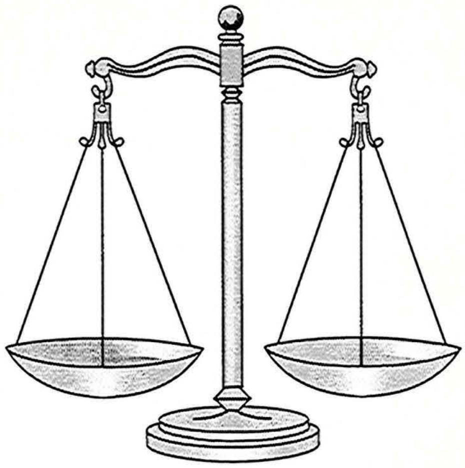
# New Zealand's National Climate Database (CLIDB): audit report on the MTHLY_STATS table

John Sansom
Allan Penney



NIWA Technical Report 62
ISSN 1174-2631
1999

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

John Sansom
Allan Penney

NIWA Technical Report 62
1999

**Published by NIWA**
**Wellington**
**1999**

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

**ISSN 1174-2631**
**ISBN 0-478-08483-8**



Citation: 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.

*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


| | Page |
|---|---|
| Abstract | 5 |
| Introduction | 5 |
| Typographical conventions | 6 |
| The DATA_AUDIT table | 6 |
| The MTHLY_STATS table | 7 |
| Summary of checks | 7 |
| Audit results | 9 |
| Details and results of Check A.1 — are all stations valid? | 9 |
| Details and results of Check A.2 — are all codes valid? | 9 |
| Details and results of Checks A.3, A.4, and B.7 — are all reliabilities and origins valid? | 9 |
| Details and results of Check B.1 and B.2 — are all dates valid? | 11 |
| Details and results of Check B.3 — are all records within the life of the station? | 12 |
| Details and results of Check B.4 — are all extreme dates within the month of the data? | 13 |
| Details and results of Check B.5 — are all statistics values reasonable? | 14 |
| Details and results of Check B.6 — are all sources for statistics reasonable? | 16 |
| Details and results of Check C.1 — are all the smallest and largest statistics values reasonable? | 17 |
| Details and results of Check C.2 — are all statistics within a group consistent? | 18 |
| Details and results of Check C.3 — are all statistics, when compared to nearby stations, reasonable? | 23 |
| Details and results of Check C.4 — are all statistics records without gaps? | 29 |
| Details and results of Check D.1 — are all statistics records long enough? | 34 |
| Details and results of Check D.2 — are all monthly and annual statistics consistent? | 37 |
| Summary and Conclusion | 37 |
| Reference | 38 |

# Abstract

**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.**

The auditing of the dataset within New Zealand's National Climate Database which contains monthly climate statistics, such as total rainfall accumulation or mean temperature, is described. Each record in the dataset consists of the place of observation, the month and year of observation, the type of data observed, the data itself, and some minor attributes. All the attributes were checked individually and in groups so that any invalid values were found; consistency between different data types 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 for each data type was assessed through the number of years of observation and consistency of reporting during those years.

Errors were found and these occasionally arose through errors in other data tables. A total of 171 212 changes were made, which represents 0.12% of the total number of records in the 17 tables involved. Many of the deletions and amendments were not of vital importance as often either the data changed were relatively mundane or it was not the primary data that were changed but values in the minor attributes. Nevertheless, the values were wrong and have now been corrected and, in some respects, these changes could be thought more valuable since the subtlety of many of the errors kept them so well hidden that only the auditing found them.

Apart from the changes to data, some changes to the processes used in the regular production of statistics were made and then all the statistics were recalculated from the base data in other tables. However, not all the statistics had been previously derived this way since many had been taken directly from paper records. There were 3 716 249 of these before the recalculation and now there are 1 754 816, which represent 22% of the total.

# Introduction

This report is the first 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.

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 *key* which, rather than being the data itself, are details about the "where", "when", and "what" of the data. The key defines each row such that no two rows have the same key; for example, for a particular point (first part of key) at a particular time (second part of key) there is only one value for the temperature and thus only one row is required. Thus from row to row the values in the columns constituting the key are independent, but it may well be that values in the other columns are not independent; 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 with 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 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 keys that are 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 MTHLY_STATS table

The **MTHLY_STATS** table contains monthly climate statistics. Its column names and the types of data they hold are:

| Column name | Null? | Type |
|---|---|---|
| AGENT_NO | NOT NULL | NUMBER(6) |
| STATS_YEAR | NOT NULL | NUMBER(4) |
| STATS_MONTH | NOT NULL | NUMBER(2) |
| STATS_CODE | NOT NULL | VARCHAR2(3) |
| STATS_VALUE | | NUMBER(10,4) |
| STATS_DATE_EXTREME | | DATE |
| ORIG_OBS_ORIGIN | | VARCHAR2(1) |
| STATS_REL | | VARCHAR2(1) |

Just as ORACLE ensures a column will hold only 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 time given by STATS_YEAR and STATS_MONTH which together give the local month at the place concerned; and the type of statistic given by STATS_CODE for which details are given in **STATS_CODE**. The remaining columns constitute the significant data with STATS_VALUE being the primary data since a row without this contains no information. All the other columns could be null, but ORIG_OBS_ORIGIN should usually be present to carry the lowest grade source of data that went into calculating STATS_VALUE. The remaining columns will often be null unless:

- for STATS_DATE_EXTREME the statistic is one for which a date can be associated (e.g., for the highest temperature in a month not only the temperature can be stored in a row but also the day, if available, on which it occurred);
- for STATS_REL if the statistic is not based on a full set of totally reliable values then a "*" is stored in this column.

A full description of **MTHLY_STATS** is given in 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.**

A.2. STATS_CODE: The entries in this column should all represent valid statistics codes i.e., they should all appear as STATS_CODEs in **STATS_CODE**.

A.3. STATS_REL: Only NULL (i.e., empty) or "*" are allowed

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".

B. Multiple column checks

B.1. STATS_YEAR, STATS_MONTH: The entries in these columns when combined for each row should form a legitimate date that is no later than the current date.

B.2. STATS_YEAR, STATS_MONTH, STATS_CODE: The earliest date for a given STATS_CODE should be reasonable. (Just finding the earliest overall date would ignore the later start of the observation of some data types.)

B.3. STATS_YEAR, STATS_MONTH, AGENT_NO: The earliest and latest dates should not be before the station opened or after it closed.

B.4. STATS_YEAR, STATS_MONTH, STATS_DATE_EXTREME: The date in STATS_DATE_EXTREME should be within the month given by STATS_MONTH, STATS_YEAR.

B.5. STATS_VALUE, STATS_CODE: For a particular STATS_CODE, values in STATS_VALUE should fall within a given range which can be set from STATS_CODE_UNITS in **STATS_CODE.**

B.6. AGENT_NO, STATS_CODE: For a given AGENT_NO, and hence through **LAND_STATION** a given STTY_STATION_TYPE, only certain types of STATS_CODE might be expected.

B.7. ORIG_OBS_ORIGIN, STATS_CODE: For a given STATS_CODE not all ORIG_OBS_ORIGINs are available.

C. Between row checks

C.1. For each STATS_CODE and STATS_MONTH, find the largest and smallest STATS_VALUE. (N.B. Since these would depend heavily on the AGENT_NO, the extremes are compared to the mean for each AGENT_NO).

C.2. For particular sub-sets of STATS_CODE and a given STATS_YEAR, STATS_MONTH and AGENT_NO, the STATS_VALUEs must be consistent, e.g., for the same place and month the warmest temperature must be higher than the lowest.

C.3. For the same STATS_CODE, STATS_YEAR and STATS_MONTH, the STATS_VALUEs should not be too different for AGENT_NOs that are physically close to each other.

C.4. For a given STATS_CODE and AGENT_NO, there should be a continuous dataset with no gaps from the row with the earliest STATS_MONTH, STATS_YEAR to that with the latest.

D. Other checks

D.1. For a given STATS_CODE and AGENT_NO, the length of record should be adequate.

D.2. If a given STATS_CODE, STATS_YEAR and AGENT_NO combination has rows for all possible values of STATS_MONTH then a row calculated from those rows for that combination must exist in **ANNUAL_STATS.**

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.1, B.3, B.4, B.7, C.2 and D.2 whereas B.2, B.5, B.6, C.1, and C.3 will highlight those rows that might be in error. Remaining checks (C.4 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. These last two checks are the only ones that perform any sort of temporal check since with monthly data the correlation from one month 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 **MTHLY_STATS** it must be known to which place the data in the row apply. A list of places where observations, and hence statistics of 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 **MTHLY_STATS** as a code for the station, thus, all the AGENT_NOs in **MTHLY_STATS** must appear in **LAND_STATION**. This held true, and so all stations were valid.

## Details and results of Check A.2 — are all codes valid?

For any row in **MTHLY_STATS** it must be known what type of statistic is held in that row. A list of the valid statistics codes with a full definition, their units of measurement and their type (mean, minimum etc.) is held in **STATS_CODE**. The list is indexed by STATS_CODE which is used in **MTHLY_STATS** as a code for the type of statistic, thus, all the STATS_CODEs in **MTHLY_STATS** must appear in **STATS_CODE**. This held true, and so all codes were valid.

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

For any row in **MTHLY_STATS** it ought to (but not *must*) be known what is the origin of the base data (i.e., that from which the statistic had been calculated). Here "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 **MTHLY_STATS**. This held true, and so all origins were valid. Also, if a set of base data is deficient in some way, but not to such an extent that precludes the calculation of a statistic value, then a "∗" is stored in STATS_REL, otherwise the column is left empty (i.e., NULL). It was found that either STATS_REL was NULL or contained a "∗", and so all reliabilities are valid.

As an extension to the basic checks just detailed and in order to determine whether some origins were invalid for the type of statistic concerned, counts were made for various statistics groups of the numbers with different origins and reliabilities. The latter were counted so that for any statistics group and origin the fraction which were unreliable could be found and considered as a measure of quality. The groups are:

| Group | Statistic codes | Comment |
|---|---|---|
| RAIN | 00 01 40 41 | |
| TEMP | 02 03 04 06 07 43 44 61 62 63 | |
| GMIN | 05 08 | Grass minimum temps |
| SUNS | 09 | |
| ETMP | 10 11 12 13 14 47 | Earth temps |
| WRUN | 15 18 | Wind run |
| VPRH | 16 64 65 | Vapour pressure & RH |
| RADN | 17 48 49 | Radiation |
| EVAP | 19 20 | |
| OCCS | 21 22 23 24 25 26 27 28 29 | Days of occurrence |
| MSLP | 30 | MSL pressure |
| WIND | 31 32 33 45 46 60 | Max gusts |

| | | |
|---|---|---|
| CALC | 34 35 36 37 38 39 | Various calculated stats |
| CLDS | 42 | Cloud |
| MAXR | 50 51 52 53 54 55 56 57 58 59 | Maximum rain |

| Origin | Rel | RAIN | TEMP | GMIN | SUNS | ETMP | WRUN | VPRH | RADN |
|---|---|---|---|---|---|---|---|---|---|
| D | | 1 113 491 | 169 643 | 17 066 | 5 004 | 21 006 | 11 011 | 205 585 | 3 473 |
| D | * | 6 956 | 45 302 | 7 386 | 556 | 5 090 | 156 | 51 341 | 1 147 |
| E | | . | . | . | . | 302 | . | 350 | . |
| E | * | . | . | . | . | 126 | . | 365 | . |
| F | | . | . | . | . | . | . | . | . |
| F | * | . | . | . | . | . | . | . | . |
| H | | . | . | . | . | 467 | . | 147 | . |
| H | * | . | . | . | . | 64 | . | 90 | . |
| M | | . | . | . | . | . | . | 7 899 | . |
| M | * | . | . | . | . | . | . | 1 779 | . |
| P | | . | . | . | . | . | . | . | . |
| P | * | . | . | . | . | . | . | . | . |
| S | | 4 938 | 13 637 | 316 | . | . | . | 5 623 | . |
| S | * | 94 | 16 614 | 990 | | 221 | . | 31 434 | . |
| T | | 985 569 | 726 859 | 234 203 | 49 170 | 153 372 | 53 233 | 28 015 | 8 875 |
| T | * | 7 013 | 1 548 | 274 | 558 | . | 1 | 12 | 3 |

| Origin | Rel | EVAP | OCCS | MSLP | WIND | CALC | CLDS | MAXR | TOTAL |
|---|---|---|---|---|---|---|---|---|---|
| D | | 2 796 | 43 002 | 3 864 | 70 243 | . | 3 109 | . | 1 669 293 |
| D | * | 272 | 1 163 | 626 | 94 | . | 1 205 | . | 121 294 |
| E | | . | . | . | . | . | . | . | 652 |
| E | * | . | . | . | . | . | . | . | 491 |
| F | | . | . | . | 13 890 | . | . | . | 13 890 |
| F | * | . | . | . | 59 | . | . | . | 59 |
| H | | . | . | . | . | . | . | . | 614 |
| H | * | . | . | . | . | . | . | . | 154 |
| M | | . | . | 2 110 | 2 185 | . | 7 | . | 12 201 |
| M | * | . | . | 670 | 299 | . | 2 | . | 2 750 |
| P | | . | . | . | . | . | . | 159 302 | 159 302 |
| P | * | . | . | . | . | . | . | 8 008 | 8 008 |
| S | | . | 648 | 1 766 | . | . | 614 | . | 27 542 |
| S | * | . | 191 | 2 321 | . | . | 2 870 | . | 54 735 |
| T | | 20 714 | 910 159 | 46 273 | 29 289 | 123 665 | 19 091 | 313 751 | 3 702 238 |
| T | * | . | 60 | . | . | . | . | 4 542 | 14 011 |

There were no occasions where a statistic had an origin that was invalid for the statistic concerned. Overall, T (Trentham) origin data had the lowest rate of unreliable data. However, the T data are not a growing source since they define those values which have not been changed since the data were recovered from the N.Z. Government computing centre at Trentham and loaded into CLIDB. Indeed, it is a decreasing source since most statistics can be recalculated from base data which are available in general from 1972 and a programme exists for extracting data from pre-1972 paper records and loading them into CLIDB. S (Synop) origin data had the highest rate of unreliable data but this is a relatively small data source. Of more interest are the D (Daily climate) origin data, where about 93% are reliable with the reliability ranging from 99.4% for the RAIN group down to 69.8% for the GMIN group. Some of the statistics groups have higher rates of reliability than others, but this is mainly due to the rules for calculating them being stricter with no option of calculating from a poor quality set of base data a value which can be marked as unreliable.

