



# Probability distributions for explaining hydrological losses in South Australian catchments

S. H. P. W. Gamage<sup>1</sup>, G. A. Hewa<sup>1,2</sup>, and S. Beecham<sup>1,2</sup>

<sup>1</sup>School of Natural and Built Environments, University of South Australia, Adelaide, Australia

<sup>2</sup>Centre for Water Management and Reuse (CWMR), University of South Australia, Adelaide, Australia

Correspondence to: S. H. P. W. Gamage (sithara.walpitagamage@unisa.edu.au)

Received: 20 March 2013 – Published in Hydrol. Earth Syst. Sci. Discuss.: 10 April 2013

Revised: 20 September 2013 – Accepted: 4 October 2013 – Published: 15 November 2013

**Abstract.** Accurate estimation of hydrological losses is required for making vital decisions in design applications that are based on design rainfall models and rainfall–runoff models. The use of representative single values of hydrological losses, despite their wide variability, is common practice, especially in Australian studies. This practice leads to issues such as over or under estimation of design floods. The probability distribution method is potentially a better technique to describe losses. However, a lack of understanding of how losses are distributed can limit the use of this technique. This paper aims to identify a probability distribution function that can successfully describe hydrological losses of a catchment of interest. The paper explains the systematic process of identifying probability distribution functions, the problems faced during the distribution fitting process and a new generalised method to test the adequacy of fitted distributions. The goodness-of-fit of the fitted distributions are examined using the Anderson–Darling test and the Q–Q plot method and the errors associated with quantile estimation are quantified by estimating the bias and mean square error (MSE). A two-parameter gamma distribution was identified as one that successfully describes initial loss (IL) data for the selected catchments. Further, non-parametric standardised distributions that describe both IL and continuing loss data are also identified. This paper will provide a significant contribution to the Australian Rainfall and Runoff (ARR) guidelines that are currently being updated, by improving understanding of hydrological losses in South Australian catchments. More importantly, this study provides new knowledge on how IL in a catchment is characterised.

## 1 Introduction

Hydrological losses have wide temporal and spatial variability, but are important inputs to rainfall–runoff (RR) models. In Australia, despite their variability, current practice (IEAust., 1987) often adopts representative single values of losses as input parameter for design applications. According to Australian Rainfall and Runoff (ARR), South Australian (SA) catchments are categorised as humid, and initial loss (IL) and continuing loss (CL) values for the catchments in this zone are 10 mm and 2.5 mm h<sup>-1</sup> in winter; and 25 mm and 4 mm h<sup>-1</sup> in summer (IEAust., 1987, Book 2, p. 47). The use of single representative loss values introduces large errors into event based RR model predictions (Ilahee and Imteaz, 2009).

Considering the random nature of hydrological losses, probabilistic modelling has been suggested as a better approach to overcome the problems associated with models that use single representative values of input parameters (Rahman et al., 2000; Loveridge et al., 2012; Hill et al., 2012; Rahman et al., 2002a; Nathan et al., 2003; Kuczera et al., 2006b). A joint probability approach (JPA) that incorporates probabilistic behaviours of the input variables can improve RR simulations (Golian et al., 2012), and thus improve estimation of major flood flows that are required for the design and operation of large water infrastructure (Haddad et al., 2010a; Rahman et al., 2000, 2002b; Nathan et al., 2003; Kuczera et al., 2006a). In addition, JPAs support more accurate design flood estimations (Haddad and Rahman, 2005; Caballero et al., 2011), streamflow forecasting (Wang et al., 2011) and runoff-yield accounting (Liang et al., 2008). The most common parameters included in JPAs include initial soil moisture

**Table 1.** Distributions tested for Australian catchments.

Distribution	Catchments tested	Recommended for IL	Recommended for CL
Four-parameter beta distribution	10 Victorian catchments (Rahman et al., 2002a)	Yes	No
	15 Queensland catchments (Tularam and Ilahee, 2007)	Yes	No
	4 Victorian catchments (Ishak and Rahman, 2006)	Yes	Yes
Exponential	4 Victorian catchments (Ishak and Rahman, 2006)	No	Yes
Two-parameter gamma	4 Victorian catchments (Ishak and Rahman, 2006)	No	Yes
	5 NSW catchments (El-Kafagee and Rahman, 2011)	Yes	Yes

content, rainfall duration and intensity, and surface runoff (Golian et al., 2012; Singh et al., 2012; Wang et al., 2011; Liang et al., 2008). However, it is useful to incorporate the joint response of initial losses and total losses into rainfall runoff simulation to improve model accuracy, as the losses have wide variability (Haddad and Rahman, 2005). Although there has been significant research in Australia on the development and application of the JPAs for design flood estimation (Haddad and Rahman, 2005), there are limited probability distributions identified for hydrological losses. Therefore, the probability distributions of the loss parameters of interest need to be identified. Table 1 summarises the distribution functions of those found to be suitable for some Australian catchments. However, none of the recommended distributions were able to adequately describe the loss data of SA catchments. For hydrological analysis, regionalised methods provide more accurate estimation than generalised methods. Therefore, it is essential to identify suitable probability distributions for particular regions. As there are no studies available of South Australian loss estimation, this study will provide important information for South Australian-based hydrological modelling.

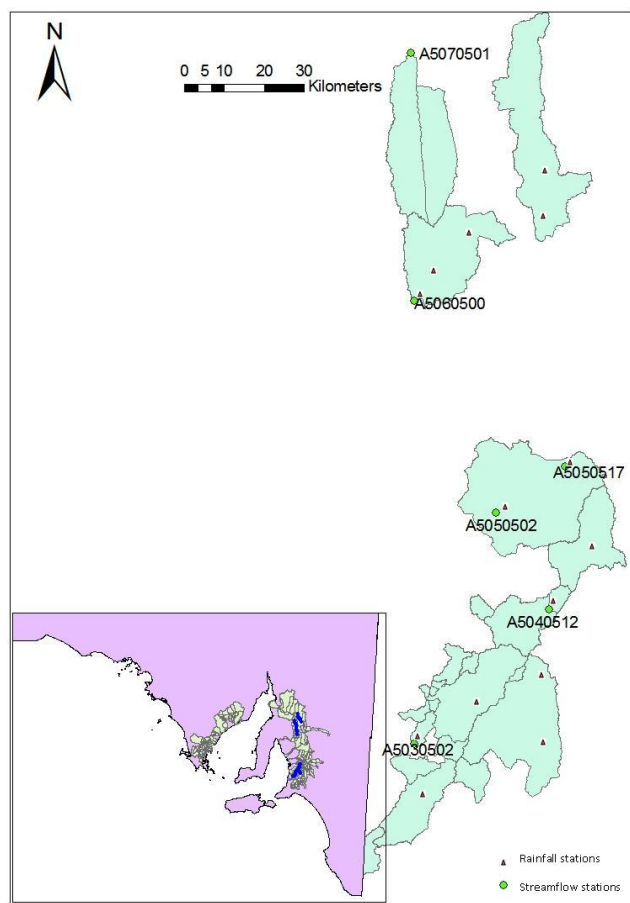
Although probabilistic methods are more attractive ways of dealing with random losses, the process of identifying probability functions needs to overcome a number of issues. For example, researchers have to identify the suitable probability function among a large number of theoretical distributions. Goodness-of-fit tests have to be used for this purpose and commonly used goodness-of-fit tests do not accurately support all the distributions (Romeu, 2003; Lugannani and Rice, 1980). Even the modern computer programs (e.g. EasyFit) that are used to identify probability distributions, use approximation methods that will cause errors in the end results. This paper highlights the issues of the commonly used Anderson–Darling test, when applied to the well-known

two-parameter gamma distribution. It is therefore necessary to have a generalised statistical goodness-of-fit test that can be accurately used for most of the distribution functions.

