Use of near-real-time satellite precipitation data and machine learning to improve extreme runoff modeling

Paul Muñoz[1], Gerald Augusto Corzo Perez[2], Dimitri Solomatine[3], Jan Feyen[4], and Rolando Célleri[1]

[1]Universidad de Cuenca
[2]UNESCO-IHE Institute for Water Education
[3]UNESCO-IHE Delft Institute for Water Education
[4]Katholieke Universiteit Leuven

November 22, 2022

Abstract

Extreme runoff modeling is hindered by the lack of sufficient and relevant ground information and the low reliability of physically-based models. The authors propose to combine precipitation Remote Sensing (RS) products, Machine Learning (ML) modeling, and hydrometeorological knowledge to improve extreme runoff modeling. The approach applied to improve the representation of precipitation is the object-based Connected Component Analysis (CCA), a method that enables classifying and associating precipitation with extreme runoff events. Random Forest (RF) is employed as a ML model. We used 2.5 years of nearly-real-time hourly RS precipitation from the PERSIANN-CCS and IMERG-early run databases (spatial resolutions of 0.04 o and 0.1 o , respectively), and runoff at the outlet of a 3391 km 2-basin located in the tropical Andes of Ecuador. The developed models show the ability to simulate extreme runoff for the cases of long-duration precipitation events regardless of the spatial extent, obtaining Nash-Sutcliffe efficiencies (NSE) above 0.72. On the contrary, we found an unacceptable model performance for a combination of short-duration and spatially-extensive precipitation events. The strengths/weaknesses of the developed ML models are attributed to the ability/difficulty to represents complex precipitation-runoff responses.

Use of near-real-time satellite precipitation data and machine learning to improve extreme runoff modeling.

Paul Muñoz[1,2], Gerald Corzo[3], Dimitri Solomatine[3], Jan Feyen[4], and Rolando Célleri[1,2].

[1] Departamento de Recursos Hídricos y Ciencias Ambientales, Universidad de Cuenca, Cuenca 010150, Ecuador.
[2] Facultad de Ingeniería, Universidad de Cuenca, Cuenca 010150, Ecuador.
[3] Hydroinformatics Chair Group, IHE Delft Institute for Water Education, 2611 AX Delft, The Netherlands.
[4] Faculty of Bioscience Engineering, Catholic University of Leuven, Leuven 3001, Belgium.

Abstract

Extreme runoff modeling is hindered by the lack of sufficient and relevant ground information and the low reliability of physically-based models. The authors propose to combine precipitation Remote Sensing (RS) products, Machine Learning (ML) modeling, and hydrometeorological knowledge to improve extreme runoff modeling. The approach applied to improve the representation of precipitation is the object-based Connected Component Analysis (CCA), a method that enables classifying and associating precipitation with extreme runoff events. Random Forest (RF) is employed as a ML model. We used 2.5 years of nearly-real-time hourly RS precipitation from the PERSIANN-CCS and IMERG-early run databases (spatial resolutions of 0.04° and 0.1°, respectively), and runoff at the outlet of a 3391 km^2-basin located in the tropical Andes of Ecuador. The developed models show the ability to simulate extreme runoff for the cases of long-duration precipitation events regardless of the spatial extent, obtaining Nash-Sutcliffe efficiencies (NSE) above 0.72. On the contrary, we found an unacceptable model performance for a combination of short-duration and spatially-extensive precipitation events. The strengths/weaknesses of the developed ML models are attributed to the ability/difficulty to represents complex precipitation-runoff responses.

Keywords: Extreme runoff; Machine Learning; PERSIANN-CCS; IMERG-early-run; Feature Engineering; Tropical Andes.

1. Introduction

Physically-based precipitation-runoff models used in water management describe the physical processes that occur in a system (basin) by using balance and conservation equations (Clark et al., 2017). However, those models demand extensive data and might be subject to overparameterization, limiting its operational value (Mosavi et al., 2018; Young, 2002). As a solution, during the last decades, a data-driven approach, using Machine Learning (ML) techniques, gained popularity among hydrologists (Bontempi et al., 2012; Chang et al., 2019; Galelli and Castelletti, 2013; Mosavi et al., 2018). An important reason responsible for the increasing interest is the fact that ML exploits the available and relevant information (e.g., precipitation, past runoff) to find relations to the target variable (i.e., runoff) without requiring knowledge about the underlying physical processes. Among ML techniques, the Random Forest (RF) algorithm is mostly used for normal and extreme runoff modeling due to its simplicity (few parameters to calibrate), higher accuracy when compared to other ML techniques, the robustness of the model, and its capacity to deal with small size samples and complex data structures (Biau and Scornet, 2016; Breiman, 2001; Contreras et al., 2021; Li et al., 2016a; Li et al., 2020; Muñoz et al., 2018, 2021; Orellana-Alvear et al., 2020; Papacharalampous and Tyralis, 2018; Tyralis et al., 2019; Wang et al., 2015).

In terms of data availability, in many regions, ground precipitation networks are either inexistent or scare, and if available mostly with extremely low areal density. This is especially true for mountainous regions, such as the tropical Andes, where the remoteness of the study areas and budget constraints limits the development of accurate precipitation-runoff models. Fortunately, continuous development of Remote Sensing (RS) products, e.g., space-based satellites have dramatically enhanced the quantity (spatiotemporal resolution) and quality of areal precipitation observations. However, RS precipitation obtained from a single sensor (satellite) hardly provides accurate estimations (Hong et al., 2019). This has stimulated the development of multi-satellite precipitation products such as the NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) (Huffman et al., 2015), and the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) (Hsu et al., 1997). IMERG and PERSIANN products are characterized by quasi-global coverage, free access, high spatiotemporal resolutions, and in continuous development (Tang et al., 2016). Given previous, both RS products are nowadays widely used in hydrometeorological applications including tracking of precipitation anomalies (Nguyen et al., 2014; Sakib et al., 2021), precipitation early-warning systems (Sorooshian et al., 2014), and flood forecasting and mapping (Belabid et al., 2019; Nguyen et al., 2015).

Once the issue of data availability is solved the arising research questions are: a) is precipitation well represented by RS data? and b) can RS precipitation be properly assimilated during the learning process (in the context of extreme runoff ML models)? The first research question is mandatory for the cases when the interest lies in providing accurate precipitation estimations. This can be achieved by validating RS products with ground precipitation estimations, see for instance the studies of Laverde-Barajas et al. (2019) and (Li et al., 2016b). While, the second issue can be addressed regardless the validation of the precipitation, for the cases when precipitation is merely an estimator for the modeling of another variable (e.g., precipitation is an estimator in precipitation-runoff models). In this case, the methodology consists of applying a feature engineering strategy to RS precipitation data enabling a better ML precipitation assimilation during the learning process, improving ultimately extreme runoff model efficiencies. In addition, ML precipitation assimilation can be improved by building runoff models able to discriminate between different precipitation event types (Laverde-Barajas et al., 2020). This is because different precipitation events produce different runoff responses as a result of various runoff generation processes, mainly infiltration and saturation excess (Gutiérrez-Jurado et al., 2019).

Precipitation events can be distinguished by applying object-based methods to RS imagery (Davis et al., 2006; Laverde-Barajas et al., 2019; Li et al., 2016b; Peña-Barragán et al., 2011; Vogels et al., 2020). A simple yet effective object-based method is the Connected Component Analysis (CCA) employed by Laverde-Barajas et al. (2019). The CCA includes a physical description of precipitation events (centroid, extension area, etc.), as well as key meteorological attributes (intensity, duration, volume, etc.). These characteristics are then used for classifying precipitation events which can be contrasted with their associated runoff responses.

In this context, the objective of this study was to develop specialized (smart) ML extreme runoff models for a 3393-km^2 basin in Ecuador. We used a feature engineering methodology to improve the areal representation of the precipitation and to maximize runoff model efficiencies by identifying and classifying precipitation events associated with extreme hydrological events.