## Details and results of Check B.1 and B.2 — are all dates valid?

For any row in **MTHLY_STATS** it must be known to which month and year the data in the row apply. STATS_MONTH must lie between 1 and 12 inclusive and when combined with STATS_YEAR should form a date no later than the current date. This was found to hold, and so all 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 some types of observations were started earlier than others, there is no fixed threshold either for the earliest date. However, the earliest date for each statistic can be found and, as can be seen below, these dates are reasonable and lie between November 1852 for code 00 (total rainfall accumulation) and November 1991 for code 47 (mean 50 cm earth temperature). If applicable, the date of the earliest extreme is shown and that this is never any earlier than the date of the earliest data for the same code. Those cases where the earliest date is much earlier than the earliest extreme data date arise because keeping these in CLIDB is a relatively recent innovation which requires base data for their determination and such data are not always available on CLIDB.

| Code | Earliest data | Earliest extreme | Code | Earliest data | Earliest extreme |
|---|---|---|---|---|---|
| 00 | 185211 | | 31 | 195301 | |
| 01 | 190001 | | 32 | 195301 | |
| 02 | 185306 | | 33 | 193909 | |
| 03 | 185306 | | 34 | 193001 | |
| 04 | 185301 | | 35 | 193001 | |
| 05 | 186301 | | 36 | 193001 | |
| 06 | 185306 | 19481229 | 37 | 193001 | |
| 07 | 185301 | 19481217 | 40 | 190001 | |
| 08 | 186207 | 19481214 | 41 | 190001 | 19000106 |
| 09 | 190502 | | 42 | 186203 | |
| 10 | 195201 | | 43 | 194812 | 19481202 |
| 11 | 193606 | | 44 | 194812 | 19481209 |
| 12 | 193606 | | 45 | 196912 | 19691225 |
| 13 | 192006 | | 46 | 196912 | 19691225 |
| 14 | 191001 | | 47 | 199111 | |
| 15 | 186501 | | 50 | 194001 | 19400126 |
| 16 | 192801 | | 51 | 194001 | 19400126 |
| 17 | 194801 | | 52 | 194001 | 19400126 |
| 18 | 193712 | 19880406 | 53 | 194001 | 19400126 |
| 19 | 194105 | | 54 | 194001 | 19400126 |
| 20 | 195112 | | 55 | 194001 | 19400126 |
| 21 | 186701 | | 56 | 194001 | 19400126 |
| 22 | 186401 | | 57 | 194001 | 19400126 |
| 23 | 186201 | | 58 | 194001 | 19400126 |
| 24 | 186701 | | 59 | 194001 | 19400126 |
| 25 | 186301 | | 60 | 197107 | |
| 26 | 186301 | | 61 | 194812 | |
| 27 | 193001 | | 62 | 194812 | 19481202 |
| 28 | 186301 | | 63 | 194812 | 19481209 |
| 29 | 194108 | | 64 | 194812 | |
| 30 | 185310 | | 65 | 194812 | |

## Details and results of Check B.3 — are all records within the life of the station?

The life of the station is from the time it opened until it closed or until the current date. The dates of opening and closing for each station are held in **LAND_STATION** together with the status of a station with regard to it being open or closed. Thus for each AGENT_NO the earliest and latest records within **MTHLY_STATS** can be found for each STATS_CODE and an error noted if the earliest is before the station opened or if the latest is after the station closed. There were 2820 occasions when the record was outside the life of the station and 860 stations were involved.

A random sample of the errors was 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 a station's life were valid and amend the date of the station's opening and/or its closing 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. Thus, with such a large number of errors to deal with and no overall solution available, no changes were made to either **MTHLY_STATS** or **LAND_STATION** but entries were made in **SITE_CHANGES** as indicated below.

A much shortened list of the date anomalies found, in which only the first and last AGENT_NOs are retained, is given below:

| AGENT_NO | Code | Station start | Data start | Data end | Station end |
|---|---|---|---|---|---|
| | | First station with any date anomalies | | | |
| 1005 | 00 | 19310701 | 19310101 | 19730930 | 19730930 |
| 1005 | 09 | 19310701 | 19350101 | 19731231 | 19730930 |
| | | 858 stations omitted | | | |
| | | Last station with any date anomalies | | | |
| 12328 | 02 | 19951110 | 19941201 | 19970630 | |
| 12328 | 03 | 19951110 | 19941201 | 19970630 | |
| 12328 | 04 | 19951110 | 19941201 | 19970630 | |
| 12328 | 06 | 19951110 | 19941201 | 19970630 | |
| 12328 | 07 | 19951110 | 19941201 | 19970630 | |
| 12328 | 15 | 19951110 | 19950101 | 19970630 | |
| 12328 | 17 | 19951110 | 19941201 | 19970630 | |
| 12328 | 18 | 19951110 | 19950101 | 19970630 | |
| 12328 | 23 | 19951110 | 19941201 | 19970630 | |
| 12328 | 31 | 19951110 | 19950101 | 19970630 | |
| 12328 | 32 | 19951110 | 19950101 | 19970630 | |
| 12328 | 33 | 19951110 | 19941201 | 19970630 | |
| 12328 | 43 | 19951110 | 19941201 | 19970630 | |
| 12328 | 44 | 19951110 | 19941201 | 19970630 | |
| 12328 | 45 | 19951110 | 19950101 | 19970630 | |
| 12328 | 46 | 19951110 | 19950101 | 19970630 | |
| 12328 | 60 | 19951110 | 19950101 | 19970630 | |
| 12328 | 61 | 19951110 | 19941201 | 19970630 | |
| 12328 | 62 | 19951110 | 19941201 | 19970630 | |
| 12328 | 63 | 19951110 | 19941201 | 19970630 | |

It can be seen for AGENT_NO 1005 that for statistic code 00 data are available before the station opened and for code 09 data remained available after the station closed. The consequent rows inserted into **SITE_CHANGES** to note these anomalies and those for AGENT_NO 12328 are given below.

| AGENT_NO | Data date | Description |
|---|---|---|
| 1005 | 19310101:0000 | STATS CODES 00, began BEFORE station opened on 19310701 |
| 1005 | 19731231:0000 | STATS CODES 09, continued AFTER station shut on 19730930 |
| 12328 | 19941201:0000 | STATS CODES 02,03,04,06,07,17,23,33,43,44,61,62,63, began BEFORE station opened on 19951110 |
| 12328 | 19950101:0000 | STATS CODES 15,18,31,32,45,46,60, began BEFORE station opened on 19951110 |

## Details and results of Check B.4 — are all extreme dates within the month of the data?

For any row in **MTHLY_STATS** with a non-null STATS_DATE_EXTREME, the date in that column must be within the month represented by STATS_YEAR, STATS_MONTH. There were 57 occasions when the extreme date was not within the month of the data and these fell into three groups:

- 8 New Zealand station-months, 4 of which were statistics code 06 (highest maximum temperature) and the other 4 were code 43 (lowest maximum temperature);
- 31 Pacific Island station-months with code 45 (highest maximum gust); and,
- 18 New Zealand station-months with code 45.

For the first group, all were corrected by recalculating codes 06 and 43 for the months concerned but for two of them when the prior month was also included in the recalculation the error returned. On close examination of the program implementing the calculation, it was found that in extracting an extreme from a series of months for one station the date and value of the extreme were not always cleared between months. Then, if the value of the extreme for the new month was the same as the prior one, the date was not reset so that although the value was correct the date was not. The program was amended.

For the second group, the extreme dates were all on the first day of the next month and the stations all lay east of the International Dateline. Such stations have a negative time offset from UTC which when applied to the extreme date would bring it back into the correct month. Thus it appeared that a UTC rather than local date was being stored for the extreme date for code 45. On close examination of the program implementing the calculation, this was found to be the case and the program was amended.

For the third group, the error was traced back to the recovery of maximum gust data from the N.Z. Government Computing Centre at Trentham when a default time of 11:59 was used for the date. This was supposed to ensure that the correct local day remained when the data were transferred to **MAX_GUST**. However, daylight saving was not allowed for and at such times 11:59 UTC represents 00:59 local on the next day. Thus, during times of daylight saving an incorrect selection of candidate days was made in the calculation of code 45 statistics because OBS_DATE in **MAX_GUST** was incorrect. To correct this, the appropriate OBS_DATEs were amended with all amendments being properly recorded in **AUD_MAX_GUST**. After recalculating all code 45 statistics, three errors remained due to missing data not allowing their recalculation but these were deleted through the **MTHLY_STATS** maintenance screen and so were moved into

**AUD_MTHLY_STATS**. (N.B. The new statistics calculating procedures implemented after this recalculation of code 45 are such that those three errors would be deleted by the new procedure.)

## Details and results of Check B.5 — are all statistics values reasonable?

For every STATS_CODE, **STATS_CODE** supplies the units of measurement, and error bounds for STATS_VALUE can be set through those units. For example, knowing that the statistic is one whose units are "day" implies that the statistic may be zero but never negative and cannot be greater than 31 for January or 28 for February in a non-leap year. However, some of the upper and lower bounds are not so definite; consider those with "Celsius" as the units. CLIDB holds temperature data for both Antarctica and tropical Pacific islands, thus, any bounds set must attempt to trap errors without highlighting too many legitimate values. In the table below which displays the error bounds, for both "mm" and "hPa" the bounds also depend on the actual statistic rather than just its units, *viz*: code 35 (Priestly-Taylor PEM) can at times be slightly negative; code 30 (mean MSL pressure at 9 am) has much larger values than either code 16 (mean vapour pressure) or code 37 (Penman saturation deficit). Also, except for code 46 (direction of maximum gust, which may be missing when the speed of the maximum gust is not), no other codes should have an empty STATS_VALUE column in a row.

| Unit | Statistic codes | Min | Max |
|---|---|---|---|
| mm | 00 19 20 34 36 41 50 51 52 53 54 55 56 57 58 59 | 0 | 9999 |
| day | 01 21 22 23 24 25 26 27 28 29 31 32 40 60 | 0 | L |
| Celsius | 02 03 04 05 06 07 08 10 11 12 13 14 38 39 43 44 47 61 62 63 65 | –60 | 50 |
| hours | 09 | 0 | L X 24 |
| km | 15 18 | 0 | 9999 |
| hPa | 16 37 | 0 | 35 |
| MJ $m^{-2}$ | 17 48 49 | 0 | 999 |
| hPa | 30 | 960 | 1040 |
| m $s^{-1}$ | 33 45 | 0 | 99 |
| mm | 35 | –20 | 9999 |
| eighths | 42 | 0 | 8 |
| degrees | 46 | 0 | 360 |
| percent | 64 | 0 | 100 |

N.B. L = Length of month concerned in days

The results of searching for out-of-range values are tabulated below by STATS_CODE with the initial count of errors, some explanation where possible of how the errors arose, and what action was taken. After taking as much remedial action as possible, another search was made and the tabulation shows the consequent count of errors and comments on whether further action is necessary.

| Code | Count before | Remedial action | Count after | Further action |
|---|---|---|---|---|
| 00 | 16 | These were either NULL or contained 22860 which is the old missing code with an inch to mm conversion. Most were deleted. | 3 | These were NULLs and were associated with accumulations from one month to the next. |
| 02 | 18 | For V88053 all 12 months for 1971 values were still in °F rather than °C. | 6 | Only Antarctic stations left, values are OK. |
| 03 | 1 | — | 1 | See 02 |
| 04 | 46 | For J81100 Nov 1929 was still in °F rather than °C. | 45 | See 02 |

| | | | | |
|---|---|---|---|---|
| 07 | 76 | — | 76 | See 02 |
| 09 | 2 | For G13231 2 values were wrong and were amended. | — | — |
| 10 | 2 | These were NULLs due to NULLs in the base data which were deleted and the statistic recalculated for the appropriate period. | 8 | These were different ones also from NULLs in the base data. A further 2225 NULLs at 5 cm were deleted from **EARTH_TEMP** but further recalculation not done until the new statistics procedures were implemented. The NULLs were introduced during the recovery of data from Trentham. |
| 15 | 1 | NULLs as for code 10 | 2 | More NULLs again as for code 10 with 669 NULLs now deleted from **SURFACE_WIND**. In this case **LAND_DATA_CAT** with 131 in LADA_LAND_DATA and just 1 in ACTUAL_COUNT had 326 rows deleted since they were for stations without data. Also DATA FAULT 1997/023 was raised. |
| 16 | 113 | There were 86 negative values which were simply deleted. One value, H40041 Mar 1951, had been incorrectly entered and was amended. | 26 | These were for tropical Pacific island with high values of up to 41 hPa and are probably correct. |
| 20 | 4 | These were NULLS, they were deleted and the statistic recalculated. | 2 | These were NULLs and were associated with accumulations from one month to the next. |
| 30 | 5 | Of these three were amended and two deleted. | — | — |
| 34 | 9 | These were negative values and were not touched. | 9 | May need some action at a later time. |
| 35 | 19 | — | 19 | These ranged down to –25 i.e., a little below the lower bound of –20 and are probably correct. |
| 37 | 34 | These were all negative values. One value, H40041 Mar 1951, had been incorrectly entered and was amended (see 16). | 33 | Mainly only slightly negative with only four values below –1. May need some action at a later time. |
| 40 | 1 | J72000 Jul 1982 was wrong and was deleted. | — | — |
| 42 | 7 | These were NULL values but the base data had many missing values so they were simply deleted. | — | It was also found that during the recovery of data from Trentham the missing code indicator had been left as 9 rather than changed to NULL. Thus 103232 rows were amended in **CLOUD_SYSTEM** and then 53666 of those were deleted as only containing NULLs. |
| 44 | 3 | — | 3 | Only Antarctic stations, values OK. |
| 64 | 143 | These were 3 NULLs and 140 negative values. They were all deleted. | — | — |

Overall 500 potential range errors were detected, but of these 233 were most probably not errors so 267 errors were corrected. Also about 50 000 rows in **CLOUD_SYSTEM** were amended and about 55 000 rows in **CLOUD_SYSTEM, EARTH_TEMP** and **SURFACE_WIND** were deleted.

