Establishment of a Quality Controlled Rainfall Database for the Revision of the Intensity-Frequency-Duration (IFD) Estimates for Australia

J.H. Green[1], K. Xuereb[1] and L. Siriwardena[2]
[1]Climate and Water Division
Bureau of Meteorology
Melbourne Vic 3008
AUSTRALIA
[2]Department of Civil and Environmental Engineering
The University of Melbourne
Melbourne Vic 3010
AUSTRALIA

Abstract: *The revision of the Intensity-Frequency-Duration (IFD) estimates for Australia needed a quality controlled database of daily-read and continuous rainfall data to be established. This required identifying all authorities who are collecting rainfall data, the extraction and archiving of the data, and the quality controlling of the data to a consistent standard.*

The large amount of data meant that manual techniques for quality controlling the data were not feasible necessitating the development of a largely automated procedure. The adopted approach in this study incorporates the automatic identification of suspect data using various techniques including domain checking; Barnes analysis; precipitation multi-sensor comparison, and the checking of station site metadata. Automated procedures were adopted for the identification and disaggregation of accumulated totals, infilling of missing data, and the correction of time shifts. The resultant high quality database is suitable not only for the IFD revision but also for a range of hydrologic analyses.

Keywords: *Rainfall data, quality control*

1. INTRODUCTION

As part of the revision of Australian Rainfall & Runoff (AR&R) (Institution of Engineers, 1987) the Australian Bureau of Meteorology (the Bureau) is revising the Intensity-Frequency-Duration (IFD) design rainfall estimates. In contrast to the current IFD estimates which were based primarily on Bureau of Meteorology operated daily read and continuous rainfall stations, the revised estimates are required to be based on daily read and continuous rainfall data collected nationally from all agencies. Further, it is a requirement that the rainfall data be quality controlled to a standard suitable for its application to the IFD Revision and 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 AR&R Revision projects, including the derivation of temporal and spatial patterns, necessitated the quality controlling of all rainfall data.

The rainfall data from various sources have been previously quality controlled to differing degrees using a range of quality control coding systems. The result of this was that it was difficult to establish a baseline of quality controlled data using existing procedures. In view of this, it was decided that the quality controlling of the rainfall data would be applied to all data, regardless of source, and would be predicated on the assumption that no checking or correcting of the data had been undertaken previously. Where quality controlling of the data had been undertaken in a systematic fashion, for example using the HYDSTRA Time Series Data Management tools developed by Kisters Pty Ltd and adopted by large water authorities, there would be efficiencies in terms of a reduction in the amount of suspect data identified and requiring checking and correction.

In light of the scope of the quality controlling 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. Section 2 discusses the data sets that have been quality controlled; Section 3 the methods adopted for quality controlling the daily read rainfall data; and Section 4 the quality controlling of the continuous rainfall data.

2. DATABASE

A review has been undertaken of all sources of rainfall data (both continuous and daily read) across Australia with the following data sources being identified:

- The Bureau of Meteorology Australian Data Archive for Meteorology (ADAM)
- Non-Bureau – collected by organisations identified in the Water Regulations 2008

A catalogue of the Bureau's rainfall station network has been prepared to facilitate the identification of both temporal and spatial gaps within the network that will be supplemented by data from other sources. Further details of the rainfall data available from each of the above sources are provided below.

2.1. ADAM

In 2009, the Bureau of Meteorology's ADAM database contained daily read rainfall data from nearly 20 000 stations (both open and closed) from 1800 and nearly 1 500 continuous rainfall stations – using both Dines tilting syphon pluviograph (DINES) and Tipping Bucket Rain Gauge (TBRG) instrumentation. The location of these raingauges and the period of record are shown in Figure 1(a) for the daily read stations and Figure 1(b) for the continuous rainfall stations.

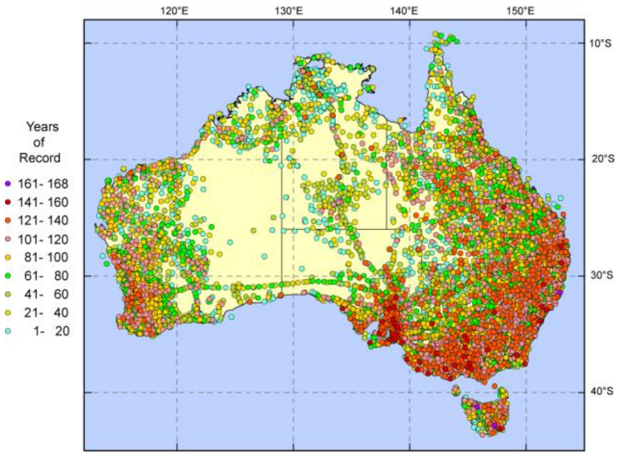

Figure 1(a) Location of Bureau daily read raingauges and period of record.

