Quality Controlling Daily Read Rainfall Data for the Intensity-Frequency-Duration (IFD) Revision Project

Janice Green
IFD Revision Project Manager, Climate and Water Division, Bureau of Meteorology, Canberra, Australia

Fiona Johnson
Hydrologist, Climate and Water Division, Bureau of Meteorology, Sydney, Australia

Deacon McKay
Student Hydrologist, Climate and Water Division, Bureau of Meteorology, Sydney, Australia

Scott Podger
Student Hydrologist, Climate and Water Division, Bureau of Meteorology, Sydney, Australia

Michael Sugiyanto
Student Hydrologist, Climate and Water Division, Bureau of Meteorology, Sydney, Australia

Lionel Siriwardena
Research Assistant, Environmental Hydrology and Water Resources, University of Melbourne, Melbourne, Australia

The Intensity-Frequency-Duration (IFD) Revision Project, undertaken by the Bureau of Meteorology, utilised data from nearly 20,000 daily read rainfall stations to derive revised IFD design rainfall estimates. Automated quality controlling software was developed to assist with the task of quality controlling the records from the thousands of read daily rainfall stations. The software detected date shifts and accumulated totals (both identified and unidentified) and flagged other possible suspect data as gross errors for further checking. The flagged errors were checked manually using the Bureau's Quality Monitoring System and information from nearby neighbours, station metadata and paper based rainfall records as well as information from remote sensing such as satellite images and radar scans. Performance statistics were prepared to present a summary of the results of the manual checks of possible gross errors across Australia. These performance statistics will provide useful information for practitioners for projects where automated quality controlling of data is required.

1. INTRODUCTION

The Australian Bureau of Meteorology (the Bureau) has recently completed a revision of the Intensity-Frequency-Duration (IFD) design rainfall estimates. For the derivation of revised IFDs for durations of one to five days, the revision utilised data from the Bureau's network of daily read rainfall stations. In 2011, the Bureau's Australian Data Archive for Meteorology (ADAM) contained daily read rainfall data from nearly 20 000 stations (both open and closed) from 1800 onwards. The location of these raingauges and their period of record are shown in Figure 1.

In addition to ensuring that as much data as possible was used, it was also necessary that the rainfall data be quality controlled (QC'd) to a standard suitable not only for its application to the IFD Revision but also to the derivation of the associated temporal and spatial patterns. While for the purposes of

the IFD Revision Project by itself it would have been sufficient to quality control the largest rainfall events for each of a range of durations, the application of the database for other of the Australian Rainfall and Runoff (AR&R) (Institution of Engineers, 1987) Revision projects, including the derivation of temporal and spatial patterns, necessitated the quality controlling of all daily read rainfall data.

In light of the scope of the quality controlling (QCing) requirements and the volume of data needing to be quality controlled, automated procedures have being developed for the identification of suspect data and, as far as possible, the correction of these data. However, the quality controlling of the data can only be automated so far and a significant amount of data was required to be manually checked.

Section 2 discusses the methods adopted for the automated quality controlling of the daily read rainfall data; Section 3 presents the manual QCing that was the quality controlling of the continuous rainfall data; and Section 4 discusses the results of the performance statistics that were derived in order to assess whether there were any spatial or temporal trends in the occurrence of errors.

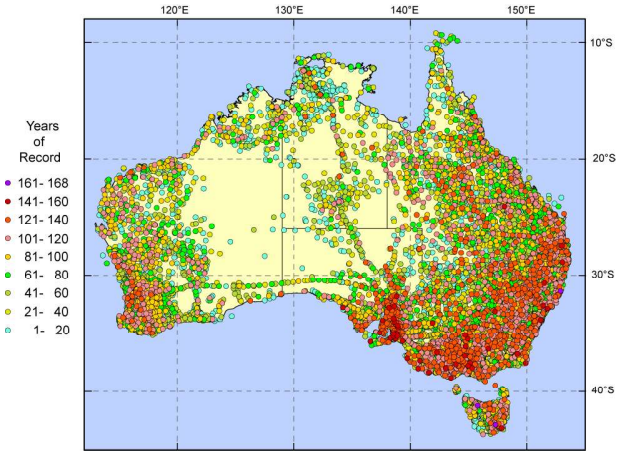

Figure 1 Location of Bureau daily read raingauges and period of record

2. AUTOMATED QCING

As part of the establishment of a quality controlled database for the revision of the IFDs, a significant amount of work has been undertaken into the adoption, adaptation, and, where necessary, development of automated procedures for the quality controlling of daily read rainfall data. These include those listed; the development and trialling of each of these procedures are described below.

