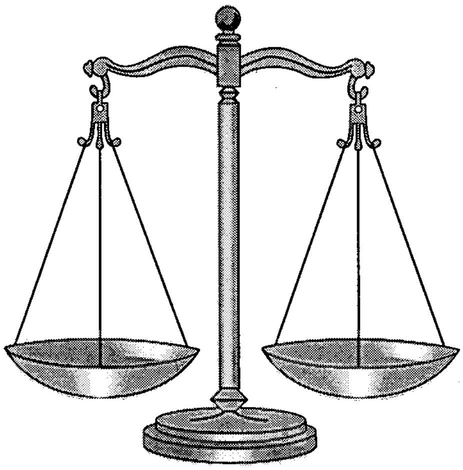
New Zealand's National Climate Database (CLIDB): audit report on the RAIN table

John Sansom
Allan Penney



NIWA Technical Report 65
ISSN 1174-2631
1999

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

John Sansom
Allan Penney

NIWA Technical Report 65
1999

Published by NIWA
Wellington
1999

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

ISSN 1174-2631
ISBN 0-478-08488-9

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

The auditing of the dataset within New Zealand's National Climate Database which contains rainfall observations is described. Each row in the dataset consists of the place of observation, the date and time of observation, the observing interval, the amount of rain accumulated over the interval, and some minor attributes. All the attributes were checked individually and in groups so that any invalid values were found; consistency between different observing intervals for the same place and time 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.

Errors were found and, excluding a few changes to other tables, a total of 2 118 583 changes were made, which is 5.7% of the total number of rainfall observations. However, only 0.4% of the daily observations were changed, only 3.6% of the hourly ones, but 45.0% of the synoptic ones. These percentages depended on just a few extensive changes; thus 95% of the changes to daily observations came from 96 220 insertions from a reclassification of synoptic reports, and 5445 deletions from Antarctic stations. Also 95% of the changes to hourly observations came from the transfer of 209 639 observations accredited to Auckland City station rather than Albert Park, and the amendment of the times on a further 61 361 observations. Finally 90% of the changes to synoptic observations came from the deletion of 439 037 rows with times not at the standard reporting times, another 731 015 deletions where contemporary hourly observations made the poorer quality synoptic data redundant, and 359 604 amendments to correct data that had been recovered with errors from CLIDB's predecessor. The need for these extensive changes to the most noticeable errors could have been found at any time and it is, perhaps, the other more particular changes which were the most valuable since the subtlety of many of those errors kept them so well hidden that only the auditing was likely to find them.

Apart from the changes to the data, some changes were also made to the programs which process the raw synoptic data, the scripts which catalogue CLIDB's data, and to the procedures which check the consistency between rainfall observations and rain rate data.

Introduction

This report is the second in a series which will document the auditing of particular data tables within New Zealand's National Climate Database (i.e., 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.

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 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 worse 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 previous runs 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 RAIN table

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

Column name	Null?	Type
AGENT_NO	NOT NULL	NUMBER(6)
OBS_DATE	NOT NULL	DATE
FREQUENCY	NOT NULL	VARCHAR2(1)
ORIG_OBS_ORIGIN		VARCHAR2(1)
AMOUNT		NUMBER(5,1)
PERIOD		NUMBER(10,4)
RELIABILITY		VARCHAR2(1)
STGD_STATE_OF_GRND		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**; the UTC date-time given by OBS_DATE; and reporting frequency (i.e., daily, hourly) given by FREQUENCY. The remaining columns constitute the significant data with AMOUNT being the primary data since a row without this contains no information. Apart from PERIOD, all the other columns could be null, although ORIG_OBS_ORIGIN should usually be present. STGD_STATE_OF_GRND is rarely used and will not be audited.

A full description of **RAIN** 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 rainfall observation.

A.2. OBS_DATE: Should not be later than the current date.

A.3. FREQUENCY: The entries in this column should all represent valid frequencies, i.e., they should all appear as CODEs in **CODE** when CODE_TYPE is “FREQ”.

A.4. ORIG_OBS_ORIGIN: The entries in this column should all represent valid origins, i.e., they should all appear as CODEs in **CODE** when CODE_TYPE is “ORIG”.

A.5. PERIOD: Should be present and greater than zero.

A.6. AMOUNT: Should be present and non-negative.

A.7. RELIABILITY: Only NULL (i.e., empty) or “*” are allowed.

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 before a raingauge was installed or after one was removed.
   - B.2. OBS_DATE, FREQUENCY: For a given FREQUENCY the earliest date should be reasonable and all observations at the correct times.
   - B.3. FREQUENCY, PERIOD: The FREQUENCY and the PERIOD should be consistent.
   - B.4. FREQUENCY, AMOUNT: For a given FREQUENCY the largest AMOUNTs should be reasonable.

C. Between row checks
   - C.1. For a given FREQUENCY and AGENT_NO, the OBS_DATE and PERIOD should be such that the previous observation at that FREQUENCY and station was made no later than PERIOD hours before OBS_DATE.
   - C.2. For a given FREQUENCY and AGENT_NO, the greatest AMOUNT should not be excessive.
   - C.3. For a given AGENT_NO, OBS_DATE and FREQUENCY, the AMOUNT accumulated over the associated PERIOD should be consistent with the sum of any AMOUNTs at a "higher" FREQUENCY over the same accumulation period.
   - C.4. For the same OBS_DATE and PERIOD with FREQUENCY of "D", the AMOUNTs should not be too different for AGENT_NOs that are physically close to each other. (FREQUENCYs "higher" than "D" are not considered due to rainfall's high variability).
   - C.5. For a given FREQUENCY and AGENT_NO, there should be a continuous dataset with no gaps from the row with the earliest OBS_DATE to that with the latest.

D. Other checks
   - D.1. For a given FREQUENCY and AGENT_NO, the length of record should be adequate.
   - D.2. For FREQUENCY of "D", any AGENT_NO should not have an excessive number of PERIODs greater than 24 h. Also few PERIODs should be such that PERIOD hours before OBS_DATE is in one local month and OBS_DATE in another.
   - D.3. For a given AGENT_NO, the FREQUENCY "D" rows that make up a complete local month should have associated rows in **MTHLY_STATS**.
   - D.4. For a given AGENT_NO and OBS_DATE, a FREQUENCY "D" row may have an equivalent row in **RAIN_RATE** in which case AMOUNT and TOTAL (i.e., a column in **RAIN_RATE**) should be equal.

The checks above operate at three levels, i.e., finding absolute errors, identifying possible errors, and measuring quality. Thus the A checks all search for absolute errors as do B.3, C.1, C.3, and D.4 whereas B.1, B.2, B.4, C.2, C.4, and D.3 will highlight those rows that might be in error. Remaining checks (C.5, D.1, and D.2) may uncover some errors but it is more likely that any gaps in a record or any short records or excessive accumulations are due simply to lack of data, and these checks will highlight the poorer records. Checks C.5 and D.1 are the only ones that perform any sort of temporal check since with rainfall data the correlation from one day or hour to the next is small and, thus, the expectation that consecutive values should be of a similar size — as in a true temporal check — is not valid.

Audit results

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

For any row in **RAIN** 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 **RAIN** as a code for the station, thus, all the AGENT_NOs in **RAIN** 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 rainfall. In the tabulation below this applied only to the "Anemometer Only" type, but 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 the case for the six "Anemometer Only" stations that had reported rainfall, i.e., B75572/1566, C84173/2135, D87862/2839, E05282/3223, E93482/3504, H32231/4813. A check of these six stations revealed that they had been legitimate observers of rainfall but had ceased, continuing as "Anemometer Only" stations.

RAIN (STANDARD)	1 976
CLIMAT (STANDARD)	432
CLIMAT/SYNOP	130
RAIN/SYNOP	52
CLIMAT (PRIVATE)	7
RAIN (PRIVATE)	122
REGIONAL COUNCIL	101
WATER SCIENCES	38
ANEMOMETER ONLY	6
SYNOP ONLY	156
AWS (SYNOP AND METAR)	107
EDR	18
CLITEL	23
LIMITED CLIMAT	6
SPECIAL STATION	9

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

For any row in **RAIN** it must be known at what date and time the observation was made and these dates should not be later than the current date. This was found to hold and so all the latest dates were valid. Unlike the latest date when the current date provides an error threshold, there is no natural threshold for the earliest date. Also, since observations of some FREQUENCYs were started earlier than those of others, there is no fixed threshold either for the earliest date. However, the earliest dates for each FREQUENCY can be found and, as can be seen below, these dates are reasonable with daily data being available from February 1862 and the others from about 1960.

Frequency	Earliest data
D (i.e., Daily Observations)	18620201
H (i.e., Hourly Observations)	19600110
S (i.e., Synoptic Observations)	19610331

As noted above, the times of observation are also constrained since: FREQUENCY "D" rows should have a time equivalent to 0900 Local; FREQUENCY "S" rows should be at one of the main synoptic reporting times, i.e., 0000, 0600, 1200, 1800 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, 0500, 1100, 1700 UTC; and, FREQUENCY "H" 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 all hourly observations are correct, but errors were found in the times of other observations.

For daily observations the distribution with hour was:

New Zealand		Non New Zealand	
Local hour	Number	Local hour	Number
0000	3	0000	5
0300	9	0200	10
0600	204	0300	1
0700	14	0500	3
0800	81	0600	8
0900	23 437 762	0800	5
1000	1	0900	2 234 225
1200	37	1100	11
1500	34	1200	7
1800	122	1300	1
2100	5	1400	3
		1500	2
		1700	2
		1800	11
		2000	5
		2300	21

i.e., the vast majority at 0900 Local. However, there were 605 error cases, of which 510 were for New Zealand and all originated from synoptic reports. For New Zealand, 51 stations were involved but 4 of these had 371 (about 75%) of the errors and the times of all but 4 of these were changed to 0900 Local and the remaining 510–367=143 rows were deleted. For outside New Zealand, 29 stations were involved and the 95 rows were deleted.

In a preliminary run for synoptic observations it was found that two Fijian stations had been classified as New Zealand stations through their DAYL_DAYLIGHT_AREA in **LAND_STATION** being set to the wrong value. After correcting this error — and ensuring no other such errors exist in **LAND_STATION** — the following distribution with hour for synoptic observations resulted:

Hour (Local for NZ else UTC)	New Zealand	Non New Zealand
0000	373 253	422 162
0100	6 103	307
0200	1	57
0300	9 553	75 987
0400	12	316
0500	189	127
0600	444 000	421 079
0700	7 110	318
0800	2	40
0900	108 389	64 921
1000	51	269
1100	190	149
1200	448 594	378 444
1300	6 220	298

1400	3	42
1500	11 971	62 131
1600	132	278
1700	216	118
1800	438 323	395 300
1900	6 411	285
2000	0	79
2100	8 838	93 148
2200	10	313
2300	154	143

i.e., most at 0000, 0600, 1200, 1800, but still 464 881 error cases of which 165 555 were for New Zealand from 228 stations and 299 326 for outside New Zealand from 139 stations.

The many New Zealand rows at the hour after the true reporting time may have been due to a problem with daylight saving. All such reports were found to be during periods of daylight saving and of the 25 844 only 91 occurred after a report at the correct time. These 91 were deleted and the OBS_DATEs on the 25 844 were decreased by 1 hour. The remaining 139 711 were checked for any observations with a PERIOD of 24 which might indicate that a FREQUENCY "D" observation had been archived incorrectly. However, the PERIODs were mainly 6 h and so they were deleted.

A check of the PERIODs associated with the 299 326 non-New Zealand error cases showed that most of these were 6 h but about 4000 had 24 h and 500 had 30 h. These 500 all originated from 7 Fijian automatic weather stations which were really reporting 1 h rainfalls every 3 h. Their reporting procedure was correct but our decoding of the synop message was incorrect since WMO code 4019 (WMO 1995) provides a suitable code for that and also for 2, 3, 9 and 15 h rainfalls. Since observations with these PERIODs appear to be relatively rare, any resulting datasets would be sparse and not useful: therefore, it was decided that such observations would not be archived in future and any already archived — they would have been archived with PERIODs greater than 24 h — should be deleted.

All the other non-New Zealand error cases were also deleted since they were at non-standard times with a PERIOD that overlapped an observation at a standard time (i.e., an error situation checked for by C.1) and the observation at the standard time was preferred. Such situations could also occur in future so a pre-processor to RMSSYNOP (i.e., the checking and archiving program which processes the raw data loaded into **RMS_SYNOP**) was amended so that rainfalls for non-New Zealand stations with non-standard times are deleted from **RMS_SYNOP** and so are not entered into **RAIN**.

A subsequent check after these program amendments showed that Cook Island stations were still having observations from non-standard hours entered into **RAIN**. The was because in **LAND_STATION** nine Cook Island stations had been given a value in DAYL_DAYLIGHT_AREA of '03' rather than the '01' common to all other non-New Zealand places. These values were amended and a further 454 rows deleted from **RAIN**. This check also highlighted the fact that PERIODs of over 24 h were still being archived at standard hours. Further amendments were made to the RMSSYNOP program and another 2302 rows deleted from **RAIN**.

