Statistical Modelling of Rainfall Intensity-Frequency-Duration Curves Using Regional Frequency Analysis and Bayesian Hierarchical Modelling

Sylvia Soltyk
Curtin University, Perth, Australia
E-mail: sylvia.soltyk@postgrad.curtin.edu.au

Michael Leonard
Research Associate, University of Adelaide, Adelaide, Australia
E-mail: michael.leonard@adelaide.edu.au

Aloke Phatak
Research Team Leader, CSIRO Computational Informatics, Perth, Australia
E-mail: Aloke.Phatak@csiro.au

Eric Lehmann
Research Scientist, CSIRO Computational Informatics, Perth, Australia
E-mail: Eric.Lehmann@csiro.au

The Intergovernmental Panel on Climate Change (IPCC, 2007) has predicted an increase in extreme rainfall due to climate change, which may also lead to an increase in natural hazards such as flooding. These hazards can result in damage to infrastructure and agriculture, and may even result in injury or loss of life. Consequently, there is a need for accurate analysis and projection of extreme rainfall and its potential impacts. For example, understanding the relationship between rainfall intensity, frequency, and duration is important for the design and safety of infrastructure so that it can withstand extreme rainfall events. This relationship is described graphically by intensity-frequency-duration (IFD) curves. Estimating IFD curves and their associated uncertainty as accurately as possible is critical as it may help reduce the human and economic impacts that result from extreme rainfall events.

In this paper, we examine two methods for modelling extreme rainfall spatially: regional frequency analysis (RFA) and a Bayesian hierarchical model (BHM). We produce IFD estimates from both methods and compare the results. We find that for some locations, the RFA and BHM estimates are similar, and for other locations, they are different. We discuss the importance of uncertainty estimates and demonstrate the flexibility of the BHM for producing such measures of uncertainty.

1. INTRODUCTION

Understanding the relationship between rainfall intensity (how much), frequency (how often), and duration (over what length of time) is important for the design and safety of infrastructure so that it can withstand extreme rainfall events. This relationship can be described graphically by intensity-frequency-duration (IFD) curves. The Intergovernmental Panel on Climate Change (IPCC) has predicted an increase in extreme rainfall in the future for most regions of the world, including Australia (IPCC, 2007). Therefore, estimating IFD curves and their associated uncertainty for both current and future climates is important for helping us to adapt to the potential impacts of climate change, including the social, human and economic impacts that may result from extreme rainfall and flooding.

By definition extreme events are rare, and thus analysis of extreme rainfall is based on small datasets. Moreover, estimates of extreme rainfall are often required at locations with no direct observations from rainfall gauges. However, since rainfall is a spatial process it is possible to use various statistical techniques to 'borrow strength' from neighbouring stations (rainfall gauges) to increase the accuracy and precision of estimates. We explore the use of two such methods: regional frequency analysis (RFA) and Bayesian hierarchical modelling (BHM). RFA is widely used in the hydrological literature and by meteorological organisations such as the Australian Government Bureau of Meteorology (BoM) and United States of America National Weather Service (see NOAA Atlas 14, 2011 and Green et al.,

2012). However, there is a growing body of work that uses spatial statistical models within a Bayesian hierarchical modelling framework for modelling rainfall and weather extremes; such models incorporate a spatial process that allows us to borrow strength from neighbouring locations, and uncertainty estimates arise naturally from the Bayesian framework (see for example Coles and Casson, 1998, Cooley et al., 2007, and Davison et al., 2012). On the other hand, most regional frequency analyses are multistep procedures, and there is currently no coherent method for obtaining estimates of the uncertainty that arises from all of the steps within the procedure, although resampling methods are often used to obtain partial estimates of uncertainty (NOAA Atlas 14, 2011).

Our objective in this paper therefore is to introduce spatial Bayesian modeling to the hydrological community; to point out the differences and similarities between RFA and BHM; and to illustrate them and compare their results using a common dataset. The analyses described here are meant to be illustrative, not exhaustive; nevertheless, they are plausible analyses using the two methods. In Section 2 we describe the methodology of RFA and BHM we use to model rainfall extremes. The data used in our analysis consists of pluviometer data from 242 rainfall gauges in and around the Sydney region of Australia and is described in Section 3. In Section 3 we also present the results from our analyses, and show IFD curves for both gauged and ungauged locations using both methods. Finally, in Section 4 we draw some general conclusions about the methods from the results that have been obtained, point out some limitations and then highlight directions for future research.