- Infilling of missing data
- Disaggregation of accumulated daily rainfall totals
- Detection of suspect data and identification of unflagged accumulated totals and of time shifts in daily rainfall data
- Identification of gross errors

2.1. Infilling of missing data

A comparison was undertaken of a number of simple infilling methods for missing data in daily rainfall records. Seven methods to infill missing rainfall data in daily rainfall records based on an estimate from the records at the nearby gauges were considered. These methods were:

(i) data at the nearest gauge assigned directly to the target gauge;
(ii) the inverse distance weighting method;
(iii) method (ii) modified to be representative of a point estimate;
(iv) method (ii) modified to account for long term variations of rainfall at nearby gauges;
(v) adoption of the closest of the five nearest gauges which has the highest daily correlation;
(vi) adoption of closest the gauge with a daily regression relationship with zero constant ($y=mx$);

(vii) method based on scaling of the probability distribution of daily rainfalls of the target gauge and the nearest gauge in the vicinity.

Further details of these methods can be found in Siriwardena and Weinmann (1996); Zucchini and Sparks (1984) and Bureau of Meteorology (2008).

The evaluation of different methods was based on the ability of the method to estimate rainfalls from the data at nearby gauges as close as possible to the actual data at the target site. This was assessed by comparing the mean and standard deviation of the estimated and the recorded data as well as using two performance indices; that is, coefficient of efficiency and root mean square error between the recorded and estimated data. The evaluation was also based on the comparison of exceedance probability plots of estimated against the recorded in which the ability to estimate extreme rainfalls was particularly examined.

The method using the inverse distance weighted average of the nearest three gauges to estimate the rainfall at the target site was shown to produce consistent and best results across all sites tested; this method is outlined further in the following section. Although this method has a tendency to produce slightly lower estimates and slightly lower extreme rainfalls for some sites, the results of this method were consistently superior to the results of the other methods. Based on the outcome of this investigation, the following method was adopted for the infilling of missing data. This approach was adopted using all gauges for which there was a valid data recording, including a record of zero.

- The maximum distance within which the gauges are used for estimating missing data was set of 50km, although this was increased in sparsely gauged areas.
- If there is only one gauge within the specified distance use the data at that gauge.
- Otherwise, use the nearest up a maximum of three gauges to estimate the missing data using the inverse distance weighting method using equation (1). If there are gauges within a distance of 0.5 km set that at a 0.5 km distance.
- If there is not a single gauge with the specified distance, flag that appropriately to indicate 'not possible' and write all such incidences to a log file.

2.2. Disaggregation of accumulated daily rainfall totals

This is primarily based on the automated disaggregation procedure adopted by Siriwardena and Weinmann (1996) for Victorian data. In this approach, if only a single gauge with valid data is found within a distance of 3 km from the target site, the rainfall pattern of that gauge is used to disaggregate the accumulated data at the target site; if more gauges are available within 3km distance equal weighting of up to three gauges is used. Otherwise, daily rainfalls at the target site over the accumulated period are first estimated from the three nearest gages using the inverse distance weighting method using equation (1). The pattern of the estimated rainfall over the accumulated period is then applied to disaggregate the accumulated data at the target site using equation (2).