1. Study area and Dataset

1.1 Study area

The Jubones Basin, located in the tropical Andes of Ecuador, was selected as the study area (Figure 1) and covers an area of ~3391 km^2 upstream of the Minas-San Francisco hydroelectric dam, with an elevation ranging between 1250 to 3920 m above sea level. The climatology of the basin is extremely variable due to the presence of the Andean mountain range, trade winds, and ocean currents from the Pacific Ocean. A distinction can be made between at least 4 rainfall regions, including a semi-arid region. The basin climate ranges from humid to arid, with average annual rainfall ranging spatially from 350 to 1170 mm.

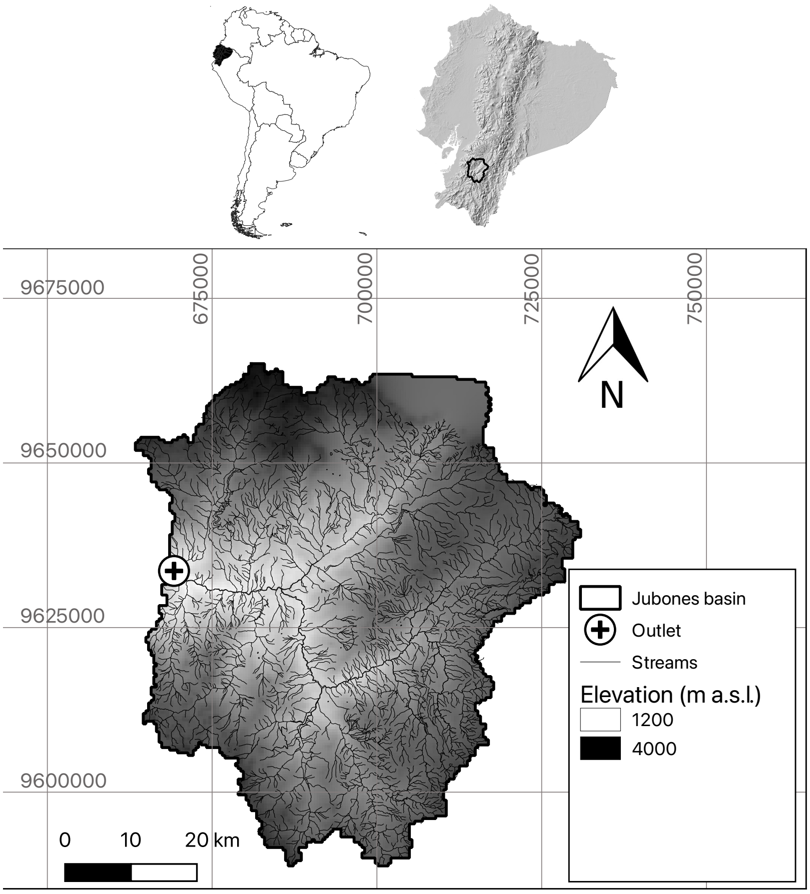


Figure 1. The Jubones basin in the Tropical Andes of Ecuador, South America.

1.2 Dataset

The dataset comprises hourly satellite-derived precipitation covering the Jubones basin, and hourly runoff data collected at the hydrological station, situated in the outlet of the basin, consisting of the Minas-San Francisco hydropower dam. Since the dam was completed in 2018, lasted the study period ~2.5 years, from 18 November 2018 to the 31st of March 2021.

Precipitation from Remote Sensing (RS) products

Precipitation information was retrieved from two near-real-time multi-satellite sources, the IMERG-early run, and the PERSIANN-Cloud Classification System (CCS) precipitation subproducts. Data were derived at hourly intervals. The main difference between both precipitation sources is the spatial resolution. The PERSIANN-CCS possesses the highest spatial resolution for the region (0.04° x 0.04°), being the result of infrared imagery processing and cloud classification using artificial neuronal networks (Hong et al., 2004). Whereas the IMERG-early run approach interpolates various microwave precipitation estimates and delivers data with a spatial resolution of 0.1° x 0.1°.

Figure 2 shows the mean annual precipitation of the Jubones basin, measured by the PERSIANN-CCS (728.5 mm) and IMERG-early run (727.2 mm) precipitation satellite subproducts. For this plot, we used hourly information for 2019 and 2020 (see also the precipitation plot in Figure 3). We found differences of 1.3 and 116 mm between the mean and the maximum annual precipitation obtained from the PERSIANN-CCS and the IMERG-early run subproducts, respectively. This is attributed to the spatial resolution difference and to the measurement principle of each satellite subproduct.

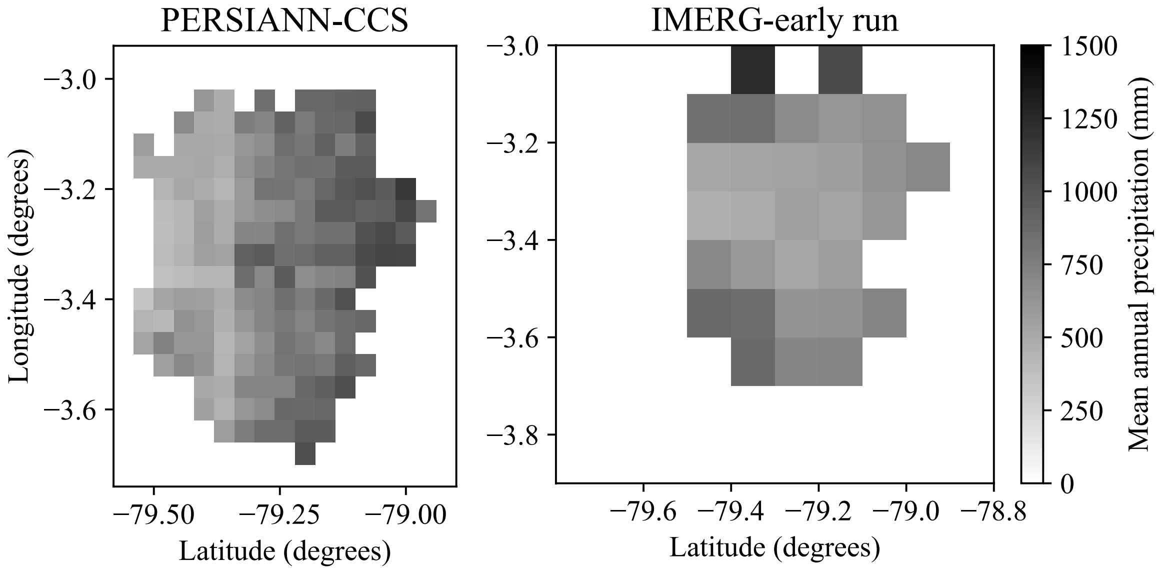


Figure 2. Mean annual precipitation in mm (2019 and 2020) measured by the PERSIANN-CCS and the IMERG-early run precipitation satellite subproducts.

Runoff at the entrance of the MSF hydropower dam

Hourly time series of runoff at the outlet of the Jubones basin were derived from the server of the Corporación Eléctrica del Ecuador (CELEC EP, https://www.celec.gob.ec/), the company that manages the Minas-San Francisco hydropower dam. Figure 3 depicts the runoff information for the study period. Figure 3a shows the hourly time series, whereas Figure 3b the corresponding probability of exceedance from which 55 nearly-independent peak flow events were selected based on peak-over-threshold values (red dots in Figure 3a). The peak flow events selection was done using the WETSPRO tool (Willems, 2009). Exceedance probability analysis reveals that for the study period the runoff magnitudes of 103.5 and 159.4 $m^3.s^{-1}$ are exceeded with probabilities of 10% and 5%, respectively. These probabilities, which correspond to the 90 and 95% quartiles, served to determine extreme hydrological runoff events for the development of the extreme precipitation-runoff models.

