RainIDF: automated derivation of rainfall intensity–duration–frequency relationship from annual maxima and partial duration series

Kian Boon Chang, Sai Hin Lai and Othman Faridah

ABSTRACT

RainIDF, a software tool for derivation of rainfall intensity–duration–frequency (IDF) relationship is developed as an Excel add-in by using Visual Basics for Applications (VBA). The tool is integrated with two of the most widely used statistical distributions for determination of IDF relationship: the generalized extreme value (GEV) distribution for annual maxima series, and the generalized Pareto (GPA) distribution for partial duration series (PDS). It provides automated distribution fitting for rainfall data in the form of annual maxima or PDS for multiple intervals, solving and plotting of rainfall IDF curves. RainIDF uses the Solver add-in function in Excel to solve the coefficients of the empirical IDF formula in one step. The methodology built into RainIDF is discussed and rainfall IDF relationships for several stations in Peninsular Malaysia are derived and compared. RainIDF is available for download on GitHub (http://github.com/kbchang/rainidf) as an Excel add-in.

Key words | annual maxima series, Excel, generalized extreme value distribution, generalized Pareto distribution, partial duration series, rainfall intensity–duration–frequency

Kian Boon Chang
Sai Hin Lai (corresponding author)
Othman Faridah
Department of Civil Engineering,
Faculty of Engineering Building,
University of Malaya,
50603 Kuala Lumpur,
Malaysia
E-mail: *laish@um.edu.my*

INTRODUCTION

The rainfall intensity–duration–frequency (IDF) relationship is an important tool for the determination of design rainfall in water resources structural design, urban stormwater management, flood modeling, etc. In some cases, depth–duration–frequency (DDF) and intensity–duration–area–frequency (IDAF) relationships are used, which serve the same purpose as the IDF relationship. Generally, there are two types of rainfall data series that are widely applied for derivation of IDF relationship: annual maxima series (AMS) and partial duration series (PDS). There are different methods or approaches to derive rainfall IDF relationship (e.g. Koutsoyiannis & Baloutsos 2000; Overeem *et al.* 2008; Palynchuk & Guo 2008; Ben-Zvi 2009; Madsen *et al.* 2009; Van de Vyver & Demarée 2010; De Michele *et al.* 2011), where the selection of good fitting probabilistic distributions for the target region is very important.

RainIDF is developed by using Visual Basics for Applications (VBA) and can be installed as an Excel add-in. Basically, input data in the form of annual maxima or PDS for multiple intervals are inserted into an Excel worksheet and the RainIDF add-in will fit the data with the corresponding probabilistic distribution (generalized extreme value (GEV) or generalized Pareto (GPA) distribution), list out all statistical parameters and optimization procedures to obtain the empirical IDF formula. Besides, it also takes the advantage of Excel's chart plotting functionality to plot the IDF curves automatically based on the derived empirical IDF formula. The output return periods or annual recurrence intervals (ARI) for the IDF relationship derived are 3-month, 6-month, 9-month, 1-year, 2-year, 5-year, 10-year, 20-year, 50-year and 100-year. For other return periods or ARI, it has to be performed manually based on the parameters of the fitted distribution.

doi: 10.2166/hydro.2013.192

EXTRACTION OF ANNUAL MAXIMA AND PARTIAL DURATION SERIES

Before extracting annual maxima or PDS, the data must be filtered to exclude years with a reasonable amount of missing data. Besides, inappropriate data values due to certain malfunction errors of the recording rain gauge or hydrological database software have to be carefully identified. A way to identify this type of invalid data is by comparing the depth and the duration of the rainfall event. For the rainfall data used in this study, we have identified some problematic years where all the rainfall events have the same duration, which are then excluded from analysis.

The extraction of AMS is fairly simple and straightforward. Annual maximum rainfall for a particular duration or interval is obtained by selecting the largest value of rainfall depth for that particular duration in each year. Besides using annual maximum rainfall for derivation of IDF relationship, some researchers have also studied the trends of the annual maximum rainfall (e.g. Adamowski & Bougadis 2003; Kuo *et al.* 2011). Meanwhile, PDS (also known as the peak-over-threshold or POT approach) consists of all the rainfall events that exceed a certain threshold value. The use of PDS is common in flood analysis, until recent studies show an increasing popularity of using PDS in rainfall analysis, especially for derivation of IDF relationship (e.g. Beguería 2005; Palynchuk & Guo 2008; Ben-Zvi 2009).

Before extracting PDS, it is recommended to determine individual events from rainfall data, as this will increase the independence of the extracted data. To identify an individual rainfall event, a minimum inter-event time can be used. If the dry period between two wet periods is equal or more than the minimum inter-event time, they are considered as two separated events. A minimum inter-event time of 6 hours is commonly used (e.g. Guo & Adams 1998; Palynchuk & Guo 2008). A method to select minimum inter-event time based on a coefficient of variation (CV) near 1 is proposed (Restrepo-Posada & Eagleson 1982), but is found to be inappropriate (Powell *et al.* 2007). In this study, a minimum inter-event time of 6 hours is adopted for separation of rainfall events.

