Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: https://www.tandfonline.com/loi/thsj20

Robustness of flood-model calibration using single and multiple events

J. E. Reynolds, S. Halldin, J. Seibert, C.Y. Xu & T. Grabs

To cite this article: J. E. Reynolds, S. Halldin, J. Seibert, C.Y. Xu & T. Grabs (2020) Robustness of flood-model calibration using single and multiple events, Hydrological Sciences Journal, 65:5, 842-853, DOI: 10.1080/02626667.2019.1609682

To link to this article: https://doi.org/10.1080/02626667.2019.1609682

Published online: 17 May 2019.

HYDROLOGICAL SCIENCES JOURNAL
2020, VOL. 65, NO. 5, 842–853
https://doi.org/10.1080/02626667.2019.1609682

SPECIAL ISSUE: HYDROLOGICAL DATA: OPPORTUNITIES AND BARRIERS

Robustness of flood-model calibration using single and multiple events

J. E. Reynolds[a,b], S. Halldin[a,b,c], J. Seibert[a,d,e], C.Y. Xu[a,f] and T. Grabs[a]

[a]Department of Earth Sciences, Uppsala University, Uppsala, Sweden; [b]Centre of Natural Hazards and Disaster Science (CNDS), Uppsala, Sweden; [c]Centre for Climate and Safety (CCS), Karlstad University, Karlstad, Sweden; [d]Department of Physical Geography, Stockholm University, Stockholm, Sweden; [e]Department of Geography, University of Zurich, Zurich, Switzerland; [f]Department of Geosciences, University of Oslo, Oslo, Norway

ABSTRACT

Lack of discharge data for model calibration is challenging for flood prediction in ungauged basins. Since establishment and maintenance of a permanent discharge station is resource demanding, a possible remedy could be to measure discharge only for a few events. We tested the hypothesis that a few flood-event hydrographs in a tropical basin would be sufficient to calibrate a bucket-type rainfall–runoff model, namely the HBV model, and proposed a new event-based calibration method to adequately predict floods. Parameter sets were chosen based on calibration of different scenarios of data availability, and their ability to predict floods was assessed. Compared to not having any discharge data, flood predictions improved already when one event was used for calibration. The results further suggest that two to four events for calibration may considerably improve flood predictions with regard to accuracy and uncertainty reduction, whereas adding more events beyond this resulted in small performance gains.

ARTICLE HISTORY
Received 16 August 2018
Accepted 5 March 2019

EDITOR
A. Castellarin

GUEST EDITOR
C. Cudennec

KEYWORDS
floods; rainfall–runoff model; event-based calibration; value of information; ungauged basins; tropical climate

1 Introduction

Floods give rise to problems and disasters in many basins around the world, especially in developing countries, where discharge data are not available. Reliable and accurate prediction of discharge in ungauged basins by means of conceptual rainfall–runoff models then becomes a difficult task. Given that these models are empirical rather than physical representations, their parameters cannot be directly estimated from measurements or basin characteristics and, thus, some calibration is needed. One approach to overcome this limitation is parameter regionalization through statistical or process-based methods (Blöschl *et al.* 2013). However, one might argue that, if time and resources were available, the best approach would be to perform field measurements, preferably including time periods representative of the processes occurring in the basins to be modelled (Seibert and Beven 2009, Tada and Beven 2012). The latter would help to constrain model parameters, to improve the predictive ability of rainfall–runoff models and, ultimately, to improve decision making in flood- or water-resources management. In general, discharge data used in calibration should be as long as possible, but it is interesting to study in more detail how valuable smaller amounts of data might be.

Many studies dealing with parameter and performance variability caused by the length or quantity of the calibration data have reported good modelling predictions when short discharge time series are used (Sorooshian *et al.* 1983, Yapo *et al.* 1996, Brath *et al.* 2004, McIntyre and Wheater 2004, Seibert and Beven 2009, Tada and Beven 2012, Melsen *et al.* 2014). Although the latter is encouraging, results from these studies vary considerably from site to site and are difficult to generalize as a unique solution for data-scarce conditions. Explanations of these differences could be related to the balance between calibration-data needs and basin-, climate- and model-complexity, dominant flood types (Sikorska *et al.* 2015), and to the way calibration data were distributed in time for each study case (i.e. either as continuous or discontinuous) (Parajka *et al.* 2013).

When continuous periods of discharge data have been used for calibration, between three months and eight years of data have been reported as sufficient to achieve parameter sets that can provide good simulations and optimal performance as discussed in the following. Sorooshian *et al.* (1983) studied the influence of calibration data variability and length on model reliability in the temperate Leaf River basin (Mississippi, USA). They report that one full hydrological cycle of daily observations, preferably one wet year, was a minimum requirement to secure an adequate activation of all the parameters during calibration and, therefore, representation of the various phenomena occurring in a basin. Harlin (1991) states that for one daily bucket-type rainfall–runoff model and three basins in Sweden (i.e. two snow basins with clear seasonal patterns and one basin with a humid marine climate), performance improvement beyond two years was limited and after six years was not significant. Anctil *et al.* (2004) analysed the impact of the length of observed records on the performance of two runoff models, an artificial neural network and a conceptual rainfall–runoff model, in the temperate Seine River basin (Paris, France) dominated by rain floods. Their analysis showed that the best model performance was reached more or less evenly in both models with three and five years of daily calibration data. Brath *et al.* (2004) calibrated a spatially-distributed model at an hourly resolution using individual flood events and several scenarios of

CONTACT J. E. Reynolds eduardo.reynolds@geo.uu.se

continuous periods equal to or shorter than one year to simulate floods in the temperate Reno River basin (north-central Italy). They validated the results against 14 independent flood events and found the best model performance when parameters were calibrated against 12 months of data, but suggest at least three months of data to obtain good results. Tada and Beven (2012) explored the most effective way to extract the information content of short observation periods in three temperate basins in Japan. They tested calibration periods ranging between four and 512 days, generated by randomly selecting their starting day, and reported good simulations in validation for all data length scenarios, even down to four days, but performance range was larger and also with poor simulations when the calibration period had a small number of observations. Furthermore, they state that it may be difficult to identify *a priori* those best performing short periods before a field-campaign plan is implemented in an ungauged basin.

Few studies have explored the possibilities of model calibration using non-continuous discharge records. Based on their assumptions, two types of studies can be distinguished: the first assumes only a limited number of individual spot discharge observations to be available (Perrin *et al.* 2007, Seibert and Beven 2009, Pool *et al.* 2017), and the second assumes continuous discharge records to be available but only for a limited number of events (Brath *et al.* 2004, McIntyre and Wheater 2004, Tan *et al.* 2008, Seibert and McDonnell 2013).

Within the first group, Perrin *et al.* (2007) found that 350 discontinuous calibration days randomly sampled out of a longer dataset including dry and wet conditions were sufficient to reach robust parameter values in 12 basins with different hydrological and climatic conditions in the USA (ranging from semi-arid to very wet). Seibert and Beven (2009) investigated how many discharge measurements are needed for the calibration of 11 seasonally snow-covered basins in Sweden and reported that model performance increased largely when the number of randomly sampled days increased from two to 16, while little further improvement was seen beyond 32 daily observations. Pool *et al.* (2017) explored an optimal strategy for sampling runoff and the potential of calibrating a rainfall–runoff model with only 12 daily discharge observations during a year in 12 basins located in regions with temperate and snow climates across the eastern USA. They report that different sampling strategies have different information value for runoff prediction since they found that strategies including high-flow conditions simulated better hydrographs, whereas strategies including low and mean values predicted better flow–duration curves.

Within the second group, McIntyre and Wheater (2004) tested the effects of using different subsets of daily continuous and event-based data for calibration of an in-river phosphorus model in the upper part of the Hun River (China). When event-based data were used, they found results less biased in some cases. Their results suggest that only small amounts of data may be necessary if data are sampled on an event basis rather than at fixed intervals. Brath *et al.* (2004) report that simulations obtained after individually calibrating 10 flood events provide, as an average, better model performance, in terms of volume and peak errors, than a simulation obtained after calibrating on a single event. The latter was not the case in terms of time-to-peak errors, since they found better performance when only using one event in calibration. Tan *et al.* (2008) investigated the feasibility of calibrating a rainfall–runoff model using 10 representative storm flows in a tropical basin located in Singapore. They calibrated the events individually using data with a 5-min temporal resolution and averaged the optimal parameter sets found to obtain an event-based parameter set. Furthermore, they compared continuous- and event-based calibration in validation against 106 events and reported that the former is more reliable in predicting runoff volume, whereas the latter is better in predicting the overall shape of the hydrograph, peak flow and time to peak. Seibert and McDonnell (2013) explored the value of limited streamflow measurements and soft data in the temperate Maimai basin (New Zealand). They reported that 10-min discharge data from one event, or at least 10 measurements sampled during high flows, were as informative for calibration of a simple conceptual model as three months of continuous discharge data. Generally, in those studies with multiple events available, the parameter sets that fit best each event were selected first in calibration and then tested individually or by an averaging procedure in validation. This approach resulted in parameter sets that contained only information for an individual event rather than that for all the events available as a whole.