The distribution of losses can be estimated by either parametric or non-parametric methods. In the parametric method, parameters are estimated by equating theoretical moments of the distribution (location, scale and shape) to sample estimated moments such as mean, standard deviation and skewness. Method of moments (MOM), maximum likelihood (ML) and probability weighted moments (PWM) are commonly used methods to estimate sample estimated moments of a random variable (Trefry et al., 2004; Haktanir et al., 2010). Non-parametric methods, on the other hand do not require a distributional assumption. Non-parametric methods are accurate, uniform, and in particular can provide improved estimates of the distribution tail (Adamowski, 1989).

Non-parametric distributions of loss values are available for certain Australian catchments. For example, IL and CL values of 22 selected Victorian catchments were expressed as a proportion of the median loss value (Nathan et al., 2003). A similar approach was used for 48 rural catchments in Queensland (Ilahee, 2005) and for five catchments in the Darling Ranges in Western Australia (Waugh, 1991). Distributions of the standardised losses of these three studies were found to be largely consistent. This indicates that the shape of the standardised distribution (by median) is the same despite the data being derived from very different hydro-climatic regions from across Australia (Hill, 2010). Therefore, if the median loss rate can be estimated accurately then the standardised distribution can be applied to estimate the distribution of losses for any given catchment. In this study, a similar non-parametric method was tested for modelling hydrological losses of selected SA catchments.

The paper demonstrates extraction of IL and CL data for four selected SA catchments using time series of flow



**Fig. 1.** Location map of the study area.

and rainfall data and selection of suitable distribution function for the extracted data using both parametric and non-parametric methods. In the parametric method, Monte Carlo simulation is conducted to assess performance of the loss estimators by using bias and mean square error, and the results are validated by using two independent test catchments. The paper introduces a generalised statistical test combined with the Anderson–Darling and Kolmogorov–Smirnov (K–S) tests for testing goodness-of-fit of a fitted distribution to hydrological loss data. Subsequently, this new method is validated using the two-parameter gamma distribution. The non-parametric method used in this study involves standardising losses with the median value and investigating the trend in standardised data.

## 2 Catchment selection and data

The characteristics considered in selecting the catchments for this study are catchment regulation, size, land-use type and record lengths for available rainfall and streamflow data. The selected catchments were unregulated and had no major land-use changes during the period of their gauge record

lengths. As the selected catchments were within the small to medium size range, it can be assumed that the temporal patterns of the pluviograph data provide representative temporal patterns for the whole catchment (Ilahee, 2005). The definition of small to medium size catchments is arbitrary and is considered to have an upper limit of 1000 km<sup>2</sup> in area (Haddad et al., 2010b). The record length of data should be at least 10 years for adequate empirical analysis (Boni et al., 2007; Jingyi and Hall, 2004; Kumar and Chatterjee, 2005). The four selected catchments for the analysis – Scott Bottom (A5030502), Mt Pleasant (A5040512), Yaldara (A5050502) and Penrice (A5050517) – and two selected catchments for the validation – Rhynie (A5060500) and Spalding (A5070501) – all satisfied these conditions. A location map of the selected catchments is given in Fig. 1 and summary details of the geographic, climatic and meteorological data for each catchment are provided in Table 2. The catchment rainfall and streamflow data were collected from the Department of Environment, Water and Natural Resources (DEWNR), South Australia. The four selected stations – A5030502, A5040512, A5050502 and A5050517 – having high quality data for 37, 35, 47 and 30 yr, respectively.

Number of rainfall and runoff events selected for this study is presented in Table 3.

## 3 Methodology

### 3.1 Loss calculation

For each of the selected catchments, rainfall events that produce a reasonable amount of runoff were extracted for this study using the HYDSTRA (KISTERS, 2008) program. In this study, the threshold runoff is selected as 0.01 mm h<sup>-1</sup>. The IL, which is defined as the amount of rainfall that occurs before the start of runoff, was calculated using Eq. (1).

$$IL = \sum_{i=1}^n I_i, \quad (1)$$

where  $n$  is the duration in hours from the start of the storm burst to the start of the surface runoff (rainfall excess) and  $I_i$  is rainfall in mm in the  $i$ th hour.

Measured streamflow data at a gauged station usually comprises quickflow (QF) (rainfall excess) and baseflow (BF) components. For hydrological loss estimations, only the QF is of interest. Therefore, the BF needs to be separated from the original total streamflow data prior to loss calculation. Nathan and McMahon (1990) compared the Lyne and Hollick method of BF separation with several other rigorous algorithms and concluded that it was simple to use, yet produced as good results as the alternatives. Hence, in this study, the Lyne and Hollick algorithm, which is in-built in the HYDSTRA (KISTERS, 2008) program, was used for BF separation.

The total rainfall (TR) resulting from a rainfall event can be expressed by Eq. (2) and hence this can be rearranged as in

**Table 2.** Geographical, climatic and meteorological data of the study catchments.

Characteristics	Scott Bottom (A5030502)	Mt Pleasant (A5040512)	Yaldara (A5050502)	Penrice (A5050517)
River	Scott Creek	Torrens	North Para	North Para
Area (km <sup>2</sup> )	27	26	384	118
Annual Rainfall (mm)	69–74	80–103	93–102	63–67
Elevation at gauging station (m)	205	415	145	285
Evaporation (mm day <sup>-1</sup> )	1.38	1.69	1.69	1.69

**Table 3.** Selected rainfall events.

Catchments	Rainfall period considered	No. of events selected
Scott Bottom (A5030502)	1991–2010	200
Mt Pleasant (A5040512)	1989–2011	227
Yaldara (A5050502)	1985–2011	200
Penrice (A5050517)	1986–2011	185
Rhynie (A5060500)	1985–2011	208
Spalding (A5070501)	1992–2011	142

Eq. (3) to calculate the CL, which is defined as the average loss in mm h<sup>-1</sup> over the remaining duration of the rainfall event.

$$TR = IL + CL \times \Delta t + QF \quad (2)$$

$$CL = \frac{TR - IL - QF}{\Delta t}, \quad (3)$$

where TR, IL and QF are in mm, CL is in mm h<sup>-1</sup> and  $t$  is the time (in h) elapsed between the start of the surface runoff and the end of the rainfall event.

### 3.2 Parametric method for describing IL

In the parametric approach, distribution parameters are estimated by equating sample estimated moments of a random variable to moments of a known theoretical distribution. In this study, the observed IL data series,  $X(1), \dots, X(n)$  is the sample of a random variable of interest. Identification of a theoretical distribution that can reasonably describe the observed loss data was conducted using Anderson–Darling and Q–Q plot methods which are described in the following.

