International Conference on Hydroinformatics

8-1-2014

Data Mining For Design Flood Prediction

James E. Ball

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

DATA MINING FOR DESIGN FLOOD PREDICTION

JAMES E BALL (1)
(1): School of Civil and Environmental Engineering, University of Technology Sydney, PO Box 123, Broadway, NSW, 2007, Australia

Design flood estimation remains a problem for many professionals involved in the management of rural and urban catchments. Advice is required regarding design flood characteristics for many design problems including the design of culverts and bridges necessary for cross drainage of transport routes, the design of urban drainage systems, the design of flood mitigation levees and other flood mitigation structures, design of dam spillways, and many environmental flow problems. When a risk based approach is adopted as the design paradigm, there is a need to predict both the magnitude of the hazard and the exceedance probability of the hazard. In this context, prediction of design flood characteristics becomes a problem in hydroinformatics and particularly in the assessment of patterns in the available data. Presented herein will be a discussion of data and hydroinformatics, and the relationship of both data and hydroinformatics to the design flood problem.

INTRODUCTION

Design flood estimation remains a problem for many professionals involved in the management of rural and urban catchments. Advice is required regarding design flood characteristics for many design problems including the design of culverts and bridges necessary for cross drainage of transport routes, the design of urban drainage systems, the design of flood mitigation levees and other flood mitigation structures, design of dam spillways, and many environmental flow problems.

While the flood characteristic of most importance depends on the nature of the problem under consideration, but typically it is one of the following:

- Flood flow rate – typically, it is the peak flow rate of the flood hydrograph that is the desired design flood hydrograph characteristic;
- Flood level – similar to the flood flow rate, it is the peak flood level during the flood hydrograph that is the commonly desired design flood hydrograph characteristic;
- Flood rate of rise – this design flood characteristic is a concern when planning is undertaken for operational floods;
- Flood volume – this design flood characteristic becomes a concern when the design flood volume is a major factor in the design problem. This situation occurs when storage of a significant portion of a flood hydrograph is used as part of a flood management system; or
- System failure – the usual design flood problem is located at a single point. There are a number of design problems, however, where the issue becomes one of multiple points within a system. Typical examples of these problems include urban drainage systems

where the individual components of the system are not statistically independent which is a common assumption, and transportation routes with multiple cross drainage structures of one or more river systems.

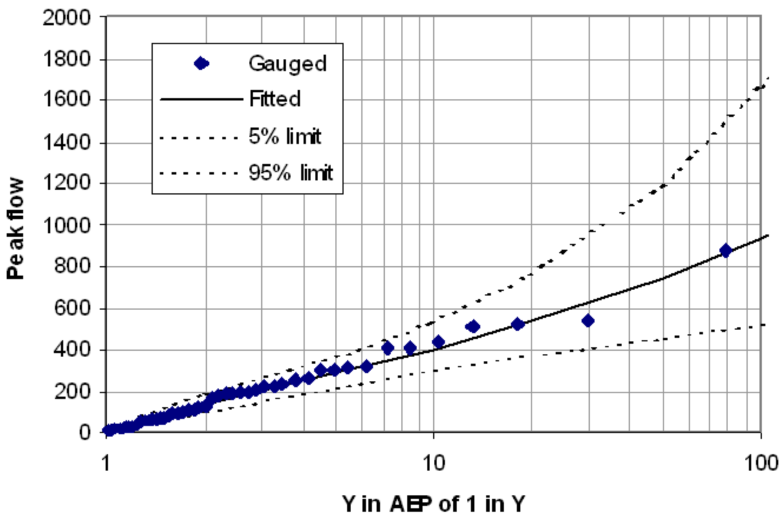

Figure 1. Probability Plot for Styx River at Jeogla

An important aspect of estimating design flood characteristics is the need to consider both the hazard arising from the magnitude of the design flood characteristic and the likelihood of that hazard occurring or being exceeded. In other words there is a need to consider the relationship between the magnitude and the exceedance probability of a design flood characteristic. An example of this relationship is shown in Figure 1. From a design flood perspective, therefore, it is the analysis of data that is fundamental to the estimation of the quantile for the desired flood characteristic.