In summary, the general understanding in the literature is that adequate calibration depends on the value or information content of the datasets rather than on their length or quantity, thereby implying that good model predictions might be achieved by using discharge measurements from a limited number of events. Previous studies focus on topics related to the minimum data requirements for model calibration, in terms of time-series length, quantity and information content, for example, on knowing how many runoff measurements are needed to obtain parameter sets similar to those obtained from longer records, on developing methods to locate the most informative parts of hydrographs, and on exploring methods for optimal sampling strategies. Our study tests the hypothesis that adequate model calibration in a data-scarce environment is possible when hydrographs of only a few events are available, and assesses the value of such events in terms of quantity and information content. Furthermore, a new calibration method based on limited discharge data is proposed to improve the robustness of flood predictions compared to the scenario of no discharge data being available. Calibration in this method is event based, but it takes into account the time series of all the events available as one. The investigation was carried out for a tropical basin in Panama using the generalized likelihood uncertainty estimation (GLUE) framework (Beven and Binley 2014). Our research questions are:

(1) How many high-flow events are needed for model calibration to achieve an adequate flood prediction?
(2) How much predictive performance is gained by increasing the number of high-flow events for calibration?
(3) Is the information content of individual extreme flood events equally informative for calibration?

2 Material and methods

2.1 Study site

The study area is the Boqueron River basin in Panama. The basin has a drainage area of 91 km^2 and is mostly covered with forest. The differences in elevation can range from several metres to nearly 900 m. The basin has a tropical climate with a wet and a dry season. Rainfall mainly occurs between May and December in the form of thunderstorms with high intensity, and is normally convective and orographic. The mean annual rainfall in the basin is around 3800 mm year^{-1}, of which 72% leaves the basin as river discharge.

Based on hourly rainfall data from four stations, areal rainfall was calculated for the period 1997–2011 using the Thiessen polygon method. River stage is recorded continuously in a natural cross-section at the Peluca station and is stored every 15 min. Hourly maximum-annual discharge for 27 years (1985–2011) and 15 years of continuous discharge data (1997–2011) are available. Long-term daily mean values of potential evaporation were estimated using daily pan evaporation data from the Tocumen station, located 36 km southeast of the basin. The continuous rainfall–runoff data available in this study were quality-controlled by Reynolds *et al.* (2017).

The events used in this study were identified using a threshold value. All events above the median annual flood (489 m^3 s^{-1} or 19. 4 mm h^{-1}, recurrence interval of 2.33 years) were chosen. This resulted in the selection of 10 events for the period between November 2002 and December 2010.

The length of the discharge time series for each event was defined as follows: (a) the start was the time step at which the precedent rainstorm started, and (b) the end was when its rainstorm had ended and the percentage change in the recession varied for 10 consecutive hourly time steps by less than 5%, or when the percentage change had a positive increase because of the occurrence of a new storm. Generally speaking, the 10 flood events had fast responses and were triggered by rainfall events with large volumes, high rainfall intensities and relatively short durations (usually less than 24 h with the exception of two events). The length of the events varied between 18 and 51 h, rainfall storms between 6 and 35 h, rainfall maxima between 31 and 96 mm h^{-1}, total rainfall depth between 137 and 573 mm, total runoff depth between 100 and 547 mm, discharge peak between 489 m^3 s^{-1} (19.4 mm h^{-1}) and 1029 m^3 s^{-1} (40.9 mm h^{-1}), initial discharge between 4 m^3 s^{-1} (0.2 mm h^{-1}) and 57 m^3 s^{-1} (2.3 mm h^{-1}), time delay (lapse between mass centroid of rainfall and discharge peak) between 0.9 and 7.2 h, and they occurred throughout the entire rainy season (May–December) (Table 1).

2.2 Model

A simple bucket-type hydrological model, the HBV model (Bergström 1976) was used in this study. The model uses rainfall, air temperature and potential evaporation as input to simulate river discharge at the basin scale. Applications of the model have shown general good results in basins with different hydrological and climatological conditions (Häggström *et al.* 1990, Seibert 1999, Li *et al.* 2014, Reynolds *et al.* 2018, Wang *et al.* 2019). The HBV model has low data requirements and low computational demands. This allows performing a large amount of model runs, which was important in this study.

The software HBV-light (v. 4.0.0.17[1]) allows several model structures, but the one chosen for this study was the standard one, which was spatially set-up in a lumped way. Model descriptions are given by Bergström (1992) and Seibert and Vis (2012).

2.3 Experimental design

To assess how many events are sufficient to achieve reliable and accurate flood simulations in data-scarce conditions, we started with the assumption that discharge data were available for a limited numbers of events. This number of events (M_p) was varied from one to five (out of the total number of events). Then, the HBV model was calibrated for all possible event combinations for a given M_p. Calibrated parameters were subsequently used in validation to simulate all individual events and to assess their predictive ability. Each event combination had its own number of behavioural parameter sets that resulted in specific model performances in validation. The median of those performances, referred to here as median model performance, was

Table 1. Characteristics of the 10 selected flood events.

Event ID	1	2	3	4	5	6	7	8	9	10
Length (h)	29	40	18	36	51	36	28	44	47	32
Rainfall depth (mm)	176.2	137.3	215.5	138.2	282.3	137.7	151.6	184.3	573.1	265.1
Rainfall duration (h)	11	14	6	13	26	8	15	12	35	20
Rainfall peak (mm h^{-1})	71.4	35.8	95.7	54.9	34.0	55.4	31.0	50.9	77.6	53.0
Mean rainfall intensity (mm h^{-1})	16.0	9.8	35.9	10.6	10.9	17.2	10.1	15.4	16.4	13.3
Discharge peak (m^3 s^{-1})	821.3	536.2	517.3	708.5	746.2	488.8	518.2	561.6	1,028.9	822.7
Runoff peak (mm h^{-1})	32.6	21.3	20.6	28.2	29.7	19.4	20.6	22.3	40.9	32.7
Runoff depth (mm)	154.8	112.8	104.0	179.9	277.2	100.0	125.8	148.0	547.4	227.3
Initial discharge (m^3 s^{-1})	13.4	16.4	3.9	56.6	13.8	6.5	16.5	8.2	15.2	3.9
Initial runoff (mm/h)	0.5	0.7	0.2	2.3	0.5	0.3	0.7	0.3	0.6	0.2
Time delay (h)	0.9	5.3	2.7	1.1	1.5	2.5	5.4	2.2	7.2	2.3
Date of occurrence (dd-mm-yy)	30-11-02	27-05-06	07-07-06	18-11-07	28-12-07	10-08-08	18-11-09	07-11-10	08-12-10	26-12-10
Day of year	334	147	188	322	362	223	322	311	342	360

[1]http://www.geo.uzh.ch/en/units/h2k/Services/HBV-Model.html.

computed for each event combination. Finally, median model performances achieved in validation were compared for different values of M_p. The median value of accuracy for every value of M_p is referred to as the median of the median model-performance values. All possible combinations of one to five events were tested, which resulted in 637 different calibration set-ups (10 for $M_p = 1$, 45 for $M_p = 2$, 120 for $M_p = 3$, 210 for $M_p = 4$, and 252 for $M_p = 5$).

To assess if the information content of individual flood events was equally informative for calibration, median model performances in validation were compared based on the types of events used in calibration for each value of M_p. The 10 calibration events were first classified into three types based on their characteristics (Section 2.4), and each of the 637 calibration set-ups was classified in three larger groups according to its event-type combination for every value of M_p (Fig. 1). These larger groups represent the information content of the three types of event identified. If the calibration set-up had an event-type combination that had most of the events of one type, it was assumed that this set-up was only representative of that type of event, regardless of whether another type was included in it. For example, if the calibration set-up included two Type-1 events and one Type-2 event (i.e. event-type combination 1–1–2), then this set-up was assigned to the group that represented the information content of the Type-1 events for $M_p = 3$. When the majority of events in any calibration set-up was equal in quantity for Type-1 and Type-2 events (i.e. combinations 1-2, 1-2-3, 1-1-2-2 and 1-1-2-2-3), this set-up was attributed to the groups that represented the information content of the two types of events, so equal weight was given to each group.

