

UCUENCA

Universidad de Cuenca

Facultad de Ingeniería

Doctorado en Recursos Hídricos

Towards the improvement of machine learning peak runoff forecasting by exploiting ground- and satellite-based precipitation data: A feature engineering approach


Trabajo de titulación previo a la obtención del título de Doctor (PhD) en Recursos Hídricos

Autor:

Paul Andrés Muñoz Pauta


Director:

Rolando Enrique Céleri Alvear

ORCID:  0000-0002-7683-3768

Tutor:

Johanna Marlene Orellana Alvear

ORCID:  0000-0002-6206-075X

Cuenca, Ecuador

2023-05-10

Resumen

La predicción de picos de caudal en sistemas montañosos complejos presenta desafíos en hidrología debido a la falta de datos y las limitaciones de los modelos físicos. El aprendizaje automático (ML) ofrece una solución al permitir la integración de técnicas y productos satelitales de precipitación (SPPs). Sin embargo, se ha debatido sobre la efectividad del ML debido a su naturaleza de "caja negra" que dificulta la mejora del rendimiento y la reproducibilidad de los resultados. Para abordar estas preocupaciones, se han propuesto estrategias de ingeniería de características (FE) para incorporar conocimiento físico en los modelos de ML, mejorando la comprensión y precisión de las predicciones. Esta investigación doctoral tiene como objetivo mejorar la predicción de picos de caudal mediante la integración de conceptos hidrológicos a través de técnicas de FE y el uso de datos de precipitación in-situ y SPPs. Se exploran técnicas y estrategias de ML para mejorar la precisión en sistemas hidrológicos macro y mesoescala. Además, se propone una estrategia de FE para aprovechar la información de SPPs y superar la escasez de datos espaciales y temporales. La integración de técnicas avanzadas de ML y FE representa un avance en hidrología, especialmente para sistemas montañosos complejos con limitada o nula red de monitoreo. Los hallazgos de este estudio serán valiosos para tomadores de decisiones e hidrólogos, facilitando la mitigación de los impactos de los picos de caudal. Además, las metodologías desarrolladas se pueden adaptar a otros sistemas de macro y mesoescala, beneficiando a la comunidad científica en general.

Palabras clave: picos de caudal, inundaciones, aprendizaje automático, ingeniería de características, Andes

Abstract

Peak runoff forecasting in complex mountain systems poses significant challenges in hydrology due to limitations in traditional physically-based models and data scarcity. However, the integration of machine learning (ML) techniques offers a promising solution by balancing computational efficiency and enabling the incorporation of satellite precipitation products (SPPs). However, debates have emerged regarding the effectiveness of ML in hydrology, as its black-box nature lacks explicit representation of hydrological processes, hindering performance improvement and result reproducibility. To address these concerns, recent studies emphasize the inclusion of FE strategies to incorporate physical knowledge into ML models, enabling a better understanding of the system and improved forecasting accuracy. This doctoral research aims to enhance the effectiveness of ML in peak runoff forecasting by integrating hydrological concepts through FE techniques, utilizing both ground-based and satellite-based precipitation data. For this, we explore ML techniques and strategies to enhance accuracy in complex macro- and meso-scale hydrological systems. Additionally, we propose a FE strategy for a proper utilization of SPP information which is crucial for overcoming spatial and temporal data scarcity. The integration of advanced ML techniques and FE represents a significant advancement in hydrology, particularly for complex mountain systems with limited or inexistent monitoring networks. The findings of this study will provide valuable insights for decision-makers and hydrologists, facilitating effective mitigation of the impacts of peak runoffs. Moreover, the developed methodologies can be adapted to other macro- and meso-scale systems, with necessary adjustments based on available data and system-specific characteristics, thus benefiting the broader scientific community.

Keywords: peak runoff, flash floods, machine learning, feature engineering, Andes

Index of contents

Chapter one: introduction.	10
1.1 Data sources	12
1.2 Peak runoff forecasting models	13
1.3 Feature engineering in peak runoff forecasting.....	14
1.4 Aim of the research	15
1.4.1 Work packages (WPs)	15
1.4.2 Outline of the thesis.....	16
1.5 Study areas	18
1.5.1 A meso-scale hydrological system: the Tomebamba catchment	18
1.5.2 A macro-scale hydrological system: the Jubones basin.....	20
Chapter two: methodological framework for developing machine learning flash flood forecasting models.	24
2.1 Aim and objectives	25
Objectives:.....	25
2.2 Review of machine learning (ml) techniques	25
2.1.1 Logistic regression	26
2.1.2 K-nearest neighbors	27
2.1.3 Naïve bayes.....	27
2.1.4 Random Forest	28
2.1.5 Multi-layer perceptron	29
2.3 Methodology for developing ml peak runoff forecasting models	29
2.3.1 Statistical lag analyses	30
2.3.2 Model hyperparameterization	31
2.3.3 Feature space reduction	32
2.4 Evaluation of machine learning peak runoff forecasting models.....	34
2.4.1 Evaluation of ML qualitative models.....	34
2.4.2 Evaluation of ML quantitative models.....	37
Chapter three: exploration of quantitative and qualitative machine learning flash flood forecasting using ground-based precipitation data.....	40
3.1 Aim and objectives	41
Objectives:.....	41
3.2 Qualitative ml flash flood forecasting	41
3.2.1 Introduction.....	41
3.2.2 Dataset and processing	42
3.2.3 Methodology	43
3.2.4 Results and discussion	46
3.3 Quantitative flash flood forecasting.....	55
3.3.1 Introduction.....	55
3.3.2 Dataset and processing	56
3.3.3 Methodology	56
3.3.4 Results and discussion	58
3.4 Summary and conclusions.....	61
Chapter four: a feature engineering strategy for exploiting of satellite-based precipitation data in machine learning models.	65