2. METHODOLOGY

2.1. Generalized extreme value theory

The generalized extreme value (GEV) distribution is often used to model rainfall extremes, in particular, annual maxima (Coles, 2001). For a fixed duration d, the distribution of annual maximum rainfall Y can be described by a GEV distribution with cumulative distribution function (CDF)

$$F(y;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \quad \xi \neq 0, \tag{1}$$

where the location parameter $\mu \in (-\infty,+\infty)$, scale parameter $\sigma > 0$, shape parameter $\xi \in (-\infty,+\infty)$, and $1+\xi(y-\mu)/\sigma > 0$. When $\xi = 0$, the GEV distribution in equation (1) reduces to the Gumbel distribution.

Koutsoyiannis et al. (1998) define $\tilde{\mu} = \mu/\sigma$, and show that both $\tilde{\mu}$ and ξ are approximately constant over different durations. The scale parameter σ, however, is duration dependent, and Koutsoyiannis et al. (1998) use the following relationships to model this dependence:

$$\text{(a)} \quad \sigma_d = \frac{\sigma}{(d+\theta)^{\eta}}, \qquad \text{(b)} \quad \sigma_d = \frac{\sigma \cdot d}{(d+\theta)^{\eta}} \ . \tag{2}$$

When rainfall intensity (depth/duration) is modelled, equation (2a) is used; when rainfall depths are used, σ_d is modelled via equation (2b). Modelling the scale parameter in this way allows us to model rainfall maxima across different durations using a single GEV distribution at the cost of only two additional parameters, θ and η. Once estimation of all the parameters has been carried out, we can calculate the return level or intensity for a given return period, N, using the quantile function obtained from the inverse of equation (1).

2.2. Regional Frequency Analysis

RFA combines data from stations into regions with similar characteristics, such as geographical location, in order to increase the sample size and hence the accuracy and precision of extreme rainfall estimates. These regions contain a set of stations whose frequency distributions are approximately the same (after the data have been scaled). To combine the data into suitably defined regions we investigated four methods: (i) fixed regions based on a clustering algorithm, (ii) a region of influence

with the nearest (Euclidean distance) 10 stations, (iii) a region of influence with all of the stations in the nearest 50km, and (iv) a region of influence with at least 120 years of data. After exploratory analysis of the four methods (Q-Q plots), method (ii) was deemed the most suitable and was used for the analysis. Thus, for each station, the data from nearest 10 stations were used to compute the parameter estimates of the chosen regional frequency distribution via the L-moments algorithm proposed by Hosking and Wallis (1997). The regional frequency distribution used to model the data is the GEV distribution of equation (1). It describes the distribution of the data at each station after the data is scaled by the at-site scaling factor. This at-site scaling factor is also known as the 'index flood' of Dalrymple (1960). The index flood is assumed to be the first sample L-moment at each station $\ell_1^{(s)}$, and the observed data $Y_{s,t}$ for station $s=1,\ldots,S$, and year $t=1,\ldots,T_s,$ are scaled by the index flood. For a given region containing a number of stations, and with each of these stations having n_s observations, the rescaled data are used to estimate the sample L-moment ratios at each station, s: $t^{(s)}=\ell_1/\ell_2$, $t_3^{(s)}=\ell_3/\ell_2$, and $t_4^{(s)}=\ell_4/\ell_2$, where ℓ_1, ℓ_2, ℓ_3, and ℓ_4 are the first four sample L-moments. The sample L-moment ratios are calculated based on the GEV distribution of equation (1). Note that RFA does not combine durations using the relationship in equation (2); instead, each duration is modelled separately. The sample L-moment ratios are used to calculate the regional average L-moment ratios t^R, t_3^R, and t_4^R using the expression

$$t_r^R=\frac{\sum_{s=1}^{N} n_s t_r^{(s)}}{\sum_{s=1}^{N} n_s}, r=3,4. \tag{3}$$

Since the index flood is assumed to be the mean of the frequency distribution at each site, the mean of the rescaled data for each site is 1. Thus, $t^R=1$. The distribution is fitted by equating the L-moment ratios to the regional average L-moment ratios (similar to that of the method of moments). Thus, the regional frequency distribution parameters are defined by these regional L-moments.