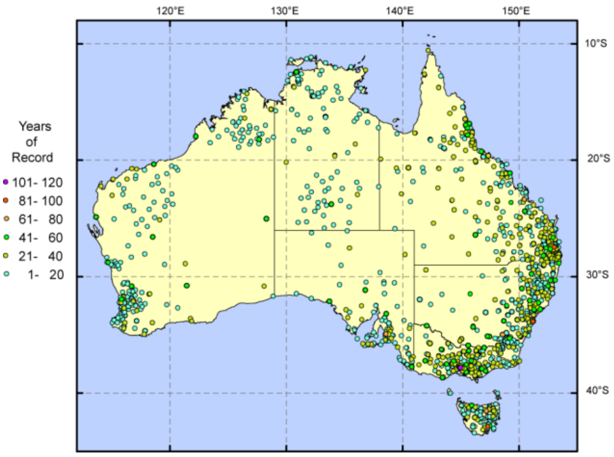

Figure 1(b) Location of Bureau continuous raingauges and period of record.

2.2. Non-Bureau sources

Under the Commonwealth *Water Act 2007*, the Bureau of Meteorology has responsibility for compiling and delivering comprehensive water information across Australia. The Water Regulations 2008 identified approximately 260 'persons' who are required to give to the Bureau, water information that they have in their possession, custody, or control. Of the 260 non-Bureau data custodians, 74 have indicated that they possess data from daily read rainfall stations and 45 that they have data from continuous rainfall stations. Based on the information from the catalogue it is expected that the following additional rainfall data will be received from non-Bureau sources.

- ~ 350 daily read rainfall stations
- ~ 2175 continuous rainfall stations

The location of these raingauges and the period of record are shown in Figure 2(a) for the daily read stations and Figure 2(b) for the continuous rainfall stations.

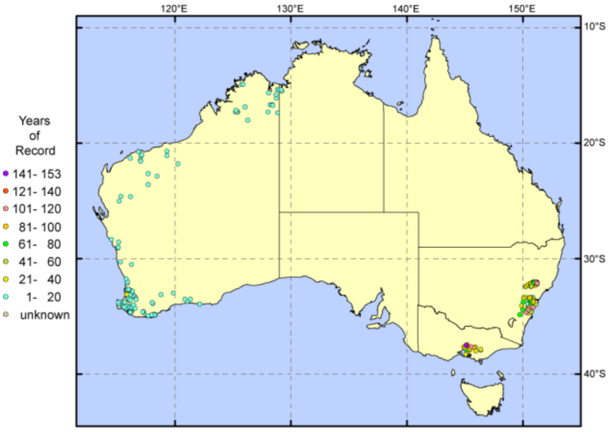

Figure 2(a) Location of non-Bureau daily read raingauges and period of record.

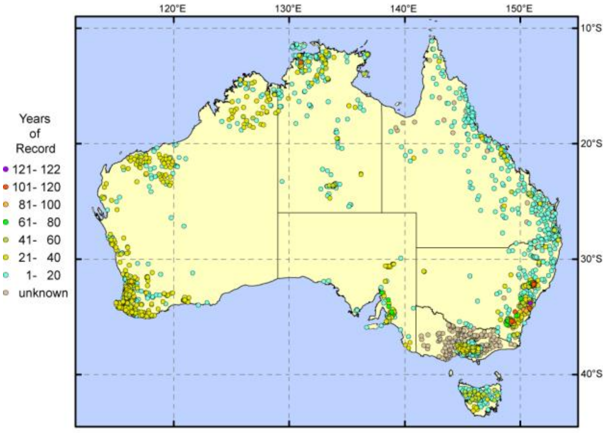

Figure 2(b) Location of non-Bureau continuous raingauges and period of record.

3. DAILY READ RAINFALL DATA

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:

- 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

The development and trialling of each of these procedures are described below.

3.1. 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 ranged from the simplest of assigning the data at the nearest gauge to a method based on scaling of probability distributions of daily rainfalls (Siriwardena and Weinmann, 1996; Zucchini and Sparks, 1984; 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. 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:

- Set the maximum distance within which the gauges are uses for estimating missing data
- 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 them appropriately to indicate 'not possible' and write all such incidences to a log file.

3.2. 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

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

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.

3.3. 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 artefacts 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

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.

3.3.1. 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 but with a two day shift.

3.3.2. 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.

3.3.3. 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.