## Details and results of Check B.6 — are all sources for statistics reasonable?

In the A.4 and B.7 checks the statistics were divided into groups. These same groups are used in this check where a search is made to locate any statistics that have been attributed to a station which is of such a type that it would not be expected to have reported that type of statistic. Stations often change their type while open or they may shut and some time later one of a different type may open sufficiently close by to merit the re-use of the closed stations number. Thus, in the table below, which shows counts of station type by statistic type, it is generally difficult to find any definite errors. However, the check did highlight one station (that below under RAIN in the UNKNOWN row) with a NULL in **LAND_STATION**'s STTY_STATION_TYPE. This was a station in the Caroline Islands (J33400) which has now been assigned to the CLIMAT/SYNOP group.

| | RAIN | TEMP | GMIN | SUNS | ETMP | WRUN | VPRH | RADN |
|---|---|---|---|---|---|---|---|---|
| RAIN (STANDARD) | 2 421 | 54 | 36 | 17 | 10 | 13 | 37 | 1 |
| CLIMAT (STANDARD) | 464 | 437 | 344 | 104 | 177 | 133 | 393 | 34 |
| CLIMAT/ SYNOP | 132 | 119 | 62 | 40 | 36 | 13 | 109 | 26 |
| RAIN/SYNOP | 54 | 14 | 10 | 1 | 3 | 1 | 25 | 2 |
| CLIMAT (PRIVATE) | 11 | 7 | 6 | 2 | 3 | 3 | 5 | 1 |
| RAIN (PRIVATE) | 251 | . | . | . | . | . | . | . |
| REGIONAL COUNCIL | 111 | 3 | 3 | 1 | . | . | 3 | . |
| WATER SCIENCES | 48 | 1 | 1 | . | 1 | 1 | 1 | . |
| ANEMOMETER ONLY | 6 | 5 | 4 | . | . | . | 5 | . |
| SYNOP ONLY | 79 | 92 | 23 | 10 | 4 | 2 | 76 | 6 |
| AWS (SYNOP/METAR) | 65 | 57 | 16 | 3 | 7 | 45 | 73 | 35 |
| EDR | 18 | 19 | 15 | 1 | 15 | 17 | 13 | 16 |
| CLITEL | 16 | 21 | 15 | 1 | 13 | 20 | 13 | 20 |
| LIMITED CLIMAT | 7 | 2 | . | . | . | 1 | . | . |
| SPECIAL STATION | 9 | 8 | 3 | 5 | 2 | 1 | 6 | 1 |
| UNKNOWN | 1 | . | . | . | . | . | . | . |

| | EVAP | OCCS | MSLP | WIND | CALC | CLDS | MAXR | TOTAL |
|---|---|---|---|---|---|---|---|---|
| RAIN (STANDARD) | 9 | 47 | 9 | 5 | 9 | 22 | 84 | 2 774 |
| CLIMAT (STANDARD) | 97 | 376 | 24 | 38 | 72 | 136 | 120 | 2 949 |
| CLIMAT/ SYNOP | 19 | 106 | 76 | 47 | 27 | 70 | 32 | 914 |
| RAIN/SYNOP | . | 10 | 4 | 6 | 2 | 19 | 4 | 155 |
| CLIMAT (PRIVATE) | 2 | 6 | . | . | 1 | . | 1 | 48 |
| RAIN (PRIVATE) | . | 1 | . | . | . | . | 2 | 254 |
| REGIONAL COUNCIL | . | 3 | . | . | . | 1 | 16 | 141 |
| WATER SCIENCES | . | 1 | . | . | . | . | 3 | 57 |
| ANEMOMETER ONLY | . | 5 | . | 47 | . | . | 1 | 73 |
| SYNOP ONLY | 3 | 60 | 81 | 28 | 8 | 53 | 17 | 542 |
| AWS (SYNOP/METAR) | 1 | 52 | 79 | 59 | 2 | 5 | 7 | 506 |
| EDR | . | 19 | 1 | 13 | . | . | | 147 |
| CLITEL | 3 | 21 | . | 20 | 1 | 4 | 4 | 172 |
| LIMITED CLIMAT | 2 | 2 | . | . | . | . | 2 | 16 |
| SPECIAL STATION | 1 | 8 | 1 | 1 | 2 | 2 | 2 | 52 |
| UNKNOWN | . | . | . | . | . | . | . | . |

A few other entries were investigated but no errors were found.

## Details and results of Check C.1 — are all the smallest and largest statistics values reasonable?

A gross check on statistics values was performed by Check B.5 above where for a given STATS_CODE values were simply required to lie between certain limits. In this check the largest and smallest values for a given set of STATS_CODE, STATS_MONTH, and AGENT_NO will be compared to the mean value for that set. This recognises that for a given code much of the variability is likely to be due to geographical position and time of year, after which the independent variability remains and can be measured by the range and compared to the mean.

The fraction (Mean–Minimum)/(Maximum–Minimum) was taken as a measure of the smallness of the minimum. Thus, if a minimum is an outlier (i.e., is much smaller than might be expected) then the value of this fraction is closer to unity than otherwise since for an outlier the minimum is relatively far from the mean and (Mean–Minimum) is relatively close to (Maximum–Minimum). For most codes the rows in **MTHLY_STATS** that may be of most interest are those where this ratio is a maximum, i.e., a candidate row's STATS_VALUE must be a minimum for a given set of STATS_CODE, STATS_MONTH, and AGENT_NO and the ratio just defined for that set must be a maximum over all possible STATS_MONTHs and AGENT_NOs with the given STATS_CODE.

Similarly, (Maximum–Mean)/(Maximum–Minimum) was taken as a measure of the largeness of the maximum. Again, for most codes the rows in **MTHLY_STATS** that may be of most interest are those where this ratio is a maximum. The ratios will highlight potential errors for many, but not all, codes since those which have distributions with one long tail will tend to have their mean closer to the other extreme and so produce large values of the ratio even for acceptable data. Taking this into account the actual errors found by this check were:

| Code | AGENT_NO | Month | Count | Mean | Maximum | MAX_RATIO |
|---|---|---|---|---|---|---|
| | | First search for large maxima | | | | |
| 00 | 6250 | 9 | 68 | 174.3 | 2 411.5 | 0.94 |
| 09 | 4244 | 11 | 38 | 251.7 | 498.2 | 0.80 |
| 18 | 3477 | 9 | 23 | 655.8 | 3 787.0 | 0.89 |
| 41 | 3629 | 6 | 26 | 40.5 | 401.4 | 0.92 |
| 54 | 5866 | 4 | 25 | 10.3 | 83.4 | 0.90 |
| 55 | 5866 | 4 | 25 | 20.3 | 250.0 | 0.93 |
| 56 | 5866 | 4 | 24 | 33.3 | 500.0 | 0.94 |
| 57 | 5866 | 4 | 25 | 55.2 | 999.9 | 0.95 |
| 58 | 1427 | 6 | 30 | 110.6 | 1 999.8 | 0.95 |
| 59 | 1427 | 6 | 30 | 148.9 | 2 999.7 | 0.96 |
| | | Second search for large maxima | | | | |
| 18 | 2212 | 4 | 15 | 268.7 | 999.0 | 0.87 |
| 41 | 4559 | 11 | 25 | 38.6 | 360.0 | 0.90 |

| Code | AGENT_NO | Month | Count | Minimum | Mean | MIN_RATIO |
|---|---|---|---|---|---|---|
| | | First search for small minima | | | | |
| 07 | 6061 | 3 | 36 | –25.5 | 20.8 | 0.94 |
| 08 | 5812 | 11 | 38 | –16.0 | –1.4 | 0.86 |
| 19 | 5814 | 8 | 20 | 0.0 | 24.1 | 0.71 |
| 23 | 6237 | 10 | 24 | 30.0 | 31.0 | 0.96 |
| 23 | 6237 | 12 | 24 | 30.0 | 31.0 | 0.96 |
| 62 | 4153 | 7 | 4 | –14.5 | 0.5 | 0.75 |
| | | Second search for small minima | | | | |
| 07 | 6044 | 1 | 67 | 3.2 | 21.6 | 0.91 |
| 19 | 5683 | 9 | 27 | 0.0 | 35.5 | 0.70 |

where "Second search" were those found after the errors from the "First search" were corrected. The remedial actions taken for each of these errors are shown below.

| Code | STN_ID | DATE | |
|---|---|---|---|
| 00 | M99500 | 9/1922 | Email to Bureau of Meteorology to get correct Lord Howe value. |
| 09 | G13231 | 11/1913 | Incorrectly punched from original paper record. |
| 18 | E15102 | 9/1995 | Incorrect value, also 2/1997 was corrected. |
| 41 | E95464 | 6/1991 | Incorrect rainfall as compared to adjacent stations. |
| 54—59 (see 1 below) | | | |
| 18 | C85314 | 4/1997 | Accumulation code treated as the observation. |
| 41 | H23631 | 11/1991 | Rainfall of 36.0 mm entered as 360.0 mm. |
| 07 | J77600 | 3/1992 | Temperature should not have been negative. |
| 08 | I68431 | 11/1914 | Temperatures for 11–12/1914 from original paper record were 32 °F and 33 °F but had been entered as 3.2 °F and 3.3 °F. |
| 19 | I68433 | 8/1968 | Original paper record showed "Frozen" thus data were really missing and the zero that had been entered was deleted (see 2 below). |
| 23 | L66500 | 10/1989 | All days had a screen frost and the value was amended to 31. |
| 23 | L66500 | 12/1984 | Confirmed 13/12/1984 had temp 0.8 °C so no change made. |
| 62 | G04601 | 7/1997 | Incorrect daily minimum (N.B. mean = (max+min)/2). |
| 07 | J76200 | 1/1992 | The value came from a SYNOP report and was possibly 23.2 °C (see 3 below). |
| 19 | I59891 | 9/1967 | Same as for I68433 8/1968 above. |

1. The values indicated that a daily total of around 999 mm had been picked up from **RAIN_RATE** in which 999.9 is used in TOTAL as a default value in STATUS=1 records, i.e., those days for which neither a manual total nor chart total was available. It appeared that some such days had been entered at the digitising stage as if only the chart were missing (i.e., a STATUS=2 record) and a manual total of 999.9 mm was available. On scanning **RAIN_RATE** 37 such cases were found and these were amended to be STATUS=1 records. The original raw data files were also amended so that a re-run would not put back the errors. Finally the erroneous records in **MTHLY_STATS** were deleted and code 50 (and consequently codes 51–59) was recalculated for the months involved.
2. During the recovery of data from Trentham, values for STATS_CODEs 19 and 20 had had those months which were marked as missing loaded as zeros into CLIDB. A further 35 months with a zero were found and on checking the original paper records each such month was noted as either "Incomplete - Frozen" or "Frozen". They were deleted.
3. There is a major problem with the loading of **MAX_MIN_TEMP** data for the Pacific Islands from SYNOP and DAILY CLIMATE messages. DATA FAULT 1996/017 has details of a possible fix, but it is a complex issue.

## Details and results of Check C.2 — are all statistics within a group consistent?

For a particular place and time (i.e., for a given set of STATS_YEAR, STATS_MONTH and AGENT_NO) the statistics within a group of STATS_CODEs must be consistent. For example, the total rain in a month must be no less than the number of days on which at least 1 mm of rain fell; or, because accumulations do not prevent the calculation of the month's total rain but can prevent the counting of 1 mm rain-days, then if the total is absent the number of rain-days should also be absent but not necessarily *vice versa*. The groups used are shown below.

| Group | Statistics code with abbreviated descriptions | | |
|---|---|---|---|
| RAIN | 00 (Total rain) | 01 (No. of days with at least 1 mm) | 40 (No. with 0.1 mm) |
| | 41 (Max 1 day fall) | 57 (Max 24 hr fall) | 59 (Max 72 hr fall) |
| TEMP | 02 (Mean) | 03 (Max) | 04 (Min) |
| | 06 (Xmax) | 07 (Xmin) | 43 (Low-max) |
| | 44 (High-min) | 62 (Low-mean) | 63 (High-mean) |
| GMIN | 05 (Gmin) | 07 (Xmin) | 08 (Xgmin) |
| | 22 (No. ground frosts) | 23 (No. air frosts) | |
| WIND | 31 (Gale days) | 32 (Storm days) | 45 (Max speed) |
| | 46 (Max direction) | 60 (Strong wind days) | |
| MAXR | 50 (10min max) | 51 (20min max) | 52 (30min max) |
| | 53 (1hr max) | 54 (2hr max) | 55 (6hr max) |
| | 56 (12hr max) | 57 (24hr max) | 58 (48hr max) |
| | 59 (72hr max) | | |
| OTHER | 15 (Windrun) | 17 (Global radiation) | 18 (Xwindrun) |
| | 49 (Diffuse radiation) | | |

These are much the same as the groups used before and generally, but not strictly, within each group the base data are the same for all the members. The OTHER group collects together two small groups.

In the tabulations below (one for each group), the first column gives criteria defining the possible inconsistencies. These are generally straightforward and will not be explained further. The "Ref" column provides a link to the notes below the table, and the other columns are the counts of the number of inconsistencies found initially and then after corrections were made. The "After (C)" refers to a count in which all records in **MTHLY_STATS** have been included whereas the "After" count excludes those records whose ORIG_OBS_ORIGIN was "C". Records of that type are usually those that have been entered or amended through CLIDB's **MTHLY_ STATS** maintenance screen and so have not been calculated from base data. For example, for a given station and month there may be insufficient data to automatically calculate the extreme maximum temperature for the month. However, it may have been noticed that the missing value was for a time when nearby stations had cold rather than warm temperatures so a valid value can be found from what data are available and entered through the maintenance screen. Then, to preserve this value from subsequent automatic deletion, its ORIG_OBS_ORIGIN is set to "C".

Some of the consistency checks involve ensuring the presence of one member of a group given that another member is present. However, the earlier values were entered directly as monthly statistics from the original paper records rather than as daily values from which the monthly statistics were subsequently calculated. Thus, for those there can be no requirement that one is present just because another is present. To prevent possibly spurious inconsistencies being noted, such checks were performed only when relevant daily data are available according to **LAND_DATA_CAT**.

## Rain group

| | Before | Ref | After | After (C) |
|---|---|---|---|---|
| Total rain < No. of days with at least 1 mm | 6 | 4 | 0 | 0 |
| Total rain < (No. of days with at least 0.1 mm)/10 | 5 | 1,4 | 0 | 0 |
| Total rain < Max 1 day fall | 35 | 1,4 | 0 | 0 |
| Total rain < Max 72hr fall | 20 | 2,5 | 16 | 16 |
| Total rain NULL but not Max 1-day or No. of 0.1 or 1 mm days | 12 342 | 3 | 0 | 0 |
| No of 0.1 mm days < No. of 1 mm days | 0 | | 0 | 0 |
| Max 24 hr fall < Max 1 day fall | 124 | 2,5 | 104 | 104 |
| Total number of station-months with an error | 12 518 | | 120 | 120 |

1. The pre-1991 total rains had been rounded to the nearest millimetre and some allowances must be made for this.
2. For codes 57 and 59 the period concerned is only required to end in the month concerned and could have started in the prior month; thus, only those which also started in that month were compared to other statistics.
3. Of these 10 433 were for Fiji Meteorological Service stations for which daily data were available. Also spot checks of other stations often showed that daily data were available. Thus a recalculation of code 00 was made and the number of errors dropped from 12 342 to 64 and these were reduced to zero by simply deleting the appropriate code 01, 40, and 41 statistics. Ten of the 64 had been cases where the total rain was NULL and another 119 such records were found and deleted.
4. These inconsistencies were cured by the recalculation done in item 3.
5. Over the past few years the manual cross-check between incoming data for entry in **RAIN** and that for entry in **RAIN_RATE** has been suspended. Consequently, some of the differences between those tables had been passed on to **MTHLY_STATS,** but on directly checking **RAIN** and **RAIN_RATE** about 4500 differences were found. About half of these were differences of 0.5 mm or less and the records in **RAIN** were just changed to agree with those in **RAIN_RATE.** The remainder had the pluviograph examined and a decision made as to which was correct. There were fewer cases in which **RAIN** was correct and these were treated by amending the raw digitised data, re-loading it into **RAIN_RATE** and recalculating codes 50–59. For most cases, **RAIN_RATE** was correct and its data were transferred to **RAIN,** but codes 00, 01, 40, and 41 were not at that time recalculated hence the 120 errors that remained.
6. **N.B.** To ensure that the situation detailed in item 5 does not re-occur, a procedure was developed which every month performs a consistency check between **RAIN** and **RAIN_RATE** for the latest data and any errors are corrected as they occur. The procedure will also insert data into **RAIN** from **RAIN_RATE** where they are missing from the former but present in the latter. On its first run this procedure added 6568 rows to **RAIN.**

