in each column can be checked to ensure that they are all within the expected range or set. Also dependencies may exist between columns such that for a given value in one column another column's values may be further restricted in its range.

Generally, in a table some of the columns hold the *key* which, rather than being the observation itself, are details about the "where", "when", and "what" of the observation where the latter includes such attributes as the frequency with which the observations were made or, for example, the particular type of radiation measurement recorded. The key defines each row such that no two rows have the same key; for example, for a particular point (first part of key) at a particular time (second part of key) there is only one value for the temperature and thus only one row is required. Thus from row to row the values in the columns constituting the key are independent, but it may well be that values in the other columns are not independent. The set of rows for a given "where" and "what" constitute a *record* for the particular station and particular type of observations. Further to the example above, for another row at a slightly earlier or later time the temperature should be not too different. This example highlights temporal dependency; the other main dependency for climate data is spatial.

Typographical conventions

Table names are printed in **BOLD UPPERCASE**, column names in PLAIN UPPERCASE, and extractions from the tabulations in a sans serif typeface. The names of other objects stored in CLIDB are also printed in **BOLD UPPERCASE**.

The DATA_AUDIT table

The auditing process is implemented by a script, which often calls subsidiary scripts, held on the CLIDB machine in a sub-directory to /clidb/adm/audit. The total process consists of a series of sub-processes, or procedures, each of which can be started by setting the environmental variable AUDIT_TYPE to the appropriate value before submitting the script as a batch job. The results of each procedure are added to a log file in /clidb/adm/audit.

For the simpler procedures, the only result is whatever is put into the log file, but for others only a sample of the result is put there while the full set of results is kept in **DATA_AUDIT**. (The "sample" referred to usually contains those results which are, or may be, the worst cases.) The structure of **DATA_AUDIT** is given below where it should be noted that the comment that a column is "NOT NULL" implies that it is a part of the key and a row is not allowed unless the whole key is present. Since it is intended to be used for all procedures within all audits, the primary key columns TABLE_NAME and ACTION will respectively carry what table is being audited and which particular audit action is being performed. Then, since all data tables within CLIDB are keyed at least by AGENT_NO and OBS_DATE, these will also be part of the key, but only some data tables are also keyed by FREQUENCY and thus it cannot be part of the key in **DATA_AUDIT**. Similarly a further column is occasionally required to complete the key in some tables (e.g., RDTN_RADIATION in **RADIATION**) and this is covered by TYPE.

Column name	Null?	Type
TABLE_NAME	NOT NULL	VARCHAR2(20)
ACTION	NOT NULL	VARCHAR2(10)
AGENT_NO	NOT NULL	NUMBER(6)
OBS_DATE	NOT NULL	DATE
FREQUENCY		VARCHAR2(2)
TYPE		VARCHAR2(1)

Thus, either the results of a specific audit procedure are put in the log file or, when it is in progress, a row is inserted into **DATA_AUDIT** for each occurrence of whatever is being sought in the table being audited. The details of these occurrences can be recovered, since it is the key that is recorded and the worst cases can then be put in the log file. All entries into **DATA_AUDIT** are made through PL/SQL scripts called from the main auditing script with each of these performing a distinct action. When such a script is started, it removes from **DATA_AUDIT** any entries it may have made in the previous run before making new entries, and then generally a view is created through which errors, or potential errors, in the table being audited can be seen.

In practice, complications often arise that require a less than straightforward use of **DATA_AUDIT**. Then a view based on **DATA_AUDIT** is created from which the required results can be queried in a straightforward way. The initial intention was that the only additional table that would be required within CLIDB to hold audit results would be the **DATA_AUDIT** table, but experience soon proved that not all the views created produced quick results when queried and in those cases the view was replaced by a table.

The SURFACE_WIND table

The **SURFACE_WIND** table contains wind data. Its column names and the types of data they hold are:

Column name	Null?	Type
AGENT_NO	NOT NULL	NUMBER(6)
OBS_DATE	NOT NULL	DATE
FREQUENCY	NOT NULL	VARCHAR2(1)
ORIG_OBS_ORIGIN		VARCHAR2(1)
SPEED		NUMBER(7,4)
DIRECTION		NUMBER(3)
PERIOD		NUMBER(10,4)
SPEED_REL		VARCHAR2(1)
DIRECTION_REL		VARCHAR2(1)
SPEED_SD		NUMBER(7,4)
DIRECTION_SD		NUMBER(7,4)

Just as ORACLE ensures a column will only hold data of the defined type, so it ensures a complete key will be present in each row. Moreover, by maintaining a unique index for the table on the key, ORACLE also ensures that more than one row with the same key will not occur.

The key contains: the place given by the AGENT_NO for which details are held in **LAND_STATION**; the UTC date-time given by OBS_DATE; and the reporting frequency (i.e., daily, hourly) given by FREQUENCY. The remaining columns constitute the significant data with SPEED being the primary data since a row without this contains no information. Apart from PERIOD, all the other columns could be null, although ORIG_OBS_ORIGIN should usually be present since it relates to the message type with which observations were transferred from their point of measurement to the procedures that loaded them into CLIDB. There are two distinct types of wind data kept in this table which are distinguished by the value of FREQUENCY: if "D" then the data are daily wind runs; if "H" then they are hourly winds. The wind runs are held as the mean speed over PERIOD, which is at least 24 h, and DIRECTION is always null, although a mean direction over one or more days could be defined. The hourly winds are the mean wind speed over either the hour or 10 minutes before the OBS_DATE and DIRECTION is often present; occasionally the standard deviations (SPEED_SD, DIRECTION_SD) are available for FREQUENCY "H" data. A full description of **SURFACE_WIND** was given by Penney (1999).

Summary of checks

A. Single column checks
- A.1. AGENT_NO: The entries in this column should all represent valid stations, i.e., they should all appear as AGENT_NOs in **LAND_STATION**. The stations should also be of the appropriate type, i.e., STTY_STATION_TYPE should be appropriate for wind observations.
- A.2. OBS_DATE: Must not be later than the current date.
- A.3. FREQUENCY: The entries in this column should all represent valid frequencies, i.e., they should all appear as CODEs in **CODE** when CODE_TYPE is "FREQ".
- A.4. ORIG_OBS_ORIGIN: The entries in this column should all represent valid origins, i.e., they should all appear as CODEs in **CODE** when CODE_TYPE is "ORIG".
- A.5. SPEED_REL: Only NULL or "*" are allowed.
- A.6. DIRECTION_REL: Only NULL or "*" are allowed.
- A.7. PERIOD: Should be present and greater than zero.
- A.8. SPEED: Should be present and non-negative.
- A.9. DIRECTION: Can be NULL, but if present must be inclusively between 0 and 360 or be 990.
- A.10. SPEED_SD: Usually NULL, but if present must be greater than zero and not too large.
- A.11. DIRECTION_SD: Usually NULL, but if present must be between 0 and 360.

B. Multiple column checks
- B.1. AGENT_NO, OBS_DATE: The earliest and latest dates should not be before the station opened or after it closed or be inconsistent with any information in **ANEM_HIS**.
- B.2. OBS_DATE, FREQUENCY: For a given FREQUENCY the earliest date should be reasonable and all observations at the correct times.
- B.3. OBS_DATE, ORIG_OBS_ORIGIN: For a given ORIG_OBS_ORIGIN the earliest date should be reasonable and all observations should be at the correct times.
- B.4. FREQUENCY, PERIOD: The FREQUENCY and the PERIOD should be consistent.
- B.5. FREQUENCY, DIRECTION: For FREQUENCY "D", DIRECTION must be NULL.
- B.6. SPEED, DIRECTION: If SPEED is zero then DIRECTION is also zero, i.e., calm conditions are represented as SPEED=0 & DIRECTION=0. If SPEED is more than zero then, if DIRECTION is not NULL, it cannot be zero since zero is reserved for calms and for northerlies DIRECTION=360.
- B.7. DIRECTION, DIRECTION_SD: If DIRECTION is NULL then DIRECTION-_SD must also be NULL.

C. Between row checks
- C.1. For a given FREQUENCY and AGENT_NO, the OBS_DATE and PERIOD should be such that the previous observation at that FREQUENCY and station was made no later than PERIOD hours before OBS_DATE.
- C.2. For a given FREQUENCY and AGENT_NO, the greatest SPEED should not be excessive.

C.3. For a given AGENT_NO, OBS_DATE and FREQUENCY, the mean SPEED over the associated PERIOD should be consistent with the mean of any SPEEDs at other FREQUENCYs over the same period.

C.4. For FREQUENCY "H" data at a given AGENT_NO, the SPEED should not be too different from its value just before or just after its OBS_DATE.

C.5. For FREQUENCY "H" data at a given AGENT_NO, the DIRECTION should not be too different from its value just before or just after its OBS_DATE.

C.6. For the same OBS_DATE, PERIOD and FREQUENCY the SPEEDs should not be too different for AGENT_NOs that are physically close to each other.

C.7. For a given FREQUENCY and AGENT_NO, there should be a continuous dataset with no gaps from the row with the earliest OBS_DATE to that with the latest.

D. Other checks

D.1. For a given AGENT_NO and FREQUENCY, the length of record should be adequate. (For FREQUENCY "H" the time of the day must also be considered).

D.2. For FREQUENCY "D", any AGENT_NO should not have an excessive number of PERIODs greater than 24 h. Also few PERIODs should be such that PERIOD hours before OBS_DATE is in one local month and OBS_DATE in another.

D.3. For a given AGENT_NO, the rows that make up a complete local month should have associated rows in **MTHLY_STATS**.

The checks above operate at three levels — finding absolute errors, identifying possible errors, and measuring quality. Thus the A checks all search for absolute errors as do B.4, B.5, B.6, B.7, C.1, and C.3 whereas the other B checks and C.2, C.4, C.5, C.6, and D.3 will highlight those rows that might be in error. Remaining checks (C.7, D.1, and D.2) may uncover some errors, but it is more likely that any gaps in a record or any short records or excessive accumulations are due simply to lack of data, and these checks will highlight the poorer records. For most of the checks to find possible errors it is not possible to set absolute rules. So, for example, in C.2 it can be said only that the "SPEED should not be excessive" and not that it should lie below a certain limit because it can vary greatly with the AGENT_NO.

Audit results

Details and results of Check A.1 — are all stations valid?

For any row in **SURFACE_WIND** it must be known to which place the data in the row apply. A list of places where observations are possible is held in **LAND_STATION** together with full information on their positions, etc. The list is indexed by the AGENT_NO which is used in **SURFACE_WIND** as a code for the station, thus, all the AGENT_NOs in **SURFACE_WIND** must appear in **LAND_STATION**. This was found to hold, and so all stations were valid.

A search was made to locate any observation that had been attributed to a station which is of such a type that it would not be expected to have reported wind observations. In the tabulation below this applied to "RAIN (STANDARD)", "REGIONAL COUNCIL", and "WATER SCIENCES" from all of which only rain observations might be expected, but it can be seen that 56 of these had some wind data. However, stations often change their type while open or may shut and some time later one of a different type may open sufficiently close by to merit the re-use of the closed station's number. This was found to be the case for 52 of the 56 stations. Of the other four, I50114/5281 was found to have only one wind record, which was deleted, and F11692/3803, F11693/3804, F11782/3817 only had wind data so were changed from being "RAIN (STANDARD)" stations to "ANEMOMETER ONLY".

RAIN (STANDARD)	51
CLIMAT (STANDARD)	384
CLIMAT/SYNOP	135
RAIN/SYNOP	40
CLIMAT (PRIVATE)	5
REGIONAL COUNCIL	4
WATER SCIENCES	1
ANEMOMETER ONLY	89
SYNOP ONLY	179
AWS (SYNOP AND METAR)	175
EDR	19
CLITEL	37
LIMITED CLIMAT	3
SPECIAL STATION	14

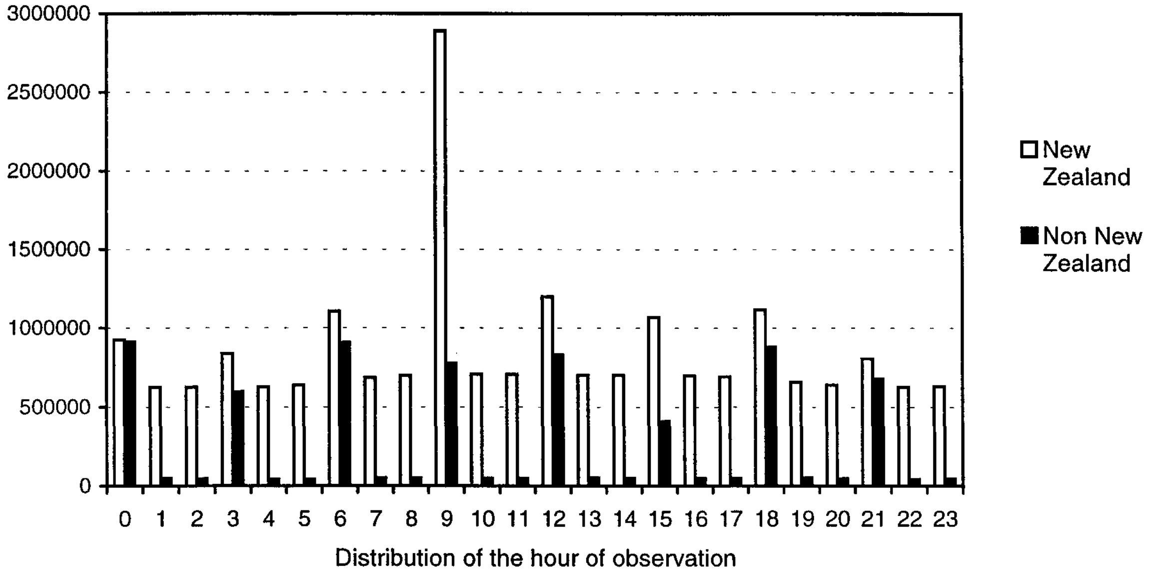
Details and results of Checks A.2, B.2, and B.3 — are all observation dates and times valid?

For any row in **SURFACE_WIND** it must be known at what date and time the observation was made and these dates should not be later than the current date. This was found to hold and so all the latest dates were valid. Unlike the latest date when the current date provides an error threshold, there is no natural threshold for the earliest date. Also, since observations of some ORIG_OBS_ORIGINs or FREQUENCY were started earlier than those of others, there is no fixed threshold either for the earliest date. However, the earliest dates for each ORIG_OBS_ORIGIN or FREQUENCY can be found and, as can be seen below, these dates are reasonable with the overall earliest being January 1891. This was for reports from H32641/4881 and no data from any other stations had a date earlier than 1928. Data for other ORIG_OBS_ORIGINs or FREQUENCYs started at dates that were appropriate to the source concerned.

ORIG_OBS_ORIGIN	Earliest data
D (i.e., Daily Reports)	18910101
H (i.e., Hourly CLITEL Reports)	19900930
E (i.e., EDR Reports)	19840701
F (i.e., Autographic Form Reports)	19391231
S (i.e., Synoptic Reports)	19481231
M (i.e., METAR Reports)	19611130
FREQUENCY	
D (i.e., Daily Windrun Observations)	18910101
H (i.e., Hourly Wind Observations)	19271231

The times of observation are also constrained since: ORIG_OBS_ORIGIN "D" and FREQUENCY "D" rows should have a time equivalent to 0900 Local; ORIG_OBS_ORIGIN "S" rows should be at one of the synoptic reporting times, i.e., 0000, 0300, 0600, 0900, 1200, 1500, 1800, 2100 UTC — except in New Zealand where during periods of daylight saving the local time of synoptic observations is not changed and so the reporting times are 2300, 0200, 0500, 0800, 1100, 1400, 1700, 2000 UTC; and, other ORIG_OBS_ORIGIN rows should all be on the hour, i.e., the minute and second part of the time is zero. All rows had times on the hour, thus the times for observations with ORIG_OBS_ORIGINs of "H", "E", "F", and "M" and FREQUENCY "H" are probably correct, but errors were found in the times of synoptic observations. Also a few ORIG_OBS_ORIGIN "D" and FREQUENCY "D" observations at stations outside New Zealand were found to be at other than 0900 Local, and so had incorrect times.

Before the erroneous synoptic times were corrected the distribution with hour was:



The larger number of observations at synoptic hours can be easily seen but for New Zealand the effect of hourly reports from automatic stations decreases the relative contribution from synoptic observations. Also, the larger number at the time of the daily climate report can be seen for New Zealand.

Many of the incorrect synoptic observation times at New Zealand stations were for just 3 days — 23–25 October 1986. They were corrected by changing the time to the nearest correct synoptic observation time where possible, but if a row already existed at that time the row with the incorrect time was deleted. Deletion was done for 324 rows and 2943 rows had the time modified. Stations outside New Zealand were dealt with in a similar manner with 2152 rows being deleted and 5984 rows modified.