#### Anderson–Darling (A–D) test

In the Anderson–Darling test procedure, the fit of an observed cumulative distribution function to an expected cumulative distribution function is compared. If the estimated A–D

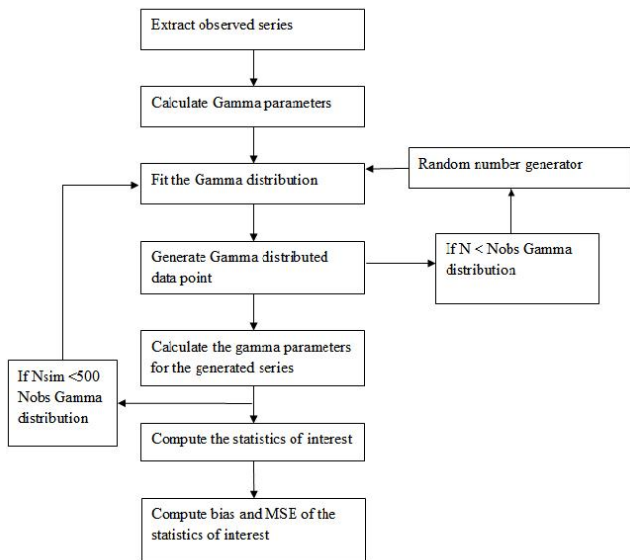
test statistic at the 95 % confidence level is greater than the critical value of the selected distribution function, the distribution function is accepted as a suitable candidate to describe the observed data.

In general, critical values of the A–D test statistic depend on the specific distribution being tested. However, tables and formulae of critical values for many distributions are not currently available (Romeu, 2003). Additionally, for some distributions, such as the two-parameter gamma distribution, finding the critical value of the A–D statistic is very complicated (Lugannani and Rice, 1980). Currently, formulae of critical values are available for the normal, log-normal, exponential and Weibull distributions (Romeu, 2003). The A–D test that is implemented in several modern computer programs such as “EasyFit” (MathWave Technologies) often uses the same critical values for every distribution. These critical values are calculated using the approximation formula, and depend on the sample size only. Therefore, in this study the A–D test is carried out for only three distributions: normal, exponential and Weibull. For the other distributions, the Q–Q plot was used as an alternative comparison method.

#### Q–Q plots

The Q–Q plots provide a graphical assessment of “goodness-of-fit” and indicate whether or not the selected sample could have come from the selected target distribution. In this study, Q–Q plots were drawn for the observed IL and CL data series to test whether the data series are derived from seven selected theoretical distributions, namely normal, log-normal, Pareto, Weibull, gamma and exponential for which Q–Q plots can be drawn using the SPSS statistical software.

In a Q–Q plot, if one or both of the axes is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are uniquely defined and can be obtained by inverting the CDF (Ledolter and Hogg, 2010). However, for the observed data with an unknown distribution, the quantile estimations for constructing the Q–Q



**Fig. 2.** Flowchart for data fitting, simulation and evaluation. Nobs – number of observations; Nsim – number of simulations; MSE – mean square error.

plots need to be done using an appropriate plotting position formula. The plotting position formula used in this study is given in Eq. (4).

$$\frac{i - 0.5}{n} = f(q_i), \tag{4}$$

where  $n$  is the sample size,  $i$  is rank of an observation in the sample organised in an ascending order and  $f(q_i)$  is the quantile of the observed data.

The  $x$  axis of the Q–Q plots consists of order statistics,  $x(1) \leq x(2) \leq \dots \leq x(n)$  with a theoretical CDF,  $F(x) = P(X \leq x)$ . The  $y$  axis of the Q–Q plots consists of quantiles of observed data.

After constructing the Q–Q plots, the theoretical distribution that provides a reasonable fit can be selected. The selected distribution then needs to be further investigated to assess possible estimation errors.

The Q–Q plot analyses confirms that, of the five distributions investigated, the gamma distribution is the most appropriate to describe the observed loss data. Hence, the methodology of the parametric modelling is now described with particular reference to the gamma distribution. The sequential steps involved in the parametric modelling are presented in Fig. 2.

### 3.2.1 Fitting and testing the gamma distribution

#### Gamma distribution

For a two-parameter continuous probability distribution function (PDF) with a shape parameter ( $k$ ) and a scale parameter ( $\theta$ ), while the shape parameter can be denoted by

$\alpha = k$ , an inverse scale parameter can be denoted by  $\beta = 1/\theta$  (Freund and Johnson, 2010).

The PDF of the gamma distribution is given in Eq. (5) and the gamma function  $\Gamma(\alpha)$  is given in Eq. (6).

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \tag{5}$$

where  $0 \leq x \leq \infty$ , and parameters  $\alpha > 0$  and  $\beta > 0$ .

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \tag{6}$$

$\Gamma(\alpha)$  is a generalised factorial that can be shown as  $\Gamma(\alpha) = (\alpha - 1)!$ , if  $\alpha$  is a positive integer. The gamma function for the arguments ( $\alpha$ ) between 0 and 1 can be found from standard mathematical tables.

#### Inverse gamma distribution

The inverse gamma distribution is a two-parameter family of continuous probability distributions on the positive real line, which is the distribution of the reciprocal of a variable that follows gamma distribution. The inverse gamma distribution’s PDF is given in Eq. (7).

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (x)^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right), \tag{7}$$

where  $x > 0$  and  $\alpha$  and  $\beta$  are the shape and scale parameters, respectively.

#### Parameter estimation

Once the distribution is selected, the next step is to estimate its parameters. The most commonly used methods for determining parameters of a PDF include Method of Moments (MOM), Maximum Likelihood (ML) and Probability Weighted Moments (PWM). In this study, MOM was adopted to estimate parameters because of its simplicity and ease of use. Estimation of distribution parameters involves equating theoretical moments of the distribution to the sample estimated moments. For the gamma distribution, the first two theoretical moments are given in Eqs. (8) and (9) (Freund and Johnson, 2010).

$$\mu = \alpha\beta \tag{8}$$

$$\sigma^2 = \alpha\beta^2 \tag{9}$$

Replacing  $\mu$  and  $\sigma^2$  in Eqs. (8) and (9) by sample estimated  $\sigma^2$  and  $s^2$ , the estimates of the gamma parameters are given in Eqs. (10) and (11).

$$\hat{\alpha} = \frac{\bar{x}^2}{s^2} \tag{10}$$

**Table 4.** Anderson–Darling statistics for normal, exponential and Weibull distributions.

Distribution	$p$ Value	A–D <sup>2</sup>	95% critical value
Normal	IL 0.02430	163.84	166.97
	CL 0.00034	240.25	242.25
Exponential	IL 0.00156	272.25	272.90
	CL 0.00005	445.21	446.27
Weibull	IL 0.00143	784.00	793.91
	CL 0.00001	900.00	911.38

$$\hat{\beta} = \frac{s^2}{\bar{x}} \quad (11)$$

### 3.2.2 Simulating data

Although the true distribution of loss data is not known, it is still interesting to understand the errors associated with the estimated quantiles, when data are derived from a gamma distribution. Hence, in this study, the estimated gamma distribution using the observed data series was assumed to be the true distribution, and quantiles estimated from the observed data were assumed to be the true quantiles in estimating bias and MSE associated with the estimated quantiles. The estimated quantiles in this study correspond to non-exceedance probabilities of 0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.999. Probability values of 0 and 1 were intentionally avoided to eliminate mathematical errors when applying the gamma inverse function. With the use of a uniform random number generator (discussed in the next section), 500 simulated gamma samples of losses of the same size as the observed series were generated. For each of the generated samples, quantiles at 10 selected non-exceedance probabilities were calculated. Hence, the simulated quantile estimation involved (1) calculating gamma parameters for each of the 500 simulated series, (2) applying a gamma inverse function for the series in step 1 to obtain simulated loss quantiles at each selected non-exceedance probability, and (3) estimating 500 simulated loss quantiles for each non-exceedance probability. The sequences of these steps are shown in Fig. 2.