There are many approaches for the estimation of design flood characteristics with the appropriateness of these approaches dependent on the problem. Most of these approaches can be framed as an exercise in data mining since the available data is analysed to develop the desired relationship between the hazard arising from the magnitude of the flood characteristic and the likelihood of that hazard occurring.

Within Australia, Australian Rainfall and Runoff (Pilgrim [12]) is the primary document providing guidance on suitable approaches for estimation of design flood characteristics. While the 1987 edition of Australian Rainfall and Runoff and its predecessors have served the engineering profession and the general community well, in the period since release of this edition, a number of developments have arisen that necessitate the production of a new edition. These developments include the many recent advances in knowledge regarding flood processes, the increased computational capacity available to the profession, and the rapidly expanding body of information about climate change and the influence of potential future climate states on the relationship between flood magnitude and flood frequency. Consideration of these developments resulted in the decision by Engineers Australia to prepare another edition of Australian Rainfall and Runoff.

An outcome of this revision has been recognition of changes in approaches to design flood estimation and the relevance to the problem of design flood estimation of the development of the concept of hydroinformatics. This has led to recasting the design flood prediction problem as a problem in hydroinformatics. The purpose of this paper is to provide an overview of the concept of design flood estimation being a problem in hydroinformatics.

APPROACHES FOR ESTIMATION OF DESIGN FLOOD QUANTILES

Background

As discussed earlier, the design flood problem is one of estimating the relationship between the hazard arising from the magnitude of the design flood characteristic and the likelihood of that hazard occurring or being exceeded. Hence there is a need to consider the design flood problem from a statistical viewpoint. This contrasts with the analysis of the flood characteristics of a historical event where a deterministic viewpoint is appropriate.

There are two alternative situations when design flood characteristics are required; these are:

- Suitable historic data is available; and
- Suitable historic data is not available.

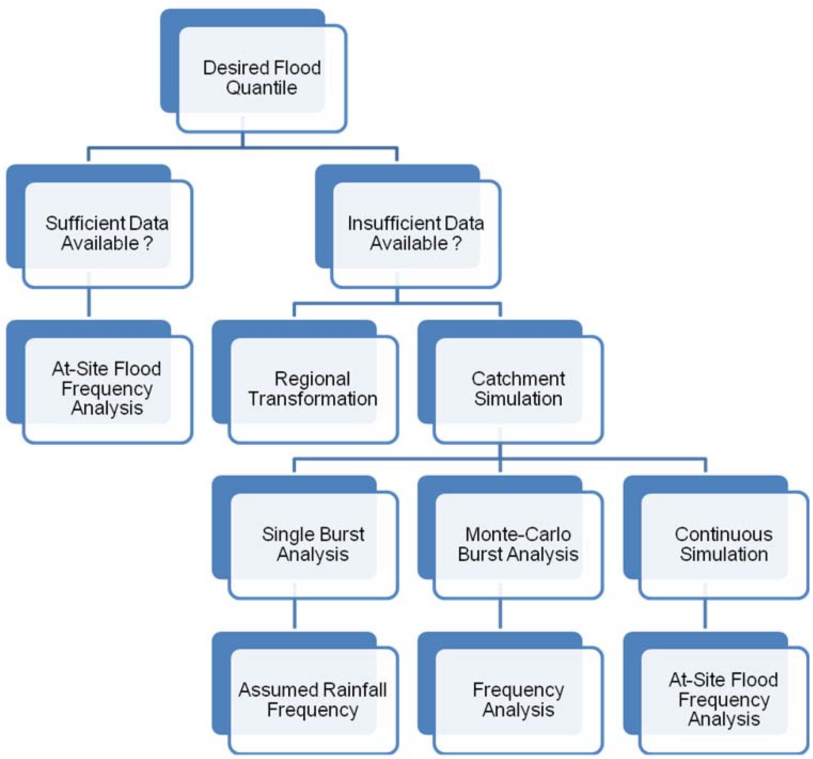

Figure 2. Design Flood Estimation Approaches (after Ball *et al.* [1])