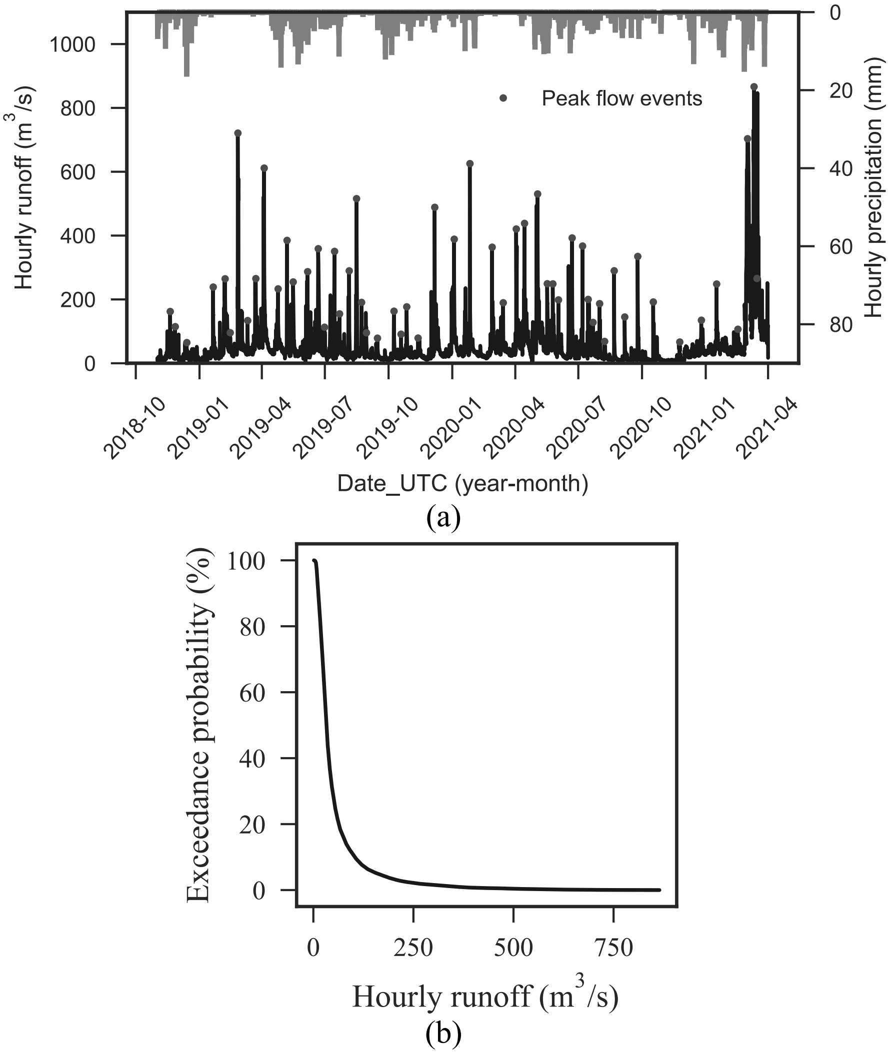


Figure 2. (a) Runoff and precipitation (PERSIANN-CCS) time series at the outlet of the Jubones basin. Peak flow events are displayed as red dots. (b) Exceedance probability for the study period (18/11/2018 to 31/03/2021).

2. Methodology

2.1 Determination of nearly independent peak hydrological events

The extreme hydrological events from the complete runoff time series were derived by applying the following two criteria: i) extreme hydrological events must exceed the 90% quartile values (98.8 $m^3.s^{-1}$), and ii) such events must be nearly independent. For meeting both criteria of independence, we used the WETSPRO time series tool (Willems, 2009), which splits runoff series in nearly independent peak and low flow events following a peak-over-threshold approach. The WETSPRO has two parameters to be calibrated, the inter-event time and peak height. In summary, we selected extreme hydrological events with a definition of independence controlled by the recession time and peak height difference of two consecutive runoff events.

2.2 Object-based Connected Component Analysis

Once extreme hydrological events were selected, the next step was to analyze their correspondent precipitation imagery from the highest-resolution satellite subproduct, the PERSIANN-CCS. The precipitation analysis was done by applying a feature engineering strategy based on an object-based Connected Component Analysis (CCA) algorithm. The CCA algorithm is fully detailed in Laverde-Barajas et al. (2019). We implemented the CCS algorithm through the scikit-image processing package in Python® version 3.7 (der Walt et al., 2014). The approach consists of the following steps (see also Figure 4):

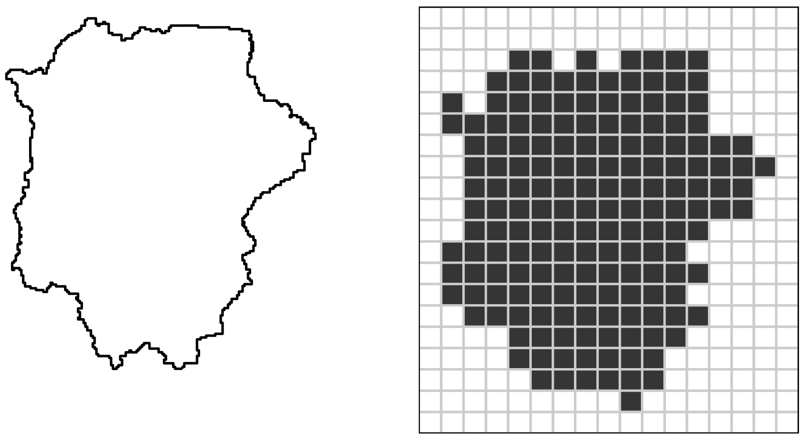
i. Clipping of the precipitation imagery to the Jubones basin (Figure 4a).

ii. Identification and localization of precipitation objects (latitude, longitude, see Figure 4b). For this we defined a precipitation threshold volume of 0.5 mm, i.e., precipitation objects with an associated precipitation volume of less than 0.5 mm are trimmed-off. This was done on a trial-and-error basis validated with precipitation objects observed in randomly selected precipitation events. The target was to remove noise from precipitation imagery and keep only clear precipitation objects in the precipitation imagery (Figure 4c).

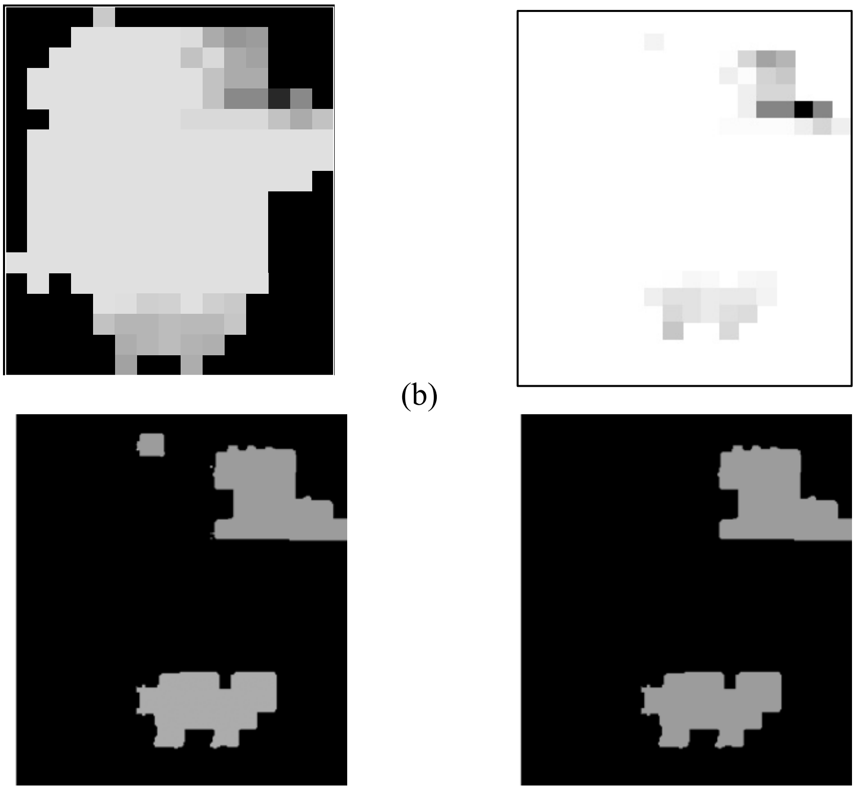
iii. Filtering of the identified precipitation objects according to size criteria. Similarly, we found and used a number-of-pixels threshold of 6, according to a trial-and-error procedure with the same target as employed in step (ii).

iv. Morphologically the identified and filtered precipitation objects were closed, by . applying as a final procedure a dilatation-and-erosion algorithm for refining precipitation objects (Figure 4d).

v. Retrieval of physical (centroid and extension area) and hydrometeorological attributes (volume of precipitation, maximum intensity, precipitation duration) from the precipitation objects defined in step (iv). For the duration of the precipitation, we defined that two precipitation objects are considered consecutive (i.e., belong to the same event) when the time between their appearance is shorter than 2 hours. This threshold was calibrated on a trial-and-error basis.



