Hydrological Sciences Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/thsj20

Flood prediction using parameters calibrated on limited discharge data and uncertain rainfall scenarios

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) Flood prediction using parameters calibrated on limited discharge data and uncertain rainfall scenarios, Hydrological Sciences Journal, 65:9, 1512-1524, DOI: 10.1080/02626667.2020.1747619

To link to this article: https://doi.org/10.1080/02626667.2020.1747619

HYDROLOGICAL SCIENCES JOURNAL
2020, VOL. 65, NO. 9, 1512–1524
https://doi.org/10.1080/02626667.2020.1747619

Flood prediction using parameters calibrated on limited discharge data and uncertain rainfall scenarios

J. E. Reynolds[a,b,c], S. Halldin[a,b,d], J. Seibert[b,e,f], C.Y. Xu[a,g] 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]Swedish Meteorological and Hydrological Institute, Norrköping, Sweden; [d]Centre for Societal Risk Research (CSR), Karlstad University, Karlstad, Sweden; [e]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden; [f]Department of Geography, University of Zurich, Zurich, Switzerland; [g]Department of Geosciences, University of Oslo, Oslo, Norway

ABSTRACT

Discharge observations and reliable rainfall forecasts are essential for flood prediction but their availability and accuracy are often limited. However, even scarce data may still allow adequate flood forecasts to be made. Here, we explored how far using limited discharge calibration data and uncertain forcing data would affect the performance of a bucket-type hydrological model for simulating floods in a tropical basin. Three events above thresholds with a high and a low frequency of occurrence were used in calibration and 81 rainfall scenarios with different degrees of uncertainty were used as input to assess their effects on flood predictions. Relatively similar model performance was found when using calibrated parameters based on a few events above different thresholds. Flood predictions were sensitive to rainfall errors, but those related to volume had a larger impact. The results of this study indicate that a limited number of events can be useful for predicting floods given uncertain rainfall forecasts.

ARTICLE HISTORY
Received 15 April 2019
Accepted 19 February 2020

EDITOR
Castellarin

ASSOCIATE EDITOR
I. E.M. Mendiondo

KEYWORDS
floods; rainfall forecasts; rainfall–runoff modelling; event-based calibration; ungauged basins; value of information

1 Introduction

Reliable and accurate representation of rainfall is important for hydrological modelling and flood forecasting. Precipitation data are used as input to hydrological models to represent surface hydrological processes, since the hydrograph response is closely related to storm precipitation characteristics, such as amount, intensity and duration (Linsley 1967, Singh 1997). Rainfall–runoff models are sensitive to precipitation input data and if these inputs do not characterize the true precipitation inputs correctly, there is no empirical- or physically-based model that could be able to produce accurate streamflow simulations (Beven 2001, Kobold 2007, Chintalapudi *et al.* 2014, Wang *et al.* 2017).

A problem that arises when predicting floods in small to median sized basins with short concentration times (i.e. floods occurring at sub-daily time scales or in the order of a few hours) is that peak discharges tend to occur as a result of a localized rainfall event and times to peak may be too short for raising adequate warnings based on real-time rainfall observations (Ferraris *et al.* 2002). In such cases, the only option is to raise warnings of potential flooding based on rainfall forecasts. Rainfall forecasts derived from numerical weather-prediction (NWP) models with an ensemble prediction system (EPS) are commonly used as input to operational flood-forecasting models (Wetterhall *et al.* 2011). The accuracy ahead of time of the rainfall forecasts plays a large role on the possibility of increasing forecast lead time (Beven 2001, 2009, He *et al.* 2009, Wetterhall *et al.* 2011). Precipitation-forecast skills of EPSs have been previously evaluated (Buizza *et al.* 1999, Buizza and Hollingsworth 2002, Cloke and Pappenberger 2008). Buizza *et al.* (1999) showed that EPSs could be useful in Europe since they can return skilful predictions of low, i.e. lower than 2 mm (12 h)$^{-1}$ up to forecast day six, and high precipitation amounts, i.e. 2–10 mm (12 h)$^{-1}$ up to forecast day four. More into extremes, Buizza and Hollingsworth (2002) investigated the performance of EPSs for predicting three severe storms in Europe, and their findings showed that EPSs could give early indications of them. Kobold and Sušelj (2005) assessed the quality of precipitation forecasts generated by the European Centre for Medium-Range Weather Forecasts (ECMWF) for large events in Slovenia, a country characterized by torrential streams and fast runoff. In their assessment, they found that forecasts considerably underestimated the amount of observed precipitation by an average of 60%.

Although the accuracy of rainfall forecasts has improved with new technologies and methods, rainfall forecasts still are the main source of uncertainty in flood forecasting, which limits the usability of hydrological models in operational applications (Kobold 2007). Several studies have explored the effects of coupling NWP models as input to hydrological-hydraulic models for predicting floods in real time (Ferraris *et al.* 2002, De Roo *et al.* 2003, Bartholmes and Todini 2005, Kobold and Sušelj 2005, Verbunt *et al.* 2006, Xuan *et al.* 2009). Kobold and Sušelj (2005) showed in a sensitivity analysis of hydrological models to rainfall errors that the deviation in runoff is much larger than the deviation in rainfall and concluded that an error in the rainfall input to hydrological models could result in a high runoff deviation. Kobold

CONTACT J. E. Reynolds eduardo.reynolds@smhi.se Department of Earth Sciences, Uppsala University, Uppsala, Sweden
[1]US Geological Survey: http://hydrosheds.cr.usgs.gov/datadownload.php?reqdata=3demg Retrieved 7March 2016.

(2007) reported that an error of +10% in the amount of rainfall resulted in a 17% overestimation of the peak flood in Slovenian rivers. Similarly, Verbunt *et al.* (2006) coupled NWP-hydrological models to predict runoff for alpine tributaries in the Rhine basin and found that rainfall forecasts overestimated precipitation in the basins with higher elevations which resulted in overestimations of runoff peaks. Bartholmes and Todini (2005) coupled a distributed hydrological model with several European meteorological models (ranging from the limited area models to the ECMWF EPS) to predict floods in the River Po, in the north of Italy. They found that the quantitative precipitation forecasts were unreliable because the predicted discharge generally underestimated the observed peak discharge and the time to peak. De Roo *et al.* (2003) developed a European flood forecasting system by coupling ECWMF EPS forecasts with LISFLOOD models and found that the ensemble members tended to underestimate river flows. As shown in literature, rainfall-forecast errors in volume and intensity are still significant and can consequently lead to poor discharge predictions, quantitatively speaking. Despite this limitation, rainfall forecasts have been shown to be useful for real-time flood predictions as they enable deriving early qualitative indications for the possible occurrence of high-flow events. However, while these indications may be sufficient for issuing flood warnings the accuracy of rainfall forecasts still needs further improvement.

Lack of discharge data in data-scarce basins complicates flood forecasting by means of hydrological models even more. In ungauged basins, models can only be selected by a priori perception of the main processes, by reading the landscape or by using a model structure from a similar gauged basin (Parajka *et al.* 2013). After a suitable model structure is chosen, model parameters need to be estimated. If a physically-based model is chosen, its parameters could be estimated a priori or directly from measurements, basin characteristics and remote sensing. However, data demands of physically-based models may be too large to overcome in data-scarce conditions. On the other hand, conceptual rainfall–runoff models have little to moderate data demands but their parameters cannot be estimated a priori or directly because these are more empirical than physical representations (Parajka *et al.* 2013), and thus, some calibration is needed.

When data are lacking for calibration, one approach to overcome this limitation is by regionalization, i.e. transferring calibrated parameter values from gauged basins to ungauged basins (Blöschl *et al.* 2013, Yang *et al.* 2018, 2019a, 2019b). The concept of regionalization is based on the notion of similarity, where basins are considered to be similar if they behave hydrologically similar (i.e. similar runoff-generation processes, land use, soil type, rainfall regime and seasonality). There are several approaches for transferring calibrated model parameters from gauged to ungauged basins (Parajka *et al.* 2013): (a) parameter averaging or simulated discharge averaging in a group of similar gauged basins (Goswami *et al.* 2007, Yang *et al.* 2019b); (b) spatial interpolation (e.g. geostatistics); (c) using similarity measures (e.g. spatial proximity) (Kokkonen *et al.* 2003); (d) regression analysis between parameters and basin characteristics (Seibert 1999, Xu 1999, Merz and Blöschl 2004); and (e) regional calibration, where the coefficients of the relationships between parameters and basin characteristics are calibrated rather than the model parameters themselves (Fernandez *et al.* 2000, Szolgay *et al.* 2003).