Details and results of Checks A.3, A.4, A.5, and A.6 — are all frequencies, reliabilities, and origins valid?

For any row in **SURFACE_WIND** it must be known over what PERIOD the wind was observed before its mean level was estimated. However, PERIOD is expected to fall into a few groups (*see* results for B.4) and these are labelled by FREQUENCY. A list of the valid frequencies with a full description is held in **CODE** where CODE_TYPE is "FREQ" and only these should appear in **SURFACE_WIND**. This was found to hold, and so all frequencies were valid.

For any row in **SURFACE_WIND** it ought to (but not *must*) be known what is the origin of the observation. A list of the valid origin types with a full description is held in **CODE**, where CODE_TYPE is "ORIG" and only these should appear in the ORIG_OBS_ORIGIN column of **SURFACE_WIND**. This was found to hold, and so all origins were valid.

If a speed observation is deficient in some way then a "*" is stored in SPEED_REL, otherwise the column is left empty (i.e., NULL). Similarly, if a direction observation is deficient in some way then a "*" is stored in DIRECTION_REL, otherwise the column is left empty. It was found that SPEED_REL and DIRECTION_REL were either NULL or contained a "*", and so all reliabilities were valid.

The tabulation below shows what combinations of frequencies, origins, and reliabilities occurred and how frequent each combination was. The only origin that FREQUENCY "D" data had was "D", but for FREQUENCY "H" data a quality ranking applies to the different origins due to the amount of precision with which the particular message type conveys the wind measurement and to the amount of quality control applied before being stored in **SURFACE_WIND**. The ranking is that the best data have an origin of "H" which indicates it was received directly as m/s with one decimal place precision. The next best has origin "E", then "F", "M", "S", and finally "D" has the poorest quality since it is received as an estimate using the Beaufort force scale.

Frequency	Origin	Speed reliability	Direction reliability	Count
D	D	*		602
D	D			943499
H	D	*	*	78
H	D	*		26
H	D		*	104
H	D			1943940
H	E	*	*	2
H	E	*		10
H	E		*	2
H	E			648739
H	F	*	*	14
H	F	*		32
H	F		*	20
H	F			12180903
H	H	*		1
H	H		*	1
H	H			841372
H	M	*	*	9
H	M	*		123
H	M		*	131
H	M			3961586
H	S	*	*	10
H	S	*		13
H	S		*	2
H	S			7774248

All of the FREQUENCY–ORIG_OBS_ORIGIN combinations above are valid. The table provides a measure of quality since it can be seen that only a few of the rows for a given combination have a reliability of "*". The tabulation above was for all stations: the one below compares New Zealand stations to others by showing for each set of stations the percentage contributions made to the total number of rows from each of the different origins.

	Percentage of rows for different origins					
	D	E	F	H	M	S
New Zealand	8	3	54	4	19	12
Elsewhere	5	0	17	1	0	77

The zero for "E" data outside New Zealand is exactly correct but not the zero for "M" since a few such rows do exist for those occasions when ORIG_OBS_ORIGIN "F" (i.e., observations estimated from anemometer charts or Forms) data were not available. The tabulation shows that most non-New Zealand wind data derive from synoptic reports while for New Zealand most have been taken off anemometer forms. However, little new "F" type data are being entered into CLIDB, but the "H" data from automatic climate stations are a growing source.

Details and results of Checks A.7 and B.4 — are all periods valid and consistent with the associated frequency?

For any row in **SURFACE_WIND** it must be known over what PERIOD the wind was observed before its mean level was estimated, thus PERIOD should be non-NULL and greater than zero for all values of FREQUENCY. This was found to hold, but FREQUENCY also constrains the value of PERIOD, since when FREQUENCY is "D" PERIOD must be a multiple of 24 and for "H" it must be either 0.1667 h (i.e., 10 minutes) or unity. Two exceptions were found: station C75731/2101 had a PERIOD of 76 h and this was changed to 72 h; station E15102/3477 had a PERIOD of 0.667 h and this was changed to 0.1667 h.

Also, for station B77691/1686 a FREQUENCY "D" row was found with a PERIOD of 1152 h, which is much more than a month: the row was deleted.

Details and results of Checks A.8 and A.10 — are all speeds and their standard deviations valid and reasonable?

The essential data in any row of **SURFACE_WIND** is the value of SPEED, so SPEED should be non-null and non-negative. This was found to hold.

In the audit of **RAIN** (Sansom & Penney 1999), FREQUENCY also had implications with regard to AMOUNT since the largest values for "D" FREQUENCY rows should have been larger than those for "S" or "H" rows. But in this audit, FREQUENCY has no implications for SPEED, as it is a mean rather than an accumulation, so all rows were checked for SPEEDs over 99 m/s. Four instances were found: station I68182/5778 had two speed of about 105 m/s in 1986 and there was another higher than expected speed associated with each one — four rows were deleted; station J82200/6086 had two occasions when the speeds were equivalent to 888 kt — one was deleted and the other changed to zero.

SPEED_SD can be null but if present it must be non-negative and not too large. This was found to hold since the largest value was 19.2 m/s and this was considered acceptable.

Details and results of Checks A.9, A.11, B.5, and B.7 — are all directions and their standard deviations valid for the associated frequency?

FREQUENCY "D" rows hold windrun data which have no associated DIRECTION value, thus for those rows DIRECTION should always be null. This was found to hold.

However, FREQUENCY "H" rows would usually have a value for DIRECTION as well as for SPEED, but the value is restricted to lie inclusively between 0 and 360 or have the value 990, which is used to indicate that although conditions were not calm no definite direction could be ascribed to the wind. Two errors were found: E1427A/3375 on 3 August 1992 at 0300 had a direction of 630 which was changed to 330; I68533/5823 on 5 March 1995 at 0600 had a direction of 380 which was changed to 280.

If DIRECTION is null then DIRECTION_SD must also be null which was found to hold, but if DIRECTION_SD is present then it must lie inclusively between 0 and 360 and this was also found to hold.

Details and results of Check B.1 — are all records within the time that the station was open?

The dates of opening and closing for each station are held in **LAND_STATION** and the dates of the installation and the removal of anemometers from some of the stations are held in **ANEM_HIS**. Thus for each AGENT_NO the earliest and latest rows within **SURFACE_WIND** can be found and an error noted if the earliest is before the station opened or if the latest is after the station closed. Also for stations with information in **ANEM_HIS** the data dates can be compared to the date when a anemometer was installed or was removed.

In the auditing of **MTHLY_STATS** (Sansom & Penney 1999a) and **RAIN** (Sansom & Penney 1999b) many hundreds of date inconsistencies had been found, but no consistent way of treating the problem had been apparent. The simplistic treatment of accepting any data outside the station dates as valid and amending the date of the station's opening or closing to accommodate the excess data was not adopted. Such treatment could have validly solved the problem in most cases, but in the remainder would have covered up the more serious error of data having been allocated to the wrong station. Thus, since such a large number of errors had to be dealt with and with no overall solution available, no changes were made to the data tables or to **LAND_STATION**, but entries were made in **SITE_CHANGES**.

In this audit only 104 date inconsistencies were found and these were corrected by determining from **LAND_DATA_CAT** for each station what the general start and end dates were for all types of observations and modifying the start and end dates in **LAND_STATION**, **SITE_CHANGES**, and **ANEM_HIS**. Also some rows that had been put into **SITE_CHANGES** during previous audits were no longer valid and were deleted.

Using the Beaufort scale, in which wind speeds are estimated from the visual effects of the wind, does not require an anemometer and, although it is current practice to note the use of Beaufort in **ANEM_HIS**, early data often preceded, by a large margin, the installation of an anemometer and any record in **ANEM_HIS**. Thus, such inconsistencies had to be accepted and some where data continued after the removal of an anemometer also had to be accepted. There were also occurrences of the reverse situation where an anemometer was installed some time before the station was opened and then data started or was removed sometime after the data stopped and then the station closed. Such inconsistencies might just be mistakes which could be rectified by adjusting the anemometer dates, but those dates might be correct and the charts or rolls from the anemometer may well be in the paper store but not have had any data taken from them. In such cases, rows similar to those in the examples below were inserted into **SITE_CHANGES** and imply that the inconsistencies could be resolved through the existence of charts or rolls: if there are any in the disputed period the station's open or close dates should be amended, but if none then the anemometer dates can be changed. No time was available to search for the charts or rolls.

AGENT_NO	Anem. date	Description
1428	19390601:0000	CODE 132 Anem installed BEFORE stn-opened/wind-data-began - any charts/rolls?
1428	19631231:0000	CODE 132 Anem removed AFTER stn-closed/wind-data-ended - any charts/rolls?

The tabulation below summarises the changes that were made.

LAND_STATION start date altered	44
LAND_STATION end date altered	30
SITE_CHANGES dates altered	74
SITE_CHANGES warnings deleted	42
SITE_CHANGES warnings inserted	23
ANEM_HIS date altered	38

Details and results of Check B.6 — are all speeds and directions consistent?

Reports concerning the individual SPEED and DIRECTION checks have been given above, now the check on the small dependency between them will be described. Conventionally, in a report of a wind observation the direction for calm conditions is given a zero as well as the speed, and so zero is not available to indicate a northerly direction for which 360 must be used. Thus if SPEED is zero then DIRECTION must be zero and if SPEED is greater than zero then DIRECTION must not be zero. Of the 44 408 errors found, most were examples of a non-zero direction with a zero speed, but some non-zero speeds with zero directions were also found. The number of errors that occurred with each ORIG_OBS_ORIGIN were:

D 100 These were all cases with SPEED=0, DIRECTION>0 and it was assumed that these should all represent calms.

H 2496 Again these were all cases with SPEED=0, DIRECTION>0. The tabulation below shows a few ORIG_OBS_ORIGIN "H" observations of DIR and SPD with the actual speed as measured to two decimal places also shown.

Station	Year	Mon	Day	Hour	DIR	SPD	Raw speed (m/s)
B76838	2000	2	24	2000	65	0	0.09
B76838	2000	2	24	2100	0	0	0.00
B76838	2000	2	24	2200	299	0	0.01
B76838	2000	2	25	0500	297	0	0.02
B76838	2000	2	25	0600	293	0	0.05

It can seen that, although the speeds were all zero, the direction was only zero when the raw speed was exactly zero, i.e., the program for formatting the raw data did not follow the rules and needed amending.

M 361 223 of these were for E1427A/3375 which had SPEED>0 but DIRECTION=0. Temporally neighbouring observations had directions of 360 or 010, and the rounding of directions such as 003 to the nearest 10 was producing 000 rather than 360. Thus it was assumed that these all represent northerlies with a direction of 360. The data logging program at the station was amended to prevent the error occurring again.
All other cases had SPEED=0, DIRECTION>0 and were assumed to be calms.

S 41 451 The worst stations were automatic weather stations (AWS) which seem to report a wind direction at all times even when the wind speed was 0 and it was again assumed that these should be calms.
A few Pacific AWS stations reported SPEED>0 but DIRECTION=0, they should probably be 360.

These errors were corrected by simply changing DIRECTION to zero if SPEED was zero — 43 979 rows amended — or changing zero DIRECTIONs to 360 if SPEED was greater than zero — 429 rows amended. For both types of change DIRECTION_REL was set to * in the rows that were amended. To ensure that such errors do not occur again and to deal with some other problems, a general procedure to be used by all archiving procedures was introduced.

The new procedure was called **WRITE_SURFACE_WIND**, and was required for incorporation into the following archiving procedures; RMSDYCLI, RMSEDR, RMSHOURLY, RMSMETAR, RMSSYNOP, COPIDYCL, and DEUPDATE. It has the following specification:

- exit if no incoming speed data
- exit if PERIOD or FREQUENCY invalid
- get any existing data from CLIDB for the incoming place and time
- exit if wind data existing in CLIDB are better than the incoming data
- For FREQUENCY "H"
    - test SPEED and SPEED_SD against limits in **RANGES**
    - test that DIRECTION and DIRECTION_SD inclusively between 0 and 360 or DIRECTION is 990
    - exit if SPEED invalid unless DIRECTION is valid and better than the one already in CLIDB
    - if SPEED = 0 ensure DIRECTION = 0 or if SPEED > 0 and DIRECTION = 0 then set DIRECTION to 360
- For FREQUENCY "D"
    - exit if SPEED > 9999
    - if SPEED = 9999 an accumulation is indicated so SPEED is set to null and PERIOD to 24 plus the value of PERIOD from the row for the immediately prior observation; that row is then deleted
- insert or amend a row in CLIDB.

Details and results of Check C.1 — are there any overlapping observations?

For a given FREQUENCY and AGENT_NO the OBS_DATE and PERIOD define the interval over which the SPEED was estimated. These intervals must not overlap — the observations are independent and complete. For FREQUENCY "H" this was always the case since all OBS_DATEs were on the hour, but for FREQUENCY "D" 158 errors were found. Of these, 143 were from just four stations with most of the errors concentrated in just eight different months between May 1986 and May 1987 and also in March 1991. The cause of this clustering of error cases could not be determined, apart from noting that the stations were all EDRs, but the errors were mostly that PERIOD was twice what it should be. Thus, PERIOD was amended for those stations and a further 14 amendments for other stations were also made and 1 OBS_DATE was amended. The changes made are summarised in the tabulation below.

AGENT_NO	No. of PERIODs amended	Dates and comments
2006	19	Dec 1986
2006	15	Mar 1987
2006	10	May 1987
2112	13	Mar 1991
3147	19	May and Jun 1986
3147	18	Feb 1987
3147	13	Apr 1987
5535	19	May 1986
5535	17	Jul 1986
	14	Miscellaneous stations and dates
3460	0	No PERIODs amended but 1 date was altered

Details and results of Check C.2 — are all the highest speeds reasonable?

Only gross checks on SPEEDs were performed in A.8 and A.10 above, but in this check the largest value for a given AGENT_NO–FREQUENCY pair was compared to the mean value for that pair. The comparison was through the ratio: maximum value of SPEED over all rows for the given AGENT_NO–FREQUENCY pair divided by their mean. There is no natural division between those ratios that indicate that the maximum SPEED for the AGENT_NO–FREQUENCY pair can be accepted and those that indicate an error. If errors exist then they are more likely to be associated with the largest values of the ratio and those cases where the ratio exceeded an arbitrary threshold were examined with the intention that the threshold could be lowered and further cases sought if most of the initial ones represented errors. The threshold ratio for FREQUENCY "D" data was 6 and for "H" was 29.

For FREQUENCY "D" rows 84 potential errors were highlighted, but using as a guide hourly wind observations — often only a single observation at 0900 Local was available — 69 of these were accepted. Of these, 39 were from V88212/6604 a station run by the Fiji Meteorological Service who supplied additional information to support their correctness. Station E15011/3460 during its final period of operation from December 1998 to July 1999 had been unable to maintain a full observational programme due to the failing health of the observer and so 10 wind rows were deleted where it was impossible to assert what the PERIOD had been.

The remaining five (i.e., 84 – 69 – 10) cases highlighted a known "feature" of the way the highest values of wind run had been treated before the introduction of CLIDB. At that time an upper limit of 1999 km existed for the windrun, but on rare occasion a larger value occurred for an accumulation, i.e., the cup counter anemometer was not read for a few days during a windy period. In such situations the overall windrun was broken up and shared about equally between the rows for the day of the observation and the previous day. The tabulation below shows the five rows found by this check, indicated by a ∗, and the rows just before and after are also shown. It can be seen that the row before (or after for 3460) has a windrun of equal size and the two together would exceed 1999 km; also this other row has a large PERIOD and so a lower value of SPEED than for the picked rows. The instances in 1996 and 1999 post-date the introduction of CLIDB but arise through the use of the MNSURWND program which enables changes to be easily made to **SURFACE_WIND**, but still implemented the pre-CLIDB rule that the windrun should not exceed 1999 km. MNSURWND was modified to allow windruns of up to 4999 km to be entered.