(a)



(b)

(c) (d)

Figure 3. Precipitation identification with an object-based Connected Component Analysis Illustration of the PERSIAN-CCS 2021-12-25 05:00 UTC image. (a) Jubones basin clipping, (b) Precipitation identification in mm from PERSIANN-CCS imagery, (c) Initial identification of 7 precipitation objects (different colors) with CCA analysis, and (d) Selection of 2 precipitation objects according to object size filtering and morphological closing.

Additionally, a modular precipitation approach for the analysis of the precipitation imagery was used. For the cases when no precipitation is observed by the PERSIANN-CCS subproduct, we switched the precipitation data source to IMERG-early imagery, following a simple spatially under-sampling technique. This means that an IMERG-early run cell of size 0.1x0.1^{o} was directly divided into ~6.4 cells with a resolution of 0.04x0.04^{o}, matching the resolution of the PERSIANN-CCS subproduct. This modular approach assures that all extreme hydrological events are trained with an existent precipitation signal, reducing noise and improving the learning process of the further developed runoff models.

2.3 Classification of precipitation events associated with extreme hydrologic events

The hydrometeorological attributes derived from the CCA analysis are used to classify precipitation events together with their associated runoff response. For this we used the following two criteria, respectively the extension of the precipitation objects (local and spatially extensive), and the duration of the precipitation events (short and long). As a result, by defining extension and duration thresholds we could establish four precipitation event classes: i) Local and short extreme events (LSE), ii) Local and long-duration extreme events (LLE), iii) Spatially extensive extreme events (SEE), and iv) Spatially extensive and long-duration extreme events (SLE).

2.4 Event-based runoff modeling

We developed one runoff model for each precipitation event class and one model without precipitation discrimination (base model). For this, we used the ML technique known as Random Forest (RF) for regression. The RF is described in the following subsection. Moreover, the input feature space to each model was formed with hourly precipitation and runoff, as well as an indicator of the belonging precipitation class. In addition to current-time precipitation and runoff information, we used past lag information which is determined according to statistical correlation analyses: partial- and auto-correlation functions for runoff, and cross-correlation function for precipitation. The construction of the input feature space was conducted following the methodology developed in Muñoz et al. (2018), with the purpose to add only relevant information to the models and improve their efficiencies.

3.4.1 Random Forest for regression

Random Forest (RF) is a ML technique of supervised learning where the main idea is to build multiple decorrelated trees (models), in which the input feature space is related to output(s) by successively applying a set of hierarchically organized conditions (Breiman, 2001). The key to the RF algorithm is the random selection of resampled datasets from the input feature space (bagging technique), which assures decorrelation between stochastically formed models.

We implemented the RF runoff models through the scikit-learn package for ML in Python® version 3.7 (Pedregosa et al., 2011). A full explanation of the RF algorithm can be found in Breiman (2001), and can be summarized as follows:

i. Construction of each decision tree by randomly selecting several bootstrap samples from the input feature space. A process known as out-of-bag (OOB) is used for forming each bootstrap with roughly two-thirds of the input feature space. On one hand, the OOB process serves to obtain unbiased estimates of the regression, and on the other hand, it allows to estimate the importance of each feature (predictor) of the feature space in the tree construction process.

ii. Optimally splitting of the data selected in step (i) at each node of each tree. This is done by determining a maximum number of features to perform the best split from the total number of predictors in the feature space. This also avoids overfitting by assuring variety and nonexistence of duplicated models.

iii. Growth of all the trees constructed in step (i) with the splits defined in step (ii) up to a size defined either by a maximum depth parameter or a minimum number of samples expected in the final node. Control of the depth of the trees aims to reduce the structural complexity of the models, leading to model parsimony and noise reduction.

iv. Determination of the output of the model as the mean response from all regression trees.

According to Contreras et al. (2021), the most-influencing RF hyperparameters for hydrological forecasting applications are the number of trees in the forest (n_trees), the maximum number of features to perform the splits of the data (max_features), and maximum depth for pruning purposes (max_depth). For all runoff models, we determined the optimal combinations of hyperparameter following a random grid-search procedure implemented with a 10-fold cross-validation process to prevent overfitting. The measure of agreement was evaluated according to the coefficient of determination (R^2) between simulations and observations for the training subsets. Table 1 presents the domain of the selected hyperparameters which forms the search space for the optimization task.

Table 1. Search space (grid) of the RF runoff models.

Hyperparameter	Domain
n_trees*	40;800;10*
max_features	n_features,n_features$^{(1/2)}$, $\log_2$(n_features)
max_depth*	40;800;10*

* Domain defined by min, max, and increment.

3.4.2 Model evaluation

We used four goodness-of-fit metrics for evaluating the efficiencies of the four runoff models. The Nash-Sutcliffe Efficiency (NSE) coefficient was set as the reference for measuring and comparing the overall model accuracy. To complement the analysis, we relied on the Kling-Gupta Efficiency (KGE), the Percent Bias ($PBIAS$), and the Root Mean Square Error ($RMSE$) metrics. The following equations were used:

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(Q_s(i) - Q_o(i)\right)^2}{\sum_{i=1}^{n}\left(Q_o(i) - \overline{Q_o}\right)^2}$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

$$PBIAS = \frac{\sum_{i=1}^{n}(Q_o - Q_s)}{\sum_{i=1}^{n} Q_o}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_s - Q_o)^2}$$

where n is the number of instances, Q_s is the simulated runoff, Q_o is observed runoff, $\overline{Q_o}$ is the mean observed runoff, $\overline{Q_s}$ is the mean simulated runoff, r is the correlation coefficient between

Q_s and Q_o, $\alpha = \frac{\sigma_s}{\sigma_o}$ is the variability ratio, $\beta = \frac{\overline{Q_s}}{\overline{Q_o}}$ is the bias ratio, and σ is the standard deviation.

The NSE is dimensionless and ranges between $-\infty$ and 1.0, NSE = 1 being the optimal value. A limitation of NSE is the underestimation of peak flows and overestimation of low flows, in such cases the KGE is suggested (Gupta et al., 2009), with $KGE = 1$ the optimal value. Additionally, the optimal value of $PBIAS$ is 0, positive values indicate model underestimation bias and negative values overestimation bias. Finally, $RMSE$ measures how model residuals are spread out from the best fit between simulations and observations, being $RMSE$ = 0 the optimal value.

3. Results

3.1 Determination of nearly independent peak hydrological events

The WETSPRO tool for the Jubones basin was calibrated using the following parameters: inter-event time of 120 hours (i.e., consecutive extreme hydrological events must be separated by a time frame of at least 5 days), and a maximum ratio of runoff drop down of 0.6 (i.e, runoff, q, drops down in between two consecutive events to a ratio $\frac{q_{min}}{q_{max}} < 0.6$). Moreover, we considered only events exceeding the 90% quartile values of the runoff time series (98.8 $m^3.s^{-1}$). With these criteria, we obtained 55 nearly independent peak hydrological events (see Figure 3a).

4.2 Object-based Connected Component Analysis

For the 55 peak hydrological events, we firstly retrieved hourly precipitation maps from the PERSIANN-CCs and the IMERG-early run subproducts. Then, we applied the CCA algorithm with the precipitation threshold volume of 0.5 mm to derive the meteorological attributes and classify the precipitation event. The step-by-step application of the CCA algorithm for the map corresponding to the PERSIAN-CCS 2021-12-25 05:00 UTC is presented in Figure 4 (see the Methodology section).

CCA results showed that, for 15 extreme hydrological events, there was nearly or even an inexistent precipitation signal from the PERSIANN-CCS subproduct. For these 15 cases, we performed the CCA algorithm on the IMERG-early run dataset, and this resulted in a reduction of 40% of the events without any precipitation signal. In other words, although we used two precipitation satellite sources, we encountered 9 hydrological events where either no precipitation at all was observed or any precipitation object was identified according to the CCA algorithm. Therefore, these events were trimmed off, leaving 46 events available for further analyses.