If some observational data are available, most of them are usually at a daily resolution, which could be used for calibration and the resulting parameters are then used together with the rainfall forecasts for predicting floods at sub-daily resolutions. This method has been criticized because of possible inaccuracies of daily parameters (Littlewood and Croke 2008), or poor model performance at sub-daily resolutions (Bastola and Murphy 2013). However, recent studies by Reynolds *et al.* (2017) and Santos *et al.* (2018) showed relatively stable parameters across temporal resolutions, suggesting that the direct transferability of daily parameters to finer resolutions is possible.

Performing field measurements including periods representative of the main hydrological processes could be another option to overcome the lack of discharge data for model calibration (Seibert and Beven 2009, Reynolds *et al.* 2019). In an earlier study (Reynolds *et al.* 2019), we tested the hypothesis that a few high-flow events would be sufficient to calibrate a bucket-type rainfall–runoff model and our results indicate that two to four events, compared to the scenario of not having any data, could considerably improve flood predictions with regard to accuracy and uncertainty reduction. These results were encouraging, but the events used in calibration were above an extreme threshold (i.e. median annual flood, return period: 2.33 years), which their occurrence is difficult to predict and therefore, unlikely of being gauged during typical field campaigns. This raises the question of whether less extreme events above a threshold with a short return period can also be useful for flood-model calibration. From here on, it is presumed that events with a low frequency of occurrence are extreme events, whereas those with a higher frequency of occurrence are less extreme.

Rainfall-forecast uncertainties and lack of discharge data for calibration motivate further research in these fields to improve the predictability of floods and, ultimately, to be able to raise timely and adequate flood warnings in fast-response basins. This study aims to assess the influence of rainfall errors on the performance of a hydrological model in providing meaningful flood predictions when using parameters calibrated on a few events with a short return period. Here we assumed multiple scenarios of rainfall data based on real observations but with volume and duration errors. The investigation was carried out for a tropical basin in Panama and a bucket-type hydrological model, namely the HBV model, using the generalized likelihood uncertainty estimation (GLUE) framework (Beven and Binley 2014). Our research questions were: (i) Can events above a threshold with a short return period be useful in calibrating a hydrological model and providing reliable flood predictions? And (ii) Between volumetric and duration errors of the input-driving data, which of these have the most substantial impact on flood predictions?

2 Material and methods

2.1 Study site

The study area is the tropical Boqueron River basin located in Panama (see Appendix, Fig. A1). The basin is predominantly covered by forest and is characterized by sub-daily runoff responses. It covers a surface area of 91 km^2 and elevation ranges from 100 to 980 m a.s.l.[1]. Climate is wet between May and December, while the other months are relatively dry. Rainfall is convective and orographic, and normally occurs as torrential downpours. Mean annual rainfall and runoff are 3800 mm and 2728 mm, respectively (Reynolds *et al.* 2017).

Areal rainfall was calculated for the period 1997–2011 by Thiessen polygons using four stations with hourly rainfall data. River stage was recorded continuously in a natural cross-section at the outlet of the basin. Hourly maximum annual discharge for 27 years (1985–2011) and 15 years of continuous 15-min discharge data (1997–2011) were available. Long-term daily mean values of potential evaporation were estimated based on daily pan evaporation from a nearby station located 36 km southeast of the basin. The continuous rainfall–runoff data used in this study were previously quality-controlled by Reynolds *et al.* (2017), but rainfall was not corrected since no information about its uncertainties was available.

Two threshold values were used to select the events used in calibration: (a) the median annual flood, which is an extreme value with a low frequency of occurrence (489 $m^3\ s^{-1}$ or 19.4 mm h^{-1}, return period: 2.33 years) and (b) a more relaxed threshold (125 $m^3\ s^{-1}$ or 5.0 mm h^{-1}, return period: 1.01 years), which has a higher frequency of occurrence than the former (Fig. 1). The first threshold is at the 50th percentile of the maximum annual discharge dataset, whereas the second is at the 1st percentile of the same dataset and it is about four times smaller than the former. Ten flood events were identified above the first extreme threshold, whereas 107 were identified above the more relaxed threshold. The events above both thresholds were selected within the period between June 2000 and December 2011. The length of the events was defined as in Reynolds *et al.* (2019). The start of each flood event was the time step at which the precedent rainstorm started, while its end was when the percentage change in the recession decreased by less than 5% for 10 consecutive hourly time steps, or when the percentage change was positive because of the occurrence of a new rainfall event. Overall, the events selected had fast responses, showed several hydrological behaviours and compromised a wide range of characteristics (Table 1).

2.2 Model

The widely known HBV model (Bergström 1976) was used in this study (software HBV-light, version 4.0.0.17[2]. The HBV model represents a class of bucket-type hydrological models commonly used for water-resource planning, operational forecasting and research. The model consists of four computational routines to simulate river discharge and it uses precipitation, air temperature and potential evaporation as input data. It has been applied successfully in many basins with different climatological conditions (Häggström *et al.* 1990, Seibert 1999, Reynolds *et al.* 2018, Osuch *et al.* 2019, Wang *et al.* 2019, Yang *et al.* 2019a). The model has low data and computational demands, which allows us to perform a large number of simulations to take into account model structure and parameter uncertainties. The standard model structure, set up in a spatially lumped way, was chosen to carry out the simulations. A detailed description of the model is given by Seibert and Vis (2012).

Table 1. Characteristic value ranges of flood events identified for thresholds with different return periods.

Return period (years)	2.33	1.01
Runoff threshold (mm h^{-1})	19.4	5.0
Number of events	10	107
Length (h)	18–51	9–53
Rainfall depth (mm)	137–573	16–573
Rainfall duration (h)	6–35	2–42
Rainfall peak (mm h^{-1})	7–96	7–96
Mean rainfall intensity (mm h^{-1})	9.8–35.9	2.6–35.9
Discharge peak ($m^3\ s^{-1}$)	489–1029	130–1029
Runoff peak (mm h^{-1})	19.4–40.9	5.2–40.9
Runoff depth (mm)	99–547	18–547
Time delay (h)	0.9–7.2	0.0–8.4

2.3 Experimental design

The influence of rainfall errors on the performance of a hydrological model in providing meaningful flood predictions after calibration to limited discharge data was assessed in two steps. The first part of the experiment was designed to test whether events above a threshold with a short return period can be useful in calibrating a hydrological model and obtaining reliable flood predictions. Following the findings in Reynolds *et al.* (2019), we started with the assumption that only three events were available and sufficient for calibration. Two sets of event combinations were generated: one set consisted of all possible combinations of three events above the threshold with a low frequency of occurrence (return period: 2.33 years), which resulted in 120 event combinations, and the other set consisted of 100 random combinations of three events between the 107 events above the threshold with a short return period

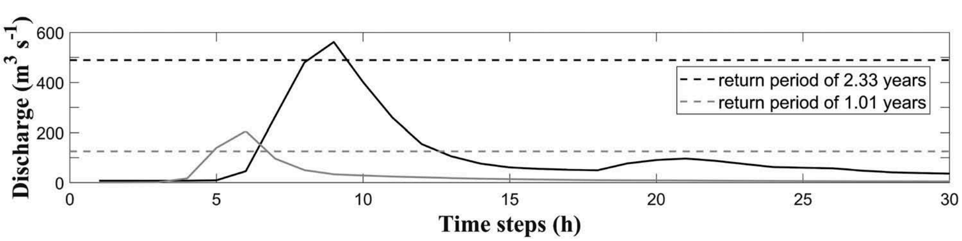

Figure 1. Typical hydrograph responses for the Boqueron River basin. The black dashed line represents the threshold with a low frequency of occurrence (489 $m^3\ s^{-1}$, return period: 2.33 years), whereas the grey dashed line represents the threshold with a high frequency of occurrence (125 $m^3\ s^{-1}$, return period: 1.01 years).