The three parameters of the fitted GEV distributions are interpolated over the domain of the stations using a thin-plate spline. The spline uses radial basis functions for the covariance and a smoothing parameter that is determined by generalized cross-validation. Latitude and longitude were used as covariates. Other covariates such as elevation were not considered. Once the parameters have been interpolated over the entire study area, the at-site quantiles can be estimated directly from the specified parameters at the location of interest.

There is currently no methodology for quantifying uncertainty at all stages of the RFA process. For example, the bootstrapping procedure outlined in Hosking and Wallis (1997) only covers the variability which arises due to the regional growth curves, and not the variability of the index flood, the variability between regions or the variability due to spatial interpolation. The NOAA has incorporated a measure of uncertainty for the parameter estimates but this does not include the uncertainty associated with the spatial interpolation (NOAA Atlas 14, 2011) and therefore likely underestimates the true uncertainty.

2.3. Bayesian Hierarchical Model

A BHM consists of three submodels, or 'levels' arranged in a hierarchy: the data model, the process model, and the parameter model. Each level contains a conditional probability distribution that describes the data, the underlying 'process', and the parameters of the process. For our model of extreme rainfall, we describe the three levels below, using the notation of Lehmann et al. (2013).

Data model

The first level of the hierarchy models the annual maximum rainfall $Y_{s,t,d}$ at station $s=1,\ldots,S$, year $t=1,\ldots,T_s,$ and for duration $d=1,\ldots,D,$ via the GEV distribution of equation (1) with the duration dependent relationship in equation (2). In our BHM, the annual maximum rainfall observations $Y_{s,t,d}$

are assumed to be independent, conditional on the GEV parameters. The data likelihood can be written as:

$$p(Y \mid \tilde{\mu}, \sigma, \xi, \theta, \eta) = \prod_{s=1}^{S} \prod_{t=1}^{T_s} \prod_{d=1}^{D} \mathrm{GEV}\left(y_{s,t,d} \mid \tilde{\mu}_s, \sigma_s, \xi_s, \theta_s, \eta_s\right) \tag{4}$$

where Y represents all rainfall maxima, the vectors $\tilde{\mu}, \sigma, \xi, \theta$ and η contain the GEV parameters at each station, and $y_{s,t,d}$ denotes the annual maximum at station s for year t and duration d. The scale parameter $\sigma_{s,d}$ is modelled using the relationship in equation (2a) or (2b) depending on whether rainfall intensities or depths are considered. To simplify the exposition below, the station subscript s will be dropped in what follows when referring to the GEV parameters, and the variable $\phi \in \{\tilde{\mu}, \sigma, \xi, \theta, \eta\}$ will be used to denote any of the GEV parameters.

Process model

In the process model, we assume that the GEV parameters vary smoothly over space. Hence, each of the parameters is modelled as the sum of a large scale trend, which might be due to covariates such as elevation, latitude, and longitude, and smaller scale variability that is spatially correlated. Thus, we write:

$$f(\phi) = X^T \beta + P(\alpha_\phi, \lambda_\phi) \tag{5}$$

where X is a vector of covariates, β is the vector of associated coefficients, and $P(\cdot)$ represents a spatially correlated, zero-mean Gaussian process with an exponential covariance function. The Gaussian process is used to model the local spatial smoothness of the GEV parameters over the spatial domain (Davison *et al.*, 2012). The exponential covariance function with sill α_χ and range λ_χ is defined as $\alpha_\chi \cdot \exp\left(-h / \lambda_\chi\right)$, where h is the distance between two stations. To ensure the GEV parameters remain within their respective ranges of values, we impose the following transformations $f(\cdot)$: $\log(\sigma)$, $\log(\theta)$, and $\mathrm{logit}(\eta)$.

Prior parameters model

The final level in the hierarchy requires the specification of prior distributions of the parameters β_ϕ, α_ϕ and λ_ϕ.Similar to Davison et al. (2012) we use the inverse Gamma, Gamma and multivariate normal prior distributions:

$$\alpha_\phi \sim \mathrm{InvGamma}\left(\kappa_{\alpha_\phi}, \gamma_{\alpha_\phi}\right), \quad \lambda_\phi \sim \mathrm{Gamma}\left(\kappa_{\lambda_\phi}, \gamma_{\lambda_\phi}\right), \quad \beta_\phi \sim \mathrm{MVN}\left(\mu_{\beta_\phi}, \Sigma_{\beta_\phi}\right), \tag{8}$$

where κ and γ are the shape and scale hyper-parameters of the respective distributions.