## Temperature group

| | Before | Ref | After | After (C) |
|---|---|---|---|---|
| Mean < Xmin or Min or Low-mean | 6 | 2 | 4 | 4 |
| Mean > Xmax or Max or High-mean | 562 | 2 | 12 | 12 |
| Mean is NULL but not Low-mean or High-mean | 12 | 2 | 0 | 0 |
| Max < Mean or Low-Max or Min or Xmin | 571 | 2 | 14 | 15 |
| Max > Xmax | 5 | 2 | 4 | 4 |
| Max is NULL but not Xmax or Low-max | 31 | 2 | 0 | 0 |
| Max is NULL but not Mean or Low-mean or High-mean | 809 | 1 | 22 | 22 |
| Min < Xmin | 7 | 2 | 5 | 5 |
| Min > Mean or High-min or Max or Xmax | 4 | 2 | 2 | 2 |
| Min is NULL but not Xmin or High-min | 23 | 2 | 0 | 1 |
| Min is NULL but not Mean or Low-mean or High-mean | 1 319 | 1 | 22 | 22 |
| Xmax is not the greatest | 29 | 2 | 7 | 7 |
| Xmax is NOT NULL but Max is NULL | 31 | 2 | 0 | 0 |
| Xmin is not the least | 10 | 2 | 6 | 6 |
| Xmin is NOT NULL but Min is NULL | 23 | 2 | 0 | 1 |
| Low-max < Xmin | 2 | 2 | 1 | 1 |
| Low-max > Max or Xmax | 0 | | 0 | 0 |
| Low-max is NOT NULL but Max is NULL | 0 | | 0 | 0 |
| High-min < Min or Xmin | 0 | | 0 | 0 |
| High-min > Xmax | 17 | 2 | 2 | 2 |
| High-min is NOT NULL but Min is NULL | 0 | | 0 | 0 |
| Low-mean < Xmin | 1 | 2 | 0 | 0 |
| Low-mean > Mean or High-mean or Xmax | 0 | | 0 | 0 |

| | | | | |
|---|---|---|---|---|
| Low-mean is NOT NULL but Mean or Min or Max is NULL | 17 | 2 | 0 | 0 |
| High-mean < Mean or Low-mean or Xmin | 1 | 2 | 0 | 0 |
| High-mean > Xmax | 5 | 2 | 0 | 0 |
| High-mean is NOT NULL but Mean or Min or Max is NULL | 17 | 2 | 0 | 0 |
| Total number of station-months with an error | 1 410 | | 36 | 37 |

1. The greatest source of errors was where code 02 was present but either code 03 or 04 was missing (1292 errors). Codes 02, 03, 04, 06, 07, 43, and 44 were recalculated for 1922 to 1992 and the number of errors fell to 165.
2. These errors were checked against original paper records where possible and corrections were made via the maintenance screen. However, since some of the original records could not be found some errors remain. The tabulation below shows what amendments (UPD), deletions (DEL), and inserts (INS) were made to the statistics tables; 385 changes were made and another 182 were made to **MAX_MIN_TEMP**.

| | ANNUAL_STATS | | | MTHLY_STATS | | | |
|---|---|---|---|---|---|---|---|
| Code | UPD | DEL | INS | UPD | DEL | INS | TOT |
| 02 | 4 | 18 | . | 17 | 53 | . | 92 |
| 03 | 4 | 1 | . | 7 | 1 | . | 13 |
| 04 | 7 | . | 2 | 16 | . | 2 | 27 |
| 06 | . | 7 | . | 4 | 12 | . | 23 |
| 07 | 6 | 11 | 1 | 56 | 22 | 5 | 101 |
| 43 | 1 | . | 1 | 20 | 1 | . | 23 |
| 44 | 8 | . | 2 | 49 | . | 2 | 61 |
| 62 | . | . | . | 3 | 17 | . | 20 |
| 63 | . | . | . | 8 | 17 | . | 25 |

## Grass minimum group

| | Before | Ref | After | After (C) |
|---|---|---|---|---|
| Fewer days of ground than of air frost | 995 | 1 | — | — |
| Gmin < Xgmin | 0 | | 1 | 1 |
| No ground frosts but Xgmin <= –1 | 478 | 2 | 13 | 13 |
| Ground frosts but Xgmin > –1 | 297 | 2 | 0 | 0 |
| No air frosts but Xmin < 0 | 328 | 2 | 22 | 22 |
| Air frosts but Xmin >= 0 | 115 | 2 | 2 | 2 |
| Gmin is NULL but Xgmin is NOT NULL | 31 | 2 | 0 | 0 |
| Xgmin is NULL but No. of ground frosts is NOT NULL | 280 | 2 | 7 | 7 |
| Xmin is NULL but No. of air frosts is NOT NULL | 200 | 2 | 5 | 5 |
| Total number of station-months with an error | 2 320 | | 49 | 49 |

1. Although this test seemed reasonable it was not because different threshold temperatures measured by different instruments are associated with the definitions of ground and air frost (i.e., an air frost occurs when the screen minimum is colder than zero while a ground frost occurs when the ground minimum is –1 °C or less). Thus, for example, a day with a screen minimum of –0.3 °C and a grass minimum of –0.6 °C would be counted as having had an air frost but not a ground frost. When this check was removed and after recalculating the statistics the total number of errors fell to 1102.
2. These errors were checked against original paper records where possible and corrections were made via the maintenance screen. However, since some of the original records could not be found some errors remain. The tabulation below shows what amendments (UPD), deletions (DEL), and inserts (INS) were made to the statistics tables; 1847 changes were made.

| | ANNUAL_STATS | | | MTHLY_STATS | | | |
|---|---|---|---|---|---|---|---|
| Code | UPD | DEL | INS | UPD | DEL | INS | TOT |
| 05 | 6 | 12 | 1 | 11 | 17 | 7 | 54 |
| 08 | 4 | 13 | 1 | 83 | 25 | 2 | 130 |
| 22 | 345 | 28 | 2 | 575 | 52 | 1 | 1 010 |
| 23 | 227 | 37 | 2 | 325 | 59 | 3 | 653 |

## Wind group

| | Before | Ref | After | After (C) |
|---|---|---|---|---|
| Fewer Strong wind days than Gale days | 0 | | 0 | 0 |
| Fewer Strong wind days than Storm days | 0 | | 0 | 0 |
| Fewer Gale days than Storm days | 2 | 2 | 0 | 0 |
| Strong gusts but no Strong wind days | 0 | | 0 | 0 |
| Gale gusts but no Gale days | 5 | 2 | 0 | 0 |
| Storm gusts but no Storm days | 6 | 2 | 0 | 0 |
| Strong wind days but no Strong gusts | 0 | | 0 | 0 |
| Gale days but no Gale gusts | 4 | 2 | 0 | 0 |
| Storm days but no Storm gusts | 7 | 2 | 0 | 0 |
| Max gust speed is NULL but Wind days or Max gust direction | 1 895 | 1 | 0 | 1 732 |
| Total number of station-months with an error | 1 916 | | 0 | 1 732 |

1. Recalculation did not decrease these errors which were mainly that codes 31 and 32 were present but codes 45 and 60 were not. The former were not based on full base datasets and had probably been manually estimated by considering nearby stations. To remove this major inconsistency either the rows could be deleted or the rows could have their ORIG_OBS_ORIGINs changed to "C" and STATS_REL to "*". The latter course was adopted.
2. These errors were checked against original paper records where possible and corrections were made. A total of 22 corrections were made to **ANNUAL_STATS** and **MTHLY_STATS**.

## Maximum rain group

| | Before | Ref | After | After (C) |
|---|---|---|---|---|
| 10 min max < (20 min max)/2 | 0 | | 0 | 0 |
| 10 min max > 20 min max | 0 | | 0 | 0 |
| 20 min max < (30 min max)/2 | 29 | 1 | 0 | 0 |
| 20 min max > 30 min max | 0 | | 0 | 0 |
| 30 min max < (1 h max)/2 | 0 | | 0 | 0 |
| 30 min max > 1 h max | 0 | | 0 | 0 |
| 1 h max < (2 h max)/2 | 0 | | 0 | 0 |
| 1 h max > 2 h max | 0 | | 0 | 0 |
| 2 h max < (6 h max)/3 | 0 | | 0 | 0 |
| 2 h max > 6 h max | 0 | | 0 | 0 |
| 6 h max < (12 h max)/2 | 0 | | 0 | 0 |
| 6 h max > 12 h max | 0 | | 0 | 0 |
| 12 h max < (24 h max)/2 | 0 | | 0 | 0 |
| 12 h max > 24 h max | 0 | | 0 | 0 |
| 24 h max < (48 h max)/2 | 0 | | 0 | 0 |
| 24 h max > 48 h max | 0 | | 0 | 0 |
| 48 h max < (72 h max)/2 | 24 | 2 | 0 | 0 |
| 48 h max > 72 h max | 0 | | 0 | 0 |
| Sub-24 h maxs without 24,48, or 72 h maxs | 63 | 2 | 0 | 0 |
| Total number of station-months with an error | 116 | | 0 | 0 |

1. These errors were generally small and occurred only for the older manually extracted ones rather than those taken from the digitised data in **RAIN_RATE**. The ratio of code 51 to code 52 should be no more than 2 but was up to 2.5 for 25 of these and to 3.29 for the other 4. An amended value for code 51 was taken to be half the value of code 52, but it was then necessary to estimate amended values for code 50 for five of these as half of the new code 51 value otherwise inconsistencies of the first type in the above table would occur.
2. Again these errors occurred only for the older manually extracted maximum rainfalls rather than those taken from the digitised data in **RAIN_RATE**. A PL/SQL script was written to populate a table with the required 1, 2, and 3 day accumulations which could be used to find estimates for the 24, 48, and 72 h falls.
3. The tabulation below shows what amendments (UPD), deletions (DEL) and inserts (INS) were made to the statistics tables; a total of 191 changes were made.

| | **ANNUAL_STATS** | | | **MTHLY_STATS** | | | |
|---|---|---|---|---|---|---|---|
| Code | UPD | DEL | INS | UPD | DEL | INS | TOT |
| 50 | . | . | . | 7 | . | . | 7 |
| 51 | . | . | . | 35 | . | . | 35 |
| 56 | . | . | . | 3 | . | . | 3 |
| 57 | . | . | . | 11 | . | 14 | 25 |
| 58 | . | . | . | 12 | . | 37 | 49 |
| 59 | . | . | . | 26 | . | 46 | 72 |

### Other group

| | Before | Ref | After | After (C) |
|---|---|---|---|---|
| Xwindrun < Windrun | 3 | 1 | 0 | 0 |
| Xwindrun NOT NULL but Windrun is NULL | 103 | 1 | 0 | 0 |
| Global radiation < Diffuse radiation | 0 | | 0 | 0 |
| Total number of station-months with an error | 106 | | 0 | 0 |

1. Some of these were already documented in DATA FAULT 1996/032 which was fixed before a recalculation was made. This reduced the number of errors to 26 and these were checked against original paper records and corrections made via the maintenance screen.

## Details and results of Check C.3 — are all statistics, when compared to nearby stations, reasonable?

As a preliminary step it was necessary to find for each station-code combination enough nearby stations, or buddies, to adequately cover the period over which the primary station had reported the code concerned and which were the closest to the primary station for that code. To be considered as a buddy, a station had to be within 1° of latitude and longitude for New Zealand, and 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 tabulations below are for the codes which had the best (code 00) and worst (code 30) buddy statistics. They show the counts of the primary stations in different distance-cover classes. For example, there were 591 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 stations' total rainfall record. At the other extreme, for total rainfall there were 15 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; for code 00 there were 16 such stations.

Buddy frequencies by distance and coverage for code 00, total rainfall

| | | 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 | 591 | 1 635 | 464 | 83 | 35 | 10 | 5 | 4 | 1 | 1 | 4 |
| Y | 85–95 | 47 | 142 | 57 | 9 | 2 | 4 | 3 | . | . | . | . |
| Y | 75–85 | 1 | 10 | 11 | 2 | 4 | . | 1 | . | . | . | 1 |
| Y | 65–75 | . | 9 | 7 | . | . | 2 | . | . | . | . | 1 |
| Y | 55–65 | . | 7 | 3 | . | . | 1 | . | . | . | . | . |
| Y | 45–55 | . | 1 | . | . | . | . | . | . | . | . | . |
| Y | 35–45 | 1 | . | . | 1 | . | . | . | . | . | 1 | . |
| Y | 25–35 | . | . | . | . | . | . | . | . | . | . | . |
| Y | 15–25 | . | . | . | . | . | . | . | . | . | . | . |
| Y | 5–15 | 2 | . | . | . | . | . | . | . | . | . | . |
| Y | 0–5 | . | . | . | . | . | . | . | . | . | . | 1 |
| N | 95–100 | 135 | 175 | 34 | 16 | 14 | 10 | 4 | 8 | 7 | 7 | 67 |
| N | 85–95 | 9 | 13 | 4 | . | 4 | 1 | . | 1 | . | . | 7 |
| N | 75–85 | . | . | . | . | . | . | . | 1 | . | . | 8 |
| N | 65–75 | . | 1 | . | . | 1 | . | . | . | . | . | 3 |
| N | 55–65 | . | . | . | . | . | . | . | . | . | . | 3 |
| N | 45–55 | . | 1 | . | . | . | . | . | . | . | . | 1 |
| N | 35–45 | . | . | . | . | . | . | . | . | . | . | 3 |
| N | 25–35 | . | . | . | . | . | . | . | . | . | . | 1 |
| N | 15–25 | . | . | . | . | . | . | . | . | . | . | 1 |
| N | 5–15 | . | . | . | . | . | . | . | . | . | . | . |
| N | 0–5 | . | . | . | . | . | . | . | . | . | . | 15 |

Buddy frequencies by distance and coverage for code 30, mean MSL pressure at 9am

| | | 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 | 20 | 6 | 10 | 12 | 14 | 21 | 10 | 10 | 9 | 6 | 13 |
| Y | 85–95 | . | 2 | 1 | 2 | 1 | 1 | 1 | . | . | 3 | 2 |
| Y | 75–85 | . | . | . | . | . | . | . | . | . | 1 | 2 |
| Y | 65–75 | . | . | . | . | . | . | . | . | . | . | 1 |
| Y | 55–65 | 1 | . | . | . | 1 | 1 | . | . | . | . | . |
| Y | 45–55 | . | . | 1 | . | . | . | . | . | . | 1 | 1 |
| Y | 35–45 | . | . | . | . | . | . | . | . | 2 | . | . |
| Y | 25–35 | . | 2 | . | . | . | . | . | . | . | . | 1 |
| Y | 15–25 | . | . | . | . | . | . | . | . | . | . | . |
| Y | 5–15 | 2 | . | . | . | . | . | . | . | . | . | . |
| Y | 0–5 | . | . | . | . | . | . | . | . | . | . | 10 |
| N | 95–100 | 11 | 4 | 5 | . | 3 | 3 | 1 | 2 | 2 | 3 | 58 |
| N | 85–95 | 1 | . | . | . | . | . | . | . | . | . | 6 |
| N | 75–85 | . | . | 1 | . | . | . | 1 | . | . | . | 2 |
| N | 65–75 | . | . | . | . | . | . | . | . | . | . | 3 |
| N | 55–65 | . | . | . | . | . | . | . | . | . | . | . |
| N | 45–55 | . | . | . | . | . | . | . | . | . | . | 2 |
| N | 35–45 | . | . | . | . | . | . | . | . | . | . | . |
| N | 25–35 | . | 1 | . | . | . | . | . | . | . | . | 5 |
| N | 15–25 | . | . | . | . | . | . | . | . | . | . | 1 |
| N | 5–15 | . | . | . | . | . | . | . | . | . | . | . |
| N | 0–5 | . | . | . | . | . | . | . | . | . | . | 37 |