Flood Frequency Analysis

As shown in Figure 2, where sufficient historical information is available, estimation of the desired flood parameter can be undertaken using at-site flood frequency analysis (FFA) methods (see, for example, Jin and Stedinger [7], and Kuczera [9]). In this case, the desired design flood quantile is obtained from an analysis of the statistical pattern of the recorded data. This is a classical application of a hydroinformatic technique, namely that of data mining the available historical flows to enable recognition of the statistical pattern associated with the desired flood characteristic.

It is worthwhile noting that FFA techniques are based on the calibration of a statistical model to the recorded data. As part of this calibration process, the historical data will be assessed for its reliability. This data assessment process represents a different type of hydroinformatic problem, namely that of data management.

Regional Flood Methods

Where adequate historical information is not available, as shown in Figure 2, it is necessary to use either catchment simulation techniques (sometimes referred to as Rainfall Based Techniques) or regional transformation techniques (sometimes referred to as Regional Flood Methods) to generate information (usually the design quantile) about the flood characteristic. Similar to FFA methods, implementation of these methods requires hydroinformatic techniques either in their application or in their development.

For example, the Regional Transformation Techniques developed by Mittelstadt *et al*. [11] and Haddad and Rahman [5] require the designer to apply a regression relationship between catchment characteristics and design flood flow quantiles. Development of these regression relationships requires data mining of both the available historical flows to enable recognition of the statistical pattern associated with the desired flood characteristic and the available catchment characteristics. Hence, use of a Regional Flood Method requires the application of hydroinformatic techniques to develop the method for general application.

Catchment Simulation

The third approach shown in Figure 2 for estimation of a desired design flood characteristic quantile is the use of catchment simulation. Conceptually, the aim of catchment simulation is to generate data that would have been recorded if catchment monitoring had been in place for the event, or sequence of events, being considered. It should be noted that situations where catchment simulation is needed include:

- Events of a suitable magnitude have not been recorded and there is a need to extrapolate from recorded events to more extreme flood events;
- Data has not been monitored at the point of interest and there is a need to generate data at that point for estimation of the design flood characteristics; and
- Changed management options for a catchment are being considered and there is a need to predict the impacts of the proposed changes on the design flood characteristics prior to implementation of the proposed catchment management strategy..

Alternative techniques for generation of the desired data can be categorised as:

- Single burst (either the design peak burst of an event or the design total event) technique;
- Monte-Carlo technique (see, for example, Weinmann *et al*. [14]); or
- Continuous simulation technique.

Historically, implementation of single burst techniques has been the more popular technique. The alternative approaches, however, are gaining in popularity as computing capacity increases.

Where the single burst technique has been implemented with an assumption that the frequency of the rainfall is transformed to the frequency of the resultant flood characteristic, it can be argued that the method as applied is a regional flood method where the catchment model is a complex regression model. This question becomes more relevant when the implementation of the approach requires the values of parameters to be selected on the basis of ensuring the transformation of frequencies from rainfall to flood flow is achieved. An example of this approach (commonly referred to as AEP neutrality) is provided by Hill *et al*. [6] who developed a method of estimating loss model parameters that are likely to result in the frequency of the rainfall being transferred to the frequency of the design flood flow.

Application of a catchment simulation approach, irrespective of the approach category requires calibration and validation of the model prior to use of the model for estimation of the design flood characteristics. During the calibration phase, the primary aim is the selection of parameter values that ensure the model adequately replicates the catchment response; in other words, the primary aim is estimation of the many parameters.