The most uncertain parameter when extracting PDS is the threshold value (Beguería 2005). Similar to minimum inter-event time, the threshold value also directly affects the number of rainfall events extracted. Most of the previous researches regarding PDS are applied on flood analysis. As for usage in rainfall analysis, a method of choosing threshold values is based on the result of goodness-of-fit test (Ben-Zvi 2009). Madsen *et al.* (2002) have implied common threshold values in all the studied stations, which result in the range of 2.5–3.2 for regional average number of exceedances per year. Palynchuk & Guo (2008) have chosen a threshold value of 25 mm for their study area in Toronto, Canada. Beguería (2005) has concluded that a unique optimum threshold value cannot be found. In this paper, arbitrary thresholds are first attempted, and then they are adjusted to produce the desired average number of events per year (around three or two to four events per year).

STATISTICAL METHODS

Parameter estimation

There are a few methods for fitting distributions to data, for example: MOM (method of moments), ML (maximum likelihood) and PWM (probability-weighted moments). They are used to estimate the parameters of the distributions. MOM is one of the oldest, simplest and most popular methods of estimating parameters. MOM was originally proposed by Gumbel (1941) to fit the Gumbel distribution. According to Madsen *et al.* (1997), in general, fitting of PDS data with MOM should be used for negative shapes of the distribution fitted (heavy tailed); while one should use AMS with MOM for moderately positive shapes; and ML for large positive shapes (light tailed). Heavy tail and light tail reflects the rate of increase of the physical variable when its exceedance probability declines. Heavy tailed distributions increase faster than the exponential rate, while light tailed distributions are slower. However, Ben-Zvi (2009) has concluded that all good fits of the GPA distribution and most of the good fits of the Gumbel distribution by use of the PWM are found better than those by use of the MOM. PWM has been described as a simple and efficient method for fitting distributions to data (Koutsoyiannis *et al.* 1998).

In RainIDF, the parameter estimation method integrated is known as *L*-moments (Hosking & Wallis 1997). *L*-moments are based on PWM, however *L*-moments provide a higher degree of accuracy and ease of use. As mentioned by Hosking & Wallis (1997), PWM uses weights of the cumulative distribution function (CDF) but it is difficult to interpret the moments as scale and shape parameters for probability distributions. PWM is used by the *L*-moments method to calculate parameters that are easy to interpret and also can be used to calculate parameters of the statistical distribution. Millington *et al.* (2011) found that the method of *L*-moments is easy to work with and more reliable as they are less sensitive to outliers, thus providing an advantage. Rowinski *et al.* (2001) discovered that the MOM

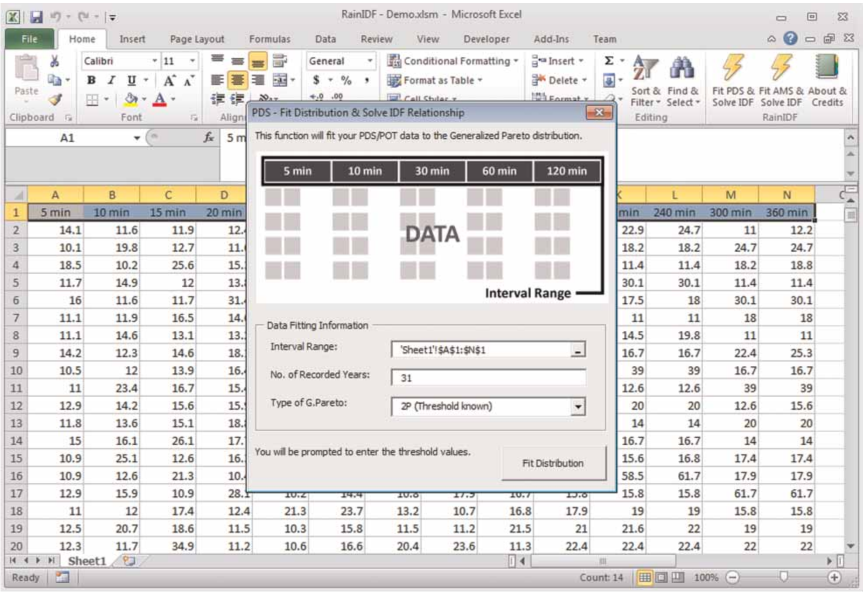

Figure 1 | Interface of RainIDF add-in with partial duration series input parameters form.