Bayesian inference

The full conditional distributions for the model variables can be derived from the posterior density. They are then used for inference on the model parameters through Markov chain Monte Carlo (MCMC) simulation. Gibbs sampling is used for the α_ϕ and β_ϕ parameters, since conjugate priors were chosen. Metropolis–Hastings (MH) steps are necessary to sample the GEV parameters and range parameters, λ_ϕ, because there are no conjugate priors for these parameters.

The BHM results in Section 3 were obtained from MCMC chains simulated over 150,000 iterations, with a burn-in of 20,000 and a thinning factor of 35. Several diagnostic plots were used to assess the convergence of the chains.

Once the parameters have been estimated through MCMC as described above, the quantile function is used to calculate the return intensities at the $(1-1/N)^{th}$ quantile. Credible intervals, which provide an estimate of the uncertainty, are calculated based on the lowest 2.5% and highest 97.5% of the MCMC chains (after burn-in and thinning).

2.4. Summary of RFA and BHM

RFA and the Bayesian spatial model we have outlined here have similar aims, but also some very important differences. Both methods make the very important assumption of *conditional independence*, that is, given the at-site parameters, extreme rainfall at adjacent sites is independent. In addition, they impose smoothness on the parameters, albeit in different ways, and also pool information from neighbouring sites to obtain more precise estimates of the GEV parameters, especially the shape parameter. In contrast to RFA, which is a multi-step procedure, the BHM is a more coherent approach that allows us to combine information from different durations and from which uncertainty estimates arise naturally from the estimation framework. The following table summarises some of the differences between RFA and BHM.

Table 1. RFA and BHM

	RFA	**BHM**
Parameter estimation	L-moments	Bayesian framework using Markov Chain Monte Carlo
Spatial extrapolation and smoothing	Dividing into regions**;** Spline smoothing	Gaussian process with an exponential covariance function
Intensity-duration relationship	None: modelled duration-by-duration	Durations are combined into one model
Uncertainty estimation	None	Uncertainty estimates (credible intervals)

3. RESULTS & DISCUSSION

3.1. Data

The rainfall maxima were extracted from pluviometer records at 242 stations located around the Sydney and Wollongong areas in New South Wales, Australia, as shown in Figure 1. The area is approximately 160 km by 340 km. The stations record rainfall depths (mm) at 5 minute intervals. Each station can have differing record lengths, which range from 7 to 41 years over the period 1959 – 2002. The pluviometer data recorded at the 5 minute intervals were then accumulated over 12 different durations: 5, 10, 15 and 30 minutes, and 1, 2, 3, 6, 12, 24, 48 and 72 hours. For each station, these 12 duration time series were used to determine the annual maxima for the corresponding year and duration. This resulted in a dataset containing 3527 years of rainfall maxima across the 242 stations. We noted that five stations had excessive 5-minute totals during certain years, and these anomalous years were removed from the respective records.

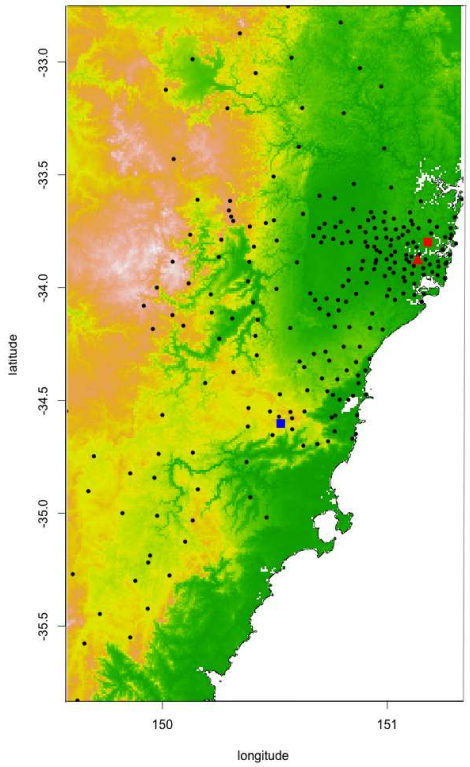

Figure 1 Spatial location of the 242 pluviometer stations. The red square represents gauged location 1, the red triangle represents gauged location 2, and the blue square represents the ungauged location.