Only one Type-3 event was identified when clustering the events, and all the calibration set-ups that included this event were considered representative of the information content of this type. Finally, after all calibration set-ups were classified, the model performance in validation was compared and assessed in terms of the information content of each type of event for every value of M_p.

2.4 Clustering events and event-type combinations

The events were characterized by: (a) rainfall depth, (b) rainfall duration, (c) rainfall peak, (d) runoff depth, (e) runoff peak, (f) day of year of flood-peak occurrence, and (g) time delay. These characteristics were normalized with respect to their standard deviation and the events were then classified into three clusters based on k-means clustering. This method partitions the observations into k clusters in which each observation belongs to the cluster with the nearest mean. The squared Euclidian distance metric was used for minimization.

The main steps in k-means clustering are: (i) select the number of clusters k to classify the observations; (ii) randomly select k observations, which are the initial clusters and means; (iii) assign each observation to the cluster closest to its mean; (iv) calculate the mean value of the cluster after a new observation has been assigned to it; (v) randomly select a new set of k observations so they are treated as the initial clusters; and repeat steps (iii)–(iv) until the sum of variation of all clusters is minimized and cluster membership does not change. The number of clusters k was chosen to be equal to three because the reduction in variation between the clusters was the greatest for this value of k, and considerably lower for higher values.

2.5 Model calibration

Behavioural parameter sets were selected from Monte Carlo simulations for each calibration set-up. One-hundred-thousand parameter sets were randomly generated assuming a uniform distribution with predefined parameter value ranges suggested from previous HBV applications (Seibert 1999, Booij 2005). Ranges of parameter values were the same for every calibration set-up (Table 2).

Simulations were run continuously for the period January 2000 to December 2010. It was assumed that the input-data time series were available to drive the model for the preceding period, but that discharge data from only one or more events were available for calibration.

There are significant trade-offs between different objective functions (Jie *et al.* 2016), but it is also recognized that multiple objective functions can improve the performance of a model by characterizing different attributes of the hydrograph (Madsen 2000). Three objective functions for evaluating the performance of the flood predictions were used in this study: (a) mean volume error of the flood events, $F_1(\theta)$, (b) mean root mean square error (RMSE) of the flood events, $F_2(\theta)$, and (c) mean peak-flow error of the flood events, $F_3(\theta)$. The first measure is an indicator of the agreement between the simulated and observed water volume (i.e. long-term water balance), the second statistic measure is an indicator of the overall agreement of the flood hydrograph, whereas the third measure is an indicator of the agreement of the flood peak.

Mean volume error of the flood events:

$$F_1(\theta) = \frac{1}{M_P}\sum_{j=1}^{M_p}\left|\frac{1}{n_j}\sum_{i=1}^{n_j}\left[Q_{\mathrm{obs,i}} - Q_{\mathrm{sim},i}(\theta)\right]\right| \qquad (1)$$

Table 2. Parameter ranges used for model calibration and for computation of the upper and lower benchmarks.

Parameter	Description	Min–Max	Unit
Soil moisture routine			
P_{FC}	Maximum soil-moisture storage	50–1000	mm
P_{LP}	Soil-moisture value above which actual evaporation reaches potential evaporation.	0.0–1.0	-
P_{BETA}	Determines the relative contribution to runoff from rainfall	0.5–6.0	-
Response routine			
P_{PERC}	Threshold parameter	0.0–19.2	mm d^{-1}
P_{ALPHA}	Non-linearity coefficient	0.1–1.9	-
P_{K1}	Storage coefficient 1	0.0024–1.2	d^{-1}
P_{K2}	Storage coefficient 2	0.0012 – 0.03	d^{-1}
Routing routine			
P_{MAXBAS}	Length of isosceles triangular weighting function	1.0–24.0	h

Mean RMSE of the flood events:

$$F_2(\theta) = \frac{1}{M_p}\sum_{j=1}^{M_p}\left[\frac{1}{n_j}\sum_{i=1}^{n_j}\left[Q_{\text{obs},i} - Q_{\text{sim},i}(\theta)\right]^2\right]^{\frac{1}{2}} \quad (2)$$

Mean peak-flow error of the flood events:

$$F_3(\theta) = \frac{1}{M_p}\sum_{j=1}^{M_p}\left|Q_{\text{obs max},j} - Q_{\text{sim max},j}\right| \quad (3)$$

where $Q_{\text{obs},i}$ is the observed runoff at time i in each event, $Q_{\text{sim},i}$ is the simulated runoff at time i in each event, n_j is the number of time steps in each flood event j, M_p is the total number of flood events in calibration, $Q_{\text{obs max},j}$ is the observed peak runoff in the flood event j, $Q_{\text{sim max},j}$ is the simulated peak runoff in the flood event j, and θ is the set of model parameters to be calibrated.

The three measures were merged into a single objective function by an aggregate measure $F(\theta)$ called the Euclidean distance (4), which gives equal weight to every measure (Madsen 2000):

$$F(\theta) = \left[\sum_{k=1}^{3}\left(F_k(\theta) + A_k\right)^2\right]^{\frac{1}{2}} \quad (4)$$

where A_k are transformation constants corresponding to different objective functions, k is the index of the objective function being transformed (i.e. 1,2,3). The value of A_k was computed as follows:

$$A_{\max} = \max\left(F_{k,\min}\right) \quad (5)$$

$$A_k = A_{\max} - F_{k,\min} \quad (6)$$

All objective functions give positive values, where values close to zero indicate best performance. The minimum of each function obtained from 100 000 model runs was considered as optimum. Parameter sets for every calibration set-up were considered behavioural if they gave an $F(\theta)$ score (4) equal to or less than $[F(\theta)_{\min} + 1]$.

Model performance scores in the evaluation period were compared to an upper and lower benchmark, as suggested by Seibert *et al.* (2018). The upper benchmark represented the best possible discharge simulation that could be achieved with data from our study basin, whereas the lower benchmark represented a simulation based on calibration from only literature information. The upper benchmark of $F(\theta)$ was the best score possible in validation from the 100 000 parameter sets previously generated. Similarly, the upper benchmarks of $F_1(\theta)$, $F_2(\theta)$ and $F_3(\theta)$ were the best possible scores obtained in validation, but only from the behavioural parameter sets selected using the $F(\theta)$ measure. For the lower benchmarks, the model was run with random parameters (here 500 sets) within typical ranges (Table 2). The discharge time series for the lower benchmark was obtained by averaging the ensemble of the 500 runoff simulations into a single mean time series. Thereafter, the model efficiencies of $F(\theta)$, $F_1(\theta)$, $F_2(\theta)$ and $F_3(\theta)$ resulting for this mean time series were calculated and used as the lower benchmarks.

3 Results

3.1 How many events are required for model calibration?

The 10 selected events were combined in all possible ways into clusters of one to five events. For every scenario of data availability, the model was calibrated and tested in validation. The median model performances obtained in validation were compared for every value of M_p to answer our first question.

Results for the $F(\theta)$ measure (Fig. 2(a,b)) show that if only one extreme event were available, this would improve predictability in comparison to the scenario of no data available at all (i.e. $F(\theta)$ values were above the lower benchmark). Having at least two events available seemed sufficient in terms of the median value of accuracy for $F(\theta)$, since it did not always improve for M_p values greater than two (Table 3). Furthermore, it was also surprising to find that median value of accuracy for $F(\theta)$ slightly decreased when three events were used instead of two (from 10.60 to 10.78 error units). However, the uncertainty ranges of $F(\theta)$ reduced considerably for a greater number of events (Table 3). This uncertainty reduction was clearly seen when moving from the scenario of having only one event available to the scenario of having four events (the range for $F(\theta)$ decreased from 1.76 to 0.44). For M_p values higher than four, no performance improvement was seen. The median values of accuracy for volume error of floods, $F_1(\theta)$, were relatively similar for all values of M_p, but uncertainty ranges were smaller when using a greater number of events (Fig. 2(c)). Furthermore, median model performance of $F_1(\theta)$ was typically below its lower benchmark, but better performance is clearly seen when the largest amount of discharge data was used for calibration. Root mean square error, $F_2(\theta)$, and peak error of floods, $F_3(\theta)$, were more sensitive to varying M_p than $F_1(\theta)$ (Fig. 2(c–f)). The accuracy and reduction in uncertainty for $F_2(\theta)$ and $F_3(\theta)$ improve for a greater number of events, but, similar to $F(\theta)$, not much improvement in performance is noticeable for M_p values higher than four (Table 3).

3.2 Assessing information content of individual extreme events

The characteristics of the 10 flood events varied considerably, but they could be classified into three types. The assessment of information content of individual events was based on the types of events used in calibration for different scenarios of data availability.