AGENT_NO	OBS_DATE	Windrun (km)	PERIOD (h)	
1883	19841030:2000	1312	408	
1883	19841031:2000	1310	24	∗
1883	19841101:2000	98	24	
1883	19990520:2100	1220	384	
1883	19990521:2100	1218	24	∗
1883	19990522:2100	53	24	
2128	19881126:2000	1645	360	
2128	19881127:2000	1644	24	∗
2128	19881128:2000	364	24	
3460	19961212:2000	1769	24	∗
3460	19961213:2000	1771	264	
3460	19961214:2000	307	24	
4647	19850516:2100	1756	528	
4647	19850517:2100	1755	24	∗
4647	19850522:2100	430	120	

Because cases other than the five discovered above may well exist, **SURFACE_WIND** was scanned for occasions where a FREQUENCY "D" row had a windrun of over 998 km and a PERIOD of over 24 h and was adjacent to a row with a windrun within 20 km of the other's. A total of 630 such rows were found of which only the one shown above for station 3460 had a PERIOD of 24 h for the earlier row, otherwise the later row had the 24 h PERIOD. The windruns and PERIODs of these pairs were added together and accredited to the row with the later OBS_DATE and the earlier row was deleted, i.e., 630 rows were amended and 630 rows were deleted.

For FREQUENCY "H" rows 138 potential errors were highlighted, 122 were from just 4 stations in the Solomon Islands, i.e., J50300/5937, J50700/5938, J52000/5941, and J52700/5942 and the rest were from New Zealand stations of which 9 were amended and 7 were deleted. The method used for the New Zealand errors was time consuming, and a scheme for quickly identifying winds that need deletion was developed. It produces a time-ordered listing of the whole record for a station in which the SPEEDs are displayed as a string of characters whose length depends on the speed value. Thus a quick time series plot can be created which has sufficient detail to determine when a significant change in the wind takes place. A typical example is given below where at 0500 on 19850520 a large but temporary increase occurs of nearly 40 kt to 65 kt — each * represents 2 kt and the numerals indicate tens of kt.

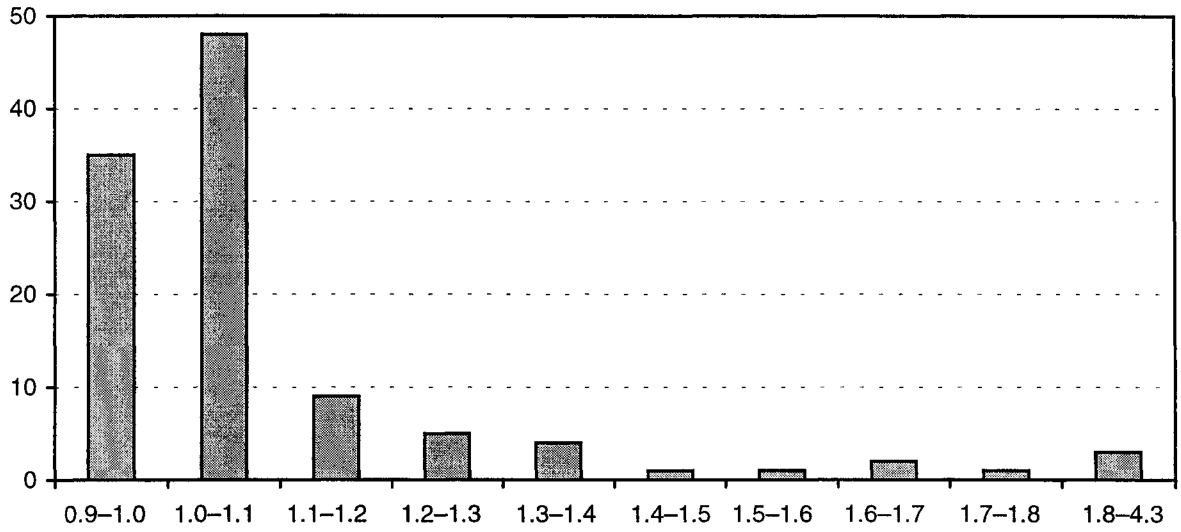
	DIRECTION	SPEED	
19850519:2300	270	08	0***
19850520:0000	270	21	0****1****2
19850520:0100	280	25	0****1****2**
19850520:0200	280	14	0****1**
19850520:0300	260	20	0****1****2
19850520:0400	270	25	0****1****2**
19850520:0500	190	65	0****1****2****3****4****5****6**
19850520:0600	250	27	0****1****2***
19850520:0700	280	46	0****1****2****3****4***
19850520:0800	290	42	0****1****2****3****4*
19850520:0900	270	43	0****1****2****3****4*
19850520:1000	280	35	0****1****2****3**
19850520:1100	260	33	0****1****2****3*
19850520:1200	250	33	0****1****2****3*
19850520:1300	290	25	0****1****2**
19850500:1400	300	27	0****1****2***
19850520:1500	270	27	0****1****2***
19850520:1600	280	23	0****1****2**
19850520:1700	290	30	0****1****2****3
19850520:1800	290	42	0****1****2****3****4*
19850520:1900	270	43	0****1****2****3****4*
19850520:2000	280	35	0****1****2****3**
19850520:2100	260	33	0****1****2****3*
19850520:2200	250	33	0****1****2****3*

In practice, a computer file in the form of the above was scanned and any rows judged as needing deletion were marked by the line in the file being edited. Subsequently the edited lines were extracted to give a list of the dates for which the rows in **SURFACE_WIND** should be deleted. A time series file was generated for each of the 4 Solomon Island stations where many SPEEDs over 65 kt had been found and the results are tabulated below.

Station	No. found by check	No. of deletions made by listings
J50300/5937	15	111
J50700/5938	54	100
J52000/5941	43	164
J52700/5942	10	64

Details and results of Check C.3 — are all windruns consistent with the component hourly speeds?

For a given AGENT_NO and OBS_DATE the SPEED for a FREQUENCY "D" row should be consistent with the mean of the SPEEDs from all FREQUENCY "H" rows with dates between OBS_DATE minus PERIOD and OBS_DATE. For a valid consistency check, the "H" rows should cover all the hours in PERIOD, but instances where a single hourly observation was missing were included in the estimation, for each station, of the mean over all "D" rows of the ratio of the mean speed from the "H" rows divided by the "D" row SPEED. The distribution of this ratio is shown in the figure below.



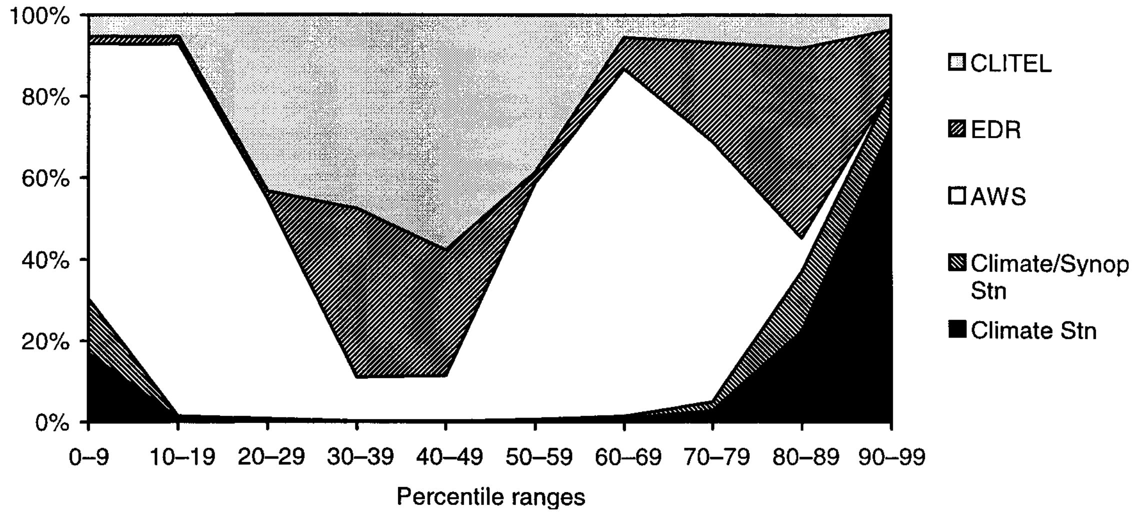
Distribution of ratio of mean wind from hourly obs to that from windrun

There were 109 stations for which contemporary "D" and "H" rows were available. The number of "D" rows at a station with matching "H" rows ranged from 48 to 10 127 with a mean of 1785 so the ratios were generally based of enough cases to yield an accurate estimate of the ratio. It can be seen from the figure above that about 80 of the stations had ratios between 0.9 and 1.1 — indeed 64 had ratios between 0.997 and 1.003 — and only 3 had ratios over 1.8. All those with a ratio over 1.25 are tabulated below.

Station	Ratio	Anemometer/station type	Mean speed (kt)	
			Anem.	Windrun
G13301 / 4260	1.257	Lambrecht		
E05622 / 3277	1.265	Dines (corrected to 6 m)	3.7	4.2
E04993 / 3147	1.280	EDR station		
E15102 / 3477	1.316	Lambrecht (corrected to 6 m)	5.6	4.6
I68174 / 5775	1.331	Lambrecht		
G12191 / 4162	1.376	Lambrecht		
J80000 / 6078	1.392	Lambrecht		
B76836 / 1646	1.454	EDR station		
C64981 / 1948	1.582	Unknown		
E14272 / 3385	1.667	Munro (corrected to 6 m)	8.5	7.8
I50102 / 5277	1.693	Lambrecht		
G22581 / 4458	1.772	Lambrecht (corrected to 6 m)	4.2	2.6
I49932 / 5243	2.405	Lambrecht		
I59235 / 5577	2.413	Lambrecht		
H31172 / 4651	4.217	Lambrecht (corrected to 6 m)	8.0	2.6

The "Mean speed (kt)" columns are taken from Reid (S.Reid, NIWA, Wellington, unpublished results) in which a closer comparison between anemometers (hourly wind instruments) and cup-counters (windrun instruments) was made; it can be seen that apart from E05622/3277 the ratio of the mean speeds give a value similar to that in "Ratio". Reid also suggests that discrepancies are often due to the anemometer and cup-counter being exposed at different heights or at some distance apart. Furthermore, different instruments have different responses at low wind speeds, and, in particular, the Lambrecht is more sensitive than the cup-counter and will consistently have a higher mean; Lambrecht anemometers were at 10 of the 15 stations in the tabulation above. Overall, the high ratios can be accounted for through exposure and instrumentation considerations.

For a particular station and day, the value of the ratio can be considered through its percentile value in the distribution of all values of the ratio. Then, disregarding those from stations with Lambrecht anemometers, the ratios can be classified by percentile and station type as in the figure below.

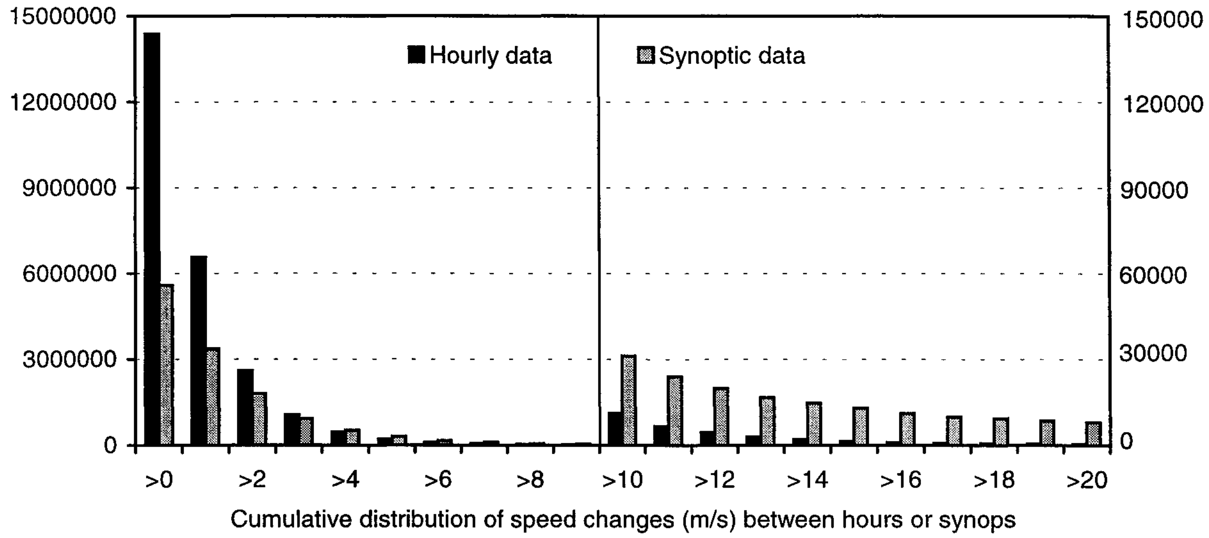


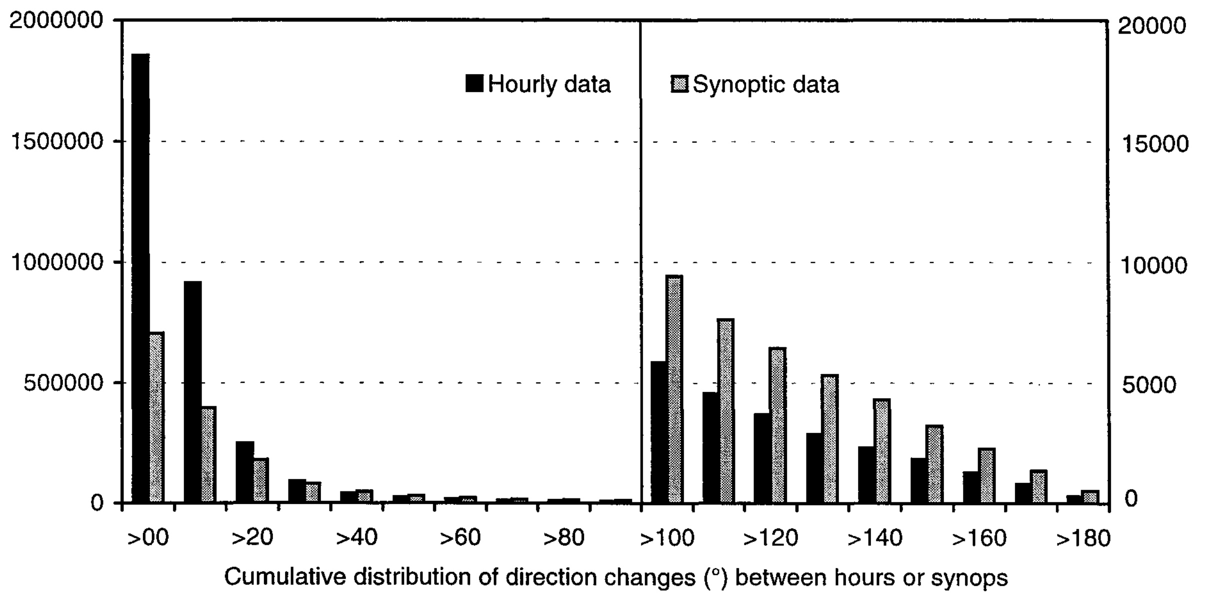
From the figure above it can be seen that most of the lowest 10% of the ratios came from automatic weather stations (AWS) while about half of those whose ratios are in the 20–60 percentile band came from CLITEL stations, i.e., NIWA's automatic climate stations. EDR (i.e., older automatic climate stations) and manual Climate stations are the major contributors for the top 30% of the ratios. However, the value of the ratio at the 90 percentile level is only 1.35 and only occasions with ratios at and above that level might be considered doubtful. Furthermore, the contributions at that level from the Climate or Climate/Synop stations came from only five stations where the instruments had different exposures whereas the EDRs have a single instrument so should always have a ratio of unity. Thus, wind observations from stations which are not of the EDR type were accepted while those from EDRs were made consistent by replacing the SPEED for FREQUENCY "D" rows by the mean of the appropriate FREQUENCY "H" rows. The number of changes made are tabulated below.

Station	Number of amendments
A53487 / 1134	488
A6487A / 1425	6
B76603 / 1609	112
B76836 / 1646	324
D9668B / 3015	308
E04993 / 3147	172
E9452A / 3551	407
I49592 / 5212	512
I50836 / 5365	521
I68182 / 5778	357
Total	3207

Details and results of Checks C.4 and C.5 — are there any excessive changes in the either the direction or the speed time series?

The wind observations at each particular AGENT_NO form time series for both the direction and speed and sample the actual continuous variation of wind at each point. There are some circumstances when wind changes rapidly in time, for example, the passage of a cold front, or the onset of a sea-breeze or of a föhn wind. However, provided the interval between observations is small enough, changes are generally small for both speed and direction although large swings in direction often take place during light winds. The figures below show that out of a total of over 20 million rows in **SURFACE_WIND** only about 1 million speeds are over 4 m/s different from the observation either 1 h before or after or 3 h before or after and, when speeds are at least 7.5 m/s, about 0.5 million directions are over 20° different. There is a change in the vertical scale from the >10 m/s and >100° classes so that the trends in the upper tail can still be seen despite the smaller numbers, which in the last class are 462 and 7937 for hourly and synoptic speeds respectively and 292 and 520 for the directions.