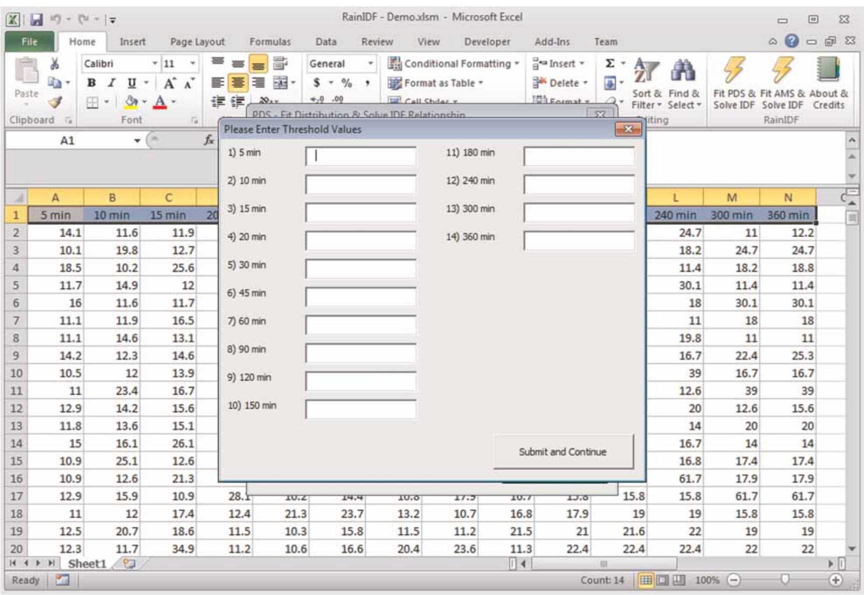

Figure 2 | Input form for entering threshold values for their corresponding interval.

techniques are only able to apply to a limited range of parameters, whereas *L*-moments can be more widely used. Therefore, the method of *L*-moments with PWM as described by Hosking & Wallis (1997) has been chosen to estimate parameters for GEV and GPA distribution in this study.

GEV distribution for annual maxima series

In recent years, more studies (e.g. Koutsoyiannis & Baloutsos 2000; Ben-Zvi 2009; Millington *et al.* 2011) have shown that GEV distribution is more appropriate than other distributions that are commonly used for fitting AMS (e.g. Gumbel and Log-Pearson Type III distributions). These studies have expressed skepticism for the appropriateness of Gumbel distribution (a family member of the GEV distribution) for rainfall extremes, which show that the Gumbel distribution tends to underestimate the largest extreme rainfall amounts. A study by Koutsoyiannis & Baloutsos (2000) shows that with a long record of annual rainfall (i.e. 136 years), the underestimation of Gumbel distribution is quite substantial (e.g. 1:2). Zalina *et al.* (2002) found that GEV distributions are the best fitted distribution for AMS in the Peninsular Malaysia.

Therefore, GEV distribution has been chosen to implement in RainIDF. The CDF and PDF (probability density function) of GEV (Hosking & Wallis 1997) are defined as:

$$F(x) = \exp\left\{-\left(1-\frac{\kappa(x-\xi)}{\alpha}\right)^{\frac{1}{\kappa}}\right\} \tag{1}$$

$$f(x) = a^{-1}\exp[-(1-\kappa)y - \exp(-y)] \tag{2}$$

where,

$$y = -\kappa^{-1}\log\left[1-\frac{\kappa(x-\xi)}{\alpha}\right],\ \text{when}\ \kappa \neq 0 \tag{3}$$

$$y = \frac{x-\xi}{\alpha},\ \text{when}\ \kappa = 0 \tag{4}$$

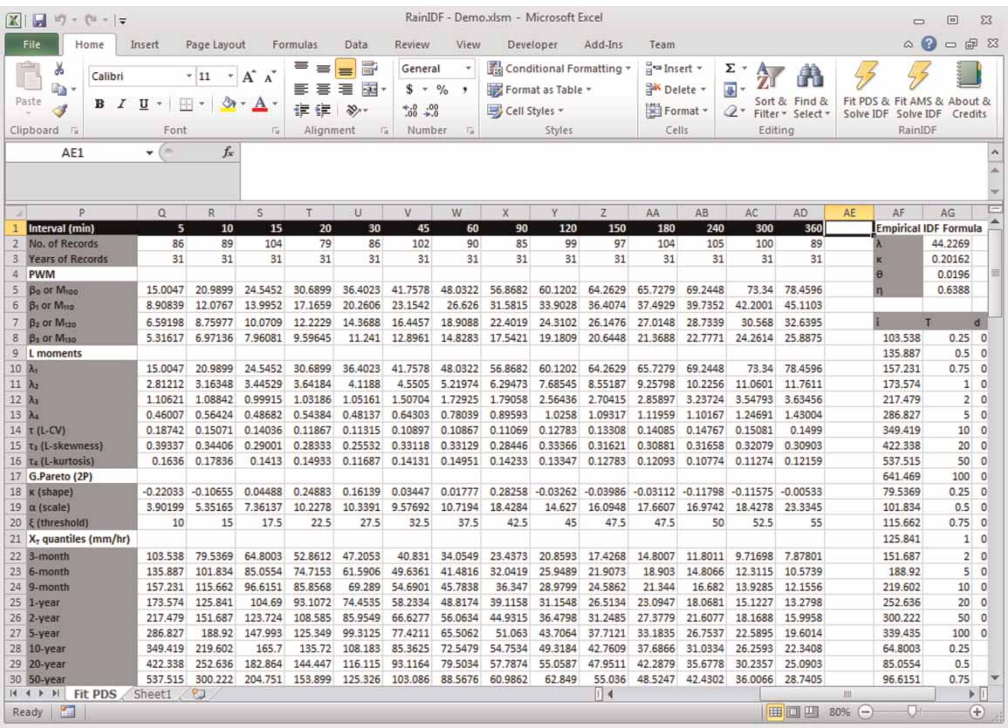

Figure 3 | Spreadsheet contains parameters of the fitted data series and IDF relationship.