### Random number generation

A reliable source of random numbers, and a means of transforming them into prescribed distributions, is essential for the success of the simulation approach. Generated random numbers,  $x$ , should belong to a domain,  $x \in [x_{\min}, x_{\max}]$ , in such a way that the frequency of occurrence (or probability density) will depend upon the value of  $x$  in a prescribed functional form  $f(x)$  (Saucier, 2000). Among the various techniques available for generating random numbers, most of the methods presume that a supply of uniformly distributed

random numbers are in the half-closed unit interval  $[0, 1)$  (Saucier 2000). The methods for random number generation include inverse transformation, composition, convolution, acceptance–rejection, sampling and data-driven techniques, techniques based on number theory and Monte Carlo simulation (Saucier, 2000). If the inverse form of a distribution function ( $F^{-1}$ ) is not available, then the inverse transformation technique is not feasible and other techniques need to be considered. However, in this study, the inverse transformation technique can be used because the inverse gamma function is available. Also, the inverse transformation is a simple, efficient and commonly used technique (Saucier, 2000).

### 3.2.3 Evaluating quantiles

As mentioned earlier, the estimated quantiles were evaluated by comparing bias and MSE estimated using 500 simulated samples. In addition, probability plots were constructed to determine the validity of the gamma distribution for describing the loss data.

#### Bias

The bias (or bias function) of an estimator is the difference between expected value of the estimator and the true value of the parameter being estimated, and is given in Eq. (12).

$$\text{Bias} = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta, \quad (12)$$

where  $E[\ ]$  denotes the expected value over the distribution  $P(x/\theta)$  (Lebanon, 2010).  $P(x/\theta)$  is a probability distribution for observed data  $x$ , with parameter  $\theta$ .

#### Mean square error (MSE)

The MSE is a measure of the variance of error in the quantile estimator and is used to give an overall measure of accuracy. The MSE thus assesses the quality of an estimator in terms of its variation and unbiasedness. The MSE of an estimator is given in Eq. (13).

$$E(\|\hat{\theta} - \theta\|^2) = E\left(\sum_{j=1}^d (\hat{\theta}_j - \theta_j)^2\right) = [\text{bias}(\hat{\theta})]^2 + \text{var}(\hat{\theta}) \quad (13)$$

#### Confidence interval (CI)

In this study, the level of confidence is set as 95 %, which reflects a significance level of 0.05. For the gamma distribution, exact confidence intervals are difficult to construct and the available methods for finding confidence interval of the gamma distribution are very complex (Fay and Feuer, 1997; Banneheka, 2012). Therefore, in this study, a simple approximation was undertaken to determine the CI. The upper confidence level (UCL) and lower confidence level

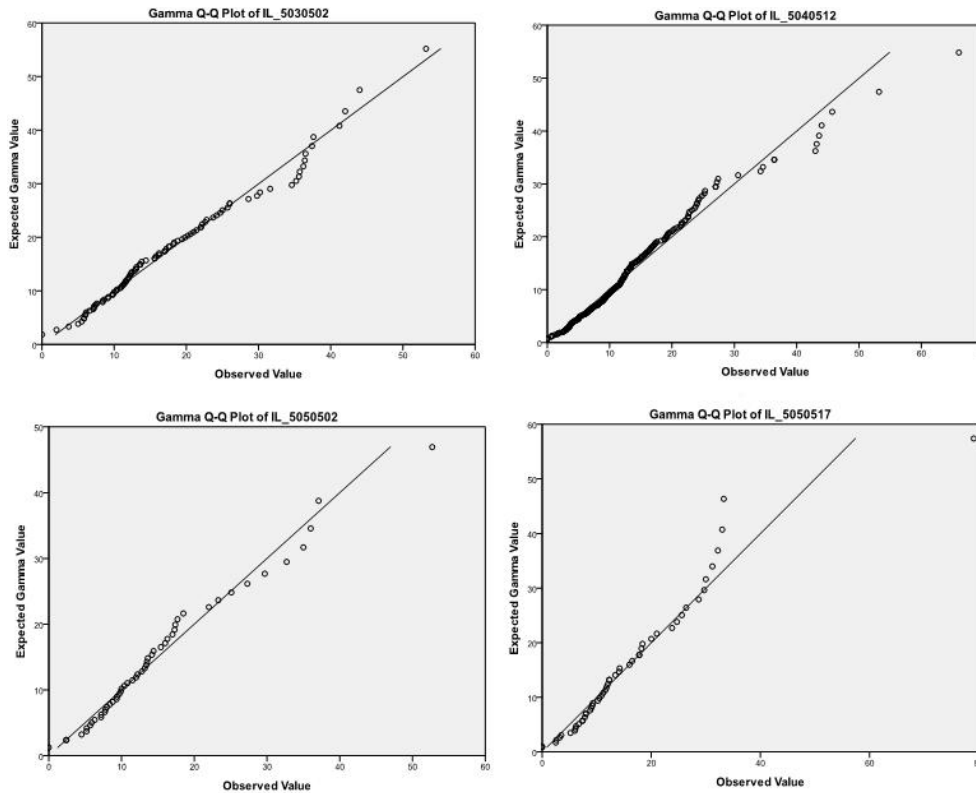


Fig. 3. Gamma Q–Q plots for the observed IL data.

(LCL) were calculated as  $UCL = f(x_u)$  and  $LCL = f(x_l)$  respectively, where  $x_u = 0.95n$ ,  $x_l = 0.05n$  and  $n$  is the sample size. As the study generated 500 simulated series,  $n = 500$ .

### 3.3 Generalised method to test the gamma distribution

As mentioned in Sect. 3.2, goodness-of-fit tests such as the A–D test can only be used for limited distributions. This is also the case for the K–S test. However if the observed data transform to the CDF of the candidate distribution, the original data set can be tested as a candidate distribution using the A–D or K–S using the method described below. This method is particularly suitable for distributions that have inverse functions and for distributions whose parameters cannot be generalised. In this study, this method is introduced based on the two-parameter gamma distribution. However it can also be modified for other distributions as well.

The theory for the introduced method is as follows.

- Consider the set of loss data as  $X_1, X_2, \dots, X_n$
- Fit the data set to the two-parameter gamma distribution and compute parameters  $\alpha$  and  $\beta$
- Calculate
- $\hat{\alpha}$  and  $\hat{\beta}$
- using the maximum likelihood method

- Make the hypothesis that  $H_0: X_1, X_2, \dots, X_n$  are drawn from  $f(x|\hat{\alpha}, \hat{\beta})$
- Compute  $Y_i = F(X_i|\hat{\alpha}, \hat{\beta})$
- where  $i = 1, 2, \dots, n$  and  $F(X|\alpha, \beta)$  is the CDF given by Eq. (14):

$$F(X|\alpha, \beta) = \int_0^x f(y|\alpha, \beta) dy. \tag{14}$$

- The theorem is that if  $H_0$  is true then  $Y_i$  are independent and uniformly distributed on  $[0,1]$
- Compute the empirical distribution of  $Y_1, Y_2, \dots, Y_n$  as given in Eq. (15)

$$G_n(x) = \frac{1}{n} \sum_1^n I(Y_i \leq x), \quad 0 \leq x \leq 1 \tag{15}$$

- This is compared with the true distribution  $G_n(x) = x$

In the Kolmogorov–Smirnov test, let the component  $\max |G_n(x) - G(x)| = \max |G_n(x) - x|$ ,  $0 \leq x \leq 1$ . This is simple to calculate as this maximum occurs at  $x = Y_1, Y_2, \dots, Y_n$ .