$$RE_{js} = \frac{\sum_{k=1}^{n} \frac{RN_{jk}}{d_k}}{\sum_{k=1}^{n} \frac{1}{d_k}} \qquad (1)$$

where RN_{jk} = precipitation at the nearby gauge k on day j
d_k = distance from gauge k to the target gauge s
RE_{js} = estimated precipitation at the target gauge s on day j
n = the number of accumulated days being considered

$$P_{js} = \frac{(\sum_{j=1}^{m} P_{js}) RE_{js}}{\sum_{j=1}^{m} RE_{js}} \quad (2)$$

where: $\sum_{j=1}^{m} P_{js}$ = precipitation at the target site accumulated over m days

RE_{js} = estimated precipitation at the target site *s* on day *j*

P_{js} = precipitation at the target site *s* on day *j*

A modified version of the procedure used by Siriwardena and Weinmann (1996) has been adopted. In the modified approach a validity check for the accumulated totals at the nearby gauges has been introduced to avoid to the use of inconsistent records in disaggregation.

2.3. Identification of suspect data

Procedures for the disaggregation of accumulated daily read rainfall totals and the infilling of missing daily read data have previously been developed (Siriwardena and Weinmann, 1996). However, no similar methods have been developed for the identification of unflagged accumulated totals in daily rainfall data or for the identification of time shifts in daily rainfall data. As with much other quality controlling of rainfall data, the checking for these two artifacts has previously been undertaken manually on an event by event basis.

An automated approach has been developed to, firstly, detect suspect data, and, secondly, identify the type of data error. The approach developed involves detecting possible errors in the daily rainfall records by examining the probability of observing a residual in clean data and then to classify the detections with appropriate flags indicating the probable cause of the errors. The residual cut-off value (RC) and the standardised residual cut-off (SRC) were then compared to the absolute residual and the standardised residual values calculated from the daily read database using the following tests:

- RC criterion was assumed to have failed when:
  - absolute residual > RC
- SRC criterion was assumed to have failed when:
  - absolute residual > RC98 and standardised residual > SRC (where RC98 is the 98th percentile of the absolute residual)

The screened data were assigned the following codes:

- Both RC & SRC criteria failed: 3
- RC criterion only failed: 2
- SRC criterion only failed: 1
- None of the criteria failed: 0
- Missing rainfall records: -1

Any data with a code greater than zero indicated the detection of suspect data with potential errors which were tested for one of the following three causes.

2.3.1. Identification of time shifts

If the data were identified as being suspect, they were first tested to determine if there was a date shift in the data caused either by the reading being recorded on the wrong date by the reader or entered incorrectly in ADAM. The test for time shift was undertaken using the following procedure:

- The recorded data during the event was shifted by one day in either direction
- Absolute residuals between the recorded and interpolated values were calculated for each day of the event for the three cases where the recorded data had been shifted both way and had

not been shifted
- For each of the three cases the sum of the residual was calculated

It was considered that a time shift had been detected if the sum of the residuals improved from the shifted to the no-shift position. The test was repeated using the same procedure with a two day shift.

2.3.2. Identification of unflagged accumulated totals

If the error had not been identified as a data shift, the data were checked for the possibility of an unflagged 2-day or 3-day accumulation using the following criteria:

- The recorded rainfall in the preceding day was zero
- The sum of the interpolated rainfalls on the preceding two days was more than 3 mm
- The suspect data were recorded during the Christmas period
- The suspect data were recorded during the Easter period
- The suspect data were recorded on a Monday

It was considered that an unflagged accumulated total had been detected if one or more of the criteria were met.

2.3.3. Identification of gross errors

If the suspect data had not been identified as either a time shift error or an unflagged accumulated total, it was flagged as being a gross error requiring more detailed, manual checking. Gross errors are random errors due to manual recording such as recording the amount incorrectly. In addition, measurement errors and errors due to instrument malfunction can also be considered under the same category. It should be noted that these data are inconsistent with the weighted average calculated from the neighbouring records and are not necessarily in error. All errors failed to identify as either date shifted or unflagged accumulation were classified as gross errors and flagged with 'GE'.