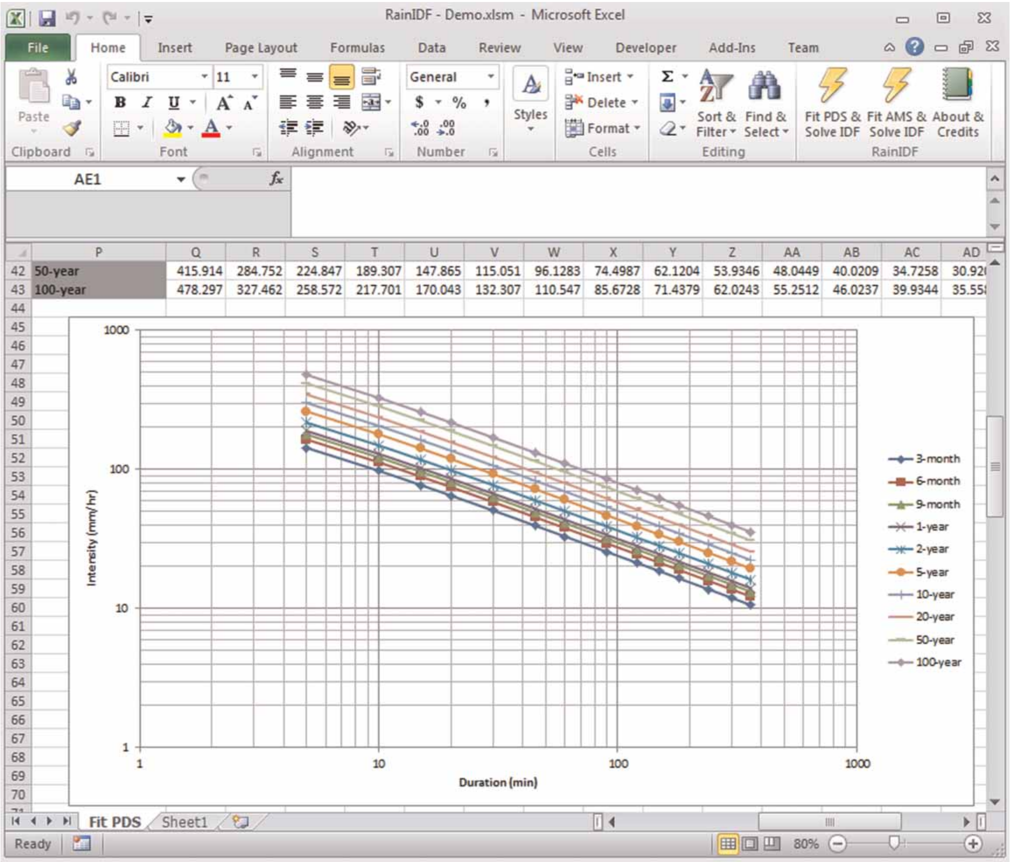

Figure 4 | Rainfall IDF curves plotted automatically with RainIDF Excel add-in.

x is the random variable of interest, ξ is the location parameter, α is the scale parameter and κ is the shape parameter.

The extreme quantile, X_T of the corresponding return period, T and duration from the AMS can be computed by using the inverse CDF of the GEV distribution:

$$X_T = \xi + \frac{\alpha\left\{1 - \left(-\ln\left(1 - \frac{1}{T}\right)\right)^{\kappa}\right\}}{\kappa}, \text{ when } \kappa \neq 0 \tag{5}$$

$$X_T = \xi - \alpha \ln\left(-\ln\frac{1}{T}\right), \text{ when } \kappa = 0 \tag{6}$$

It is worth noting that the GEV distribution turns into Gumbel distribution when the shape parameter, κ, is equal to zero. Gumbel distribution is often chosen for its ease of use, since it only consists of two parameters (without the shape parameter). However, by implementing the automated distribution fitting function in RainIDF, all three parameters of GEV distribution can be estimated easily and thus eliminates the two-parameter advantage of the Gumbel distribution.

GPA distribution for partial duration series

The GPA distribution is one of the most popular distributions used for partial duration or POT analysis (e.g. Beguería 2005; Palynchuk & Guo 2008; Ben-Zvi 2009). The CDF and PDF of GPA distribution as defined by Hosking & Wallis (1997) are:

$$F(x) = 1 - e^{-y} \tag{7}$$

$$f(x) = \alpha^{-1} e^{-1(1-\kappa)y} \tag{8}$$

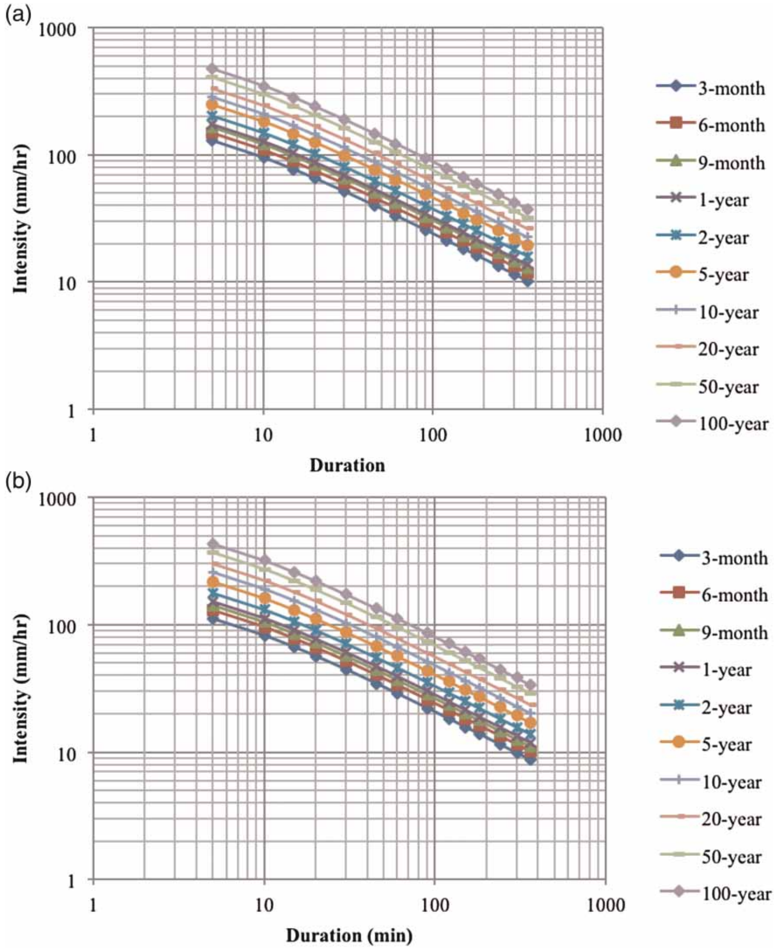

Figure 5 | IDF curves derived from station 2330009 in Johor. (a) Partial duration series. (b) Annual maxima series.

where,

$$y = -\kappa^{-1} \log\left\{1 - \frac{\kappa(x-\xi)}{\alpha}\right\}, \text{ when } \kappa \neq 0 \tag{9}$$

$$y = \frac{(x-\xi)}{\alpha}, \text{ when } \kappa = 0 \tag{10}$$

ξ is the location parameter, α is the scale parameter and κ is the shape parameter. Special cases: $\kappa = 0$ is the exponential distribution with two parameters; $\kappa = 1$ is the uniform distribution on the interval $\xi \leq x \leq \xi + \alpha$.

The location parameter, ξ is actually the threshold of the data series. The threshold value is usually known when fitting PDS to the GPA distribution. In this case, the two parameters (2-P) GPA distribution is used for fitting PDS, where only the scale and shape parameters are estimated with *L*-moments. Given that the average number of events per year λ is known with the corresponding threshold x_0, the quantile of a specific duration with *T*-year return period can be calculated by:

$$X_T = x_0 + \frac{\alpha}{\kappa}\left[1 - \left(\frac{1}{\lambda T}\right)^{\kappa}\right], \text{ when } \kappa \neq 0 \tag{11}$$

$$X_T = x_0 + \alpha \, \text{In}\left(\frac{1}{\lambda T}\right), \text{ when } \kappa = 0 \tag{12}$$

The 2-P GPA distribution has a different formula for parameter estimation with *L*-moments compared to the three parameters (3-P) GPA distribution. Although the 2-P GPA distribution is preferred for fitting PDS, 2-P and 3-P are both

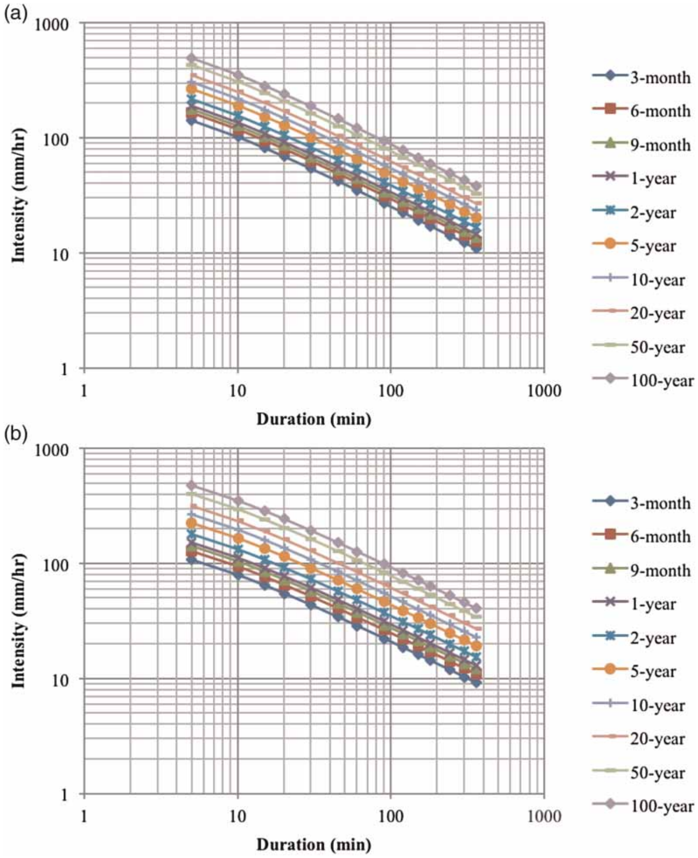

Figure 6 | IDF curves derived from station 3628001 in Pahang. (a) Partial duration series. (b) Annual maxima series.

included in RainIDF. The 2-P GPA distribution requires the user to specify the threshold values for each data series, while the 3-P GPA distribution estimates the location parameter from the data series.