Details and results of Checks A.3, A.4, and A.7 — are all frequencies, reliabilities, and origins valid?

For any row in **RAIN** it must be known over what PERIOD the rainfall accumulated before the amount was measured. However, PERIOD is expected to fall into a few groups (*see* results for B.3) and these are labelled by FREQUENCY. A list of the valid frequencies with a full description is held in **CODE** where CODE_TYPE is "FREQ" and only these should appear in **RAIN**. This was found to hold, and so all frequencies were valid.

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

For any row in **RAIN** 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. A list of the valid origin types with a full description is held in **CODE** where CODE_TYPE is "ORIG" and only these should appear in the ORIG_OBS_ORIGIN column of **RAIN**. This was found to be mostly the case, and so most origins were valid apart from those denoted by "Ref 1–3" below.

Frequency	Origin	Reliability	Count	Ref
D	D		25 594 900	
D	D	*	13 637	
D	S		134 564	
D	S	*	76	
D	P		4 324	
D	P	*	530	
D	Q		2 997	
D	Q	*	3	
D	^	*	1	1
H	M		6 569 880	
H	M	*	6 936	
H	E		706 684	
H	E	*	662	
H	H		446 832	
H	H	*	1	
H	F		61 361	2
S	S		3 842 892	
S	S	*	26	
S	D		6	3

1. This was for I68182/5778 on 1 Jan 1987 and the origin was changed from "^"to "D" to align it with the station's usual reporting practice.
2. Origin "F" implies the data were derived from paper forms. Although it is a valid origin type, it was unexpected for hourly rain data and was found to be for A54733/1283 for the years 1979–85 inclusive. However these **RAIN** observations had been specially collected and entered from forms (along with data for **SCREEN_OBS** and **EARTH_TEMP**) so they were valid although DATA FAULT 1994/013 was outstanding regarding these data. The fault was that when recovered from Trentham and placed in CLIDB it was assumed that data times were NZDT when they were NZST. The times were changed to correct the **RAIN** part of the fault.
3. This was for E04991/3145 for 10–19 Oct 1991 and, although "D" is a valid origin type, it should not occur with "S" FREQUENCY data. The six records had identical values to those with a FREQUENCY of "D" and an ORIG_OBS_ORIGIN of "D" and so they were deleted.

The table provides an extension to the basic checks since the combinations of FREQUENCY and ORIG_OBS_ORIGIN were also considered. Furthermore, the table provides a measure of quality since it can be seen that, generally, only about 0.1% of the rows for a given FREQUENCY–ORIG_OBS_ORIGIN pair have a RELIABILITY of "*". However, that only 1 H–H and 26 S–S rows have a "*" reflects more the fact that these types of observation are as-received and subject to little quality control rather than that they are more reliable.

Details and results of Checks A.5 and B.3 — are all periods valid and consistent with the associated frequency?

For any row in **RAIN** it must be known over what PERIOD the rainfall accumulated before the amount was measured, thus, PERIOD should be non-NULL and greater than zero for all values of FREQUENCY. However, 349 error cases were found and all of these were for FREQUENCY "S" rows, 14 with a PERIOD of NULL and 335 with zero. The worst station was J76200/6044 which had 69 errors.

Since the earliest date for these error cases was October 1991, the problem was attributed to the checking and archiving program RMSSYNOP, which processes the raw data loaded into **RMS_SYNOP.** Rows for **RAIN** had their PERIOD calculated from IND_PRECIP_PERIOD by multiplying it by 6 but without checking its value first, hence NULL and zero values resulted if IND_PRECIP_PERIOD was NULL or zero. The program was modified to accept only indicator codes 1–4, and flag any others as errors. The 349 error rows were deleted since if the original data had not contained a legitimate period is was unlikely that is had contained correct rain amounts.

But FREQUENCY also constrains the value of PERIOD since when FREQUENCY is "D" PERIOD must be a multiple of 24 and for "S" a multiple of 6 but no more than 24 and for "H" it must be unity. One exception was found: station E94311/3523 had a PERIOD of 364 h and this was changed to 336 h. (Some exceptions for "S" rows might have initially existed, i.e., those with PERIODs over 24 h, but those rows had already been deleted.)

Details and results of Checks A.6 and B.4 — are all amounts valid and reasonable for the associated frequency?

The essential data in any row of **RAIN** is the value of AMOUNT, so AMOUNT should be non-NULL and non-negative for all values of FREQUENCY. However, 320 error cases were found: 190 with an AMOUNT of NULL of which 114 were for Fijian stations and 76 for New Zealand stations and all with a FREQUENCY of "D"; and 130 with negative values all for station A54733/1283 and all with a FREQUENCY of "H".

The NULLs occur for daily data because of the way accumulations are handled in the current archiving program for manual data and the way they were handled by the recovery program that populated **RAIN** in CLIDB from the data held at Trentham. An accumulation occurs when a reading of the raingauge is missed for one or more days so that when it is finally read the observation applies to all the days since the last reading and not just to the 24 h before the reading. The archiving programs deal with monthly sets of daily data and as they work through the month and come upon a day with no rain data a NULL record is made in **RAIN**. This is done since the expectation is that a day with data will soon be found and the number of NULL records before that day will preserve the length of the accumulating period after which they can be deleted. However, if an accumulation crosses a month boundary, a row is left in **RAIN** with an AMOUNT of NULL. The Fijian cases were simply deleted but the New Zealand ones were examined more closely by looking for the next row, then: if no row occurred within the next month, the NULL row was deleted – 24 cases; if the next row was a dry day, the NULL was amended to a zero – 11 cases; or, if the next record was not zero, the PERIOD from the NULL row was added to that of the next row to extend the accumulation period – 41 cases.

The negative AMOUNTs for station A54733/1283 came from a special dataset at Trentham in which estimated values were held as negatives, but this had been overlooked when the data were recovered and put into CLIDB. Thus, the negative values were simply made positive and RELIABILITY was set to "✳".

But FREQUENCY also has implications with regard to AMOUNT since the largest values for "D" FREQUENCY rows should be larger than those for "S" or "H" rows. All "D" and "S" rows with AMOUNT greater than 999 mm and all "H" rows with AMOUNT greater than 99 mm were considered.

There were 24 such "D" rows of which 19 originated from synoptic reports and had an AMOUNT of 999.X mm where X was from 1 to 8. In these reports rainfall of 0.1 mm is encoded as 9991 and 0.2 mm as 9992 etc. and the 19 large values were each reduced by 999 mm. Of the other 5 large values, one was deleted and the other 4 accepted as they had large PERIODs and so were long accumulations. There were no "S" rows with AMOUNT greater than 999 mm, but there were 16 "H" rows with AMOUNT greater than 99 mm. Of these 16, 11 had AMOUNT reduced to 0 mm, 2 to a smaller value, 2 were deleted and only 1 was accepted, i.e., 100.6 mm between 10 and 11 pm on 16 February 1966 at A64761/1410.

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

The dates of opening and closing for each station are held in **LAND_STATION** and the dates of the installation and the removal of a raingauge from a station are held in **RNGAUGE_HIS**. Thus for each AGENT_NO the earliest and latest records within **RAIN** can be found and an error noted if the earliest is before the station opened or a raingauge was installed or if the latest is after the station closed or the raingauge was removed.

Some of the 505 errors found were examined in detail but no consistent way of treating the problem became apparent, except the simplistic one which would be to accept that any data outside the station or gauge dates were valid and to amend the date of the station's opening or closing or the gauge's installation or removal to accommodate the excess data. This may well 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. A sample of the date anomalies found is given below to illustrate the different types of anomalies that occurred.

AGENT_NO	Station start	Raingauge start	Data start	Data end	Raingauge end	Station end
1011	19670901	19670901	19670801	19960331		19970404
1126	19220601	19220601	19220601	19710630	19710531	19710531
2005	19691001	19691001	19691001	19891130	19890831	19890901
4734	19810131	19810201	19801201	19981031		
7337	19841231		19850201	19920531		19910501

The anomaly types are tabulated below where it can be seen that the most frequent are occasions when the data date was at variance with a date common to both the station opening and the gauge installation or the closing/removal. Anomalies with different dates exist, but are only 10% of those with common dates. A final error that occurred was where data existed but no gauge dates were available.

		Example
Type of anomaly	Number	AGENT_NO
Data began before station/raingauge opened	301	1011
Data continued after station/raingauge shut	169	1126
Data began before station opened	15	4734

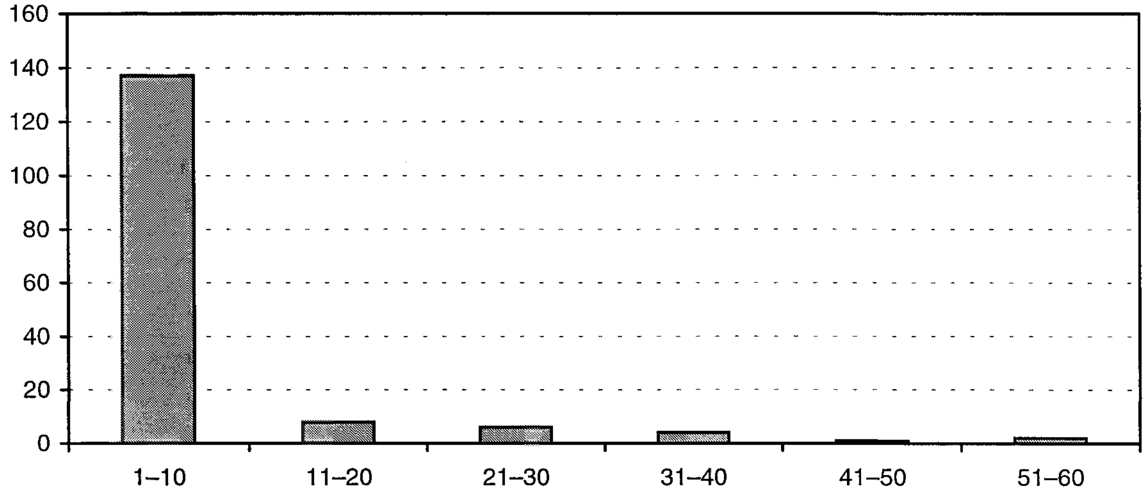
Data continued after station shut	10	2005
Data began before raingauge installed	15	4734
Data continued after raingauge removed	6	2005
Data present but no raingauge given	13	7337

With such a large number of errors to deal with and no overall solution available, no changes were made to **RAIN**, **RNGAUGE_HIS**, or **LAND_STATION** but entries were made in **SITE_CHANGES** as indicated below. In **LAND_DATA_CAT** rows where LADA_LAND_DATA has a value of "181" contain information on daily rainfall, hence, to assist the retrieval of these rows it has been used to form the time part of "Data Date".

AGENT_NO	Data Date	Description
1011	19670801:1810	RAIN 181 began BEFORE station and RAIN-GAUGE opened on 19670901
1126	19710630:1810	RAIN 181 continued AFTER station and RAIN-GAUGE shut on 19710531
4734	19801201:1810	RAIN 181 began BEFORE station opened on 19810131
2005	19891130:1810	RAIN 181 continued AFTER station shut on 19890901
4734	19810201:1810	RAIN-GAUGE installed AFTER RAIN 181 began on 19801201
2005	19890831:1810	RAIN-GAUGE removed BEFORE RAIN 181 ended on 19891130
7337	19850201:1810	RAIN-GAUGE not given for RAIN 181 data 19920531

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

For a given FREQUENCY and AGENT_NO the OBS_DATE and PERIOD define the interval over which the AMOUNT accumulated. These intervals must not overlap — the observations are independent and complete. For FREQUENCY "H" this was always the case since all OBS_DATEs were on the hour, but for FREQUENCYs "S" and "D" errors were found, 152 262 and 857 respectively.



