Proc. 3rd International Conference on Water and Environmental Engineering (iCWEE2022)
27-30 Nov 2022, Sydney, Australia
ISBN: 978-0-6456692-0-6

Streamflow Data Preparation for Trend Analysis: A Case Study for Australia

S M Anwar Hossain[1] and Ataur Rahman[2]

[1]PhD candidate, Western Sydney University, Kingswood, Australia
[2]Professor, Western Sydney University, Kingswood, Australia

Corresponding author's E-mail: 18916573@student.westernsydney.edu.au

Abstract

The accuracy of the outcome of flood quantile estimation and trend analysis largely depends on the quantity and quality of the available streamflow data. Various types of quality issues in streamflow data can be found such as gaps in the data series, a short record length of the data, and outliers in the data. The outcome of streamflow data analysis can be influenced by the ways in which the missing values and outliers are processed. Therefore, it is of utmost importance to scrutinise the collected streamflow data series to ensure that they are suitable for trend analysis. This study considered Australian catchments and collected annual maximum flood (AMF) series data from a large number of stations all over Australia. This study uses several steps to check data quality to minimise errors in collected AMF data. The collected AMF data has the quality code against each data point assigned by the streamflow gauging authority. These data points are checked and if appropriate, gauging stations with poor-quality coded data are excluded from this study. The presence of any outliers in the AMF series is identified. In this study, multiple Grubbs-Beck tests are adopted. Gaps in the AMF series are infilled using different methods based on the availability of the monthly instantaneous maximum (IM) data, monthly maximum mean daily (MMD) data of the missing year and the availability of the missing year's AMF data of nearby stations. Where IM and MMD data for the missing year are available, the gap is filled by comparing the IM data with MMD data of the same station for the year with data gaps. Where annual maximum mean (AMD) flow is available and AMF data is missing then missing AMF data is estimated using regression between the AMD series against the AMF series of the same station. The coefficient of determination R^2 of this regression is found between 0.9 to 0.99. Where IM and MMD are not available, simple linear regression is used between the station of missing AMF and the nearby station's AMF where the missing year's AMF is available to fill up the gap. A total of 676 stations are initially selected each having a minimum record length of 20 years. For in-filling gaps in AMF series, priority is given to the first approach followed by the second and third as appropriate. Finally, 307 stations are selected with minimum record of 50 years. Twenty-three (7%) out of 307 stations are found to have missing points/gaps. These data will be used to examine trends and in non-stationary flood frequency analysis.

Keywords: Data Preparation, Missing Data, AMF, Trend Analysis, FFA.

1. INTRODUCTION

The availability of a reliable source of metrological and hydrological data is a fundamental requirement for the modeling of different metrological and hydrological processes such as trend analysis, at-site stationary flood frequency analysis (FFA), non-stationary FFA and regional flood frequency analysis (RFFA).

Even with the advancement of science and technology, most meteorological and hydrologic observed data series (such as streamflow, rainfall and temperature) suffer from missing data. There are different

reasons for this, such as measuring instruments may break suddenly, measuring instruments may become faulty or data transmission from the instrument to the computer may be interrupted. The use of the hydrological time series with missing data can result in errors that exhibit temporal and spatial patterns (Stooksbury et al., 1999). Missing data points in observed time series data can cause biases in estimates of the relations between two or more variables (Pigott 2001). For analysis of the detection of temporal trends in time series data, it is required to have complete records on all observation stations (Hann et al., 2013). The accuracy of trend analysis at each individual stream gauging station also depends on the record length. Higher record length can capture multiple and long-term climate variability cycles whereas a shorter record length of annual maximum flood (AMF) data may provide misleading trends. Therefore, the selection of a cut-off record length at an individual site is a very important step in trend analysis. The record length should be as long as possible while retaining enough gauging sites in the study area to make the results more meaningful. The presence of outliers in data significantly affects statistics of flood data such as the average and standard deviation of a sample, resulting in overestimated or underestimated values (Kwak and Kim, 2017). The data also needs to satisfy basic assumptions of homogeneity and stationarity. Different approaches are available to treat the missing data or data gaps. The use of different approaches for infilling gaps and treating outliers can change the results of trend analysis and FFA. Therefore, adequate treatment of missing data and outliers is crucial for analysis (Kwak and Kim, 2017). This study presents Australian AMF data preparation so that the final data set can be used for trend analysis.