Thus it appears for code 00 that most stations are well covered by buddies lying within 15 km while for code 30 although the coverage was still excellent buddies were generally further away. However, MSL pressure (i.e., code 30) is less spatially variable than rainfall and having wider spaced stations should not undermine the spatially checking. Such tabulations were made for every code and all showed a similar form. A more general set of statistics for each code is given below:

| Co- | No. of | Number with given number of buddies | | | | | | % Cover | | | Dist. to buddy (km) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| de | 1ry stns | Nil | 1 | 2 | 3 | 4 | 5 | Min. | Avg. | Max. | Min. | Avg. | Max. |
| 00 | 3 719 | 16 | 3 703 | 1 815 | 805 | 360 | 172 | 11 | 98 | 100 | 0 | 19 | 666 |
| 01 | 2 983 | 7 | 2 976 | 1 394 | 586 | 262 | 113 | 11 | 98 | 100 | 0 | 19 | 539 |
| 02 | 800 | 17 | 783 | 353 | 134 | 52 | 20 | 11 | 97 | 100 | 0 | 48 | 599 |
| 03 | 791 | 17 | 774 | 344 | 127 | 51 | 21 | 11 | 98 | 100 | 0 | 46 | 599 |
| 04 | 838 | 32 | 806 | 350 | 130 | 52 | 22 | 11 | 98 | 100 | 0 | 50 | 657 |
| 05 | 559 | 16 | 543 | 256 | 102 | 42 | 18 | 21 | 98 | 100 | 0 | 29 | 530 |
| 06 | 763 | 13 | 750 | 328 | 119 | 53 | 24 | 11 | 98 | 100 | 0 | 39 | 539 |
| 07 | 817 | 47 | 770 | 337 | 121 | 54 | 25 | 11 | 97 | 100 | 0 | 49 | 655 |
| 08 | 553 | 15 | 538 | 252 | 100 | 38 | 20 | 21 | 98 | 100 | 0 | 28 | 227 |
| 09 | 200 | 15 | 185 | 84 | 29 | 11 | 4 | 9 | 95 | 100 | 0 | 67 | 600 |
| 10 | 65 | 10 | 55 | 15 | 2 | 0 | 0 | 24 | 93 | 100 | 5 | 53 | 185 |
| 11 | 257 | 7 | 250 | 115 | 40 | 14 | 4 | 44 | 98 | 100 | 0 | 36 | 207 |
| 12 | 232 | 5 | 227 | 103 | 38 | 11 | 2 | 31 | 98 | 100 | 0 | 35 | 199 |
| 13 | 239 | 5 | 234 | 107 | 45 | 18 | 4 | 30 | 98 | 100 | 0 | 40 | 315 |
| 14 | 171 | 6 | 165 | 67 | 23 | 7 | 3 | 44 | 97 | 100 | 0 | 43 | 269 |
| 15 | 265 | 31 | 234 | 106 | 37 | 12 | 4 | 31 | 96 | 100 | 0 | 46 | 281 |
| 16 | 772 | 38 | 734 | 311 | 104 | 35 | 13 | 11 | 98 | 100 | 0 | 40 | 539 |
| 17 | 149 | 25 | 124 | 41 | 6 | 0 | 0 | 27 | 95 | 100 | 0 | 52 | 119 |
| 18 | 245 | 16 | 229 | 95 | 26 | 5 | 1 | 30 | 96 | 100 | 0 | 46 | 281 |
| 19 | 23 | 5 | 18 | 8 | 1 | 1 | 0 | 22 | 87 | 100 | 6 | 53 | 123 |
| 20 | 154 | 9 | 145 | 70 | 24 | 9 | 2 | 44 | 96 | 100 | 0 | 46 | 207 |
| 21 | 568 | 9 | 559 | 232 | 72 | 22 | 6 | 37 | 98 | 100 | 0 | 49 | 616 |
| 22 | 556 | 16 | 540 | 254 | 95 | 35 | 15 | 22 | 98 | 100 | 0 | 29 | 199 |
| 23 | 786 | 47 | 739 | 318 | 117 | 43 | 19 | 11 | 97 | 100 | 0 | 50 | 596 |
| 24 | 594 | 9 | 585 | 262 | 97 | 38 | 18 | 18 | 98 | 100 | 0 | 43 | 539 |
| 25 | 596 | 9 | 587 | 259 | 94 | 38 | 21 | 23 | 97 | 100 | 0 | 43 | 539 |
| 26 | 595 | 9 | 586 | 259 | 98 | 42 | 19 | 22 | 98 | 100 | 0 | 43 | 539 |
| 27 | 557 | 9 | 548 | 195 | 49 | 7 | 2 | 43 | 98 | 100 | 0 | 46 | 616 |
| 28 | 588 | 10 | 578 | 243 | 86 | 32 | 14 | 23 | 98 | 100 | 0 | 44 | 539 |
| 29 | 487 | 10 | 477 | 158 | 36 | 0 | 0 | 56 | 99 | 100 | 0 | 42 | 539 |
| 30 | 323 | 47 | 276 | 104 | 29 | 5 | 0 | 10 | 93 | 100 | 0 | 122 | 599 |
| 31 | 218 | 27 | 191 | 50 | 14 | 1 | 0 | 22 | 95 | 100 | 0 | 50 | 263 |
| 32 | 218 | 27 | 191 | 50 | 14 | 1 | 0 | 22 | 95 | 100 | 0 | 50 | 263 |
| 33 | 315 | 33 | 282 | 75 | 11 | 3 | 0 | 16 | 96 | 100 | 0 | 46 | 305 |
| 34 | 88 | 5 | 83 | 40 | 14 | 3 | 0 | 22 | 93 | 100 | 1 | 50 | 122 |
| 35 | 120 | 9 | 111 | 54 | 19 | 3 | 0 | 69 | 97 | 100 | 1 | 52 | 451 |
| 36 | 78 | 5 | 73 | 33 | 8 | 1 | 0 | 22 | 92 | 100 | 1 | 52 | 123 |
| 37 | 120 | 9 | 111 | 54 | 19 | 3 | 0 | 69 | 97 | 100 | 1 | 52 | 451 |
| 40 | 2 977 | 7 | 2 970 | 1 394 | 580 | 262 | 113 | 11 | 98 | 100 | 0 | 19 | 539 |
| 41 | 2 990 | 7 | 2 983 | 1 413 | 608 | 268 | 118 | 11 | 98 | 100 | 0 | 18 | 539 |
| 42 | 695 | 29 | 666 | 276 | 74 | 24 | 8 | 19 | 97 | 100 | 0 | 43 | 616 |
| 43 | 701 | 10 | 691 | 273 | 89 | 28 | 11 | 11 | 98 | 100 | 0 | 40 | 539 |
| 44 | 756 | 45 | 711 | 283 | 91 | 26 | 12 | 11 | 98 | 100 | 0 | 52 | 596 |
| 45 | 210 | 35 | 175 | 35 | 6 | 0 | 0 | 32 | 94 | 100 | 0 | 51 | 263 |
| 47 | 13 | 9 | 4 | 0 | 0 | 0 | 0 | 72 | 92 | 100 | 60 | 87 | 106 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 48 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | — | — | — | — | — | — |
| 49 | 6 | 4 | 2 | 0 | 0 | 0 | 0 | 54 | 70 | 86 | 46 | 46 | 46 |
| 50 | 259 | 8 | 251 | 103 | 36 | 6 | 2 | 16 | 97 | 100 | 0 | 40 | 451 |
| 51 | 259 | 8 | 251 | 103 | 36 | 6 | 2 | 16 | 97 | 100 | 0 | 40 | 451 |
| 52 | 297 | 7 | 290 | 105 | 32 | 7 | 2 | 16 | 97 | 100 | 0 | 37 | 451 |
| 53 | 297 | 7 | 290 | 105 | 31 | 7 | 2 | 16 | 97 | 100 | 0 | 37 | 451 |
| 54 | 297 | 7 | 290 | 104 | 31 | 7 | 2 | 16 | 97 | 100 | 0 | 37 | 451 |
| 55 | 297 | 7 | 290 | 106 | 31 | 7 | 2 | 16 | 97 | 100 | 0 | 37 | 451 |
| 56 | 297 | 7 | 290 | 107 | 31 | 7 | 2 | 16 | 97 | 100 | 0 | 37 | 451 |
| 57 | 297 | 6 | 291 | 111 | 36 | 8 | 2 | 16 | 97 | 100 | 0 | 37 | 451 |
| 58 | 297 | 6 | 291 | 112 | 36 | 8 | 2 | 16 | 97 | 100 | 0 | 37 | 451 |
| 59 | 297 | 6 | 291 | 112 | 36 | 8 | 2 | 16 | 97 | 100 | 0 | 37 | 451 |
| 60 | 210 | 35 | 175 | 35 | 6 | 0 | 0 | 32 | 94 | 100 | 0 | 51 | 263 |
| 61 | 708 | 11 | 697 | 273 | 86 | 20 | 10 | 11 | 98 | 100 | 0 | 39 | 539 |
| 62 | 688 | 18 | 670 | 262 | 83 | 22 | 6 | 11 | 98 | 100 | 0 | 50 | 539 |
| 63 | 688 | 18 | 670 | 262 | 83 | 22 | 6 | 11 | 98 | 100 | 0 | 50 | 539 |
| 64 | 757 | 38 | 719 | 284 | 71 | 15 | 3 | 11 | 98 | 100 | 0 | 39 | 539 |
| 65 | 830 | 64 | 766 | 302 | 80 | 12 | 2 | 11 | 98 | 100 | 0 | 44 | 616 |

Having established a set of buddies for each station-code, the largest contemporary difference was found for every primary-buddy-code and compared to the mean contemporary difference for the same primary-buddy-code combination, i.e., the ratio MaxDifference/MeanDifference was formed. From all of these ratios the 5% for each code with the largest values were printed out for inspection and possible correction. The tabulation below summarises the numbers of primary-buddy-code combinations involved (e.g., for the rainfall statistics the 809 such combinations found represent 5% of the total number of primary-buddy pairs with a rainfall statistic).

| | Codes | Counts |
|---|---|---|
| Rainfall | 00 01 40 41 | 809 |
| Temperature | 02 03 04 06 07 43 44 61 62 63 | 402 |
| Grass-minimums | 05 08 22 23 | 128 |
| Sunshine & radiation | 09 17 48 49 | 13 |
| Earth-temperature | 10 11 12 13 14 47 | 51 |
| Windrun | 15 18 33 | 43 |
| 9am values | 16 30 42 64 65 | 154 |
| Evaporation | 19 20 34 35 36 37 | 28 |
| Days of occurrence | 21 24 25 26 27 28 29 | 189 |
| Wind gusts | 31 32 45 46 60 | 35 |
| Maximum rainfall | 50 51 52 53 54 55 56 57 58 59 | 154 |
| Total | | 2 006 |

The total here of 2006 possible errors was too large for every one to be closely examined, so only the 2%, rather than 5%, with the largest differences were examined any further. For these, listings, wherever possible, of daily values were printed out to allow a visual check of the daily values in the database.

This examination quickly revealed that values for codes 43 and 44 (the lowest maximum and highest minimum for the month) were sometimes being calculated when they should not be. For example, when a multi-day maximum is present within a month, it is still valid to include such periods in a scan for the highest maximum for the month, but invalid in a scan for the lowest maximum. For this latter statistic to be sensible a maximum for every day of the month must be available. A similar problem existed for the highest minimum. The statistics calculation procedures were modified to prevent both the lowest maximum and highest minimum statistics being calculated for station-months that included multi-day maximum or minimum temperatures.

This was the only procedural error highlighted, but sometimes it was obvious, or suspected, that one or two daily values had been miscopied from the original paper forms. Thus the daily value listings, and the monthly statistics difference listings, were used to compare the values held in **MTHLY_STATS** with those on the original paper forms. Corrections were then made to the database daily or monthly statistics data as appropriate. The tabulation below summarises how many changes were made to both the base data and to **MTHLY_STATS**.

| | Daily values | | Monthly values | |
|---|---|---|---|---|
| | Deleted | Amended | Deleted | Amended |
| Rainfall | 916 | 214 | 7 | 4 |
| Temperature | 55 | 192 | 2 | 6 |
| Grass-minimums | 0 | 12 | 47 | 7 |
| Sunshine & radiation | 41 | 61 | 0 | 3 |
| Earth-temperature | 5 | 4 | 0 | 4 |
| Windrun | 1 | 11 | 0 | 3 |
| 9am Values | 178 | 67 | 0 | 5 |
| Evaporation | 2 | 1 | 1 | 0 |
| Days of occurrence | 11 376 | 0 | 2 497 | 0 |
| Wind gusts | 0 | 3 | 4 | 0 |
| Maximum rainfall | 0 | 0 | 0 | 2 |

Some example of errors found and corrections made are given below.

## Rainfall

- Large number of raindays compared to buddy values
  For defining a rainday observers should use 0.1 mm as a threshold amount, but some have used a "Trace", i.e., some rain was observed but measured to be under 0.05 mm.
- Low total for month compared to buddy values
  Station opening part way through the month or telemetry problem.
- Large daily rainfalls interspersed with several dry days whereas rain reported at buddy
  These were candidates for unacknowledged accumulated falls but it is hard to prove unless comments were made on the forms. One station had only a couple of readings but noted "in hospital between Sept 19 to Oct 28th 1993". Another, for the period 3–23 May 1942, had "away on Home Guard instruction course". A Pacific Island station for seven months reported rainfall every seven days with zeroes inbetween; its maximum 1 day rainfalls were useless.
- Values needed adjusting
  In two cases values were in millimetres on the form but were converted again to millimetres.
  In another "taken with new measuring glass" which was for a 5 inch gauge though the gauge was a 10 inch gauge.
- Genuine high rainfalls which looked too large compared to buddy values
  387.4 mm at D14481 on 27 December 1939 — confirmed in memo from the lighthouse keeper about damage to roads etc.
  188 mm at H40272 on 1 December 1952 — confirmed by comments about floodwaters, and noted as the record 24 h fall for 1907–52.

## Temperature

- B75381 only occasionally reset the minimum thermometer from late December 1970 to the end of February 1971. New lower minimums have been accepted, but those equal to the previous day's minimum have been deleted.

- Some long periods of accumulation (occasionally thousands of hours long) were found in the daily data. These were introduced during the recovery of data from Trentham. DATA FAULT 1998/030 was raised, and actioned to remove periods of 672 hours or more.

### Sunshine & radiation

- Dunedin data from 1914 were held as I50951 although that station began recording sunshine only from December 1947 and earlier data belong to I50852 and I50856. Further checks also showed that temperature data attributed to I50951 belonged to I50852, I50855, I50856, and I50857. DATA FAULT 1998/032 was raised, the data moved, and **LAND_STATION** and **SITE_CHANGES** amended.
- G13222 and G13231 both had the same sunshine data over a considerable period. DATA FAULT 1998/035 was raised and the data accredited to the correct station and removed from the other.
- I59065 for December 1995 to October 1996 was low by a factor of ten. DATA FAULT 1998/033 was raised and the corrected raw data were re-archived. This was possible because the station is part of NIWA's network of automatic climate stations (CLITEL) and so the raw data were still available.

### 9 am values

- Wet-bulb wicks can dry out if not properly maintained leading to the dry-bulb and wet-bulb temperatures being identical and giving a false RH of 100% and consequently monthly values that are too humid compared to buddies.
- A few frequently occurring values for pressure, such as 981.6, appeared in many synop reports during July and August 1978. This was when synops were first automatically stored in computer files.