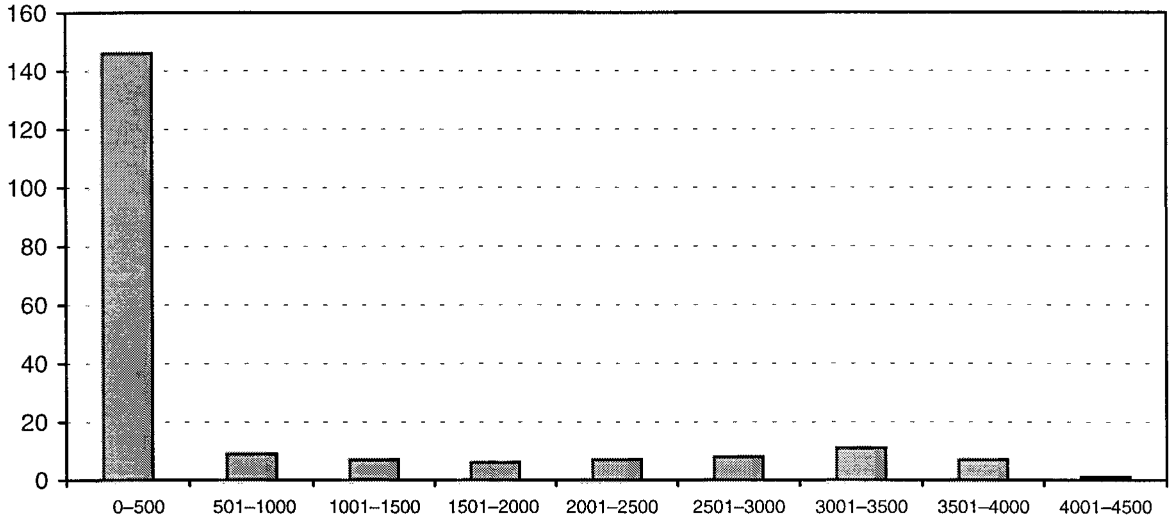
Distribution across stations of daily overlaps

The 857 daily overlaps were spread over 155 stations and were distributed among them as in the figure above which shows that most of the stations had under 11 overlaps. The most daily overlaps at one station was 58 at C74283/2006. The major cause was where in a series of rows from a non-synoptic origin the gaps left by accumulations over two or more days had been filled in by synoptic data. This would have arisen during the recovery of data from Trentham. Since synoptic data are of lesser quality, the accumulation gaps would have been days of missing data and so the synoptic data were deleted. This affected 47 stations and 652 rows were deleted, but 457 overlaps still remained; the apparent excess was because often more than one synop had been overlapped by a single accumulation.

The remaining overlaps fell into four groups:

- in a series of daily observations each had a PERIOD of 48 h despite there being a row for each day;
- within an accumulation over several days additional dry days had been inserted;
- the accumulation over a series of days had been attributed to the first of the days rather than the last; and,
- during a quality control procedure a series of daily readings, which included an accumulation, had been all put back one day but the reading(s) which were then covered by the accumulation were not deleted.

These types of errors were not trapped during the calculation of monthly statistics until the procedures were changed in early 1998 and most of the errors were for data prior to that time. The errors were corrected manually, with 358 rows being updated and 338 rows deleted from 112 stations. Furthermore, with the removal of all 857 overlaps an additional 447 station-months of monthly statistics were calculated.



Distribution across stations of synoptic overlaps

The 152 262 synoptic overlaps were spread over 202 stations and were distributed among them as in the figure above which shows that most of the stations had under 501 overlaps. The most synoptic overlaps at one station was 4370 at J75300/6029. Since by the time this check was performed the times of synoptic reports had been set to the standard hours at 6 h intervals (*see* results for A.2 and B.2 above), overlaps could occur only for PERIODs of 12,18 or 24 h. For those with 24 h an investigation was made to determine if they could be changed to FREQUENCY "D" observations. This uncovered a major error in the transfer during 1991 of the synoptic rainfalls from Trentham into CLIDB.

At Trentham, synoptic rainfalls had been coded using 990 for 0.0 mm and 991 for 0.1 mm up to 999 for 0.9 mm then 001 for 1 mm, 002 for 2 mm, etc. The codes 990–999 had been decoded correctly but it been assumed that 001 was for 0.1 mm etc. Thus for FREQUENCY “S” rows within CLIDB with dates earlier than October 1991, AMOUNTs of zero are correct and AMOUNTs over 1 can be corrected by multiplying them by 10 but an AMOUNT of 0.1 mm, say, may be correct or it may really be 1 mm. These doubtful cases might be resolved through comparisons with the appropriate “H” rows, but, if the “H” rows exist then the “S” rows can be deleted. Similar comparisons with “D” rows might also be possible for those “S” rows with a PERIOD of 24 h, but, if a “D” row is available then, again, the “S” row can be deleted, so such comparisons were not attempted directly but are implied in the following.

- If during the time OBS_DATE minus PERIOD to OBS_DATE for a FREQUENCY “S” row there were any FREQUENCY “H” data or the interval was covered by some data in **RAIN_RATE** (i.e., FREQUENCY “H” rows could be generated), then it was deleted. This was done for both pre-1992 and post-1991 data with a total of 731 015 FREQUENCY “S” rows being deleted.
- For all FREQUENCY “S” rows with a PERIOD of 24 h, an attempt was made to update the OBS_DATE to the nearest 0900 Local and the FREQUENCY to “D”. If this update succeeded, extra daily data resulted, but if it failed then a daily observation must have been already present and the “S” row was deleted. With successful updates RELIABILITY was set to “✳” for non-zero AMOUNTs and, for pre-1992 data, it was also necessary to update AMOUNTs of 1 mm or more to AMOUNT ×10, but AMOUNTs between 0.1 mm and 0.9 mm were left unchanged and to show the doubt over the value RELIABILITY was set to “?”. Overall, 96 217 rows were updated to make new daily rows and 48 313 rows were deleted. Furthermore, the RMSSYNOP program was amended so that any future incoming synoptic observation with a PERIOD of 24 h will be archived, if possible, as a daily observation.
- Remaining pre-1992 synoptic data AMOUNTs were updated by multiplying by 10 where AMOUNT was 1 mm or more and setting RELIABILITY to “✳” (49 688 cases) or “?” (309 916 cases) as appropriate.

After these changes only 17 438 synoptic overlaps were left.

A final means of rescuing synoptic overlaps was where the AMOUNTs in both the overlapped and overlapping rows were zero, when the overlapping PERIOD was cut back to just fill the interval between OBS_DATE of the overlapped and OBS_DATE of the overlapping; 3441 rows were amended in this way leaving 13 997 overlaps. These 13 997 rows were just those synoptic rows in **RAIN** with PERIODs of 12 or 18 h and so they were deleted. However, it is not possible to stop 12 or 18 h PERIOD synoptic data being archived in future as they may be legitimate, so overlaps may still occur in future.

Details and results of Check C.2 — are all the largest amounts reasonable?

A gross check on AMOUNTs was performed in B.4 above, where depending on FREQUENCY values above a certain level were checked. In this check the largest value for a given AGENT_NO-FREQUENCY pair will be compared to the mean value for that pair. Since PERIOD may vary within a given FREQUENCY, AMOUNT was standardised by dividing it by PERIOD and multiplying the result by 24 or 6 according to FREQUENCY being “D” or “S” respectively. Also AMOUNTs of zero were not included in deriving the mean.

Initially, the 5% of stations with the largest ratios of the maximum AMOUNT to the mean were found. There were 181 stations: 157 with FREQUENCY “D”, 18 with “S”, and 6 with “H”. Almost all the “D”s were for data with a synoptic origin and for most of these, and the “S” stations themselves, the maximum found was closely followed by another large value. But “How large was too large?” — Tomlinson (1980), which includes maps showing the spatial variation of maximum rainfalls during specified durations and return periods, was used as a guide. For New Zealand the area of largest maxima is Fiordland and, since CLIDB contains synoptic rainfall records back to about 1960, maxima with a return period of 40 years were estimated giving 500, 720, and 870 mm for PERIODs of 6, 12, and 18 h respectively. For 24 h the value was close to the 999 mm which had already been used in check B.4. These values were used as upper limits and any larger AMOUNTs were deleted; the numbers deleted are shown below.

PERIOD (h)	Range of AMOUNTs deleted (mm)		Number
	Minimum	Maximum	
6	501	988	449
12	722	989	74
18	900	987	8

From the initial list of maximum to mean ratios, it could be seen that most of these were over 100, i.e., the maximum was at least 100 times greater than the mean. Thus to capture not only the largest at a station but any other large ratios at the same station, all observation above a given ratio ("Max. Ratio" in the table below) were examined. The numbers found for each FREQUENCY and the remedial actions taken are tabulated below. A general methodology for dealing with the synoptic cases was to compare the total of the synoptic observations for a day to the equivalent FREQUENCY "D" row since it would usually be subject to more thorough quality control procedures. At times the daily observation itself was suspect, in which case it was checked against nearby stations.

Max. Ratio	Freq.	No.	Action taken
100	D	19	Ten were for an Antarctic station and were deleted because the data were incomplete for that month. Eight were in the Pacific Islands and again were deleted because the data for the month were incomplete. The remaining case I59236/5578 was compared to nearby stations and it was obvious that the value of 900.0 mm should be 0.0 mm
	H	1	I50921/5397 had one hourly value of 84.6 mm during April 1963. No other station nearby had hourly data, but a check of the daily value for I50921/5397 and nearby stations revealed that the day was dry. All other hours for the day were 0.0 mm so 84.6 mm was replaced by 0.0 mm.
	S	57	28 were changed to 0.0 mm, 4 were changed to a non-zero value, and 25 were deleted. Also seven daily amounts were corrected to an estimated amount in line with nearby stations.
90	D	5	All five were deleted.
	H	0	
	S	38	Nine were changed to 0.0 mm, 3 were changed to a non-zero value, and 26 were deleted: 22 of these were for Antarctic stations. Also seven daily amounts were corrected to estimated amounts after comparison with nearby stations.
70[1]	D	15	Only one was a New Zealand station: of the others 10 were Pacific and 4 Antarctic. Three had been deleted by the Antarctica process[1], which was done between running the check and doing these corrections, four others were deleted, seven were changed to 0.0 mm, and one was changed to a non-zero value.
	H	3	One had been accepted in an earlier run and the two new cases were also accepted because there was no other evidence to reject them.
	S	91	Two were accepted as correct, 29 were changed to 0.0 mm, 15 were changed to a non-zero value, and 45 were deleted, including 10 from Antarctic stations. Also 20 daily records were corrected to estimated amounts after comparison with nearby stations.

50	D	30	Four were accepted (3 New Zealand, 1 Pacific), and 26 (25 Pacific, 1 Antarctic) were deleted.
	H	9	Three had been accepted in earlier runs and two new cases were also accepted; four were deleted.
	S	185	106 were from New Zealand stations, 55 were Pacific, and 24 from Antarctica. Of the New Zealand ones, 16 were accepted, 43 were deleted, 32 were made 0.0 mm, and the remaining 15 changed to a non-zero value; also 20 daily records were either deleted or corrected to estimated amounts after comparison with nearby stations. Of the Pacific ones, 15 were accepted, 25 were deleted, and 15 changed to a non-zero value. All 24 from the Antarctic were deleted.
50[2]	D	6	Four were already accepted and two were deleted.
	H	5	All five were already accepted.
	S	61	28 were from New Zealand with 16 previously accepted, 22 from the Pacific with 15 previously accepted, and 11 from Antarctica. Of the 12 new New Zealand cases, 4 were accepted, 3 deleted, 2 made 0.0 mm, and the remaining 3 changed to a non-zero value. Of the 7 new Pacific cases, 1 was accepted, 4 were deleted, and 2 were changed to a non-zero value. All 11 new cases from Antarctica were deleted.

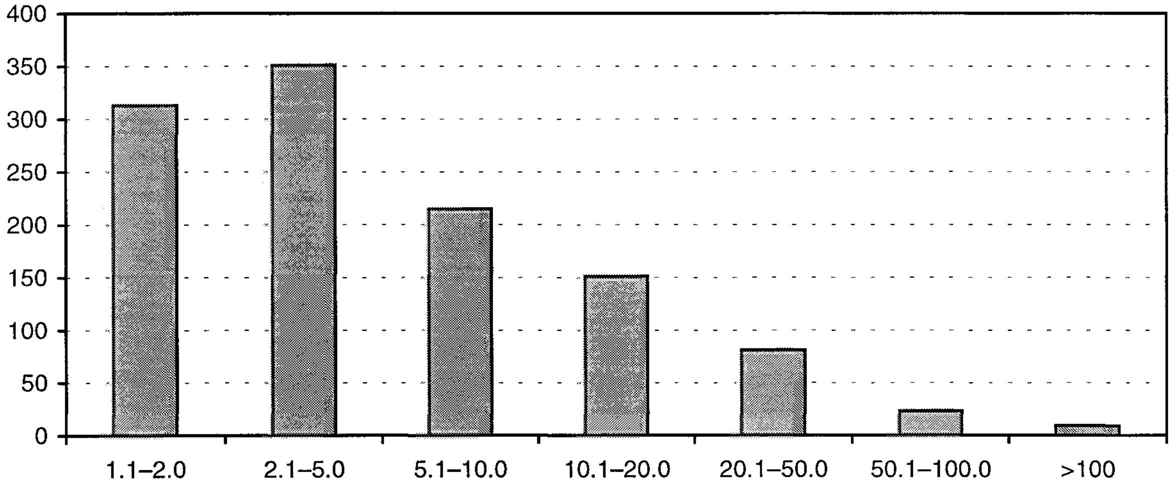
1. There seemed to be too many rainfall reports from Antarctica where, at the same time, the station records were often patchy. If a station regularly reported rainfalls then it was probably intentional, whereas those stations with rather patchy records may well have never intentionally reported rain, rather occasional rain reports may have occurred through errors in the synoptic messages. All rainfall records from Antarctica stations were examined to determine how many rows there were for each station and what percentage these were of the maximum number that could have been archived while the station was open. The shorter and patchier records were then deleted, i.e., 5445 FREQUENCY "D" rows from 30 stations and 2993 "S" rows from 23 stations. The remaining "D" records all had about 1000 rows and were about 50% complete. For the "S" records the worst became 525 rows and 19% complete and the best 8198 rows and 77% complete.
2. A second run with the maximum ratio set to 50 could be done because in the previous run some deletions were made and so new cases could arise — if all large values had been accepted then a second run would not have produced any new cases. Furthermore, the ratio was not lowered to 30, say, for this run since, although there would have been more incorrect values uncovered, there would also have been many more cases to check and many of those would have been acceptable values. Because of time constraints it was deemed prudent to stop after this second run since subsequent runs might well have continued to find a similar number of error cases in each run — the process would never end.