PARAMETER ESTIMATION

All three generic techniques involve the use of models and hence the issue of identification of the parameter values for these models becomes important. It is worthwhile noting that the term model in the context of this discussion refers to both process based models of catchment response to storm events and statistical models used for describing the relationship between the flood characteristic quantile and its magnitude through application of FFA techniques or through application of regression models commonly employed in regional flood frequency estimation techniques (regional methods). Irrespective of how the need for the parameter estimation is generated, the process of parameter estimation is one of ensuring that employment of the model and selected parameter values reflect the patterns within the available historic data.

While the need for estimation of parameter values is needed for any flood estimation technique, the focus herein will be on the estimation of parameter values for use in catchment modelling techniques. Fundamental to this discussion is the concept that all predictions obtained from systems of models for prediction of catchment response to either individual storm events or sequences of storm events will contain residuals, or differences between the predicted and recorded values. While these residuals arise from multiple sources during the modelling process, the principle sources, following Kuczera *et al*. [10] can be classified as:

- Process errors;
- Structural errors;
- Parameter errors; and
- Data errors.

The *process errors* arise from the difference between the conceptual process incorporated in the modelling system and the actual process within the catchment; in other words, the process errors arise from the need for representation of physical processes in a mathematical formulation. The magnitudes of these errors are influenced directly by the degree of simplification within the modelling system. For example, use of one-dimensional and two-dimensional river models result in differing simplifications and different errors. It is worth noting, however, that additional complexity in the process model may not result in a reduction in residuals due to the increased number of parameters necessary for the more complex model.

Structural errors are the result of the manner in which the various process models are combined to provide the catchment modelling system. In many situations, alternative structures are available. For example, shown in Figure 3 are two alternative conceptual systems presented by Ball *et al*. [1] for incorporation of rainfall losses in a catchment model. As illustrated through this example, structural errors are linked to conceptualisation of processes in the catchment model. Hence, the distinction between a process model and a structural error is diffuse and, in many cases, difficult to quantify. This difficulty is shown also by considering the study of Umakhanthan and Ball [13] who investigated the influence of the rainfall model (the spatial and temporal distribution of rainfall across the catchment). Errors in the rainfall model could be classified as either a process error or a structural error depending on the viewpoint of the analyst.

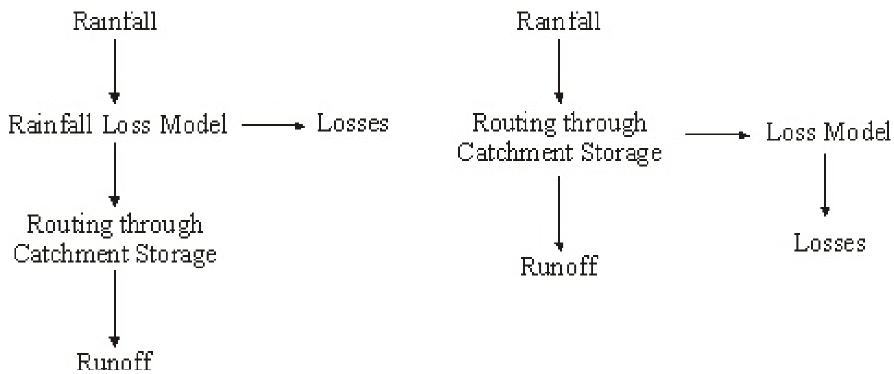

Figure 3. Alternative Rainfall Loss Models (after Ball *et al*., 2011)

The third form of error are the *parameter errors*. For discussion of parameter errors, there are two alternative cases that need consideration. The first of these cases is where sufficient moniotored data is available for estimation of the parameter values using data from the catchment. The second case occurs where there is insufficient data available for estimation of parameter values using data from the catchment and hence estimation of parameter values is based on regional relationships and other inference models. An example of the use of inference models for estimation of parameter values is presented by Choi and Ball [3] who investigated the use of land-use classifications for prediction of the imperviousness of subcatchments.