2. DATA PREPARATION

The AMF data preparation can be divided into a number of steps: (a) selection and refinements of initial candidate catchments and selection of minimum record length; (b) filling up gaps in data by estimating missing data points; (c) outlier identification; (d) time trend analysis (Haddad and Rahman, 2010). The above steps are discussed in the method adopted for data preparation.

3. METHOD ADOPTED FOR DATA PREPARATION

The following methods are used in data preparation for this study.

3.1. Selection of catchments

To meet the objectives of the trends analysis and non-stationary FFA, several criteria need to be identified before starting the data preparation task which will guide in selecting sets of initial candidate catchments. These include factors such as record length, catchment size, regulation, urbanisation, land use change, quality of data, and climate change.

Record length: The AMF record length of a stream gauging station should be long enough to represent the trend in data that can be correlated with indices and the underlying probability distribution with reasonable accuracy. In practice, the record length of AMF data in many stations is not long enough to characterise the trend and the representative probability distribution. Therefore, enough stations with a reasonable minimum record length are to be considered. More stations allow for capturing more spatial information about the study area. The reasonable long record length increases the accuracy of trend analysis and flood quantile estimation using FFA. To establish the relationship of a trend with different indices and to achieve a more appropriate fitting of probability distribution with AMF series, the gauging data record length should be relatively long. The level of uncertainties and standard errors becomes lesser with the increase of data length as compared to that of shorter length (Griffis and Stedinger, 2007). In this study, the minimum AMF record length is selected as 50 years.

Catchment size: When the catchment size becomes larger, the hydrological response of the catchment changes. The streamflow behaviour and the frequencies of occurrence of flood change significantly when the catchment size becomes larger (Haddad, 2013). ARR 2019 Book 3 (Ball et al., 2019) recommended considering small to medium-sized catchments with an upper limit of 1000 km^2 for RFFA.

Considering the above, initially, 676 stations are selected with at least 20 years of record and finally, 307 stations are selected with a minimum record length of 50 years. The average and maximum record lengths are 60 and 103 years respectively. Figure 1 shows the plot of AMF record length for the selected Queensland (QLD) stations.

