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

John Sansom
Allan C. Penney

NIWA Technical Report 80
ISSN 1174-2631
2000

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

John Sansom

Allan C. Penney

NIWA Technical Report 80
2000

Published by NIWA
Wellington
2000

Inquiries to:
Publication Services, NIWA,
PO Box 14-901, Wellington, New Zealand

ISSN 1174-2631
ISBN 0-478-23206-3

Citation: Sansom, J. & Penney, A. C. 2000:
New Zealand's National Climate Database (CLIDB):
audit report on the SCREEN_OBS table.
NIWA Technical Report 80. 39 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 SCREEN_OBS table. ***NIWA Technical Report 80.*** **39 p.**

The auditing of the dataset within New Zealand's National Climate Database which contains temperature and humidity observations is described. Each row in the dataset consists of the place of observation, the date and time of observation, the air temperature, a number of different parameters all expressing measures of humidity, 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 temperature and humidity 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 **SCREEN_OBS** was 379 932, which is 1.9% of its total number of rows. The largest contribution (297 419) was where the temperature was under −10 °C, and was not changed, but WET_BULB and DEWPOINT (types of humidity measures) were set to NULL. This implemented a new rule which recognises that at cold temperatures both WET_BULB and DEWPOINT become poor means of conveying humidity measurements. There were a number of other major changes, none of which applied to New Zealand stations. For example, nearly 18 000 were for 28 Antarctic stations where time consistency had been used to detect rows with large differences from their temporal neighbours. Also, about 7000 deletions were for just three Pacific island stations for the rows which had a DEWPOINT of 0 °C and the temperature was a multiple of 5 °C. Finally, about 5000 amendments were made to the time of observations for some non-New Zealand stations where the times needed to be moved either back or forward 1 h. 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_SCREEN_OBS** was written to follow all the rules, old and new, which apply to the insertion and amendment of data into **SCREEN_OBS**. This new procedure was incorporated into the nine procedures that are regularly used to insert new data or amend existing data.

Introduction

This report is the third in a series which will document 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, etc.) 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 records. 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 from its full range.

Generally in a table some of the columns hold the *primary key* which, rather than being the data itself, are details about the "where", "when", and "what" of the data. The primary 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. 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 a spatial one.

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 primary 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 SCREEN_OBS table

The **SCREEN_OBS** table contains temperature and humidity 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
DRY_BULB		NUMBER(5,1)
WET_BULB		NUMBER(5,1)
DEWPOINT		NUMBER(5,1
RELATIVE_HUMIDITY		NUMBER(4,1)
DRY_BULB_REL		VARCHAR2(1)
WET_BULB_REL		VARCHAR2(1)
ORIG_DRY_BULB		VARCHAR2(1)
ORIG_WET_BULB		VARCHAR2(1)

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** and the UTC date-time given by OBS_DATE. The remaining columns constitute the significant data with DRY_BULB being the primary data since a row without this contains no information. All the other columns could be null, although ORIG_DRY_BULB should usually be present. Of course, the humidity data (WET_BULB, DEWPOINT and RELATIVE_HUMIDITY) are often present and if one of these is present then the other two and ORIG_WET_BULB should also be present. However, it was realised during the auditing process described in this report that at low temperatures both WET_BULB and DEWPOINT become poor means of conveying humidity measurements and the presence of all three humidity measures is now only required for temperatures of –10 °C or higher while at lower temperatures RELATIVE_HUMIDITY can be present by itself.

A full description of **SCREEN_OBS** 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 temperature and humidity observations.
   - A.2. OBS_DATE: Must not be later than the current date.
   - A.3. DRY_BULB_REL: Only NULL or "*" are allowed.
   - A.4. WET_BULB_REL: Only NULL or "*" are allowed.
   - A.5. ORIG_DRY_BULB: 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.6. ORIG_WET_BULB: The entries in this column should all represent valid origins.
   - A.7. DRY_BULB: Must be present.
   - A.8. RELATIVE_HUMIDITY: Can be NULL, but if present then must lie between 0 and 100.

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 **TEMP_HIS**.
   - B.2. OBS_DATE, ORIG_DRY_BULB (or ORIG_WET_BULB): For a given ORIG_DRY_BULB (or ORIG_WET_BULB), all observations should be at the correct times.
   - B.3. AGENT_NO, DRY_BULB: For a given place the DRY_BULB should be reasonable.
   - B.4. AGENT_NO, DRY_BULB, OBS_DATE: For a given place, time of day, and time of year the DRY_BULB should lie within a restricted range.
   - B.5. AGENT_NO, RELATIVE_HUMIDITY: For a given place the RELATIVE-_HUMIDITY should be reasonable.
   - B.6. AGENT_NO, RELATIVE_HUMIDITY, OBS_DATE: For a given place, time of day, and time of year the RELATIVE_HUMIDITY should lie within a restricted range.
   - B.7. DRY_BULB, RELATIVE_HUMIDITY: If RELATIVE_HUMIDITY is present then DRY_BULB must also be present.
   - B.8. WET_BULB, DEWPOINT, RELATIVE_HUMIDITY: If any one of these is present then the other two must also be present and consistent with each other and the particular value of DRY_BULB *unless DRY_BULB is lower than –10 °C in which case only RELATIVE_HUMIDITY should be present.*

C. Between row checks
   - C.1. For a given AGENT_NO, the DRY_BULB should not be too different from its value just before or just after its OBS_DATE.
   - C.2. For a given AGENT_NO, the RELATIVE_HUMIDITY should not be too different from its value just before or just after its OBS_DATE.
   - C.3. For a given OBS_DATE, the DRY_BULBs should not be too different for AGENT_NOs that are physically close to each other.
   - C.4. For a given OBS_DATE, the RELATIVE_HUMIDITYs should not be too different for AGENT_NOs that are physically close to each other.
   - C.5. For a given 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 time of the day, the length of record should be adequate.

D.2. For a given AGENT_NO, the rows that make up a complete local month should have associated rows in **MTHLY_STATS**.

The italicised part of B.8 was a condition added during the auditing and the check was initially performed ignoring the italicised part. 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.7 and B.8 whereas the other B checks and C.1, C.2, C.3, C.4, and D.2 will highlight those rows that might be in error. Remaining checks (C.5 and D.1) may uncover some errors, but it is more likely that any gaps in a record or any short records 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 B.3 it can only be said that the "DRY_BULB should be reasonable" and not that it should lie between certain limits because they vary greatly with the AGENT_NO.

Audit results

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

For any row in **SCREEN_OBS** 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 **SCREEN_OBS** as a code for the station, thus, all the AGENT_NOs in **SCREEN_OBS** 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 temperature or humidity. In the tabulation below this applied to "RAIN (STANDARD)", "REGIONAL COUNCIL", "WATER SCIENCES", and "ANEMOMETER ONLY", but 50 of these 58 stations had a considerable number of rows in **SCREEN_OBS** and only 8 had under 50 rows. 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 assumed to be the case for the 50 stations while the 8 with only a little data were checked and in all cases it was probable that the data had been misplaced and so it was deleted. The stations were B75253/1540, C84661/2156, E05542/3271, F12071/3832, I50111/5278, I50114/5281, I57761/5436, and J80800/6081 with a total of 130 rows being deleted.

RAIN (STANDARD)	47
CLIMAT (STANDARD)	391
CLIMAT/SYNOP	133
RAIN/SYNOP	37
CLIMAT (PRIVATE)	4
REGIONAL COUNCIL	4
WATER SCIENCES	1
ANEMOMETER ONLY	6
SYNOP ONLY	170
AWS (SYNOP AND METAR)	173
EDR	19
CLITEL	25
SPECIAL STATION	9

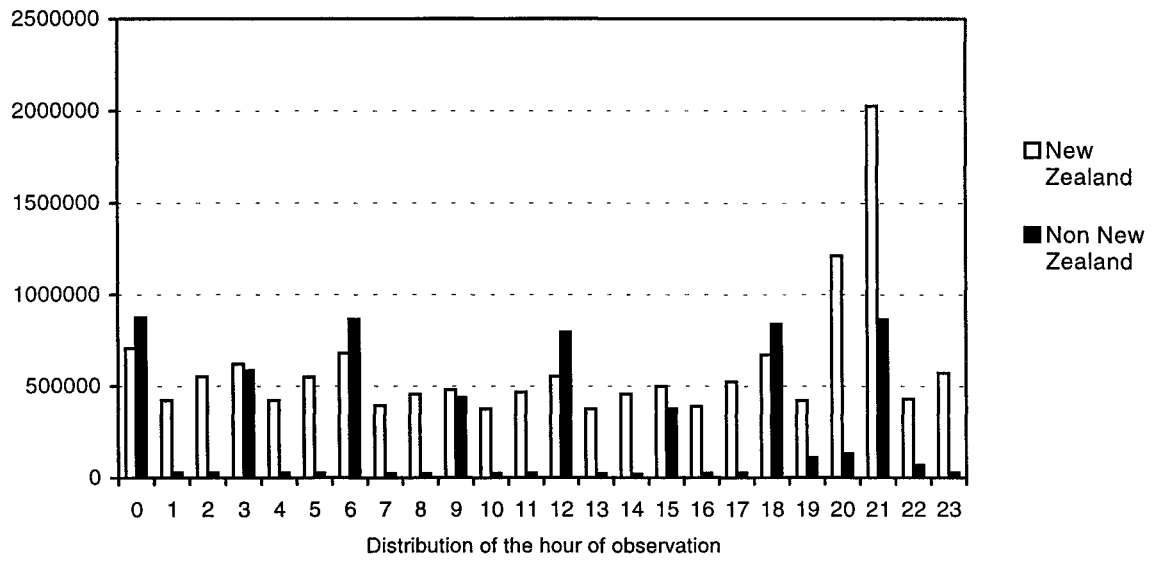
Details and results of Checks A.2 and B.2 — are all observation dates and times valid?

For any row in **SCREEN_OBS** it must be known at what date and time the observation was made and these dates should not be later than the current date. One observation was found with a date of 7 March 2104 at 2 a.m. for H31883/4764; it was deleted. 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_DRY_BULBs were started earlier than those of others, there is no fixed threshold either for the earliest date. However, the earliest dates for each ORIG_DRY_BULB can be found and, as can be seen below, these dates are reasonable with the overall earliest being July 1890. This was for daily climate reports from H32641/4881 and no data from any other stations had a date earlier than 1928. Data of other ORIG_DRY_BULBs start at dates that are appropriate to the source concerned.

ORIG_DRY_BULB	Earliest data
D (i.e., Daily Reports)	18900630
H (i.e., Hourly CLITEL Reports)	19940630
E (i.e., EDR Reports)	19840701
F (i.e., Auotographic Form Reports)	19711231
S (i.e., Synoptic Reports)	19481231
M (i.e., METAR Reports)	19591231

The times of observation are also constrained since: ORIG_DRY_BULB "D" rows should have a time equivalent to 0900 Local; ORIG_DRY_BULB "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_DRY_BULB 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_DRY_BULBs of "H", "E", "F", and "M" are probably correct, but errors were found in the times of other observations.

The distribution with hour was:



The larger number of observations at synoptic hours can be easily seen for non New Zealand stations but not for New Zealand ones where the effects of daylight saving and the hourly reports from automatic stations mask the contribution from synoptic observations. However, the larger number at the time of the daily climate report can be seen for New Zealand. Before checking that all rows with an ORIG_DRY_BULB of "S" were at synoptic reporting times and that all "D" rows were at 0900 Local, observations at possibly erroneous reporting times were found. These were where stations were not reporting every hour but had observations at non-synoptic times, although reports at 0900 Local were ignored.

Hour (Local for NZ else UTC)	First run	ORIG_DRY_BULB (second run)						Third run
		D	H	E	F	S	M	S
0100	781	0	2	23	1	708	47	225
0200	180	0	0	29	1	111	39	33
0400	818	0	0	24	6	760	28	269
0500	2 020	0	0	24	6	137	1 853	53
0700	4 476	0	0	26	9	905	3 536	252
0800	7 644	1	0	64	10	75	7 004	33
1000	9 743	0	1 716	159	32	685	7 151	273
1100	8 755	0	19	217	72	167	8 280	60
1300	4 466	0	13	172	28	729	3 524	278
1400	4 064	0	10	148	23	76	3 807	13
1600	3 754	0	7	105	8	722	2 912	212
1700	2 459	2	4	67	4	120	2 262	50
1900	1 094	3	1	24	1	614	448	237
2000	280	18	0	22	0	128	112	64
2200	454	5	0	27	0	332	90	119
2300	834	0	0	19	0	187	628	104
No of stations	483	13	13	12	13	149	72	132

In the first run the results were not stratified by ORIG_DRY_BULB but it was noticed that many New Zealand stations had one or two observations at 0800 Local and that only the dates 21 February 1975 and 5 March 1976 were concerned. These dates were the end of daylight saving for the years concerned suggesting that, for those years, the data had been processed as if daylight had lasted for one more day than it actually had. An hour was added to the observation time in 494 rows.

In the second run the results were stratified by ORIG_DRY_BULB and the following action taken.

D All 29 rows from the 13 stations had their observation time amended to be equivalent to 0900 Local.