4. CONTINUOUS RAINFALL DATA

Automated procedures have also been developed for the quality controlling of the continuous rainfall data. These procedures use comparisons with other data sources including the Australian Water Availability Project (AWAP) gridded data; daily read rainfall stations; automatic weather stations (AWS); and synoptic stations to identify and correct spurious and missing data.

The quality controlling of continuous rainfall data is considerably more complicated than for daily read rainfall data due to the significantly larger amounts of data resulting from the much shorter timesteps at which the rainfall data are recorded. For Bureau operated DINES stations, the pluviograph traces were digitised using a ‘break-point’ analysis procedure where a value was digitised at each change of gradient in the trace. For the TBRG stations the data are stored as tips (mostly 0.2mm) which are timestamped to a resolution of one second. In order to reduce the amount of continuous rainfall data that needed to be quality controlled to a manageable volume, it was decided to adopt a subset of the largest rainfall events based on the Partial Duration Series (PDS). The PDS was created by extracting the highest rainfall records equal to three times the number of effective years of record at each site.

The procedures described below have been applied to the Bureau’s continuous rainfall data and a quality controlled database of the Bureau’s continuous rainfall records has been established. This database consists of over 500 000 of the three largest events (on average) from each year of record for the 779 of the Bureau’s continuous rainfall stations that have eight years or more of data.

4.1. Quality Control of Partial Duration Series values

Quality control of the PDS values was initially done by comparing the PDS values with gridded values of daily rainfall at the nearest grid point to the gauge location. PDS values were flagged whenever they exceeded the daily gridded rainfall. The gridded daily rainfall datasets have been prepared as part of the Australian Water Availability Project (AWAP) (Jones *et al*, 2007). An example of the AWAP gridded rainfall dataset for 25 December 2008 is shown in Figure 3.

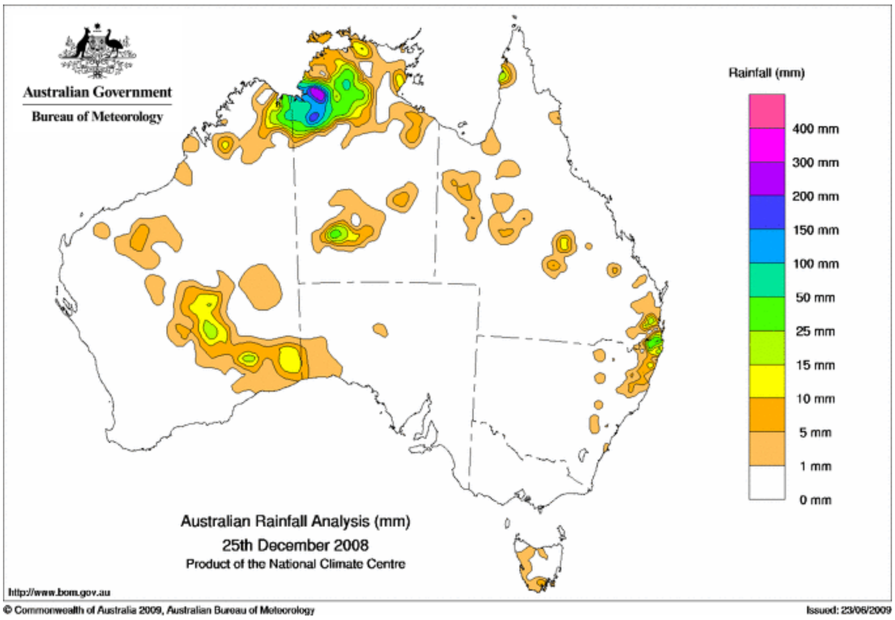

Figure 3 Example of AWAP Gridded Daily Rainfall Data

The total number of PDS values comes to over 500,000. Over 60% of these values were flagged after comparison with gridded daily rainfall data. Careful examination of the output showed that many of the flagged values exceeded the gridded daily rainfall value but not necessarily the actual daily rainfall value at a co-located or nearby site. This is because the gridding procedure creates a smoothed surface that can often under-estimate the actual values. Comparison with the daily rainfall at a co-located site was incorporated next into the quality control procedure.