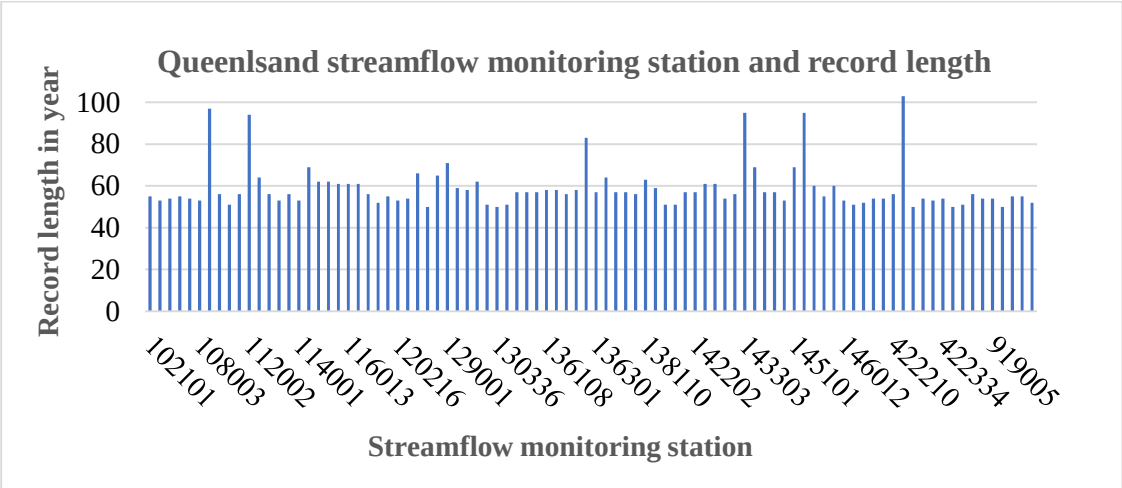

Figure 1. AMF record lengths of the selected Queensland stations

3.2. Estimation of missing data points

Despite the advancement of science and technology, missing data are still very common in metrological and hydrological time series datasets. Therefore, one of the fundamental requirements of different hydrological analysis such as trend analysis and FFA is to make appropriate decisions about these missing AMF data points. Number of methods are used for gap-filling such as interpolation from nearby stations, multiple infilling methods (Harvey et al., 2012), and statistical methods (Gedney et al., 2006). Gao et al. (2018) discussed different methods to deal with missing hydrological data. Data interpolation/fill-up techniques can be simple to complex statistical approaches. Each method has advantages and disadvantages. One of the commonly used methods is to correlate records (by linear regression analysis) among the stream gauging station having missing data points and one or more nearby stream gauging stations (Haddad and Rahman, 2010; Lopes et al., 2016; Lavers et al., 2010; Hannaford and Buy, 2012). This correlation by linear regression approach can be suitable to fill up a few scattered missing observations, to improve estimates of the variances and mean at the station having short record length, and to increase the shorter record length of a station's AMF data (Stedinger et al, 1993). The average standard error arising from the interpolation of a missing record using data from other stations can be very high compared to interpolating a missing record with the flow observation of the same gauging station (Fontaine, 1986).

In this study, gaps in the AMF series are infilled using different methods based on the availability of the monthly instantaneous maximum (IM) data, monthly maximum mean daily (MMD) data of the missing year, and the availability of the missing year's AMF data of nearby station (Rahman et al, 2009; Rahman, 1997; Haddad et al, 2008). (a) Where IM and MMD data of the missing year are available the gap is filled by comparing the IM data with MMD data of the same station for the year with data gaps. (b) Where annual maximum mean (AMD) flow is available and AMF data is missing then missing AMF data is estimated using regression between the AMD series against the AMF series of the same station. (c) Where IM and MMD are not available, simple linear regression is used between the station of missing AMF and the nearby station's AMF where the missing year's AMF is available to fill up the gap. For the in-filling gap in AMF series priority is given to the first approach followed by the second and third as appropriate.

3.3. Identification of outliers and their treatments

The outlier is the extremely high or low flow record compared to the remaining records in a streamflow data series. AMF series can contain low outliers and according to Bulletin 17B, low outliers are small values in a data series that deviate greatly from the trend of the rest of the data (IACWD, 1982). Extreme outliers (low or high) in AMF series substantially affect the flood frequency estimates (Viglione et al., 2013). Low outliers in AMF are important to identify and treat to avoid artificially low skewness in FFA using three-parameter probability distributions with AMF (Haddad and Rahman, 2010). However wrong treatment of high outliers may result in lower flood quantile estimates and should be retained in the FFA unless it is due to a data error (Haddad and Rahman, 2010). Different methods are available to identify outliers in time series data. Usually, a statistical significance test is used to identify outliers in a data set (Kottegoda, 1984). The Grubbs & Beck test (GBT, 1972) is a widely used method to identify thresholds of high and low outliers by applying a one-sided 10% significance level. The generalization of the GBT is called the multiple Grubbs-Beck test (MGBT). This study uses MGBT.

3.4. Trend test and sensitivity analysis

AMF data for the flood frequency analysis method should be stationary, homogeneous, and consistent. If AMF time series data exhibit trends or jumps, the data series is said to have inconsistency and non-homogeneity (Yevjevich & Jeng, 1969). Inconsistency in the AMF data is the variation in the systematic error (data error consistently changes in the same direction) and non-homogeneity is a change in the statistical properties of the data series (Haddad and Rahman, 2010). In this study, Mann-Kendall ((Kendall, 1970) and modified Mann-Kendall statistical trend test (Hamed and Rao, 1998) are adopted to detect trends in the AMF series data.