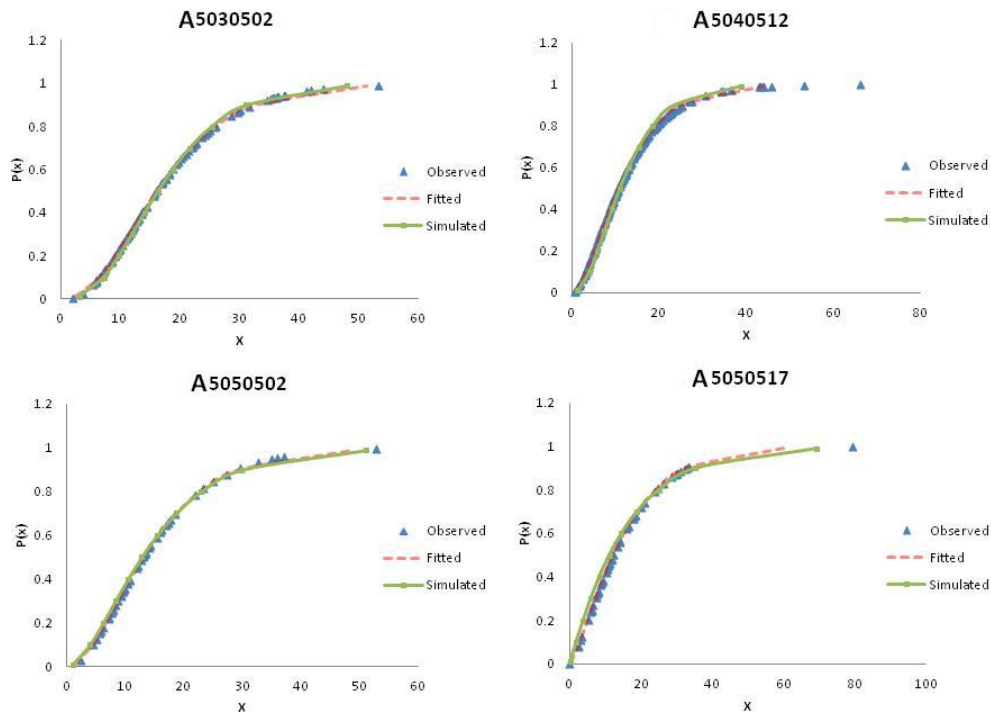


Fig. 4. Probability plots of the study catchments.

Therefore, the max is

$$\max |G_n(Y_i) - Y_i|, 1 \leq i \leq n. \tag{16}$$

In the Anderson–Darling test,

$$S = \sum_{k=1}^n \frac{2k-1}{n} [\ln(G(Y_k)) - \ln(1 - G(Y_{n+1-k}))], \tag{17}$$

where  $Y_1 \leq Y_2 \leq \dots \leq Y_n$ .

Therefore,  $S = \sum_{k=1}^n \frac{2k-1}{n} [\ln y_k - \ln(1 - Y_{n+1-k})]$ , which again is simple to calculate.

The advantage of this theory is that as the observed data set  $X_1, X_2, \dots, X_n$  are transposed to  $Y_1, Y_2, \dots, Y_n$ , the  $Y$  series has to be tested for uniform distribution. In most of the available goodness-of-fit tests, critical values and  $p$  values are available for the uniform distribution, therefore the test can be carried out accurately. If the  $Y$  series follows a uniform distribution, then the  $X$  series should be from the gamma distribution.

This method can also be adapted for other distributions, in which case parameters of the particular distribution should be determined. Then, the CDF of the particular distribution should be determined and the resulting  $Y$  series can then be tested for uniform distribution.

### 3.4 Validating the selected distribution

In order to demonstrate the applicability of the selected distribution to other parts of SA, the recommended

two-parameter gamma distribution was tested for two test catchments, which follow the same criteria described in Sect. 2. The two test catchments, A5060500 and A5070501, both belong to the SA wet/humid region, according to ARR. Probability plots, bias and MSE estimates, and the introduced goodness-of-fit tests were investigated for these two test catchments.

### 3.5 Non-parametric method

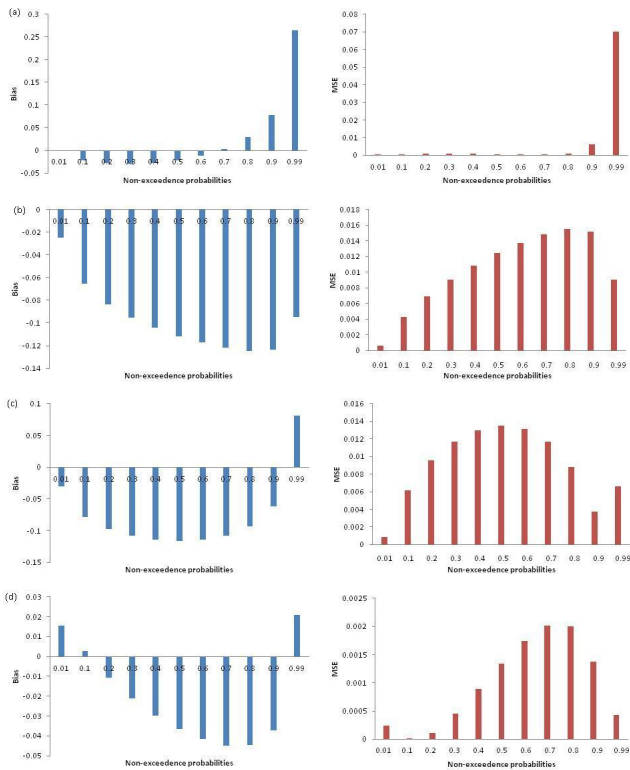
A non-parametric method, as the name implies, has no dependency on parameters. Non-parametric plots were drawn for all the catchments with the y axis as the standard storm IL and CL (fraction of the median) and the x axis as the proportion of the sample value exceeded (%). The plots were also compared with similar studies carried out for other parts of Australia.

## 4 Results and discussion

### 4.1 Parametric method

One objective of this study is to identify a suitable parametric distribution that can describe IL and CL data of the selected SA catchments. The A–D test was carried out to investigate whether the IL and CL data follow the normal, exponential or Weibull distributions. The A–D test statistics are presented in Table 4.





**Fig. 5.** Bias and MSE for simulated samples for catchments (a) A5040512, (b) A5030502, (c) A5050502 and (d) A5050517.

As all the  $p$  values of the A–D tests are less than 0.05, and  $A-D^2$  is less than the critical value, the null hypothesis (samples follow the above mentioned distribution) can be rejected. Therefore, the IL and CL data are not well fitted by normal, exponential or Weibull distributions.

Instead, the two-parameter gamma distribution was selected from six theoretical distributions using Q–Q plots. Figure 3 shows the gamma Q–Q plots for the observed IL data for the four selected catchments. Parameters of the gamma distribution that were calculated for the observed set of IL data are presented in Table 5.