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

(i.e. 1.01 years). The HBV model was calibrated for both sets of event combinations, and their flood predictive ability was tested in validation using a rainfall scenario with a quality as good as that of real-time observations. A fixed number of behavioural parameter sets was selected and tested in validation for each event combination. The median value of model performance values in validation was computed for each event combination and the resulting values for the two sets of event combinations were compared, together with an upper and lower benchmark (Seibert *et al.* 2018). The upper benchmark represented the best model performance that could be achieved with the data of the study basin, while the lower benchmark represented what could be achieved if only information of parameter value ranges was available. For the lower benchmark, a mean discharge time series was computed from runoff simulations generated with 500 random parameter sets. Thereafter, model efficiency resulting from this mean discharge time series was used as the lower benchmark for comparison.

In the second part of the experiment, the predictive ability of the calibrated parameters from the second set of event combinations (i.e. the one based on events above the threshold with a short return period, i.e. 1.01 years) was further tested using several rainfall scenarios as input to the model. The rainfall scenarios were based on real-time observations but included (artificially generated) duration and volume errors. In this part of the experiment, the median of the median values (MMV) of model performances obtained for all the rainfall scenarios was compared to answer the question concerning the effects on flood predictions caused by uncertainties of the rainfall data.

2.4 Rainfall scenarios

The calibrated parameter sets were tested in validation for several rainfall scenarios. These were based on real-time rainfall observations but several scenarios of volume and duration errors were added to the data. Nine possible scenarios of volume errors were considered (i.e. – 50%, – 37.5%, – 25%, – 12.5%, 0%, +25%, +50% +75% +100%), each of them with nine possible duration-error scenarios (same percentages as for volume error). When adding the duration errors to the rainfall observations, the centre of mass of each rainfall event was used as a reference to increase and reduce its duration in both directions. Rainfall volume and distribution of each event were the same for every duration-error scenario. Only rainfall duration and intensities were modified based on the duration change of each scenario. The latter resulted in 81 possible rainfall scenarios with different degrees of uncertainty.

2.5 Identification of floods for validation

The 107 identified hydrographs included events with different characteristics, but the calibrated parameters were only tested for predicting the more extreme events and therefore, these events were grouped to separate the extreme events from the low–medium ones. First, the hydrographs were characterized by quantitative descriptors including: rainfall depth, rainfall duration, rainfall peak, runoff depth, runoff peak, day of year of peak-flood occurrence and time delay (lapse between rainfall centre of mass and discharge peak). Subsequently, the values of these descriptors were normalized with respect to their standard deviation and the hydrographs were then divided into four clusters based on *k*-means clustering. The latter is a clustering technique frequently used in hydrology (Dettinger and Diaz 2000, Parajka *et al.* 2010, Kingston *et al.* 2011), which partitions the observations into *k* clusters by minimizing the sum of squared distances between the observations and the cluster means. Each observation was assigned to the cluster closest to its mean and the squared-Euclidian-distance metric was used for minimization. By choosing a number of clusters *k* equal to four, the more extreme hydrographs were clearly identified into two clusters, whereas the low–medium events were found in the other two. In total, 13 flood events were identified (including the ten events identified above the median annual floods) and used to assess the predictive ability of the calibrated parameters in validation.

2.6 Model calibration

Based on 100,000 Monte Carlo simulations, behavioural parameter sets were selected for each event combination. Monte Carlo simulations were based on parameter sets randomly generated assuming a uniform distribution with predefined parameter-value ranges used in previous HBV applications in many basins worldwide (Seibert 1999, Booij 2005, Reynolds *et al.* 2019) (Table 2). The chosen parameter value ranges for assessing the value of data are, thus, rather general and not specifically suited for the kind of tropical region in which the study basin is located. The model was run continuously from January 1999 until December 2011, assuming that input-data time series were available for the preceding period and that discharge data from only three events were available for calibration.

To characterize different attributes of the hydrograph, we used three objective functions: mean volume error of the events, $F_1(\theta)$; mean root mean square error (RMSE) of the events, $F_2(\theta)$; and mean peak-flow error of the events, $F_3(\theta)$. The first function indicates the agreement between the simulated and observed water volume, the second indicates the overall agreement of the hydrograph, and the third indicates the agreement of the peak flow. The objective functions are calculated as follows:

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

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

$$F_3(\theta) = \frac{1}{M_p} \sum_{j=1}^{M_p} \left| Q_{\mathrm{obsmax},j} - Q_{\mathrm{simmax},j} \right| \quad (3)$$

where $Q_{\mathrm{obs},i}$ is the observed runoff at time i in each event j; $Q_{\mathrm{sim},i}$ is the simulated runoff at time i in each event j; n_j is the number of time steps in each event j; M_p is the total number of events in calibration and validation; $Q_{\mathrm{obs\ max},j}$ is the observed

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

peak runoff in event j; $Q_{\text{sim max},j}$ is the simulated peak runoff in event j; and θ is the set of model parameters to be calibrated.

An aggregate measure $F(\theta)$ (Eq. 4) was used to combine the three objective functions (1)–(3) into one (Madsen 2000). This aggregate measure gives equal weight to every objective function and was used to select the behavioural parameter sets.

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

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

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

$$A_k = A_{\max} - F_{k,\min} \tag{6}$$

For each of the objective functions used in calibration (Eqs. (1)–(4)), a value of zero corresponds to a perfect fit and values increase with decreasing performance. The 100 best parameter sets, with respect to $F(\theta)$, for each event combination were retained and considered as behavioural.

2.7 Model evaluation

Four additional measures were used in validation to assess the goodness-of fit of the 13 simulated flood hydrographs, which have been widely applied in flood-model calibration and flood forecasting (Madsen 2000, Chahinian and Moussa 2009, Jie *et al.* 2016): relative volume error, V_E; relative peak-flood error, P_E; relative time-to-peak error, T_E; and Nash-Sutcliffe efficiency, R_{eff}. The $F_1(\theta)$ relates to V_E, $F_2(\theta)$ to R_{eff} and $F_3(\theta)$ to P_E. The additional measures are calculated as follows:

$$V_E = \frac{1}{M_P}\left[\sum_{j=1}^{M_p}\left|\frac{\sum_{i=1}^{n_j}\left[Q_{\text{obs},i} - Q_{\text{sim},i}(\theta)\right]}{\sum_{i=1}^{n_j} Q_{\text{obs},i}}\right|\right]100\% \tag{7}$$

$$P_E = \frac{1}{M_p}\left[\sum_{j=1}^{M_p}\left|\frac{Q_{\text{obsmax},j} - Q_{\text{simmax},j}}{Q_{\text{obsmax},j}}\right|\right]100\% \tag{8}$$

$$T_E = \frac{1}{M_P}\sum_{j=1}^{M_p}\left|T_{\text{obs},j} - T_{\text{sim},j}\right| \tag{9}$$

$$R_{\text{eff}} = 1 - \left[\frac{\sum_{j=1}^{M_p}\sum_{i=1}^{n_j}\left[Q_{\text{sim},i}(\theta) - Q_{\text{obs},i}\right]^2}{\sum_{j=1}^{M_p}\sum_{i=1}^{n_j}\left[Q_{\text{obs},i} - \overline{Q_{\text{obs}}}\right]^2}\right] \tag{10}$$

where $T_{\text{obs},j}$ is the time of occurrence of the observed peak flood in event j;$T_{\text{sim},j}$ is the time of occurrence of the simulated peak flood in event j; and $\overline{Q_{\text{obs}}}$ is the mean observed runoff. The other notation is as previously described.

For V_E, P_E and T_E (Eqs. (7)–(9)), a value of 0% corresponds to a perfect fit and values increase with decreasing performances. For R_{eff} (Eq. (10)), a value of 1 corresponds to a perfect fit. A R_{eff} of 0 means that the simulation is as accurate as the mean of the observed data, whereas $R_{\text{eff}} < 0$ means that the observed mean is a better predictor than the model.