3.2. Intensity-frequency-duration (IFD) curves

Figures 2–5 present intensity-frequency-duration (IFD) curves obtained using the BHM and RFA: the blue

line corresponds to the posterior mean rainfall intensity for a 50-year return period from the BHM with the credible intervals (uncertainty) in the lighter blue, and the black dots represent the rainfall intensity from RFA for a 50-year return period. The red lines (BHM) and green dots (RFA) show the intensity-duration relationship for a 100-year return period. The horizontal axis represents the duration and the vertical axis the rainfall intensity in mm/h. The plots are displayed on a log-log scale.

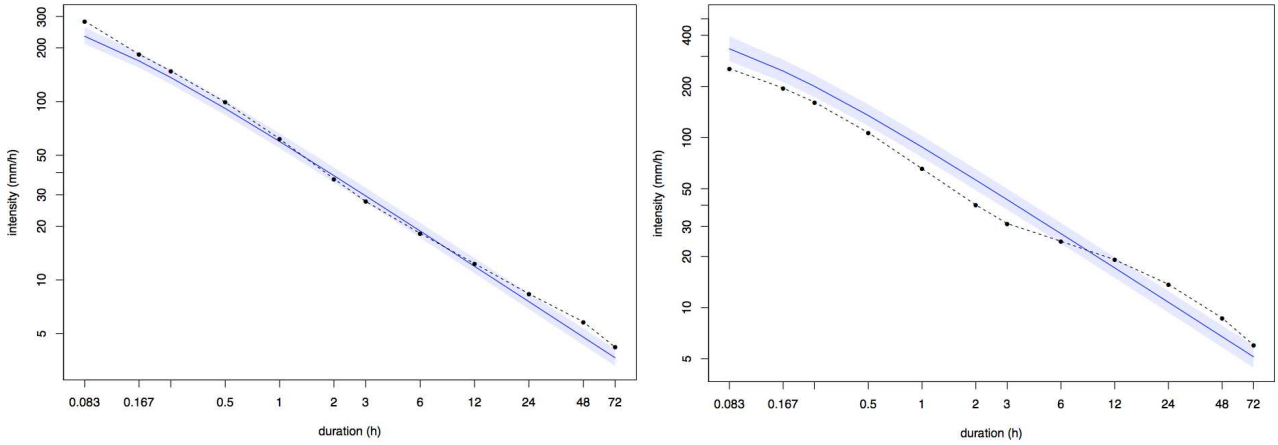

Figure 2a and 2b 50-year IFD curve for gauged location 1 and gauged location 2, respectively

Figure (2a) shows an IFD curve corresponding to a 50-year return period for a gauged location with 35 annual maximum observations (represented by the red square in Figure 1). For this gauged location, the rainfall intensity estimates from RFA and the BHM are similar. For example, for the 24-hour duration, the BHM estimates 7.55mm/hour and RFA estimates 8.36mm/hour of rainfall. Figure (2b) shows an IFD curve corresponding to a 50-year return period for a gauged location with 7 annual maximum observations (represented by the red triangle in Figure 1). For this gauged location, the rainfall intensity estimates from RFA and the BHM are different. For example, for the 24-hour duration, the BHM estimates 10.75mm/hour and RFA estimates 13.72mm/hour of rainfall, which lies outside the 95% credibility interval of the BHM estimate.