4.1	Aim and objectives	67
	Objectives:.....	67
4.2	Feature engineering strategies.....	67
4.2.1	Object-based connected component analysis (CCA)	67
4.2.2	Classification of precipitation events leading to peak runoffs.....	69
4.3	Implementation of fe strategies for peak runoff modeling	70
4.3.1	Dataset and processing	70
4.3.2	Methodology	71
4.3.3	Results	74
4.3.4	Discussion	81
4.3.5	Conclusions	82
Chapter five: feature engineering strategies for exploiting ground- and satellite-based precipitation data and for adding process-based hydrological knowledge.		84
5.1	Aim and objectives	85
	Objectives.....	85
5.2	Peak runoff forecasting in a precipitation ungauged macro-scale hydrological system	85
5.2.1	Dataset processing.....	86
5.2.2	Methodology	87
5.2.3	Results	93
5.2.4	Discussion	98
5.3	Flash flood forecasting exploiting ground- and satellite-based precipitation data in a meso-scale hydrological system.....	100
5.3.1	Dataset processing.....	100
5.3.2	Methodology	101
5.3.3	Results and discussion	106
5.4	Summary and conclusions.....	114
Chapter six: summary, conclusions and feature work.		117
References		122

Index of figures

Figure 1.1. Work packages of the doctoral research and their associated thematic chapters.....	17
Figure 1.2. The Tomebamba catchment in the southern Ecuadorian Andes. Location of ground-based precipitation stations (Toreadora, Virgen del Cajas, and Chirimachay), and runoff at the outlet (Sayausí).....	19
Figure 1.3. (a) PERSIANN-CCS coverage and mean annual precipitation over the Tomebamba catchment. (b) Comparison between annual precipitation measured by ground-based products (average of three rain gauges over microcatchments M1, light blue line) and the PERSIANN-CCS (average over M1, dark blue line). The remaining gray lines depict the PERSIANN-CCS precipitation for microcatchments M2-M6.....	20
Figure 1.4. The Jubones basin in the Tropical Andes of Ecuador, South America (UTM coordinates).....	21
Figure 1.5. Mean annual precipitation measured by the PERSIANN-CCS and the IMERG-ER satellite products for the study period from January 2019 to June 2022 (Jubones basin, Ecuador).....	22
Figure 2.1. Step-wise methodology scheme for developing ML peak runoff forecasting models..	30
Figure 3.1. Time series of precipitation (Toreadora) and discharge (Matadero-Sayausí). Horizontal dashed lines indicate the mean runoff and the currently employed flood alert levels for labeling the Pre-alert and Alert flood warnings classes.....	43
Figure 3.2. Methodologic scheme for the development and testing of ML flash flood forecasting models.....	44
Figure 3.3. (a) Autocorrelation function (ACF) and (b) Partial-autocorrelation function (PACF) of the Matadero-Sayausí. (Tomebamba catchment) discharge series. The blue hatch indicates in each case the correspondent 95% confidence interval.....	47
Figure 3.4. Pearson's cross-correlation comparison between the Toreadora (3955 m a.s.l.), Virgen (3626 m a.s.l.), and Chirimachay (3298 m a.s.l.) precipitation stations and the Matadero- Sayausí discharge series. Note the grey horizontal line at a fixed correlation of 0.2 for determining the number of lags.....	47
Figure 3.5. <i>F1 scores</i> per flood warning state (No-alert, Pre-alert, and Alert) for all combinations of ML techniques and lead times. The brightest and dashed lines in each case (color coding) represent the scores for the test subset.....	53
Figure 3.6. Methodology scheme for parsimonious model development.....	59
Figure 3.7. Empirical extreme value distribution of peak flows (flash floods).....	61
Figure 3.8. Comparison of nearly independent peak flow maxima.....	61

Figure 4.1. Precipitation identification with an object-based Connected Component Analysis (CCA) Illustration of the PERSIAN-CCS 2021-12-25 05:00 UTC image. (a) Jubones basin boundary, (b) Precipitation identification in mm from the PERSIANN-CCS product, (c) Identification of three precipitation objects with the CCA, and (d) Final identification of two precipitation objects after object size filtering and morphological closing.....69

Figure 4.2. (a) Hourly runoff and precipitation (PERSIANN-CCS) time series at the outlet of the Jubones basin. Peak flow events are displayed as dots. (b) Exceedance probability for the study period (18/11/2018 to 01/04/2021).....71

Figure 4.3. Illustration of the precipitation-retrieval modular approach using PERSIANN-CCS and IMERG-ER 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.....75

Figure 4.4. Meteorological precipitation information was retrieved from 46 extreme hydrological events: (a) maximum intensity, (b) duration, (c) total volume, and (d) maximum area.....76

Figure 4.5. Localization of precipitation object centroids (blue dots) associated with extreme hydrological events in the Jubones basin.....77

Figure 4.6. 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).....78

Figure 4.7. 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.....82

Figure 5.1. Mean annual precipitation measured by the PERSIANN-CCS and the IMERG-ER satellite products for the study period from January 2019 to June 2022 (Jubones basin, Ecuador).....88

Figure 5.2. Scheme of the methodology for developing peak runoff forecasting models, (a) extreme peak runoff selection and subflow separation, (b) satellite precipitation processing, and (c) forecast modeling approach.....88

Figure 5.3. (a) Directflow and baseflow separation from the total flow time series at the outlet of the Jubones basin. Peak flow events selected with the WETSPRO tool are displayed as blue dots. (b) Exceedance probability of total flow for the study period (01/01/2019 to 13/06/2022).....90

Figure 5.4. Meteorological precipitation information retrieved from 81 extreme hydrological events: (a) maximum intensity, (b) event duration, (c) total volume, and (d) maximum area.....94

Figure 5.5. Comparison of the scatter plots of the observed and forecasted runoff for the base model and the specialized models for the 1-hour lead time.....97

Figure 5.6. Hourly runoff (total flow) of the Sayaus. station, and its subflow components (baseflow and directflow). 156 nearly-independent peak flows are displayed as green dots. Study period from Jan/2015 to May/2021..... 104

Figure 5.8. Model structures for baseflow and directflow RF forecasting models.....105

Figure 5.9. Comparison of PERSIANN-CCS and average ground-based (microcatchment M1) histograms of hourly precipitation. Considering the asymmetry of the data, the histograms were split up into different size class bins.....109