### Days of occurrences

- A new form for climate observers to complete with their daily observations was introduced in January 1972. This had a provision for recording the occurrence of various phenomena during a day. But since only some stations recorded these "days of occurrence", there was considerable difficulty determining whether a station had zero occurrences (i.e., none of the phenomena occurred during the month) or it was a station not recording occurrences. An indication had to be made on the form before data-entry that the occurrences were not being recorded since the default condition was that occurrences were being monitored but none were observed. Over 1972–73, the stations which were consistently reporting these occurrences were determined and the errors flagged in this audit often reflected the indecision, in that initial period, about a station's ability.

### Miscellaneous

- All climate data elements for C75832 for November 1976 had been entered from the December 1976 climate return form so the November 1976 form was submitted to the data-entry process.
- In **LAND_STATION** the longitude for the automatic station at Invercargill was corrected.

Obviously, many of the inspected values were correct and, if this checking procedure were to be performed again, they would re-appear 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 **STATS_DIFFS** which was created by this checking procedure and has the following structure:

| Column name | Null? | Type |
|---|---|---|
| SCODE | | VARCHAR2(2) |
| AGENT_NO | | NUMBER |
| BUDDY | | NUMBER |
| DIST | | NUMBER |
| OBS_DATE | NOT NULL | DATE |
| PERC | | NUMBER |
| P_VALUE | | NUMBER |
| B_VALUE | | NUMBER |

For each SCODE, AGENT_NO, and BUDDY the values with the greatest difference occurred at OBS_DATE and are held in P_VALUE and B_VALUE and PERC holds the percentile of this set's maximum to mean difference. For example, those with PERC equal to one for a given SCODE are the 1% of all the AGENT_NO-BUDDY pairs for that SCODE which have the greatest relative difference. Thus, on a rerun the contents of **STATS_DIFFS** can be moved to **OLD_STATS_DIFFS**, say, before being over-written and rows common to both tables (except PERC which may change between runs) can be ignored.

## Details and results of Check C.4 — are all statistics records without gaps?

Ideally, for a given AGENT_NO and STATS_CODE there should be no breaks in the record from when it started until either the present day or when the station closed or ceased to report the base data associated with the given code. Such a situation is extremely rare since missing data occur at even the best stations.

This check is more a quality check than a search for errors and the following tabulation shows the distribution of the number of gaps with statistics code. For example, for code 00 there were 1319 stations that had no gaps in their records and there were 2164 stations that had between one and nine gaps in their records etc.

| | Counts of stations classified by the number of gaps (i.e., 1–9 etc.) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Code | 0 | 1–9 | 10–19 | 20–29 | 30–39 | 40–49 | 50–59 | 60–69 | 70–79 | 80–89 | >89 |
| 00 | 1 319 | 2 164 | 188 | 39 | 8 | . | 1 | . | . | . | . |
| 01 | 291 | 1 282 | 594 | 315 | 189 | 113 | 78 | 61 | 20 | 17 | 22 |
| 02 | 119 | 484 | 138 | 44 | 12 | 2 | 1 | . | . | . | . |
| 03 | 151 | 503 | 108 | 24 | 4 | 1 | . | . | . | . | . |
| 04 | 159 | 554 | 99 | 18 | 4 | 2 | . | . | . | . | . |
| 05 | 83 | 338 | 104 | 27 | 7 | . | . | . | . | . | . |
| 06 | 127 | 482 | 117 | 28 | 7 | 2 | . | . | . | . | . |
| 07 | 135 | 505 | 130 | 35 | 9 | 3 | . | . | . | . | . |
| 08 | 71 | 324 | 117 | 30 | 8 | 2 | 1 | . | . | . | . |
| 09 | 63 | 102 | 13 | 3 | . | . | . | . | . | . | . |
| 10 | 28 | 31 | 2 | . | . | . | . | . | . | . | . |
| 11 | 72 | 159 | 14 | 2 | . | . | . | . | . | . | . |
| 12 | 55 | 152 | 18 | 2 | . | . | . | . | . | . | . |
| 13 | 103 | 118 | 10 | 1 | . | . | . | . | . | . | . |
| 14 | 64 | 94 | 8 | . | . | . | . | . | . | . | . |
| 15 | 59 | 141 | 49 | 7 | . | . | . | . | . | . | . |
| 16 | 137 | 535 | 85 | 14 | 1 | . | . | . | . | . | . |
| 17 | 32 | 101 | 10 | 1 | . | . | . | . | . | . | . |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 18 | 27 | 127 | 63 | 17 | . | 1 | . | . | . | . | . |
| 19 | 5 | 13 | 1 | 3 | . | . | . | . | . | . | . |
| 20 | 31 | 100 | 11 | 1 | . | . | . | . | . | . | . |
| 21 | 207 | 356 | 5 | . | . | . | . | . | . | . | . |
| 22 | 58 | 281 | 141 | 52 | 13 | 8 | 1 | 1 | . | . | . |
| 23 | 99 | 483 | 146 | 37 | 16 | 4 | 1 | . | . | . | . |
| 24 | 227 | 359 | 8 | . | . | . | . | . | . | . | . |
| 25 | 220 | 364 | 11 | 1 | . | . | . | . | . | . | . |
| 26 | 221 | 363 | 10 | 1 | . | . | . | . | . | . | . |
| 27 | 213 | 342 | 2 | . | . | . | . | . | . | . | . |
| 28 | 193 | 386 | 8 | 1 | . | . | . | . | . | . | . |
| 29 | 243 | 243 | 1 | . | . | . | . | . | . | . | . |
| 30 | 82 | 200 | 31 | 4 | 2 | . | . | . | . | . | . |
| 31 | 36 | 113 | 61 | 7 | 1 | . | . | . | . | . | . |
| 32 | 36 | 113 | 61 | 7 | 1 | . | . | . | . | . | . |
| 33 | 118 | 133 | 10 | 1 | . | . | . | . | . | . | . |
| 34 | 19 | 55 | 12 | 2 | . | . | . | . | . | . | . |
| 35 | 24 | 76 | 18 | 1 | 1 | . | . | . | . | . | . |
| 36 | 15 | 48 | 13 | 2 | . | . | . | . | . | . | . |
| 37 | 24 | 76 | 18 | 1 | 1 | . | . | . | . | . | . |
| 40 | 284 | 1273 | 600 | 320 | 190 | 111 | 77 | 65 | 19 | 15 | 22 |
| 41 | 385 | 1539 | 524 | 260 | 130 | 57 | 42 | 26 | 12 | 5 | 8 |
| 42 | 219 | 191 | 31 | 11 | 1 | . | . | . | . | . | . |
| 43 | 119 | 456 | 95 | 25 | 4 | 2 | . | . | . | . | . |
| 44 | 135 | 464 | 117 | 30 | 7 | 3 | . | . | . | . | . |
| 45 | 23 | 87 | 73 | 22 | 4 | 1 | . | . | . | . | . |
| 46 | 23 | 87 | 73 | 22 | 4 | 1 | . | . | . | . | . |
| 47 | 2 | 5 | 5 | . | . | . | . | . | . | . | . |
| 48 | 0 | 4 | . | . | . | . | . | . | . | . | . |
| 49 | 0 | 3 | 1 | . | . | . | . | . | . | . | . |
| 50 | 58 | 89 | 72 | 33 | 6 | 1 | . | . | . | . | . |
| 51 | 58 | 88 | 72 | 33 | 7 | 1 | . | . | . | . | . |
| 52 | 57 | 106 | 76 | 44 | 13 | 1 | . | . | . | . | . |
| 53 | 58 | 104 | 79 | 42 | 13 | 1 | . | . | . | . | . |
| 54 | 58 | 104 | 78 | 44 | 11 | 2 | . | . | . | . | . |
| 55 | 59 | 106 | 84 | 34 | 13 | 1 | . | . | . | . | . |
| 56 | 60 | 107 | 82 | 36 | 11 | 1 | . | . | . | . | . |
| 57 | 92 | 169 | 29 | 5 | 2 | . | . | . | . | . | . |
| 58 | 94 | 174 | 22 | 5 | 2 | . | . | . | . | . | . |
| 59 | 95 | 173 | 25 | 2 | 2 | . | . | . | . | . | . |
| 60 | 23 | 87 | 73 | 22 | 4 | 1 | . | . | . | . | . |
| 61 | 95 | 301 | 89 | 31 | 5 | 2 | 1 | . | . | . | . |
| 62 | 43 | 142 | 147 | 76 | 48 | 25 | 15 | 4 | 3 | . | . |
| 63 | 43 | 142 | 147 | 76 | 48 | 25 | 15 | 4 | 3 | . | . |
| 64 | 148 | 518 | 79 | 11 | 1 | . | . | . | . | . | . |
| 65 | 223 | 507 | 65 | 9 | 2 | . | . | . | . | . | . |

It is not easy to assess **MTHLY_STATS**'s quality from the above tabulation and the one below provides some further details. For each code the number of stations with a record and how many of those are perfect is shown. Stations with only a few gaps (up to three, say) are also of high quality and the numbers of such stations are shown together with the percentage that they and the perfect ones are of the total. This percentage is two or three times that of the perfect ones alone and ranges from 15% to 90%, but most codes have 35% to 75% of the stations with at most three gaps. The maximum number of gaps for each code is also shown to range from 9 to 159, but with most between 20 and 70. The last three columns deal with the worst stations for each code showing what percentage of the total have records under 50% complete and what the equivalent number is for that and for under 10% complete. It can be seen that about half the codes have less than 4% of the stations with records under 50% complete and most have more than 90% of the stations better than that; also only one in every 200 to 300 stations has a record less than 10% complete.

| Code | Total no. of stns. | No. with no gaps | % of total | Number with given number of gaps 1 | 2 | 3 | % with 0,1,2, or 3 gaps | Max. no. of gaps | % of total <50 % com-plete | Number with given limit of completion (%) <50 | <10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 00 | 3 719 | 1 319 | 35.5 | 743 | 421 | 287 | 74.5 | 51 | 2.6 | 98 | 7 |
| 01 | 2 983 | 291 | 9.8 | 203 | 200 | 179 | 29.3 | 158 | 9.4 | 280 | 16 |
| 02 | 800 | 119 | 14.9 | 116 | 75 | 63 | 46.6 | 59 | 3.5 | 28 | 0 |
| 03 | 791 | 151 | 19.1 | 134 | 83 | 69 | 55.2 | 47 | 2.4 | 19 | 0 |
| 04 | 838 | 159 | 19.0 | 147 | 98 | 74 | 57.0 | 42 | 3.2 | 27 | 0 |
| 05 | 559 | 83 | 14.8 | 81 | 47 | 42 | 45.3 | 38 | 1.6 | 9 | 0 |
| 06 | 763 | 127 | 16.6 | 115 | 76 | 57 | 49.1 | 45 | 4.8 | 37 | 2 |
| 07 | 817 | 135 | 16.5 | 122 | 83 | 68 | 49.9 | 49 | 8.9 | 73 | 3 |
| 08 | 553 | 71 | 12.8 | 65 | 46 | 43 | 40.7 | 50 | 3.3 | 18 | 1 |
| 09 | 200 | 63 | 31.5 | 41 | 8 | 17 | 64.5 | 25 | 3.5 | 7 | 0 |
| 10 | 65 | 28 | 43.1 | 13 | 6 | 2 | 75.4 | 14 | 1.5 | 1 | 0 |
| 11 | 257 | 72 | 28.0 | 42 | 32 | 25 | 66.5 | 23 | 1.6 | 4 | 1 |
| 12 | 232 | 55 | 23.7 | 32 | 29 | 17 | 57.3 | 25 | 0.9 | 2 | 0 |
| 13 | 239 | 103 | 43.1 | 46 | 25 | 14 | 78.7 | 22 | 0.4 | 1 | 1 |
| 14 | 171 | 64 | 37.4 | 35 | 18 | 10 | 74.3 | 15 | 0.6 | 1 | 1 |
| 15 | 265 | 59 | 22.3 | 43 | 39 | 23 | 61.9 | 25 | 2.3 | 6 | 0 |
| 16 | 772 | 137 | 17.7 | 143 | 107 | 71 | 59.3 | 34 | 5.2 | 40 | 4 |
| 17 | 149 | 32 | 21.5 | 27 | 24 | 12 | 63.8 | 23 | 4.0 | 6 | 1 |
| 18 | 245 | 27 | 11.0 | 27 | 32 | 20 | 43.3 | 41 | 2.4 | 6 | 0 |
| 19 | 23 | 5 | 21.7 | 1 | 3 | 2 | 47.8 | 27 | 0.0 | 0 | 0 |
| 20 | 154 | 31 | 20.1 | 33 | 20 | 12 | 62.3 | 26 | 3.2 | 5 | 0 |
| 21 | 568 | 207 | 36.4 | 142 | 76 | 52 | 84.0 | 15 | 7.9 | 45 | 3 |
| 22 | 556 | 58 | 10.4 | 46 | 37 | 41 | 32.7 | 63 | 4.9 | 27 | 1 |
| 23 | 786 | 99 | 12.6 | 99 | 88 | 58 | 43.8 | 52 | 9.9 | 78 | 4 |
| 24 | 594 | 227 | 38.2 | 135 | 74 | 57 | 83.0 | 18 | 5.9 | 35 | 2 |
| 25 | 596 | 220 | 36.9 | 142 | 81 | 52 | 83.1 | 20 | 6.0 | 36 | 2 |
| 26 | 595 | 221 | 37.1 | 142 | 77 | 52 | 82.7 | 21 | 6.4 | 38 | 2 |
| 27 | 557 | 213 | 38.2 | 137 | 74 | 52 | 85.5 | 15 | 10.4 | 58 | 7 |
| 28 | 588 | 193 | 32.8 | 140 | 88 | 56 | 81.1 | 26 | 7.5 | 44 | 4 |
| 29 | 487 | 243 | 49.9 | 108 | 43 | 47 | 90.6 | 16 | 7.6 | 37 | 5 |
| 30 | 323 | 82 | 25.4 | 51 | 39 | 22 | 60.1 | 34 | 5.9 | 19 | 0 |
| 31 | 218 | 36 | 16.5 | 17 | 18 | 15 | 39.4 | 37 | 9.6 | 21 | 1 |
| 32 | 218 | 36 | 16.5 | 17 | 16 | 15 | 38.5 | 37 | 9.6 | 21 | 1 |
| 33 | 315 | 118 | 37.5 | 37 | 27 | 16 | 62.9 | 25 | 6.0 | 19 | 0 |
| 34 | 88 | 19 | 21.6 | 13 | 11 | 5 | 54.5 | 23 | 0.0 | 0 | 0 |
| 35 | 120 | 24 | 20.0 | 17 | 11 | 11 | 52.5 | 30 | 0.0 | 0 | 0 |
| 36 | 78 | 15 | 19.2 | 11 | 10 | 4 | 51.3 | 23 | 1.3 | 1 | 0 |
| 37 | 120 | 24 | 20.0 | 17 | 11 | 11 | 52.5 | 30 | 0.0 | 0 | 0 |
| 40 | 2 977 | 284 | 9.5 | 203 | 200 | 173 | 28.9 | 159 | 9.4 | 279 | 15 |
| 41 | 2 990 | 385 | 12.9 | 299 | 223 | 222 | 37.8 | 126 | 5.7 | 170 | 12 |
| 42 | 695 | 219 | 31.5 | 53 | 44 | 19 | 48.2 | 35 | 6.5 | 45 | 3 |
| 43 | 701 | 119 | 17.0 | 101 | 82 | 58 | 51.4 | 45 | 5.4 | 38 | 2 |
| 44 | 756 | 135 | 17.9 | 100 | 81 | 65 | 50.4 | 46 | 9.9 | 75 | 3 |
| 45 | 210 | 23 | 11.0 | 14 | 12 | 14 | 30.0 | 40 | 16.2 | 34 | 0 |
| 46 | 210 | 23 | 11.0 | 14 | 12 | 14 | 30.0 | 40 | 16.2 | 34 | 0 |
| 47 | 13 | 2 | 15.4 | 1 | 0 | 0 | 23.1 | 15 | 0.0 | 0 | 0 |
| 48 | 4 | 0 | 0.0 | 0 | 0 | 0 | 0.0 | 9 | 25.0 | 1 | 0 |
| 49 | 6 | 0 | 0.0 | 0 | 1 | 0 | 16.7 | 10 | 16.7 | 1 | 0 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 259 | 58 | 22.4 | 22 | 14 | 14 | 41.7 | 42 | 2.7 | 7 | 1 |
| 51 | 259 | 58 | 22.4 | 20 | 14 | 16 | 41.7 | 42 | 2.7 | 7 | 1 |
| 52 | 297 | 57 | 19.2 | 25 | 15 | 12 | 36.7 | 42 | 4.0 | 12 | 0 |
| 53 | 297 | 58 | 19.5 | 22 | 17 | 11 | 36.4 | 41 | 4.0 | 12 | 0 |
| 54 | 297 | 58 | 19.5 | 25 | 13 | 14 | 37.0 | 41 | 4.4 | 13 | 0 |
| 55 | 297 | 59 | 19.9 | 26 | 13 | 14 | 37.7 | 42 | 4.0 | 12 | 0 |
| 56 | 297 | 60 | 20.2 | 27 | 14 | 10 | 37.4 | 43 | 4.0 | 12 | 0 |
| 57 | 297 | 92 | 31.0 | 40 | 33 | 26 | 64.3 | 34 | 2.0 | 6 | 0 |
| 58 | 297 | 94 | 31.6 | 39 | 33 | 28 | 65.3 | 33 | 2.0 | 6 | 0 |
| 59 | 297 | 95 | 32.0 | 40 | 33 | 26 | 65.3 | 32 | 2.0 | 6 | 0 |
| 60 | 210 | 23 | 11.0 | 14 | 12 | 14 | 30.0 | 41 | 16.2 | 34 | 0 |
| 61 | 708 | 95 | 13.4 | 65 | 37 | 45 | 34.2 | 57 | 4.2 | 30 | 0 |
| 62 | 688 | 43 | 6.3 | 16 | 21 | 20 | 14.5 | 72 | 14.0 | 96 | 7 |
| 63 | 688 | 43 | 6.3 | 16 | 21 | 20 | 14.5 | 72 | 14.0 | 96 | 7 |
| 64 | 757 | 148 | 19.6 | 140 | 103 | 73 | 61.3 | 34 | 5.4 | 41 | 4 |
| 65 | 830 | 223 | 26.9 | 150 | 99 | 71 | 65.4 | 34 | 5.3 | 44 | 3 |