Quality control criteria were applied in stages after examining the output carefully at the end of each stage. Each stage was added if its adoption resulted in less than 10% incorrect values remaining undetected. The stages trialed were as follows:

The PDS values were flagged if:

(i) the PDS value exceeded the gridded daily rainfall value on the same day
(ii) the PDS value exceeded the daily rainfall value on the same day at the same site
(iii) the PDS value exceeded the daily rainfall value on the next day at the same site
(iv) the PDS value exceeded the sum of daily rainfall values at the same site on the same day and the day after plus a 10 mm buffer
(v) the PDS value exceeded the sum of daily rainfall values at the same site on the same day and the day before plus a 10 mm buffer

Criterion (iii) was added because it was found that often a subdaily event straddled two days; that is, the event started on one day and finished on the next day. Daily rainfalls are given from 9 am to 9 am whereas PDS values are generated using a sliding window. Therefore a rainfall amount which fell mostly in day one could have an end date on day 2, whereas the 9 am to 9 am rainfall would show most of the rainfall fell on the first day.

Criterion (iv) was added because several times an event would occur partly on one day and partly on the next. The 10 mm buffer was added due to variations between readings from different instruments. Criterion (v) was added for the same reason as criterion (iv) and the same buffer as in (iv) was applied. A common error in the daily rainfall database is a shift of the rainfall of one or more days. The rainfall amount would be correct but the date would be incorrect. Criteria (iii) to (v) act to eliminate the detection of such events.

For continuous rainfall sites with no co-located daily site, daily rainfall data were extracted for the nearest 10 or so sites. The automated procedures flagged PDS values if criteria (i) to (v) were satisfied for any one of the sites within a 10 km radius from the continuous rainfall. For continuous rainfall stations having no daily rainfall sites within 10 km, the daily rainfall data for 10 nearest sites was extracted and the quality control procedure was performed manually. The buffer was increased to 20 mm to allow for natural spatial variation of rainfall.

4.2. Manual Quality Controlling of Partial Duration Series values

After the application of the automated procedures described above, there remained approximately 2500 events that required manual quality controlling. The manual quality control procedure adopted involved comparing 9 am to 9 am continuous rainfalls with daily (also 9 am to 9 am) rainfalls in order to decide whether the flagged value was correct or not. Continuous and daily read rainfall records were extracted for three months including the month during which the event occurred, the month before and the month after the event for the same location. The daily rainfall record for the month for the nearest 10 daily sites was also extracted. For continuous rainfall sites with no co-located daily site, the continuous rainfall record was compared with the daily rainfall record of the nearest site. Most of the time, comparison with the daily data for the same (or nearest) daily site and other nearby daily and continuous rainfall sites was enough to make a decision as to whether the continuous rainfall value was correct or not. The continuous rainfall value was not modified in any way. The comparison with daily values was made in order to assess whether it was valid or not.

For sites located in more remote areas where the distance to the nearest daily site is greater than 20 km, other data were examined. In particular, the Bureau’s site metadata database, SitesDB, which contains details of station data, including maintenance records, details of equipment upgrades and inspections, Automatic Weather Station (AWS) data, satellite images, weather charts and synoptic data were used. A PDS value was rejected only if it was considered to be incorrect with a very high degree of probability. Generally, it is considered that a daily rainfall reading is more likely to be correct than a continuous rainfall reading. This is because continuous rainfall stations are more susceptible to malfunction than the standard daily rain gauge. In about 40 instances though, the daily rainfall reading was found to be incorrect, through comparison with nearby daily sites. If the reason to reject a PDS value was not compelling, it was retained. Such situations arose in isolated areas for events when AWS, satellite images, weather charts and synoptic data are not available.

5. CONCLUSION

As part of the revision of the IFDs a quality controlled rainfall database has been established. Automated procedures have been developed for the quality controlling of daily read rainfall data collected by both the Bureau of Meteorology and other agencies across Australia. The quality controlling procedures that have been developed include:

- 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

These procedures have been applied to the over 20,000 daily read rainfall stations that are or have been operated by the Bureau.

Automated procedures have also been developed for the quality controlling of the continuous rainfall data. These procedures use comparisons with other data sources including the AWAP gridded data; daily read rainfall stations; automatic weather stations; and synoptic stations to identify and correct spurious and missing data. The procedures have been applied to the Bureau’s continuous rainfall data and a quality controlled database of the Bureau’s continuous rainfall records has been established. This database consists of over 500 000 of the three largest events (on average) from each year of record for the 779 Bureau continuous rainfall stations with eight years or more of data.

6. ACKNOWLEDGMENTS

The revision of the IFD estimates that is being undertaken by the Bureau of Meteorology is a team effort and the input of the individual team members is gratefully acknowledged: Catherine Beesley, Shirrah Comeadow. Fiona Johnson, Jeanette Meighen, Garry Moore, Brian Taylor, and Cynthia The.

7. REFERENCES

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

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

Jones, David A., William Wang and Robert Fawcett, (2007). *Climate Data for the Australian Availability Project.* National Climate Centre, Australian Bureau of Meteorology.

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, Report 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.