Figure 5.10. Scatter density of PERSIANN-CCS estimates and corresponding ground-based precipitation at (a) daily and (b) monthly scales. Period of analysis from Jan/2015 to May/2021.....108

Figure 5.11. Precipitation-event characteristics derived from the CCA: (a) event duration, (b) maximum intensity plotted with different class widths, (c) areal extension, and (d) total volume plotted with different class widths.....109

Figure 5.12. Scatter density plot of timeseries of forecasted total flow for the testing periods: (ad) plots for lead times of 1, 4, 8, and 12 hours, respectively.....113

Figure 5.13. (a) Empirical peak value distribution, (b) Comparison of nearly independent peak flow maxima.....114

Index of tables

Table 2.1. Model hyperparameters of the most-employed ML techniques for peak runoff forecasting.....	32
Table 3.1. Model hyperparameters and their ranges/possibilities for tuning.....	45
Table 3.2. Input feature space composition (number of features) for all ML models of the Tomebamba catchment.....	48
Table 3.3. Model hyperparameters and the number of principal components used for each specific model (ML technique and lead time).....	49
Table 3.4. The number of samples and relative percentage for the entire dataset and the training and test subsets.....	50
Table 3.5. Models' performance evaluation on the test subset. Bold fonts indicate the best performance for a given lead time.....	51
Table 3.6. Random Forest most-relevant model hyper-parameters and their search domain for tuning.....	58
Table 3.7. Input feature space composition of the RF models and their parsimonious versions (4, 8, 12, and 24-hour lead time) for the Tomebamba catchment.....	59
Table 3.8. Model performance of the RF models and their parsimonious versions (4, 8, 12, and 24-hour lead time).....	60
Table 4.1. Search space (grid) of the RF runoff models.....	74
Table 4.2. RF hyperparameterization of extreme runoff models.....	79
Table 4.3. The number of events and efficiencies on test subsets of runoff models specifically developed for different precipitation events.....	79
Table 5.1. Search space of the RF hyperparameters.....	92
Table 5.2. RF hyperparameterization of the forecasting models for the 1-hour lead time.....	94
Table 5.3. Model efficiencies (LOOCV evaluation framework) for the base and specialized forecasting models across lead times.....	96
Table 5.4. Efficiency metrics for precipitation.....	103
Table 5.5. Optimal combination of RF hyperparameters for the baseflow and directflow forecasting models across lead times.....	110
Table 5.6. Forecasting performances for the baseflow, directflow, and total flow models across increasing lead times.....	111

List of abbreviations

ANN: Artificial Neural Networks

CCS: Cloud Classification System

CELEC-EP: Empresa Pública de la Corporación Eléctrica del Ecuador

DL: Deep Learning

ER: Early Run

ETAPA-EP: Empresa Pública Municipal de Telecomunicaciones, Agua Potable, Alcantarillado y Saneamiento de Cuenca

FE: Feature Engineering

IMERG: Integrated Multi-satellite Retrievals for GPM

LULC: Land Use Land Cover

ML: Machine Learning

MSF: Minas San-Francisco

PERSIANN: Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks

RF: Random Forest

SPP: Satellite Precipitation products

SVM: Support Vector Machines

Chapter one: introduction.

Floods resulting from peak runoffs are the most frequent and destructive natural disasters worldwide [1]–[4]. They have major impacts on society, including human losses, increased health risks, and disruptions to water and sewer services, as well as on the economy, with losses of agricultural production and damage to infrastructure and transportation networks. Floods also have significant ecological consequences, altering hydro-geomorphic conditions and causing changes to river and floodplain habitats, as well as biodiversity loss [5].

With much concern, recent studies worldwide have associated the increasing frequency and severity of peak runoff and flood events with land use land cover (LULC) changes (e.g., deforestation and urbanization), and climate change [3], [6]–[8]. For these reasons, peak runoff forecasting has globally become an emerging field of research, and its applications are of major importance for water management, risk analysis, and resilience enhancement [9], [10].

Floods can be classified according to their generation mechanisms into long- and short-precipitation floods [11], [12]. In any case, the response time between a precipitation event and its associated flood response can be on the scale of minutes, hours, or even longer [13]. Based on these concepts, flash floods are defined as peak runoffs that develop less than six hours after a precipitation event with little or no forecast lead time [14]. Thus, the key to building resilience to flash floods is to sufficiently anticipate the event itself and provide accurate forecasts for decision-making.

However, although crucial, the development of flood anticipation (forecasting) models is still a major challenge within the scientific community, especially for complex hydrological systems (e.g., catchment). Complex hydrological systems are hereafter defined as meso-scale ($10 \text{ km}^2 \leq \text{area} \leq 1000 \text{ km}^2$) and macro-scale ($> 1000 \text{ km}^2$) mountain systems whose flash flood response is the result of extremely variable yet poorly monitored driving forces. In short, meso-scales hydrology refers to the study of hydrological processes, where local land use and topography play a significant role. Whereas, macro-scale hydrology deals with the large-scale hydrological processes that occur at a regional or global level, where the effects of local land use and topography are less significant.

The main driving forces for peak runoff (including flash floods) formation are for instance precipitation, soil and LULC information, soil moisture (humid areas), and topography [15], [16].

Overall, macro-scale systems depict higher heterogeneity on the peak runoff main driving forces when compared to meso-scale mountain systems. This is because larger mountain systems commonly encompass multiple climates and terrain features, which in combination, lead to highly variant peak runoff driving forces. Consequently, the difficulty to monitor these driving forces induce spatial and/or temporal data scarcity issues. On one hand, spatial data scarcity relates to representability of the information. For instance, it is known that in complex regions such as the tropical Andes, the characterization of precipitation patterns is limited by inexistent or insufficient ground precipitation networks (rain gauges) [17]–[20]. And on the other hand, temporal data scarcity is referred to the cases when there is insufficient dataset extension for describing general processes, e.g., the use of less than a hydrological year for a water balance model. A solution to deal with both spatial and temporal data limitation is the exploitation of data derived from remote sensing estimates such as satellite precipitation products (SPPs). The situation is similar when it comes to soil moisture, which is even more complicated because the remote sensing data available only provides imagery at a temporal resolution of one day, which is not suitable for sub-daily peak runoff forecasting.