DERIVATION OF IDF RELATIONSHIP WITH RAINIDF

By installing RainIDF add-in on Excel 2007 or 2010 (Windows PC), the rainfall IDF relationship can be computed automatically based on input data series (annual maxima or partial duration). It is straightforward to derive IDF relationship from AMS, while for PDS there are some required parameters such as number of recorded years and threshold value (if the 2-P GPA distribution is chosen). Figure 1 shows the interface of RainIDF in Excel, where the RainIDF menu buttons are located at the top right corner of the home tab. The first step towards generation of IDF relationship is to import annual maxima or PDS into Excel spreadsheet, with the header containing the interval value of the data series in minutes. The input data can be in any format (e.g. .txt (text) files and .csv (comma separated values) files) as long as they can be imported (or copied and pasted) into the Excel spreadsheet. By selecting the header range (see Figure 1), RainIDF can identify and locate all the data series below the header range (up to 30 sets of data series), and obtain the interval information of the data series from the header. In Figure 1, the headers of the PDS are selected for generation of IDF relationship, where the 2-P GPA distribution is selected.

Since the 2-P GPA distribution is chosen, the user will be prompted to enter the corresponding threshold value based on the selected range of headers (Figure 2). If the 3-P GPA distribution is selected, this step is skipped, as the threshold or location parameters will be estimated from the data series by using the *L*-moments method. The user may choose to use a set of common threshold value of all rainfall stations (which

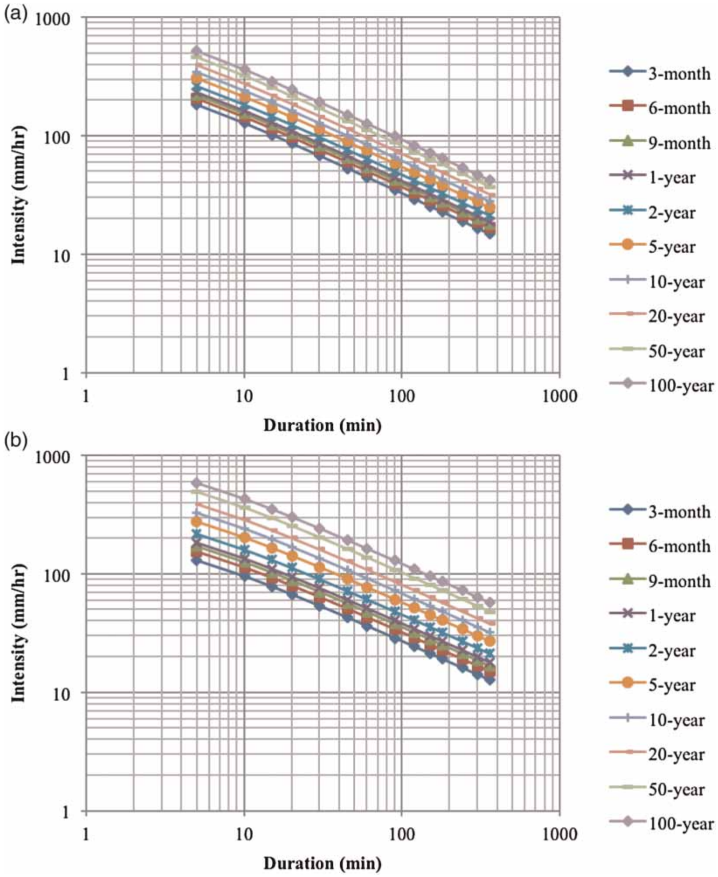

Figure 7 | IDF curves derived from station 6019004 in Kelantan. (a) Partial duration series. (b) Annual maxima series.

produces varied average number of events per year), or arbitrary threshold values that requires adjustment for different stations to produce a desired average number of events per year (e.g. two to four events). After the threshold values are entered, RainIDF will automatically filter data value that is lower (if there is any) than its corresponding threshold value. RainIDF creates a new spreadsheet (Figure 3) containing all the important parameters (such as PWM, *L*-moments, distribution parameters and quantiles), coefficients of empirical IDF formula and IDF curves (Figure 4). The empirical IDF formula used by Bernard (1932) is known as:

$$i = \frac{\lambda T^{\kappa}}{(d+\theta)^{\eta}} \tag{13}$$

where i is the rainfall intensity (mm/hour) of the corresponding d-duration (hour) and T-year return period. All parameters must be positive values and $0 < \eta < 1$.