The differences after which error cases predominate is not known, but it can be assumed that the larger the difference the more likely an error, so the rows with the largest differences were selected for investigation. A brief examination of these for direction strongly suggested that examining the time series for direction in itself would not clearly indicate whether a direction was in error and so only speeds were examined further. The levels at which the class memberships fell below 10 000 were found, then the AGENT_NOs and OBS_DATEs for all such occurrences. These differences were divided into "blips" and "steps" where the former implies that a change was followed immediately by a compensatory change while the latter implies that a more permanent change took place. From the total of 15 802 blips and steps, taking the 5% of each with the largest differences gave 146 different stations.

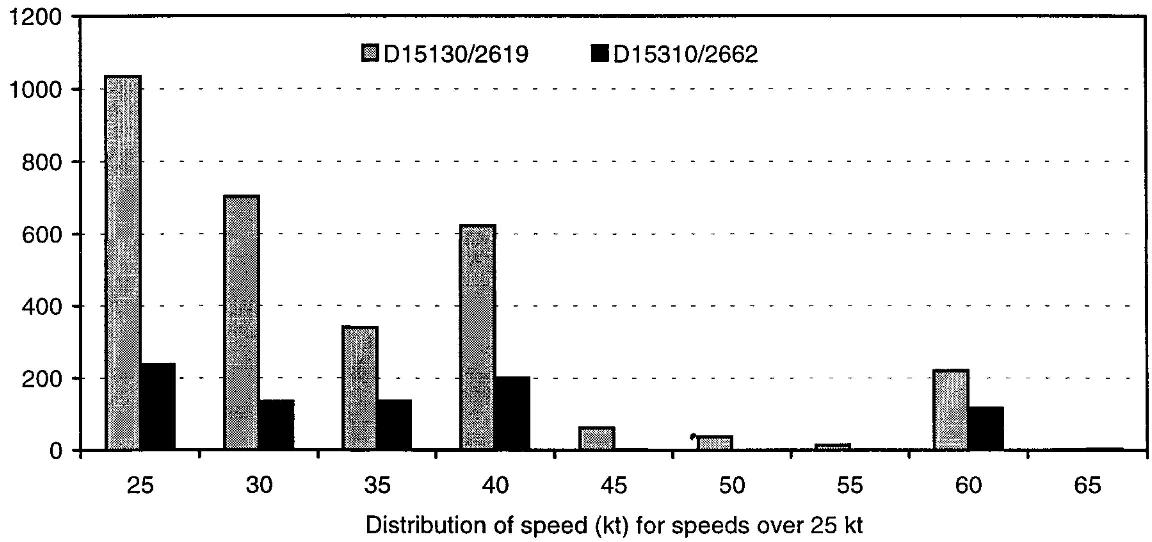
For each of these stations a listing was created in the style used for the C.2 Check (i.e., the speeds from the wind rows were displayed in time-order with their values indicated as a number of ✳s). The listing had any large differences marked and were used in much the same way as before with the marked differences in the computer file either left or unmarked if thought acceptable, but other rows were sometimes marked during the inspection process. Some of the marked rows in the listing also had new values for the speed added, in which case the relevant row in **SURFACE_WIND** was amended rather than deleted.

In a second run the largest steps and blips indicated that another 104 stations needed examining and in a third run another 31 were examined. All changes to **SURFACE_WIND** were noted in **AUD_ SURFACE_WIND**, and the numbers involved are tabulated below.

Run	No. of stations	No. of deletions	No. of amendments
1	146	8075	64
2	104	1142	36
3	31	3016	3

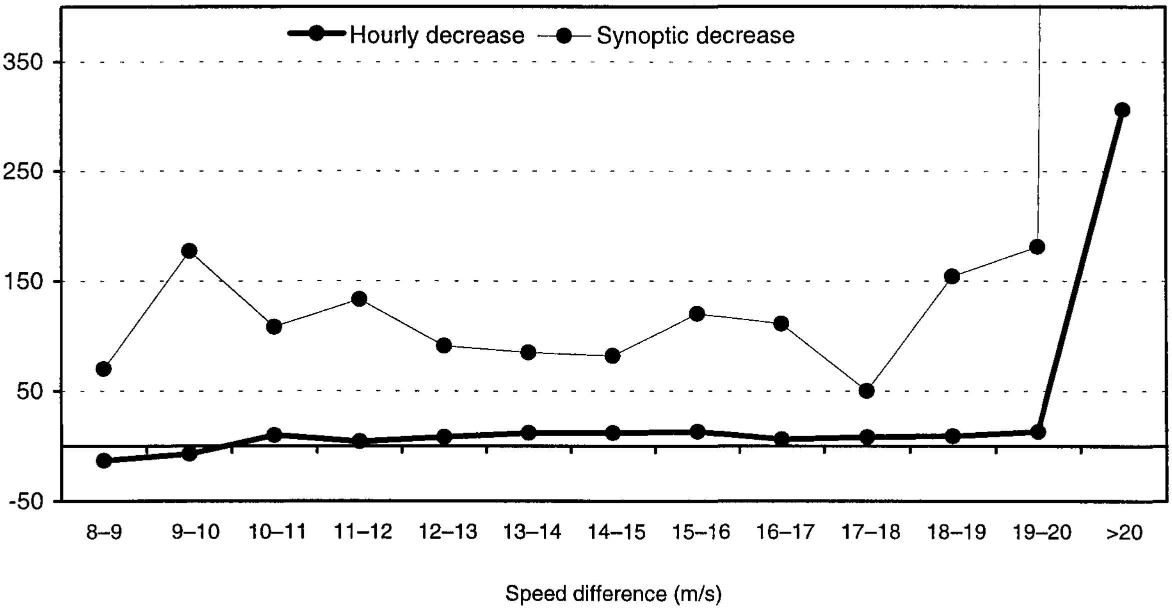
Some special cases that were dealt with are given below.

- For D15130/2619 and D15310/2662 the distribution of speeds are given in the figure below where it can be seen that the frequencies for both 40 and 60 kt are higher than might be expected. These result from the difficulty associated with the extraction of high wind speeds from the charts that are produced by Lambrecht anemometers. The charts show windrun accumulation as a rising trace which is brought back to a baseline after every 10 km and the mean wind speed for an hour is found by estimating the windrun during the hour. At high speeds the trace is returned to the baseline frequently and the bias to over-estimating the frequencies of 40 and 60 kt results.



- Twelve Antarctic stations contributed 2667 deletions (i.e., 22% of the total from 4% of the stations) and all but 11 of the 2667 were rows which had a speed of 61, 62, 63, 64, or 65 kt. It is probable that some systematic, but unknown, coding error was made when these observations were encoded as synoptic reports.
- The Pacific island AWS station J76700/11111 contributed 120 deletions (i.e. 1% from 1 station out of 281) which had all been either 57 or 69 kt from 170°.

The distributions of speed and direction changes from hour-to-hour and synop-to-synop was redetermined after the third run and the figure below shows the change in numbers for speed of the various classes. For hourly data in classes for differences over 10 m/s there were only small decreases except for the last class while for synoptic data in all cases the class-membership decreased by about 100 but by 6729 in the last class. For direction there was little change for hourly data but a decrease of about 100 in each class for synoptic data.



Details and results of Check C.6 — are all winds, when compared to nearby stations, reasonable?

As a preliminary step it was necessary to find, for each station, enough stations, or buddies, to adequately cover the period over which the primary station had reported wind and which were the closest to the primary station. To be considered as a buddy, a station had to be within 1° of latitude and longitude for New Zealand (5° elsewhere) of the primary station and had to be contemporary with at least 30% or 5 years of its record. The nearest such candidate buddy was taken to be the first one and further buddies were selected in order of distance from the primary, provided at least a further year was added to the coverage and until at least 90% coverage was reached, but no more than five buddies were noted for any station-code combination.

How well does this buddy system work? The tabulation below shows the counts of primary stations in different distance-cover classes. For example, for UTC hour 00 there were 58 primary stations in New Zealand each with its furthest away buddy nearer than 5 km and whose buddies covered at least 95% of the primary station's wind record. At the other extreme for that hour there were 22 stations outside New Zealand for which the coverages were under 95% and the furthest buddies were over 95 km away. However, the tabulation does not include those primaries for which no buddies could be found; there were 98 such stations for UTC hour 00.

Hr	N.Z.?	Cover(%)	Distance (km) of the most distant buddy										
			<5	5–15	15–25	25–35	35–45	45–55	55–65	65–75	75–85	85–95	>95
00	Y	≥95	58	70	59	45	36	31	22	23	6	5	5
00	Y	<95	9	3	2	8	2	.	2	3	3	1	6
00	N	≥95	13	7	10	2	8	3	1	1	.	1	94
00	N	<95	3	1	1	1	.	.	.	.	2	2	22
03	Y	≥95	57	63	58	40	43	28	21	19	6	5	5
03	Y	<95	6	2	4	4	1	.	3	3	3	.	8
03	N	≥95	10	2	7	1	5	2	.	2	.	2	68
03	N	<95	4	2	.	1	.	.	.	.	2	1	22
06	Y	≥95	52	68	48	33	35	17	19	25	8	7	8
06	Y	<95	9	4	3	6	2	.	1	1	2	1	6
06	N	≥95	13	8	9	2	7	2	1	2	.	1	93
06	N	<95	4	2	.	1	.	.	.	.	2	2	23
09	Y	≥95	38	57	38	18	33	21	20	18	6	10	11
09	Y	<95	6	3	6	2	1	1	2	2	5	1	9
09	N	≥95	8	2	6	1	3	2	1	3	.	2	42
09	N	<95	4	1	1	1	.	.	.	.	2	1	24
12	Y	≥95	41	56	46	18	29	25	21	24	6	10	9
12	Y	<95	10	3	6	5	.	.	.	3	3	.	5
12	N	≥95	12	8	9	2	5	2	.	.	.	2	91
12	N	<95	5	1	.	1	.	.	.	.	2	1	21
15	Y	≥95	38	56	41	15	35	25	23	20	4	8	9
15	Y	<95	4	3	6	5	1	1	3	1	1	2	7
15	N	≥95	6	2	4	2	3	1	3	1	.	2	41
15	N	<95	5	1	1	1	.	.	.	.	2	1	25
18	Y	≥95	55	68	45	35	42	23	25	26	8	10	9
18	Y	<95	6	3	4	6	3	.	.	3	2	1	5
18	N	≥95	12	8	8	2	6	2	1	1	.	2	92
18	N	<95	3	1	.	1	.	.	.	.	2	1	22
21	Y	≥95	125	218	136	90	58	22	11	10	9	2	3
21	Y	<95	9	14	12	11	4	.	.	3	2	2	6
21	N	≥95	21	13	19	9	7	3	3	2	2	2	83
21	N	<95	5	3	1	2	.	.	.	.	2	2	20
D	Y	≥95	14	37	39	31	26	20	13	15	4	5	9
D	Y	<95	.	2	4	1	3	3	4	2	.	.	2
D	N	≥95	1	4	1	1	1	1	1	2	.	.	2
D	N	<95	1	1	.	.	.	.	1	1	.	.	0

The tabulation above shows buddy counts not only for synoptic observation hours but also for the windrun observations which are indicated by a D in the Hr column. For these over half the stations within New Zealand have a buddy within 35 km providing better than 95% coverage, few had buddies over 65 km away, and only two were in the worst distance-cover class. Few windrun stations exist outside New Zealand and those that do are evenly spread amongst the distance-cover classes.

The tabulation above also shows that for New Zealand and UTC hour 21, which is the hour of the daily climatological observation, most buddies were within 5–15 km of primary stations with a better than 95% coverage, few had buddies over 55 km away, and only six were in the worst distance-cover class. For other hours in New Zealand, most buddies also lay 5–15 km from the primary but the counts in the under 5 km and 15–25 km classes were generally of the same order, and a fairly even frequency of 30–40 primaries with buddies within each 10 km distance class existed out to about 75 km. Outside New Zealand the difference between hour 21 and the other hours was not marked, and most stations had buddies lying over 95 km away. Generally, this applied to 150–200 stations of which about a fifth were in the worst cover class, but for most hours about 30 primaries had buddies no more than 25 km away. Some further statistics regarding the buddies are tabulated below.

Hr	No. of 1ry stns	Number with given number of buddies						% Cover			Dist. to buddy (km)		
		Nil	1	2	3	4	5	Min.	Avg.	Max.	Min.	Avg.	Max.
00	669	98	571	165	46	16	4	19	96	100	0	79	658
03	618	108	510	142	43	11	2	12	95	100	0	78	658
06	622	95	527	147	41	14	4	17	96	100	0	83	658
09	535	123	412	122	31	9	2	15	93	100	0	80	598
12	589	107	482	129	41	11	2	21	95	100	0	97	658
15	538	129	409	122	33	10	1	18	94	100	0	81	601
18	652	109	543	163	51	14	3	23	96	100	0	87	658
21	1 040	94	946	361	110	38	9	12	97	100	0	49	658
D	265	13	252	116	29	5	3	16	98	100	0	43	482

Having established a set of buddies, the largest contemporary differences for windrun and, at each synoptic hour, for speed were found for every distinct primary-buddy pair. These were compared to the mean contemporary differences for the same hour, or windrun, and primary-buddy pair, i.e., the ratios MaxDifference/MeanDifference were calculated. The numbers involved are tabulated below.

Hour	Number of primary-buddy pairs
00	802
03	708
06	733
09	576
12	665
15	575
18	774
21	1 464
D	405

For the "D" class the 10% of primary-buddy pairs with the largest ratios were examined since those observations were potentially the most likely to be errors; for the other classes the top 5% were examined. This gave 217 distinct occasions which were examined by listing out from CLIDB for the station and time concerned the wind observation and the six observations either side of the given time, together with observations from neighbouring stations at the same times. By inspecting the listings it was decided whether observations were consistent with those nearby in space and time, or an amended value should be estimated, or the value should be removed.

The checking procedure was run three times and because some inspected values were correct they reappeared on subsequent runs but did need not be re-examined for error. Those that did not require correction were remembered from one run to the next through **RWIND_DIFFS** for windruns and **HWIND_DIFFS** for hourly winds. These tables were created by this checking procedure and have the following structure.

Column name	Null?	Type
HR (only for **HWIND_DIFFS**)		VARCHAR2(2)
AGENT_NO		NUMBER
BUDDY	NOT NULL	NUMBER
DIST		NUMBER
OBS_DATE	NOT NULL	DATE
P_WRUN or P_ SPEED		NUMBER
B_WRUN or B_ SPEED		NUMBER
PERC		NUMBER

For each HR, AGENT_NO, and BUDDY the values with the greatest difference occurred at OBS_DATE and are held in P_WRUN/SPEED and B_WRUN/SPEED, while PERC holds the percentile of this combination's maximum to mean difference. For example, those with PERC equal to 1 are the 1% of all the HR, AGENT_NO, BUDDY combinations which have the greatest relative difference. Thus, on a re-run the contents of **R/HWIND_DIFFS** can be moved to **OLD_R/HWIND _DIFFS**, say, before being over-written and rows common to both tables (except PERC which may change between runs) can be ignored.

The consequence changes made to **SURFACE_WIND** are tabulated below.