An additional limiting issue in peak runoff forecasting is the selection of an adequate forecasting technique in terms of input data demand, model efficiency, and computational cost criteria. For instance, employment of the most complex forecasting model might be unfeasible due to extensive input data demand, lack of efficiency, or simply because the required computation time hinders the lead time. Therefore, the appropriate forecasting model must meet parsimony concepts and optimize its accuracy (performance metrics), for instance, by improving the representation of peak runoff governing processes from the available input data. In practice, representation improvement can be achieved by the application of strategies such as input variable selection, preprocessing of input data, and derivation of new information for facilitating the data assimilation of forecasting models.

From the previous revision of state-of-the-art flash flood forecasting, we identified two clear research niches that aim to improve the efficiency of peak runoff forecasts. The first one is related to the process of forecasting with the latest techniques (models) together with strategies for improving forecasting efficiencies in complex meso- and macro-scale hydrological systems. The second niche is related to find ways to exploit SPP information for dealing with spatial and temporal data scarcity. In the following paragraphs, we expand on these niches.

1.1 Data sources

Peak runoff forecasting models demand information on at least two variables: runoff and precipitation (i.e., precipitation-runoff models). Runoff data has to be obtained from in situ measurements, whereas, precipitation information can be retrieved either from ground-based or SPPs products.

The most employed sources of precipitation are rain gauges and weather radars yet they demand purchase, installation, and maintenance costs for continuous monitoring [25]. The problem gets exacerbated when the interest lies in mountain regions with complex topography. This is because the spatial characterization precipitation demands highly-dense monitoring networks, and even worst, the most traditional and affordable rain gauges (tipping bucket) have shown important deficiencies in measuring certain types of precipitation such as drizzle, causing important precipitation underestimations/overestimations [20], [21]. To overcome these issues, recent advances in remote sensing technology such as freely-available SPPs are becoming popular since they provide spatial precipitation data that, in principle, could produce more efficient forecasting models [25]. However, the primary challenge is validating/correcting satellite data in cases where there are no ground-based precipitation networks available.

Among SPPs, we highlight the NASA Global Precipitation Measurement (GPM), Integrated Multi-satellite Retrievals for GPM (IMERG) [22], and the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) [23]. IMERG and PERSIANN products offer quasi-global coverage, free access, high spatiotemporal resolutions, and short latency times adequate for flash flood forecasting and real-time operation. For instance, the spatial resolution of the PERSIANN-Cloud Classification System (CCS) product is 0.04° (i.e., pixels of $\sim 4.4 \times 4.4$ km). The PERSIANN-CCS delivers global precipitation images each hour (1-hour temporal resolution), and they are available to the public with a latency time of 1 hour [24]. For these reasons, these SPPs have yielded a growing body of literature for hydrometeorological applications [25]. Current applications include tracking precipitation anomalies [26], [27], precipitation early-warning systems [28], and flood forecasting and mapping [29], [30]. Thus, the combined use of SPPs and ground-based data represents an opportunity for developing flash flood forecasting models in regions with data scarcity issues such as the Andes.

1.2 Peak runoff forecasting models

Two main paradigms can be followed for the development of peak runoff forecasting models. These are the physics-based and data-driven paradigms. The physics-based paradigm aims at including mathematical equations for describing the physical processes that govern flash flood generation processes in a system [31]. Nevertheless, the use of traditional physically-based models to forecast peak runoffs is either restricted to data rich regions or leads to significant uncertainties for regions with complex biophysical characteristics [31]. This is due to data scarcity and extreme spatiotemporal variability of the driving forces. Furthermore, even with the increasing availability of SPPs, it remains mandatory a validation/correction with ground information before its usage. Moreover, overall the use physically-based models demands intensive computation, and leads to overparameterization issues and higher uncertainties for data poor regions. This complicates the use of physics-based models for peak runoff and real-time applications [9], [32]–[38].

Contrary to the physics-based paradigm, the data-driven one approaches peak runoffs phenomena as stochastic processes whose distribution probability can be directly derived from historical data. In other words, the data-driven paradigm can be used without requiring knowledge about the underlying physical processes in a system. Among traditional data-driven approaches for peak runoff modeling, we highlight the autoregressive moving average (ARMA) [39], autoregressive integrated moving average (ARIMA) [40], and multiple linear regressions (MLR) [41]. However, although these traditional models have provided improved generalization power and computational costs when compared to physically-based models, a major improvement is still required given their unsuitability for peak runoff and real-time [37]. The main reasons are lack of accuracy, high complexity regarding model structure (i.e., dependence of initial parameter values to inputs), and elevated computational costs that hinder the temporal forecast window or lead time for operational hydrology.

To overcome the shortcomings of traditional data-driven models, extensive research during the last decades has focused on the development and use of advanced data-driven models, e.g., machine learning (ML) [6], [34], [37], [42]–[48]. Particularly during the last decade, ML approaches have increased their popularity among hydrologists, mostly since the forecasting ability of a model is dependent on how much the modeler can exploit from relevant input information to find relations to the target variable (i.e., runoff) [37]. Moreover, since there is no assumption on a global function

describing the data, ML techniques are particularly relevant for problems of non-stationarity, missing features (estimators), and systematic measurement errors [42].

In this sense, the use of ML represents both a challenge and an opportunity. First, the challenge is to select the optimal ML technique for short-term peak runoff forecasting (flash floods). Different ML methods have been explored in pursue of a better performance in conventional hydrological problems. It is now common to find new studies that use artificial neural networks (ANNs), specialized ANNs also known as deep learning (DL), support vector machines (SVMs), and random forest (RF) [37], [47]–[54]. Second, the use of ML techniques is a great opportunity for exploiting SPPs, especially in cases where validation/correction is not possible due to a lack of ground monitoring networks. This is because SPP data in ML models are merely estimators of another target variables, i.e., runoff. Thus, the premise is that systematic errors of the estimators can be absorbed by ML models. This opportunity is pertinent for complex regions such as the tropical Andes, where the installation of ground monitoring networks is restricted by its topography [17], [18].