3.2.1 Flood types

The k-means clustering algorithm divided the 10 events into three groups: Type-1, Type-2 and Type-3 (Fig. 3). Type-1 events typically occurred during the last two months of the year during which also the most extreme monthly rainfall totals occurred (ETESA 2018). Type-1 events were further characterized by relatively large discharge peaks (between 700 and 850 $m^3\ s^{-1}$) and runoff depth (between 150 and 300 mm), short time delays (less than 2.2 h) and long rainfall durations (between 11 and 26 h). In contrast, Type-2 events occurred throughout the entire rainy season and were characterized by relatively low discharge peaks (below 600 $m^3\ s^{-1}$), low runoff depth (below 150 mm), long time delays (between

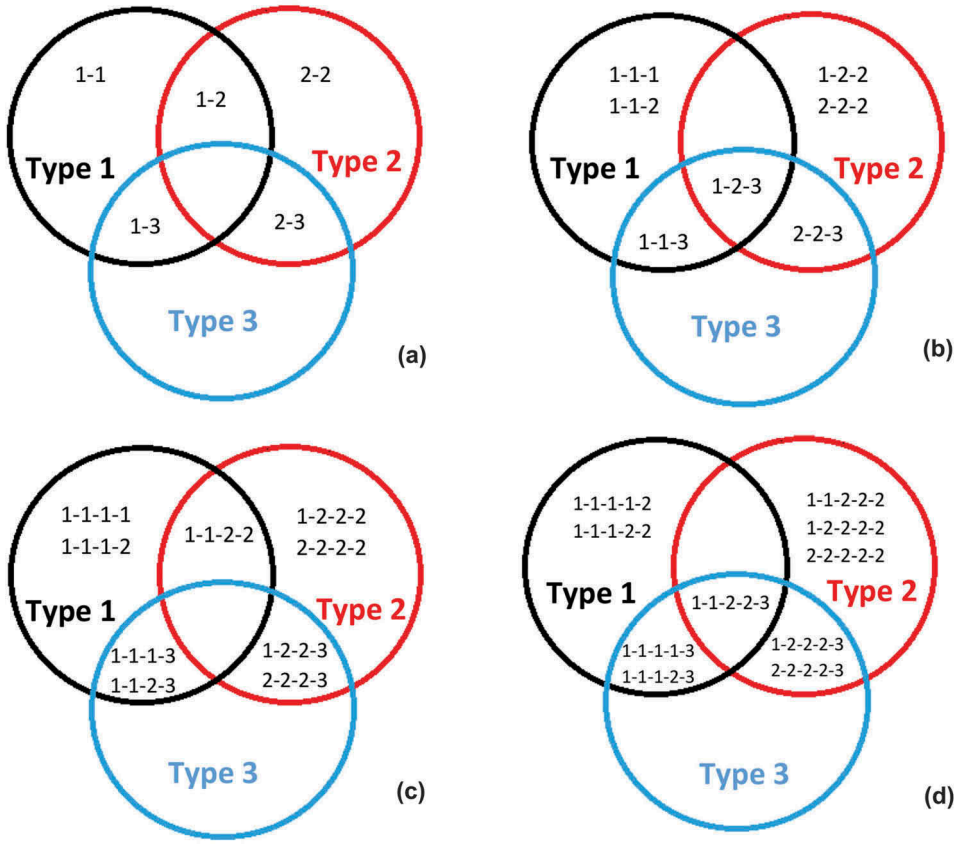

Figure 1. Classification of calibration set-ups, based on their event-type combination, to groups representing the information content of each type of event when M_p equals (a) two, (b) three, (c) four and (d) five. For every M_p value, each circle represents the information content of one type of event and the serial numbers inside them indicate the event-type combinations of the calibration set-ups included in each group.

2.2 and 5.5 h) and short rainfall durations (between 6 and 15 h). The single Type-3 event occurred in the last month of the rainy season of year 2010 and was the most extreme of all 10 events. It was characterized by the largest discharge peak (1029 m^3 s^{-1}), the largest rainfall and runoff depths (573 and 547 mm, respectively), the longest rainfall duration (35 h) and the longest time delay (7.2 h). The distinction between the three event types is visible by comparing both their hydrographs and event characteristics (Fig. 3, Table 1). It is worth noting that antecedent conditions (represented as discharge before the onset of each event) showed small correlations with the aforementioned event characteristics (non-parametric Kendall correlation tests returned ρ coefficients of $\leq \pm 0.29$).

3.2.2 Information content of individual flood events based on type

When the 10 events had been classified into three types and when the 637 calibration set-ups had been grouped based on their event-type combination, median model performances in validation were compared with respect to the type of events used in calibration for every scenario of data availability. Flood predictions using the $F(\theta)$ measure improved when at least one event was available for calibration regardless of the event type (Fig. 4(a)). The comparison of the $F(\theta)$ measure obtained when using only a single event ($M_p = 1$) shows the highest median accuracy results for the Type-3 event. The median value of accuracy was found to be better for Type-2 than for Type-1 events, but the uncertainty was smaller for the latter (Table 4). This was the case for all the objective functions used in calibration (Fig. 4(a–d)). However, given the low number of events for each type (i.e. four Type-1 events, five Type-2 events and only one Type-3 event), it is difficult to generalize the findings when $M_p = 1$.

The results for $M_p = 2$ varied considerably for any type of event, similarly to those for $M_p = 1$. Calibration set-ups in the Type-2 group resulted in higher median values of accuracy and reduced uncertainty compared to set-ups in the Type-1 group (Fig. 4(e–h), Table 4). Calibration set-ups that included the Type-3 event resulted in larger uncertainties for estimating event volume (quantified by $F_1(\theta)$, Fig. 4(f)). However, these set-ups in the Type-3 group were also associated with considerably smaller uncertainties for estimating the hydrograph shape (quantified by $F_2(\theta)$, Fig. 4(g)), and peak flows (quantified by $F_3(\theta)$, Fig. 4(h)).

When $M_p = 3$, uncertainty ranges for all objective functions, except for $F_1(\theta)$, were relatively similar for the three types of events but with different median values of accuracy (Fig. 4(i–l), Table 4). For $M_p > 3$, median values of accuracy and uncertainty ranges were relatively similar for any type of event used in calibration (Fig. 4(m–t), Table 4). Our results suggest that when less than three events are available, the type