In order to be able to prioritise the amount of manual QCing that was required, a screening algorithm was introduced to reduce the percentage of 'false' identification of these errors or suspect data. If the difference between the daily rainfall and nearest gauges is within 0.5 times the absolute residual (difference between the recorded and interpolated) for at least for two gauges of the nearest five gauges then the value is assumed acceptable given that the difference is not less than 10mm.

The suspect gross errors were further categorised as high probability (GEH) and low probability based on a simple algorithm. If the difference between the recorded rainfall and the estimated rainfall was greater than a factor (read from the parameter file) times the residual cutoff value, subjected to a 25mm threshold, the recorded value was flagged as 'GEH'. All the other values are flagged as 'GEL'. This factor is currently set at 1.35 based on a range of sensitivity analysis tests. On average, application of this factor results in a 40:60 split of GEH to GEL errors.

3. MANUAL QCING

The manual QCing of the GEH flagged gross errors was facilitated through the use of the Quality Monitoring System (QMS) developed by the Bureau's National Climate Centre. QMS is a suite of programs that is used to check, analyse, edit and monitor the data in the Climate database, ADAM. As the daily read data had already been checked and, where necessary, corrected using the automated procedures described in Section 2, the functionalities of QMS that were used were the ability to map the suspect value in relation to nearby stations and the linking to GIS data from other systems including RADAR, Satellite and Mean Sea Level Pressure (MSLP) Analysis.

The approach that was adopted for the manual QCing of the flagged data is summarised below:

- The rainfall of nearby stations was checked visually to provide an indication of rainfall over the general area.

- The available charts (Radar, MSLP, Satellite) were viewed to give guidance about rainfall over the area and at the target station.
- If available multi-source precipitation information (for example, continuous rainfall stations, Automatic Weather Stations) were compare to the daily read values.
- The Bureau's meta database, Sitesdb, was viewed for information relating to the target station such as malfunctions of the gauge.
- The electronic copy of F68 form (the form used to record daily read rainfall totals) was viewed (if available) to determine if the data matched the original data in the F68.

Figure 3 gives an example of the manner in which QMS facilitates the comparison to nearby stations and the viewing of RADAR images to verify the high rainfall total recorded at the target (shown in red). The manually checked GEH data were assigned one of the following classifications:

- OK – data is supported by neighbours or not sufficient evidence to reject
- OK_MOD – data is OK once manually modified to match TBRG or F68
- WRONG – data is not supported by neighbours and type of error could not be determined
- SHIFT – data is OK once shifted by X days either forward or back
- ACC – data should have been flagged as an accumulation
- OTHER – problems with data are systemic or unclear and flagged for further review by National Climate Centre