Run	Accepted	Amended	Deleted
1	66	81	70
2	87	58	58
3	45	47	25
Total	198	186	153

Apart from these individual changes other suspect data were found from the listings. In Check C.2 above some windruns on two consecutive days had been amalgamated since they were actually accumulations, now it could be seen that occasionally the amalgamation should have been over three days — and even four in two instances. For example:

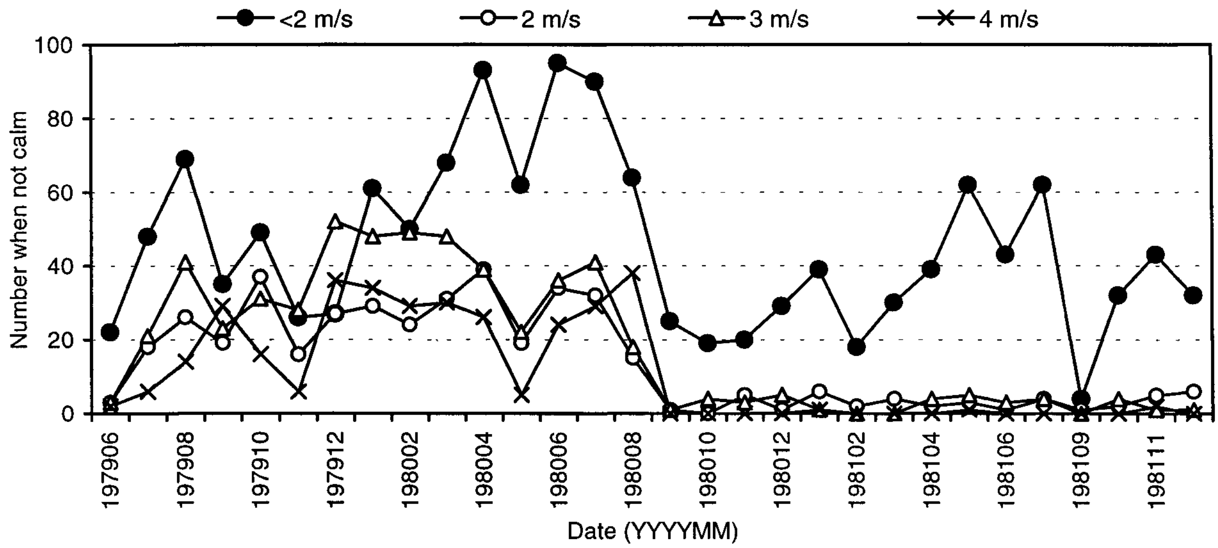
Situation before Check C.2 ➡ Situation "now", i.e., after Check C.2

OBS_DATE	AMOUNT	PERIOD		OBS_DATE	AMOUNT	PERIOD
YYYYMMDD	X	96		Record deleted	–	–
YYYYMMDD+1	X	24	➡	YYYYMMDD+1	2X	120
YYYYMMDD+2	X	24		YYYYMMDD+2	X	24

Situation "now", i.e., after Check C.2 ➡ Situation after present correction

OBS_DATE	AMOUNT	PERIOD		OBS_DATE	AMOUNT	PERIOD
YYYYMMDD+1	2X	120	➡	Record deleted	–	–
YYYYMMDD+2	X	24		YYYYMMDD+2	3X	144

where X is large enough for this check to highlight day YYYYMMDD+2 as having a value much larger than the mean and the fully correct situation is also shown. Many such instances were found for I59722/5666 where the windrun was read only once a week and, not only split into three equal sections, but also into two unequal ones with 1000 km in one and the remainder in the other. A total of 40 PERIODs were amended and 42 unnecessary rows deleted; nine stations were involved.



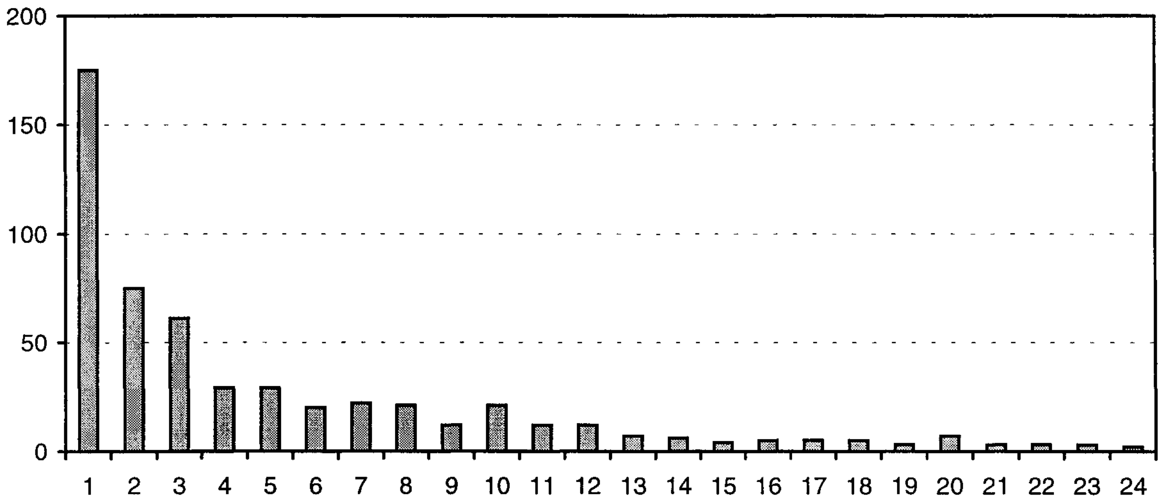
The AWS station E95467/3632 often reported calm while the manual station E95465/3630 less than 1 km away reported significant wind. The figure above shows a time series of the number of times that is was not calm at E95465/3630 when it was calm at E95467/3632 and these counts have been stratified by the wind speed. Clearly, the frequency of speeds below 2 m/s does not vary greatly, while for other speeds it is clear that before September 1980 there were many discrepancies between the stations, but from that time they agreed. An entry in **SITE_CHANGES** mentions the sheltered nature of E95467/3632 and so the 3053 rows at that station up to September 1980 were deleted.

Similar, but less pronounced, situations occurred at three other AWS/manual station pairs. Different periods were involved and rather than delete all rows during affected periods only calms at the AWS when the manual was reporting at least 5 kt were deleted. The stations, periods, and numbers deleted are tabulated below.

AWS station	Manual station	Period of deletions	Number of deletions
C84174 / 2136	C84173 / 2135	Dec 1979–Feb 1980	298
C84174 / 2136	C84173 / 2135	Mar 1981–Apr 1981	98
D78752 / 2710	D78751 / 2709	Jun 1980–Mar 1981	339
I69482 / 5893	I69481 / 5892	Dec 1983	10
Total			745

Details and results of Check C.7 — are all wind records without gaps?

Wind is continuous in time but, apart from the traces on anemometer charts, its observations are taken at discrete time intervals. In CLIDB all OBS_DATEs are on the hour and observations at hourly intervals are available from some places. More common are the synoptic observations taken every 3 h and the climatological ones taken at 0900 Local which also include the daily windrun observations. There are always gaps, but only for an hourly record is a break of 1 h necessarily a gap since for synoptic records that hour might not have been a synoptic hour. Thus, observations missing at synoptic hours are necessary for a gap to exist in a synoptic record and at 0900 Local in a climatological record. Ideally, for a given AGENT_NO there should be no breaks in the particular type of record from when it started until either the present day or when the station closed. This is extremely rare since missing data occur at even the best stations. Thus, rather than a search for errors, this check is more a quality check in which the "completeness" of the station records is examined.

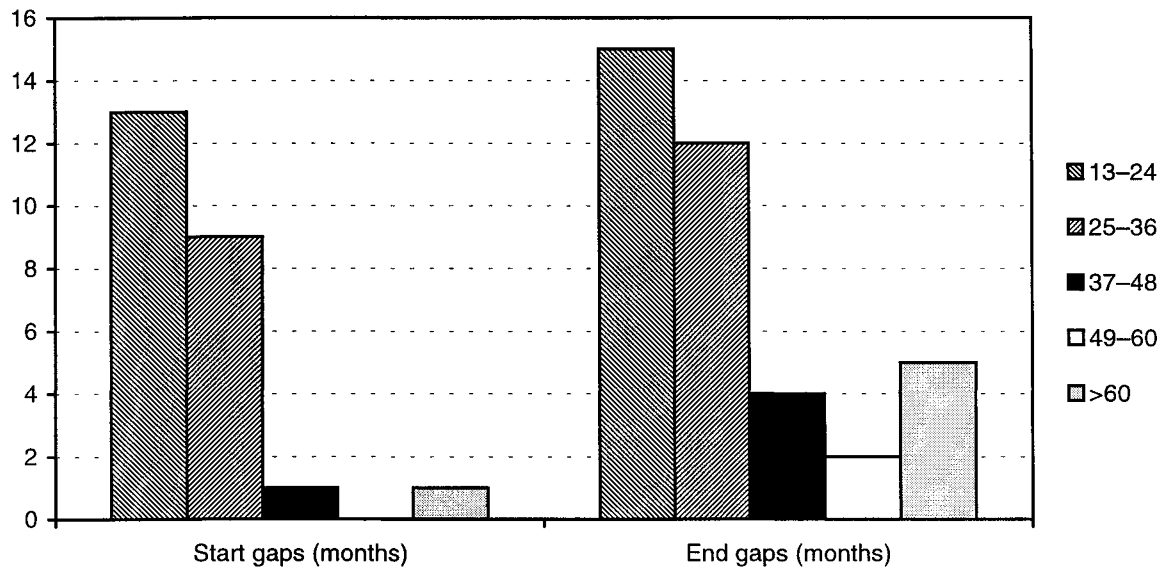


Distribution of short record lengths

However, as an initial step extremely short records were discarded. The distribution of short records is shown in the figure above where it can be seen that 175 stations had record lengths of just one row and 75 of two rows, etc. Such records are not easy to explain or correct and usually do not represent real observations, rather they originated from reports with coding errors. Thus all the rows that constituted records of less than 25 rows were deleted, i.e., a total of 2731 rows which were all FREQUENCY "H" and came from 607 station and hour of observation combinations.

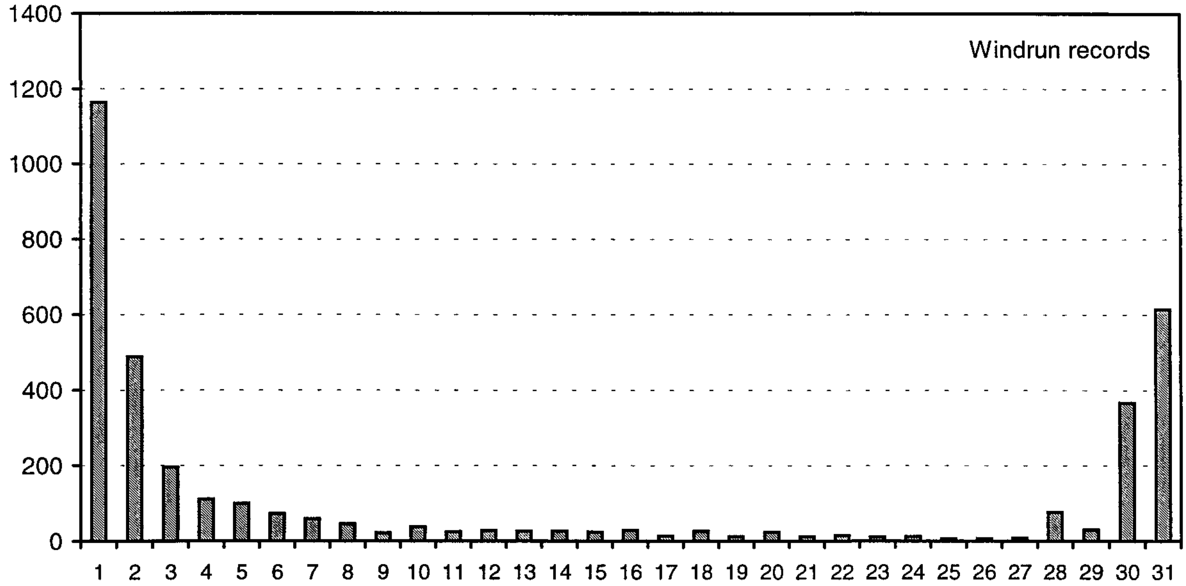
Having removed the shortest records, stations where the completeness was small were examined, and, although most of these would have to be accepted as due to missing data, there were two types of error that it might be possible to correct. First, if data from a station are wrongly attributed to another station which had been closed for some time, then this closed station has its record incorrectly extended but, since a large gap occurs in the record just before the last data, its completeness is low. The second error is the same in principle, with data from a different station attributed to another station but this time before it was opened. A slight variation to these errors is where the station to which the data were attributed was correct but a wrong date was used. At this stage it is not necessary to differentiate between hourly, synoptic, and climatological records or FREQUENCYs "D" and "H".

The only gaps considered were those where the period covered between the gap and the end of the record was less than 2 months. Such gaps before or after the real station record could be of any length from many years down to just a few — or even nil — days. However, as far as completeness of record and ease of error detection is concerned, long gaps are the most significant and so in the figures below only gaps of at least a year are included. There were 224 shorter gaps at the start of records and 264 at the end bringing the total number of such gaps to 248 and 302 respectively. The records being referred to in the figure below were for different station and hour of observation combination and the 62 records involved only 37 stations. Nearly all the gaps were due to one or two early or late reports of synoptic origin, however, there were some cases where a whole month or more of data was concerned. A total of 461 rows was deleted which included 246 from just 4 Antarctic stations, 92 from 3 Fijian stations, and 93 from New Zealand stations of which half were from E05361/3236.



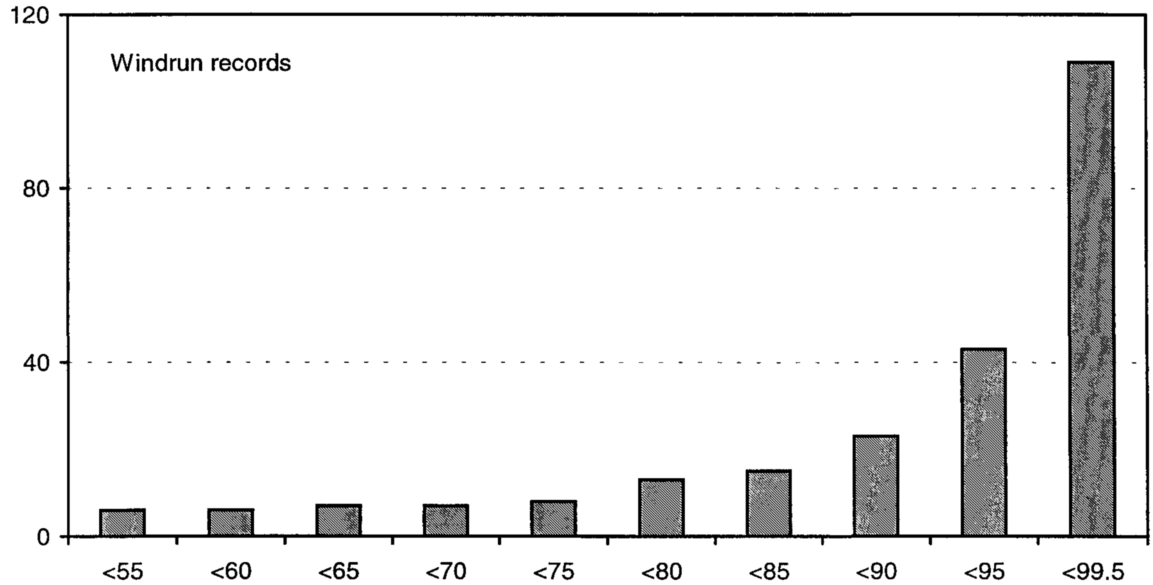
Distributions of gaps at the ends of station records

The remainder of this section is a description of the state of the windrun, climatological, synoptic, and hourly records after the changes described above had been made. The figure below is for windrun observations and shows the distribution of missing days per station-month. There were nearly 1200 station-months that had a single day missing, about 500 with two days, which could be either together or apart, etc. These numbers are a significant fraction of the total number of windrun observations, which is equivalent to 32 800 station-months of which 3700 have some days missing. The counts for the 28–31 day classes are much larger than for all but the first few classes. If the numbers for classes 28 and 29 are taken together as representing February, then the numbers for classes 30 and 31 are about four and seven times larger. There are four months of the year with 30 days and seven with 31 days, thus the higher numbers for classes 28–30 and not just those for class 31 are due to complete months being missing — a total of 1100 whole station-months are missing.



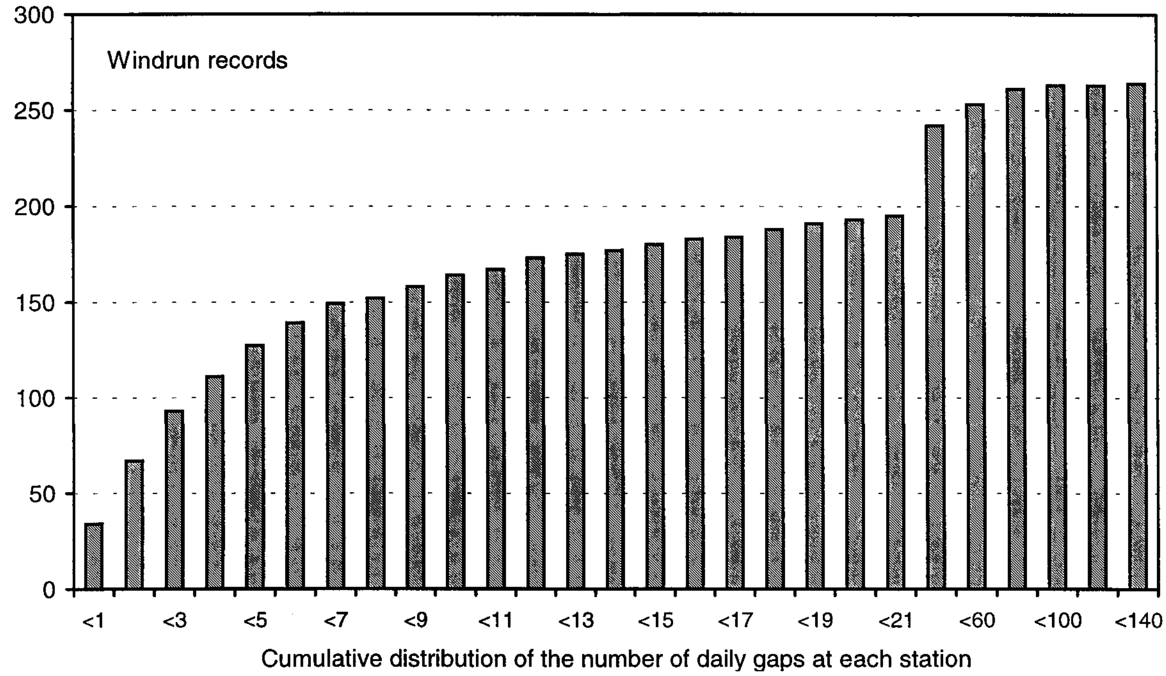
Distribution of missing days per station-month

How are these missing days spread among the stations? CLIDB has 264 stations with records of windrun observations and 73 of these have near perfect records. The cumulative distribution of the percentage complete of the records at the other 191 stations is shown below and shows that about 23 of these have records less than 90% complete and only 5 records are under 50% complete.