H The 1716 rows at 1000 Local were mainly from just four New Zealand stations (A53986/16137, B75382/1551, D06433/12636, F12839/16826) and were from dates during periods of daylight saving. The stations had been organised to provide observations at 0900 NZST rather than at 0900 NZDT, so no action was taken.

E The automatic devices that provide these observations do occasionally fail and then a complete set of 24 rows for a day will not be available. No action taken.

F Similar to "E", some of the record is lost when the trace on the autographic chart is unreadable. No action taken.

S The dates of these synoptic observations nearly all fell between October 1991 and March 1993, i.e., from the time that CLIDB began until it was realised that part of the synoptic message which was supposed to contain the time of observation had been used to retain the time of

receipt at MetService. The program that extracts observations from the MetService database of synoptic observations had been amended in March 1993 to accredit observations to the right time, but the times of reports already received had not been amended. The times were now amended.

- For New Zealand stations with observations at 0700 Local and missing observations at 0900 Local, the time was amended to 0900 Local — 74 such rows were amended
- Three New Zealand stations with observations at 0700 Local had observations at 0900 Local with a "D" ORIG_DRY_BULB — 111 rows were deleted.
- Non New Zealand stations with observations at non-synoptic hours typically had large numbers of observations 1 h after a synoptic hour and smaller counts 1 h before. In both cases the observation time was amended to the nearest synoptic hour if the observation at that time was missing — 5111 rows were amended.

A third run was made at this point after a minor modification to the program which had assumed that the time of the daily climate observation was also a synoptic hour. The result of not assuming that is show in the tabulation above under "Third run".

- Amending the time to the nearest synoptic hour was repeated — 729 more rows were amended.
- The remaining 1546 rows were deleted.

M Only METARs from automatic stations provide 24 observations a day whereas those from airports which are staffed only during daylight hours often provide only 6–12 observations a day. It can be seen in the tabulation that it is mainly the daylight hours that are concerned. No action was taken.

The checks above did not explicitly find the instances where a "D" ORIG_DRY_BULB is not at 0900 Local or a "S" is not at a synoptic hour. An explicit check was made and 69 "D"s and 1329 "S"s at wrong times were found. Of these 2 "D"s and 436 "S"s had their observation times amended so that "holes" in the record were filled. The 67 remaining "D"s were all for days whose other observations had an "M" ORIG_DRY_BULB and, since a "D" row has better data than a "M" row, these were transferred to their correct times and the "hole" left behind filled with an estimate. The 893 remaining "S" rows were deleted.

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

For any row in **SCREEN_OBS** it ought to (but not *must*) be known what is the origin of the observation where "origin" relates to the message type with which observations are transferred from their point of measurement to the procedures that load them into CLIDB. It is possible that, for a particular row, the DRY_BULB value and the humidity data (WET_BULB, etc.) were received through different messages and so an origin column is available for both the dry and wet data. 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_DRY_BULB and ORIG_WET_BULB columns of **SCREEN_OBS**. This was found to hold.

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

The tabulation below shows what combinations of origins and reliabilities occurred and how frequent each combination was. A quality ranking applies to the different origins due to the amount of precision with which the particular message type conveys the temperature or humidity measurement and to the amount of quality control applied before being stored in **SCREEN_OBS**. The ranking is that the best data have an origin of "D" which indicates it was received with three significant figures and was checked for consistency with reading from maximum and minimum thermometers. The next best has origin "H", then "E", "F", "S" and finally "M" has the poorest quality since it is received as whole degrees and would have had little quality control.

Dry Origin	Wet Origin	Dry Reliability	Wet Reliability	Count	Ref
D	D	*	*	71	
D	D	*		298	
D	D		*	15	
D	D			2 755 489	
D	E			917	
D	M			2 398	
D	S	*		16	
D	S		*	48	
D	S			100 969	
D		*	*	2	1
D		*		21	
D				92 545	
E	D			9	2
E	E	*	*	2	
E	E	*		9	
E	E		*	2	
E	E			793 951	
E	M			35 815	
E	S			936	
E		*		10	
E				19 623	
F				652 909	
H	H			305 855	
H	S			1	
H				42 728	
M	D	*		2	2
M	M	*	*	6	
M	M	*		37	
M	M		*	1	
M	M			6 327 411	
M		*		6	
M				496 281	
S	D		*	1	2
S	D			5	2
S	M	*		1	
S	M			590	
S	S	*	*	1 066 219	
S	S	*		336	
S	S		*	1 229 401	
S	S			4 609 268	
S		*	*	1	1
S		*		902 673	
S			*	5	1
S				1 056 645	

Most of the combinations above are valid apart from those indicated under "Ref" where:

1. shows those occasions where a WET_BULB_REL of "*" occurred when ORIG_WET_BULB was null and so there was probably no humidity data in the rows concerned. It was confirmed that for these 8 rows no humidity data were present and the WET_BULB_REL for those rows was amended to null.
2. shows those occasions where ORIG_WET_BULB indicated that the humidity data had a better origin that the temperature. This condition was initially thought to be incorrect, but during the joint Check A.8, B.5, B.7, B.8, it became apparent that the condition could well be valid. However, it was found that these 17 rows were error cases because their humidity data had been added to an existing row with a DRY_BULB value without ORIG_DRY_BULB being amended to "D". Such amendments were made.

The table provides a measure of quality since it can be seen that, generally, only 0.1% at most of the rows for a given ORIG_DRY_BULB–ORIG_WET_BULB pair have reliabilities of "*". Indeed, for rows with an ORIG_DRY_BULB of "M" the rate is only 1 row in 100 000, but this reflects more that these observations are as-received and subject to little quality control rather than that they are more reliable. The exception is for data with an ORIG_DRY_ORIGIN of "S" where about 1 in 3 rows have a reliability of "*" due to much of the early synoptic data having been reported in whole degrees.

Details and results of Check B.1[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 thermometers from some of the stations are held in **TEMP_HIS**. Thus for each AGENT_NO the earliest and latest records within **SCREEN_OBS** 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 **TEMP_HIS** the data dates can be compared to the date when a thermometer 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**. This time only 58 date inconsistencies were found and these are tabulated below.

AGENT_NO	Station start	Thermo-meter start	Data start	Data end	Thermo-meter end	Station end
1024	19481201		19481130	19960617		19860101
1316	19341101		19610401	19831231		19800731
1333	19800901		19800810	19870325		19870430
1434	19620901		19611130	19901129		19901231
1691	19820601		19820630	19881030		19880331
1703	19770701		19770610	19930227		19940201
1820	19320901		19731031	19870629		19870331
1856	19751231		19711231	19990603		
1967	19630201		19950101	19980331		19980315

[1] The last single column checks (i.e., A.7 and A.8) are described below with the multiple column checks.

2005	19691001		19711231	19891129		19890901
2006	19860228	19860228	19860219	19990604		
2095	19820801		19820731	19840830		19840131
2198	19210701		19711231	19770429		19770331
2367	19860801		19860704	19980428		
2465	19650701		19950101	19970525		19970323
2692	19901031	19901008	19901107	19990603		
2871	19850501		19850131	19880529		19900101
3014	19820601		19781231	19890830		19890901
3015	19851231		19850430	19940522		19940522
3099	19870301		19860630	19891230		19891231
3142	19901031	19901011	19901108	19990603		
3445	19600901		19591231	19990603		
3551	19860501	19860401	19860410	19990604		
3716	19700801		19651129	19990603		
3814	19860201		19860121	19880413		19891231
3909	19631201		19611130	19990603		
4118	19841001		19840821	19881129		19911231
4231	19721001		19720918	19740730		19740731
4333	19491001		19711231	19860429		19860228
4458	19050101		19711231	19950330		19941231
4764	19470101	19470101	19711231	21040307		
4883	19750701		19750715	19780830		19751231
5123	19721201		19721207	19780416		19780331
5217	19900901		19880229	19940130		
5397	19621201		19621109	19990603		
5745	19871130		19871202	19930830		19901031
5998	19810801		19810721	19860227		19860228
6012	19720101		19591231	19990603		
6066	19470101		19490630	19890530		19800131
6078	19370331		19560101	19970610		19960901
6080	19411031		19560101	19970118		19960601
6095	19400630		19720101	19970115		19960601
6096	19460930		19570101	19961101		19960901
6104	19290531		19720101	19980630		19960701
6124	19131231		19851120	19961016		19960601
6125	19830630		19830614	19960531		19960531
6172	19490101		19410703	19950831		19950831
6194	19770101		19751231	19990603		
6222	19681101		19681130	19910127		19900131
6311	19801101		19800930	19860929		
7339	19911121	19911121	19911106	19990603		
7364	19920209		19920323	19941104		19940531
7450	19920428		19911231	19990603		
7519	19920604		19921204	19930824		19930814
9874	19940126		19931231	19990429		
10034	19940306	19940228	19940306	19981029	19981031	19981031
12740	19961129		19960121	19990604		
16826	19980501	19980501	19980428	19990603		

All these inconsistencies 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, occasionally, **TEMP_HIS**. In a few cases the start and end dates were correct and some rows of **SCREEN_OBS** for times outside those dates were deleted. Also some rows that had been put into **SITE_CHANGES** during the audit of **MTHLY_STATS** were no longer valid and were deleted.

The tabulation below summarises the changes that were made.

LAND_STATION start date altered	31
LAND_STATION end date altered	24
SITE_CHANGES dates altered	55
SITE_CHANGES warnings deleted	89
TEMP_HIS date altered	4
SCREEN_OBS rows deleted	29

Details and results of Checks A.7 and B.3 — are all temperatures valid and reasonable?

The essential datum in any row in **SCREEN_OBS** is the temperature in the DRY_BULB column, so it should be non-NULL. No rows with a NULL DRY_BULB were found. Also the temperature should always lie between some limits which could be arbitrarily set at –75 °C and 50 °C, but, since the extremely cold temperatures apply only to Antarctica, for places north of 60° S the lower limit can more sensibly be set at –20 °C. A total of 2398 rows with temperatures outside these limits were found and various cases arose.

- ORIG_DRY_BULB was "S" and a conversion from Fahrenheit to Celsius was required. This occurred for observations from J99700/6169 for October 1970 which was just before the time when all temperatures began to be reported in Celsius. The 117 rows concerned were amended by the usual Fahrenheit to Celsius conversion formula.
- ORIG_DRY_BULB was "S" and the temperature was less than –20 °C or over 50 °C at places north of 60° S. These all had dates after July 1978 which is when automatic storage of synop reports began and is well after the change over from Fahrenheit to Celsius. Thus, all rows with temperatures over 50 °C could be deleted, but some with values under –20 °C might be valid. There were only 12 such rows, with –23 °C being the lowest temperature, and through comparisons with observations adjacent in time to them they were all found to be incorrect. Those 12 rows and another 590 with temperatures above 50 °C were deleted.
- ORIG_DRY_BULB was "S" and the temperature was less than –75 °C or over 50 °C at places south of 60° S. Those south of 60° S all had dates after July 1978 which is when automatic storage of synop reports began. Through comparisons with observations adjacent in time, the rows with temperatures below –75 °C generally appeared to be correct. Those with temperatures above 50 °C were, of course, incorrect but could have resulted from misuse of the code used to transmit synoptic observations. This did not seem to be the case and 448 such rows were deleted.
- ORIG_DRY_BULB was "H" and the temperature was less than –20 °C. There were 10 such rows which were all for G22582/11234 on 6 and 7 December 1997 and some adjacent in time observations were also found to be in errors. A total of 28 rows were deleted.
- The remainder are tabulated below. The value for 6194, which is the South Pole station, was taken to be correct, but the other six were amended through temporal and spatial comparisons.

AGENT_NO	Data date/time	Dry bulb	Wet bulb	Dewpoint	RH	Dry origin	Wet origin
1534	19720229:2100	64.2	48.1	45.7	41.2	D	D
2282	19890114:0600	–29.0				M	
3629	19910601:2100	70.4	30.4	9.0	3.6	D	D
5450	19960820:0300	53.0	24.3	3.0	5.3	M	M
6194	19780916:2100	–77.5				D	
6352	19860816:2100	98.0	73.2	71.9	35.8	D	D
6529	19850320:2100	70.2	65.3	65.0	79.4	D	D

By chance, several values of 0 °C were noticed for some Pacific island stations. Perhaps setting the lower validity limit as –20 °C for such places had been too generous and the check was re-run for Pacific islands using 0 °C as the lower limit. The tabulation shows which stations had rows with a temperature of 0 °C or less, how many they each had and the dates of the first and last such temperatures. Only 2 of these temperatures were amended: the other 3092 were deleted.

AGENT_NO	First date/time	Last date/time	Number
5946	19990829:2100	19990829:2100	1
5953	19780728:0000	19811230:1200	1221
5954	19780727:1800	19880708:1800	1316
5971	19791226:1500	19811230:1800	546
6028	19790823:1200	19790823:1200	1
6081	19800625:0600	19800630:0600	2
6136	19800929:0300	19801228:1200	3
6139	19800928:2100	19800928:2100	1
11334	19981025:0300	19981027:0300	3

Also noticed by chance for some Pacific island stations were several values of 0 °C for DEWPOINT with associated DRY_BULB values of exactly 35 °C and 45 °C. Perhaps, 50 °C is also too generous and the check was re-run for Pacific island using 40 °C as the upper limit. The check found 2824 rows with DRY_BULB at least 40 °C, 23 stations were involved with 5 (J55000/7430, J57000/5953, J57400/5954, J59800/5971, J76700/11111) having most of them and the other 18 having only 37 such temperatures. The treatment of these 37 is tabulated below.