3 Results

3.1 Model calibration using events above thresholds with a high and a low frequency of occurrence

Two sets of event combinations of three hydrographs were generated using events above thresholds with a low and a high frequency of occurrence – return periods of 2.33 and 1.01 years, respectively. These two sets were calibrated and then tested in validation assuming a rainfall scenario with a quality as good as that of the actually observed rainfall. The median values of model performances obtained in validation for the two sets were compared to answer our question, about whether events above a threshold with a short return period are useful in calibrating a hydrological model and providing reliable flood predictions.

Model performance results obtained for the two sets of event combinations (Fig. 2) showed relatively similar predictability of floods. Regardless of the frequency of occurrence of the thresholds, having three events improved predictability in comparison to the scenario of no available data. This was because the model performance of all performance measures, except for V_E, was typically above the lower benchmarks. On average, about 3% of the event combinations based on the threshold with a low frequency of occurrence returned model performances below the lower benchmarks, whereas this number increased to 8% for the event combinations based on the threshold with a high frequency of occurrence. Nevertheless, the differences between the two sets were relatively small. Results suggest that if the quality of the rainfall forecasts was as good as that of the observations, three events above a threshold with a short return period would be useful for calibration and, therefore, for predicting floods.

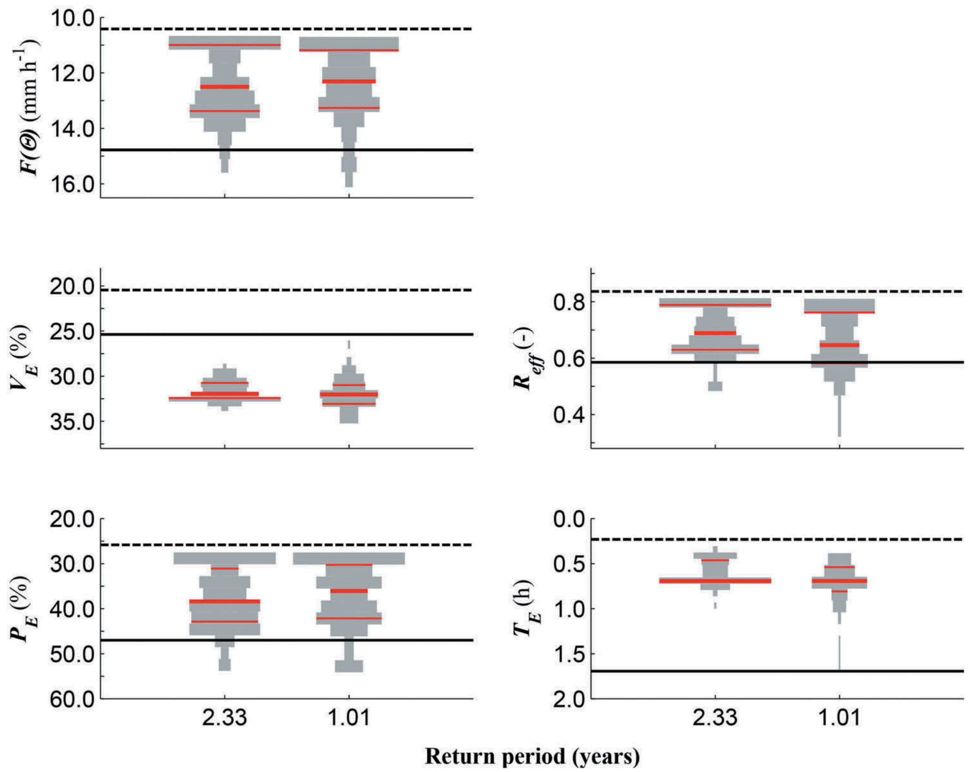

Figure 2. Histogram of median model performance in validation, visualized as violin plots, using parameters calibrated on three event hydrographs and using rainfall forcing data with a quality as good as that of real-time observations. The event combinations were based on events above thresholds with a low (return period: 2.33 years) and high frequency of occurrence (return period: 1.01 years). Black dashed lines represent the upper (optimal) benchmarks, and black solid lines represent the lower benchmarks. The red solid lines represent the 25th (top), 50th (middle) and 75th (bottom) percentiles of median model performance.

3.2 Effects on flood predictions caused by rainfall errors

The predictive ability of a model calibrated using event combinations of three hydrographs was further assessed using several rainfall scenarios. The event combinations in question were those generated with the hydrographs identified above the threshold with a short return period (i.e. 1.01 years). The MMV of model performances obtained for the different rainfall scenarios was evaluated to answer our question concerning the effects on flood predictions caused by rainfall errors, i.e. between volumetric and duration errors of the input-driving data, which of these have the most substantial impact on flood predictions?

Results for the $F(\theta)$ measure (Fig. 3) show that the predictive ability of the model was sensitive to both volume and duration errors of rainfall. Between the two rainfall errors, flood-predictions were highly sensitive to volume errors since model performance varied more with varying volume errors than with varying duration errors. The best $F(\theta)$ score was obtained when the duration error was 0% and volume error was – 12.5%. To some extent, similar results were obtained when assessing model performance in terms of peak-flood error and Nash-Sutcliffe efficiencie (Figs. 4 and 5). The best R_{eff} was found when the rainfall-duration error was 0%, as was found for the $F(\theta)$ score, but with a volume error of – 25%.

The sensitivity was considerably higher for positive volume and duration errors than for negative volume and duration errors. The sensitivity to rainfall-volume errors in the absence of duration errors was locally quantified. If one assumes a – 12.5% rainfall-volume error (i.e. associated with a 31% relative peak-flood error, the best performance obtained) an increase of up to +100% rainfall volume error leads to the highest relative peak-flood error (157%). This corresponds to an increase of peak-flood error of approximately 11% per 10% increase in rainfall-volume error (Fig. 4). On the other hand, a decrease from – 12.5% rainfall-volume error to – 50% rainfall-volume error (associated with a 46% relative peak-flood error) corresponds only to 4% increase in peak-flood error per 10% decrease in rainfall-volume error. Similarly, in the absence of volume errors, an increase from – 12.5% rainfall-duration error (associated with a 32% relative peak-flood error) to +100% rainfall-duration error (with a 64% relative peak-flood error) corresponds to 3% increase in peak-flood error per 10% increase in rainfall duration error. In contrast, a decrease from – 12.5% rainfall-duration error to – 50% rainfall-duration error (associated with a 37% relative peak-flood error) corresponds only to 1% increase in peak-flood error per 10% decrease in rainfall-duration error. In the same manner, the R_{eff} decreased notably for positive rainfall-volume errors (Fig. 5).

When assessing flood predictions in terms of relative time-to-peak errors, model performance was only sensitive to rainfall-duration errors and the best relative time-to-peak error was found when the rainfall-duration error was 0%. (Figure 6). With this respect, time-to-peak errors were more sensitive to negative than to positive duration errors. Relative time-to-peak errors increased about +0.5 h for an error of −10% in rainfall duration, whereas for an error of +10% in rainfall

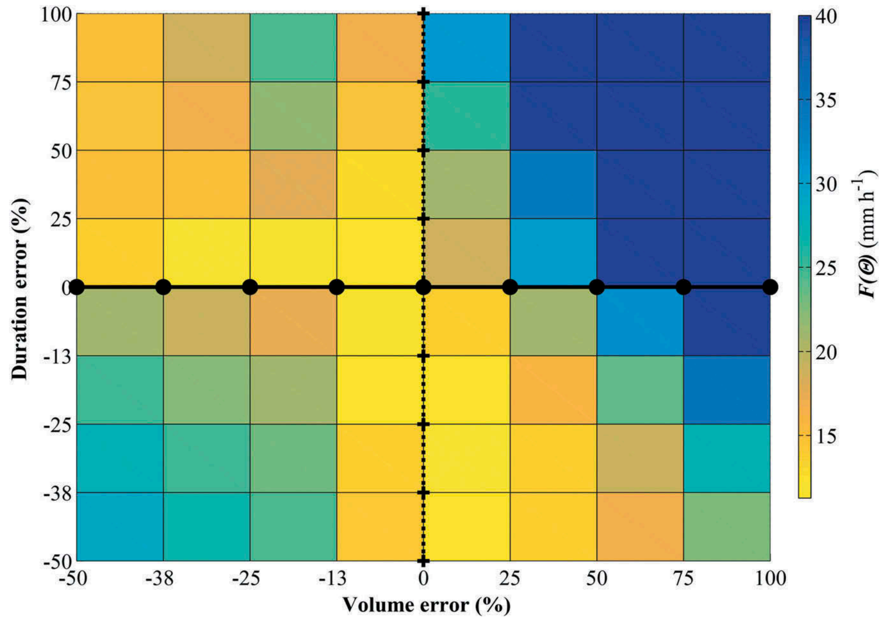

Figure 3. Two-dimensional view of the MMV of $F(\Theta)$ for several rainfall scenarios. A $F(\Theta)$ value of zero corresponds to a perfect fit and values increase with decreasing performance. The black solid line represents the vector of rainfall scenarios with only volume errors, and the black dotted line represents the vector with only duration errors.