Details and results of Check C.3 — are all daily amounts consistent with the sums of the hourly or synoptic components?

For a given AGENT_NO and OBS_DATE the AMOUNT for a FREQUENCY "D" row should be consistent with the sum of the AMOUNTs from all FREQUENCY "H" rows with dates between OBS_DATE minus PERIOD and OBS_DATE. Thus, if 24 "H" rows are available then their sum should be equal to the "D"'s AMOUNT except the actual time of daily observations is only nominally 0900 Local and some discrepancy can occur if rain was present at that time. This discrepancy can be avoided by forming two sums from the "H" rows, an "outer" sum which includes the hours before and after the observation period of the "D" row and an "inner" sum which excludes the first and last hours of that period. The "D" AMOUNT should then be equal to or less than the outer, but equal to or greater than the inner. Where some of the "H" rows are missing a comparison with the outer sum cannot be made since the missing hours might have been wet enough to exceed the "D" AMOUNT, but an error is uncovered by a comparison with the inner sum even if only a single "H" row is present and it exceeds the "D" AMOUNT.

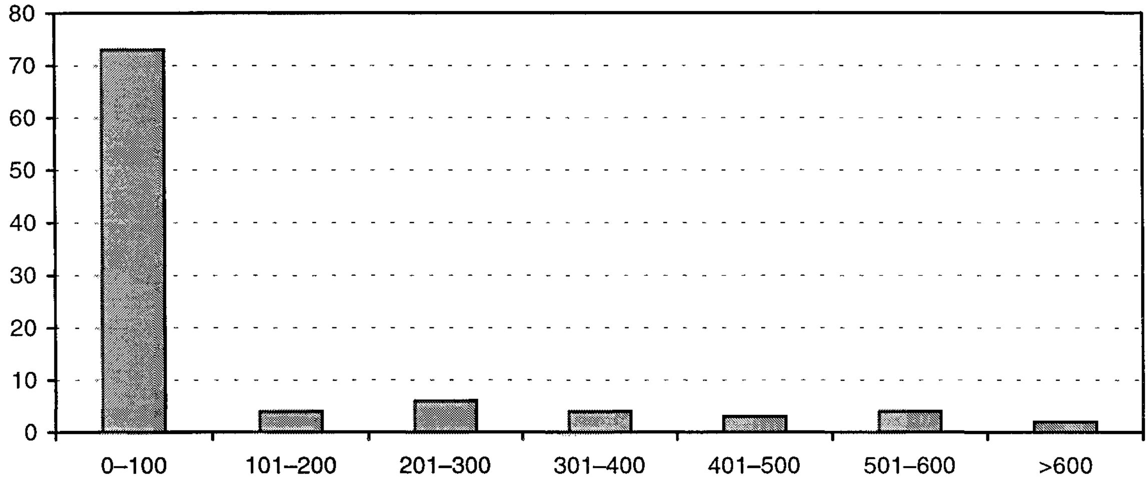
There were 12 184 differences between the daily observation and the sum of the component hourly amounts, but 9931 of these were for amounts of 1 mm or less and of the others A64878/1434 had 1112 differences. The large number for this station — the main Auckland City site — was found to be due to it never having had a gauge capable of giving hourly observations and the readings from A64871/1427 — Albert Park in central Auckland — had been used as an approximation. The hourly observations were regenerated from data in **RAIN_RATE** and archived under A64871/1427 and then the 209 639 "H" rows at A64878/1434 were deleted.

The remaining 1141 differences of over 1 mm, which were considered significant, were shared among 95 stations. The distribution of these differences using an approximate doubling scale for the class ranges is given in the figure below which shows that over half the differences were at most 5 mm.

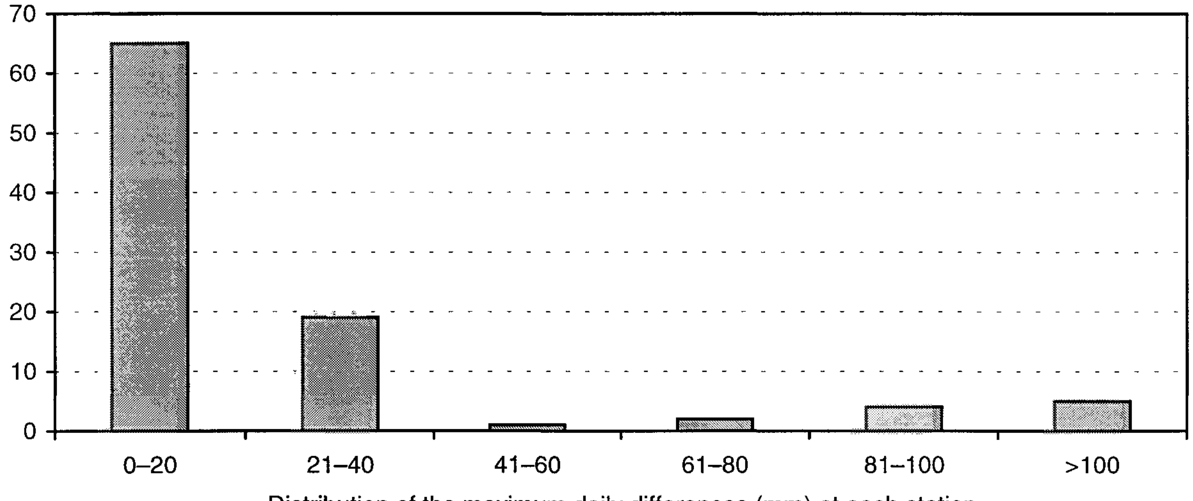


Distribution of differences (mm) between daily and hourly amounts

The total number of differences at each station and the maximum difference at each station are shown in the figure below. It can be seen that most stations had no more than 100 differences and the maximum was usually no more than 20 mm.



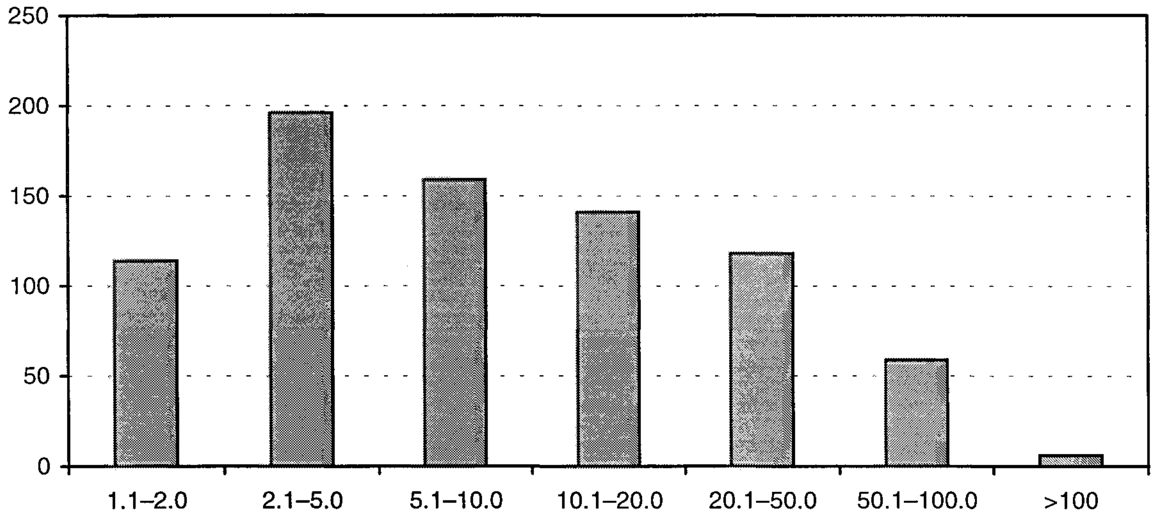
Distribution of the total number of daily differences at each station



Distribution of the maximum daily differences (mm) at each station

These significant differences needed to be resolved — was the "D" row correct? or did the sum of the "H" rows seem a better total? The main guidance for taking this decision was by a comparison with nearby stations, but for the six Pacific island stations involved there were no nearby stations. However, for one of these (J83000/6096) 141 of its 156 differences were due to all the "H" rows being zero but the equivalent "D" rows were not zero; the daily values were judged correct. In contrast, at J68000/6012 16 of its 55 differences were due to the "D" rows being on a different day to the "H" rows; the hourly values were judged correct. Now where the "D" rows were taken as correct the component "H" rows were deleted (12 934 "H" rows were deleted) and where the "H" rows were taken as correct, the "D" row's AMOUNT was changed to be the sum of the "H" rows' AMOUNTs (328 "D" AMOUNTs were changed).

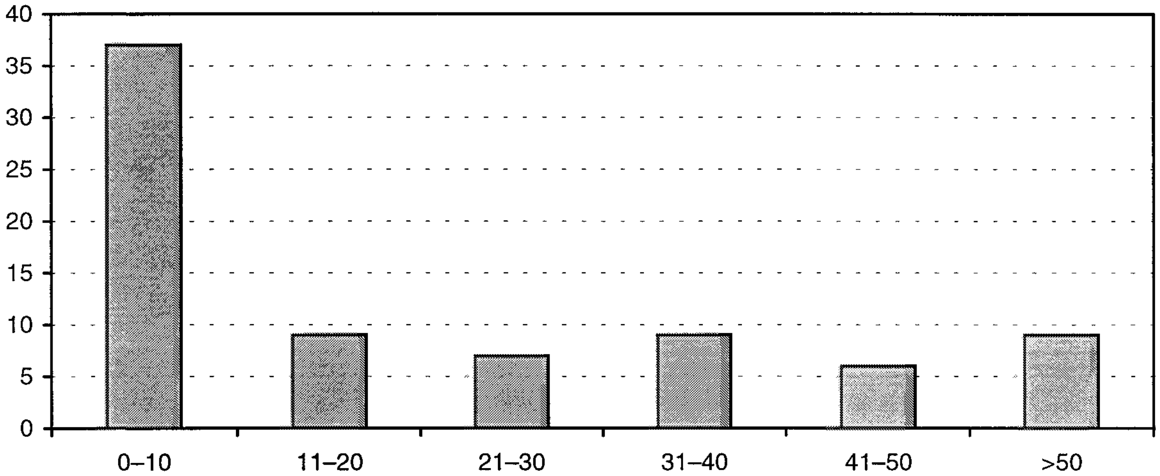
Similar considerations to those given in the first paragraph of this section apply to a comparison of "D" AMOUNTs and the sum of "S" AMOUNTs. In this case, except where 0900 Local is a synoptic reporting time, the measuring period of the "D" row is offset from the component "S" rows. Thus "outer" and "inner" sums are again required but when a row had a RELIABILITY of "?", it was not included.



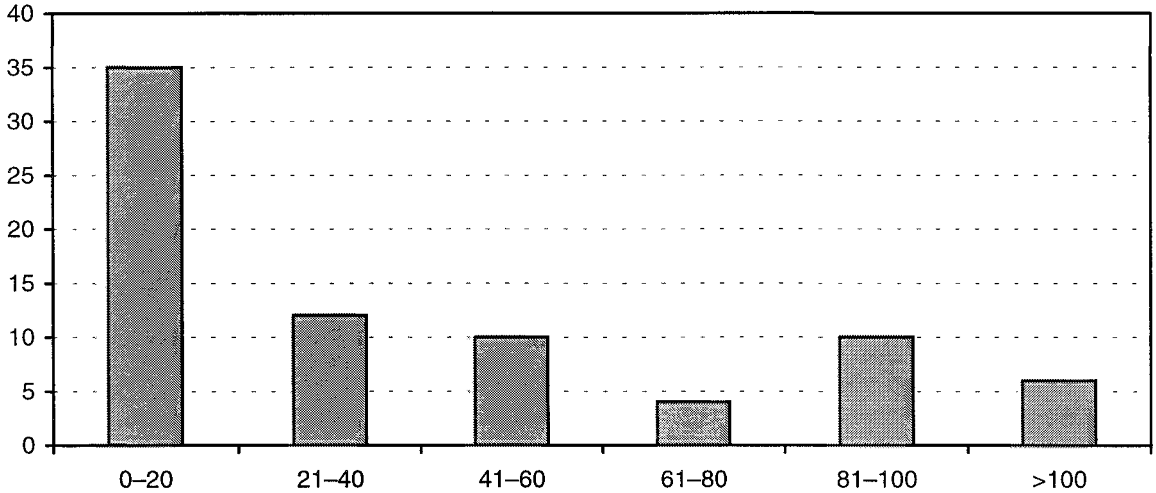
Distribution of differences (mm) between daily and synoptic amounts

There were 1630 differences between the daily observation and the sum of the component synoptic amounts, but 836 of these were for amounts of 1 mm or less. The remaining 794 differences of over 1 mm were shared among 77 stations. The distribution of these differences using an approximate doubling scale for the class ranges is given in the figure above which shows that nearly half the differences were at most 5 mm.