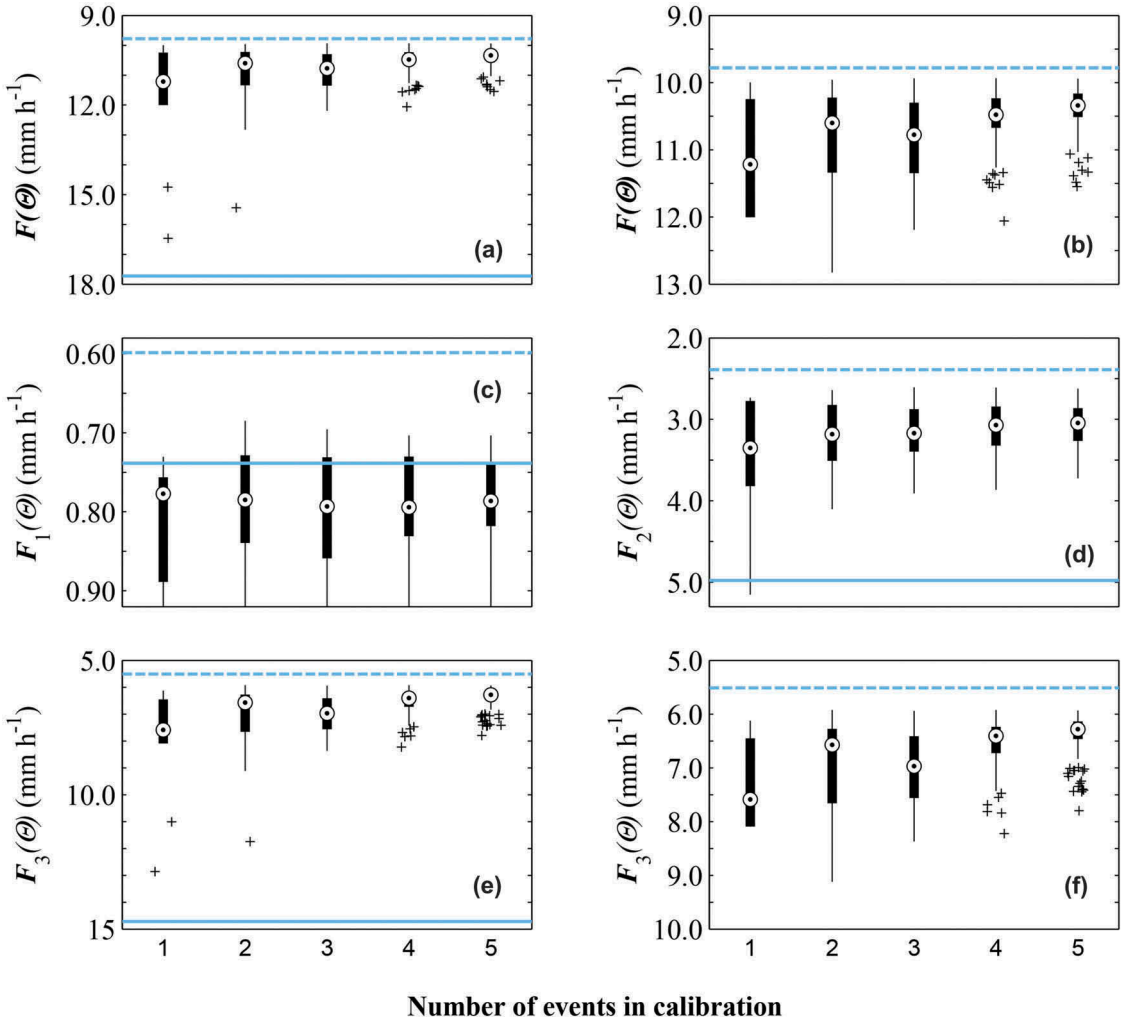

Figure 2. Boxplots of median performance in validation for different scenarios of data availability. Median model performance refers to the median values of model performance achieved by the behavioural parameter sets for each calibration set-up. The dashed (blue) lines represent the upper benchmarks, whereas the solid (blue) lines represent the lower benchmarks; (b) and (f) are zoom-ins of (a) and (e), respectively. Upper and lower benchmarks of $F_1(\theta)$ would be the equivalent of relative volume errors equal to 13% and 15% respectively; upper and lower benchmarks of $F_2(\theta)$ would be the equivalent of Nash-Sutcliffe efficiencies of 0.90 and 0.45, respectively, whereas upper and lower benchmarks of $F_3(\theta)$ would be the equivalent of relative error of peak flow of 22% and 56%, respectively.

of event matters in calibration, whereas for a greater number of events, the type of event influences prediction performance less and becomes less important.

Table 3. Summary statistics of distribution of median performance in validation for different scenarios of data availability. The range is defined as the difference between the 75th and 25th percentiles of median performance.

Measure (mm h^{-1})	Percentile	$M_p = 1$	$M_p = 2$	$M_p = 3$	$M_p = 4$	$M_p = 5$
$F(\theta)$	25th	10.25	10.22	10.30	10.23	10.16
	50th	11.22	10.60	10.78	10.48	10.34
	75th	12.00	11.34	11.35	10.67	10.51
	Range	1.76	1.12	1.05	0.44	0.35
$F_1(\theta)$	25th	0.76	0.73	0.73	0.73	0.74
	50th	0.78	0.78	0.79	0.79	0.79
	75th	0.89	0.84	0.86	0.83	0.82
	Range	0.13	0.11	0.13	0.10	0.08
$F_2(\theta)$	25th	2.77	2.82	2.88	2.84	2.86
	50th	3.35	3.18	3.17	3.07	3.05
	75th	3.82	3.51	3.40	3.32	3.27
	Range	1.05	0.69	0.52	0.48	0.40
$F_3(\theta)$	25th	6.45	6.27	6.41	6.24	6.14
	50th	7.59	6.57	6.97	6.41	6.28
	75th	8.09	7.66	7.56	6.73	6.46
	Range	1.64	1.39	1.15	0.49	0.33

The latter can be illustrated when comparing the simulated hydrograph of event no. 9 (Type-3 event) with the behavioural parameter sets selected following the calibration of the different scenarios of data availability (Fig. 5). As expected, using data from the same event for calibration would result in a good fit (Fig. 5(a)). However, using only a single event of a different type, with a faster response, resulted in considerable overestimations of peak discharge and underestimations of the time to peak (Fig. 5(b)). When two different types of events were used in calibration (i.e. $M_p = 2$, where one event had a fast response and the other a slow response), peak discharge was still overestimated and time to peak was still underestimated but both much less than in the former case (Fig. 5(c)). For the M_p cases previously described, it is worth noting that the parameter sets could reproduce the first flow peak well. When $M_p \geq 3$, the information content of the events in calibration was relatively similar, which resulted in similar simulations of the peak discharge and errors in the time to peak (Fig. 5(d–e)), but with higher accuracy than for the first two M_p cases.

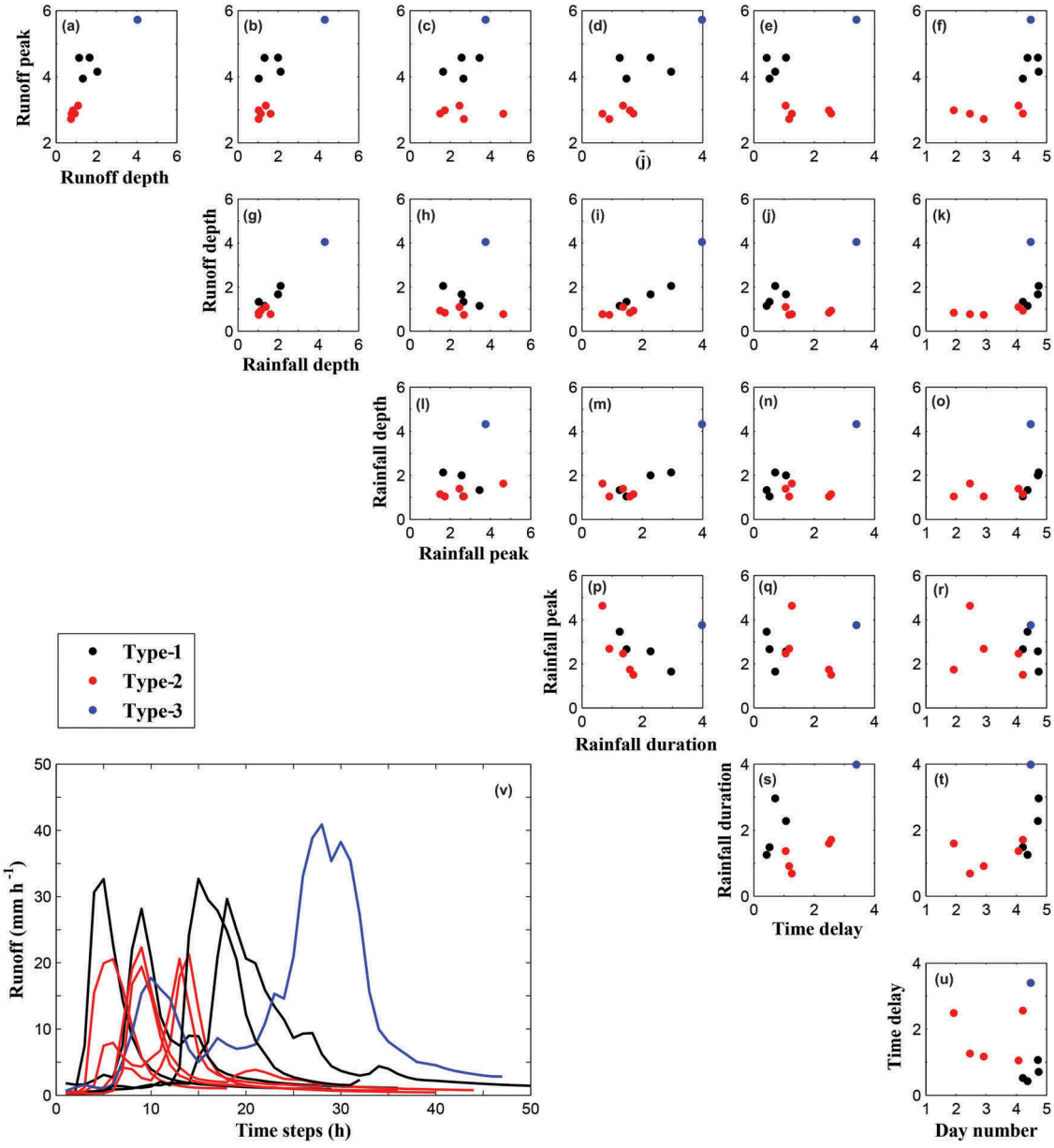

Figure 3. (a)–(u) Normalized event characteristics (unitless), and (v) discharge time series of flood events classified by their types. The event characteristics were normalized with respect to their standard deviation.