In both cases, parameter errors arise from differences between the true value of the parameter and that used in the simulation. Residuals arising from errors of this type have been the focus of significant historical research with a significant volume of this research focussed on the problem of obtaining optimal or near optimal values for the modelling system parameters. Arising from this research has been the concept of equifinality (see Beven and Binley, [2]) which can be paraphrased as there are multiple sets of parameter values that will result in similar system performance.

When values for spatially distributed parameters are sought, the concept of equifinality becomes increasingly relevant. This is illustrated in Figure 4 where the variation in the Sum of Squared Error (SSE) with variations in the conceptual width of a single subcatchment is shown. As shown in this figure, similar values of the SSE can be obtained over a wide range of values. Also illustrated in this figure is the need to consider sets of parameter values rather than the value of a single parameter. The same value of the conceptual width parameter can result in many values of the SSE; these multiple SSE values are the result of different values for the other parameters necessary for the operation of the catchment modelling system.

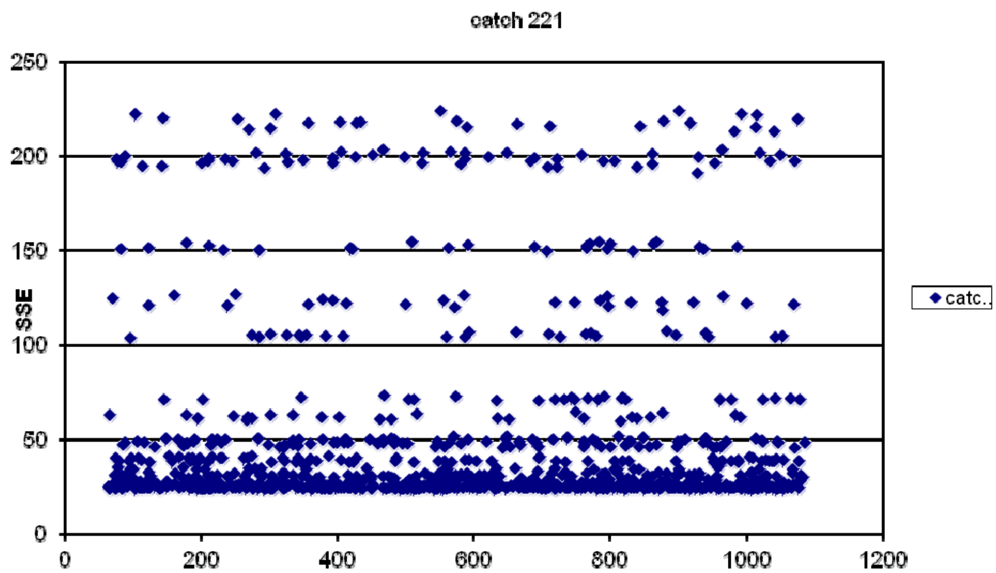

Figure 4. Variation in Sum Squared Residual with SWMM Width Parameter for 1 Subcatchment in a Catchment Model with 42 Subcatchments

The last form of error are the *data errors*. While there are many types of data errors, the characteristic of these errors are that they represent the difference between the true value of the monitored data and the value recorded in the database. An example of errors of this type are those errors arising from the need to extrapolate the rating table for a gauging station above the highest gauging to enable transformation of the recorded level to an equivalent flow; errors arising from this source were the focus of Kuczera [8]. Note that, in addition to the extrapolation error, a measurement error associated with the recording of the level also occurs.

As previously discussed, the focus of the calibration and validation of a catchment modelling system for generation of data necessary for design flood estimation is minimisation of the residual. While the parameter error generally is the major concern, the alternative forms of error cannot be neglected. The need for this is discussed by Choi and Ball [4] who showed that minimising the prediction residual can result in degraded predictions for alternative events. In other words, the parameter values that minimise the prediction residual include corrections for the non-parameter errors that occurred during the event.

If the process of calibration and validation of a catchment model is considered to be the estimation of generic parameter values that result in minimisation of the residuals, then the problem of calibration and validation becomes one of data mining.