The total number of differences at each station and the maximum difference at each station are shown in the figure below: about half of the stations had no more than 10 differences and the maximum was often no more than 20 mm.



Distribution of the total number of synoptic differences at each station



Distribution of the maximum synoptic difference (mm) at each station

Again these significant differences needed to be resolved — was the "D" row correct or did the sum of the "S" rows seem a better total? Only 12 of the stations were from New Zealand and so little guidance from nearby stations was available and in general the "D" rows were taken to be correct. Now where the "D" rows were taken as correct the component "S" rows were deleted (2966 "S" rows were deleted) and where the "S" rows were taken as correct, the "D" row's AMOUNT was changed to be the average of the outer and inner sums of the "S" rows' AMOUNTs (9 "D" AMOUNTs were changed).

Details and results of Check C.4 — are all rainfall observations, 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 rainfall 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, there were 452 primary stations in New Zealand each with its furthest away buddy nearer than 5 km and whose buddies covered more than 95% of the primary station's daily rainfall record. At the other extreme there were 50 stations outside New Zealand for which the coverages were under 5% 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 54 such stations.

		Distance (km)										
N.Z.?	Cover(%)	<5	5–15	15–25	25–35	35–45	45–55	55–65	65–75	75–85	85–95	>95
Y	95–100	452	1 306	397	91	38	9	2	7	.	2	5
Y	85–95	27	103	46	9	1	2	2	.	2	1	.
Y	75–85	1	5	3	5	2	.	.	.	.	.	1
Y	65–75	.	2	5	1	1	.	.	.	.	.	.
Y	55–65	.	3	.	1	1	.	1	.	.	.	.
Y	45–55	.	.	3	.	.	.	.	.	.	.	.
Y	35–45	1	.	.	1	.	.	.	.	.	.	.
Y	25–35	.	.	.	.	.	.	.	.	.	1	.
Y	15–25	.	.	.	.	.	.	.	.	.	.	.
Y	5–15	1	.	.	.	.	.	.	.	.	.	.
Y	0–5	.	.	.	.	.	.	.	.	.	.	20
N	95–100	99	188	32	18	6	4	3	4	9	7	73
N	85–95	5	10	5	3	2	1	.	.	.	.	4
N	75–85	1	.	.	1	.	.	.	1	.	1	2
N	65–75	.	.	.	.	.	.	.	.	.	.	3
N	55–65	.	.	.	.	.	.	.	.	.	.	1
N	45–55	.	.	.	.	.	.	.	.	.	.	4
N	35–45	.	.	1	.	.	.	.	.	.	.	2
N	25–35	.	.	.	.	.	.	.	.	.	.	.
N	15–25	.	.	.	.	.	.	.	.	.	.	.
N	5–15	.	.	.	.	.	.	.	.	.	.	.
N	0–5	.	.	.	.	.	.	.	.	.	.	50

Some further statistics regarding the buddies are tabulated below.

Number of primary stations	3 076
Number with no buddies	54
Number with 1 buddy	3 022
Number with 2 buddies	1 476
Number with 3 buddies	620
Number with 4 buddies	279
Number with 5 buddies	125
Minimum % coverage	11

Average % coverage	98
Maximum % coverage	100
Minimum distance to buddy (km)	0
Average distance to buddy (km)	19
Maximum distance to buddy (km)	546

Having established a set of buddies, the largest contemporary difference was found for every primary-buddy pair and compared to the mean contemporary difference for the same primary-buddy pair, i.e., the ratio MaxDifference/MeanDifference was formed. There were 4240 primary-buddy pairs and the 5% with the largest ratios were examined since observations made when the ratio was largest at these 212 pairs were potentially the most likely to be errors. However, 41 of these were duplicates in the sense that the same station-date was repeated with another buddy. Thus 171 cases were examined by listing out from CLIDB the daily rainfalls for the station and its neighbours for a week either side of the given date. These lists were compared to the original paper forms.

A number of different types of error were seen: the original record differed from the CLIDB value — 78 cases; values were assigned to the wrong day or month or year — 30 cases (another 40 cases were noticed in the listings); days covered by an accumulation in a later day were in CLIDB as dry days — 14 cases; and values were assigned to the wrong station — 1 case. Often the error highlighted by the MaxDifference/MeanDifference ratio affected several days and not just the day when the ratio was largest, so a total of 519 changes to either AMOUNT or PERIOD were made, 97 rows were deleted, and 2 rows were inserted.

However, 48 of the 171 cases were accepted as correct either because there was not sufficient evidence that they were in error or because remarks made on the paper forms confirmed what might otherwise appear to be erroneous values. For example:

Station/ AGENT_NO	Date	Amount (mm)	Comment on paper form
From the first run			
A54623/1256	19660216	336.0	Torrential rain during night.
B65541/1506	19710416	259.1	Most of which fell 11pm–5am thunderstorm lasted 6–7 hours. Floods, slips and washouts on the roads and landslides on grass country.
B75261/1543	19810412	405.5	Some flooding.
B75261/1543	19810413	157.7	Gales. Excessive flooding.
D06062/2499	19530127	351.5	The 'Middle Road' and Tamumu both recorded falls of 12 inches the bulk of which fell between 5pm and 7am.
D15352/2672	19290515	298.2	Bad floods. NE to Southerly.
D96471/2974	19440304	268.5	Poured in evening and night with rapid rise and fall of river. Much destruction of fences and crops etc. River rose 16 feet
D14481/2594	19391226	387.4	Rainfall on 26th did much local damage to roads and water courses also much damage was done to Mr Riddiford's homestead at Orongorongo.
F12162/3841	19540306	439.4	Overcast, low cloud, heavy rain.
H23611/4556	19230506	365.8	13 inches in 12 hours.
H23611/4556	19230507	258.3	Abnormal floods and landslides.
H32891/4949	18950625	247.7	The fall on the 25th June was the greatest since the rain gauge has been at Akaroa.
From the second run			
A53121/1037	19960718	136.5	July record. 132.9 mm fell 12.30pm–3pm.
A53442/1109	19990121	210.5	Heavy rain most of which fell 3pm–8pm.

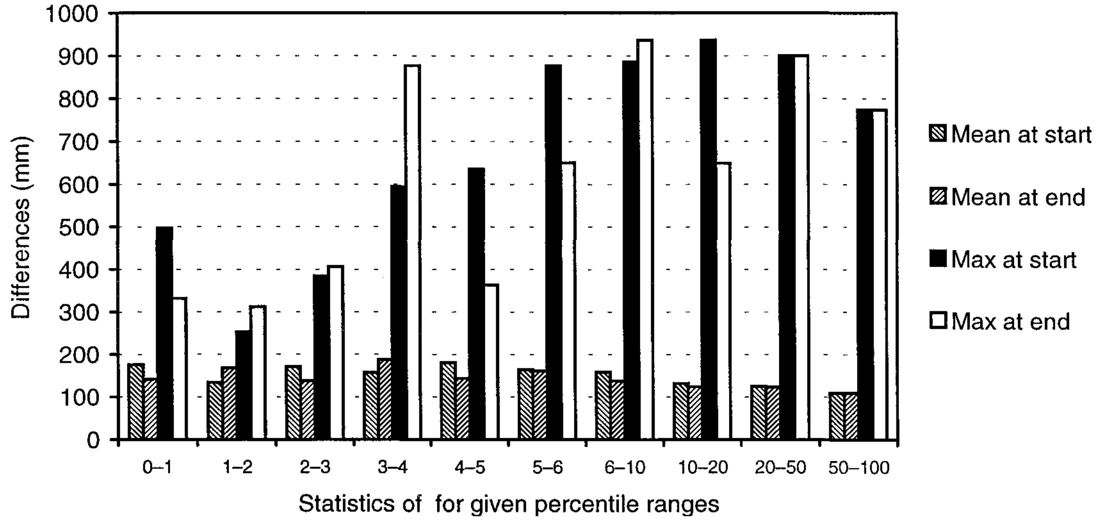
B75381/1550	19380204	419.1	There was a record rainfall of 16.5 inches in the 24 hours. This heavy rainfall was general in the district but the phenomenal fall was very local — the School of Mines distant about 3/4 mile registered something over 12 inches.
C75612/2094	19890503	126.5	Torrential rain.
H40272/4999	19521201	188.0	Heavy local damage by floodwaters.

Because some inspected values were correct, when this checking procedure is performed again, they will 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 was done through **RAIN_DIFFS** which was created by this checking procedure and has the following structure:

Column name	Null?	Type
AGENT_NO		NUMBER
BUDDY		NUMBER
DIST		NUMBER
OBS_DATE	NOT NULL	DATE
PERC		NUMBER
P_AMOUNT		NUMBER
B_AMOUNT		NUMBER

For each AGENT_NO and BUDDY the values with the greatest difference occurred at OBS_DATE and are held in P_AMOUNT and B_AMOUNT, while PERC holds the percentile of this set's maximum to mean difference. For example, those with PERC equal to 1 are the 1% of all the AGENT_NO, BUDDY pairs which have the greatest relative difference. Thus, on a re-run the contents of **RAIN_DIFFS** can be moved to **OLD_RAIN_DIFFS**, say, before being over-written and rows common to both tables (except PERC which may change between runs) can be ignored.

A second search for excessive maxima was made and the 48 cases that had already been accepted in the first run were filtered out using **RAIN_DIFFS** and **OLD_RAIN_DIFFS**. In this run there were 136 cases of which 49 were accepted, 38 found to be on the wrong day, 11 due to accumulations, and the remainder with values different from those on the paper forms. As a result, 464 changes to either AMOUNT or PERIOD were made, 56 rows were deleted and 1 row was inserted.



For percentile ranges of the MaxDifference/MeanDifference ratios at the station-buddy pair the figure above shows the mean and maximum of the actual differences both before and after changes due to this check were made to **RAIN**. The check seems to have made little difference since the mean maximum difference for all classes remains about 150 mm for up to the 10 percentile of the MaxDifference/MeanDifference ratio and about 100 mm otherwise. Also the maximum difference remains about 400 mm for the lower classes and about 700 mm for the others. This figure underlines the importance of comparing the maximum difference at a station-buddy pair with the mean of their differences since to be in the higher percentile classes the largest maxima must occur where the mean differences are also large. Such cases may be in error but are harder to highlight for checking, whereas for station-buddy pairs where the mean difference is small the maximum difference need not be so large for a potential error to be found. If absolute maximum differences had been used as the criteria for potential errors, then a greater proportion would have been found acceptable and the errors associated with the smaller maxima would have been missed.

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

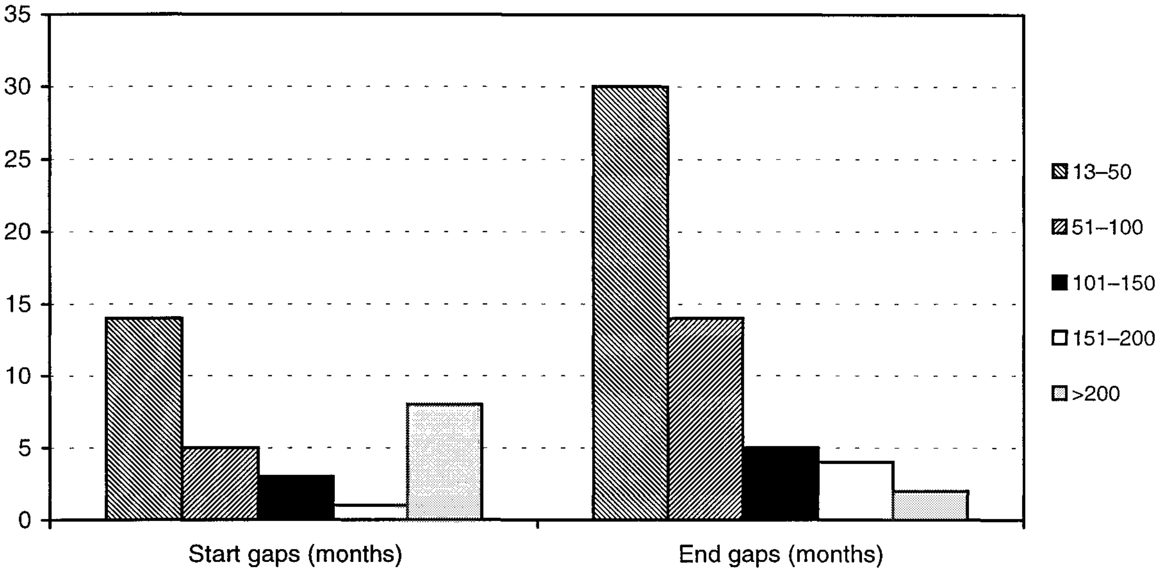
Ideally, for a given AGENT_NO there should be no breaks in the record for each FREQUENCY from when it started until either the present day or when the station closed. This is extremely rare since missing data occur at even the best stations. Thus, rather than a search for errors, this check is more a quality check in which the "completeness" of the station records is examined.