1.3 Feature engineering in peak runoff forecasting

While ML techniques have shown promising results in hydrology, there has been ongoing controversy in the field due to the black box nature of ML models, which do not explicitly represent the hydrological processes of a system and, as a result, lack physical knowledge that limits performance improvement and reproducibility of results [43]. To address this issue, current and future studies have focused on the use of FE strategies to improve input data representation and incorporate physical knowledge of the system to enhance the interpretability of ML models and improve their efficiency.

FE is a crucial component in the use of ML for hydrological modeling as it enables the creation of more meaningful input features, thereby improving the performance and interpretability of ML models. Specifically, FE involves a series of strategies such as missing data imputation, variable transformation, and feature creation, all aimed at enhancing the quality and relevance of input data. For example, in peak runoff forecasting, FE can be employed to incorporate additional information beyond precipitation and runoff, such as soil moisture, topography, and land use, to develop more specialized and accurate ML models. By leveraging FE, hydrologists can improve

their understanding of complex hydrological systems and develop more effective and reliable forecasting tools.

So far, however, there has been little discussion about ways in which available precipitation data can be exploited in ML models, or on how to incorporate essential hydrological knowledge to improve forecasting efficiencies of ML models. Some efforts are, for instance, the use of precipitation data to mimic antecedent soil moisture conditions through a proxy variable derived from precipitation data. This was done by Orellana-Alvear et al. [55] from weather radar data and can be replicated to any SPP. Other successful ways in which FE can be applied are the use of object-based methods for extracting precipitation attributes from SPPs [48], [56]–[61], runoff separation into subflow components [62]–[64], exploitation of topographic characteristics [65], the addition of stream network information [66], [67], sub-catchment modeling, and various ways to leveraging hydrological knowledge in selecting input attributes [68].

1.4 Aim of the research

The aim of this research is to improve the effectiveness of machine learning peak runoff forecasting by utilizing feature engineering techniques that exploit both ground- and satellite-based precipitation data, while also incorporating process-based hydrological knowledge. The study will focus on two complex mountain systems that are representative of meso- and macro-scales. To achieve this objective, the research is organized into three work packages, which are developed in six thematic chapters (see Figure 1.1).

1.4.1 Work packages (WPs)

- WP1: Development of ML peak runoff forecasting models using ground-based precipitation data.
- WP2: Exploitation of SPP data with a FE strategy for ML.
- WP3: Improvement in ML peak runoff and flash flood forecasting through the use of FE strategies for exploiting ground- and satellite-based precipitation data, and for adding process-based hydrological knowledge.

1.4.2 Outline of the thesis

Chapter One serves as an introduction to the importance of flash flood forecasting for society and the scientific community. The chapter also gives an overview of the theoretical dimensions for improving the effectiveness of machine learning (ML) flash flood forecasting through the use of feature engineering (FE) strategies for modeling complex mountain hydrological systems at meso- and macro-scale with data scarcity issues.

Chapter Two, as the first part of WP1, provides a detailed methodological framework for developing and evaluating ML-based flash flood forecasting models using model parsimony criteria. This chapter lays the foundation for the development of ML models for flash flood forecasting in meso- and macro-scale hydrological systems (WP1). The framework is then applied to explore qualitative and quantitative flash flood forecasting using ground-based precipitation data, as presented in Chapter Three (second part of WP1). The qualitative and quantitative case studies in a meso-scale hydrological system demonstrate the link between academia and society, as the developed models can be immediately integrated into a flash flood forecasting system.

Additionally, Chapter Four highlights the effectiveness of FE implementation to achieve ML forecasting improvement (WP2). FE strategies, such as exploiting satellite-based precipitation data and adding physical knowledge of the system to ML models, are tested for modeling the functioning of a system.

Then, Chapter Five encompasses WP3, which explores the use of the implemented FE strategies for improving flash flood forecasts in meso- and macro-scale hydrological systems by exploiting ground- and satellite-based precipitation data and adding process-based hydrological knowledge. This chapter focuses on two case studies: a precipitation ungauged meso-scale system and a macro-scale system where SPPs complemented existing ground-based precipitation data.

The final Chapter Six provides a summary of the research findings and highlights the future directions of the field. Overall, this thesis demonstrates the effectiveness of using ML techniques and FE strategies for modeling complex mountain hydrological systems and improving flash flood forecasting.