4 Discussion

Assessing the value of small amounts of data for model calibration is essential in data-scarce conditions. This study tested the hypothesis that adequate model calibration in a data-scarce environment is possible when hydrographs of only a few events are available. Studies where the information content of short runoff records was assessed were commonly carried out at daily resolutions, except for some (Brath *et al.* 2004, Tan *et al.* 2008, Seibert and McDonnell 2013, Melsen *et al.* 2014). Here, we took advantage of the availability of long time series of hourly rainfall–runoff data in a small tropical basin, which we treated as ungauged for the purpose of the experiment. While the hourly sampling of the data was sufficient to simulate runoff at the time scale of the basin response, it also provided more information at the event scale, since the time series of the events are longer at such fine temporal resolution in comparison to the data being available at a daily resolution.

The calibration in our experiment was event-based and, to avoid making assumptions about initial conditions before the occurrence of each event, it was assumed that the entire time series of the input data was available to drive the model. Although it is known that any calibration procedure depends on the interaction between input data and model structure (Beven 2001), errors in the input data were not corrected since no information about uncertainties of the data was available. Consequently, we assumed that rainfall estimates characterized correctly the true input to the basin. Even if rainfall uncertainty was large for the periods where no discharge data were used in calibration, it is assumed that the initial conditions before the occurrence of each event did not have a large influence for generating runoff. Antecedent wetness conditions in tropical basins,

Table 4. Summary statistics of distribution of median performance in validation for different scenarios of data availability based on the type of the events used in calibration. The range is defined as the difference between the 75th and 25th percentiles of median performance. TY1, TY2, TY3: Type-1, −2, −3, respectively.

Measure (mm h^{-1})	Percentile	M_p = 1			M_p = 2			M_p = 3			M_p = 4			M_p = 5		
		TY1	TY2	TY3	TY1	TY2	TY3	TY1	TY2	TY3	TY1	TY2	TY3	TY1	TY2	TY3
$F(\theta)$	25th	10.99	10.18	10.25	10.23	10.19	10.23	10.41	10.23	10.23	10.24	10.20	10.23	10.21	10.14	10.22
	50th	11.70	10.82	10.25	10.74	10.52	10.40	10.80	10.75	10.57	10.48	10.43	10.52	10.38	10.28	10.35
	75th	13.26	13.12	10.25	11.74	11.09	11.02	11.27	11.35	11.05	10.70	10.65	10.68	10.57	10.45	10.49
	Range	2.27	2.94	0.00	1.50	0.90	0.79	0.86	1.12	0.82	0.46	0.45	0.44	0.36	0.31	0.27
$F_1(\theta)$	25th	0.78	0.75	0.76	0.73	0.73	0.74	0.77	0.72	0.73	0.74	0.73	0.73	0.77	0.73	0.73
	50th	0.83	0.76	0.76	0.80	0.75	0.79	0.80	0.77	0.79	0.81	0.78	0.81	0.80	0.78	0.79
	75th	0.90	0.90	0.76	0.88	0.83	0.84	0.85	0.87	0.86	0.83	0.83	0.83	0.83	0.81	0.82
	Range	0.12	0.15	0.00	0.15	0.10	0.10	0.07	0.15	0.13	0.09	0.10	0.10	0.06	0.08	0.09
$F_2(\theta)$	25th	3.28	2.75	3.07	2.83	2.79	2.97	2.97	2.84	2.90	2.84	2.82	2.88	2.91	2.84	2.91
	50th	3.71	2.77	3.07	3.23	3.00	3.23	3.26	3.05	3.11	3.18	3.02	3.22	3.20	2.99	3.19
	75th	4.47	4.28	3.07	3.59	3.48	3.49	3.43	3.38	3.41	3.35	3.26	3.41	3.31	3.25	3.30
	Range	1.19	1.53	0.00	0.77	0.69	0.52	0.46	0.54	0.51	0.51	0.44	0.53	0.40	0.41	0.38
$F_3(\theta)$	25th	7.11	6.42	6.13	6.36	6.20	5.99	6.43	6.36	6.32	6.27	6.20	6.19	6.16	6.11	6.10
	50th	7.93	7.40	6.13	6.69	6.47	6.12	6.85	6.87	6.47	6.41	6.40	6.29	6.28	6.22	6.19
	75th	9.55	9.28	6.13	7.95	7.55	6.59	7.45	7.56	7.44	6.69	6.75	6.64	6.47	6.42	6.40
	Range	2.43	2.87	0.00	1.59	1.35	0.60	1.03	1.20	1.12	0.42	0.56	0.45	0.31	0.31	0.30

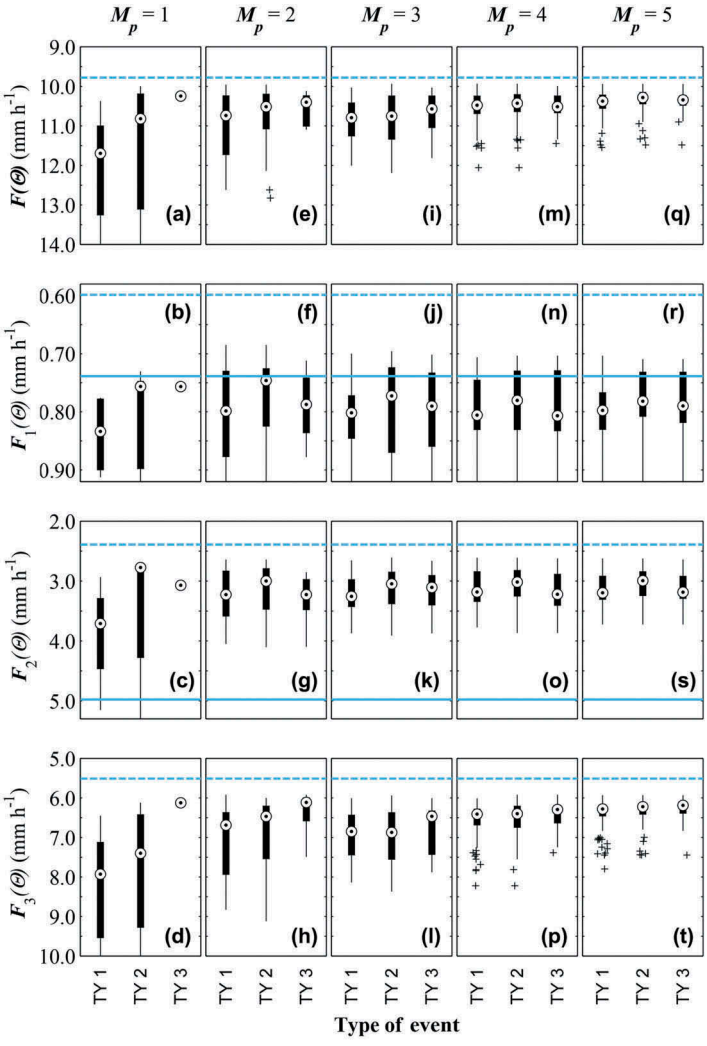

Figure 4. Boxplots of median performance in validation for different scenarios of data availability based on the type of the events used in calibration. Median model performance refers to the median values of model performance achieved by the behavioural parameter sets for each calibration set-up. The dashed (blue) lines represent the upper benchmarks, whereas the solid (blue) lines represent the lower benchmarks. Lower benchmarks of $F(\theta)$ and $F_3(\theta)$ were equal to 17.7 and 14.7 mm h^{-1}, respectively. Upper and lower benchmarks of $F_1(\theta)$ would be the equivalent of relative volume errors of 13% and 15%, respectively; upper and lower benchmarks of $F_2(\theta)$ would be the equivalent of Nash-Sutcliffe efficiencies of 0.90 and 0.45, respectively, whereas upper and lower benchmarks of $F_3(\theta)$ would be the equivalent of relative error of peak flow of 22% and 56%, respectively.