Action	Number
Rows deleted	21
Rows accepted	3
DRY_BULB lowered by 10 °C	13
Humidity also changed	2

For the others, their association with a DEWPOINT of 0 °C was exploited together with the observation that these rows seemed to have DRY_BULB values that were a multiple of 5. Counts at Pacific island stations of the occurrence of a DEWPOINT of 0 °C and a DRY_BULB of exactly 30 °C or 35 °C or 40 °C or 45 °C indicated that only J57000/5953, J57400/5954, and J59800/5971 were involved and 7199 rows were deleted. Presumably, the high temperatures at J55000/7430 and J76700/11111 are correct.

Since a lower limit of –20 °C had proved too low for the tropics it seemed more than likely that an upper limit of 50 °C was too high for Antarctica. The check was re-run for Antarctic stations using 10 °C as the upper limit and 6453 errors were found at 68 stations. Just 3 stations had a total of 4935 errors and, on the other hand, there were 40 stations all with no more than 9 errors each and with a total of 150 between them. Of these latter, 12 were accepted as correct, 114 were amended, and 24 were deleted.

The other errors found needed attention but, because the method used for the first 150 errors was time consuming, a scheme for quickly identifying temperatures that need deletion was developed. It produces a time-ordered listing of the whole record for a station in which the temperatures are displayed as a string of characters whose length depends on the temperature value. Thus a quick time series plot can be created which has sufficient detail to determine when a significant change in temperature takes place. A typical example is given below where a large but temporary drop of nearly 30 °C — each ✳ represents 2 °C — happens between 19650506 and 19650507. Later, on 19650525, a temporary increase of over 30 °C takes place to a temperature above zero — the "0" provides a benchmark for zero.

```
19650504:2100    |****|****|****
19650505:2100    |****|****|****|
19650506:2100    |****|****|****
19650507:2100    |*
19650508:2100    |****|****|****
19650509:2100    |****|****|**
19650510:2100    |****|****|**
19650511:2100    |****|****|**
19650512:2100    |****|****
19650513:2100    |****|
19650514:2100    |****|****|***
19650515:2100    |****|****|*
19650516:2100    |****|****
19650517:2100    |****|****|**
19650518:2100    |****|****|*
19650519:2100    |****|
19650520:2100    |****|****|
19650521:2100    |****|****|**
19650522:2100    |****|**
19650523:2100    |****|****|
19650524:2100    |****|****|**
19650525:2100    |****|****|****|****|****|****0**
19650526:2100    |****|****|***
19650527:2100    |****|****|****|****|***
19650528:2100    |****|****|****|****|*
19650529:2100    |****|****|****|****|*
19650530:2100    |****|****|****|**
19650531:2100    |****|****|****|*
19650601:2100    |****|****|****|*
19650602:2100    |****|****|**
19650603:2100    |****|****|***
19650604:2100    |****|****|***
19650605:2100    |****|****|**
19650606:2100    |****|****|****|
```

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 **SCREEN_OBS** should be deleted. A time series file was generated for each of the 28 Antarctic stations where many temperatures over 10 °C had been found. The tabulation below shows how many rows were deleted for each of those stations with a total of 13 569 rows being deleted overall.

AGENT_NO	Years of record	Number of rows	Number deleted	Number amended
6192	1978-1992	18404	77	0
6194	1976-1999	19987	50	2242
6196	1980-1999	14780	74	0
6198	1983-1986	818	4	0
6199	1981-1983	1247	30	0
6203	1979-1999	18338	68	0
6207	1982-1999	17152	53	0
6209	1978-1996	2420	44	0
6212	1978-1999	50319	67	0
6216	1978-1999	49302	63	0
6218	1978-1999	9359	123	0
6219	1978-1999	50155	24	0

6220	1985-1999	32252	38	0
6222	1968-1991	5013	0	5
6236	1978-1999	37300	95	0
9943	1993-1999	12805	330	0
9961	1991-1999	6159	35	0
9963	1993-1999	12556	81	0
11858	1994-1999	10949	574	0
11860	1994-1997	4158	604	0
11868	1994-1999	10790	683	1
11869	1994-1999	10392	92	1
11871	1994-1998	5702	40	0
11873	1994-1998	6388	4559	0
11877	1994-1998	6732	4065	0
15708	1994-1999	9476	529	2
15711	1995-1998	3156	1027	0
15712	1996-1999	740	134	0

The tabulation above also shows that some amendments were done, at L00900/6194 in particular. For this station most the errors occurred in the winters of 1978–81 inclusive when temperatures reported by synoptic observations ranged between –10 °C and 20 °C but those reported through the daily climate observation were –70 °C to –40 °C. At that time the synoptic reporting practice was to report negative temperatures as their absolute value plus 50 (e.g., –3 °C would be coded as 53), but how were temperatures below –50 °C reported? Whatever the answer, it was found that by changing DRY_BULB to (DRY_BULB × –1) – 50 values which agreed with the climate observations resulted.

Details and results of Checks A.8, B.5, B.7, and B.8 — are the humidity observations valid, reasonable, consistent, complete, and supported by temperatures?

Since the essential datum in any row in **SCREEN_OBS** is the DRY_BULB, the humidity data (WET_BULB, DEWPOINT, RELATIVE_HUMIDITY) should not be present without a non-NULL DRY_BULB. No rows with a NULL DRY_BULB were found so all humidity data that exist are supported by temperatures. Furthermore, using only the non-italicised text of B.8, if any one of WET_BULB, DEWPOINT, RELATIVE_HUMIDITY is present then the other two must also be present and all three must be consistent with each other and with the value of DRY_BULB. Thus, provided they are consistent, it is only necessary to check that one of WET_BULB, DEWPOINT, RELATIVE_HUMIDITY has valid and reasonable values; the easiest choice is RELATIVE_HUMIDITY which mostly lies between 0 and 100, but values between 0 and 5 are probably too small. (A further check could be that DRY_BULB is never less than WET_BULB or DEWPOINT, but it would be equivalent to the consistency and validity checks together).

A combined check for WET_BULB, DEWPOINT, and RELATIVE_HUMIDITY all being present with the latter between 5 and 100 resulted in 27 171 rows being picked as potentially in error. Of these only 326 had neither ORIG_DRY_BULB nor ORIG_WET_BULB of "S" and were dealt with as follows.

Action	Number	Comment
Humidity section deleted	140	
Humidity section recalculated	51	Most lacked relative humidity
WET_BULB amended	81	
DRY_BULB amended	31	Most were 10 °C in error
Nil action	23	Acceptable low values from L66300/6222

Of the remaining 26 845:

- 25 706 had negative values of RELATIVE_HUMIDITY;
- 648 were from New Zealand stations;
- the earliest was from March 1963 but only 40 before July 1978 which is when automatic storage of synop reports began;
- the latest was from about the time that the check was made with 471 after October 1991 which was when CLIDB was established with a new scheme for retrieving **SCREEN_OBS** data from synoptic observations.

Many of these could probably have been corrected, but considering the number involved and that most of them were for stations outside New Zealand, the simple option of deleting the humidity sections of the 26 845 rows was taken.

Consistency between the three measures of humidity was checked by recalculating the dewpoint and relative humidity for each row from WET_BULB and DRY_BULB and comparing the result to the values in DEWPOINT and RELATIVE_HUMIDITY. An inconsistency was noted if either the absolute difference between DEWPOINT and its recalculation was over 0.2 or that for RELATIVE_HUMIDITY was over 1. A total of 272 772 inconsistencies was found and the tabulation below shows how these are divided between areas (NZ is New Zealand, ANTC is Antarctica, REST is mainly Pacific Islands), ORIG_WET_BULB (Origin) and degree of inconsistency (i.e., 1C5RH refers to those rows where either the discrepancy between DEWPOINT and its recalculation was between 0.5 °C and 1.4 °C or that for RELATIVE_HUMIDITY was 2.5% to 7.4%, etc, but 9C45RH catches all inconsistencies over 8.5 °C and 42.5%).

AREA	Origin	1C 5RH	2C 10RH	3C 15RH	4C 20RH	5C 25RH	6C 30RH	7C 35RH	8C 40RH	9C 45RH	TOTAL
NZ	D	9 460	14	0	0	0	0	0	0	0	9 474
NZ	H	3 421	77	1	0	0	0	0	0	20	3 519
NZ	E	25	22	35	27	30	19	9	4	4	175
NZ	S	5 356	25	13	12	8	5	0	5	113	5 537
NZ	M	3 900	73	18	7	2	0	0	0	935	4 935
ANTC	D	1	0	0	0	0	0	0	0	0	1
ANTC	H	13 596	3 323	286	33	3	0	0	0	0	17 243
ANTC	S	199 809	23 815	4 471	1 847	981	179	145	26	16	231 289
REST	S	597	0	1	0	0	0	0	0	0	598
REST	M	1	0	0	0	0	0	0	0	0	1

Most of the inconsistencies were for Antarctic stations and most of those were under 1.4 °C or 7.4%, but significant numbers occurred for larger discrepancies and for New Zealand stations. A number of sources for these inconsistencies were found.

1. Data for **SCREEN_OBS** are transferred into CLIDB through several different archiving procedures each using a different message format as its input. The procedures all call **HUMIDITY** (a procedure stored within CLIDB) and it requires the height of the station for which it is completing the suite of humidity measures from the temperature and one of WET_BULB, DEWPOINT, or RELATIVE_HUMIDITY. However, sometimes the station's height is not known and some of the procedures dealt with this by assuming a height of zero while others passed a NULL height to **HUMIDITY** which resulted in values corresponding to a station with a height of over 1000 m. **HUMIDITY** was modified to assume a height of zero when no height is available.

2. The ranking for ORIG_DRY_ORIGIN and ORIG_WET_ORIGIN described in the section on the validity of reliabilities and origins is used to decide whether some incoming data for **SCREEN_OBS** should overwrite data already in CLIDB. A row could be created in

SCREEN_OBS with just temperature but no humidity data, and later some data with a lower ranked origin become available so that the humidity part of the row could be "filled" in. However, in this situation most of the archiving procedures ignore the already stored DRY_BULB and use the incoming temperature in **HUMIDITY** so inconsistencies can arise when DRY_BULB and the incoming temperature differ. If the difference is significant then doubt is thrown on all the data, but sometimes the difference is due only to the different precisions of the data sources. A general procedure **WRITE_SCREEN_OBS** was introduced to be used by all archiving procedures; it deals with the problem concerned here and with some others and will be described shortly.

3. For low temperatures both WET_BULB and DEWPOINT become poor means of conveying humidity measurements. They lose physical realism and are not generally used: instead many Antarctic stations report relative humidity directly. If such RELATIVE_HUMIDITY values are used to calculate WET_BULB and DEWPOINT the degree of precision to which they are held in CLIDB is not enough for a subsequent calculation from WET_BULB or DEWPOINT to yield the original value of RELATIVE_HUMIDITY. Such calculations do occasionally take place and to preserve the original values it was decided that for rows with a DRY_BULB less than –10 °C only RELATIVE_HUMIDITY would be stored and WET_BULB and DEWPOINT would always be NULL. **HUMIDITY** was modified such that when called by **WRITE-_SCREEN_OBS** all subsequent archivals in **SCREEN_OBS** would obey this new rule. Also WET_BULB and DEWPOINT were removed from 297 419 rows with 195 010 inconsistencies being resolved. The program that calculates **MTHLY_STATS** code 16 (mean vapour pressure) was also amended as was that which performs Check B.8 of this **SCREEN_OBS** audit.

The new procedure **WRITE_SCREEN_OBS** was required for incorporation into the following archiving procedures RMSDYCLI, RMSEDR, RMSHOURLY, RMSMETAR, RMSSHPSY, RMSSYNOP, RMKS_DECODE, COPIDYCL, and DEUPDATE. Originally, in these procedures if an incoming temperature was not valid but the particular humidity measure was valid and a matching temperature was available from CLIDB, then only DEUPDATE allowed the humidity data to be stored. **WRITE_SCREEN_OBS** was written to perform like DEUPDATE and has the following specification:

- exit if no incoming temperature data
- get any existing data from CLIDB for the incoming place and time
- exit if both temperature and humidity data existing in CLIDB are better than the incoming data
- test incoming data against limits in **RANGES** (except relative humidity which should be between 0 and 100)
- exit if incoming humidity data invalid and any existing CLIDB temperature data better than the incoming temperature data
- value for DRY_BULB should be no less than either the value for WET_BULB or that for DEW-POINT, except if either is only 0.5 °C bigger then reset DRY_BULB to that value (this was introduced to allow data of different precisions to be archived in the same row)
- call **HUMIDITY**
- insert or amend a row in CLIDB.