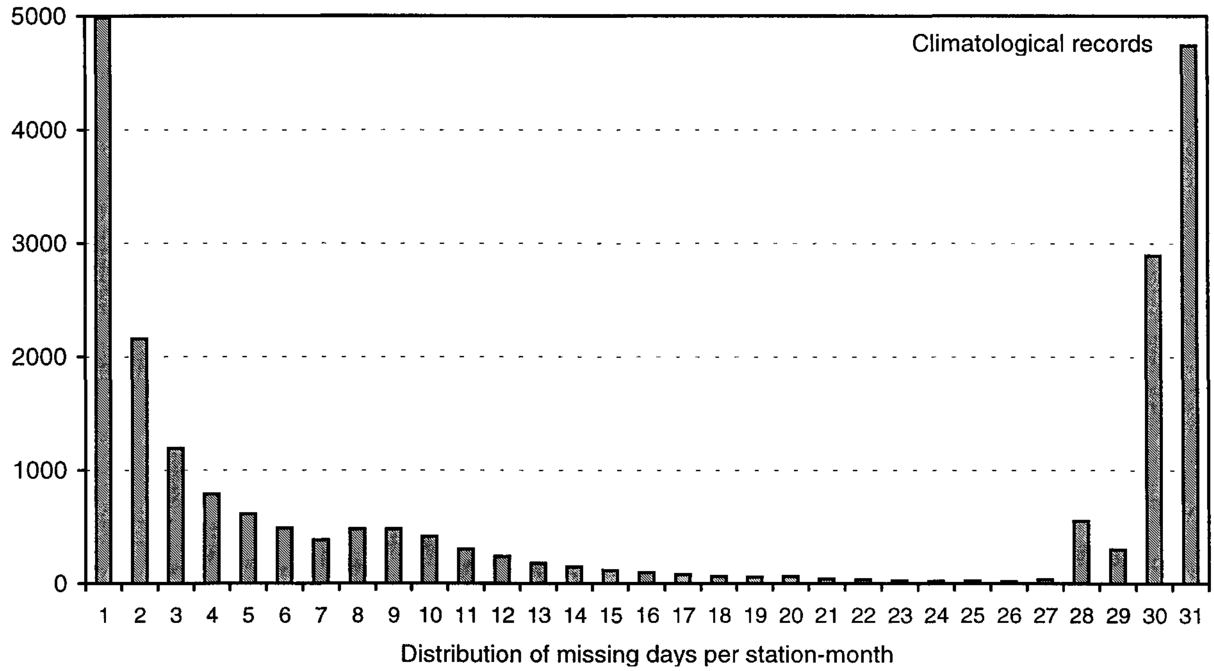


Cumulative distribution of percentage completeness of windrun records

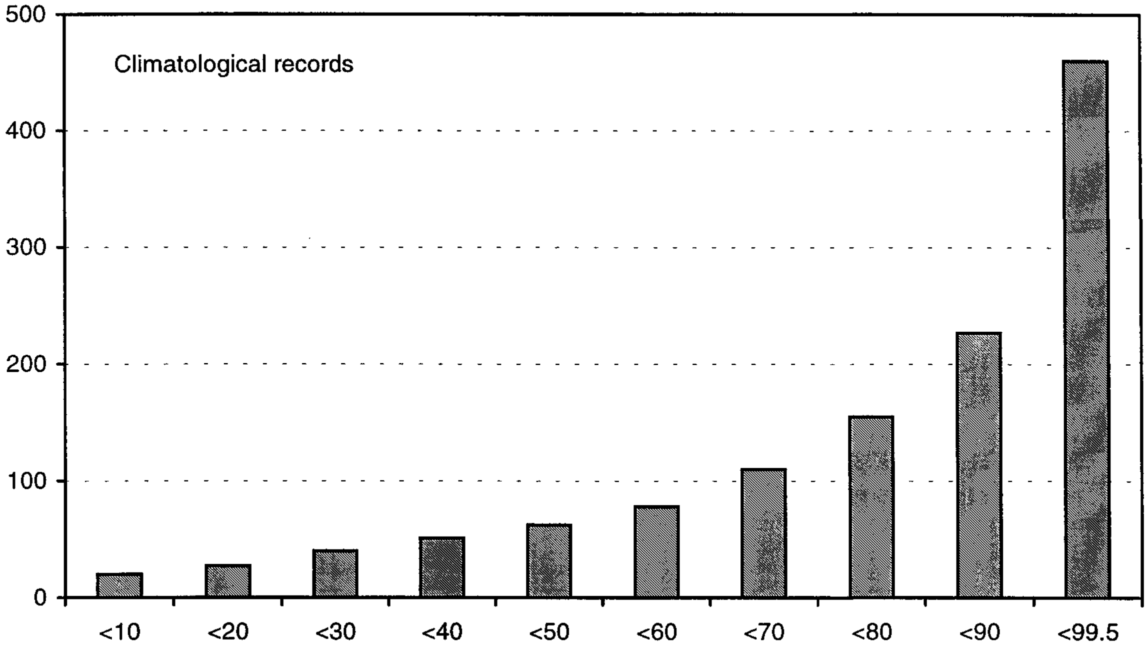
The worst stations are, of course, those where the percentage complete is small, but those with a large number of gaps, rather than just a low percentage complete, are also of poor quality. This is because many gaps are a sign that the station has been unable to keep up a programme of regular observations, whereas a few large gaps could well mean that, although the station had to be closed occasionally, it was otherwise a regular observer. Thus, the best stations are the 34 without any gaps in their windrun records (i.e., those in the <1 class in the figure), and about half of the records have fewer than five gaps. There is a change of class width after the <21 gap class.



The next three figures are similar to the last three, but show the respective distributions for the climatological (i.e., 0900 Local) observations. The figure below shows that there were nearly 5000 station-months that had a single day missing, about 2000 with two days, which could be either together or apart, etc. The total number of climatological observations is equivalent to 72 000 station-months of which 22 000 have some days missing, and total of 8500 whole station-months are missing.

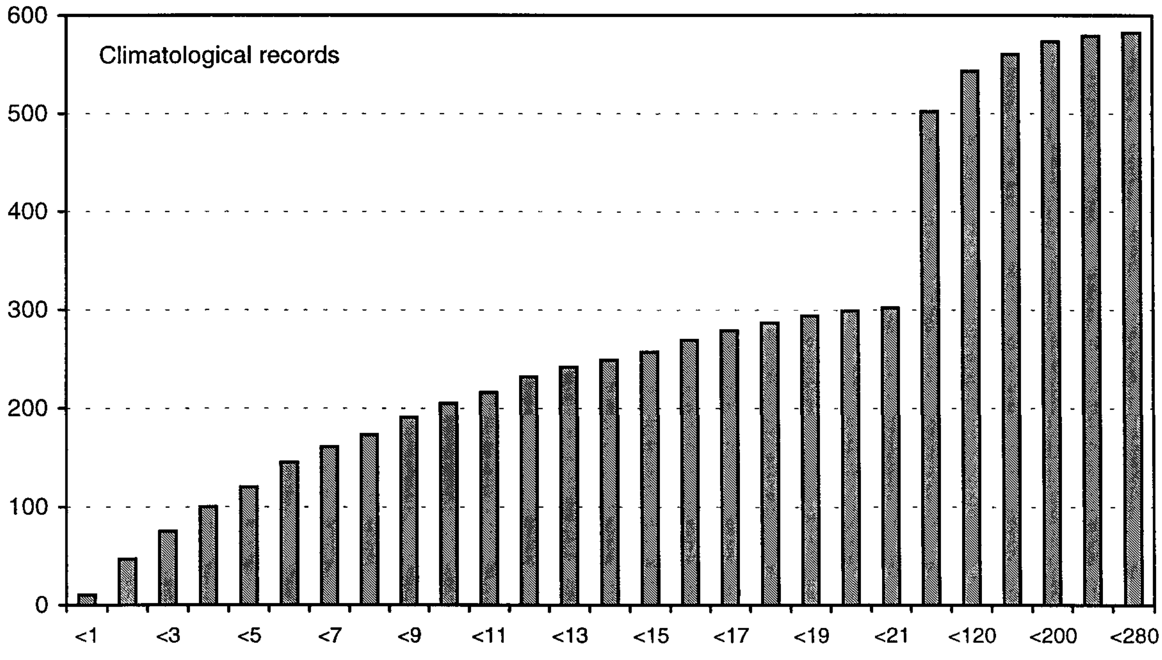


CLIDB has 584 stations with records of climatological observations and 124 of these have near perfect records. The cumulative distribution of the percentage complete of the records at the other 460 stations is shown below and shows that about half of these have records less than 90% complete, but only 60 records are under 50% complete.



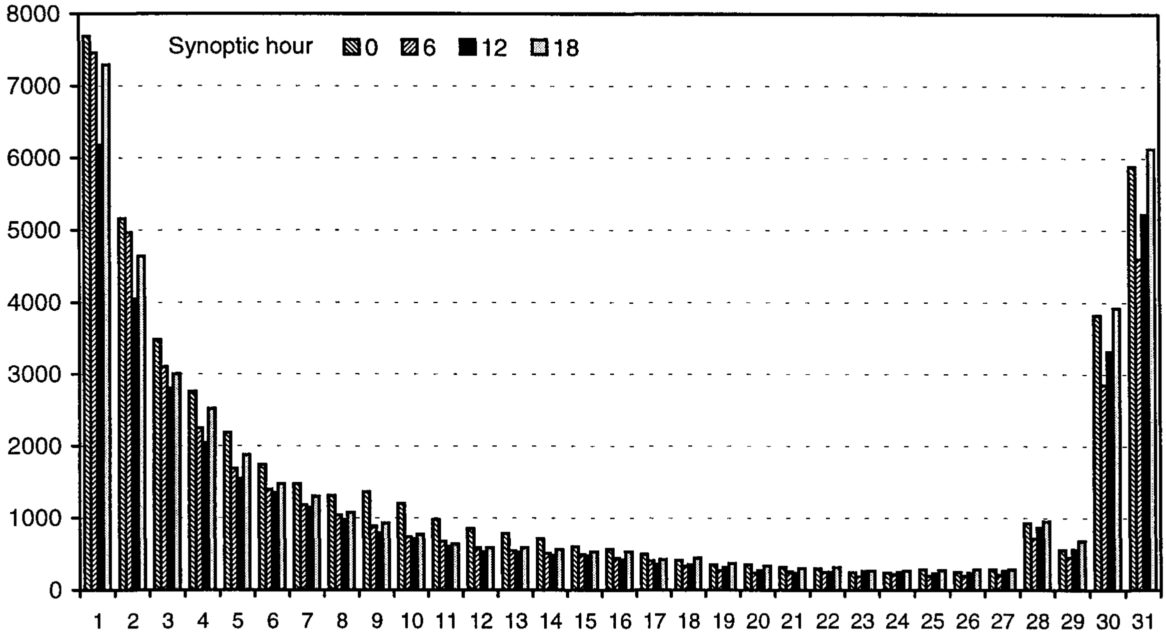
Cumulative distribution of percentage completeness of climatological records

The best climatological stations are just the 10 without any gaps in their records (i.e., those in the <1 class in the figure below), but about 100 of the records have fewer than five gaps. There is a change of class width after the <21 gap class.



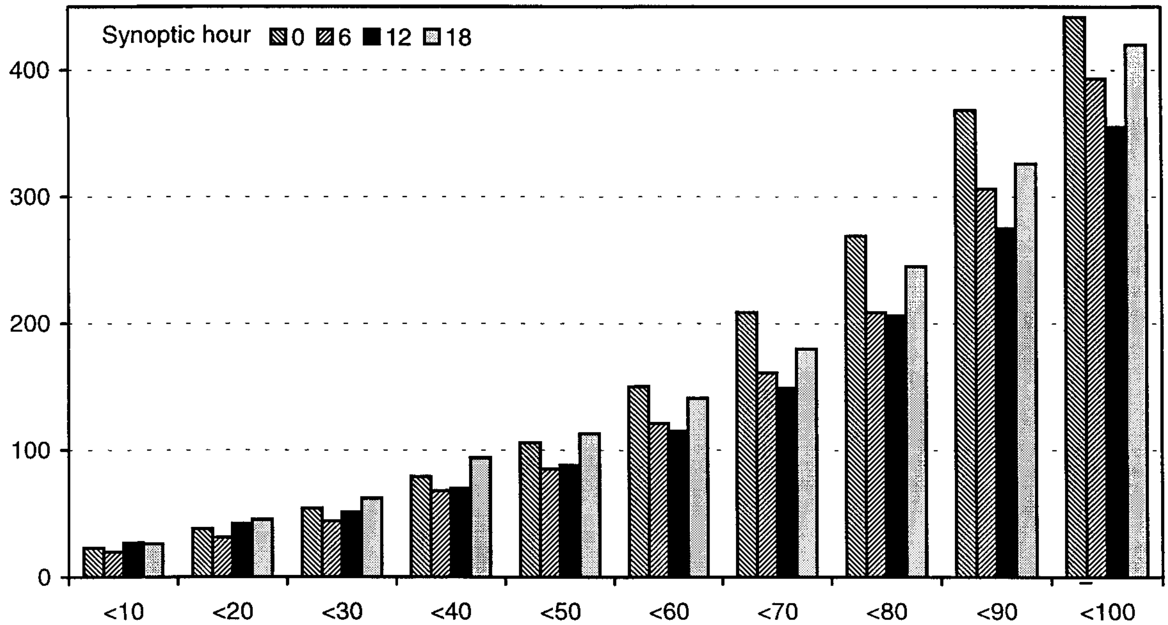
Cumulative distribution of the number of daily gaps at each station

The next three figures are similar to the last three, but show the respective distributions for synoptic records at the main synoptic hours of 00, 06, 12, and 18. It can be seen that they differ little with the hour except the numbers for 06 and 12 tend to be a little lower than for the other hours.



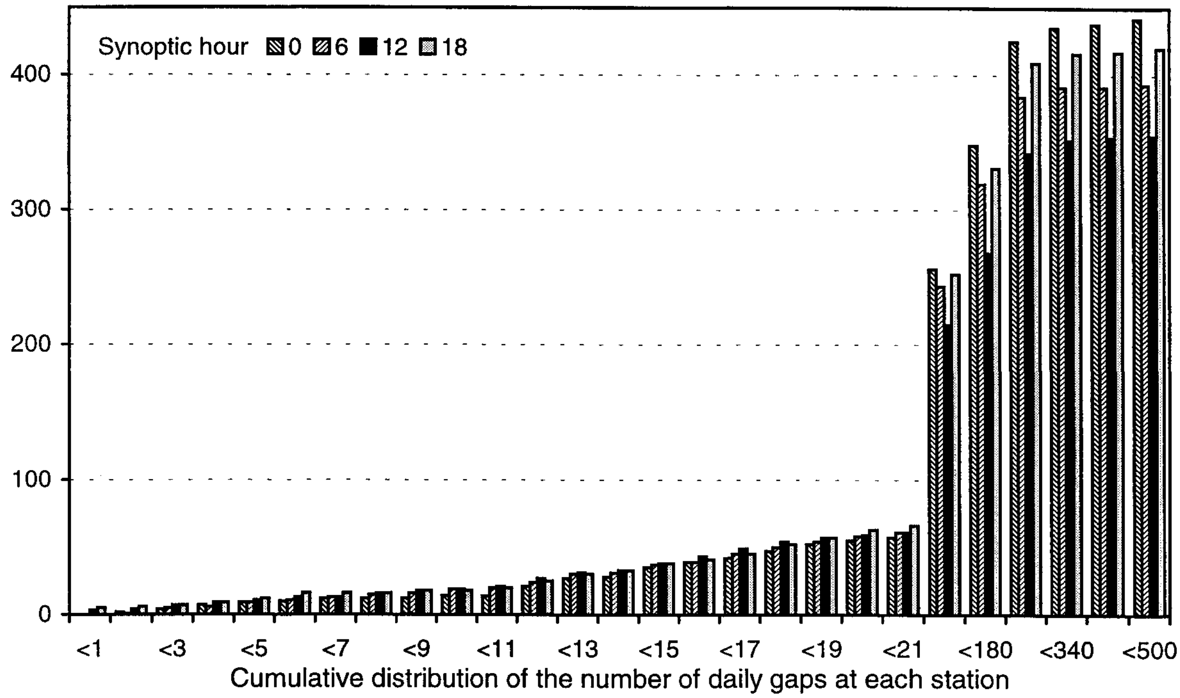
Distribution of missing days per station-month

The figure above shows a similar pattern to that of the distribution of missing days for climatological records, but the numbers in all classes from 2 to 27 inclusive are up to twice what they were. Also, there are about 400 station records for each of the hours whereas there were nearly 600 for the climatological records. Thus, even if the class memberships were identical for climatological and synoptic records, the latter would have relatively more gaps.