During this exercise it was found that the program that calculates the percentage completeness for daily observation records and places the result in **LAND_DATA_CAT** was in error. To determine the total period covered it was counting only the number of daily observation rows, and assuming these were all 24 h long, rather than summing the PERIODs from these rows. Thus, the completeness was being underestimated since, as is described in Check D.2, accumulations over more than 1 day are common. Together with others for cataloguing wind, evaporation, maximum temperature, and grass minimum temperature, the program was altered so that it would give a correct estimate of the completeness. There were many records previously noted as only 5–10% complete showing an increase to 40–50%.

However, apart from this incorrect assessment of completeness, there were still many stations where it 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.

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 over a year are included. For the shorter gaps, there were 95 at the start of records and 205 at the end bringing the total number of such gaps to 126 and 260 respectively.

Of the 33 stations with "Start gaps", the rows for the 15 which had a full month's data before the gap and those for the 12 which had just a few days before the gap, were all deleted, a total of 474 rows. For one station, I49613/5217, the data had already been re-archived under I49711/5223 and it is possible that some of the other deleted data should be re-archived under another station. If at some later stage this is found to be necessary then the deleted data can be recovered from **AUD_RAIN**. Apart from the deletions to **RAIN**, 13 of the warnings inserted into **SITE_CHANGES** during the B.1 check were deleted.

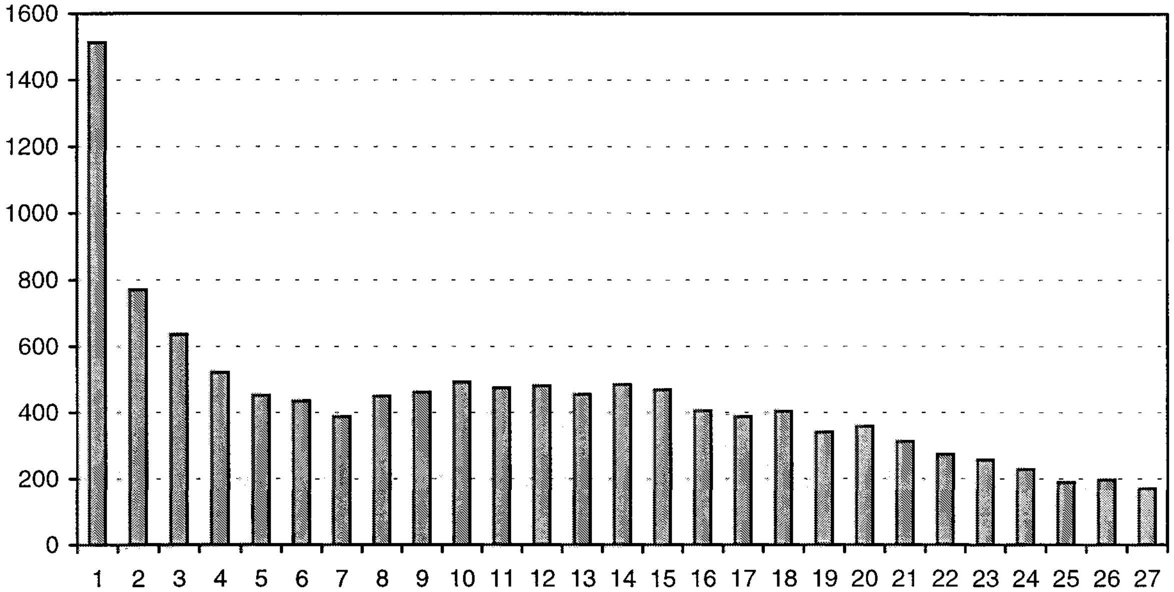


Distributions of gaps at the ends of station records

Of the 55 stations with "End gaps", the rows for the 28 which had a full month's data after the gap and for the 25 which had just a few days after the gap, were all deleted, a total of 880 rows. For two stations, C94001/2276 and E14181/3355, the data had already been re-archived under C94004/2279 and E1418E/3352 and, as before, if necessary at some later stage deleted data can be recovered from **AUD_RAIN**. Apart from the deletions to **RAIN**, 16 of the warnings inserted into **SITE_CHANGES** during the B.1 check were deleted and 15 station closing dates were changed in both **SITE_CHANGES** and **LAND_STATION**.

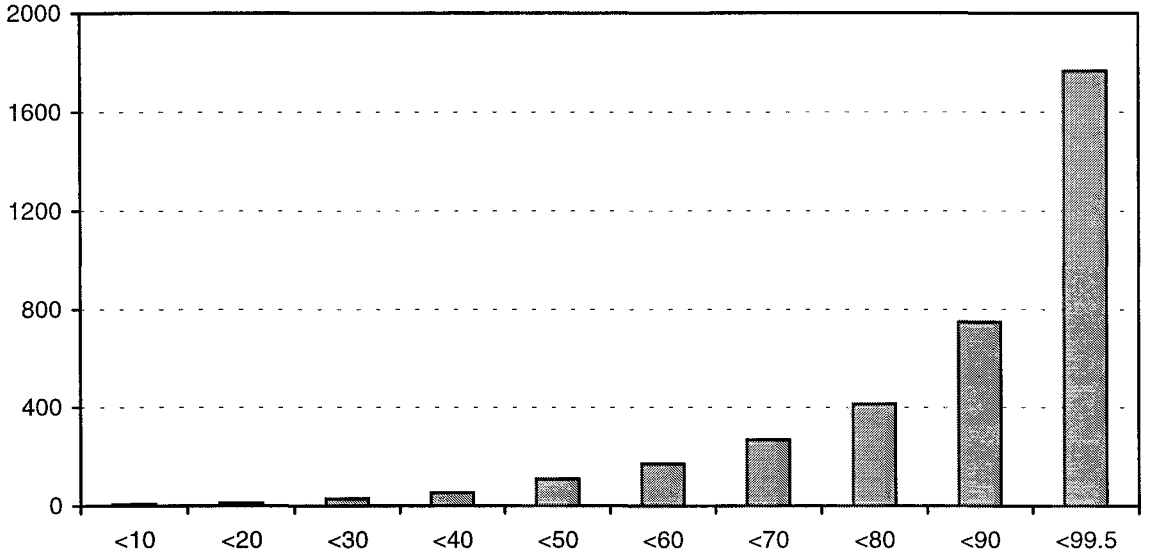
Similar effects might also apply to the hourly observation records. It was found that for each of 28 stations a single hourly report at 0000 UTC on 1 May 1992 preceded a gap of 31 months: the 28 rows were deleted. After dealing with the gaps near the ends of the records, those records which were less than 20% complete were individually investigated. There were 11 such records and they were all for daily observations. Of these, 4 were for skifield stations which reported only occasionally through the winter and their observations were largely retained (only 16 deleted) as they are from areas where the rainfall network is particularly sparse. The 17 observations from the other stations were all deleted. Two of the warnings inserted into **SITE_CHANGES** during the B.1 check were deleted and two station closing dates were changed in both **SITE_CHANGES** and **LAND_STATION**.

The remainder of this section is a description of the state of the daily and hourly records after the changes described above had been made. The figure below is for daily observations and shows the distribution of missing days per station-month where it should be noted that a day is not considered missing if it is covered by a subsequent observation whose PERIOD extends back to cover that day. There were about 1500 station-months that had a single day missing, nearly 800 with two days, which could be either together or apart, etc. These numbers are small when compared to the total number of daily observations, which is equivalent to 840 000 station-months. The counts for the 28–31 day classes are not included below because they are much larger at 3488, 1381, 18 277, and 31 422. 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 54 000 whole station-months are missing.



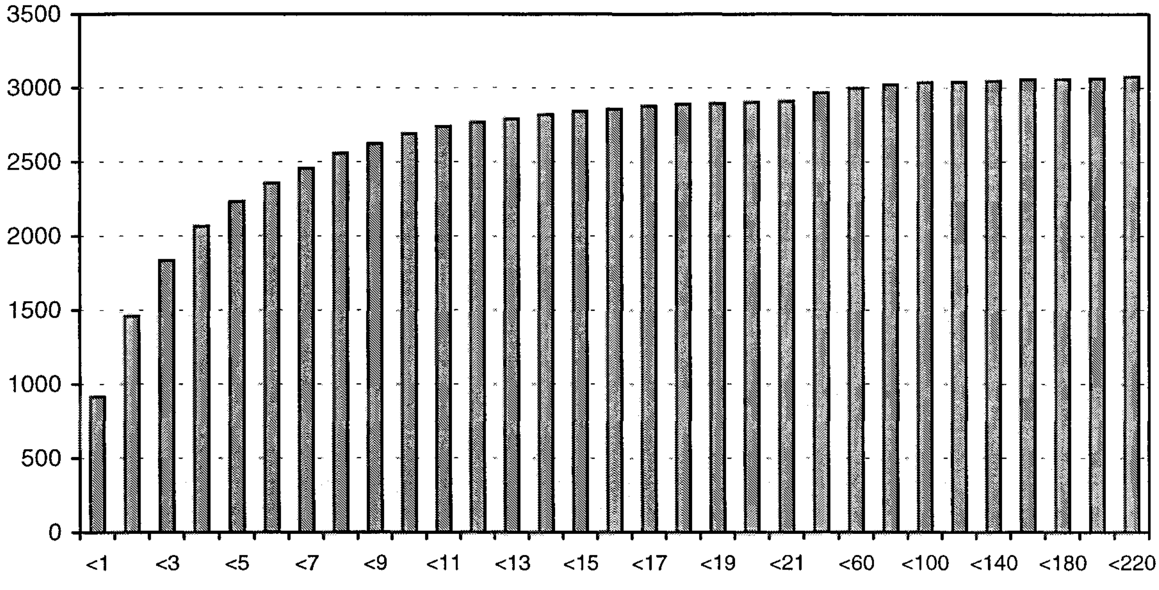
Distribution of missing days per station-month

How are these missing days spread among the stations? CLIDB has 3076 stations with records of daily rainfall observations with 1310 of these having near perfect records and the cumulative distribution of the percentage complete of the records at the other 1766 stations is shown below. It can be seen that under 800 of these have records less than 90% complete and only about 100 records are under 50% complete.



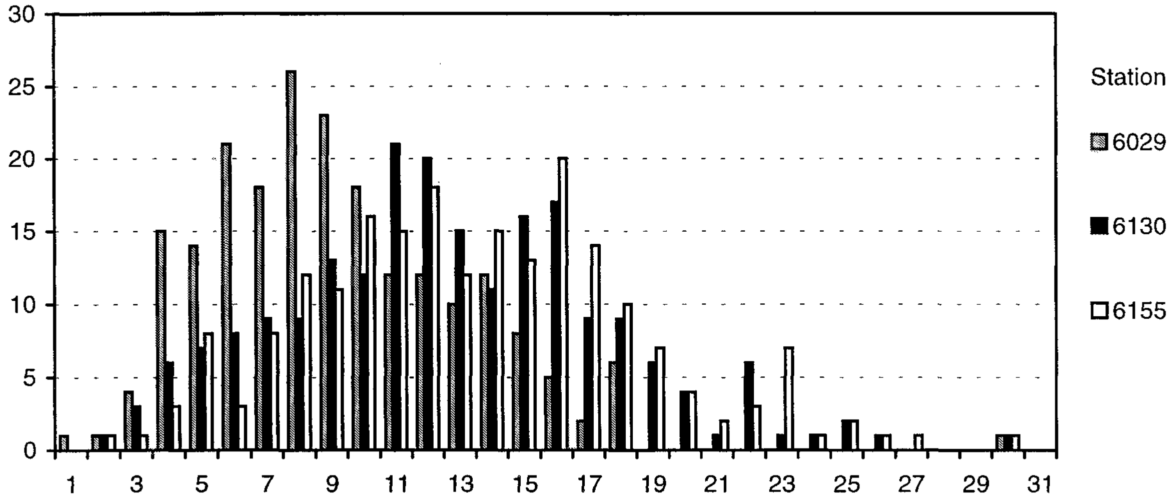
Cumulative distribution of percentage completeness of daily rain records

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 big 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 913 without any gaps in their daily rainfall records (i.e., those in the <1 class in the figure below), but stations with only a few gaps are also of high quality and if all the single gaps were filled then nearly half of the stations would be 100% complete. There is a change of class width after the <21 gap class.