After removing inconsistencies for rows where DRY_BULB was lower than –10 °C, 77 762 remained and **WRITE_SCREEN_OBS** was used to remove them. For each candidate row DRY_BULB and ORIG_DRY_BULB were retained and one of WET_BULB, DEWPOINT, or RELATIVE_HUMIDITY was retained according to ORIG_WET_BULB, i.e., WET_BULB if ORIG_WET_BULB was "D" or "E", DEWPOINT if "M" or "S", and RELATIVE_HUMIDITY if "H". The retained one was used to calculate the other two. Afterwards only 65 inconsistencies remained which were deal with as tabulated below

Action	Number
Humidity section deleted	50
WET_BULB amended	8
DRY_BULB amended	6
Nil action	1

The suite of checks was run again but with the full text of B.8 being used. The combined check for WET_BULB, DEWPOINT, and RELATIVE_HUMIDITY all being present when DRY_BULB is higher than –10 °C and only the latter for lower temperatures, and with the latter between 5 and 100 resulted in 43 rows being picked as potentially in error. They were dealt with as follows

Action	Number	Comment
Humidity section deleted	15	
Humidity section recalculated	1	It lacked relative humidity
DRY_BULB amended	2	
DRY_BULB and WET_BULB amended	1	
Nil action	24	Acceptable low values from L66300/6222 & I49592/5212

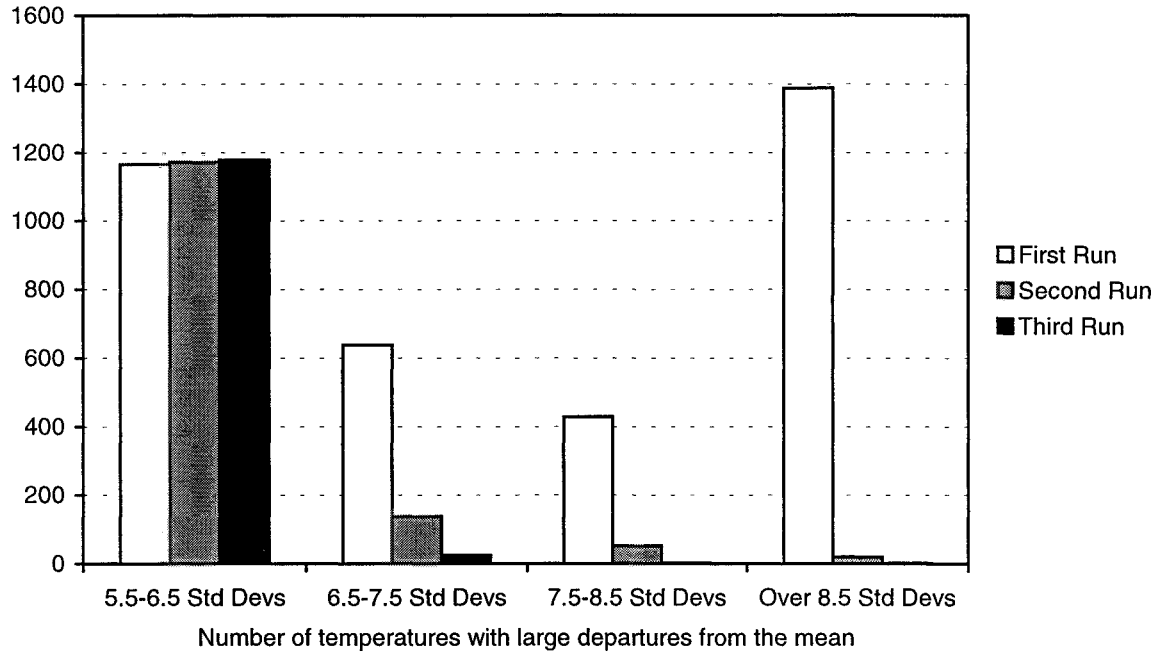
The repeat of the consistency check gave the results tabulated below in the same format as was used above. There are still a considerable number in the smallest difference class, especially for Antarctica where discrepancies arise due to the standard practice of always estimating the vapour pressure as if it were over water rather than ice. For example, an observation which was received with a temperature of –4 °C and a relative humidity of 100% would be stored in **SCREEN_OBS** with WET_BULB and DEWPOINT as –4 °C and –4.5 °C respectively. In this check the DRY_BULB and WET_BULB would have been used to calculate DEWPOINT as –4.5 °C but RELATIVE_HUMIDITY as 96.2%, i.e., an inconsistency of 3.8%. Furthermore, once observations have been entered into CLIDB the fact that one was received with a relative humidity rather than a wet bulb reading is lost. Thus, it can only be supposed that most of the small inconsistencies arose in this way. In a similar way, relative humidity measurements are received from many New Zealand automatic weather stations as an additional remark while the main coded hourly message contains a dew point. Thus, as before, once in CLIDB the fact that the humidity arrived as a relative humidity rather than as either a wet bulb or a dew point is lost and inconsistencies can arise.

Area	Origin	1C 5RH	2C 10RH	3C 15RH	4C 20RH	5C 25RH	6C 30RH	7C 35RH	8C 40RH	9C 45RH	Total
NZ	H	3 479	81	0	0	0	0	0	0	0	3 560
NZ	S	2 444	4	0	0	0	0	0	0	0	2 448
NZ	M	1 569	1	0	0	0	0	0	0	0	1 570
ANTC	H	1 291	0	0	0	0	0	0	0	0	1 291
ANTC	S	48 462	354	1	0	0	0	0	0	1	48 818
REST	S	593	0	0	0	0	0	0	0	0	593
REST	M	1	0	0	0	0	0	0	0	0	1

There is a total of 58 281 inconsistencies in the tabulation above but only 417 were new cases which had not been picked up in the first run. They were all successfully dealt with by using **WRITE_SCREEN_OBS** as before. Then the 412 individual cases with discrepancies in the 2C10RH, 3C15RH, and 9C45RH classes were examined. Most of these were accepted as valid, but three DRY_BULB values were amended and one WET_BULB.

Details and results of Check B.4 — are all temperatures reasonable for the time of day and time of year?

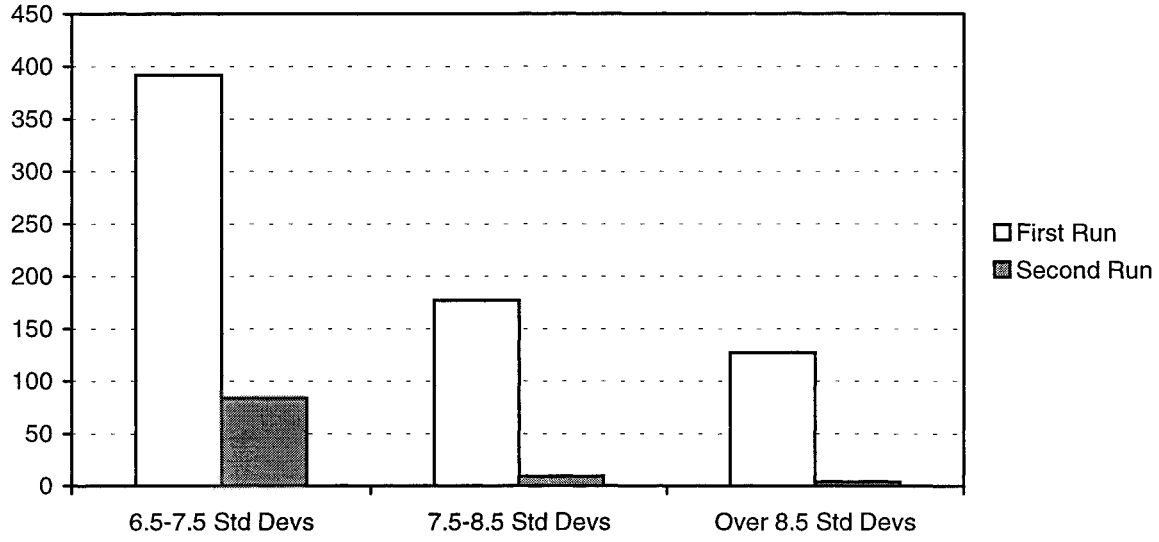
Temperature depends heavily on locality, time of day, and time of year. Some allowance for locality was made when checks A.7 and B.3 were performed, i.e., observations were classed as originating from either New Zealand, or Antarctica, or elsewhere. In this check each station will be considered individually with every temperature observation being compared to the mean at that station for the particular month of the year and hour of the day for the observation. The comparison will be made through the standard deviation at that station for the particular month of the year and hour of the day. If the temperatures were normally distributed, then the properties of the normal distribution suggest that out of a sample of 20 million — there are 20 million rows in **SCREEN_OBS** — about 9000 may have a departure from the mean of between 3.5 and 4.5 standard deviations, but only 1 with a greater departure. Thus, allowing a one standard deviation margin, all departures of six or more are likely to be errors.



The figure above shows the distribution of observations which had values which were more than 5.5 standard deviations from the mean for the particular station and month and hour of the observation. Corrections were made and a second determination of the means and standard deviations made, since some may well have changed due to the corrections made, then the large departures were found again. More corrections were made, the means, etc. re-determined and a few more corrections made. For each class and run at least 95% of the identified observations were of synoptic origin. Those of other origins were dealt with individually 26 being deleted, 101 amended, and 32 accepted of which nearly all were in the first class. Because many of the non-synoptic observations with departures from the mean of about 6 standard deviations proved acceptable, the synoptic ones in that class were also accepted. It may well be that many were in error, but the large number involved precluded individual treatment and, also because of the large numbers, synoptic observations with larger departures were just deleted without being inspected; altogether 2622 rows of synoptic origin were deleted.

Details and results of Check B.6 — are all relative humidities reasonable for the time of day and time of year?

Humidity, just like temperature, also depends heavily on locality, time of day, and time of year. In this check each station will be considered individually with every relative humidity observation being compared to the mean at that station for the particular month of the year and hour of the day for the observation. As for temperature, the comparison was made through the standard deviation at that station for the particular month of the year and hour of the day and, as before, only one observation with a departure greater than 4.5 standard deviations might be expected. Thus, allowing a one standard deviation margin, all departures of six or more are likely to be errors, but from the experience of examining temperature extremes only departures greater than 6.5 standard deviations were dealt with since many smaller ones had proved to be acceptable.



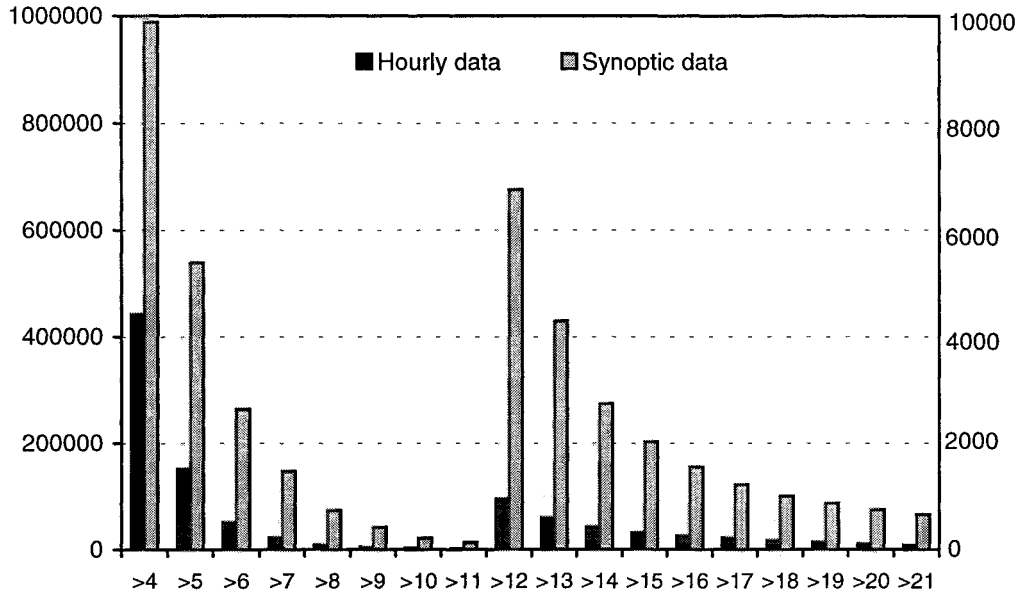
The figure above shows the distribution of observations which had values which were more than 6.5 standard deviations from the mean for the particular station and month and hour of the observation. The observations were sorted into two groups: firstly, all New Zealand observations together with those from elsewhere that were not of synoptic origin: secondly, the remainder. The actions tabulated below were taken.