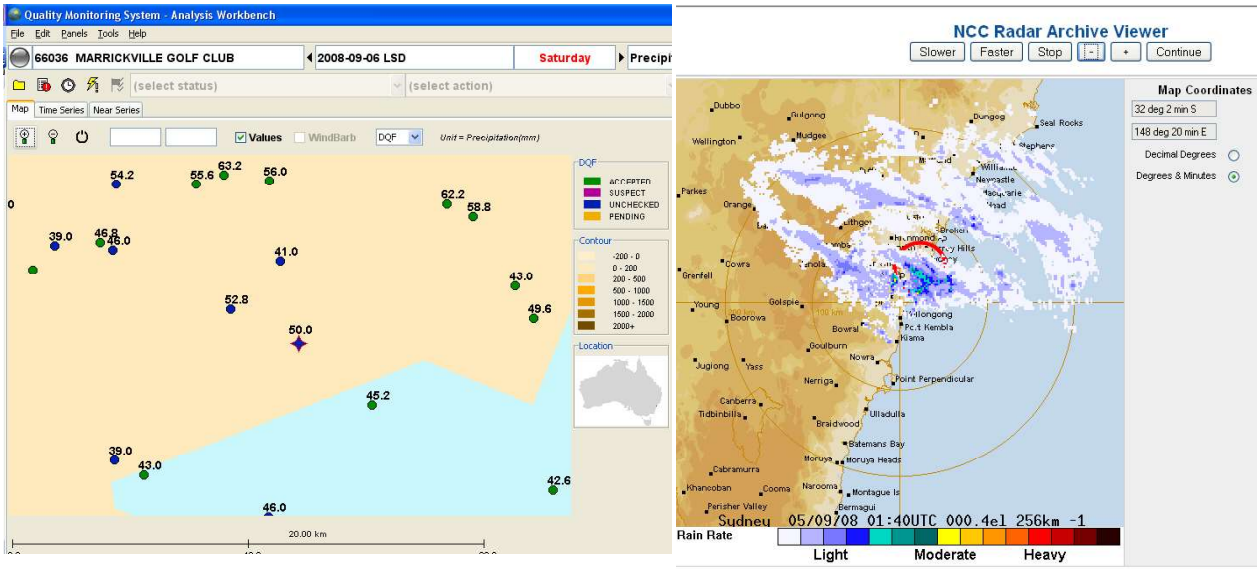

Figure 3 Example of manual QCing of suspect data using QMS

4. STATISITICS OF MANUALLY QC'D DATA

In order to provide information on the number, type and temporal and spatial distribution of the manually checked errors in the Bureau's daily read rainfall data, some performance statistics were derived. The primary objective of this was to determine whether there were any systematic errors that would need to be taken into consideration in the analyzing of the daily read data for the derivation of the IFDs. However, it is considered that these performance statistics will be of use for practitioners undertaking projects where quality controlling of data is required.

4.1. Number of GEH errors

Even with the classification of the gross errors into GEH and GEL, there still remained just under 8000 stations with GEH flagged errors that needed to be manually checked. On average each of these stations had ten errors that needed to be checked, however more than 700 stations had more than 20 errors flagged. The distribution of the number of GEH gross errors per station that needed to be manually is shown in Figure 4.

4.2. Number of 'true' errors

It was found that in general approximately 80% of the data flagged as GEH errors were actually correct and that the failure to meet the criteria in the automated checking process was due to the natural variability of rainfall in Australia. Figure 5 shows the percentage of flagged GEH errors that we accepted as true rainfalls after the manual QC checks. From this is can be seen that rainfall stations where the flagged GEHs were found to not be true rainfalls generally occur in the south of Australia. However, further investigation is required to determine whether this is due to the greater density of long term stations in this region.

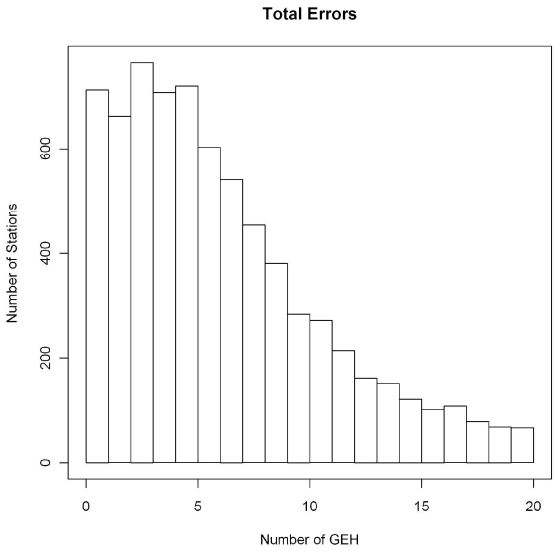

Figure 4 Example of manual QCing of suspect data using QMS

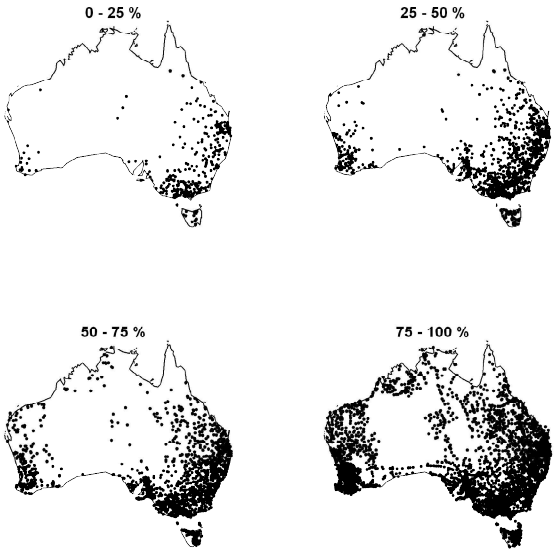

Figure 5 Percentage of flagged GEH errors accepted as true rainfalls after manual QCing