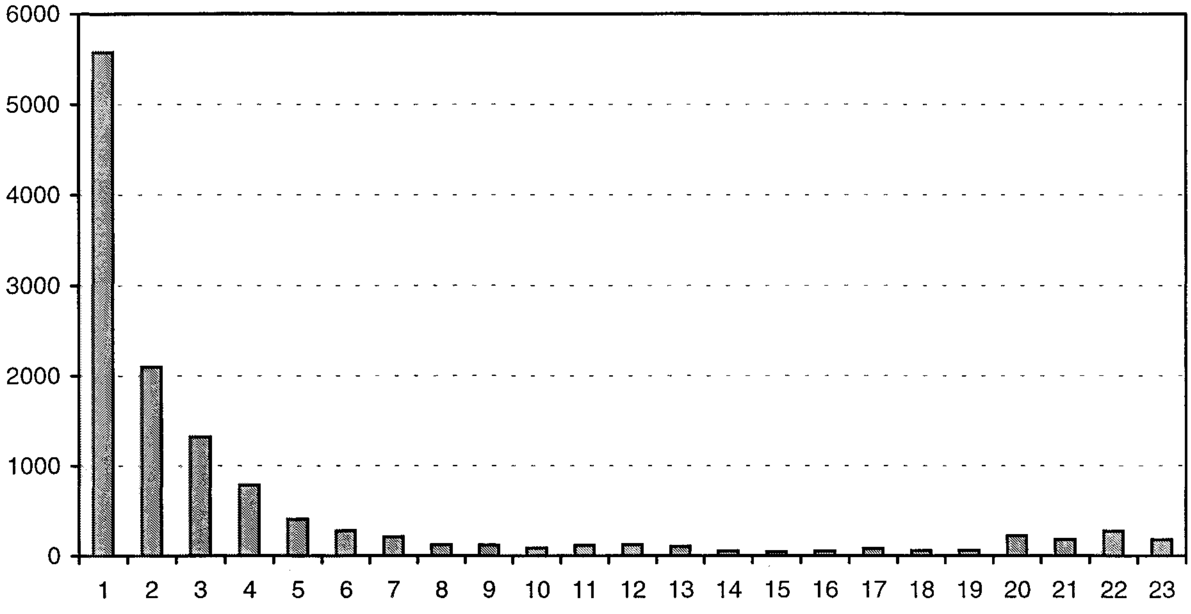
Cumulative distribution of the number of daily gaps at each station

The three stations sharing the most gaps were J75300/6029, J92500/6130, and J94300/6155 each with 209 gaps. The distribution of gaps for these stations is shown in the figure below.



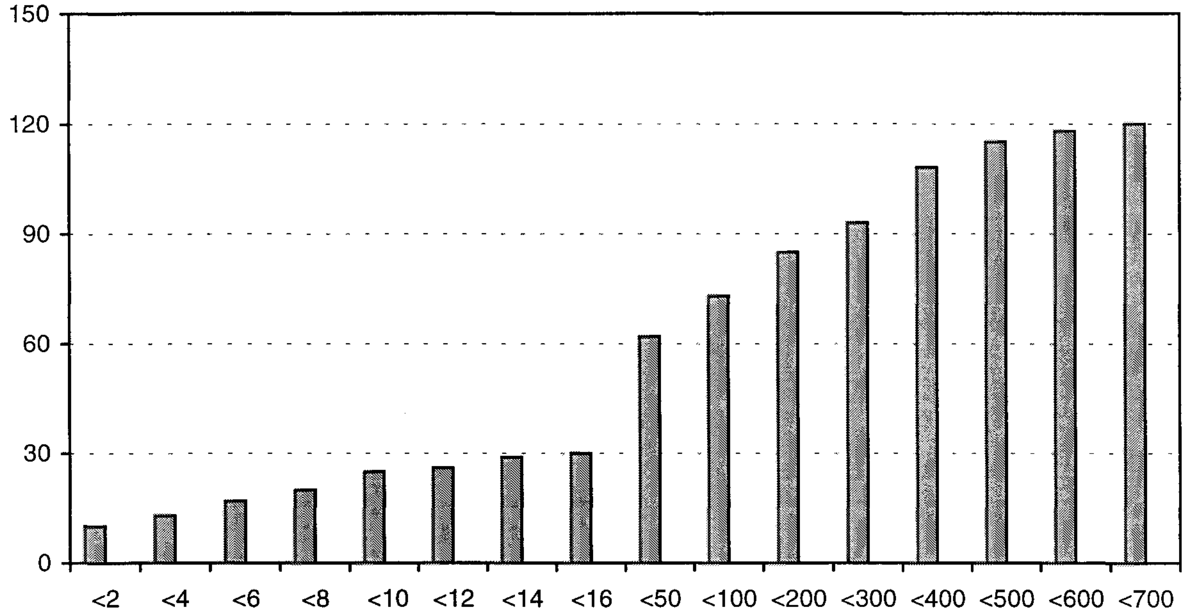
Distributions of gap sizes (days) at the three worst stations

Synoptic records were not examined for completeness since few, if any, would be complete, but records of hourly observations were examined and the figure below is for hourly observations and shows the distribution of missing hours per station-day. There were about 5500 station-days that had a single hour missing, about 2000 with 2 hours, which could be either together or apart, etc. Even these numbers are relatively small when compared to the total number of hourly observations in CLIDB, which is equivalent to 324 682 station-days, and the numbers for over 6 hours missing are smaller still. The count for the 24 h class is not included below because it is much larger at 15 565 and represents the number of whole station-days that are missing.



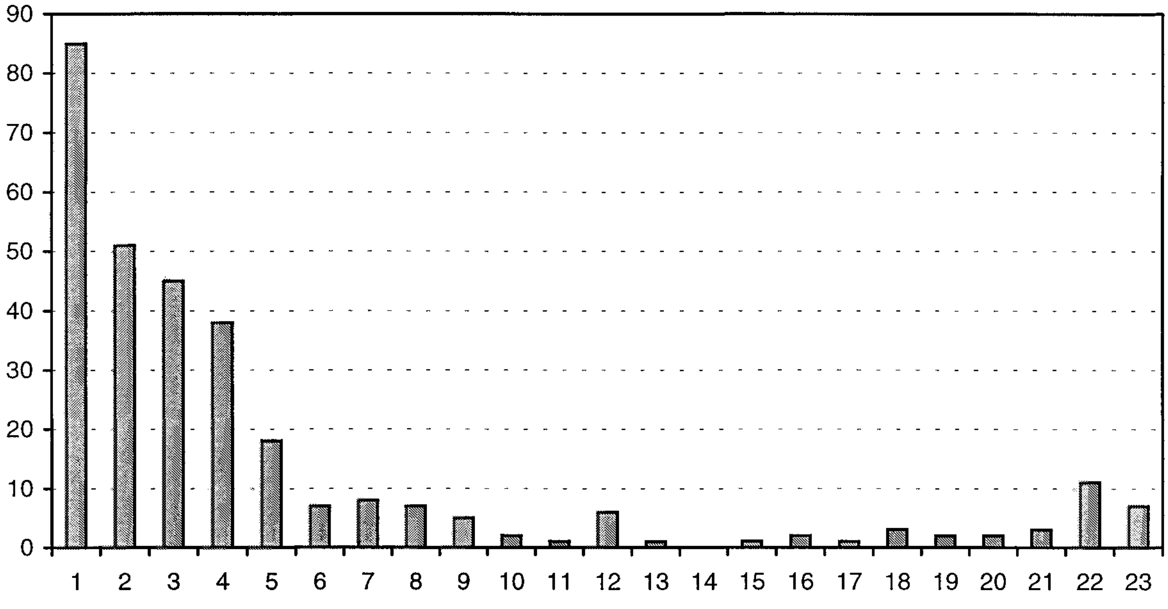
Distribution of missing hours per station-day

How are these missing hours spread among the stations? CLIDB has 120 stations with records of hourly rainfall observations with 61 of these having near perfect records. Of the other 59 stations, only 7 have records less than 80% complete with the least completeness being 55%. The worst stations are, of course, those with the lowest percentage completeness but, as with the daily records, those with a large number of gaps, rather than just a low percentage complete, are also of poor quality. There were no stations without any gaps in their hourly rainfall records, but 30 stations have fewer than 16 gaps and under half the records have over 50 gaps. The figure below shows the distribution of hourly gaps; there are changes of class width after both the <16 and <50 gap classes.



Cumulative distribution of the number of hourly gaps at each station

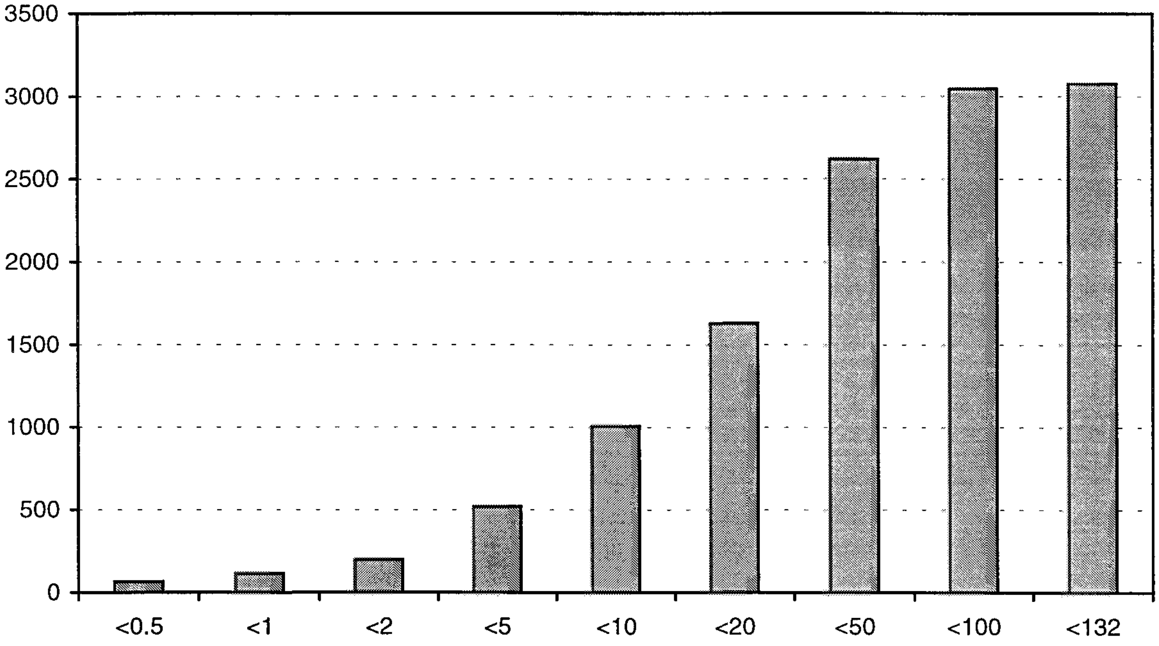
The station with the most gaps was D9668B/3015 with 627 gaps. The distribution of gaps for this station is shown in the figure below except the 321 complete days missing have been omitted.



Distribution of gap sizes (h) at station D9668B/3015

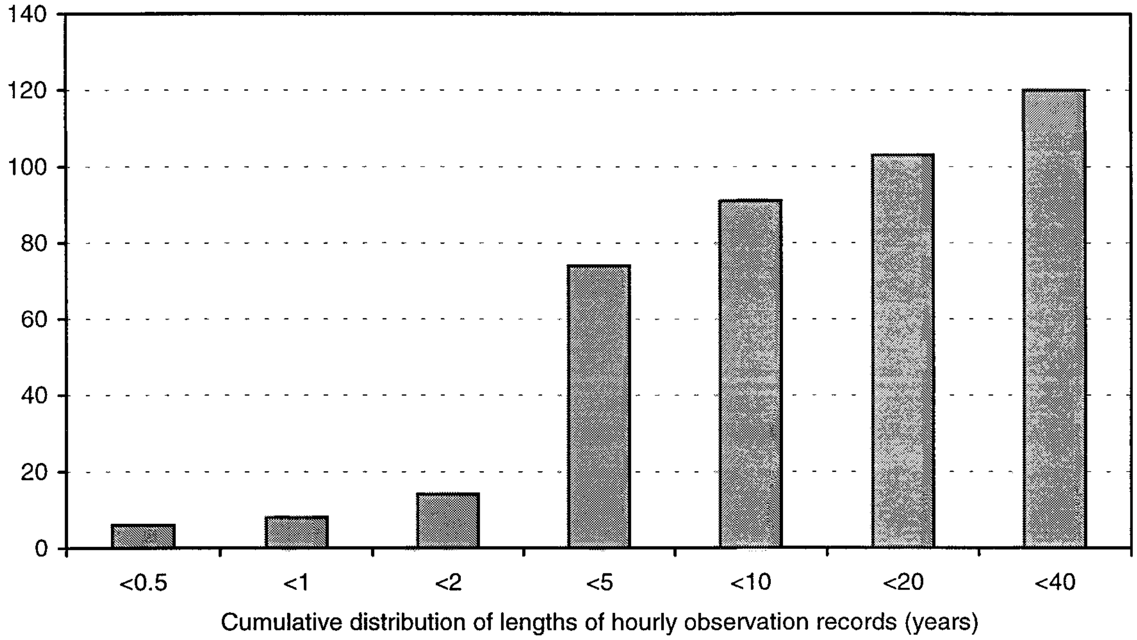
Details and results of Check D.1 — are all rain records long enough?