The validity of the precipitation modular approach is demonstrated in two extreme hydrological events (see Figure 5). For instance, for the event from 2019-07-13 20:00 to 2019-07-14 20:00 UTC, it seems evident that the highest resolution of the PERSIANN-CCS subproduct leads to a clearer precipitation-runoff relation when compared to precipitation obtained from the IMERG-early run subproduct. The opposite happened for the event from 2019-10-07 16:00 to 2019-10-08 16:00 UTC, where the PERSIANN-CCS signal was practically inexistent, and the IMERG-early run signal was used to relate precipitation with runoff.

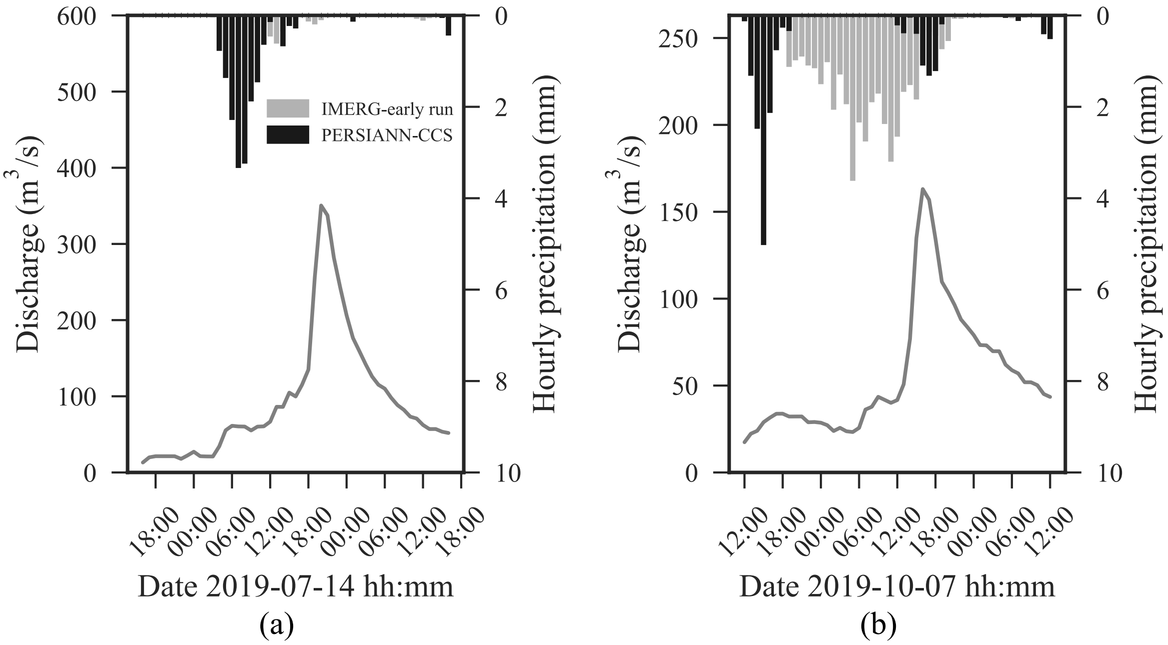


Figure 4. Illustration of the precipitation-retrieval modular approach using PERSIANN-CCS and IMER-early run data sources, respectively for the events from (a) 2019-07-13 18:00 to 2019-07-14 18:00 UTC, and (b) from 2019-10-07 12:00 to 2019-10-08 12:00 UTC.

Moreover, the precipitation objects identified with the CCA algorithm for each one of the 46 extreme hydrological events were tracked down. From this analysis, the following information was retrieved: quantity, localization (centroids) and extension of precipitation objects, precipitation duration, total precipitation volume, and precipitation maximum intensity. This information is summarized in Figure 6 and served to infer duration and extension thresholds of 7 hours and 50 km^2, respectively. These thresholds were used in the following subsection to classify the precipitation events.

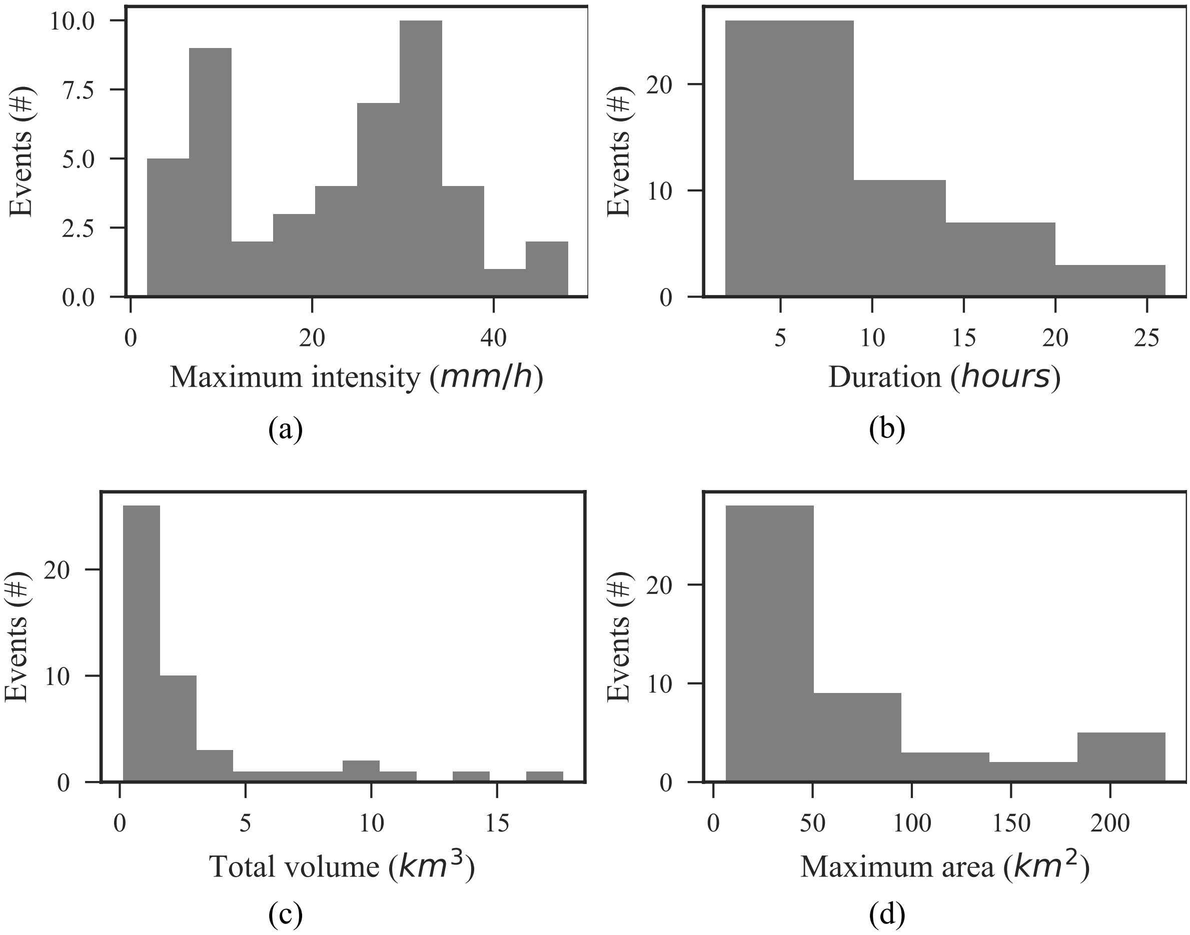


Figure 5. Meteorological precipitation information retrieved from 47 extreme hydrological events: (a) maximum intensity, (b) duration, (c) total volume, and (d) maximum area.

With respect to the localization of precipitation objects within the Jubones basin, centroid occurrence appeared to be unaffected by any physical attribute that could be derived for the basin (i.e., altitude, land use, etc.). Interestingly, no hotspot of precipitation occurrence was detected for the Jubones basin (see Figure 7). This suggests, for instance, that there is no evident orographic precipitation enhancement, and that the runoff generation process is rather driven by infiltration and saturation mechanisms before precipitation becomes streamflow.