	First run		Second run	
	Group 1	Group 2	Group 1	Group 2
Rows accepted	36	—	40	26
Rows amended	237	—	15	12
Humidity section deleted	7	410	—	2

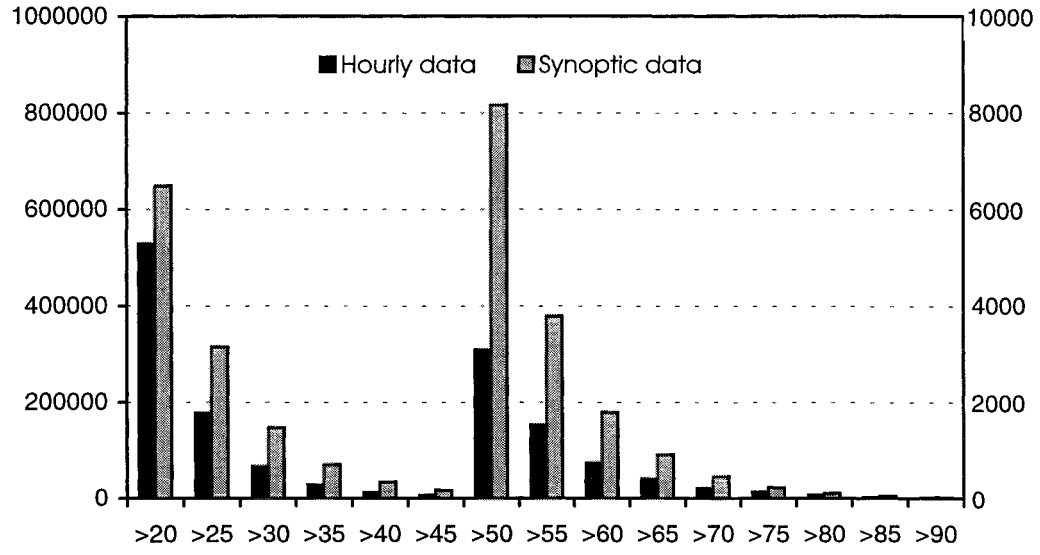
Details and results of Checks C.1 and C.2 — are there any excessive changes in the either the temperature or the humidity time series?

The temperature and humidity observations at each particular AGENT_NO form time series which sample the actual continuous variations of temperature and humidity at each point. There are some circumstances when temperature and/or humidity change 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. The figures below show that out of a total of over 20 million rows in **SCREEN_OBS** under 1.5 million temperatures are over 4 °C different

from the observation either 1 h before or after or 3 h before or after and about 1 million relative humidities are over 20% different. There is a change in the vertical scale from the >12 °C and >50% classes so that the trends in the upper tail can still be seen despite the small numbers, which in the last class are only 84 and 660 for hourly and synoptic temperatures respectively and 8 and 20 for the humidities.



Cumulative distribution of temperature changes (°C) between hours or synops



Cumulative distribution of humidity changes (%) between hours or synops

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. The levels at which the class memberships fell below 10 000 were found, then the AGENT_NOs and OBS_DATEs for all such occurrences were found with those qualifying through both temperature

and humidity being accredited to temperature. 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 13 115 blips and steps, taking the 5% of each with the largest differences gave the number of different stations involved as tabulated below.

	Temperature	Humidity	Total number of rows
Blip	19	25	96
Step	78	70	369

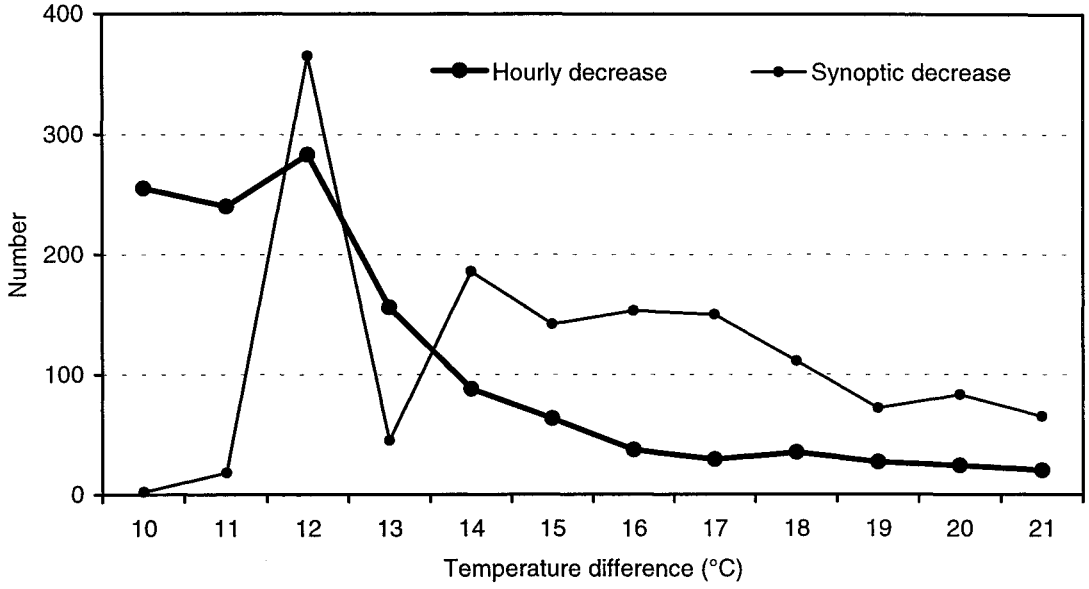
For each of the stations a listing was created in the style used for the A.7 and B.3 check (i.e., the temperature or humidity rows were displayed in time-order with their values indicated as a number of ✳s). The listing was started at a date a little before the first occurrence of a large difference and continued to a little after the last one with any large differences marked. The listings 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 of the temperature or relative humidity added, in which case the relevant row in **SCREEN_OBS** was amended rather than deleted. All changes to **SCREEN_OBS** were noted in **AUD_SCREEN_OBS**, and the numbers involved are tabulated below.

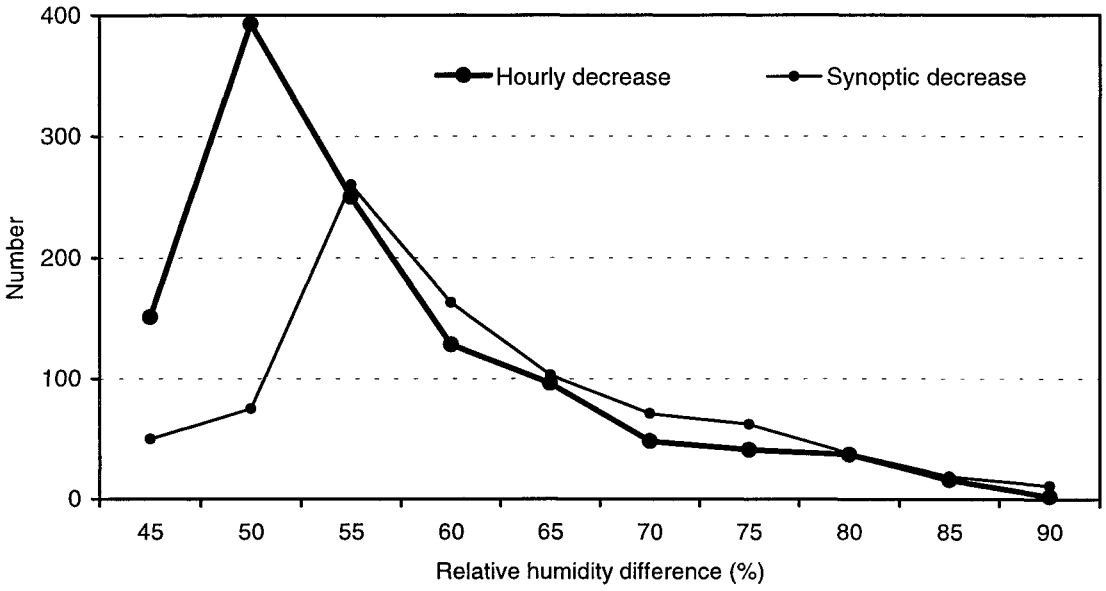
	Action	New Zealand	Elsewhere	Total
Blip	Deletions	—	—	411
Blip	Amendments	—	—	178
Temperature step	Deletions	168	16	184
Temperature step	Amendments	1005	259	1264
Humidity step	Humidity removed	154	50	204
Humidity step	Amendments	836	126	962

During this processing it was noticed that some stations showed periods, sometimes long, with suspect data. For example, temperatures of exactly and only 10 °C, 15 °C, 20 °C, 25 °C, etc for extended periods seem unlikely to be genuine. For some automatic stations it appeared that the temperature sensor was not responding with the temperature staying at some fixed value — or changing only slightly — over extended periods. The stations involved and the numbers of rows deleted are tabulated below

Station	Start	End	Number deleted
B86331/1796	Mar 1980	Dec 1981	53
D15245/2649	Feb 1979	Dec 1981	17
D87812/2833	Aug 1978	Sep 1991	52
F12211/3844	Aug 1978	Oct 1987	118
J55000/7430	Sep 1996	May 1997	742
J57000/5953	Jul 1978	Dec 1981	617
J57400/5954	Aug 1978	Dec 1981	1260
L81100/11874	Jan 1994	Apr 1996	3692
L81300/11876	Dec 1994	Apr 1996	3466

The distributions of temperature and relative humidity changes from hour-to-hour and synop-to-synop were redetermined and the figures below show the change in numbers of the various classes. In all cases the class-membership has decreased and the differences decrease more rapidly than they did before this check was made.





Details and results of Checks C.3 and C.4 — are all temperatures and humidities, 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 temperature 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 31 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 temperature record. At the other extreme for that hour there were 127 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 121 such stations for UTC hour 00.

			Distance (km) of the most distant buddy										
Hr	N.Z.?	Cover(%)	<5	5–15	15–25	25–35	35–45	45–55	55–65	65–75	75–85	85–95	>95
00	Y	≥95	31	11	19	33	30	24	23	19	10	8	9
00	Y	<95	9	.	3	2	2	5	3	.	.	1	24
00	N	≥95	14	6	10	2	7	3	2	2	.	2	93
00	N	<95	1	1	1	.	.	.	.	.	2	1	127
03	Y	≥95	25	10	19	27	33	20	27	10	6	9	7
03	Y	<95	7	1	3	1	2	2	2	2	.	.	23
03	N	≥95	10	1	7	1	4	3	1	2	.	2	74
03	N	<95	2	2	.	.	.	.	.	.	2	1	140
06	Y	≥95	25	12	19	25	26	13	20	14	9	9	11
06	Y	<95	8	.	2	1	2	3	1	.	.	1	20
06	N	≥95	13	7	10	3	7	2	3	3	.	1	101
06	N	<95	2	1	.	.	.	.	.	.	.	1	118
09	Y	≥95	18	9	11	10	17	16	19	9	2	9	15
09	Y	<95	6	.	3	.	1	.	1	1	1	.	29
09	N	≥95	7	.	7	3	4	4	1	3	.	3	54
09	N	<95	2	.	1	.	.	.	.	.	.	.	156
12	Y	≥95	14	9	13	10	21	17	18	11	4	10	11
12	Y	<95	5	1	4	.	3	2	1	.	1	.	30
12	N	≥95	12	7	11	4	6	3	2	1	.	2	97
12	N	<95	2	.	.	.	.	.	.	.	.	.	126
15	Y	≥95	12	8	12	6	18	18	22	10	3	7	10
15	Y	<95	1	.	3	.	2	2	1	.	.	3	27
15	N	≥95	6	1	3	3	3	2	3	1	.	3	52
15	N	<95	4	1	1	.	.	.	.	.	2	1	161
18	Y	≥95	27	10	21	24	30	20	18	21	11	9	13
18	Y	<95	5	1	3	.	.	3	2	2	.	1	23
18	N	≥95	13	7	9	2	7	3	1	2	.	2	89
18	N	<95	2	1	1	.	.	.	.	.	2	1	129
21	Y	≥95	99	176	143	87	62	13	3	4	4	2	3
21	Y	<95	9	18	10	4	4	1	1	1	1	.	14
21	N	≥95	23	15	20	8	6	6	3	2	3	4	95
21	N	<95	6	1	1	.	.	.	.	.	.	1	114

The tabulation above 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 45 km away, and only 14 were in the worst distance-cover class. For other hours in New Zealand, most buddies lay 40–50 km from the primary but a fairly even frequency of 15–20 primaries with buddies within each 10 km distance class existed out to about 70 km; also about 25 primaries had buddies in the worst distance-cover class. 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 third 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	540	121	419	126	23	6	0	19	95	100	0	105	658
03	488	130	358	101	19	5	2	11	94	100	0	113	658
06	493	112	381	118	21	4	0	23	96	100	0	116	658
09	422	138	284	83	15	3	0	14	91	100	0	138	617
12	458	127	331	95	18	2	0	28	95	100	0	140	658
15	412	144	268	89	16	4	0	26	91	100	0	143	617
18	517	127	390	117	24	4	0	21	95	100	0	114	658
21	967	105	862	322	92	12	3	12	98	100	0	55	658

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

Hour	Number of humidity primary-buddy pairs	Number of temperature primary-buddy pairs
00	224	256
03	194	227
06	200	232
09	144	195
12	180	208
15	134	187
18	213	247
21	510	549

From each of these hour-data-type classes the 5% of primary-buddy pairs with the largest ratios were examined since those observations were potentially the most likely to be errors. These 292 distinct occasions were examined by listing out from CLIDB for the station and time concerned the temperature and humidity 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 could be decided if observations were consistent with those nearby in space and time, or an amended value could be estimated, or it could be that the value needed to be removed. The consequence changes made to **SCREEN_OBS** are tabulated below. The table includes some additional changes which were made after a trial run of the procedure which finds the largest ratios.

Hour	Accepted	Amended	Humidity removed	Deleted
00	14	23	0	7
03	11	7	4	14
06	14	20	1	2
09	14	18	1	0
12	15	19	0	1
15	15	13	0	0
18	14	24	1	4
21	30	75	1	1
Total	127	199	6	29

Apart from these individual changes some "patches" of suspect data were found from the listings. The treatment of these is tabulated below.