For a given AGENT_NO and FREQUENCY the record should be long enough to establish the mean level and variability of the rainfall with respect to the place and frequency concerned. Longer records can be used to track any trends in the rainfall, 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 daily observation records.



Cumulative distribution of lengths of daily observation records (years)

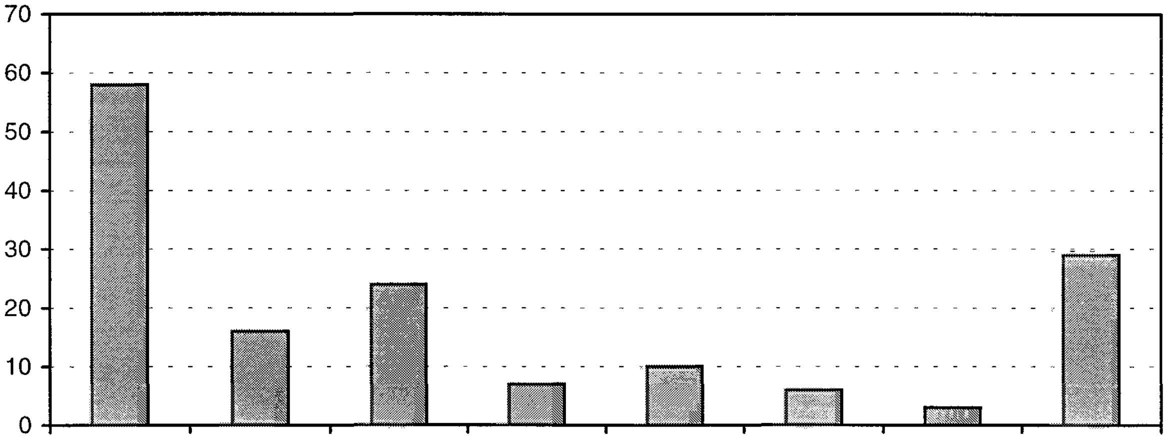
Only about 100 records are under a year long and about half of the 3076 records are over 20 years long. The longest record is from A64871/1427, which started in January 1863 and is 96% complete — monthly rainfall totals for an earlier 10 years are also available for that station. Of those records under a year which were not from stations which had opened within the last year, 21 had no more than 8 rows. These records were discarded by 38 rows being deleted from **RAIN**. From **SITE_CHANGES,** 12 rows regarding data outside the station's lifetime were also deleted, and in both **SITE_CHANGES** and **LAND_STATION** 12 opening or closing dates were changed.



The distribution of the record lengths for hourly observations is shown in the figure above. Less than 10 records are under a year long and nearly half of the 120 records are over 5 years long. The longest record is from E14387/3445, which started in October 1960 and is 77% complete. Of those records under a year which were not from stations which had opened within the last year, four had rows from just one day. These records were discarded by 25 rows being deleted from **RAIN**.

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

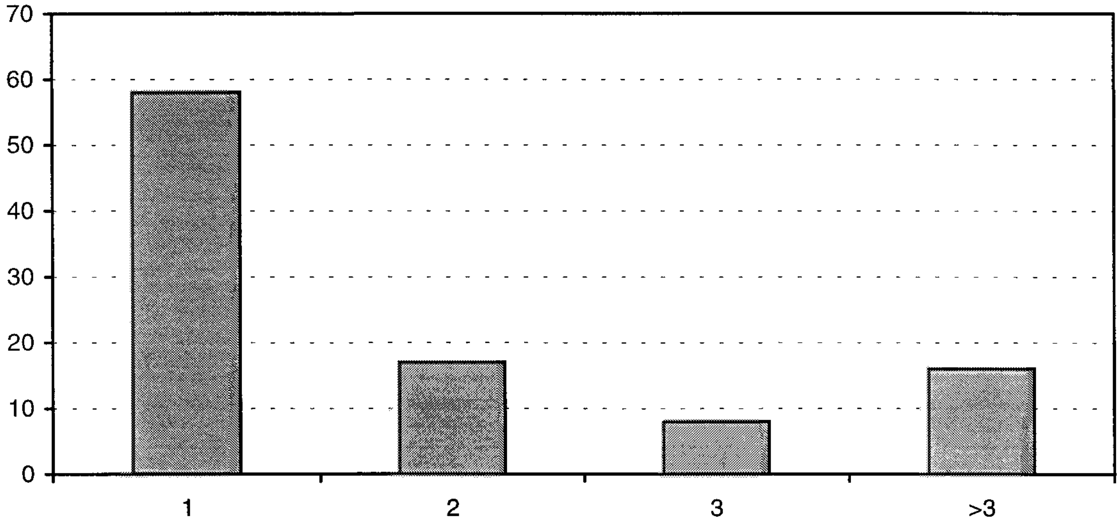
For a given AGENT_NO and FREQUENCY of "D", the number of times that an observation applies to more than one day should be small compared to the total number of daily readings for that AGENT_NO. Few stations have no accumulations, but it was found that for 95% of the stations at most 20% of their daily observations are accumulations. The percentage of accumulations for the other 5% of stations (161 stations) is distributed as in the figure below which shows a decrease of numbers as the percentage increases. However, there are still about 30 stations with 91–100% of their daily observations being accumulations, and ten of these had records that were over 99% accumulations.



Distribution of percentages of records that are accumulations

The maximum number of days of accumulation was 31 which occurred at 743 stations, but only at 107 of the 161 stations covered by the figure above. These are due to stations providing a total for the month but no daily rainfalls. The stations involved in this practice were often private rainfall stations, octapent gauge stations which were visited irregularly, or Pacific Island stations which sent in an annual summary with totals for each month. In **RAIN** the monthly total for these approximately 600 station-years of accumulated daily observations is given on the last day of the month, along with a PERIOD covering the full month.

Again for a given AGENT_NO and FREQUENCY of "D", few, if any, accumulations should be cross-month accumulations, i.e., those that start in one month and end in another. The figure below shows their distribution among stations Thus, about 60 stations had a single occasion when a daily observation had a PERIOD such that the start of the period was in the month before that of OBS_DATE. A total of 99 stations was involved with a total of 505 cross-month accumulations. Eleven of the 16 stations that had over 3 of them were from the Cook Islands with a total of 327 cross-month accumulations.



Distribution across stations of the number of cross-month accumulations

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

From the FREQUENCY "D" rows monthly summary statistics are calculated and entered into **MTHLY_STATS**. The statistics concerned are: total rainfall; number of days with at least 1 mm of rain; number with at least 0.1 mm; and the maximum one day fall. There are certain rules associated with their calculation which ensure the statistics are valid and are exactly as defined. For example, if for a set of FREQUENCY "D" rows with the same AGENT_NO and with all OBS_DATEs falling within the same local month, the maximum AMOUNT was associated with a PERIOD of over 24 h, then no maximum one day fall can be found for that month. However, PERIODs of over 24 h in the set of rows do not preclude the extraction of the one day maximum since it is only required that the maximum AMOUNT was with a 24 h PERIOD.

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

The errors reported that might indicate errors in **RAIN** are: extra days, which are seen as a negative number of missing days and result from the sum of the PERIODs of the rows for the concerned month exceeding the length of the month; rows exist where nothing is recorded for AMOUNT; a cross-month accumulation; despite an error a non-deletable statistic exists; and data with an origin not normally associated with **RAIN**. It is also possible that any cross-month accumulations had already been accepted and were the reason for the first type of error, so only the separate occurrences of either extra days or cross month accumulations are reported. The following reportable errors occurred.

AGENT_NO	Error	Count
3438	Extra days	1
4372	Extra days	1
5934	First wrong	1
6404	Extra days	1
14629	Extra days	1

Apart from that for AGENT_NO 3438, the errors were all cross-month accumulations that had occurred since Check D.2 had been performed. For E1438B/3438 the error occurred for January 1998 when an accumulation over 9 days also had rows with nil AMOUNTs within the accumulation period; the eight extra rows were deleted.

Details and results of Check D.4 — are the rainfall totals in RAIN_RATE consistent with the equivalent daily observations?

RAIN_RATE holds breakpoint data that are digitised off the pluviographs from Dine's tilting siphon automatic raingauges or are extracted from the high temporal resolution drop style gauges (i.e., the Hydras and RIGs). The Dine's gauges are always collocated with a manual gauge whose observations are archived in **RAIN**, and furthermore the pluviograph charts are changed every day around the time that the manual gauge is read. Thus, the total rainfall represented by the pluviograph, which is stored in TOTAL in **RAIN_RATE**, and the manual total (AMOUNT in **RAIN**) should be the same. Sansom

& Penney (1999) found and corrected many inconsistencies between **RAIN** and **RAIN_RATE** and set up the regular checking for consistency for new data. This check was essentially the same as that for new data but checked all the rows in **RAIN_RATE**.

There are four parts to the checking.

1. If for some AGENT_NO and OBS_DATE in **RAIN_RATE** the equivalent row in **RAIN** is missing, then such a row is inserted: no such inserts were made.
2. On the other hand, if for some AGENT_NO and OBS_DATE in **RAIN** the equivalent row in **RAIN_RATE** has STATUS of 1 or 2 (i.e., no pluviograph was available) and AMOUNT and TOTAL disagree, then TOTAL is changed to AMOUNT: 319 such changes were made.
3. If AMOUNT and TOTAL differ by at most 0.5 mm then AMOUNT is changed to agree exactly with TOTAL: such changes may well have been made but they were not counted.
4. If AMOUNT and TOTAL differ by over 0.5 mm, then the pluviograph involved is examined and a decision made as to whether **RAIN** or **RAIN_RATE** is correct: 43 such discrepancies were found. Of these 28 AMOUNTs were correct and so the pluviographs were redigitised, and, since the other 15 TOTALs were correct, the AMOUNTs were changed to agree.

The changes of 2. were made by creating pseudo-digitised data which looked like those from the digitising system when a pluviograph is missing or of such poor quality that it cannot be digitised. Before this check was made such pseudo data always resulted in STATUS 2 records even if there had been no rain on the day concerned, but on such days, with no temporal structure to the rainfall, a STATUS 0 record could equally well be created. The changes of this check were made in this way, the regular consistency checking program was changed so that future changes would be made this way, and the 13 659 STATUS 1 or 2 days with a TOTAL of zero had their STATUS changed to 0. While making these changes, it was noticed that the last of the changes to be made through pseudo data was being dropped; the program responsible was changed.

Summary and Conclusion

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

- The programs which process the raw synoptic data were changed so only rainfall at the main synoptic hours is accepted and the associated period must be 6, 12, 18, or 24 h and if 24 h then an attempt is made to create a FREQUENCY “D” row rather than an “S” row.
- The scripts which populate **LAND_DATA_CAT** with rows concerning wind run, maximum gusts, evaporation, maximum temperature, and grass minimum temperature as well as rain were amended to allow for accumulations, i.e., a true percentage completeness is now calculated by summing the PERIODs in the base tables rather than just counting the rows.
- The monthly consistency check between **RAIN** and **RAIN_RATE** now ensures that any dry days with missing charts are archived with a STATUS of 0 rather than 2.

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

Table name	Inserts	Deletions	Amendments
RAIN – Daily	96 220	8 466	2 214
RAIN – Hourly	0	12 993	271 144
RAIN – Synop	0	1 338 500	389 046
RAIN – Total	96 220	1 359 959	662 404
LAND_STATION	0	0	40
SITE_CHANGES	502	0	29
RAIN_RATE	0	0	13 978
Total	96 722	1 359 959	676 451

The grand total of changes made to **RAIN** was 2 118 583, which is 5.7% of its total number of rows. However, when broken down by FREQUENCY the percentages of rows changed are 0.4% for "D" rows, 3.6% for "H" rows, and 45.0% for "S" rows. These percentages depend on just a few large numbers; thus

- for "D", the 96 220 inserts derived from reclassifying "S" rows with a PERIOD of 24 h as "D" rows and the deletion of 5445 rows from Antarctic stations contributed the most.
- for "S", the largest contributions were the deletion of 439 037 rows with times not at the standard reporting times, another 731 015 deletions of rows where contemporary "H" rows were also available, and 359 604 amendments to correct, or mark, the data recovered with errors from Trentham.
- for "H", the largest contributions were the transfer of 209 639 rows from the Auckland City station to Albert Park and the adjustment of OBS_DATE on 61 361 which had mistakenly been given an NZDT time rather than an NZST time.

The need for these extensive 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.

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. 1999: New Zealand's National Climate Database (CLIDB): audit report on the MTHLY_STATS Table. *NIWA Technical Report 62.* 38 p.

Tomlinson, A.I. 1980: The frequency of high intensity rainfalls in New Zealand. *Water & Soil Technical Publication 19.* 36 p.

WMO 1995: Manual on Codes. International codes Vol. 1 Part A. *WMO 306.*