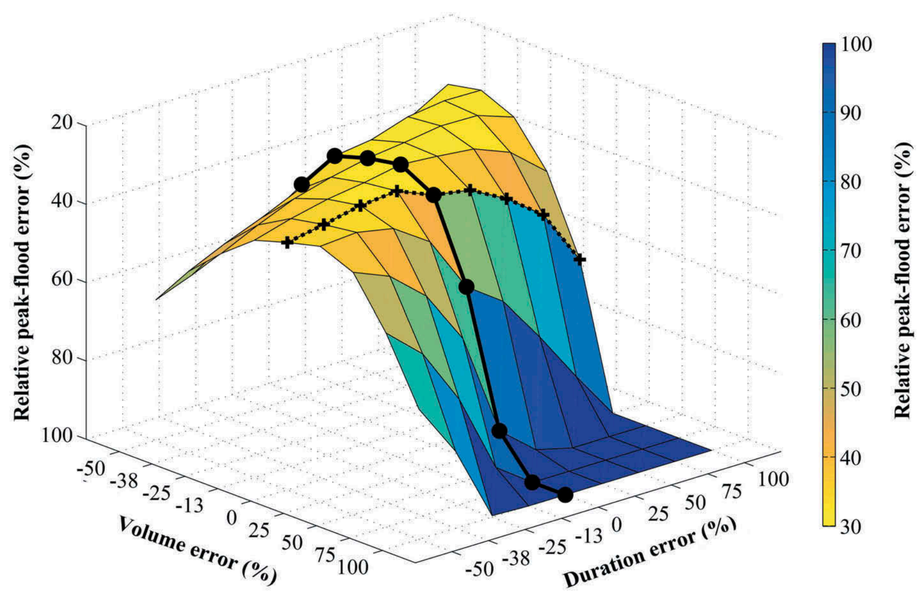

Figure 4. Three-dimensional view of the MMV of relative peak-flood errors (P_E) for several rainfall scenarios. A P_E value of 0% corresponds to a perfect fit and values increase with decreasing performance. The black solid line represents the vector of rainfall scenarios with only volume errors, and the black dotted line represents the vector with only duration errors.

duration, it only increased +0.1 h. On the contrary, when assessing flood predictions in terms of relative flood-volume errors, model performance was only sensitive to rainfall-volume errors (Fig. 7). Relative flood-volume errors were also more sensitive to positive than to negative rainfall-volume errors. An error of +10% in rainfall volume led to an increase of relative flood-volume error of 11%, whereas an error of – 10% in rainfall volume led to an increase of relative flood-volume error of only 6%. Surprisingly, the smallest flood-volume error (23%) was found for rainfall-volume errors of – 25%, as for R_{eff}.

4 Discussion

The application of operational flood-forecasting models in fast responding basins depends on the availability of discharge observations for calibration and on rainfall forecasts to be used as input. Lack of the former and uncertainty of the latter call for further studies on improving the robustness of flood-model calibration based on limited discharge data and on gaining a better understanding of the effects on flood predictions caused by rainfall errors. This study deals with the latter and it was investigated on a tropical basin, which was treated as

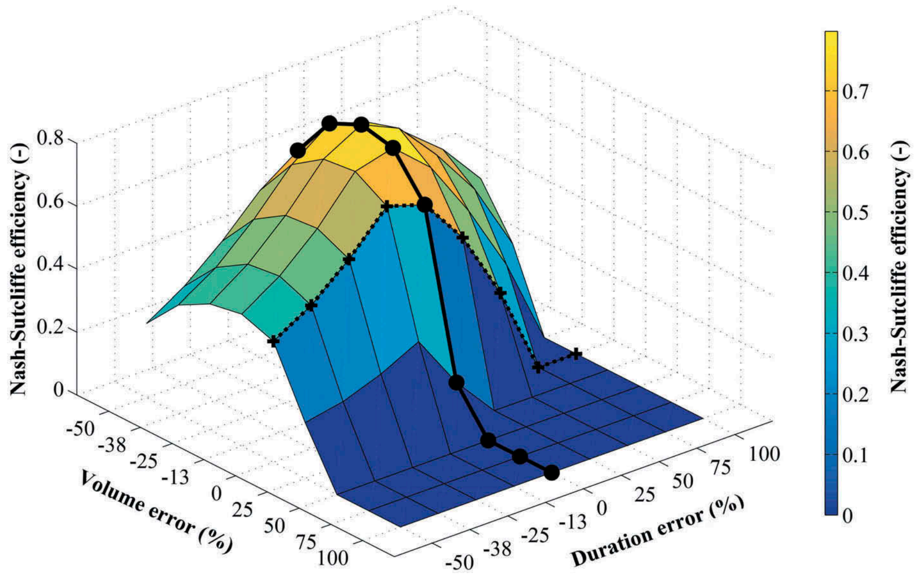

Figure 5. Three-dimensional view of the MMV of Nash-Sutcliffe (R_{eff}) efficiency for several rainfall scenarios. The black solid line represents the vector of rainfall scenarios with only rainfall-volume errors, and the black dotted line represents the vector with only rainfall-duration errors.

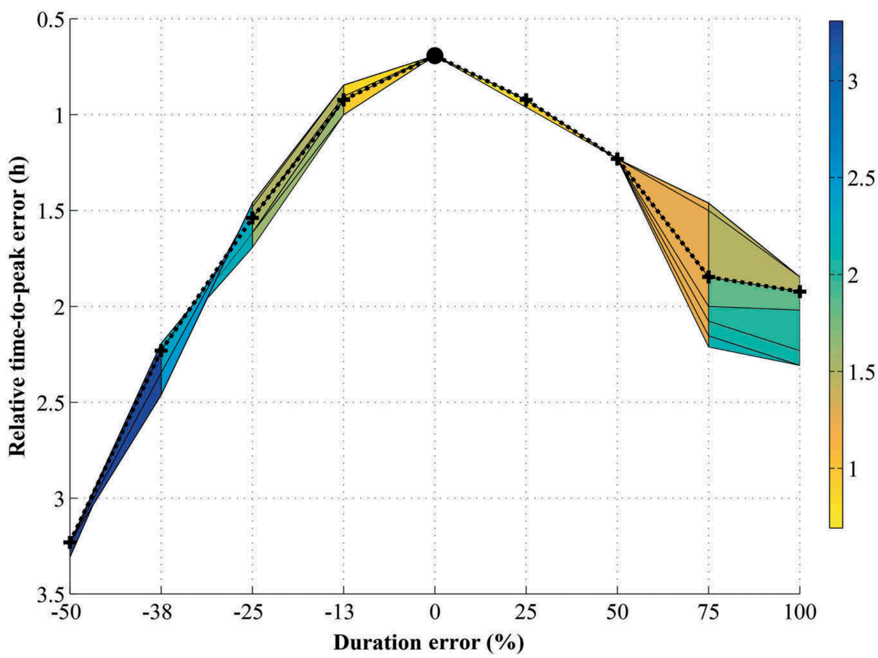

Figure 6. Median of the median values (MMV) of relative time-to-peak errors (T_E) for several rainfall scenarios. A T_E value of 0 h corresponds to a perfect fit and values increase with decreasing performance.

ungauged but had hourly rainfall–runoff data available adequate for the purpose of our study.