AGENT_NO		Action	Dates	Comment
2496	5	Rows deleted	May 1982	Isolated reports
2807	14	Rows deleted	Aug 1990	Temperature "stuck" at –2.0°C
2980	7	Humidities removed	Nov 1998	High values
3147	39	Humidities removed	30 Nov–1 Dec 1986	Wet-bulb wick dry
3147	13	Rows deleted	Mar 1987	Inconsistent temperatures
4141	9	Rows deleted	15–17 Jul 1985	0.0°C and 100%
4141	144	Rows deleted	Aug–Sep 1985	0.0°C and 100%
4394	61	Rows deleted	Jul 1978–Dec 1982	Only a few observations/month
4396	70	Rows deleted	Jun 1996	Suspect negative temperatures
4780	5	Rows deleted	Apr–May 1979	Only 5 observations in 2 months
4997	13	Rows deleted	Feb 1993	Inconsistent temperatures
5093	6	Rows deleted	12–16 Jan 1996	0.0°C and 100%
5397	16	Rows deleted	Apr 1993–Dec 1994	Only a few observations/month
5971	95	Rows deleted	Nov 1978–Nov 1981	Only 5°C,10°C,15°C etc
6156	10	Humidities removed	Nov 1999	Low values
7339	45	Humidities removed	6 Nov–10 Nov 1991	RH always 33%

Because some inspected values were correct, if this checking procedure were to be performed again, then they would reappear but need not be re-examined for error. Thus, those that did not require correction must be remembered from one auditing to the next and this can be done through **SCROBS_DIFFS** which was created by this checking procedure and has the following structure.

Column name	Null?	Type
TYPE		CHAR(1)
HR		VARCHAR2(2)
AGENT_NO		NUMBER
BUDDY		NUMBER
DIST		NUMBER
OBS_DATE	NOT NULL	DATE
PERC		NUMBER
P_ VALUE		NUMBER
B_ VALUE		NUMBER

For each TYPE, HR, AGENT_NO, and BUDDY the values with the greatest difference occurred at OBS_DATE and are held in P_ VALUE and B_ VALUE, 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 TYPE, HR, AGENT_NO, BUDDY combinations which have the greatest relative difference. Thus, on a re-run the contents of **SCROBS _DIFFS** can be moved to **OLD_ SCROBS _DIFFS**, say, before being over-written and rows common to both tables (except PERC which may change between runs) can be ignored. However, because many of the differences were acceptable, a re-run was not made.

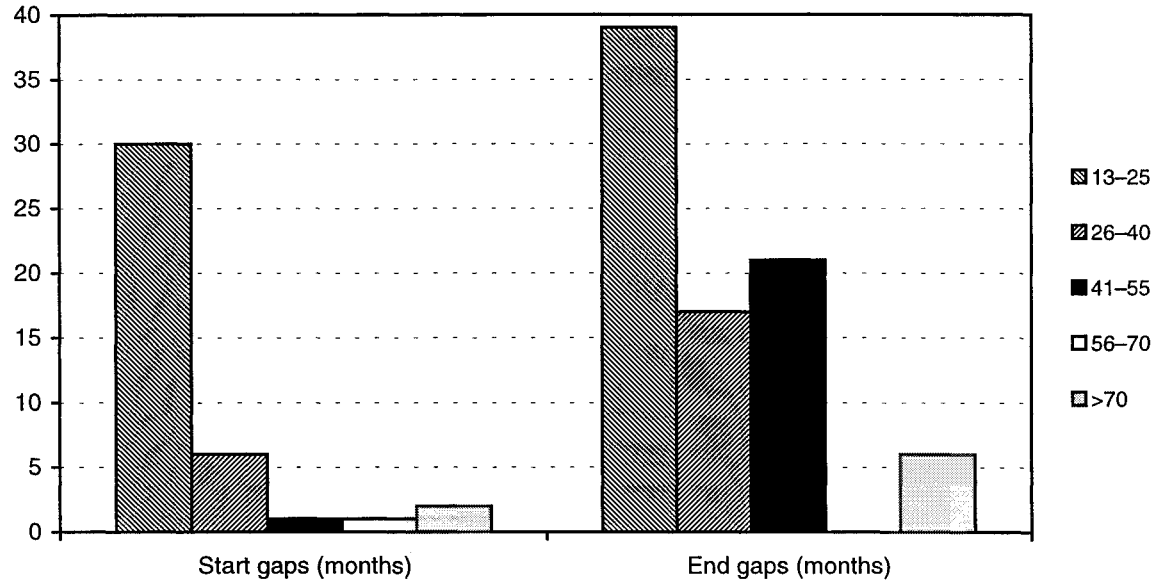
Details and results of Check C.5 — are all temperature records without gaps?

Both temperature and humidity are continuous in time but, apart from the traces on thermographs and hygrographs, their 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. Thus, 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.

There were stations where the completeness was small 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.

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 309 shorter gaps at the start of records and 340 at the end bringing the total number of such gaps to 342 and 418 respectively.



Distributions of gaps at the ends of station records

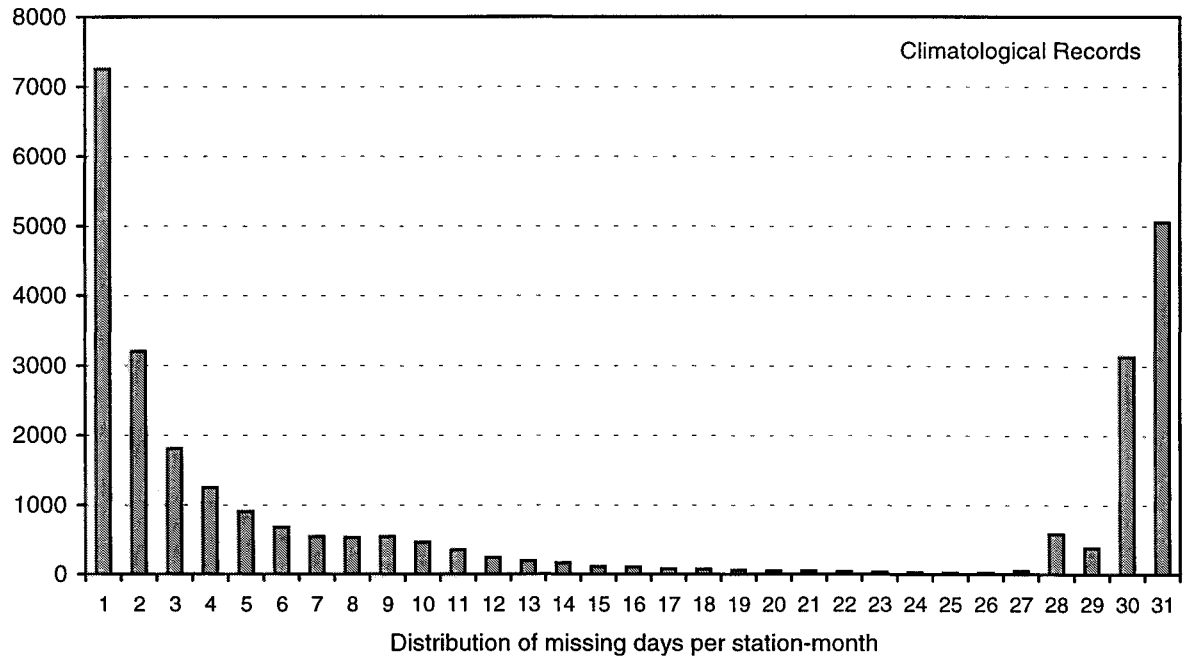
Of the 40 stations with "Start gaps", nearly all were due to a single early report of synoptic origin, although for L77400/11869 there were 92 reports before a 13 month gap to the start of the main record. A total of 125 rows was deleted. However, there were also two cases where a whole month of data was concerned.

1. I50721/5336 for January 1972 was followed by a 58 month gap. The station had opened in January 1967 with autographic instruments from which only January 1972 had had data extracted. The data were accepted.

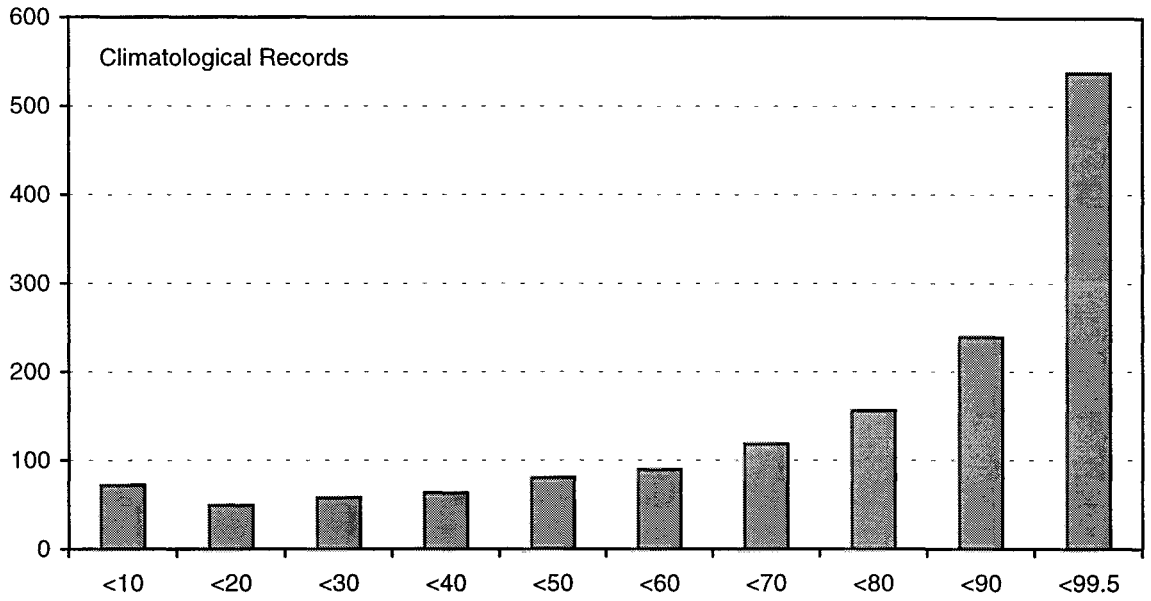
2. I49613/5217 for March 1988 was followed by a 29 month gap. The station had opened in September 1990 and a check on a previous station in its area (I49711/5223) showed that the data for March 1988 were identical at both stations. Thus the data had been incorrectly assigned to I49613/5217 and the 25 rows concerned were removed from **SCREEN_OBS** and 25 rows were also removed from each of **CLOUD_SYSTEM**, **MAX_MIN_TEMP**, **SURFACE_WIND**, **WEATHER**, and **WEATHER_PHEN**. Another 21 rows, which were dependent on those just deleted, were deleted from **MTHLY_STATS** and from **SITE_CHANGES** the associated warnings were deleted.

Of the 85 stations with "End gaps", nearly all were due to a few late reports of synoptic origin and 96 rows were deleted. However, there were two cases for Antarctica (L76800/15709 and L87300/11881) with sets of observations for June 1999. These were left as it was felt these may have come from special parties visiting the sites.

The remainder of this section is a description of the state of the climatological, synoptic, and hourly records after the changes described above had been made. The figure below is for climatological observations and shows the distribution of missing days per station-month. There were about 7200 station-months that had a single day missing, about 3200 with two days, which could be either together or apart, etc. These numbers are a significant fraction of the total number of climatological observations, which is equivalent to 86 000 station-months of which 28 000 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 9000 whole station-months are missing.

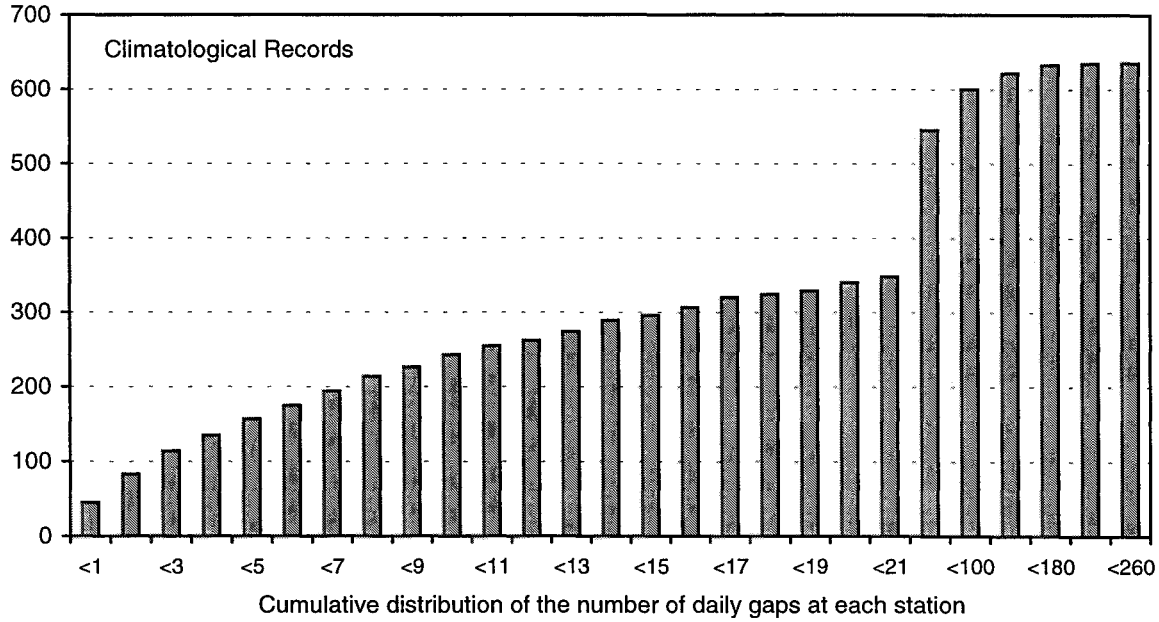


How are these missing days spread among the stations? CLIDB has 636 stations with records of climatological observations and 98 of these have near perfect records. The cumulative distribution of the percentage complete of the records at the other 538 stations is shown below and shows that as many as 250 of these have records less than 90% complete and about 80 records are under 10% complete.