The sensitivity of record length on the statistical trend of AMF data series is carried out by dividing each station's data series into 5 different record lengths (10, 20, 40, 70, and more than 70 years). Trend test is applied with these 5 data series of each station with different record lengths.

4. RESULTS

4.1. Data preparation

A total of 307 stations are finally selected from all over Australia each having a minimum of 50 years of AMF record. Out of 307 stations, 23 stations have missing data. Altogether 43 data points from 23 stations are in-filled.

Method (a) is illustrated for station 412050 in the years 2004 and 2020 in Table 1 and Table 2. It can be seen that during the missing months of IM data, the MMD flows are substantially lower than the MMD and IM flow in April and in December. Therefore, it can safely be assumed that the annual IM flow for 2004 and 2020 are the values recorded in August and September, respectively.

Table 1. AMF data gap filling for the year 2020 with method (a) for station 412050 at Queensland

2020	Jan	Feb	Mar	April	May	June	July	Aug	Sep	Oct	Nov	Dec
MMD series	213	98	33	262	663	136	1466	9863	973	1390	1092	37
IM series	4273	311	323	1059	842	286	1958	15510	1257	1634	1739	0

Table 2. AMF data gap filling for the year 2004 with method (a) for station 412050 at Queensland

2004	Jan	Feb	Mar	April	May	June	July	Aug	Sep	Oct	Nov	Dec
MMD series	92.4	1.3	4.2	1.7	9.4	71.8	56.8	105.6	842.7	664	582	9.2
IM series	195.9	1.3	4.2	0.0	9.6	78.4	60.4	116.4	1292	1086	755	227.3

In Method (b), we establish linear regression between the annual MMD flow series and the annual IM series of the same station. These regression equations show a strong correlation with R^2 greater than 0.90 and were used for filling gaps, but not to extend the overall period of record.

Method (c) is used only when IM flow and MMD flow data are not available. This is just simple linear regression between the AMF of a station with a data gap and the AMF of a nearby station where the missing year's AMF data is available to fill up the gap. The R^2 value of linear regression is found smaller in method (c) than in method (b). For example, at station 412050 in NSW R^2 is 0.82, and at station 102307 in QLD R^2 is 0.57. The minimum R^2 is found to be 0.40 at station 136112 at QLD,

Table 3 shows the change of trend with the change of record length at station 215004 in NSW, station 112002 in QLD and station 226204 in Victoria (VIC). Table 3 also shows that with the change of record length, the direction of trend in AMF changes from upward (+ve) to downward (-ve) and from downward (-ve) to upward (+ve). This shows that the trend test result is quite sensitive to the length of the AMF data.

Table 3. Mann-Kendall test results for the selected 3 stations at 10% significance level

Station ID	Record Length (year)	State	H-Value	Mann-Kendall p-Value	Significance level	MK test result	Trend in AMF data	Direction of significant trend (+ ve or -ve)
112002	10	NSW	0.000	1.000	0.10	0	Upward	
112002	20	NSW	-1.950	0.051	0.10	1	Downward	-ve downward
112002	40	NSW	-1.504	0.133	0.10	0	Downward	
112002	70	NSW	3.600	0.000	0.10	1	Upward	+ve upward
112002	94	NSW	4.456	0.000	0.10	1	Upward	+ve upward
215004	10	QLD	-0.179	0.858	0.10	0	Downward	
215004	20	QLD	0.357	0.721	0.10	0	Upward	
215004	40	QLD	3.111	0.002	0.10	1	Upward	-ve downward
215004	70	QLD	0.639	0.523	0.10	0	Upward	
215004	98	QLD	-0.510	0.610	0.10	0	Downward	
226204	10	VIC	0.179	0.858	0.10	0	Upward	
226204	20	VIC	-1.720	0.086	0.10	1	Downward	-ve downward
226204	40	VIC	-1.002	0.316	0.10	0	Downward	
226204	70	VIC	-1.060	0.289	0.10	0	Downward	
226204	92	VIC	-2.617	0.009	0.10	1	Downward	-ve downward

Figure 2 shows that the AMF trend at station 215004 in Queensland changes direction after 1975 from positive to negative and the overall trend for full record length is negative. This shows the importance of using long record lengths for trend analysis.