RainIDF calls the Solver utility function in Excel, to perform the one-step least squares method for solving and optimizing the coefficients of the empirical IDF formula. Other methods and details about the one-step least squares method are discussed in Koutsoyiannis *et al*. (1998). The coefficients of the solved empirical IDF formula are listed in the generated spreadsheet (see top right corner of Figure 3). The optimized quantiles from the empirical IDF relationship are calculated and plotted into IDF curves (Figure 4). It is worth noting that the Solver add-in embedded in Excel must be enabled in order to derive IDF relationship with RainIDF, otherwise a warning message box will appear as RainIDF fails to call the Solver add-in function.

To demonstrate RainIDF, three stations in Malaysia are chosen and their IDF relationships are derived using RainIDF for both partial duration and AMS (Figures 5–7). The data series include durations of 5, 10, 15, 20, 30, 45, 60, 90, 120, 150, 180, 240, 300 and 360 minutes. By

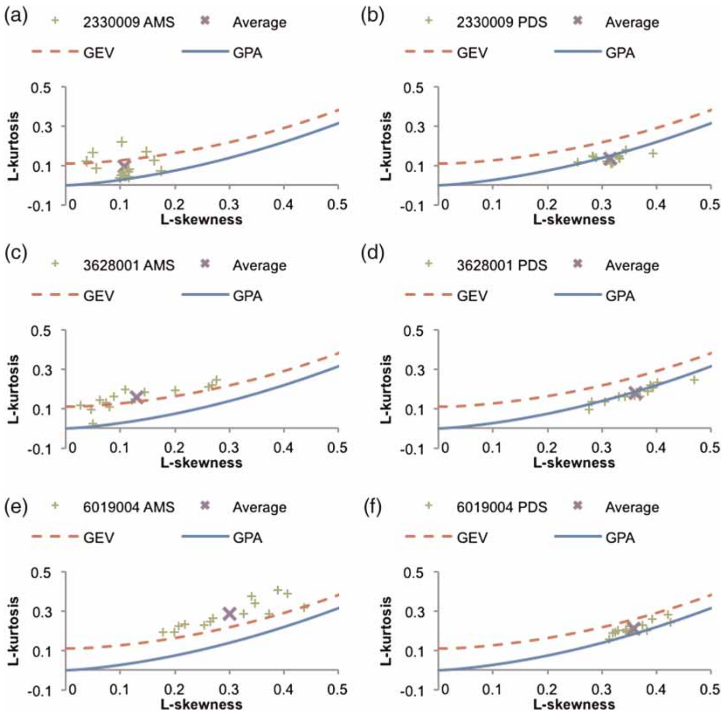

Figure 8 | *L*-moment ratio diagrams for annual maxima series (AMS) and partial duration series (PDS) data obtained from three selected rainfall stations in Peninsular Malaysia. (a) Station 2330009 (AMS), (b) Station 2330009 (PDS), (c) Station 3628001 (AMS), (d) Station 3628001 (PDS), (e) Station 6019004 (AMS), (f) Station 6019004 (PDS).

comparing IDF curves computed from AMS with IDF curves of PDS, it seems that the rainfall intensity for short return period (e.g. 5 years and below) for PDS are slightly higher than AMS throughout all the plotted durations. Note that the PDS used are around three events per year on average, compared to AMS with one event per year. It is easier to derive IDF relationship with AMS and GEV distribution as the extraction of AMS is straightforward while extraction of PDS requires extra steps (e.g. separation of rainfall events and selecting threshold values). Although the choice of minimum inter-event time and threshold values can cause slight differences for PDS results, the most important step in preparation of data is to filter the invalid or problematic data, especially for short duration rainfall such as 5 and 10 minutes.

To choose between AMS–GEV and PDS–GPA approaches, *L*-moment ratio diagram as demonstrated by Hosking & Wallis (1997) can be used to determine the goodfit of the AMS or PDS to their corresponding distributions. *L*-moment ratio diagrams for the three selected stations are plotted in Figure 8. By comparing the fitting of AMS–GEV (Figures 8(a), (c) and (e)) with the fitting of PDS–GPA (Figures 8(b), (d) and (f)), it is observed that PDS–GPA has a better fitting than AMS–GEV, as the PDS–GPA *L*-moments ratios and sample mean are closer to the population of *L*-skewness and *L*-kurtosis of the GPA distribution. In this case, the use of PDS–GPA approach is encouraged for the derivation of IDF relationship for these three stations.

CONCLUSIONS

The importance of rainfall IDF curves as design rainfall references in water resources engineering are increasing especially with the impact of climate change. RainIDF can help to speed up analysis work and thus, the large-scale derivation of the rainfall IDF relationship can be performed with ease (especially for PDS). If other IDF formula is preferred, one may use the distribution parameters and quantiles obtained via RainIDF to apply with the preferred IDF formula. Rainfall data with missing data and invalid values have to be filtered carefully before they are used to extract annual maxima or PDS, as these problematic data will affect the accuracy of the derived IDF curves.

On the other hand, if the desired return period is different from the supported return period, one may also calculate the quantiles of the desired return period manually based on the statistical parameters. Since RainIDF fits GEV distribution to AMS and GPA distribution to PDS, one should study their suitability for the selected regions as their condition of fitting to data series might vary on different regions. Goodness-of-fit tests and *L*-moment ratio diagrams are among the widely used method for testing the goodness-of-fit of distributions. RainIDF is available as an Excel add-in on GitHub (http://github.com/kbchang/rainidf) for research purpose and non-commercial usage.