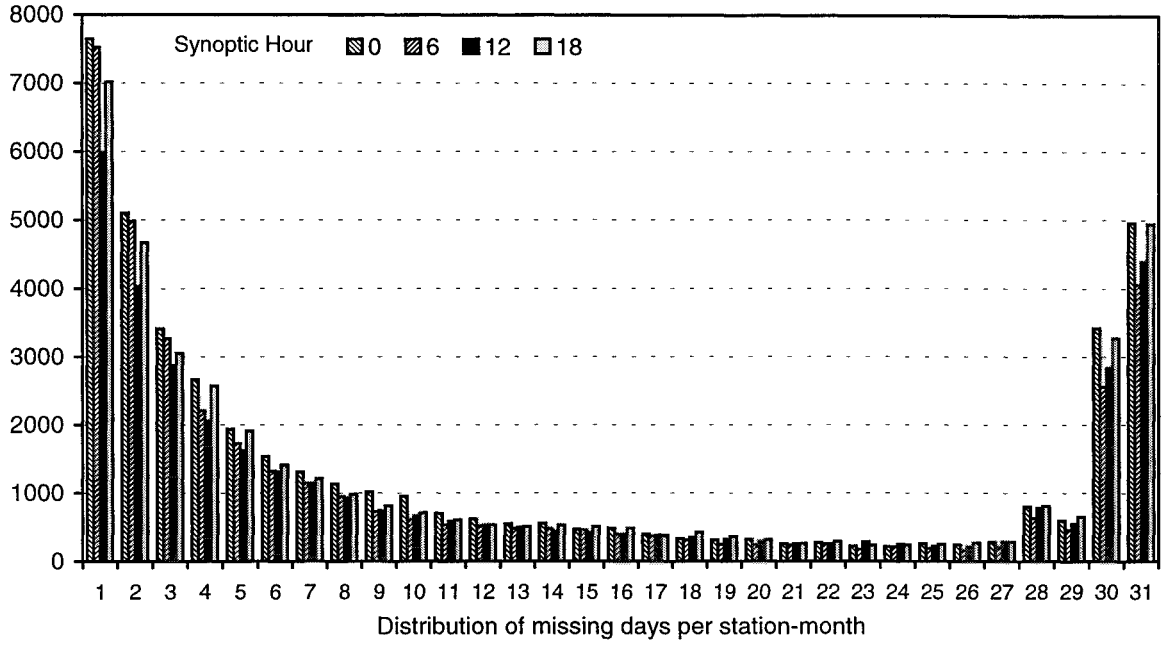
Cumulative distribution of percentage completeness of climatological records

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 45 without any gaps in their climatological records (i.e., those in the <1 class in the figure below for which months that are totalling missing have been neglected), and about half of the records have fewer than 21 gaps. There is a change of class width after the <21 gap class.

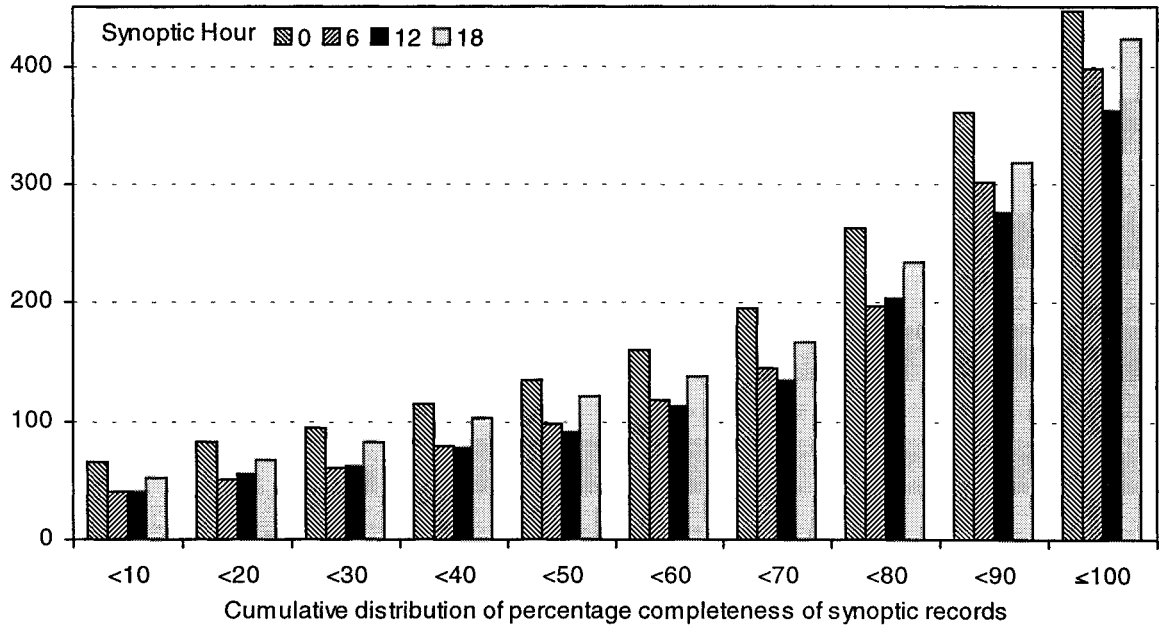


Cumulative distribution of the number of daily gaps at each station

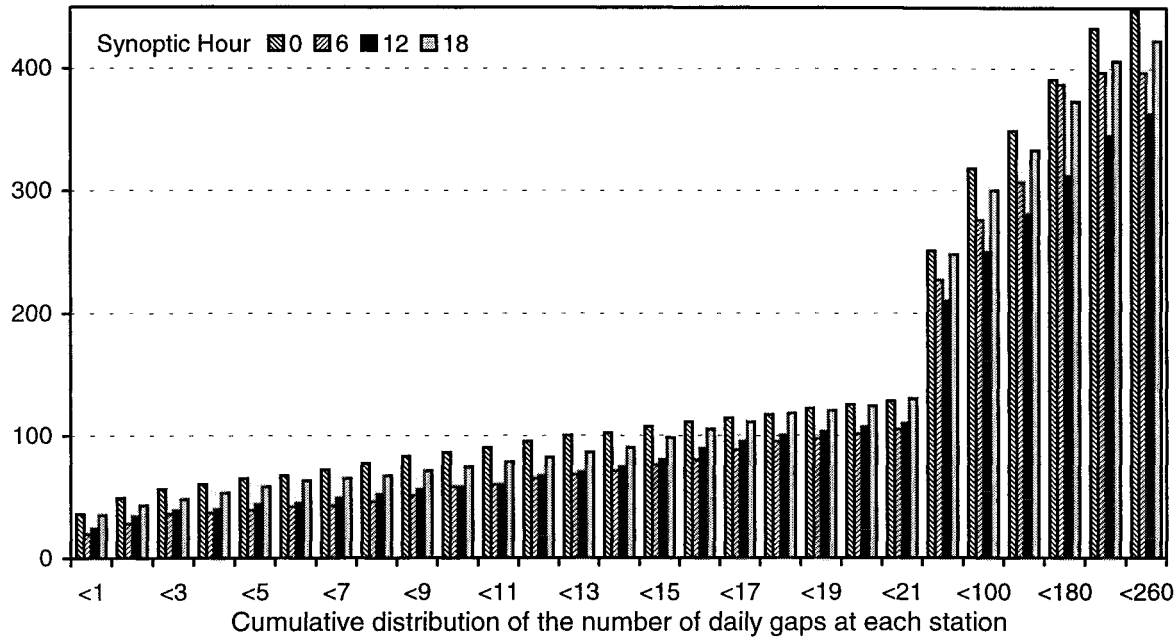
The next three figures are similar to the last three, but show the respective distributions for synoptic records at the main synoptic hours of 00, 06, 12, and 18. It can be seen that they differ little with the hour except the numbers for 06 and 12 tend to be a little lower than for the other hours.



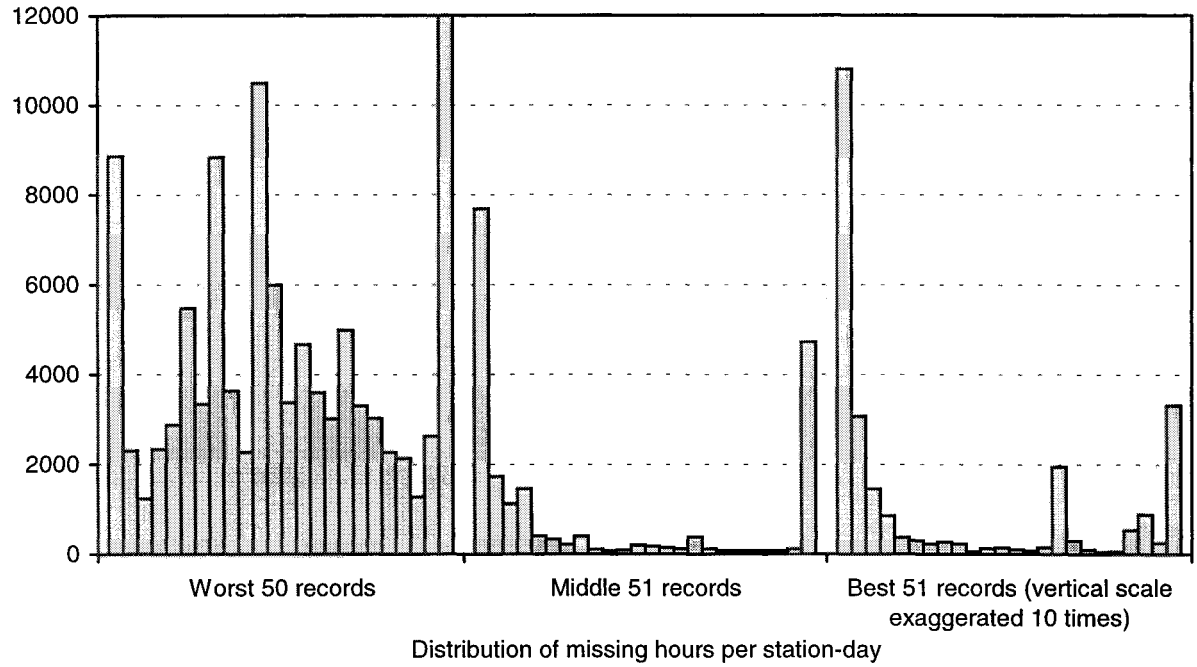
The figure above shows a similar pattern to that of the distribution of missing days for climatological records, but the numbers in all classes from 2 to 27 inclusive are up to twice what they were. Also, there are about 400 station records for each of the hours whereas there were over 600 for the climatological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



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 40% for the climatological records which can be seen from the figure showing its distribution.



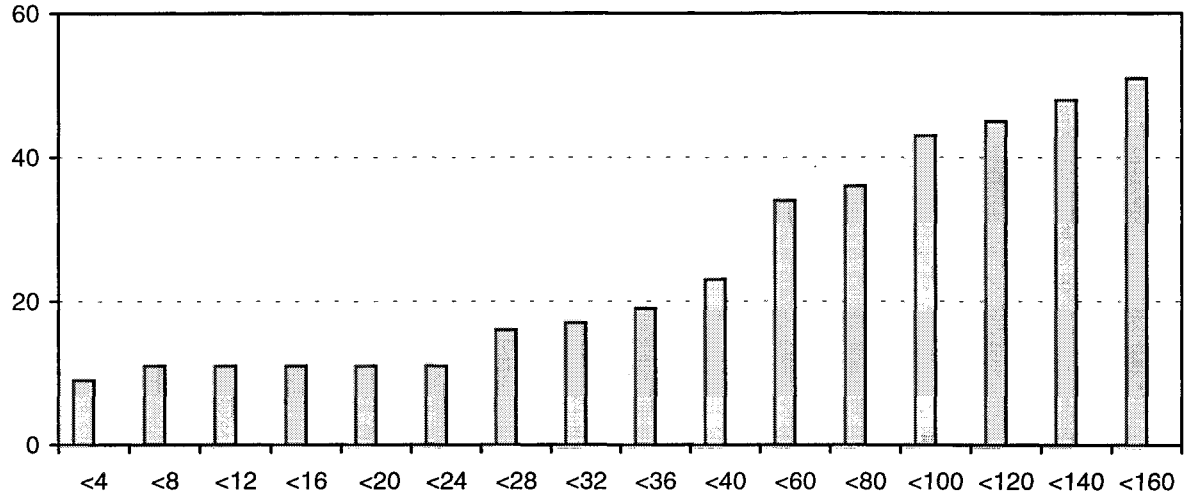
The figure above again shows that synoptic records are of a lesser quality than the climatological ones since about half of those records had under 17 gaps, whereas for synoptic records it can be seen that half of the records have up to 50 gaps.