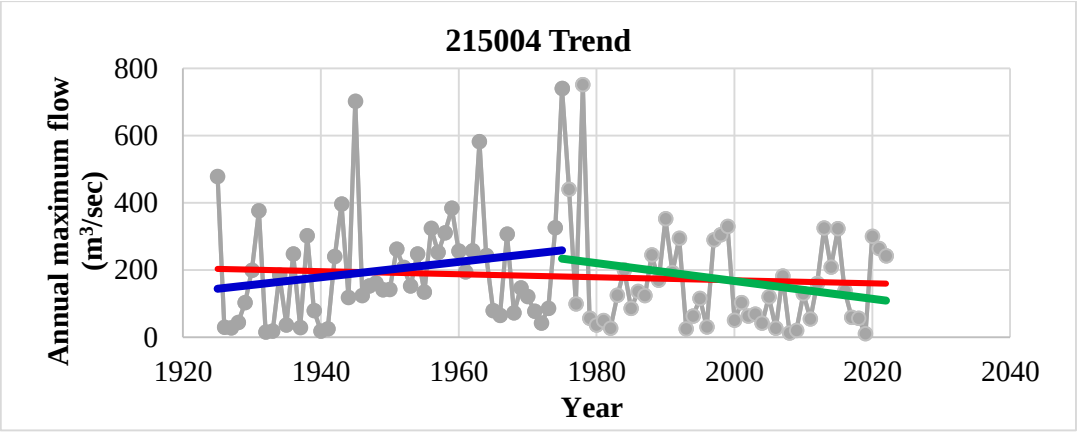

Figure 2. AMF series plot showing change of trend with record length at station 215004

5. CONCLUSIONS

This paper presents a case study detailing the streamflow data preparation procedures adopted for AMF data for all the Australian states and territories with the aim of using these AMF data for trend analysis and for non-stationary FFA. The following lessons are learnt from this study:

(a) Quality of streamflow data used for hydrological analysis are affected by different sources of uncertainties and errors (such as gaps, outliers and trends in the data).
(b) The AMF data for the selected 307 stations across Australia (having at least 50 years data length) have been prepared by filling gaps.
(c) The record length of AMF data has great impact on the result of trend analysis. A shorter record length of AMF data may not capture multiple and long-term climate variability cycles and may provide misleading trends in AMF data. Therefore, a higher record length is recommended for trend analysis.

ACKNOWLEDGMENTS

The authors would like to acknowledge Water Aauthorities of Australia for providing the streamflow data used in this study.

REFERENCES

Ball, J., Babister, M., Nathan, R., Weeks, W., Weinmann, E., Retallick, M., Testoni, I. (Editors) (2019). Australian Rainfall and Runoff: A Guide to Flood Estimation, © Commonwealth of Australia (Geoscience Australia), 2019.

Fontaine, R. A. (1986). Comparison of two stream discharge record reconstruction techniques for Eight Gaging Stations in Maine. *Selected Papers in the Hydrologic Sciences, 1986*, *2290*, 107.

Gao, Y., Merz, C., Lischeid, G., & Schneider, M. (2018). A review on missing hydrological data processing. *Environmental earth sciences*, 77(2), 1-12.

Gedney, N., Cox, P. M., Betts, R. A., Boucher, O., Huntingford, C., & Stott, P. A. (2006). A quality-controlled global runoff data set (Reply). *Nature*, *444*(7120), E14-E15.

Grubbs, F. E., & Beck, G. (1972). Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics*, *14*(4), 847-854.