In all four catchments, the higher values of the losses ( $> 30$  mm) deviate more from the  $y = x$  line. This means that extreme loss values, which are less than 5 % of the total number of IL values deviate from the gamma distribution. It is necessary to determine how much these extreme values deviate from the gamma distribution and this will be explained by using probability plots, later in this discussion.

Although the observed CL data were also checked for the same distributions, none of the Q–Q plots for the four catchments followed the  $y = x$  line.

The IL quantiles that were calculated for the selected non-exceedance probabilities are presented in Table 6. In Table 6, the values of  $\text{Sim}(x)$  were calculated using a 95 % confidence interval. The last column of the table, which provides the range of IL values, was generated considering both the

observed and simulated data. These ranges with their non-exceedance probabilities are useful for design applications.

Probability plots (PDFs for the observed, fitted and simulated IL data for each catchment) were used to demonstrate that the data simulated using the two-parameter gamma distribution match well with the observed data. The probability plots for the selected catchment are shown in Fig. 4 with a randomly selected simulated sample. Evaluating the fitted curves and observed data, it can be concluded that the IL data follow the two-parameter gamma distribution very well. In addition, these probability plots show that the simulated and the observed data are very close. Although there is a deviation of fitted, observed and simulated series in their high values, the difference is very small.

As mentioned earlier, the Q–Q plots indicate that extreme loss values do not closely follow gamma distribution as other data values. However, the probability plots indicate that the gap between the fitted (theoretical line) and simulated values is very small. Therefore the two-parameter gamma distribution is not limited to low and medium ranges of loss values, but can also be used to represent the high values. In addition, the error caused by the quantile estimation and simulation should also be quantified. Although the probability plots provide certain estimation of the error associated with the simulation, the errors have been further quantified using other methods, namely bias and MSE of the estimator.

The bias and MSE for the four selected catchments are presented in Fig. 5. Because all the values are close to zero, it can be concluded that the IL simulated using a two-parameter gamma distribution is accurate and very close to observed data. However, the values of bias and MSE can change slightly according to the different selected simulated samples.

After comparing a number of samples (observed data series with different randomly selected simulated data series), it was observed that the bias and MSE are comparatively lower in low non-exceedance probabilities. The bias and MSE have a relatively wide range in the higher non-exceedance probabilities relative to low non-exceedance probabilities. Despite this, the bias and MSE values are always close to zero. Therefore it can be assumed that the range of IL, and its probability of occurrence as shown in Table 6, can be used in design applications. However those values should only be used for hydrologically similar catchments.

In addition, the theorem introduced in Sect. 3.3 was also used to test the selected two-parameter gamma distribution. The observed IL data sets of all four catchments were transposed to the CDF of the gamma distribution and the resulting data series then tested to determine if it is from the uniform distribution. From the A–D and K–S tests, it was found that the distribution of the transposed series followed the uniform distribution for all four catchments. Therefore, this test also confirmed that the observed IL data of the selected catchments can be described by the two-parameter gamma distribution.

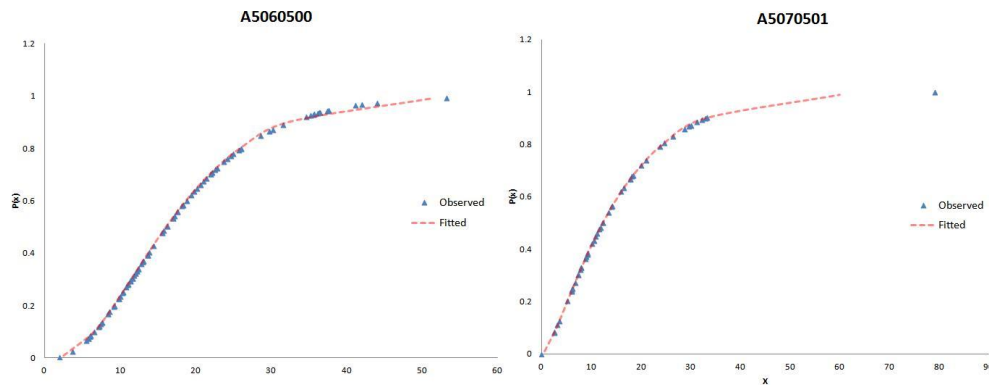


Fig. 6. Probability plots for the test catchments.

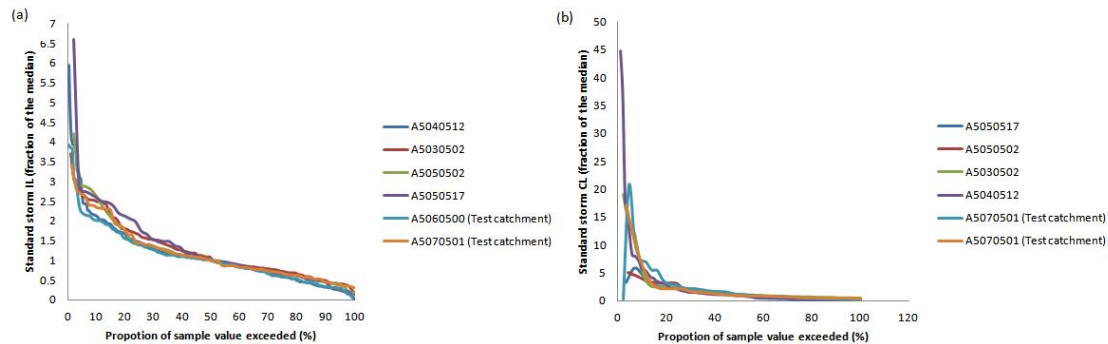


Fig. 7. IL and CL standardised by median values.

#### 4.2 Two-parameter gamma distribution for other SA catchments

In order to demonstrate that the two-parameter gamma distribution can be used to explain the IL of other SA wet/humid catchments, the observed IL data series of test catchments A5060500 and A5070501 were fitted to the two-parameter gamma distribution. Figure 6 shows the probability plots for these test catchments. When fitting the two-parameter gamma distribution, the gamma parameters ( $k$  and  $\theta$ ) were taken as the average values given in Table 5. The bias for the catchments A5060500 and A5070501 were estimated as  $-0.01946$  and  $-0.02223$  and the MSE of the catchments were calculated as  $0.000597$  and  $0.000706$ , respectively. The probability plots presented in Fig. 6, indicate that the observed IL of these test catchments follow the two-parameter gamma distribution. Also, the bias and MSE values being closer to zero further support this conclusion. Finally, by testing the observed IL data series using the introduced goodness-of-fit test, it can be stated that the two-parameter gamma distribution is suitable for SA wet/humid catchments. However, it should be noted that this study has not examined the dry catchments that exist in the SA region and therefore these could be the subject of future investigations.