The worst station for any code is that which has the most gaps rather than most missing months. 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 the station had to be closed occasionally but was otherwise a regular observer. For each code the station with the most gaps was determined and these are tabulated below. For example, for code 00 AGENT_NO 6361 had a record 406 months long of which 141 were missing and so the record was 65% complete; there were 24 single months missing, 9 periods either 2 or 3 months long, and the longest missing period was 12 months.

| Co -de | Stn. | No. of mon-ths | % com-plete | No. of missing periods for the given month lengths | | | | | | | Long-est miss-ing period | Total miss-ing months |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 1 | 2–3 | 4–6 | 7–12 | 13–24 | 25–48 | >48 | | |
| 00 | 6361 | 406 | 65 | 24 | 9 | 16 | 2 | . | . | . | 12 | 141 |
| 01 | 1164 | 1 010 | 73 | 97 | 45 | 15 | 1 | . | . | . | 7 | 278 |
| 02 | 6168 | 698 | 75 | 49 | 8 | . | . | 1 | . | 1 | 89 | 177 |
| 03 | 5545 | 319 | 72 | 29 | 12 | 5 | 1 | . | . | . | 7 | 89 |
| 04 | 5545 | 319 | 75 | 28 | 8 | 4 | 2 | . | . | . | 9 | 80 |
| 05 | 4591 | 816 | 84 | 21 | 12 | 3 | 1 | . | . | 1 | 59 | 132 |
| 06 | 4107 | 763 | 91 | 31 | 12 | 2 | . | . | . | . | 4 | 66 |
| 07 | 6168 | 698 | 76 | 37 | 10 | . | . | 1 | . | 1 | 89 | 170 |
| 08 | 4591 | 815 | 82 | 30 | 14 | 5 | . | . | . | 1 | 58 | 147 |
| 09 | 5952 | 193 | 71 | 12 | 9 | 3 | 1 | . | . | . | 10 | 56 |
| 10 | 1819 | 159 | 59 | 4 | 4 | 2 | 3 | 1 | . | . | 14 | 66 |
| 11 | 2376 | 157 | 75 | 14 | 7 | 2 | . | . | . | . | 5 | 39 |
| 12 | 2376 | 157 | 77 | 18 | 6 | 1 | . | . | . | . | 4 | 36 |
| 13 | 2376 | 157 | 81 | 15 | 7 | . | . | . | . | . | 3 | 30 |
| 14 | 1819 | 159 | 64 | 6 | 3 | 3 | 3 | . | . | . | 12 | 58 |
| 15 | 1134 | 150 | 61 | 9 | 11 | 4 | 1 | . | . | . | 9 | 59 |
| 16 | 3949 | 139 | 46 | 16 | 11 | 6 | 1 | . | . | . | 8 | 75 |
| 17 | 6044 | 231 | 48 | 12 | 5 | 2 | 1 | 2 | 1 | . | 36 | 120 |
| 18 | 5867 | 328 | 68 | 23 | 14 | 2 | 1 | . | 1 | . | 36 | 106 |
| 19 | 5683 | 326 | 88 | 18 | 7 | 2 | . | . | . | . | 4 | 40 |
| 20 | 4461 | 334 | 53 | 2 | 3 | 11 | 9 | 1 | . | . | 23 | 159 |
| 21 | 6222 | 216 | 29 | . | . | . | 13 | 2 | . | . | 22 | 155 |
| 22 | 4591 | 814 | 81 | 29 | 22 | 9 | 3 | . | . | . | 11 | 153 |
| 23 | 4630 | 301 | 63 | 29 | 13 | 9 | 1 | . | . | . | 9 | 111 |
| 24 | 5608 | 531 | 85 | 7 | 2 | 3 | 6 | . | . | . | 12 | 80 |
| 25 | 5608 | 531 | 80 | 8 | 2 | 3 | 6 | 1 | . | . | 24 | 105 |
| 26 | 5608 | 531 | 73 | 9 | 2 | 3 | 6 | . | . | 1 | 60 | 142 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 27 | 6222 | 216 | 29 | . | . | . | 13 | 2 | . | . | 22 | 155 |
| 28 | 5608 | 531 | 78 | 13 | 2 | 4 | 6 | 1 | . | . | 24 | 117 |
| 29 | 6222 | 216 | 28 | 1 | . | . | 13 | 2 | . | . | 22 | 156 |
| 30 | 3949 | 136 | 50 | 18 | 11 | 3 | 2 | . | . | . | 8 | 69 |
| 31 | 3375 | 173 | 39 | 17 | 9 | 8 | 3 | . | . | . | 12 | 106 |
| 32 | 3375 | 173 | 39 | 17 | 9 | 8 | 3 | . | . | . | 12 | 106 |
| 33 | 3223 | 243 | 40 | 8 | 8 | 2 | 5 | . | 2 | . | 34 | 146 |
| 34 | 4591 | 513 | 93 | 15 | 7 | 1 | . | . | . | . | 5 | 35 |
| 35 | 4591 | 604 | 92 | 18 | 10 | 2 | . | . | . | . | 6 | 50 |
| 36 | 4591 | 513 | 93 | 15 | 7 | 1 | . | . | . | . | 5 | 35 |
| 37 | 4591 | 604 | 92 | 18 | 10 | 2 | . | . | . | . | 6 | 50 |
| 40 | 1164 | 1 010 | 72 | 98 | 45 | 15 | 1 | . | . | . | 7 | 279 |
| 41 | 5083 | 945 | 78 | 79 | 39 | 6 | 2 | . | . | . | 8 | 210 |
| 42 | 4396 | 236 | 49 | 20 | 6 | 1 | 7 | 1 | . | . | 19 | 120 |
| 43 | 4107 | 763 | 91 | 31 | 12 | 2 | . | . | . | . | 4 | 66 |
| 44 | 6168 | 315 | 83 | 39 | 7 | . | . | . | . | . | 3 | 54 |
| 45 | 4420 | 317 | 78 | 26 | 11 | 1 | 2 | . | . | . | 9 | 70 |
| 46 | 4420 | 317 | 78 | 26 | 11 | 1 | 2 | . | . | . | 9 | 70 |
| 47 | 3551 | 146 | 87 | 11 | 4 | . | . | . | . | . | 2 | 19 |
| 48 | 6237 | 110 | 44 | . | . | 3 | 6 | . | . | . | 10 | 62 |
| 49 | 6237 | 110 | 55 | 1 | . | 8 | 1 | . | . | . | 9 | 50 |
| 50 | 2944 | 271 | 72 | 28 | 11 | 2 | . | 1 | . | . | 13 | 76 |
| 51 | 2944 | 271 | 72 | 28 | 11 | 2 | . | 1 | . | . | 13 | 76 |
| 52 | 2944 | 271 | 72 | 28 | 11 | 2 | . | 1 | . | . | 13 | 76 |
| 53 | 2944 | 271 | 72 | 27 | 11 | 2 | . | 1 | . | . | 13 | 75 |
| 54 | 2944 | 271 | 72 | 27 | 11 | 2 | . | 1 | . | . | 13 | 75 |
| 55 | 2944 | 271 | 74 | 31 | 8 | 2 | . | 1 | . | . | 13 | 72 |
| 56 | 2944 | 271 | 73 | 32 | 8 | 2 | 1 | . | . | . | 12 | 73 |
| 57 | 1896 | 263 | 60 | 25 | 6 | . | 1 | 1 | 1 | . | 36 | 107 |
| 58 | 1896 | 263 | 60 | 24 | 6 | . | 1 | 1 | 1 | . | 36 | 106 |
| 59 | 1896 | 263 | 60 | 23 | 6 | . | 1 | 1 | 1 | . | 36 | 105 |
| 60 | 4420 | 317 | 78 | 27 | 11 | 1 | 2 | . | . | . | 9 | 71 |
| 61 | 6168 | 316 | 80 | 49 | 8 | . | . | . | . | . | 2 | 65 |
| 62 | 1565 | 544 | 79 | 46 | 21 | 4 | 1 | . | . | . | 7 | 116 |
| 62 | 4107 | 762 | 83 | 47 | 20 | 3 | 1 | 1 | . | . | 14 | 128 |
| 63 | 1565 | 544 | 79 | 46 | 21 | 4 | 1 | . | . | . | 7 | 116 |
| 63 | 4107 | 762 | 83 | 47 | 20 | 3 | 1 | 1 | . | . | 14 | 128 |
| 64 | 3949 | 139 | 46 | 16 | 11 | 6 | 1 | . | . | . | 8 | 75 |
| 65 | 3949 | 139 | 48 | 18 | 9 | 6 | 1 | . | . | . | 8 | 73 |

It can be seen above that a different station does not occur for every code. Furthermore, AGENT_NO 3949 occurs four times (codes 16, 30, 64, and 65) and has some of the lowest complete-percentages. The tabulation below provides a summary of its observing history after its synoptic reports, as well as its rain observations, became available. When only rain was available, it reported without fail from February 1891, apart from a few years when it was temporarily closed, up to December 1977. The electronic archiving of all synoptic reports began during 1978 and the tabulation below illustrates the general difficulty with which statistics have been extracted from synops.

The numbers of **missing** months by year and statistic group for AGENT_NO 3949

| | GROUP | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| YEAR | RA | TE | GM | SS | ET | WR | VP | RD | EV | OC | MP | WD | CC | CL | MR |
| 1978 | 0 | . | . | . | . | . | 11 | . | . | . | 11 | . | . | 11 | . |
| 1979 | 0 | . | . | . | . | . | 6 | . | . | . | 6 | . | . | 6 | . |
| 1980 | 0 | . | . | . | . | . | 5 | . | . | . | 5 | . | . | 5 | . |
| 1981 | 0 | . | . | . | . | . | 6 | . | . | . | 6 | . | . | 6 | . |
| 1982 | 2 | . | . | . | . | . | 10 | . | . | . | 10 | . | . | 9 | . |
| 1983 | 0 | . | . | . | . | . | 5 | . | . | . | 5 | . | . | 5 | . |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1984 | 3 | . | . | . | . | . | 7 | . | . | . | 9 | . | . | 7 | . |
| 1985 | 2 | . | . | . | . | . | 6 | . | . | . | 6 | . | . | 5 | . |
| 1986 | 1 | . | . | . | . | . | 8 | . | . | . | 6 | . | . | 5 | . |
| 1987 | 1 | . | . | . | . | . | 3 | . | . | . | 2 | . | . | 2 | . |
| 1988 | 2 | . | . | . | . | . | 6 | . | . | . | 5 | . | . | 4 | . |
| 1989 | 0 | . | . | . | . | . | 6 | . | . | . | 6 | . | . | 6 | . |
| 1990 | 8 | 11 | 11 | . | . | . | 10 | . | . | 11 | 11 | . | . | 10 | . |

RA, rainfall; TP, temperature; GM, grass minimum temperature; SS, sunshine; ET, earth temperature; WR, wind run; VP, vapour pressure and relative humidity; RD, radiation; EV, evaporation; OC, days of occurrence; MP, mean sea level pressure; WD, wind; CC, calculated statistics; CL, cloud; and MR, maximum rain.

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

For a given AGENT_NO and STATS_CODE the record should be long enough to establish the mean level and variability of the climate with respect to the place and code concerned. Longer records can be used to track any trends in the climate and 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.