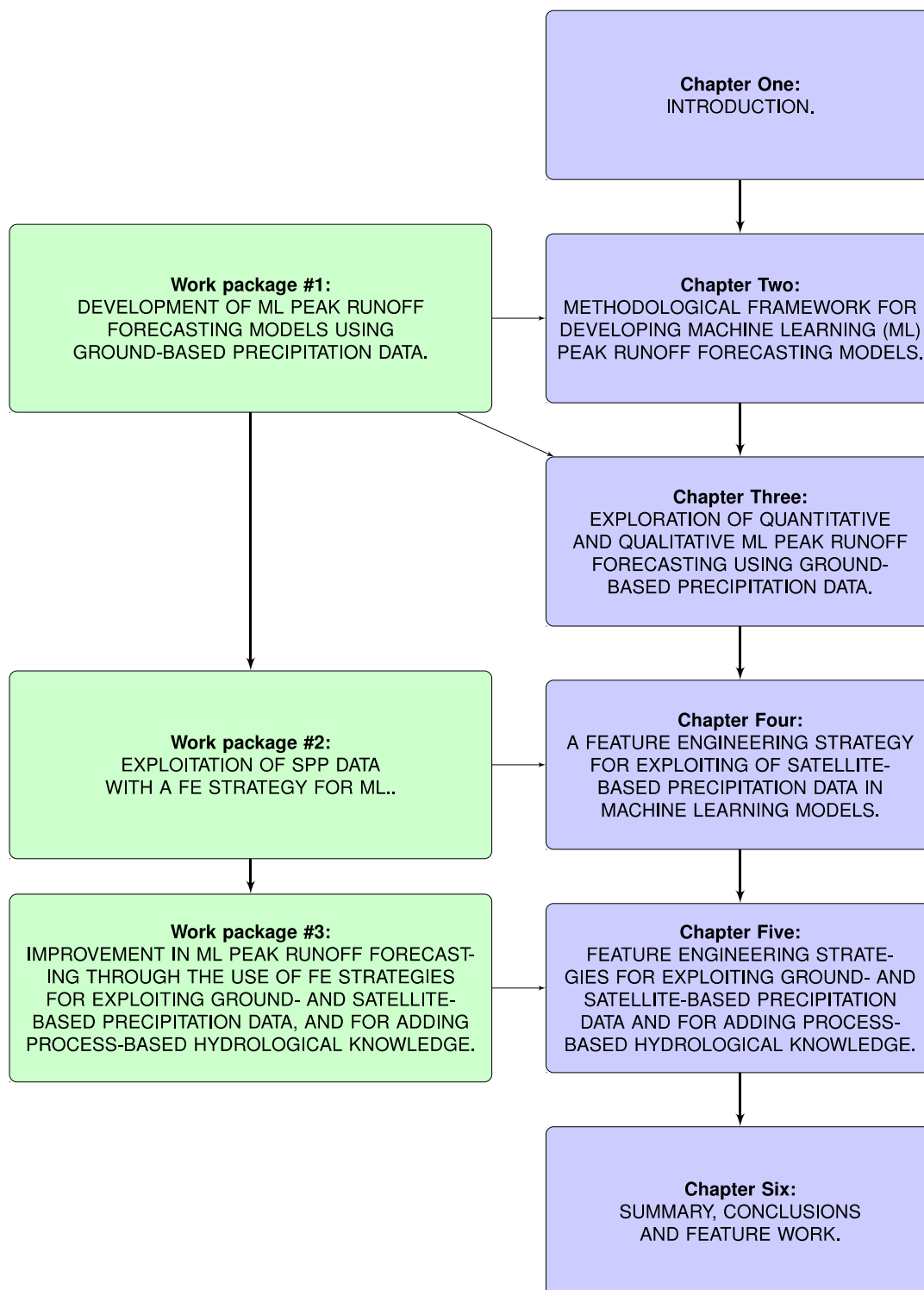


Figure 1.1. Work packages of the doctoral research and their associated thematic chapters.

1.5 Study areas

We have chosen two study areas that represent meso- and macro-scale hydrological systems in the tropical Andes of Ecuador. The first study area is the Tomebamba catchment, covering an area of 300 km², while the second one is the Jubones basin, covering an area of 4400 km². These areas are also typical examples of mountainous regions where it is often difficult to gather comprehensive hydrological data due to limited budgets and the complexity of the terrain. As a result, essential information beyond precipitation and runoff is often lacking. This is evident from the absence of operational flash flood forecasting systems in the Andes, as there are no sufficiently dense ground-based monitoring networks available, as reported in the studies of Dávila [56] and del Granado et al. [57].

Moreover, the selection of meso- and macro-scale systems for this study was based, in part, on the availability of precipitation data. For small-area systems like the Tomebamba catchment, ground-based precipitation monitoring networks can be used in combination with SPPs. However, for larger systems like the Jubones basin, SPPs may be the sole source of precipitation data for flash flood forecasting. As a result, differences in the assimilation process of input data by machine learning (ML) techniques must be carefully considered for effective application in each study area.

1.5.1 A meso-scale hydrological system: the Tomebamba catchment

The Tomebamba catchment is delineated upstream of the Matadero-Sayausí hydrological station, and it is the major water source for the city of Cuenca (third most populated city in Ecuador with 0.6 million inhabitants). The Tomebamba catchment is located in the southeastern flank of the Ecuadorian Andes discharging to the Amazon river and ultimately to the Atlantic Ocean. The catchment area is approximately 300 km², with an elevation range from 2700 to 4400 m a.s.l. (Figure 1.2). Moreover, the Tomebamba is part of the Cajas National Park, which was declared by UNESCO as a World Biosphere Reserve in 2013. Based on the main streams, the Tomebamba can be divided into 6 micro catchments (M1-M6), whose approximate areas are 93.2, 51.4, 73.3, 60.8, 12.7, and 8.6 km².

Paramo vegetation covers 70% of the Tomebamba catchment, followed by native woody species, pastures, and crops among other land cover use. Numerous lakes can be found in the western part of the microcatchments, especially in M2 and M3. Soils in the study area are the result of

volcanic ash accumulation (including andosols and histosols) and are characterized by a high-water retention capacity associated with organic and clay composition [58], [59]. The climate at the Tomebamba is mainly influenced by continental air masses from the Amazon basin [60]. Precipitation depicts a bimodal regime with peaks during March-May and October; the mean (maximum) annual precipitation is around 1110 (1210) mm. Most of the precipitation falls as drizzle [61], with intensities less than $2 \text{ mm}\cdot\text{h}^{-1}$ in more than 95 % of events in the upper [20] and lower parts of the catchment. The average temperature and relative humidity are 6.9 degrees Celsius and 92.1 %, respectively.