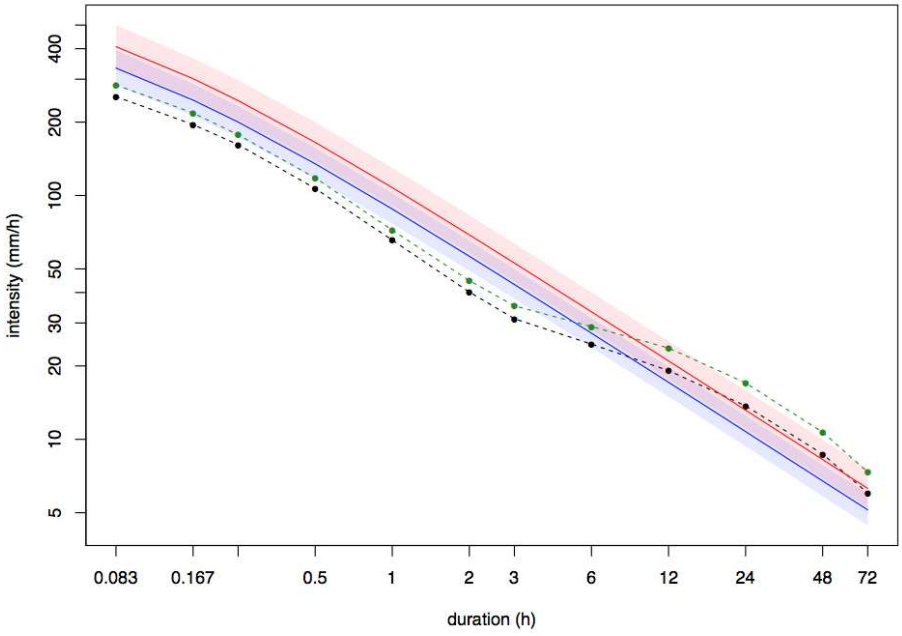

Figure 3 50-year and 100-year IFD curve for gauged location 2

Figure (3) shows the IFD curve for the same gauged location as in Figure (2b), but with the 100-year return period in red for the BHM and green for RFA. The change in intensity between the return periods with respect to duration is different for RFA compared to the BHM, which is a result of the different fitting

techniques. The estimates for the 50- and 100-year return periods are similar, but with the BHM we can assess the difference with respect to the uncertainty in the individual estimates. Furthermore, the RFA estimates at this location are below the BHM interval at shorter durations and above at longer durations. Further work is required to understand the sources of uncertainty between the two methods and why they arrive at different estimates for this gauged location.

Figure (4) shows an IFD curve corresponding to a 50-year return period for an ungauged location (represented by a blue square in Figure 1). The intensity estimates from RFA and the BHM are similar for most of the durations. It is expected that the uncertainty at ungauged locations would be greater (wider credible intervals) when compared to a gauged location since it is necessary to extrapolate to a location where there is no data available. The results of the BHM demonstrate that the framework is consistent with this expectation and is apparent when comparing the IFDs in Figures (2a) and (2b), which are both gauged, to those in Figure (4).

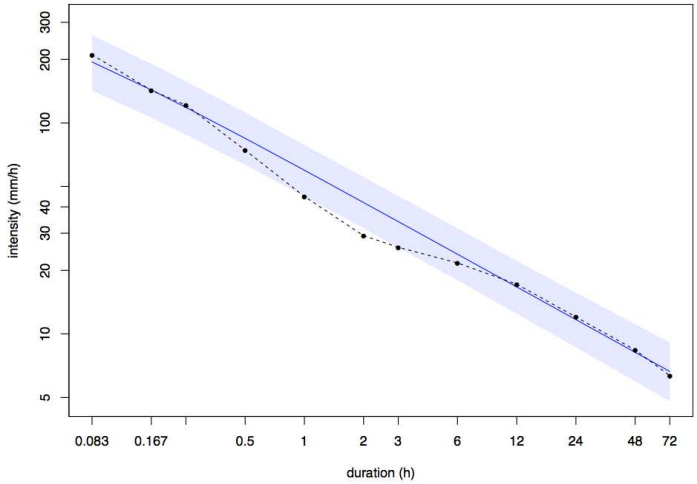

Figure 4 50-year IFD curve for the ungauged location

Figure (5) shows the same gauged location as Figure (2b) with a return period of 50 years. The magenta points represent the intensity estimates using the Australian Government Bureau of Meteorology's (BoM) 2013 IFD tool (http://www.bom.gov.au/water/designRainfalls/revised-ifd/). The difference between the BoM's curve and both RFA and the BHM is likely due to a difference in methodology and the fact that the BoM has used more pluvio data and has also incorporated daily data in the IFD estimate (BoM "New IFDs: Rainfall Data System", 2013).