The following tabulation shows the cumulative distributions of record length for each code as percentiles. For example, remembering the percentiles have been rounded to the nearest percent, for code 00 under 0.5% of 3721 (i.e., about 18) stations have a record only 1 month long while about 18% (i.e., 670) have records of at most 5 years. For 1 month the percentile ranges from less than 0.5 to 5 while, generally, 80% have records of at least 5 years, but only 20% have records longer than 20–30 years.

| | Percentile for given record length | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Months | | | Years | | | | | | |
| Code | 1 | ≤3 | ≤6 | ≤1 | ≤2 | ≤5 | ≤10 | ≤20 | ≤40 | Total |
| 00 | 0 | 1 | 2 | 4 | 7 | 18 | 34 | 54 | 78 | 3 721 |
| 01 | 1 | 2 | 3 | 6 | 11 | 24 | 40 | 60 | 83 | 2 985 |
| 02 | 1 | 2 | 3 | 6 | 11 | 23 | 42 | 62 | 86 | 801 |
| 03 | 1 | 2 | 3 | 6 | 11 | 23 | 45 | 66 | 86 | 792 |
| 04 | 1 | 2 | 3 | 6 | 10 | 25 | 48 | 67 | 87 | 840 |
| 05 | 1 | 3 | 4 | 6 | 11 | 23 | 40 | 62 | 86 | 559 |
| 06 | 1 | 2 | 3 | 6 | 12 | 25 | 47 | 66 | 86 | 764 |
| 07 | 1 | 4 | 6 | 11 | 17 | 30 | 50 | 68 | 88 | 818 |
| 08 | 1 | 3 | 5 | 8 | 13 | 24 | 40 | 61 | 86 | 553 |
| 09 | 2 | 2 | 4 | 5 | 9 | 17 | 33 | 56 | 80 | 200 |
| 10 | 2 | 3 | 5 | 8 | 20 | 37 | 72 | 88 | 98 | 65 |
| 11 | 2 | 3 | 4 | 7 | 13 | 27 | 46 | 71 | 92 | 257 |
| 12 | 2 | 2 | 4 | 8 | 15 | 26 | 44 | 69 | 91 | 232 |
| 13 | 0 | 1 | 3 | 5 | 10 | 21 | 41 | 66 | 88 | 239 |
| 14 | 2 | 3 | 4 | 6 | 13 | 27 | 51 | 73 | 90 | 171 |
| 15 | 0 | 3 | 3 | 8 | 15 | 38 | 63 | 79 | 92 | 266 |
| 16 | 1 | 3 | 5 | 8 | 14 | 27 | 47 | 78 | 93 | 774 |
| 17 | 2 | 3 | 5 | 11 | 18 | 41 | 75 | 89 | 99 | 148 |
| 18 | 1 | 4 | 5 | 9 | 16 | 43 | 68 | 84 | 97 | 246 |
| 19 | 4 | 4 | 4 | 4 | 9 | 13 | 26 | 48 | 100 | 23 |
| 20 | 1 | 2 | 2 | 6 | 12 | 27 | 51 | 83 | 99 | 154 |
| 21 | 3 | 4 | 7 | 10 | 18 | 30 | 46 | 69 | 96 | 568 |
| 22 | 1 | 4 | 5 | 9 | 13 | 26 | 42 | 62 | 87 | 556 |
| 23 | 1 | 3 | 5 | 9 | 15 | 30 | 50 | 71 | 89 | 787 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 24 | 2 | 3 | 5 | 7 | 14 | 28 | 45 | 66 | 89 | 594 |
| 25 | 2 | 4 | 5 | 7 | 14 | 28 | 45 | 66 | 90 | 596 |
| 26 | 2 | 3 | 5 | 7 | 14 | 27 | 45 | 66 | 89 | 595 |
| 27 | 3 | 4 | 7 | 13 | 21 | 37 | 55 | 93 | 99 | 557 |
| 28 | 2 | 3 | 5 | 9 | 16 | 31 | 48 | 76 | 93 | 588 |
| 29 | 2 | 3 | 6 | 10 | 17 | 34 | 53 | 100 | 100 | 487 |
| 30 | 3 | 6 | 8 | 11 | 17 | 27 | 46 | 70 | 87 | 324 |
| 31 | 4 | 8 | 10 | 16 | 24 | 44 | 70 | 81 | 98 | 219 |
| 32 | 4 | 8 | 10 | 16 | 24 | 44 | 70 | 81 | 98 | 219 |
| 33 | 2 | 9 | 14 | 21 | 34 | 68 | 84 | 92 | 98 | 316 |
| 34 | 0 | 0 | 1 | 2 | 2 | 7 | 23 | 43 | 84 | 88 |
| 35 | 0 | 0 | 0 | 0 | 1 | 6 | 17 | 44 | 78 | 120 |
| 36 | 0 | 0 | 1 | 3 | 3 | 8 | 27 | 46 | 85 | 78 |
| 37 | 0 | 0 | 0 | 0 | 1 | 6 | 17 | 44 | 78 | 120 |
| 40 | 1 | 2 | 3 | 6 | 11 | 24 | 40 | 60 | 83 | 2 979 |
| 41 | 1 | 2 | 3 | 5 | 10 | 22 | 38 | 58 | 82 | 2 992 |
| 42 | 1 | 3 | 4 | 8 | 13 | 28 | 49 | 88 | 96 | 695 |
| 43 | 1 | 2 | 4 | 7 | 13 | 29 | 50 | 78 | 94 | 702 |
| 44 | 1 | 4 | 7 | 12 | 19 | 35 | 54 | 82 | 94 | 757 |
| 45 | 5 | 9 | 11 | 18 | 26 | 46 | 73 | 92 | 100 | 211 |
| 46 | 5 | 9 | 11 | 18 | 26 | 46 | 73 | 92 | 100 | 211 |
| 47 | 0 | 0 | 0 | 15 | 31 | 31 | 62 | 100 | 100 | 13 |
| 48 | 0 | 0 | 0 | 0 | 0 | 25 | 100 | 100 | 100 | 4 |
| 49 | 0 | 0 | 0 | 0 | 0 | 17 | 100 | 100 | 100 | 6 |
| 50 | 2 | 2 | 4 | 7 | 13 | 25 | 42 | 70 | 95 | 259 |
| 51 | 2 | 2 | 4 | 7 | 13 | 25 | 42 | 71 | 95 | 259 |
| 52 | 1 | 2 | 3 | 7 | 12 | 26 | 44 | 73 | 96 | 297 |
| 53 | 1 | 2 | 3 | 7 | 12 | 26 | 44 | 73 | 96 | 297 |
| 54 | 1 | 2 | 3 | 7 | 12 | 26 | 44 | 73 | 96 | 297 |
| 55 | 1 | 2 | 3 | 7 | 12 | 26 | 44 | 73 | 96 | 297 |
| 56 | 1 | 2 | 3 | 7 | 12 | 26 | 44 | 73 | 96 | 297 |
| 57 | 1 | 2 | 3 | 6 | 12 | 26 | 41 | 69 | 96 | 297 |
| 58 | 1 | 2 | 3 | 6 | 12 | 26 | 41 | 69 | 96 | 297 |
| 59 | 1 | 2 | 3 | 6 | 12 | 26 | 41 | 69 | 96 | 297 |
| 60 | 5 | 9 | 11 | 18 | 26 | 46 | 73 | 93 | 100 | 211 |
| 61 | 1 | 2 | 4 | 7 | 13 | 28 | 50 | 79 | 94 | 709 |
| 62 | 2 | 5 | 7 | 12 | 19 | 35 | 58 | 85 | 95 | 689 |
| 63 | 2 | 5 | 7 | 12 | 19 | 35 | 58 | 85 | 95 | 689 |
| 64 | 1 | 3 | 5 | 9 | 15 | 29 | 49 | 84 | 99 | 759 |
| 65 | 2 | 3 | 4 | 8 | 15 | 34 | 51 | 84 | 99 | 832 |

To examine the situation in more detail, a listing was made of all those stations which had a record no longer than the 5 percentile value for the particular code or whose record was no more than a year. The listing contained 3410 lines but it was found that 635 resulted from an incomplete calculation of codes 61, 62, and 63 and another 145 from an incomplete one of code 42. Thus, 2630 short records remained and these divided among the different statistics groups as tabulated below in the "Before" column. No systematic reasons for these numbers could be found and this formed part of the motivation for recalculating all monthly and annual statistics using the newly developed calculation procedures. This resulted in a small overall improvement, but not one for every statistics group as can be seen from the "After" column.

| | Codes | Before | After |
|---|---|---|---|
| Rainfall | 00 01 40 41 | 748 | 614 |
| Temperature | 02 03 04 06 07 43 44 61 62 63 | 460 | 557 |
| Grass-minimums | 05 08 22 23 | 210 | 181 |
| Sunshine & radiation | 09 17 48 49 | 33 | 26 |
| Earth-temperature | 10 11 12 13 14 47 | 70 | 58 |

| | | | |
|---|---|---|---|
| Windrun | 15 18 33 | 58 | 100 |
| 9am Values | 16 30 42 64 65 | 267 | 254 |
| Evaporation | 19 20 34 35 36 37 | 30 | 32 |
| Days of occurrence | 21 24 25 26 27 28 29 | 369 | 358 |
| Wind gusts | 31 32 45 46 60 | 174 | 159 |
| Maximum rainfall | 50 51 52 53 54 55 56 57 58 59 | 211 | 181 |
| Totals | | 2 630 | 2 520 |

During the recalculation not all errors were reported, but those that were not related to the base data were reported and are summarised below. The 150 origin errors were corrected.

| Error message | No. of errors | Date range of the errors |
|---|---|---|
| DATA ERRORS – C STAT REMAINS | 2 924 | 1956–96 |
| NO DATA SUPPORT – STAT DELETED | 476 | 1981–96 |
| BAD ORIGIN – STAT INSERTED | 150 | 1992–96 |

Obviously many short periods of statistics still exist due to missing base data from poor stations, or due to genuinely short period stations, or due to data which have been wrongly dated or identified. Improvement may result from correcting the latter and two cases were dealt with: firstly, that where the statistics were for times either before the station opened or after it closed; and, secondly, that where a station-code combination had a record of only 1 month.

There were 32 stations with statistics outside the station's period of operation. Each was investigated to determine if the statistics really belonged to the station (i.e., the station's start or end date was wrong) or whether the data belonged to another station. There was also the possibility of the year or month being incorrect. Corrections were then made by: moving or deleting daily data; moving or deleting statistics; amending stations' start or end dates as appropriate in **LAND_STATION**; and deleting warning entries from **SITE_CHANGES** about statistics outside the station's period of operation. The tabulation below summarises the deletions and amendments that were made.

| | |
|---|---|
| Daily records deleted | 785 |
| Daily records moved to another station | 249 |
| Monthly statistics deleted | 112 |
| Monthly statistics moved to another station | 263 |
| Annual statistics moved to another station | 1 |
| **LAND_STATION** start date altered | 12 |
| **LAND_STATION** end date altered | 14 |
| **SITE_CHANGES** record with "BEFORE START" message deleted | 8 |
| **SITE_CHANGES** record with "AFTER END" message deleted | 24 |

There were 321 station-code combinations which had only 1 month of data, but only 106 different station-months were concerned. As in the previous case, these statistics might be genuine but some might have been misplaced and really be part of another station's record. Furthermore, the base data might well have already been moved to the correct station, in which case the ORIG_OBS_ORIGIN of such statistics from 1972 to 1991 would be "T" since all statistics have been recalculated and base data should be available in that period and only these without base data would have a "T". Only 10 station-months were of this type and the 19 statistics involved were deleted. The remainder would need to be dealt with individually and, since experience showed this to be time-consuming, no further action was taken.

## Details and results of Check D.2 — are all monthly and annual statistics consistent?

If for a given combination of AGENT_NO, STATS_YEAR, and STATS_CODE a value is available for each possible STATS_MONTH, then a row should exist in **ANNUAL_STATS** which is consistent with those rows in **MTHLY_STATS.** This is really a check on **ANNUAL_STATS** since all its rows are calculated from the rows in **MTHLY_STATS**. A complete recalculation of **ANNUAL_STATS** was made before running the audit procedure. However, the procedure was aborted after only 25% of the stations had been checked since many inconsistencies were found.

| | |
|---|---|
| Annual value but no monthly | 1 350 |
| Monthly value but no annual | 1 851 |
| Values differ | 66 |
| Extreme dates differ | 681 |
| Reliabilities differ | 41 |
| Origins differ | 30 772 |

It can be seen that an enormous number of discrepancies were found and the run was aborted since it was possible that a program error was giving nonsense answers. However, the fault lay in the programs that extract annual statistics from the monthly statistics having not been kept up to date as new data origins had arisen. The 65 programs that calculated the annual statistics were replaced by a single program which, through an input parameter, can calculate all of them. All annual statistics were then recalculated.

Running the audit procedure to completion produced a total of 68 errors of which 66 were for station 5446 for codes 50–59 inclusive. These codes are based on **RAIN_RATE** and the station had just had digitised data loaded into CLIDB. The audit highlighted that, although the monthly values are automatically calculated shortly after **RAIN_RATE** is loaded, the annual statistics are not automatically done. Some operational programs were amended so that in future any new digitised data will trigger the annual statistics calculation for codes 50–59. Codes 50–59 were recalculated for years 1969–1975.

Only two other errors remain: one station had a null ORIG_OBS_ORIGIN which required some amendment to base data. The cause of the other error was not found, but it was for 1974 and disappeared during the recalculation for 1969–75 mentioned above.

# Summary and Conclusion

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

| Table name | Inserts | Deletions | Amendments |
|---|---|---|---|
| **ANNUAL_STATS** | 12 | 127 | 615 |
| **CLOUD_SYSTEM** | 0 | 53 666 | 49 666 |
| **EARTH_TEMP** | 0 | 2 230 | 4 |
| **EVAPORATION** | 0 | 2 | 1 |
| **LAND_STATION** | 0 | 0 | 27 |
| **MAX_GUST** | 0 | 0 | 18 513 |
| **MAX_MIN_TEMP** | 0 | 55 | 388 |
| **MTHLY_STATS** | 13 595 | 3 396 | 3 442 |
| **PRESSURE** | 0 | 89 | 34 |
| **RADIATION** | 0 | 21 | 30 |
| **RAIN** | 6 568 | 916 | 2 714 |

| | | | |
|---|---|---|---|
| **RAIN_RATE** | 0 | 0 | 2 037 |
| **SCREEN_OBS** | 0 | 89 | 33 |
| **SITE_CHANGES** | 904 | 0 | 0 |
| **SUNSHINE** | 0 | 20 | 31 |
| **SURFACE_WIND** | 0 | 670 | 11 |
| **WEATHER_PHEN** | 0 | 11 376 | 0 |
| Total | 21 079 | 72 657 | 77 546 |

The grand total of changes made was 171 212 which represents 0.12% of the total rows in 17 tables. Many of the deletions and amendments were not of vital importance as often either the data changed were relatively mundane or it was not the primary data that were changed but rather values in supporting columns. For example, the changes in **CLOUD_SYSTEM** corrected for missing values, and for **MAX_GUST** a small time change was made. Nevertheless the values were wrong and have now been corrected and, in some respects, these changes could be thought more 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 data, the following changes to the processes used in the regular production of statistics were also made.

- Algorithms for calculating codes 06, 43, 44, and 45 (N.B. during the auditing new programs were written for the calculation of all statistics so these would have been changed anyway, but errors of principle were found through the audit in the new programs for codes 43 and 44).
- A system for maintaining an ongoing consistency between **RAIN_RATE** and **RAIN** was established.
- A new program for calculating all annual statistics to replace the existing 65 separate programs was established.
- This new annual statistics program was introduced into CLIDB's standard monthly production cycle so that any changes made to monthly statistics from years which should have complete monthly sets would cause the calculation of annual statistics for those year.

These new monthly and annual statistics calculation programs were used to recalculate all the statistics from the base data. However, not all the statistics that were recovered from Trentham have base data (i.e., those that were extracted from paper forms without the daily values also being extracted) and these would remain in CLIDB with an ORIG_OBS_ORIGIN of "T". There were 3 716 249 of these in **MTHLY_STATS** before the recalculation; there are now 1 754 816, 22% of the total.

# Reference


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

*NIWA*

*Taihoro Nukurangi*

National Institute of
Water and Atmospheric
Research Limited