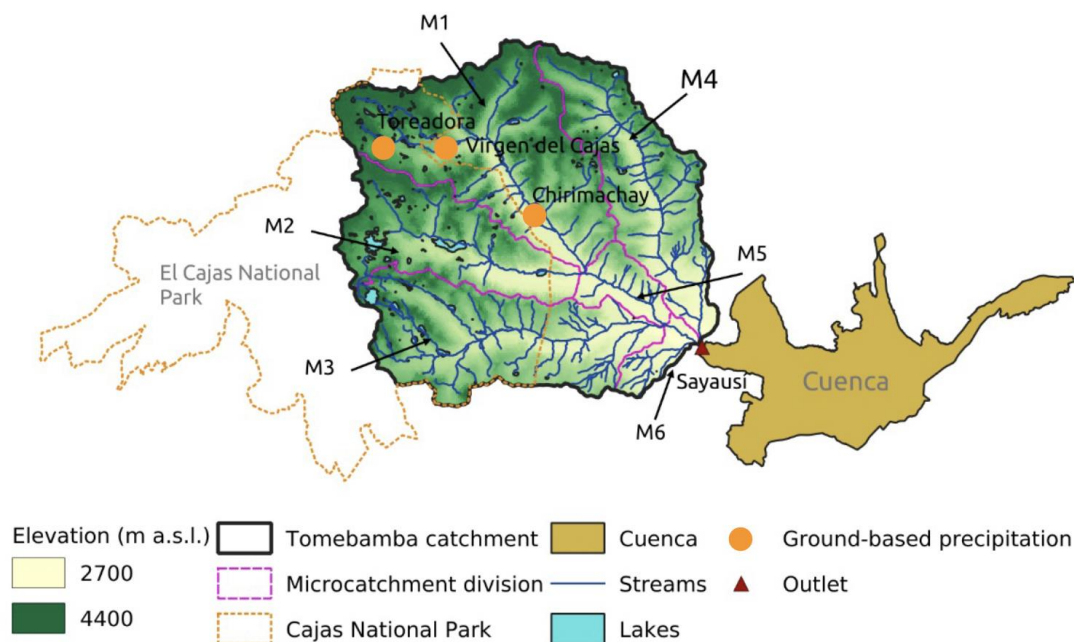


Figure 1.2. The Tomebamba catchment in the southern Ecuadorian Andes. Location of ground-based precipitation stations (Toreadora, Virgen del Cajas, and Chirimachay), and runoff at the outlet (Sayausí).

Dataset

The dataset comprises hourly information on two variables, runoff measured at the outlet of the catchment (Figure 1.2) and precipitation within the catchment for the period January/2015 to May/2021. Runoff time series for the Sayausí station were obtained from the drinking water facility of Cuenca, the Empresa Pública Municipal de Telecomunicaciones, Agua Potable, Alcantarillado y Saneamiento de Cuenca (ETAPA-EP).

Precipitation data were retrieved from ground and satellite sources. Ground estimates were acquired from three rain gauges installed in the upper and middle parts of the catchment, Toreadora at 3395, Virgen del Cajas at 3626, and Chirimachay at 3298 m a.s.l. These rain gauges are located within the microcatchment M1. On the other hand, satellite estimates of precipitation were retrieved from the PERSIANN-CCS database, resulting in 15 pixels-based information for the Tomebamba catchment. Figure 1.3 shows the PERSIANN-CCS coverage over the study catchment as well as a comparison between the annual cumulated precipitation measured by the satellite- and ground-based products for the study period.

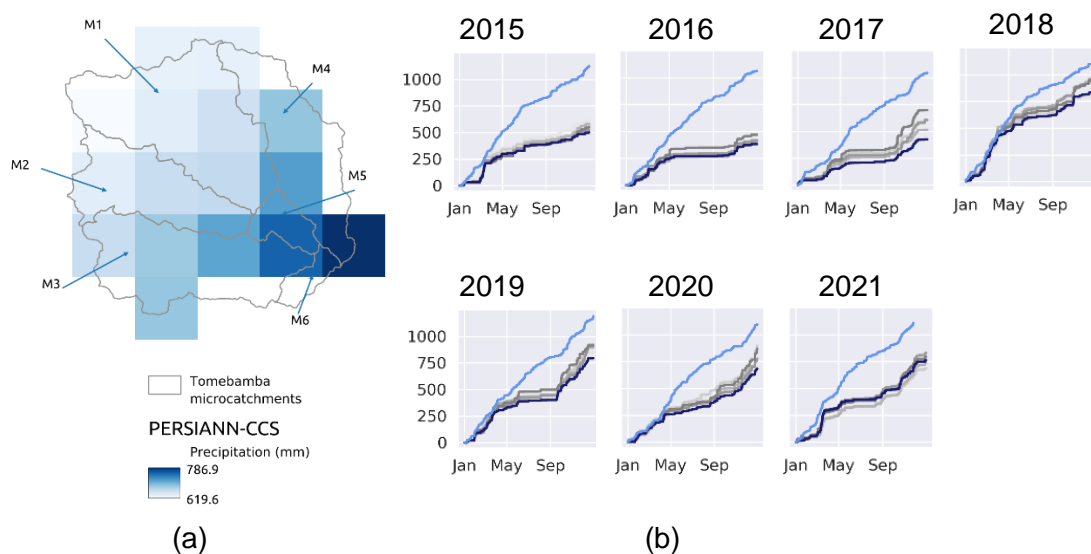


Figure 1.3. (a) PERSIANN-CCS coverage and mean annual precipitation over the Tomebamba catchment. (b) Comparison between annual precipitation measured by ground-based products (average of three rain gauges over microcatchments M1, light blue line) and the PERSIANN-CCS (average over M1, dark blue line). The remaining gray lines depict the PERSIANN-CCS precipitation for microcatchments M2-M6.

1.5.2 A macro-scale hydrological system: the Jubones basin

The Jubones basin is located in the tropical Andes of Ecuador, covering an area of 3391 km² upstream of the Minas-San Francisco (MSF) hydroelectric dam (Figure 1.4). The MSF was constructed and started operating in late 2018. The elevation of the Jubones basin ranges between 1250 to 3920 m above sea level. The climatology of the basin is governed by local topography, the presence of the Andean Mountain range, trade winds, and ocean currents from

the Pacific Ocean. As a result, the spatial distribution of the climatology is very variable, depicting tropical to semi-arid climates according to the Köppen-Geiger classification [62]. As a result, mean annual precipitation in the basin is extremely variable in space, ranging from 290 to 925 mm. Similarly, reported mean annual temperature of the basin ranges from 15 to 28 degrees Celsius [63], yet it is also expected a high variability across the altitudinal gradient.