#### 4.3 Non-parametric method

A non-parametric method of describing losses was also developed. The non-parametric method used in this study involves standardising both IL and CL with their median values. Median values have been used for standardisation in similar studies conducted in other parts of Australia including southeastern Australian catchments (Hill, 1996), Queensland catchments (Ilahee 2005) and southwestern Australian catchments (Waugh, 1991). Using the median values for the standardisation in this study allows the distributions of losses across the different catchments to be directly compared, as shown in Fig. 7. Figure 7a shows the distribution of standardised IL while Fig. 7b shows the distribution of standardised CL for both study and test catchments. Figure 7a and b indicates that both IL and CL have consistency when standardised by median values. The shapes of the non-parametric distributions presented in this study are consistent with similar studies conducted for other regions of Australia (Nathan and Weinmann, 2004; Ilahee, 2005; Nathan et al., 2003; Waugh, 1991). In particular, Nathan et al. (2003) show that the shape of the standardised IL distribution does not change with location. Thus despite the data being derived from very different hydro-climatic regions across Australia, the results clearly

**Table 5.** Estimated gamma distribution parameters.

Station No.	A5040512	A5030502	A5050502	A5050517	Average
Shape parameter ( $k$ )	1.798	2.714	2.015	1.450	1.99
Scale parameter ( $\theta$ )	0.142	0.153	0.137	0.094	0.131

**Table 6.** Estimated IL quantiles for selected non-exceedance probabilities.

$P(x)$	A5030502		A5040512		A5050502		A5050517		Range
	obs(x)	sim(x) LCL-UCL	obs(x)	sim(x) LCL-UCL	obs(x)	sim(x) LCL-UCL	obs(x)	sim(x) LCL-UCL	
0.01	2.53	1.52–3.50	0.85	0.46–1.21	1.39	0.78–2.11	0.56	0.46–0.96	1–3
0.1	6.53	4.49–7.90	3.24	3.03–5.31	4.48	3.43–5.92	2.94	2.02–3.29	3–7
0.2	9.15	7.70–10.01	5.11	4.09–5.67	6.73	6.17–7.92	5.12	4.07–5.71	5–9
0.3	11.46	10.00–12.65	6.88	6.05–9.27	8.79	8.02–9.81	7.31	6.24–7.89	7–11
0.4	13.72	13.66–16.00	8.71	7.67–11.87	10.87	10.79–11.82	9.64	7.56–9.92	9–14
0.5	16.10	15.03–16.90	10.69	9.64–11.34	13.10	14.01–16.92	12.26	9.16–13.01	11–16
0.6	18.73	17.67–19.61	12.96	11.90–13.82	15.61	15.53–20.32	15.32	11.21–16.92	13–19
0.7	21.86	20.80–23.59	15.73	14.66–17.09	18.64	20.56–25.01	19.12	14.00–21.72	16–22
0.8	25.93	25.88–27.01	19.42	17.34–21.03	22.65	25.57–33.21	24.29	19.17–26.37	19–26
0.9	32.34	29.31–33.46	25.40	23.32–28.71	29.07	29.01–33.21	32.85	30.72–37.84	25–33
0.99	51.34	45.37–58.32	43.87	40.77–50.92	48.57	40.59–55.92	59.99	49.90–72.13	44–60

show that while the magnitude of losses may vary between different catchments, the shape of the distribution does not.

In addition, it can be concluded that the variation of IL and CL values across the four catchments are higher for the proportion of sample exceedance that are less than 10 %. This variation is higher for the CL than the IL values. In this study, the standardised CL varied from 2 to 45, and IL values varied from 0 to 8 while in other similar studies (Nathan and Weinmann, 2004; Ilahee, 2005; Nathan et al., 2003; Waugh, 1991) CL varied from 1 to 14 and IL varied from 0 to 8 (Nathan and Weinmann, 2004; Ilahee, 2005; Nathan et al., 2003; Waugh, 1991). A major limitation of this method is that the distribution is highly dependent on the median value and for this a large sample size is required.

## 5 Conclusions

This paper investigates both parametric and non-parametric methods to describe hydrological losses. The two-parameter gamma distribution was successfully fitted for observed IL data. For each catchment, the parameters were estimated and the IL values were simulated from the two-parameter gamma distribution. The simulated data compare very well with the observed data, with some tendency to overestimate the occurrence of higher losses. The parameters and CDF of the gamma distribution can be used to find the frequency distribution and can be used to estimate the probability of occurrence of IL in design applications. The problems associated with the existing goodness-of-fit tests were also discussed and new goodness-of-fit test theory was introduced. Although the new theory described in this paper is based on

the two-parameter gamma distribution, it can easily be developed for other distributions. In addition, this method can be used for any set of hydrological data, to test whether or not the data can be fitted to a particular distribution. However, for the CL component, none of the parametric distributions seems to fit the observed data satisfactorily. The non-parametric method tested, which is the standardised distribution of both IL and CL over median values, exhibits a remarkable degree of consistency with other studies. These standardised values can therefore be used in design applications. The presented work contributes to ongoing research on losses in Australia for updating *Australian Rainfall Runoff (ARR) – A Guide for Flood Estimation in Australia*. Applicability of the introduced methods for different hydrological regions is a potential area for further research.

*Acknowledgements.* The authors would like to thank Associate Professor John van der Hoek of the University of South Australia for his advice on the statistical analyses.

Edited by: N. Verhoest

## References

- Adamowski, K.: Monte Carlo comparison of parametric and non-parametric estimation of flood frequencies, *J. Hydrol.*, 108, 295–308, doi:10.1016/0022-1694(89)90290-4, 1989.
- Banneheka, B.: Confidence intervals for the median of a gamma distribution, *Vidyodaya J. Science*, 15, 123–129, 2012.