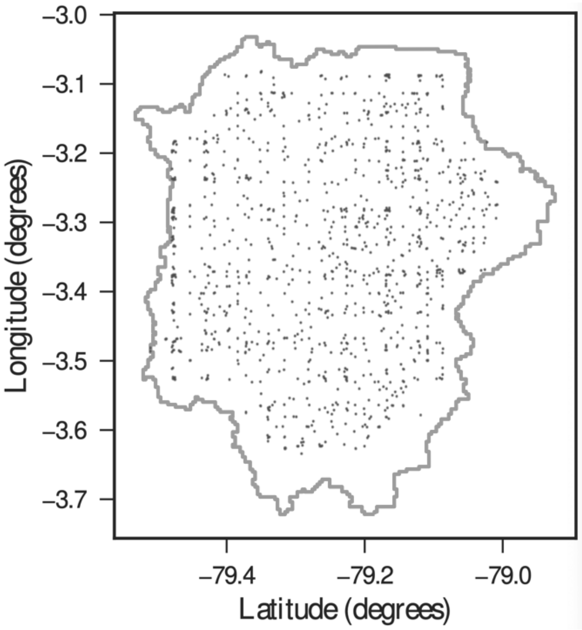


Figure 6. Localization of precipitation object centroids (green dots) associated with extreme hydrological events in the Jubones basin.

4.3 Classification of precipitation events associated with extreme hydrologic events

The combination of duration and extension thresholds of 7 hours and 50 km^2 served to define four precipitation classes. We determined 24 extreme hydrological events for the LSE precipitation class, 5 for the LLE, 7 for the SEE, and 10 for the SLE. Figure 8 depicts the visual discrimination between precipitation classes, from which it is apparent that the majority of extreme hydrological events occurred as a result of short duration and spatial local (LSE) precipitation events, and long duration and spatially extensive events (SLE).

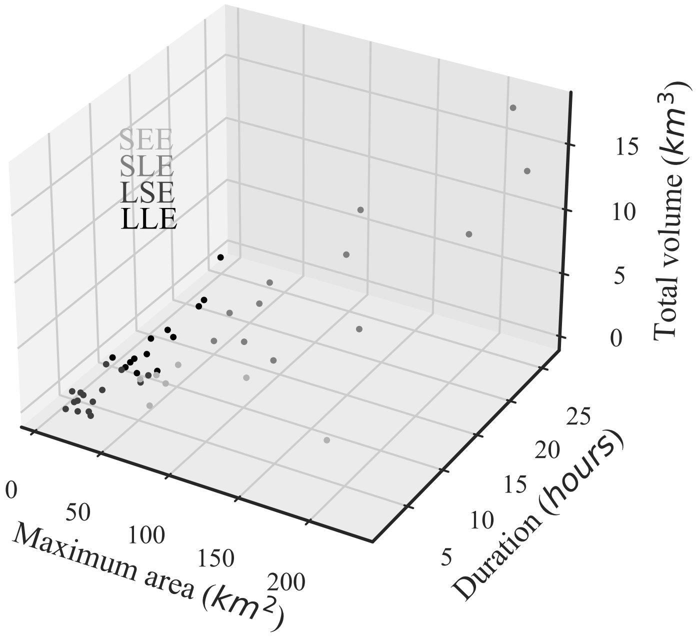


Figure 7. Precipitation classes associated with extreme hydrological events: Local and short extreme events (LSE), Local and long-duration extreme events (LLE), Spatially extensive extreme events (SEE), and Spatially extensive and long-duration extreme events (SLE).

4.4 Event-based runoff modeling

First, we defined the dimension of the input feature space of all extreme runoff models as a combination of current time precipitation together with past precipitation and past runoff data influencing current rime runoff. In this regard, results from partial- and auto-correlation functions for runoff suggest using past lags (hours) from 1 up to 12 lags, with a 95% confidence level for both correlation functions. Similarly, the cross-correlation function for precipitation determined 13 past lags (hours) of precipitation with correlations higher than 0.2. These results are congruent with the concentration-time of the Jubones basin, which was estimated at 11 hours by averaging the concentration times found with the equations of Giandotti, Johnstone, and the U.S. Army Corps of Engineers (equations recommended for the basin area, see de Almeida et al. (2014)).

Once the input feature space was defined, we constructed RF models for each precipitation class and the base model. For the model training and testing of each model, we assigned 70% of the events for training and the remaining 30% for testing. For instance, there were 46 events available for the LSE precipitation class; therefore, we assigned 32 events for training and 14 for testing. Moreover, since the objective was to simulate the hydrographs corresponding to each event, we used a time frame of 24 hours before and after peak events. Concerning RF hyperparameterization, Table 2 presents the optimized combination of hyperparameters for each runoff model. The coefficient of determination between simulations and observations for the training subsets of each model was always higher than 0.91.

Table 2. RF hyperparameterization of extreme runoff models.

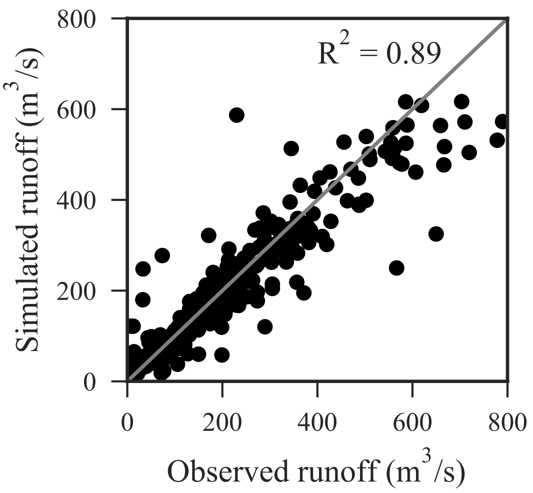
Hyperparameter	None	LSE	LLE	SLE	SEE
n_trees*	300	280	250	300	300
max_features	$2100^{(1/2)}$	$\log_2(2100)$	$\text{n\_features}^{(1/2)}$	$2100^{(1/2)}$	$\log_2(2100)$
max_depth*	200	200	150	180	200

Table 3 summarizes the number of events used for developing extreme runoff models, and a comparison of the NSE coefficients obtained for each precipitation class and the base model. It is apparent from this table that LSE and especially SEE precipitation events are causing decay in the overall NSE-value of 0.83 (see also Figures 9b and 9d). Surprisingly, LSE presents the majority of extreme hydrological events, and it seems contradictory that for LSE events, the higher number of events for training did not result in a higher NSE. This suggests that there are physical processes not well represented in the input feature space that disturbs the learning process of the RF models, as further discussed.

Table 3. Number of events and efficiencies on test subsets of runoff models specifically developed for different precipitation events.

Precipitation class	# Total Events (Test)	NSE	KGE	PBIAS	RMSE
None	46 (14)	0.83	0.85	4.49	55.38
LSE	24 (7)	0.67	0.71	-1.45	35.00
LLE	5 (2)	0.72	0.74	-23.94	41.76
SEE	7 (3)	-1.93	-0.48	-61.44	60.44
SLE	10 (3)	0.90	0.94	-2.72	69.09

From the data in Figure 9, we can infer the spectrum of the runoff magnitudes modeled for each precipitation class. What is striking from the subfigures in Figure 9 is that regardless of the spatial extension, short-duration precipitation events (LSE and SEE classes) caused the lowest extreme runoff magnitudes at the outlet of the Jubones basin. Now, since we developed models for extreme runoff, we maximized the efficiencies for the highest runoff magnitudes. Therefore, it is evident that the lowest NSE coefficients for the LSE and SEE classes are found. Physically, this finding may be explained by the fact that the runoff response of short-duration events is somehow softened by the infiltration and saturation processes. This means that the volume of precipitation that becomes streamflow is somehow lower when compared to long-duration precipitation classes (LLE and SLE). If we now turn to the modeling of all extreme hydrological events (Figure 9a), we can infer that the learning process is biased towards lower runoff magnitudes, and the results for the highest magnitudes are more spread out. However, the bias for long-duration events was reduced by classifying precipitation types before the modeling task (Figures 9c and 9e).