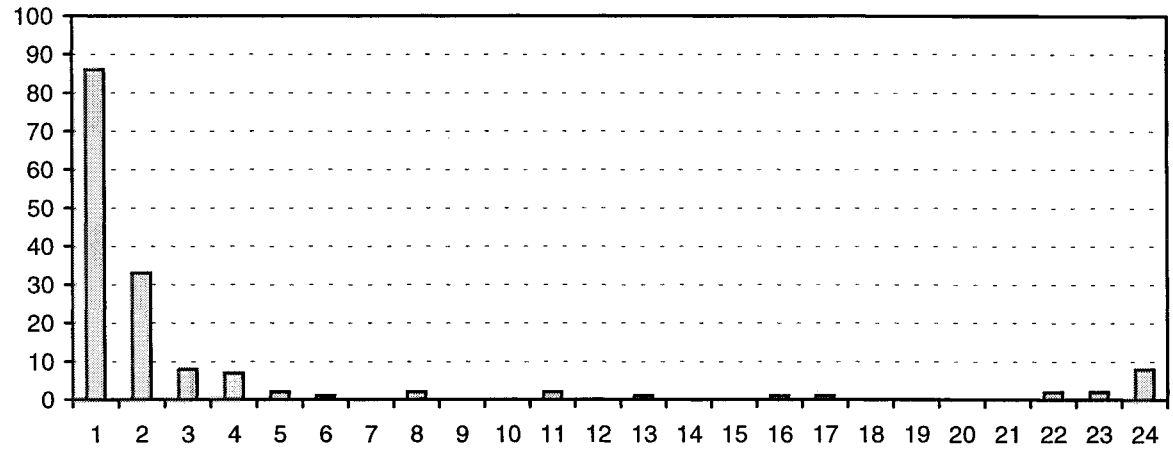
The records of hourly observations were examined and the figure above shows distributions of missing hours per station-day. There were 152 stations with hourly records and in the figure these have been split into three groups and a distribution shown for each group. The left hand scale applies to the first two groups, but it is 10 times too big for the third group so, for example, for the best records about 1000 station-days had a single hour missing and about 300 had 2 hours, which could be either together or apart, etc. This "best" group represents stations which always reported hourly, although the count of 200 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 but the 1 or 2 h classes. This is especially so for the "worst" group where the number at 89 590 is off the scale of the figure and is a significant fraction of the total of 342 300 station-days covered by this group. The "best" group covered 89 268 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 51 such stations with records of hourly observations with 44 of these having records at least 98% complete. The worst stations are, of course, those with the lowest percentage completeness but, as with the climatological and synoptic records, those with a large number of gaps, rather than just a low percentage complete, are also of poor quality. There was one station without any gaps in its hourly observations, 11 stations had fewer than 7 gaps, and under half the records have over 40 gaps. The figure below shows the distribution of hourly gaps; there is a change of class width after the <40 gap class.



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

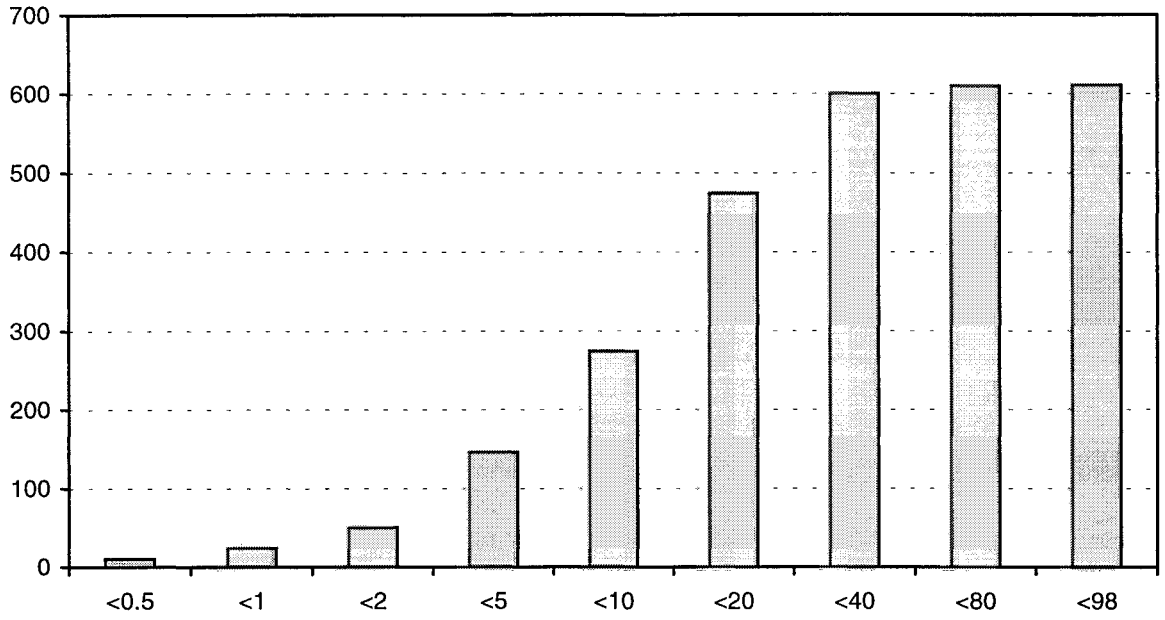


Distribution of gap sizes (h) at station C94012/2283

The station with the most gaps was C94012/2283 with 156 gaps, and the distribution of gaps for this station is shown in the figure above.

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

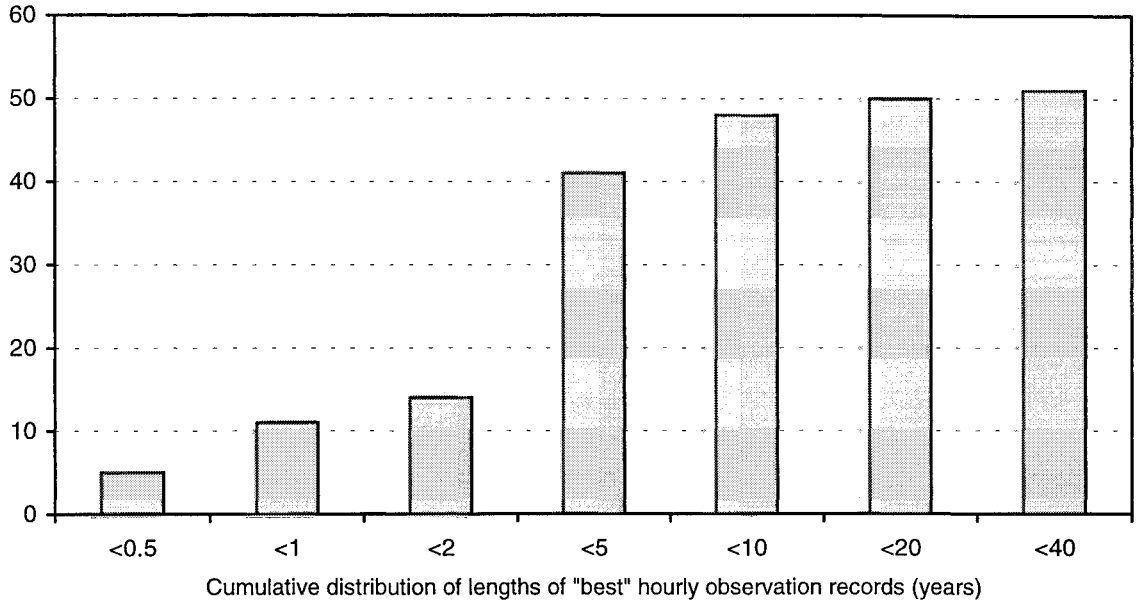
For a given AGENT_NO the climatological and hourly records — synoptic records need not be considered — should be long enough to establish the mean level and variability of the temperature and humidity for the place concerned. Longer records can be used to track any trends, while short records, although 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 climatological observation records.



Cumulative distribution of lengths of climatological observation records (years)

Only 24 records are under a year long and nearly half of the 636 records are over 10 years long. The longest record is from H32641/4881, which lasted from July 1890 until December 1987 and is 99% complete. Of those records under a year which were not from stations which had opened within the last year, four had no more than 19 rows and another two each had about 450 rows, but the data were for other stations. These six records were discarded by 929 rows being deleted from **SCREEN_OBS**.

The distribution of the record lengths for the "best" hourly observations is shown in the figure below. About 10 records are under a year long and about half are over 5 years long. The longest record is from A53021/1024, which lasted from December 1948 until April 1985 and is nearly perfect. Of those records under a year which were not from stations which had opened within the last year, two had some hours with just a few rows each. The records for those were discarded by 60 rows being deleted from **SCREEN_OBS**, thus those stations ceased to have hourly records but still had synoptic records.



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

Details and results of Check D.2 — are monthly temperature statistics consistent with the daily observations upon which they are based?

From the rows with an OBS_DATE equivalent to 0900 Local, monthly summary statistics are calculated and entered into **MTHLY_STATS**. The statistics concerned are: mean vapour pressure; mean 9 a.m. relative humidity; mean 9 a.m. temperature; total Penman potential evapotranspiration (PET); total Priestley-Taylor PET; and, total Penman open water evaporation. There are certain rules associated with their calculation which ensure the statistics are valid and are exactly as defined. For example, for the 9 a.m. rows from a particular AGENT_NO with the OBS_DATEs falling within a particular local month, if there are more than 10 days without an observation, then none of the mean vapour, mean 9 a.m. relative humidity and mean 9 a.m. temperature can be found for that month.

In the example, no statistics are possible and their absence is not an error. Rather this check should look for instances where a statistic exists despite the **SCREEN_OBS** 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 **SCREEN_OBS** and some of these would fail through lack of data or other legitimate reasons that do not occur because an error exists in **SCREEN_OBS** itself. But there are some failures which could be associated with errors in **SCREEN_OBS**, and this check captured those potential errors.

The errors reported that might indicate errors in **SCREEN_OBS** are: rows exist where nothing is recorded for DRY_BULB; despite an error a non-deletable statistic exists; and data exist with an origin not normally associated with **SCREEN_OBS**. However, no errors of these types were found.

Summary and Conclusion

The grand total of changes made to **SCREEN_OBS** was 379 902, which is 1.9% of its total number of rows. The tabulation below shows that the largest contribution was in the "Remove Humidity" class in which the DRY_BULB value of the row was not changed, but WET_BULB and DEW-

POINT were always set to NULL, and sometimes RELATIVE_HUMIDITY, ORIG_WET_BULB, and WET_BULB_REL were as well. However, in only 9% of the 325 199 cases was the RELATIVE_HUMIDITY, etc, removed and the other cases were those with a DRY_BULB lower than −10 °C and the new rule given in check B.8 was applied. For both deletions and amendments, about half the cases came from just two sources each.

- 13 569 deletions were for 28 Antarctic stations where time-ordered listings had been used to detect rows with large differences from their temporal neighbours.
- 7199 deletions were for just three Pacific island stations for the rows which had a DEWPOINT of 0 °C and the DRY_BULB was a multiple of 5 °C.
- 5111 amendments were made to the OBS_DATE of rows from a number of non-New Zealand stations. The rows were of synoptic origin but the times were either an hour before or after a synoptic reporting time and no row was present for that time.
- 2242 amendments were made at L00900/6194 mostly for the winters of 1978–81 inclusive when temperatures reported by synoptic observations ranged between −10 °C and 20 °C but those reported through the daily climate observation were −70 °C to −40 °C. Changing DRY_BULB to (DRY_BULB × −1) − 50 gave values which agreed with the climate observations.

The changes made to CLIDB are summarised in the following tabulation.

Table name	Deletions	Amendments	Remove humidity
SCREEN_OBS	42 256	12 447	325 199
LAND_STATION	0	55	—
SITE_CHANGES	95	55	—
Other tables	146	4	—
Total	42 497	12 561	325 199

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.

- The **HUMIDITY** procedure was modified to assume a height of zero for the calculation if no height is given.
- The program that calculates **MTHLY_STATS** code 16 (mean vapour pressure) was amended to use RELATIVE_HUMIDITY rather than WET_BULB when calculating vapour pressure because for DRY_BULBs under −10 °C the humidity data are now only available in RELATIVE-_HUMIDITY.
- A new procedure **WRITE_SCREEN_OBS** was written to follow all the rules, old and new, which apply to the insertion and amendment of data into **SCREEN_OBS**.
- The new procedure **WRITE_SCREEN_OBS** was incorporated into the following archiving procedures RMSDYCLI, RMSEDR, RMSHOURLY, RMSMETAR, RMSSHPSY, RMSSYNOP, RMKS_DECODE, 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.