Griffis, V. W., & Stedinger, J. R. (2007). Log-Pearson type 3 distribution and its application in flood frequency analysis. I: Distribution characteristics. *Journal of Hydrologic Engineering*, *12*(5), 482-491.

Haddad, K. (2013). *Regional flood frequency analysis in the range of small to large floods*: development and testing of Bayesian regression-based approaches. Doctor of Philosophy Thesis, Western Sidney University, Australia.

Haddad, K., Rahman, A., Weinmann, P. E., Lambert, M., Daniell, T., & Leonard, M. (2008). Streamflow data preparation for regional flood frequency analysis: Important Lessons from a case study. In *Water Down Under: Proceedings of the Water Down Under 2008 Conference, incorporating 31st Hydrology and Water Resources Symposium and 4th International Conference on Water Resources and Environment Research, held in Adelaide, SA., 14-17 April 2012*.

Haddad, K., Rahman, A., Weinmann, P. E., Kuczera, G., & Ball, J. (2010). Streamflow data preparation for regional flood frequency analysis: Lessons from southeast Australia. *Australasian Journal of Water Resources*, 14(1), 17-32.

Hamed, K. H., & Rao, A. R. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of hydrology*, *204*(1-4), 182-196.

Hannaford, J., & Buys, G. (2012). Trends in seasonal river flow regimes in the UK. *Journal of Hydrology*, *475*, 158-174.

Harvey, C. L., Dixon, H., & Hannaford, J. (2012). An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. *Hydrology Research*, *43*(5), 618-636.

Hydrology Subcommittee. (1982). Bulletin 17B Guidelines for determining flood flow frequency.

Interagency Advisory Committee on Water Data (IACWD) (1982). *Guidelines for Determining Flood Flow Frequency*, Bulletin #17B of the Hydrology Subcommittee, US Geological Survey, Reston, VA.

Kendall, M. G. (1970). *Rank Correlation Methods*, 4th edition, Griffen, London, 202 pp.

Kottegoda, N. T. (1984). Investigation of outliers in annual maximum flow series. *Journal of Hydrology*, *72*(1-2), 105-137.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology*, *70*(4), 407-411.

Lavers, D., Prudhomme, C., & Hannah, D. M. (2010). Large-scale climate, precipitation and British river flows: Identifying hydroclimatological connections and dynamics. *Journal of Hydrology*, *395*(3-4), 242-255.

Lopes, A. V., Chiang, J. C. H., Thompson, S. A., & Dracup, J. A. (2016). Trend and uncertainty in spatial-temporal patterns of hydrological droughts in the Amazon basin. *Geophysical Research Letters*, *43*(7), 3307-3316.

Rahman, A. (1997). *Flood Estimation for ungauged catchments: A regional approach using flood and catchment characteristics* (Doctoral dissertation, Monash University).

Rahman, A., Haddad, K., Kuczera, G., & Weinmann, P. E. (2009). Regional flood methods for Australia: data preparation and exploratory analysis. *Australian Rainfall and Runoff Revision Projects, Project*, *5*.

Stedinger, J. R., Vogel, R. M. & Foufoula-Georgiou, E. (1993). "Frequency analysis of extreme events", *Handbook of Hydrology*, Chapter 18, Maidment, D. R. (editor), McGraw-Hill, New York.

Stooksbury, D. E., Idso, C. D., & Hubbard, K. G. (1999). The effects of data gaps on the calculated monthly mean maximum and minimum temperatures in the continental United States: A spatial and temporal study. *Journal of Climate*, *12*(5), 1524-1533.

Viglione, A., Merz, R., Salinas, J. L., & Blöschl, G. (2013). Flood frequency hydrology: 3. A Bayesian analysis. *Water Resources Research*, *49*(2), 675-692.

Yevjevich, V. M., & Jeng, R. I. S. (1969). *Properties of non-homogeneous hydrologic time series* (Doctoral dissertation, Colorado State University. Libraries).