(a)

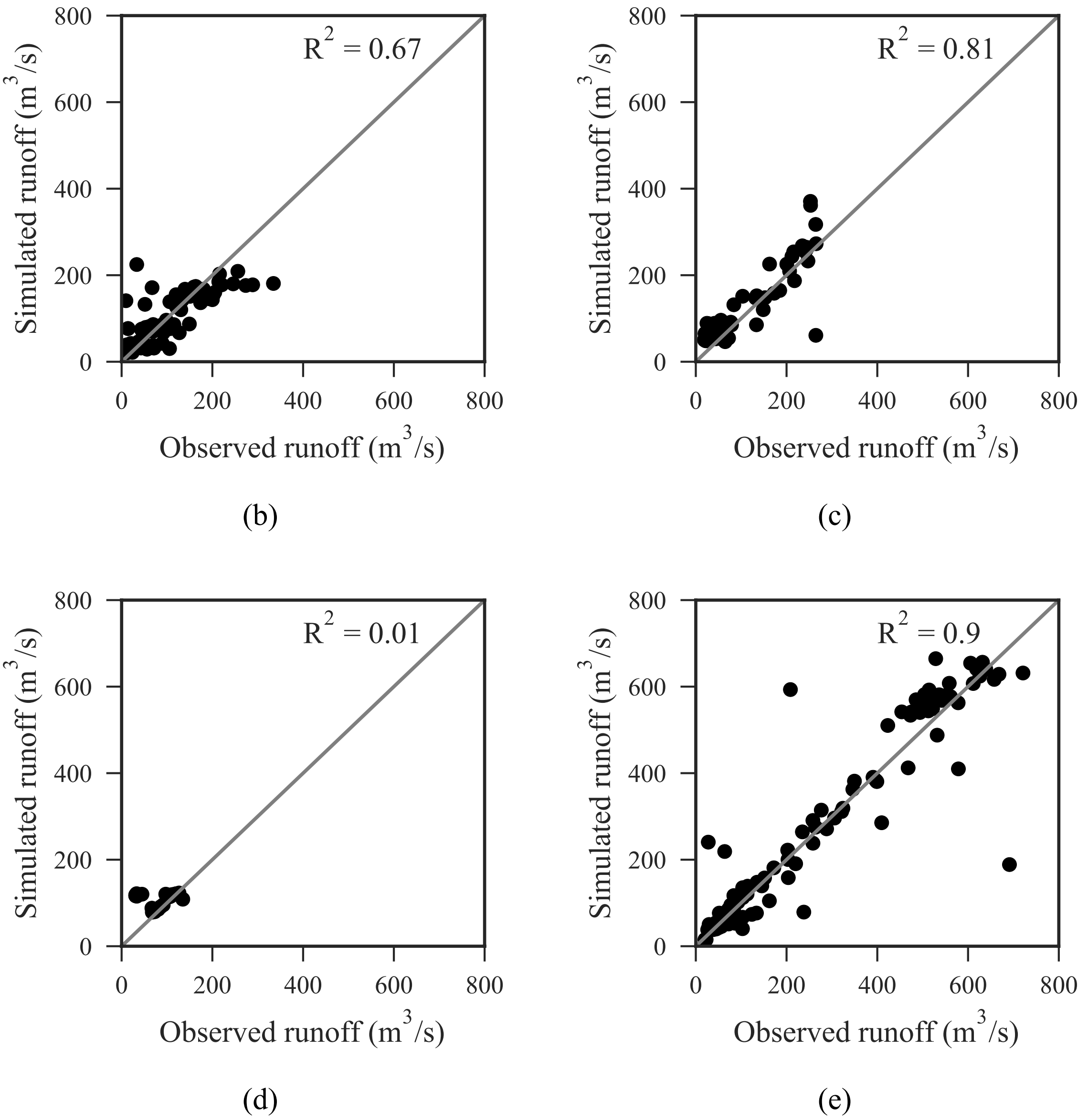


Figure 8. Scatter plot between extreme runoff observations and simulations for (a) No-precipitation event classification, (b) LSE events, (c) LLE events, (d) SEE events, and (e) SLE events.

4. Discussion

In this study, specialized (smart) extreme runoff models were developed for a 3391-km^2 representative basin of the Ecuadorian tropical Andes. The efficiencies of the developed ML models are comparable and outperformed the ones obtained with traditional physically-based models such as HEC-RAS (see the study of Belabid et al. (2019)), wflow-sbm (see Laverde-Barajas et al. (2020)), and the hydrologic-hydraulic HiResFlood-UCI model (see Nguyen et al.

(2015)). Particular to this finding is that unlike physically-based models, data-driven runoff models exploit precipitation satellite data without prior ground validation. Therefore, this study represents a solution for the cases when ground precipitation networks are scarce or even inexistent.

The specificities of our extreme runoff models were delineated for four precipitation-event types based on a combination of their duration and spatial extension (LSE, LLE, SEE and SLE). Developing specialized models served to identify the hidden strong-and-weak points of the base runoff model without precipitation classification. For instance, this approach could be used in the study of Belabid et al. (2019), where they obtained, in some cases, unacceptable runoff efficiencies (negative NSE).

For the Jubones basin, the vast majority of extreme hydrological events are the result of local and short-duration (LSE) precipitation events. In addition, we found that the centroids of LSE-associated objects were well distributed across the Jubones basin. These results indicate that small precipitation volumes are concentrated on many small different land use areas, characterized by a variety of specific runoff generation processes. Therefore, even for a discriminated LSE precipitation event, multiple precipitation-runoff responses can mislead the learning process of RF models. This explains the lower model efficiencies of LSE events (NSE=0.67) in comparison to SLE (0.90) and LLE (0.72) events. The opposite occurred for the case of long-duration and spatially extensive events (SLE), which were associated with the most extreme runoff magnitudes. For such events, even though we had less than half of the events available for LLE, model efficiencies reached the maximum (NSE=0.90). The LLE runoff model was clearly optimized for extreme runoff magnitudes (KGE=0.94). Physically, this is explained by the fact that the RF learning process becomes straightforward after a greater portion of the basin is saturated, and any additional precipitation volume is directly converted into streamflow. The major difficulty comes from the modeling of extreme runoff triggered by spatially-extensive and short-duration precipitation events (SEE).

The efficiencies of the developed and tested models highlighted the advantage of developing specialized extreme runoff models but also revealed the need to include additional information on antecedent soil saturation and its dynamic along with extreme hydrological events. This is particularly required for short-duration precipitation events (SEE and LSE), where the runoff generation process is governed by the antecedent saturation state of the basin. Foregoing is the reason why short-duration and non-extreme precipitation intensities can trigger extreme hydrological events. Given this, we encourage the approach employed by Massari et al. (2018) where they used satellite soil moisture observations to improve extreme runoff forecasting. Moreover, unveiling the limitations of runoff modeling for the Jubones basin opens the path for future research focused on exploring additional ML algorithms. We recommend, for instance, exploration of additional ML algorithms for the modeling of LSE and SEE events, and to come up with a superior model consisting of an ensemble of specialized runoff models.

5. Conclusions

This study exploits the possibility of using two near-real-time satellite precipitation sources (without ground validation) for the development of smart extreme runoff models for a 3391-km^2 basin. Smart models are characterized by the use of a ML algorithm with prior data assimilation enhancement under hydrometeorological criteria. For dealing with complex precipitation-runoff response and the optimization of the runoff model efficiency a straightforward feature engineering methodology was used. The major finding emerging from this study is that improvement of the representation of precipitation maximizes the efficiency of extreme runoff models. In addition, precipitation discrimination also served to unveil the precipitation-runoff scenarios misleading the learning process of RF extreme models.

In general, we found that the spatial extension of precipitation events made no significant difference in the learning process of RF models when they occurred for long-duration periods. In fact, these particular events produced the highest runoff magnitudes at the outlet of the basin. Physically, the success in modeling such precipitation events is attributed to a clear precipitation-runoff signal resulting from a gradual soil saturation process before precipitation is turned into runoff. This signal served to improve the learning process of RF models by reducing noise and maximizing model efficiencies. In terms of input data, the present study intentionally used and tested two near-real-time precipitation satellite sources, the PERSIAN-CCS and IMERG-early run subproducts. We used a modular framework of precipitation data acquisition that reduced 40% of precipitation events with nearly- or even inexistent precipitation signal.