4.1 Can events above a threshold with a short return period be useful in calibrating a hydrological model and providing reliable flood predictions?

In a previous study (Reynolds *et al.* 2019), we showed that using few events in calibration improved flood predictions in comparison to the scenario of no data at all. The applicability of the latter study relies on the availability of hydrographs with a low frequency of occurrence (return period: 2.33 years). However, such hydrographs are challenging to gauge in practice during field campaigns. Hence, we further extend the methodology of the previous paper in this study by using a few events above two thresholds with different frequencies of occurrence for flood model calibration. The previous paper answered the question about how many high-flow events are needed for flood model calibration. The study presented in the current paper, on the other hand, explores if using few events above a threshold with a short return period could be useful for predicting floods when additionally uncertainties in rainfall

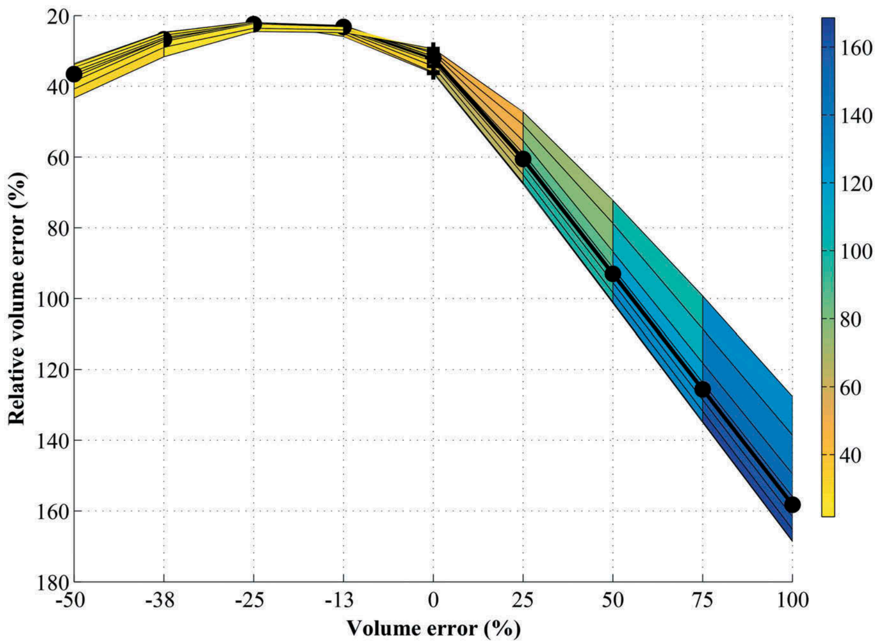

Figure 7. Median of the median values (MMV) of relative volume errors (V_E) for several rainfall scenarios. A V_E value of 0% corresponds to a perfect fit and values increase with decreasing performance.

data are accounted for (by using multiple scenarios of rainfall data with volume and duration errors).

Comparison of model-performance results in validation showed relatively similar flood-prediction skills for both thresholds, although the one found for the threshold with the highest frequency of occurrence resulted in a larger number of event combinations that were slightly less informative than when no data were available (i.e. cases when model performance was worse than the lower benchmark). Still, around 90% of event combinations based on the latter threshold showed to be equal or more informative than when no data were available and therefore, proved to be useful for flood-model calibration and prediction. This is promising for real flood-forecasting applications in data-scarce and budget-limited conditions, but at the same time, it raises the question of whether the return period of the threshold could be even shorter than those tested here and still be useful for flood predictions.

Model calibration was event-based but the model was run continuously to avoid the risk of assuming erroneous initial conditions before the occurrence of each event. For the threshold with the shortest return period, the median of the median values (MMV) for relative volume error (V_E) and relative peak-flow error (P_E) in calibration were 9% and 15% respectively, but they increased to 32% and 36% in validation. Flood predictions in validation were somewhat overestimated, which may explain the model performances found for those objective functions. This is further addressed in the next section where model performance was compared and discussed for multiple scenarios of rainfall data. Regardless of the threshold, all event combinations resulted in relative volume errors under the lower benchmark. The latter was not surprising, as a previous study by Tan *et al.* (2008) showed that event-based calibration performs better in predicting flood hydrograph, peak flow and time to peak than continuous-based calibration, whereas the latter performs better than the former in predicting flood-volume errors. Besides, the lower benchmark of relative volume error was relatively high and difficult to match.

It might be argued that the threshold with the shortest return period (125 $m^3\ s^{-1}$ or 5 mm h^{-1}) is extreme compared to average conditions (the 50th percentile of the time-series discharge data is about 4.68 $m^3\ s^{-1}$ or 0.19 mm h^{-1}). However, this threshold was only about 12% relative to the largest peak flood of the events selected for evaluation. This fraction is low compared to recommendations given in a comparable study of different objective functions by Jie *et al.* (2016). There, threshold values of peak flow between 40% and 70% were suggested as suitable for flood-model calibration of multi-objective functions. Overall, our results suggest that if rainfall forecast could be as good as that of real-time observations, flood-model calibration based on hydrographs of a few events above a threshold with a short return period could improve the predictability of floods. Furthermore, our findings support previous authors' findings that calibration data needs may be less if data are taken in an event-based way since they provide sufficient information to constrain model parameters (McIntyre and Wheater 2004, Seibert and McDonnell 2013, Reynolds *et al.* 2019).

4.2 Between volumetric and duration errors of the input driving data, which of these have the most substantial impact on flood predictions?

In tropical regions where rainfall is highly variable and influenced by orographic and convective conditions, rivers are characterized by fast runoff responses triggered by high rainfall intensities. Hydrological models coupled with rainfall forecasts can offer warnings of potential floods in basins with short concentration times. However, errors in the rainfall forecasts are still large and affect the capability of hydrological models for simulating floods accurately. Errors on flood predictions are mainly the result of rainfall-forecast errors associated with the average basin precipitation (Kobold 2007).

To assess the effects on flood predictions caused by rainfall errors, 81 rainfall scenarios based on real-time observations with errors in duration and volume were used as input to a model that was already calibrated on few events. Results showed the flood predictions were sensitive to both volume and duration errors of the rainfall. Furthermore, there seemed to be an interplay between the two errors since good modelling performances could be found for large errors of both, meaning that errors of one may compensate for the errors of the other. At the same time this interplay was found to be complex. It could be assumed that the model is already constrained because of using only hydrographs of a few events in calibration. However, it is not known to what extent this limitation could make the model more sensitive to rainfall errors, which needs to be further investigated.

In general, flood-model predictions showed to be more sensitive to rainfall-volume errors than to rainfall-duration errors. This was as expected because peak-flow errors, $F_3(\theta)$, which are not sensitive to the time to peak, were larger and had a bigger influence on the aggregate measure $F(\theta)$ than the other two objective functions (i.e. $F_1(\theta)$ and $F_2(\theta)$). The latter would also explain why the $F(\theta)$ surface is more similar to the surface of P_E than to the surface of R_{eff}. For some rainfall scenarios (i.e. positive rainfall-volume errors and negative rainfall-duration errors), it was found that $F(\theta)$ indicated a good performance which, however, was not seen in R_{eff} (Figs. 3 and 5). Similar findings were reported by Jie *et al.* (2016), who concluded that using larger thresholds of peak flows in calibration could contribute to a good performance of peak flows at the expense of simulating worse the global shape and volume of the flood hydrographs. Time-to-peak errors were only sensitive to rainfall-duration errors as expected. The latter function is a measure of how good was the timing of the simulated peak discharge against the one observed. The timing of the simulated peak discharge is controlled by the occurrence of intense rainfall, which was modified when adding rainfall duration errors.