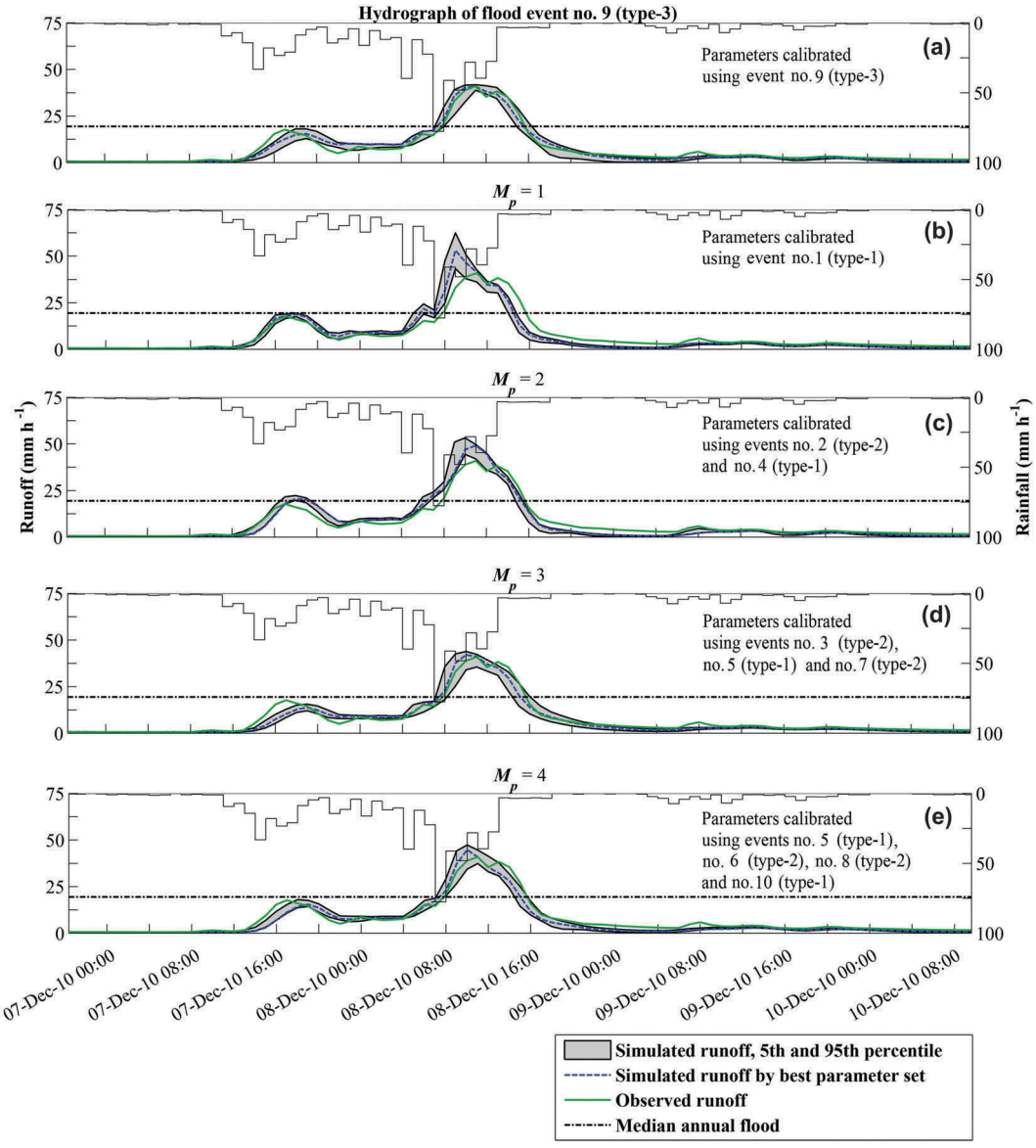

Figure 5. Simulations for the Boqueron River basin at Peluca for an extreme flood event (Type-3): (a) when discharge data of such an event were available for calibration; (b–e) when the number of events used in calibration (M_p) was equal to one, two, three and four, respectively. The simulated runoff corresponds to the 5th and 95th percentiles of model simulations obtained with the behavioural parameter sets for each scenario.

as the one used in this study, tend to be more random and not relevant for producing floods, since they are often dominated by infiltration-excess overland flow (Rosbjerg *et al.* 2013). The latter was supported by the fact that we did not find a high correlation between initial discharge and other flood event characteristics.

Our findings suggest that continuous observations from at least two events may be sufficient to calibrate a model for flood prediction. While increasing the number of calibration events leads to increased performance and reduced uncertainty, the effect seems to level off already after four calibration events. This strengthens the argument that calibration-data needs rely on the information content in observations rather than on the pure number of individual observations. The results are in line with the findings of McIntyre and Wheater (2004) that fewer calibration data are necessary if data are taken in an event-based way. One explanation for this could be that continuous measurements of individual events at a high temporal resolution may provide sufficient information for constraining model parameters as suggested by Seibert and McDonnell (2013). Our results are also in line with the findings of Tan *et al.* (2008), in which runoff volume predictions were better when the discharge data used in calibration were the largest in terms of quantity. Contrary to previous event-based studies, we calibrated our model using multiple events at once, depending on the scenario of data availability, rather than calibrating our model for each individual event.

Although there is a lot of information to be gained from individual events, not all events were equally informative for all the scenarios of data availability. When one or two events were used for calibration, the results obtained for all measures varied considerably for any type of event, in terms of median accuracy and uncertainty ranges. This is because it is highly unlikely that one single event could include sufficient information for adequate model calibration and for accurate prediction of all types

of floods, even though it can provide better model predictions than without data. When the number of events used for calibration was equal to or greater than three, it seemed that the information content of individual events became relatively equal and less influential in the predictions, regardless of the characteristics and magnitude of them. It is assumed that by doing the calibration of the model using multiple events at once, parameter sets selected as behavioural include sufficient information to predict different types of events and not just one type as it happens when one or two events are used. The low number of events identified for each type resulted in few calibration set-ups compound only of events of one type, which makes it difficult to generalize any assessment of their information content for every scenario of data availability. To overcome the latter, a larger number of calibration set-ups, based on their event-type combination, were considered representative of the information content of a given type of event, even if they included events of other types. It might be argued that using calibration set-ups that include multiple types of events to represent the information content of a specific type of event would have a large influence on the results. However, an additional experiment in which only the calibration set-ups compound of one type of event were used, showed similar results in terms of information content of the events as we did here, although with smaller performance ranges possibly because of the small number of calibration set-ups with such characteristic.

The results reported herein are based on one tropical basin and one bucket-type rainfall–runoff model. It can be assumed that calibration-data needs in small and flashy basins located in regions dominated by rain floods will be relatively similar to the one found in this study. Data needs are expected to be higher in large basins because of the effects of rainfall spatial variability on hydrograph response. For the latter case, different scenarios of rainfall activity for the same areal rainfall, such as storm movement across the basin or localized rainfall events, are expected to result in different shapes of the hydrograph and therefore, more events may be needed to cover such variability. One can also expect that calibration-data needs will be larger in basins dominated by snowmelt floods in spring and rain floods in autumn. Similarly, data needs are also expected to depend on model complexity. Here, a simple and lumped bucket-type hydrological model with eight free parameters was used in our experiment. Using a more complex model would mean more parameters to calibrate and, hence, longer time series of runoff data may be needed for adequate identification of the parameters which may not be feasible in data-scarce conditions. In either case, calibration-data requirements are expected to be relative to the complexity of the basins, climate and models, as well as by the dominant flood types for each case. Our study case was a small basin with a fast response dominated by rainfall floods which could explain the small amount of data needed for improving the prediction of floods compared to the scenario of no data.

It is suggested in the literature that sampling high flows in combination with recession data may be the most informative sampling strategy for hydrograph prediction (Pool *et al.* 2017). In our experiment, it was assumed that the discharge time series of the most extreme flood events were recorded and available for calibration. In practice, measuring these events is difficult because their probability of occurrence is low and it is not known in advance when these events will occur. Additionally, there is a higher risk for equipment to be damaged or lost when used to monitor extreme events. In a typical field campaign, it is most likely that a small to medium event would be gauged rather than an extreme one. The results of our experiment give an idea of what could be expected or achieved at best if such data were available.

It was also assumed that discharge data were available for short periods for each flood event. This implies that stage measurements were translated to river discharge by applying a rating curve, which is typically not available in ungauged basins. Although nowadays, thanks to modern technology, it is possible for hydrologists to obtain accurate and continuous level and flow-velocity measurements to estimate river discharge, even without a rating curve. If only water-level data could be made available, an alternative could be to rely on calibration methods where these data are used to constrain model parameters (Jian *et al.* 2017).

Model performance based on a limited number of events was compared against upper and lower benchmarks to take into account what could and should be possible with the data. This approach has been recognized as useful when there is an interest in comparing the value of gauging a few discharge measurements against the value of long-term gauging (Pool *et al.* 2017). Here, median model performance was typically better than the lower benchmarks in most cases, which showed the value of adding more events in calibration in comparison to when no data were available. Our findings raise interesting questions on what would be the minimum characteristics of an event required for calibration, and if it would be possible to identify these events before they occur. This would be valuable knowledge for water-resources managers with limited time and resources to optimize field campaigns that will result in the most informative data for model calibration and, therefore, for accurate flood prediction.