All in all, the knowledge gained from the functioning of the basin, the proposed feature engineering methodology, and the evaluation of nearly-real-time satellite precipitation sources provides hydrologists with the tools for the future development of real-time runoff forecasting models. In addition, this study can be used to assist decision-makers in the fields of flood forecasting, water resources management, optimization of hydropower generation, and many more.

References

Belabid, N., Zhao, F., Brocca, L., Huang, Y., Tan, Y., 2019. Near-real-time flood forecasting based on satellite precipitation products. Remote Sens. 11, 252.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bontempi, G., Taieb, S. Ben, Le Borgne, Y.-A., 2012. Machine Learning Strategies for Time Series Forecasting., in: EBISS. pp. 62–77.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chang, L.-C., Chang, F.-J., Yang, S.-N., Kao, I., Ku, Y.-Y., Kuo, C.-L., Amin, I., others, 2019. Building an Intelligent Hydroinformatics Integration Platform for Regional Flood Inundation Warning Systems.

Clark, M.P., Bierkens, M.F.P., Samaniego, L., Woods, R.A., Uijlenhoet, R., Bennett, K.E., Pauwels, V.R.N., Cai, X., Wood, A.W., Peters-lidard, C.D., 2017. The evolution of process-based hydrologic models : historical challenges and the collective quest for physical realism 3427–3440.

Contreras, P., Orellana-Alvear, J., Muñoz, P., Bendix, J., Célleri, R., 2021. Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment. Atmosphere (Basel). 12, 238.

Davis, C., Brown, B., Bullock, R., 2006. Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. Mon. Weather Rev. 134, 1772–1784.

de Almeida, I.K., Almeida, A.K., Anache, J.A.A., Steffen, J.L., Alves Sobrinho, T., 2014. Estimation on time of concentration of overland flow in watersheds: A review. Geociencias 33, 661–671.

der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., 2014. scikit-image: image processing in Python. PeerJ 2, e453.

Galelli, S., Castelletti, A., 2013. Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. Hydrol. Earth Syst. Sci. 17, 2669–2684.

Gupta, H. V, Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91.

Gutiérrez-Jurado, K.Y., Partington, D., Batelaan, O., Cook, P., Shanafield, M., 2019. What triggers streamflow for intermittent rivers and ephemeral streams in low-gradient catchments in Mediterranean climates. Water Resour. Res. 55, 9926–9946.

Hong, Y., Hsu, K.-L., Sorooshian, S., Gao, X., 2004. Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. J. Appl. Meteorol. 43, 1834–1853.

Hong, Y., Tang, G., Ma, Y., Huang, Q., Han, Z., Zeng, Z., Yang, Y., Wang, C., Guo, X., 2019. Remote sensing precipitation: Sensors, retrievals, validations, and applications. Obs. Meas. Li, X., Vereecken, H., Eds 1–23.

Hsu, K., Gao, X., Sorooshian, S., Gupta, H. V, 1997. Precipitation estimation from remotely sensed information using artificial neural networks. J. Appl. Meteorol. 36, 1176–1190.

Huffman, G.J., Bolvin, D.T., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., Yoo, S.-H., 2015. NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). Algorithm Theor. Basis Doc. Version 4, 26.

Laverde-Barajas, M., Corzo, G., Bhattacharya, B., Uijlenhoet, R., Solomatine, D.P., 2019. Spatiotemporal analysis of extreme rainfall events using an object-based approach, in: Spatiotemporal Analysis of Extreme Hydrological Events. Elsevier, pp. 95–112.

Laverde-Barajas, M., Perez, G.A.C., Chishtie, F., Poortinga, A., Uijlenhoet, R., Solomatine, D.P., 2020. Decomposing satellite-based rainfall errors in flood estimation: Hydrological responses using a spatiotemporal object-based verification method. J. Hydrol. 591, 125554.

Li, B., Yang, G., Wan, R., Dai, X., Zhang, Y., 2016. Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China. Hydrol. Res. 47, 69–83. https://doi.org/10.2166/nh.2016.264

Li, J., Hsu, K.-L., AghaKouchak, A., Sorooshian, S., 2016. Object-based assessment of satellite precipitation products. Remote Sens. 8, 547.

Li, M., Zhang, Y., Wallace, J., Campbell, E., 2020. Estimating annual runoff in response to forest change: A statistical method based on random forest. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2020.125168

Massari, C., Camici, S., Ciabatta, L., Brocca, L., 2018. Exploiting satellite-based surface soil moisture for flood forecasting in the Mediterranean area: State update versus rainfall correction. Remote Sens. 10, 292.

Mosavi, A., Ozturk, P., Chau, K.W., 2018. Flood prediction using machine learning models: Literature review. Water (Switzerland) 10, 1–40. https://doi.org/10.3390/w10111536

Muñoz, P., Orellana-Alvear, J., Célleri, R., 2021. Application of a Machine Learning Technique for Developing Short-Term Flood and Drought Forecasting Models in Tropical Mountainous Catchments, in: Integrated Research on Disaster Risks. Springer, pp. 11–35.

Muñoz, P., Orellana-Alvear, J., Willems, P., Célleri, R., 2018. Flash-flood forecasting in an andean mountain catchment-development of a step-wise methodology based on the random forest algorithm. Water (Switzerland) 10. https://doi.org/10.3390/w10111519

Nguyen, P., Sellars, S., Thorstensen, A., Tao, Y., Ashouri, H., Braithwaite, D., Hsu, K., Sorooshian, S., 2014. Satellites track precipitation of super typhoon Haiyan. Eos, Trans. Am. Geophys. Union 95, 133–135.

Nguyen, P., Thorstensen, A., Sorooshian, S., Hsu, K., AghaKouchak, A., 2015. Flood forecasting and inundation mapping using HiResFlood-UCI and near-real-time satellite precipitation data: The 2008 Iowa flood. J. Hydrometeorol. 16, 1171–1183.

Orellana-Alvear, J., Célleri, R., Rollenbeck, R., Muñoz, P., Contreras, P., Bendix, J., 2020. Assessment of Native Radar Reflectivity and Radar Rainfall Estimates for Discharge Forecasting in Mountain Catchments with a Random Forest Model. Remote Sens. 12, 1986.

Papacharalampous, G.A., Tyralis, H., 2018. Evaluation of random forests and Prophet for daily streamflow forecasting 201–208.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res.

Peña-Barragán, J.M., Ngugi, M.K., Plant, R.E., Six, J., 2011. Object-based crop identification using multiple vegetation indices, textural features and crop phenology. Remote Sens. Environ. 115, 1301–1316.

Sakib, S., Ghebreyesus, D., Sharif, H.O., 2021. Performance Evaluation of IMERG GPM Products during Tropical Storm Imelda. Atmosphere (Basel). 12, 687.

Sorooshian, S., Nguyen, P., Sellars, S., Braithwaite, D., AghaKouchak, A., Hsu, K., 2014. Satellite-based remote sensing estimation of precipitation for early warning systems. Extrem. Nat. hazards, disaster risks Soc. Implic. 1, 99.

Tang, G., Long, D., Hong, Y., 2016. Systematic anomalies over inland water bodies of High Mountain Asia in TRMM precipitation estimates: No longer a problem for the GPM era? IEEE Geosci. Remote Sens. Lett. 13, 1762–1766.

Tyralis, H., Papacharalampous, G., Langousis, A., 2019. A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water (Switzerland). https://doi.org/10.3390/w11050910

Vogels, M.F.A., de Jong, S.M., Sterk, G., Wanders, N., Bierkens, M.F.P., Addink, E.A., 2020. An object-based image analysis approach to assess irrigation-water consumption from MODIS products in Ethiopia. Int. J. Appl. Earth Obs. Geoinf. 88, 102067.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., Bai, X., 2015. Flood hazard risk assessment model based on random forest. J. Hydrol. 527, 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008

Willems, P., 2009. A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models. Environ. Model. Softw. 24, 311–321.

Young, P.C., 2002. Advances in real--time flood forecasting. Philos. Trans. R. Soc. London. Ser. A Math. Phys. Eng. Sci. 360, 1433–1450.