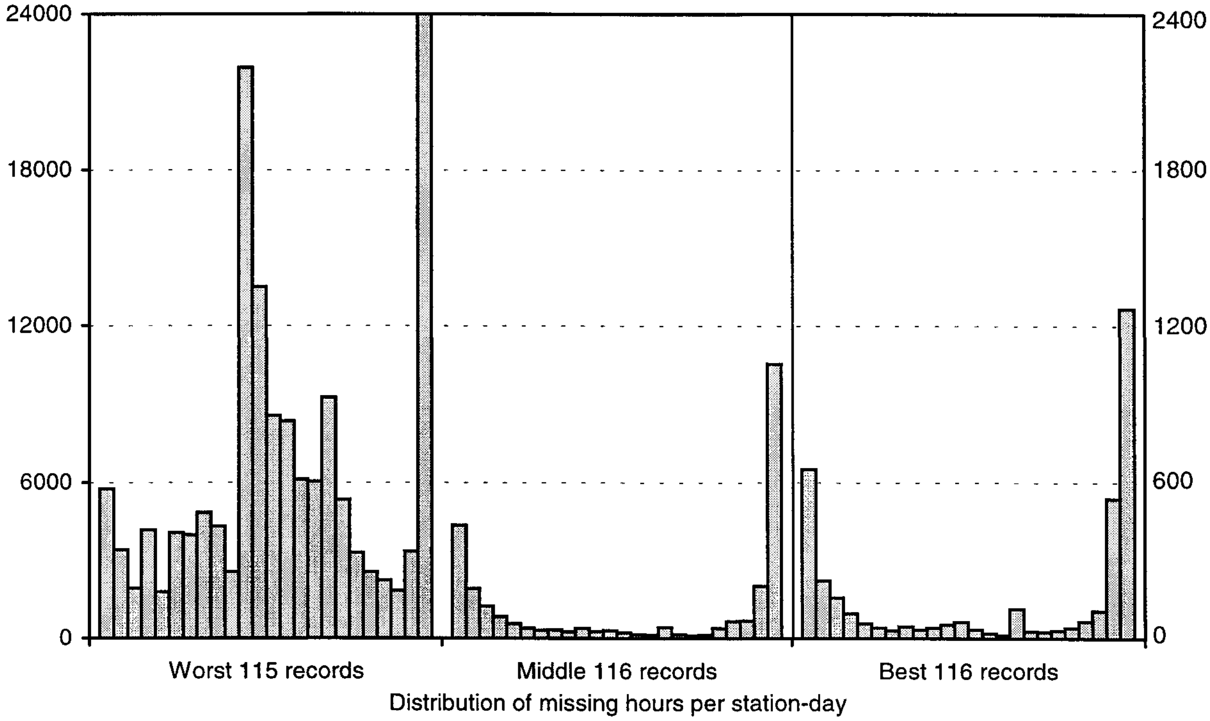


Cumulative distribution of percentage completeness of synoptic records

The figure above is slightly different to that for the cumulative distribution of the percentage completeness of climatological records in which the near perfect records were excluded. In the synoptic records there were few perfect records and the rightmost class above includes the few that there were. The figure shows that for all classes there are relatively more members than in the climatological records, e.g., the <90 class indicates that about 75% of the synoptic records are less than 90% complete compared to about half for the climatological records which can be seen from the figure showing its distribution.

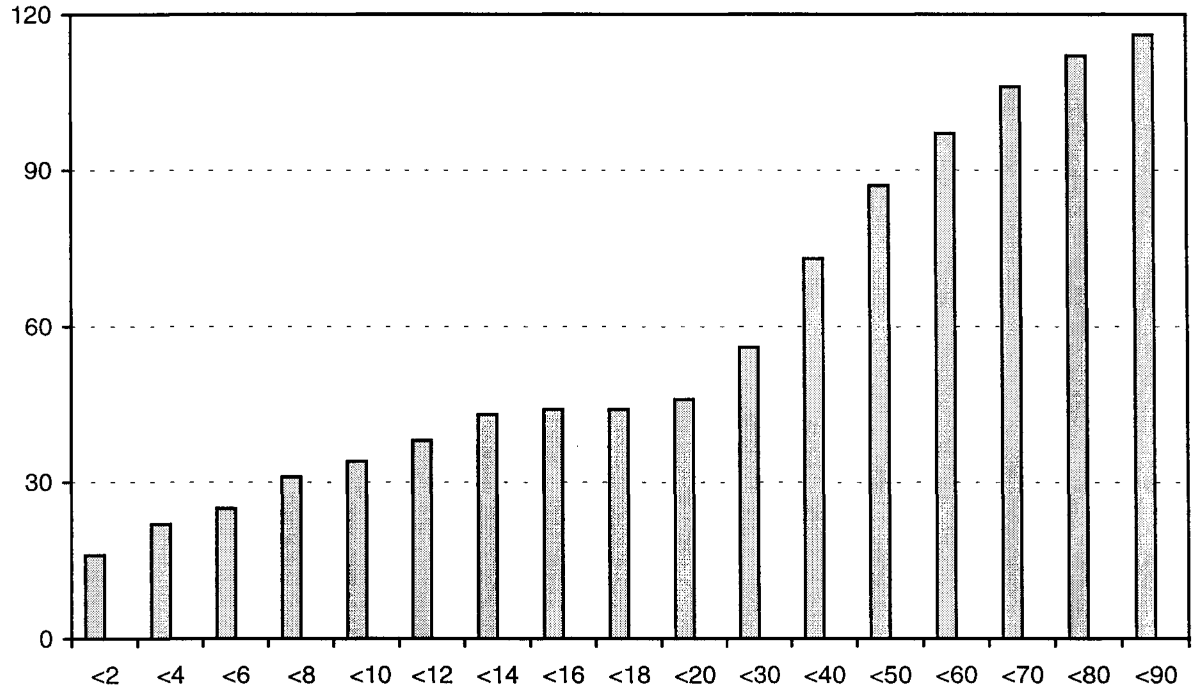


The figure above again shows that synoptic records are of a lesser quality than the climatological ones since about half of those records had under 18 gaps, whereas for synoptic records it can be seen that half of the records have up to 100 gaps.



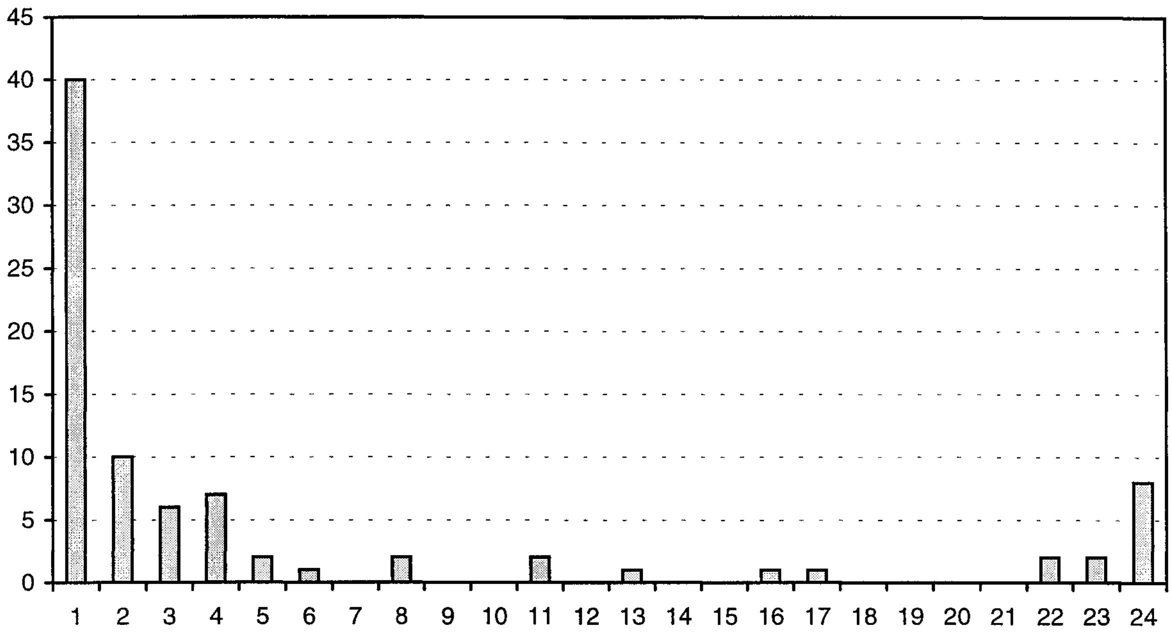
The records of hourly observations were examined and the figure above shows distributions of missing hours per station-day. There were 347 stations with hourly records and in the figure these have been split into three groups and a distribution shown for each group. The left hand scale applies to the first two groups, but it is 10 times too big for the third group so, for example, for the best records about 650 station-days had a single hour missing and about 200 had 2 hours, which could be either together or apart, etc. This "best" group represents stations which always reported hourly, although the count of 110 at 16 hours missing probably represents times when stations were reporting only at synoptic hours. On the other hand, the "worst" group represents those stations that reported only during the daytime, hence the highest frequency at 11 hours. The "middle" may include some stations belonging to the worst group, but probably consists mainly of poor quality hourly stations. The counts for the 24 h classes, which represent the number of whole station-days that are missing, are much larger than for all other classes. This is especially so for the "worst" group where the number at 257 709 is off the scale of the figure and is a significant fraction of the total of 750 000 station-days covered by this group. The "best" group covered 135 690 station-days, so even the membership of the 1 h class is relatively small when compared to the total number of hourly observations in CLIDB for that group.

How are these missing hours spread among the stations? Only the "best" stations will be considered as none of the "worst" and some of the others are not stations that are expected to report every hour. There are 116 such stations with records of hourly observations with over half of these having records at least 98% complete and only one record being under 50% complete. The worst stations are, of course, those with the lowest percentage completeness but, as with the windrun, climatological and synoptic records, those with a large number of gaps, rather than just a low percentage complete, are also of poor quality. There were four stations without any gaps in their hourly observations, 16 stations had fewer than 4 gaps, and under half the records have over 30 gaps. The figure below shows the distribution of hourly gaps; there is a change of class width after the <20 gap class.



Cumulative distribution of the number of hourly gaps at "best" station

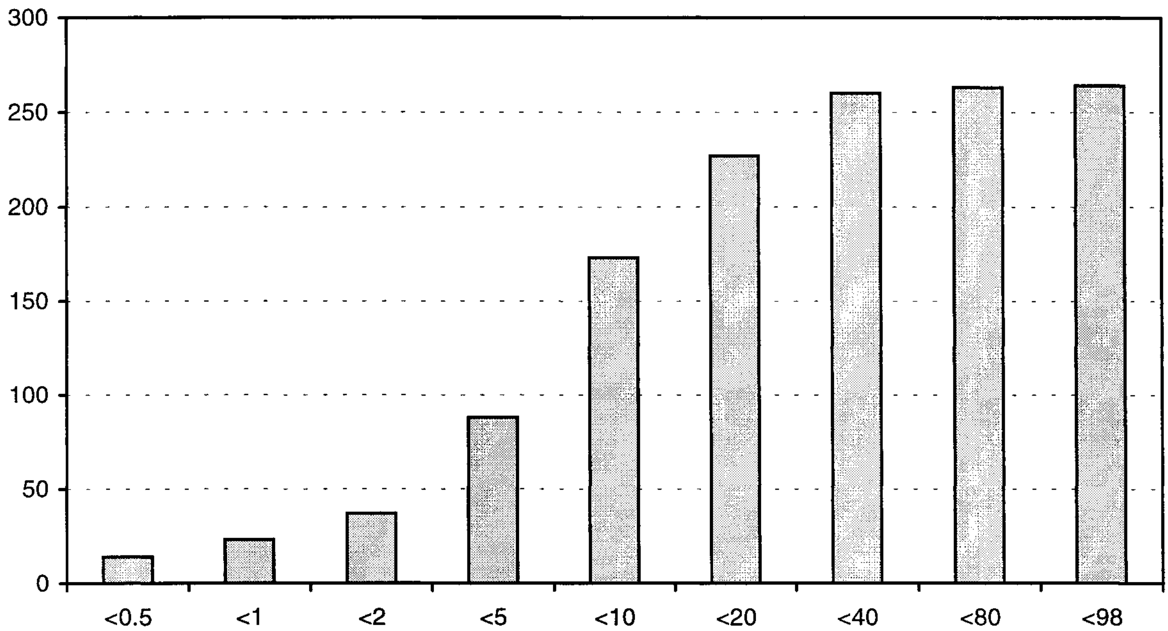
Out of the best stations, the station with the most gaps was C94012/2283 with 85 gaps, and the distribution of gaps for this station is shown in the figure below.



Distribution of gap sizes (h) at station C94012/2283

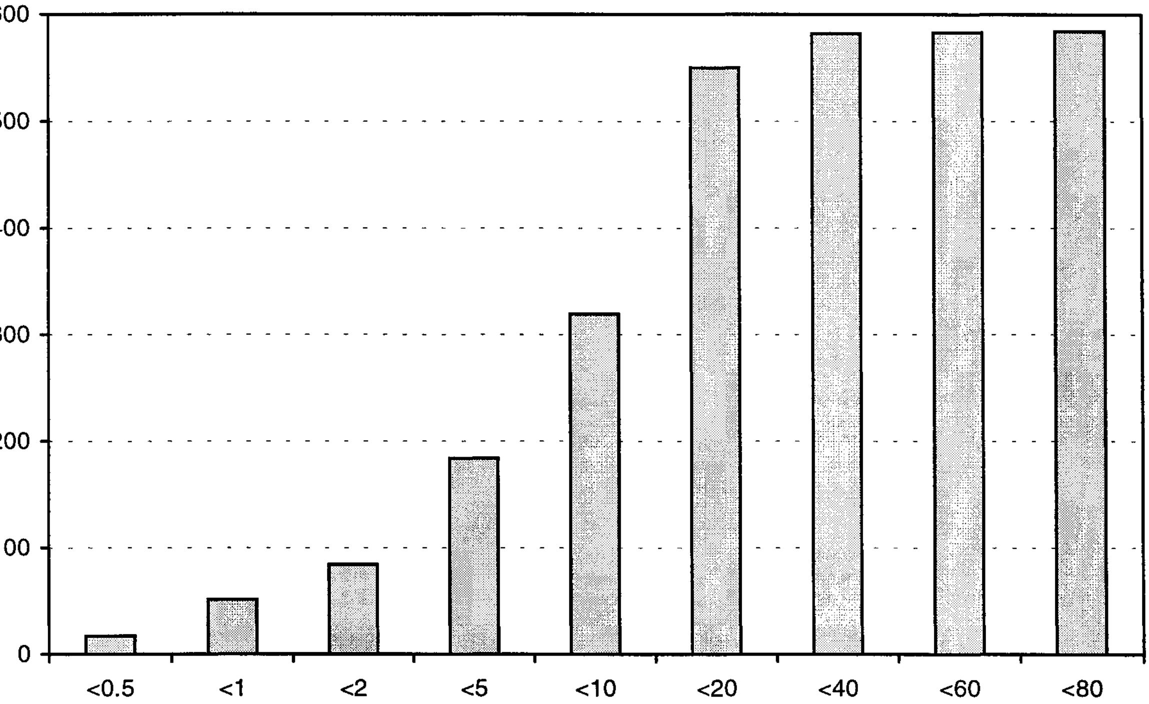
Details and results of Check D.1 — are all wind records long enough?