CONCLUSIONS

Within Australia, Australian Rainfall and Runoff (Pilgrim [12]) is the primary guideline document used for prediction of design flood quantiles. While the 1987 edition of Australian Rainfall and Runoff and its predecessors have served the engineering profession and the general community well, in the period since the release of the previous edition, a number of developments have arisen. Consideration of these developments resulted in the decision by

Engineers Australia to prepare another edition of Australian Rainfall and Runoff. As part of this revision, the problem of estimating design flood quantiles has been recast as a problem in hydroinformatics.

Analysis of approaches and techniques used for design flood estimation highlighted the conceptual consistency with the definition of a hydroinformatic problem and the focus of the alternative approaches and techniques. This consistency confirms the proposal that design flood estimation is a problem in hydroinformatics.

ACKNOWLEDGEMENT

The ideas expressed herein have been developed as part of the revision of Australian Rainfall and Runoff. This revision has been supported by the Department of Climate Change and Energy Efficiency, Geosciences Australia, and the Bureau of Meteorology. In addition, there has been extensive in-kind support from the members of Engineers Australia.

REFERENCES

[1] Ball, JE, Babister, KM, and Retallick, M, Revisiting the design flood problem, *Proceedings 34th IAHR Congress*, Brisbane, Australia, (2011), pp 31-38, ISBN 978-0-85825-868-6

[2] Beven, K, and Binley, A, The future of distributed models: model calibration and uncertainty prediction, *Hydrological Processes*, Vol. 6, (1992), pp 279-298.

[3] Choi, KS and Ball, JE, Parameter estimation for urban runoff modelling, *Urban Water*, Vol 4 No 1, (2002a), pp 31-41.

[4] Choi, KS and Ball, JE, A generic calibration approach: Monitoring the calibration, *Proc. 2002 Hydrology and Water Resources Symposium*, Melbourne, Australia, I.E.Aust., (2002b)

[5] Haddad, K and Rahman, A, Regional flood estimation in New South Wales Australia using Generalised Least Squares Quantile Regression, *ASCE, Journal of Hydrologic Engineering*, Vol. 16 No 11, (2011), pp 920-925.

[6] Hill, PI, Mein, RG and Siriwardena, L, How much rainfall becomes runoff? Loss modelling for flood estimation, *Industry Report 98/5, Cooperative Research Centre for Catchment Hydrology*, Department of Civil Engineering, Monash University, Clayton, Australia, (1998), ISSN 1039-7361

[7] Jin, M and Stedinger, JR, Flood frequency analysis with regional and historical information, *Water Resources Research*, Vol 25 No 5, (1989), pp 925-936.

[8] Kuczera, G, Correlated rating curve error in flood frequency inference, *Water Resources Research*, Vol 32 No 7, (1996), pp 2119 - 2127.

[9] Kuczear, G, Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference, *Water Resources Research*, Vol 35 No 5, (1999), pp 1551-1558.

[10] 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, *Journal of Hydrology*, Vol 331 No 1-2, (2006), pp 161-177.

[11] Mittelstadt, GE, McDermott, GE and Pilgrim, DH, Revised flood data and catchment characteristics for small to medium sized catchments in New South Wales, *Unpublished Master of Engineering Science Report*, Department of Water Engineering, School of Civil Engineering, The University of New South Wales, Sydney, NSW, (1987).

[12] Pilgrim, DH, "*Australian Rainfall and Runoff: A guide to flood estimation*", The Institution of Engineers Australia, Barton, ACT, (1987).

[13] Umakhanthan, K and Ball, JE, Rainfall Models for Catchment Simulation, *Australian Journal of Water Resources*, Vol 9 No1, (2005), pp 55-67.

[14] Weinmann, PE, Rahman, A, Hoang, TMT, Laurenson EM, and Nathan RJ, Monte Carlo simulation of flood frequency curves from rainfall - the way ahead, *Australian Journal of Water Resources*, IEAust, Vol 6 No 1, (2002), pp 71-80.