The best model performances were found when the rainfall-duration error was 0% and, more surprisingly, when the rainfall-volume error was negative. The former finding came as no surprise as the hydrograph responses are expected to be directly correlated to real-time rainfall pulses. As for the latter finding, the predicted floods events in validation were overestimated for the rainfall scenario with no volume errors, which resulted in relatively high MMV of V_E and P_E. The runoff coefficient of the events for that scenario was about 87%, but it increased to 99% and 116% for the scenarios with volume errors of – 12.5% and – 25%. It is assumed that this behaviour may be caused by errors on the observed input data (perhaps not fully representative of the spatial rainfall variability during the occurrence of such extreme events) or by model-structure errors that overestimated the volume of water mobilized during high flow events which was then indirectly compensated by negative rainfall-volume errors.

Nash-Sutcliffe, peak-flood and volume error efficiencies were more sensitive to positive than to negative rainfall-volume errors. Similar results were reported by Xu *et al.* (2006) in studying the sensitivity of the NOPEX-6 model to precipitation errors; they concluded that overestimation of precipitation affects runoff simulations more than does underestimation. The fact that the three measures behave similarly is not surprising as they correlate when the focus is only on predicting floods. As previously noted, the flows in the events chosen in validation were overestimated for the rainfall scenario without volume errors. It is assumed that by adding positive rainfall-volume errors, this overestimation became larger which deteriorated model efficiencies considerably. The opposite occurred for the rainfall scenarios with negative volume errors as model efficiencies improved until specific percentages but then they decreased at a lesser degree. Another possible explanation could be that the percentages for positive rainfall-volume errors were larger than for negative errors which resulted in worse performance and in a seemingly higher sensitivity of the performance to positive rainfall-volume errors.

This study could have benefited from the availability of rainfall forecasts to assess the effects of their errors on flood predictions, but these were not available. In spite of this limitation, we took advantage of the availability of long time series of hourly rainfall data in our study basin to create multiple rainfall scenarios with errors considered commensurable to those in practical applications. The centre of mass of the rainfall events was used as a reference to increase or decrease the duration of the events for the different rainfall scenarios. Another reference could have been used (i.e. start of the rainfall event, peak rainfall), but it was considered that the one chosen was more of a robust measure than the others in rainfall-forecast applications.

Global models, such as the Global Flood Awareness System (GloFAS), were not considered in this study because their spatial and temporal resolution is not high enough to capture local and intense precipitation in regions with high rainfall variability (e.g. GloFAS has a horizontal grid resolution of about 32 km for 10 days, increasing to 65 km from Day 11 to Day 15) (Alfieri *et al.*, 2013). For regions with high rainfall variability and complex topography, accurate representations of rainfall in time and space at fine resolution then become essential for runoff modelling and flood forecasting. In the long run, the results found in this study are expected to promote more research studies aiming to improve the accuracy of rainfall forecasts with a focus on rainfall–runoff models and, ultimately, to obtain meaningful flood forecasts. More studies in this field are needed using rainfall forecasts to assess how wrong could these forecasts be and still be useful in operational applications, even with the inherent uncertainties in the modelling process and lack of long time series of discharge data for calibration.

5 Conclusions

The overall aim of our study was to assess the influence of rainfall errors on the performance of a hydrological model in providing meaningful predictions of floods when using parameters calibrated on limited discharge data. This was achieved in two steps. First, we explored if hydrographs above a threshold with a short return period can be useful in flood-model calibration and prediction. Second, we assessed the effects of rainfall errors on a flood model calibrated with limited discharge data. The specific conclusions from our analysis are given below:

(1) Model performance was relatively similar when using calibrated parameters based on a few events above thresholds with a high and a low frequency of occurrence, although a small increase of outliers was noticed for the former parameter sets.
(2) Calibrating a model with a limited number of events proved to be also useful for predicting floods given uncertain rainfall data.
(3) Between volume and duration errors of rainfall, the former affected model performance more.
(4) Good flood predictions could be achieved even with large rainfall errors in volume and duration because of the interplay between the two errors, which seem to compensate for each other.
(5) Runoff simulations and model performance were generally more sensitive to positive than to negative rainfall-volume errors.

This methodological study was limited to one basin, one model and one calibration method. Therefore, the generality of our results needs to be further tested.

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 valuable comments and suggestions that helped us improve the manuscript.

Author contributions

J.E. Reynolds chose the experimental design, run the model simulations and he was the main responsible for writing the manuscript. S. Halldin, J. Seibert, C.Y. Xu and T. Grabs provided advice for designing the experiment, took part in the formulation of the research questions as well as in the interpretation and analysis of the results, and contributed significantly writing and preparing the manuscript for publication.

Disclosure statement

The authors declare no conflicts of interest.

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

Alfieri, L., *et al.*, 2013. GloFAS-global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, 17 (3), 1161–1175. doi:10.5194/hess-17-1161-2013
Bartholmes, J. and Todini, E., 2005. Coupling meteorological and hydrological models for flood forecasting. *Hydrology and Earth System Sciences Discussions*, 9 (4), 333–346. doi:10.5194/hess-9-333-2005
Bastola, S. and Murphy, C., 2013. Sensitivity of the performance of a conceptual rainfall–runoff model to the temporal sampling of calibration data. *Hydrology Research*, 44 (3), 484–493. doi:10.2166/nh.2012.061
Bergström, S., 1976. *Development and application of a conceptual runoff model for Scandinavian catchments*. Norrköping, Sweden: SMHI, Report No. RHO 7.
Beven, K., 2001. *Rainfall–runoff modelling: the primer*. Chichester, UK: Wiley.
Beven, K., 2009. *Environmental modelling: an uncertain future?* Abingdon-on-Thames, England, UK: Routledge.
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
Buizza, R. and Hollingsworth, A., 2002. Storm prediction over Europe using the ECMWF ensemble prediction system. *Meteorological Applications*, 9 (3), 289–305. doi:10.1017/S1350482702003031
Buizza, R., *et al.*, 1999. Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Weather and Forecasting*, 14 (2), 168–189. doi:10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2
Chintalapudi, S., Sharif, H.O., and Xie, H., 2014. Sensitivity of distributed hydrologic simulations to ground and satellite based rainfall products. *Water (Switzerland)*, 6 (5), 1221–1245. doi:10.3390/w6051221
Cloke, H.L. and Pappenberger, F., 2008. Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. *Meteorological Applications*, 15 (1), 181–197. doi:10.1002/met.58
Dettinger, M.D. and Diaz, H.F., 2000. Global characteristics of stream flow seasonality and variability. *Journal of Hydrometeorology*, 1, 289–310. doi:10.1175/1525-7541(2000)001<0289:GCOSFS>2.0.CO;2
De Roo, A.P.J., *et al.*, 2003. Development of a european flood forecasting system. *International Journal of River Basin Management*, 1 (1), 49–59. doi:10.1080/15715124.2003.9635192
Fernandez, W., Vogel, R.M., and Sankarasubramanian, A., 2000. Regional calibration of a watershed model. *Hydrological Sciences Journal*, 45 (5), 689–707. Taylor and Francis Group. doi:10.1080/02626660009492371
Ferraris, L., Rudari, R., and Siccardi, F., 2002. The uncertainty in the prediction of flash floods in the Northern Mediterranean environment. *Journal of Hydrometeorology*, 3 (6), 714–727. doi:10.1175/1525-7541-(2002)003<0714:TUITPO>2.0.CO;2
Goswami, M., O'Connor, K.M., and Bhattarai, K.P., 2007. Development of regionalisation procedures using a multi-model approach for flow simulation in an ungauged catchment. *Journal of Hydrology*, 333 (2–4), 517–531. doi:10.1016/j.jhydrol.2006.09.018
Häggström, M., *et al.*, 1990. *Application of the HBV model for flood forecasting in six Central American rivers*. Norrköping, Sweden: SMHI Hydrology, No. 27.
He, Y., *et al.*, 2009. Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. *Meteorological Applications*, 16 (1), 91–101. doi:10.1002/met.132
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
Kingston, D.G., *et al.*, 2011. Regional classification, variability, and trends of northern North Atlantic river flow. *Hydrological Processes*, 25 (7), 1021–1033. doi:10.1002/hyp.7655