For a given AGENT_NO the windrun, climatological, and hourly records — synoptic records need not be considered — should be long enough to establish the mean level and variability of the wind for the place concerned. Longer records can be used to track any trends, while short records, although still useful as observations, do suggest poor quality. But "How long is long enough?" is not a question with a definitive answer and the best course is to simply examine the distribution of the record lengths, which is shown in the figure below for windrun records.



Cumulative distribution of lengths of windrun observation records (years)

For windrun records, only 23 records were under a year long and over half of the 264 records were over 10 years long. The longest record was from H32641/4881, which lasted from January 1891 until December 1987 and is 99% complete. Most of the records of under a year, which were not from stations which had opened within the last year, were nearly fully complete and so were just short records. However, of the 13 such records 2 were under 50% complete and were recognised as special automatic stations at Auckland and Christchurch whose data is operationally accredited to the long-term manual climate stations collocated with these automatic ones. Some of the data were missing from the manual station and 58 rows were amended to transfer the data to the manual station. The remaining data were duplications and the 55 rows concerned were deleted.



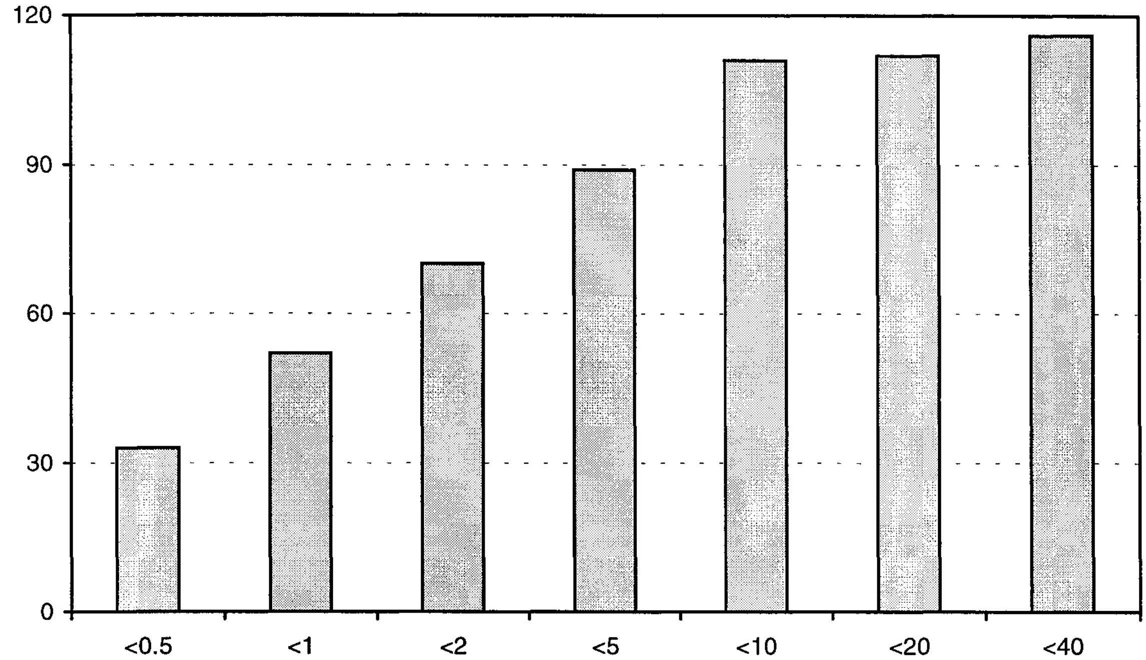
Cumulative distribution of lengths of climatological observation records (years)

For climatological records, only 51 are under a year long and nearly half of the 584 records are over 10 years long. The longest record is from H40041/4970, which lasted from January 1928 until December 1991 and is 78% complete. Of those records under a year which were not from stations which had opened within the last year, four had a total of only 168 rows, and two had only patchy records so 69 rows were deleted. However, E95352/3601 is a station on the Turoa skifield and it had a complete month so its data were kept, and I59992/5692 was a special Forest Service station that had only been open for a short time and its wind data was 70% complete so its data were also kept.

The distribution of the record lengths for the "best" hourly observations is shown in the figure below. About 50 records are under a year long and only a quarter are over 5 years long. The longest record is from H32451/4843, which opened in January 1960, is still reporting, and is nearly perfect. Of those records under a year which were not from stations which had opened within the last year, three were especially short, i.e.,

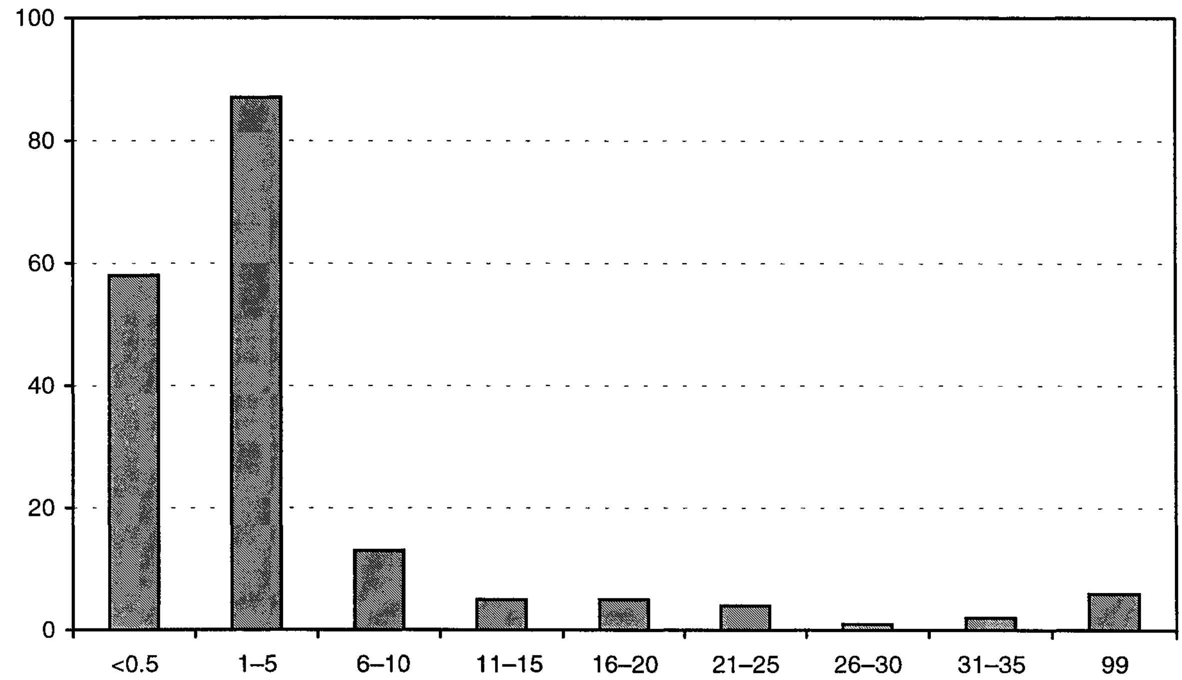
- H22871/4532 had a near perfect climatological record from January 1972 to January 1991 and six weeks of hourly observations from 1 January 1987 whose origin is not known but they were accepted as genuine;
- H31451/4677 is a station on the Mt. Hutt skifield with a patchy climate record but a perfect month (August 1983) of hourly observations: however, there were also three single observations in the 1990s which were deleted;

- H32673/4905 was a special high altitude wind station only open for six months during 1981, the record was patchy but was retained.



Cumulative distribution of lengths of "best" hourly observation records (years)

Details and results of Check D.2 — are the number of accumulations, during and between months, reasonable?

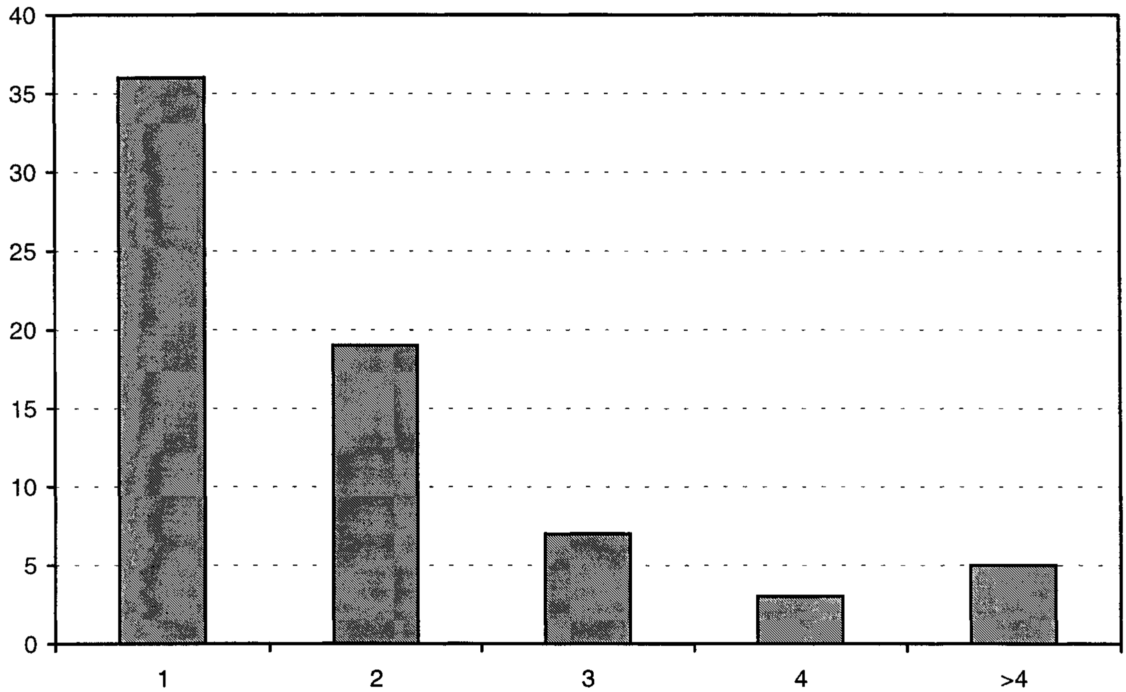


Distribution of percentages of records that are accumulations

For a given AGENT_NO and FREQUENCY of "D", the number of times that an observation applies to more than one day should be small compared to the total number of daily readings for that AGENT_NO. Nearly a third of the 263 stations have no accumulations and the percentage of accumulations for those stations with accumulations is distributed as in the figure above which shows that a further quarter of the records have only 0.5% as accumulations and that most records were at most 5% accumulations. However, there were six records that were over 99% accumulations, their details are tabulated below and show that, despite the degree of accumulation, the records were nearly all complete.

Station	Start date	End date	No. of rows	% Complete
I50084 / 5270	Jan 1975	Aug 1983	522	97
I50086 / 5272	Oct 1983	Nov 1985	132	100
I50721 / 5336	Jan 1980	Sep 1981	109	100
I59693 / 5661	Oct 1981	Jun 1986	292	100
I59722 / 5666	Jan 1977	Aug 1980	198	89
I68174 / 5775	Mar 1984	May 1988	263	100

Again for a given AGENT_NO and FREQUENCY of "D", few, if any, accumulations should be cross-month accumulations, i.e., those that start in one month and end in another. The figure below shows their distribution among stations Thus, 36 stations had a single occasion when a daily observation had a PERIOD such that the start of the period was in the month before that of OBS_DATE.,A total of 70 stations was involved with a total of 149 cross-month accumulations.



Distribution across stations of the number of cross-month accumulations

Details and results of Check D.3 — are monthly wind statistics consistent with the observations upon which they are based?

Monthly summary statistics are calculated and entered into **MTHLY_STATS**: from FREQUENCY "D" rows the statistics concerned are the mean daily windrun and the maximum daily windrun, and from FREQUENCY "H" rows the mean wind speed. These monthly statistics are also used in the estimation of the monthly Penman potential evapotranspiration and open water evaporation. There are certain rules associated with the calculation of these statistics which ensure they are valid and are exactly as defined. For example, if for a set of FREQUENCY "D" rows with the same AGENT_NO and with all OBS_DATEs falling within the same local month, the maximum SPEED was associated with a PERIOD of over 24 h, then no maximum daily windrun can be found for that month. However, PERIODs of over 24 h in the set of rows do not preclude the extraction of the maximum since it is only required that the maximum SPEED was with a 24 h PERIOD.

In the example, and in the much commoner instances where data are missing, no statistics are possible and their absence is not an error. Rather this check should look for cases where a statistic exists despite the **SURFACE_WIND** data being deficient. However, it is somewhat easier to just recalculate the statistics since erroneous ones would get deleted. During such a recalculation an attempt would be made to calculate statistics for every station-month that is represented within **SURFACE_WIND** and some of these would fail through lack of data or other legitimate reasons that do not occur because an error exists in **SURFACE_WIND** itself. But there are some failures which could be associated with errors in **SURFACE_WIND**, and this check captured those potential errors.

The errors reported that might indicate errors in **SURFACE_WIND** are: extra days, which are seen as a negative number of missing days and result from the sum of the PERIODs of the rows for the concerned month exceeding the length of the month (Extra days); rows exist where nothing is recorded for SPEED; a cross-month accumulation (First wrong); despite an error a non-deletable statistic exists (Stat remains); and data with an origin not normally associated with **SURFACE _WIND** (Bad origin). It is also possible that any cross-month accumulations had already been accepted and were the reason for the first type of error, so only the occurrence of extra days without a cross month accumulation are reported. The following reportable errors occurred.

AGENT_NO	Date of error	Type of error	Cause of error and action
1858	Sep 1990	First wrong	6-day accumulation to 4th – Nil action
2473	Mar 1991	Stat remains	Windrun for 31st missing – Statistics for codes 15, 34, and 36 were deleted
3238	Dec 1999	Bad origin	Sunshine, which is used in calculation of codes 34 and 36, on 21st had an origin of "R" – Origin was changed to "D"
4325	Oct 1987	Stat remains	Windrun for 7th missing but estimated value for statistic was acceptable – Nil action
5867	May 1994	Extra days	4-day accumulation to 3rd – Nil action
5867	Mar 2000	Extra days	Windrun periods for 11–12th overlapped – Periods were amended
6256	Nov 1979	First wrong	2-day accumulation to 1st – Nil action
6256	Feb 1981	First wrong	3-day accumulation to 1st – Nil action

Summary and Conclusion

The grand total of changes made to **SURFACE_WIND** was 77 638, which is 0.3% of its total number of rows. The changes are summarised in the following tabulation.

Table name	Deletions	Amendments	Inserts
SURFACE_WIND	19 915	57 723	0
LAND_STATION	0	77	0
SITE_CHANGES	42	74	23
Other tables	3	39	0
Total	19 960	57 913	23

For deletions and amendments, about 90% of the cases came from just three and two sources respectively.

- After examining the times series of observations for steps and blips 12 233 rows were deleted from 281 stations.
- The first 3053 rows from E95467/3632 up to September 1980 were deleted as it often reported calm while the manual station E95465/3630 less than 1 km away reported significant wind.
- Incorrect synoptic observation times were corrected by changing the time to the nearest correct synoptic observation time for 8927 rows but in 2476 other rows the nearest synoptic hour already had an observation so these rows were deleted.
- Observation of calm conditions is stored with zeros for both SPEED and DIRECTION, but in 43 979 rows a zero SPEED had a non-zero DIRECTION. These DIRECTIONs were amended to zero.

The need for the changes to the most noticeable errors could have been found at any time and it is, perhaps, the other, more particular, changes which are the most valuable since the subtlety of many of the errors kept them so well hidden that only the auditing was likely to find them.

Apart from the changes to the data, some changes to programs were also made.

- A new procedure **WRITE_SURFACE_WIND** was written to follow all the rules which apply to the insertion and amendment of data into **SURFACE_WIND**.
- The new procedure **WRITE_SURFACE_WIND** was incorporated into the following archiving procedures RMSDYCLI, RMSEDR, RMSHOURLY, RMSMETAR, RMSSYNOP, COPIDYCL, and DEUPDATE.

References

Penney, A.C. 1999: Climate database (CLIDB) user's manual. Fourth edition (revised). *NIWA Technical Report 59.* 161 p.

Sansom, J. & Penney, A.C. 1999a: New Zealand's National Climate Database (CLIDB): audit report on the MTHLY_STATS table. *NIWA Technical Report 62.* 38 p.

Sansom, J. & Penney, A.C. 1999b: New Zealand's National Climate Database (CLIDB): audit report on the RAIN table. *NIWA Technical Report 65.* 36 p.