- Boni, G., Ferraris, L., Giannoni, F., Roth, G., and Rudari, R.: Flood probability analysis for un-gauged watersheds by means of a simple distributed hydrologic model, *Adv. Water Resour.*, 30, 2135–2144, 2007.
- Caballero, W. L., Taylor, M., Rahman, A., and Shrestha, S.: Regionalisation of intensity-frequency-duration data: A case study for New South Wales, in: MODSIM 2011 – 19th International Congress on Modelling and Simulation – Sustaining Our Future: Understanding and Living with Uncertainty, 19th International Congress on Modelling and Simulation – Sustaining Our Future: Understanding and Living with Uncertainty, MODSIM2011, 21 December 2011–16 December 2011, Perth, WA, Australia, 3775–3781, 2011.
- El-Kafagee, M. and Rahman, A.: A study on initial and continuing losses for design flood estimation in New South Wales, in: MODSIM 2011 – 19th International Congress on Modelling and Simulation – Sustaining Our Future: Understanding and Living with Uncertainty, 19th International Congress on Modelling and Simulation – Sustaining Our Future: Understanding and Living with Uncertainty, MODSIM2011, 12 December 2011–16 December 2011, Perth, WA, Australia, 3782–3787, 2011.
- Fay, M. P. and Feuer, E. J.: Confidence intervals for directly standardized rates: a method based on the gamma distribution, *Stat. Medicine*, 16, 791–801, 1997.
- Freund, J. and Johnson, R. A.: *Miller and Freund's probability and statistics for engineers*, Pearson Education (US), 2010.
- Golian, S., Saghafian, B., and Farokhnia, A.: Copula-based interpretation of continuous rainfall-runoff simulations of a watershed in Northern Iran, *Can. J. Earth Sci.*, 49, 681–691, doi:10.1139/e2012-011, 2012.
- Haddad, Rahman, A., and Weinmann, P. E.: Estimation of major floods: Applicability of a simple probabilistic model, *Aust. J. Water Resour.*, 14, 117–126, 2010a.
- Haddad, K. and Rahman, A.: Regionalisation of rainfall duration in Victoria for design flood estimation using Monte Carlo simulation, in: MODSIM05 – International Congress on Modelling and Simulation: Advances and Applications for Management and Decision Making, Proceedings, International Congress on Modelling and Simulation: Advances and Applications for Management and Decision Making, MODSIM05, 12 December 2005–15 December 2005, Melbourne, VIC, Australia, 1827–1833, 2005.
- Haddad, K., Rahman, A., Weinmann, P., Kuczera, G., and Ball, J.: Streamflow data preparation for regional flood frequency analysis: lessons from southeast Australia, *Aust. J. Water Resour.*, 14, 17–32, 2010b.
- Haktanir, T., Cobaner, M., and Kisi, O.: Frequency analyses of annual extreme rainfall series from 5 min to 24 h, *Hydrol. Process.*, 24, 3574–3588, doi:10.1002/hyp.7759, 2010.
- Hill, P. I.: Empirical analysis of data to derive losses: methodology, programs and results/P. I. Hill, U. K. Maheepala [and] R. G. Mein, Working document (Cooperative Research Centre for Catchment Hydrology); 96/5., available at: <http://nla.gov.au/nla.cat-vn1810945>, edited by: Mein, R. G., Maheepala, U. K., and Cooperative Research Centre for Catchment, H., Cooperative Research Centre for Catchment Hydrology, Clayton, Vic., 1996.
- Hill, P. I.: Project 6: Loss models for catchment simulation, Engineers Australia, ACT, Project report, 2010.
- Hill, P. I., Graszekiewicz, Z., Sih, K., Nathan, R., Loveridge, M., and Rahman, A.: Outcomes from a pilot study on modelling losses for design flood estimation, Hydrology and Water Resources Symposium, Sydney, Australia, 19–22 November 2012, 2012.
- IEAust.: Australian rainfall and runoff: a guide to flood estimation, Rev. Edn. 1987 [i.e. 3rd Edn.], edited by: Pilgrim, D. H., Institution of Engineers, Australia, Barton, ACT, 2001.
- Ilahee, M.: Modelling Losses in Flood Estimation, Ph.D., School of Urban Development Queensland University of Technology, Queensland University of Technology, Queensland, 2005.
- Ilahee, M. and Imteaz, M. A.: Improved Continuing Losses Estimation Using Initial Loss-Continuing Loss Model for Medium Sized Rural Catchments, *Am. J. Eng. Appl. Sci.*, 2, 796–803, doi:10.3844/ajeassp.2009.796.803, 2009.
- Ishak, E. and Rahman, A.: Investigation into probabilistic nature of continuing loss in four catchments in Victoria, 30 Hydrology & Water Resources Symposium: Past, Present & Future, Launceston, 4–7 December 2006, 6–1, 2006.
- Jingyi, Z. and Hall, M.: Regional flood frequency analysis for the Gan-Ming River basin in China, *J. Hydrol.*, 296, 98–117, 2004.
- KISTERS: HYDSTRA, in: KISTERS North America, Inc, available at: <http://www.kisters.com.au/english/html/au/homepage.html> (last access: 2 April 2013), 2008.
- Kuczera, G., Kavetski, D., Franks, S., and Thyer, M.: Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, 331, 161–177, 2006a.
- Kuczera, G., Lambert, M., Heneker, T., Jennings, S., Frost, A., and Coombes, P.: Joint probability and design storms at the crossroads, *Aust. J. Water Resour.*, 10, 63–79, 2006b.
- Kumar, R. and Chatterjee, C.: Regional flood frequency analysis using L-moments for North Brahmputra region of India, *J. Hydrol. Eng.*, 10, 1–7, 2005.
- Lebanon, G.: Bias, Variance, and MSE of Estimators, available at: <http://www.cc.gatech.edu/~lebanon/notes/estimators1.pdf> (last access: 1 June 2013), 2010.
- Ledolter, J. and Hogg, R. V.: *Applied statistics for engineers and physical scientists*, Pearson/Prentice Hall, 2010.
- Liang, Z., Wang, J., Lei, Y., and Shi, Y.: A statistically-based runoff-yield model, in: IAHS-AISH Publication, Proceedings of the Chinese Prediction in Ungauged Basins(PUB)International Symposium 2006, 1 September 2008, Beijing, China, 88–94, 2008.
- Loveridge, M., Sydney, U. o. W., and Rahman, A.: Probabilistic losses for design flood estimation: a case study in New South Wales, Hydrology and Water Resources Symposium, Engineers Australia, 19–22 November 2012, Sydney, Australia, 1449–1456, 2012.
- Lugannani, R. and Rice, S.: Saddle point approximation for the distribution of the sum of independent random variables, *Adv. Appl. Probabil.*, 12, 475–490, 1980.
- Nathan, R. and McMahon, T.: Evaluation of automated techniques for base flow and recession analyses, *Water Resour. Res.*, 26, 1465–1473, 1990.
- Nathan, R., Weinmann, E., and Hill, P.: Use of Monte Carlo Simulation to estimate the expected probability of large to extreme floods, 28th International Hydrology and Water Resources Symposium: About Water; 10–13 November, Wollongong, NSW, Australia, 1.105–1.112, 2003.

- Nathan, R. and Weinmann, P.: An improved framework for the characterisation of extreme floods and for the assessment of dam safety, Proceedings of the British Hydrological Society International Conference, Imperial College, London, July, 186–193, 2004.
- Rahman, A., Weinmann, P. E., and Mein, R. G.: The probabilistic nature of initial losses and its impact on design flood estimation, Hydrology and Water Resources Symposium, Perth, Australia, 20–23 November, Sheraton Perth Hotel, Perth, Western Australia, 71–75, 2000.
- Rahman, A., Weinmann, E., and Mein, R.: The use of probability-distributed initial losses in design flood estimation, Aust. J. Water Resour., 6, 17–29, 2002a.
- Rahman, A., Weinmann, P., Hoang, T., and Laurenson, E.: Monte Carlo simulation of flood frequency curves from rainfall, J. Hydrol., 256, 196–210, 2002b.
- Romeu, J. L.: Anderson-Darling: A goodness of fit test for small samples Assumptions, in: 5, Reliability Analysis Center, NY, 1–6, 2003.
- Saucier, R.: Computer generation of statistical distributions, DTIC Document, Army Research Laboratory, available at: <http://ftp.arl.mil/random/random.pdf> (last access: 2 April 2013), 2000.
- Singh, V. P., Zhang, L., and Rahimi, A.: Probability distribution of rainfall-runoff using entropy theory, Tran. ASABE, 55, 1733–1744, 2012.
- Trefry, C. M., Watkins, D. W., and Johnson, D. L.: Development of regional rainfall intensity-duration-frequency estimates for the state of Michigan, in: Joint Conference on Water Resource Engineering and Water Resources Planning and Management 2000: Building Partnerships, Joint Conference on Water Resource Engineering and Water Resources Planning and Management 2000, 30 July 2000–2 August 2000, Minneapolis, MN, USA, 2004.
- Tularam, G. A. and Ilahee, M.: Initial loss estimates for tropical catchments of Australia, Environ. Impact Assess. Rev., 27, 493–504, doi:10.1016/j.eiar.2006.12.006, 2007.
- Wang, E., Zhang, Y., Luo, J., Chiew, F. H. S., and Wang, Q. J.: Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data, Water Resour. Res., 47, W05516, doi:10.1029/2010wr009922, 2011.
- Waugh, A.: Design losses in flood estimation, International Hydrology and Water Resources Symposium, Perth, 629–630, 1991.