ACKNOWLEDGEMENTS

The authors would like to thank the editor and reviewers (especially Reviewer 1) for their comments and suggestions, which resulted in the improvement of this paper. Also, the authors would like to thank the Department of Irrigation and Drainage (DID) Malaysia for supplying rainfall data for this research. This study is part of a research project (ER029- 2011A), funded by the Ministry of Higher Education (MOHE), Malaysia.

REFERENCES

Adamowski, K. & Bougadis, J. 2003 Detection of trends in annual extreme rainfall. *Hydrol. Process.* **17** (18), 3547–3560.

Beguería, S. 2005 Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value. *J. Hydrol.* **303** (1–4), 215–230.

Ben-Zvi, A. 2009 Rainfall intensity–duration–frequency relationships derived from large partial duration series. *J. Hydrol.* **367** (1–2), 104–114.

Bernard, M. M. 1932 Formulas for rainfall intensities of long durations. *Trans. ASCE* **96**, 592–624.

De Michele, C., Zenoni, E., Pecora, S. & Rosso, R. 2011 Analytical derivation of rain intensity–duration–area–frequency relationships from event maxima. *J. Hydrol.* **399** (3–4), 385–393.

Gumbel, E. J. 1941 The return period of flood flows. *Ann. Math. Stat.* **12** (2), 163–190.

Guo, Y. & Adams, B. J. 1998 Hydrologic analysis of urban catchments with event-based probabilistic models: 1. Runoff volume. *Water Resour. Res.* **34** (12), 3421–3431.

Hosking, J. R. M. & Wallis, J. R. 1997 *Regional Frequency Analysis: An Approach Based on L-moments*. Cambridge University Press, New York.

Koutsoyiannis, D. & Baloutsos, G. 2000 Analysis of a long record of annual maximum rainfall in Athens, Greece, and design rainfall inferences. *Natural Hazards* **22** (1), 29–48.

Koutsoyiannis, D., Kozonis, D. & Manetas, A. 1998 A mathematical framework for studying rainfall intensity–duration–frequency relationships. *J. Hydrol.* **206** (1–2), 118–135.

Kuo, Y.-M., Chu, H.-J., Pan, T.-Y. & Yu, H.-L. 2011 Investigating common trends of annual maximum rainfalls during heavy rainfall events in southern Taiwan. *J. Hydrol.* **409** (3–4), 749–758.

Madsen, H., Arnbjerg-Nielsen, K. & Mikkelsen, P. S. 2009 Update of regional intensity–duration–frequency curves in Denmark: tendency towards increased storm intensities. *Atmos. Res.* **92** (3), 343–349.

Madsen, H., Mikkelsen, P. S., Rosbjerg, D. & Harremoës, P. 2002 Regional estimation of rainfall intensity–duration–frequency curves using generalized least squares regression of partial duration series statistics. *Water Resour. Res.* **38** (11), 1239.

Madsen, H., Rasmussen, P. F. & Rosbjerg, D. 1997 Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At-site modeling. *Water Resour. Res.* **33** (4), 747–757.

Millington, N., Das, S. & Simonovic, S. P. 2011 The Comparison of GEV, Log-Pearson Type 3 and Gumbel Distributions in the Upper Thames River Watershed under Global Climate Models. Water Resources Research Report, The University of Western Ontario.

Overeem, A., Buishand, A. & Holleman, I. 2008 Rainfall depth–duration–frequency curves and their uncertainties. *J. Hydrol.* **348** (1–2), 124–134.

Palynchuk, B. & Guo, Y. 2008 Threshold analysis of rainstorm depth and duration statistics at Toronto, Canada. *J. Hydrol.* **348** (3–4), 535–545.

Powell, D., Khan, A., Aziz, N. & Raiford, J. 2007 Dimensionless rainfall patterns for South Carolina. *J. Hydrol. Eng.* **12** (1), 130–133.

Restrepo-Posada, P. J. & Eagleson, P. S. 1982 Identification of independent rainstorms. *J. Hydrol.* **55** (1–4), 303–319.

Rowinski, P. M., Strupczewski, W. G. & Singh, V. P. 2001 A note on the applicability of log-Gumbel and log-logistic probability distributions in hydrological analyses: I. Known pdf. *Hydrol. Sci. J.* **47** (1), 107–122.

Van de Vyver, H. & Demarée, G. R. 2010 Construction of Intensity–Duration–Frequency (IDF) curves for precipitation at Lubumbashi, Congo, under the hypothesis of inadequate data. *Hydrol. Sci. J.* **55** (4), 555–564.

Zalina, M. D., Desa, M. N. M., Nguyen, V.-Y.-V. & Kassim, A. H. M. 2002 Selecting a probability distribution for extreme rainfall series in Malaysia. *Water Sci. Technol.* **45** (2), 63–68.

First received 1 November 2012; accepted in revised form 27 January 2013. Available online 1 March 2013