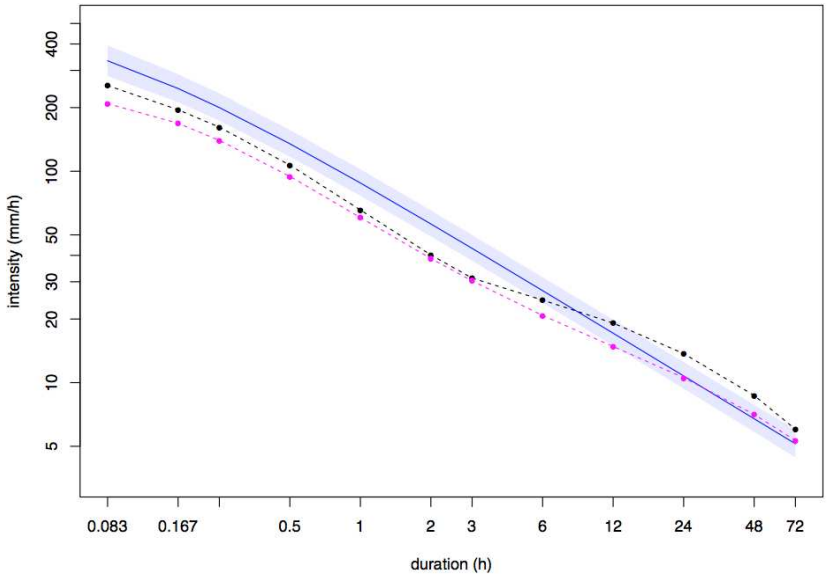

Figure 5 50-year IFD curve for gauged station 2

4. SUMMARY

In this paper, we presented the use of RFA and BHM for estimating extreme rainfall and producing IFD curves. We show that for some locations, both methods produce similar results, and for others there are different results. We also demonstrate the flexibility and coherence of the BHM in producing uncertainty estimates for the IFD curves. The BHM shows IFD relationships which have a gradually varying slope when plotted against duration, whereas the RFA curves have very different shapes. This is a consequence of combining different durations into a smooth relationship in the BHM and fitting durations separately in RFA.

In both RFA and BHM we incorporate simplifications such as the assumption of stationarity over time, and the assumption of conditional independence. Further work on the BHM will include the incorporating time-varying parameters, the addition of climate drivers via covariates in the process model, dependence between durations, and allowing for anisotropy. Further comparison of the two approaches will facilitate greater understanding of their respective strengths and weaknesses and a better appreciation of uncertainty in IFD estimation.

5. ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the Australian Government through Geoscience Australia, and the substantial in-kind support provided by the members of Engineers Australia. We would also like to thank Rex Lau and Joanne Chia of CSIRO Computational Informatics for their input and expertise. Special thanks to Mark Palmer, and to Sydney Water for the provision of pluviometer data.

6. REFERENCES

Australian Government Bureau of Meteorology (BoM). *New IFDs: Rainfall Data System,* viewed 30 October 2013, < http://www.bom.gov.au/water/designRainfalls/revised-ifd/>.

Coles, S. (2001*). An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer, London.

Coles, S. and Casson. (1998). *Extreme value modelling of hurricane wind speeds*. Structural Safety 20, 283–296.

Cooley, D., Nychka, D., and Naveau, P. (2007). *Bayesian spatial modeling of extreme precipitation return levels.* Journal of the American Statistical Association, 102, 824–840.

Dalrymple, T. (1960). *Flood frequency analyses.* Water Supply Paper 1543-A, U.S. Geological Survey, Reston, Virginia.

Davison, A.C., Padoan, S.A. and Ribatet, M. (2012). *Statistical Modeling of Spatial Extremes.* Statistical Science, 27(2), 161-186.

Green, J., Xuereb, K., Johnson, F., Moore, G. and The, C. (2012*). Implications of the Revised Intensity-Frequency-Duration (IFD) Design Rainfall Estimates.* Hydrology Water Resources Symposium, Perth, 24 – 27 February 2014.

Hoksing, J.R.M. and Wallis, J.R. (1997). *Regional Frequency Analysis: An approach based on L-moments.* Cambridge University Press, New York.

Intergovernmental Panel on Climate Change (IPCC) (2007). *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,* M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson, Eds., Cambridge University Press, Cambridge, UK, 976pp.

Koutsoyannis, D., D. Kozonis, and A. Manetas (1998). *A mathematical framework for studying rainfall intensity-duration-frequency relationships*. Journal of Hydrology, 206, 118–135.

Lehmann, E.A., Phatak, A., Soltyk, S., Chia, J., Lau, R. and Palmer, M. (2012). *Bayesian hierarchical modelling of rainfall extremes.* MODSIM: 20th International Congress on Modelling and Simulation, Adelaide, 1 – 6 December 2013.

NOAA: *Precipitation-Frequency Atlas of the United States*. (2011). NOAA Atlas 14, Volume 1, Version 5.0, Bonnin, G.M., Martin, D., Lin, B., Parzybok, T., Yekta, M. and Riley, D., NOAA, National Weather Service, Silver Spring, Maryland.