5 Conclusions

In this study, we explored the value of discharge time series of a limited number of events, in a data-scarce environment, for adequacy of model calibration. It was shown that if a few event hydrographs could be made available, good model performance could be achieved for predicting floods in basins dominated by rainfall floods. The specific conclusions from our analysis are

(1) Flood predictions will improve even if only one event hydrograph is available for calibration.
(2) Using two events for calibration may be sufficient, but accuracy and reduction in uncertainty may improve if data from more events can be made available. No significant improvements are gained after more than four events.
(3) When three or more events were used for calibration, the information content of individual events impacted model simulations less regardless of their magnitude.

This study was limited to one basin and one model. While the generality of our results needs to be further tested in other basins and with other models, these results are encouraging and call for the development of field methods towards making event hydrographs available in ungauged basins.

Acknowledgements

The authors thank the Panama Canal Authority (ACP), who provided the rainfall and discharge data from the Boqueron River basin, and the Department of Hydrometeorology at Empresa de Transmisión Eléctrica, S.A. (ETESA), who provided the pan-evaporation data. The authors also thank the two anonymous reviewers for their constructive comments for improving the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was carried out within the CNDS research school, supported by the Swedish International Development Cooperation Agency (Sida) through their contract with the International Science Programme (ISP) at Uppsala University [contract number: 54100006]. The fourth author was supported by the Research Council of Norway [FRINATEK Project 274310].

ORCID

J. E. Reynolds http://orcid.org/0000-0002-7390-7290
J. Seibert http://orcid.org/0000-0002-6314-2124

References

Anctil, F., Perrin, C., and Andréassian, V., 2004. Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall–runoff forecasting models. *Environmental Modelling & Software*, 19 (4), 357–368. doi:10.1016/S1364-8152(03)00135-X

Bergström, S., 1976. *Development and application of a conceptual runoff model for Scandinavian catchments*. Norrköping, Sweden: Swedish Meteorological and Hydrological Institute, SMHI Report No. RHO 7.

Bergström, S., 1992. *The HBV model – its structure and applications*. Norrköping, Sweden: Swedish Meteorological and Hydrological Institute, SMHI Hydrology, RH No. 4.

Beven, K., 2001. *Rainfall–runoff modelling: the primer*. Chichester, UK: Wiley.

Beven, K. and Binley, A., 2014. GLUE: 20 years on. *Hydrological Processes*, 28 (24), 5897–5918. doi:10.1002/hyp.10082

Blöschl, G., *et al.*, 2013. *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139235761

Booij, M.J., 2005. Impact of climate change on river flooding assessed with different spatial model resolutions. *Journal of Hydrology*, 303 (1–4), 176–198. doi:10.1016/j.jhydrol.2004.07.013

Brath, A., Montanari, A., and Toth, E., 2004. Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model. *Journal of Hydrology*, 291 (3–4), 232–253. doi:10.1016/j.jhydrol.2003.12.044

ETESA, 2018. Empresa de Transmisión Eléctrica, S.A. Available from: http://hidromet.com.pa/clima_historicos.php?sensor=2 [Accessed 14 June 2018].

Häggström, M., *et al.*, 1990. *Application of the HBV model for flood forecasting in six Central American rivers*. Norrköping, Sweden: Swedish Meteorological and Hydrological Institute, SMHI Hydrology, No. 27.

Harlin, J., 1991. Development of a process oriented calibration scheme for the HBV hydrological model. *Hydrology Research*, 22 (1), 15–36. doi:10.2166/nh.1991.0002

Jian, J., *et al.*, 2017. Towards hydrological model calibration using river level measurements. *Journal of Hydrology Regional Studies*, 10, 95–109. doi:10.1016/j.ejrh.2016.12.085

Jie, M.-X., *et al.*, 2016. A comparative study of different objective functions to improve the flood forecasting accuracy. *Hydrology Research*, 47 (4), 718–735. doi:10.2166/nh.2015.078

Li, H., Beldring, S., and Xu, C.-Y., 2014. Implementation and testing of routing algorithms in the distributed Hydrologiska Byråns Vattenbalansavdelning model for mountainous catchments. *Hydrology Research*, 45 (3), 322–333. doi:10.2166/nh.2013.009

Madsen, H., 2000. Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. *Journal of Hydrology*, 235 (3–4), 276–288. doi:10.1016/S0022-1694(00)00279-1

McIntyre, N.R. and Wheater, H.S., 2004. Calibration of an in-river phosphorus model: prior evaluation of data needs and model uncertainty. *Journal of Hydrology*, 290 (1–2), 100–116. doi:10.1016/j.jhydrol.2003.12.003

Melsen, L., *et al.*, 2014. Catchments as simple dynamical systems: a case study on methods and data requirements for parameter identification. *Water Resources Research*, 50 (7), 5577–5596. doi:10.1002/2013WR014720

Parajka, J., *et al.*, 2013. Prediction of runoff hydrographs in ungauged basins. *In*: G. Blöschl, *et al.*, eds. *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge, UK: Cambridge University Press, 227–269. doi:10.1017/CBO9781139235761.013

Perrin, C., *et al.*, 2007. Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. *Hydrological Sciences Journal*, 52 (1), 131–151. doi:10.1623/hysj.52.1.131

Pool, S., Viviroli, D., and Seibert, J., 2017. Prediction of hydrographs and flow-duration curves in almost ungauged catchments: which runoff measurements are most informative for model calibration? *Journal of Hydrology*, 554, 613–622. doi:10.1016/J.JHYDROL.2017.09.037

Reynolds, J.E., *et al.*, 2018. Definitions of climatological and discharge days: do they matter in hydrological modelling? *Hydrological Sciences Journal*, 63 (5), 836–844. doi:10.1080/02626667.2018.1451646

Reynolds, J.E., *et al.*, 2017. Sub-daily runoff predictions using parameters calibrated on the basis of data with a daily temporal resolution. *Journal of Hydrology*, 550, 399–411. doi:10.1016/j.jhydrol.2017.05.012

Rosbjerg, D., *et al.*, 2013. Prediction of floods in ungauged basins. *In*: G. Blöschl, *et al.*, eds. *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge, UK: Cambridge University Press, 189–226. doi:10.1017/CBO9781139235761.012

Seibert, J., 1999. Regionalisation of parameters for a conceptual rainfall–runoff model. *Agricultural and Forest Meteorology*, 98–99, 279–293. doi:10.1016/S0168-1923(99)00105-7

Seibert, J. and Beven, K., 2009. Gauging the ungauged basin: how many discharge measurements are needed? *Hydrology and Earth System Sciences*, 13 (6), 883–892. doi:10.5194/hess-13-883-2009

Seibert, J. and McDonnell, J.J., 2013. Gauging the ungauged basin: the relative value of soft and hard data. *Journal of Hydrologic Engineering*, 20 (1). doi:10.1061/%28ASCE%29HE.1943-5584.0000861

Seibert, J. and Vis, M., 2012. Teaching hydrological modeling with a user--friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, 16 (9), 3315–3325. doi:10.5194/hess-16-3315-2012

Seibert, J., *et al.*, 2018. Upper and lower benchmarks in hydrological modeling. *Hydrological Processes*, 32 (8), 1120–1125. doi:10.1002/hyp.11476

Sikorska, A.E., Viviroli, D., and Seibert, J., 2015. Flood-type classification in mountainous catchments using crisp and fuzzy decision trees. *Water Resources Research*, 51 (10), 7959–7976. doi:10.1002/2015WR017326

Sorooshian, S., Gupta, V.K., and Fulton, J.L., 1983. Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: influence of calibration data variability and length on model credibility. *Water Resources Research*, 19 (1), 251–259. doi:10.1029/WR019i001p00251

Tada, T. and Beven, K.J., 2012. Hydrological model calibration using a short period of observations. *Hydrological Processes*, 26 (6), 883–892. doi:10.1002/hyp.8302

Tan, S.B., *et al.*, 2008. Performances of Rainfall–runoff Models Calibrated over Single and Continuous Storm Flow Events. *Journal of Hydrologic Engineering*, 13 (7), 597–607. doi:10.1061/(ASCE)1084-0699(2008)13:7(597)

Wang, X., *et al.*, 2019. Understanding the discharge regime of a glacierized alpine catchment in the Tianshan Mountains using an improved HBV-D hydrological model. *Global Planetary Change*, 172, 211–222. doi:10.1016/J.GLOPLACHA.2018.09.017

Yapo, P.O., Gupta, H.V., and Sorooshian, S., 1996. Automatic calibration of conceptual rainfall–runoff models: sensitivity to calibration data. *Journal of Hydrology*, 181 (1–4), 23–48. doi:10.1016/0022-1694(95)02918-4