Kobold, M. and Sušelj, K., 2005. Precipitation forecasts and their uncertainty as input into hydrological models. *Hydrology and Earth System Sciences*, 9 (4), 322–332. doi:10.5194/hess-9-322-2005
Kobold, M., 2007. Hydrological modelling of floods. *In*: *Advances in urban flood management*. Boca Raton, FL: CRC Press, 433–452.
Kokkonen, T.S., *et al*., 2003. Predicting daily flows in ungauged catchments: model regionalization from catchment descriptors at the coweeta hydrologic laboratory, North Carolina. *Hydrological Processes*, 17 (11), 2219–2238. John Wiley and Sons, Ltd. doi:10.1002/HYP.1329
Linsley, R.K., 1967. The relation between rainfall and runoff: review paper. *Journal of Hydrology*, 5, 297–311. North-Holland Publishing Co., Amsterdam. doi:10.1016/S0022-1694(67)80128-8
Littlewood, I.G. and Croke, B.F.W., 2008. Data time-step dependency of conceptual rainfall—streamflow model parameters: an empirical study with implications for regionalisation. *Hydrological Sciences Journal*, 53 (4), 685–695. doi:10.1623/hysj.53.4.685
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
Merz, R. and Blöschl, G., 2004. Regionalisation of catchment model parameters. *Journal of Hydrology*, 287 (1), 95–123. doi:10.1016/j.jhydrol.2003.09.028
Chahinian, N. and Moussa, R., 2009. Comparison of different multi-objective calibration criteria of a conceptual rainfall–runoff model of flood events. *Hydrology and Earth System Sciences Discussions*, 13 (4), 519–535. doi:10.5194/hess-13-519-2009
Osuch, M., Wawrzyniak, T., and Nawrot, A., 2019. Diagnosis of the hydrology of a small Arctic permafrost catchment using HBV conceptual rainfall–runoff model. *Hydrology Research*, 50 (2), 459–478. doi:10.2166/nh.2019.031
Parajka, J., *et al*., 2013. Prediction of runoff hydrographs in ungauged basins. *In*: G. Blöschl, *et al*., ed. *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge, UK: Cambridge University Press, 227–269. doi:10.1017/CBO9781139235761.013
Parajka, J., *et al*., 2010. Seasonal characteristics of flood regimes across the Alpine-Carpathian range. *Journal of Hydrology*, 394 (1–2), 78–89. doi:10.1016/j.jhydrol.2010.05.015
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*., 2019. Robustness of flood-model calibration using single and multiple events. *Hydrological Sciences Journal*, 1–12. doi:10.1080/02626667.2019.1609682
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
Santos, L., Thirel, G., and Perrin, C., 2018. Continuous state-space representation of a bucket-type rainfall–runoff model: a case study with the GR4 model using state-space GR4 (version 1.0). *Geoscientific Model Development*, 11 (4), 1591–1605. doi:10.5194/gmd-11-1591-2018
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., 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., *et al*., 2018. Upper and lower benchmarks in hydrological modeling. *Hydrological Processes*, 32 (8), 1120–1125. doi:10.1002/hyp.11476
Singh, V.P., 1997. Effect of spatial and temporal variability in rainfall and watershed characteristics on stream flow hydrograph. *Hydrological Processes*, 11 (12), 1649–1669. doi:10.1002/(SICI)1099-1085(19971015)11:12<1649::AID-HYP495>3.0.CO;2-1
Szolgay, J., *et al*., 2003. Regional estimation of parameters of a monthly water balance model. *Journal of Hydrology and Hydromechanics*, 51, 256–273.
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)
Verbunt, M., *et al*., 2006. Verification of a coupled hydrometeorological modelling approach for alpine tributaries in the Rhine basin. *Journal of Hydrology*, 324 (1–4), 224–238. doi:10.1016/j.jhydrol.2005.09.036
Wang, L., Meerveld, I.V., and Seibert, J., 2017. When should stream water be sampled to be most informative for event-based, multi-criteria model calibration? *Hydrology Research*, 48 (6), 1566–1584. doi:10.2166/nh.2017.197
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 and Planetary Change*, 172, 211–222. doi:10.1016/J.GLOPLACHA.2018.09.017
Wetterhall, F., *et al*., 2011. Effects of temporal resolution of input precipitation on the performance of hydrological forecasting. *Advances in Geosciences*, 29, 21–25. doi:10.5194/adgeo-29-21-2011
Xu, C.-Y., 1999. Estimation of parameters of a conceptual water balance model for ungauged catchments. *Water Resources Management*, 13 (5), 353–368. doi:10.1023/A:1008191517801
Xu, C.Y., *et al*., 2006. Evaluation of seasonal and spatial variations of lumped water balance model sensitivity to precipitation data errors. *Journal of Hydrology*, 324 (1–4), 80–93. doi:10.1016/j.jhydrol.2005.09.019
Xuan, Y., Cluckie, I.D., and Wang, Y., 2009. Uncertainty analysis of hydrological ensemble forecasts in a distributed model utilising short-range rainfall prediction. *Hydrology and Earth System Sciences*, 13 (3), 293–303. doi:10.5194/hess-13-293-2009
Yang, X., *et al*., 2019a. Dependence of regionalization methods on the complexity of hydrological models in multiple climatic regions. *Journal of Hydrology*. doi:10.1016/j.jhydrol.2019.124357
Yang, X., *et al*., 2018. Runoff prediction in ungauged catchments in Norway: comparison of regionalization approaches. *Hydrology Research*, 49 (2), 487–505. doi:10.2166/nh.2017.071
Yang, X., Magnusson, J., and Xu, C.-Y., 2019b. Transferability of regionalization methods under changing climate. *Journal of Hydrology*, 568, 67–81. doi:10.1016/j.jhydrol.2018.10.030

Appendix.

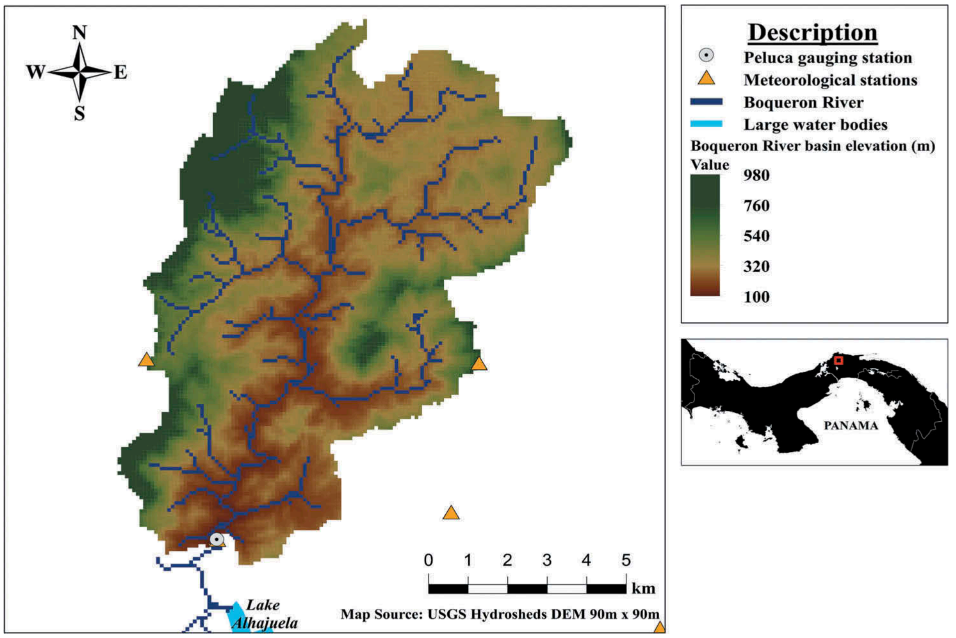

Figure A1. Location of Boqueron River Basin at Peluca in Panama. Reprinted from "Sub-daily runoff predictions using parameters calibrated on the basis of data with a daily temporal resolution" by Reynolds, J. E., Halldin, S., Xu, C. Y., Seibert, J. and Kauffeldt, A., 2017, *J. Hydrol*. 550, 399–411. doi:10.1016/j.jhydrol.2017.05.012. Copyright 2017 Elsevier. Reprinted with permission.