4.3. Type of errors

The types of errors that were found through the manual QCing of the GEH errors were summarised and are reported for Australia as a whole and according to state in Table 1. This shows that of the 20% of GEH flagged data that were considered to be in error, around half of these errors were as a result of the data being recorded on the wrong date. The remaining 50% of true errors were unflagged accumulated totals or erroneous data. It is emphasised that these statistics refer to the small sub-set of errors identified by the automated procedures that could not be corrected automatically and were also assigned a high probability of being true errors. The statistics do not reflect the overall accuracy of the daily read rainfall data collected by the Bureau of Meteorology.

4.4. Temporal and spatial distribution of errors

An analysis of the occurrence of gross errors across Australia showd that there was considerable spatial variation in the results of the manual checks. However, a temporal analysis of the identified gross errors that was also carried out found that, in general, there were no strong trends in the likelihood of data being flagged as a gross error. That is, the quality of the rainfall records had neither improved nor deteriorated over time.

Table 1 Results of manual QCing of GEH errors

	Correct	Wrong	Timeshift	Unflagged Accumulation	Other
Australia	73%	7%	11%	7%	2%
WA	82%	5%	9%	4%	1%
NT	87%	1%	7%	4%	0%
SA	77%	8%	11%	3%	1%
QLD	74%	2%	10%	5%	9%
NSW	69%	13%	9%	7%	2%
VIC	61%	5%	16%	17%	1%
TAS	75%	9%	10%	6%	0%

5. CONCLUSIONS

As part of the revision of the IFD design rainfalls undertaken by the Bureau of Meteorology, a quality controlled database of daily read rainfall data was created and used to derive the revised IFDs for durations of one to five days. The rainfall data comprises the records from the over 20 000 daily read rainfall stations contained in the Bureau's ADAM database. The data were checked using automated QCing procedures developed as part of the IFD Revision Project and, where possible, erroneous data were corrected automatically. Data that could not be corrected were flagged as being gross errors and assigned a high (GEH) or low (GEL) probability of being incorrect. These data were manually checked using the functionalities of the Bureau's QMS.

Performance statistics were derived on the manual checked data which showed that 80% of the GEH flagged values were correct and that of the remaining 20%, half of the errors were due to the data being recorded on the wrong date.

These performance statistics should be useful for other projects where quality controlled daily rainfall data are required. In addition, the availability of a systematically quality controlled set of daily read rainfall data will be of considerable use to practitioners undertaking hydrologic analyses by reducing the amount time spent on checking and correcting data and providing a consistent data set to be used across studies.

6. REFERENCES

Bureau of Meteorology (2008). Quality Monitoring System (QMS) Test Description, Version 1.0, National Climate Centre / Data Management, Bureau of Meteorology, Australia.

Green, J.H., Xuereb, X. and Siriwardena, L. (2011). "Establishment of a Quality Controlled Rainfall Database for the Revision of the Intensity-Frequency Duration (IFD) Estimates for Australia". Presented at 34th IAHR Congress, Brisbane, Qld, June 2011.

Institution of Engineers (1987). Australian Rainfall and Runoff – A Guide to Flood Estimation. Institution of Engineers, Australia, Barton, ACT, 1987

Siriwardena, L. and Weinmann, P.E. (1996). *Derivation of areal reduction factors for design rainfall in Victoria: for rainfall durations 18-120 hours*. CRC for Catchment Hydrology, 96/6, October 1996.

Zucchini, W. and Sparks, R.S. (1984). *Estimating the missing values in rainfall records*, Department of Civil Engineering, University of Stellenbosch, Water Research Commission Report No. 91/3/84.