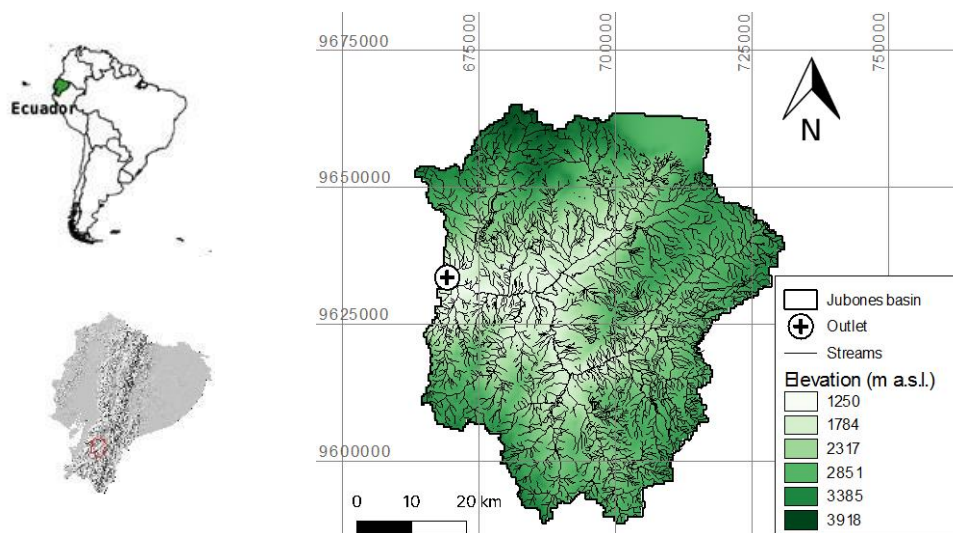


Figure 1.4. The Jubones basin in the Tropical Andes of Ecuador, South America (UTM coordinates).

Dataset

The dataset comprises ~3.5 years of hourly information on two variables, precipitation, and runoff for the period January 2019 to June 2022. Precipitation data were retrieved from two near-real-time databases, the IMERG-Early Run (ER), and the PERSIANN-Cloud Classification System (CCS) products. Data were extracted at the finest temporal resolution (30 minutes and 1 hour for the IMERG-ER and PERSIANN-CCS products, respectively) and then aggregated to the hourly time step. Apart from inner satellite image processing, the most remarkable difference between both precipitation sources is their spatial resolution. The PERSIANN-CCS presents the highest spatial resolution for the study area (0.04° , ~4.4 km), and it is the result of infrared imagery processing and cloud classification using artificial neural networks [64]. Whereas the IMERG-ER delivers 30-min maps with a spatial resolution of 0.1° (~11.1 km) using an approach based on the interpolation of multiple microwave precipitation estimates. It is worth noting the difference in the

number of pixels (timeseries) obtained with each satellite product, 174 and 30 pixels for the PERSIAN-CCS and the IMERG-ER, respectively.

Figure 1.5 compares hourly satellite precipitation measured by both satellite products in the Jubones basin, with mean (maximum) annual precipitation depths of 729 (1167) and 1532 (2759) mm, respectively. The mean annual precipitation differences of 803 and 1592 mm for the mean and the maximum precipitation are attributed to the aforementioned reasons.

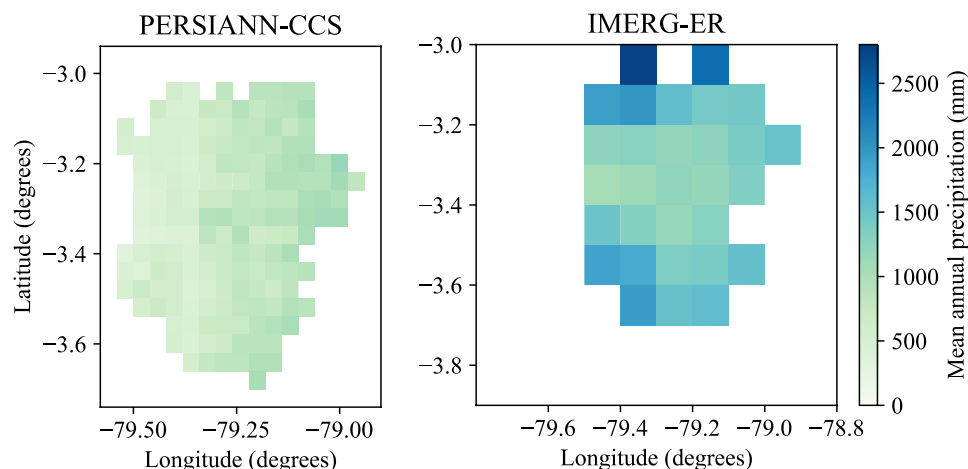


Figure 1.5. Mean annual precipitation measured by the PERSIANN-CCS and the IMERG-ER satellite products for the study period from January 2019 to June 2022 (Jubones basin, Ecuador).

To date, no ground precipitation gauges are operating in the basin. However, a precipitation comparison can be done with the study of [63] to give an idea about SPPs agreement with ground observations. In that study, daily historical data for the period 1982-1998 revealed mean annual precipitation ranging from 471 to 1106 mm in the Jubones basin, which better agrees with the obtained PERSIANN-CCS information. Although it was not possible to perform an hourly validation of the satellite precipitation with ground measurements, this was not a limiting aspect since precipitation is merely an estimator of runoff when ML techniques are employed. Instead, we exploited the spatiotemporal variability of both precipitation signals under the assumption that the overall bias of each of them remains constant for the study area.

On the other hand, hourly runoff data was collected for a hydrological station in the outlet of the basin, i.e., the entrance MSF hydropower dam (see Figure 1.4). The runoff data were facilitated

by the Corporación Eléctrica del Ecuador (CELEC EP, <https://www.celec.gob.ec/>), the company that operates the MSF hydropower dam.

Chapter two: methodological framework for developing machine learning flash flood forecasting models.

Peak runoff including flash flood forecasting can be issued with either quantitative or qualitative approaches [47], [65]–[71]. Quantitative forecasts become a regression problem for addressing hydrological tasks where peak runoff magnitudes are of importance for water resources management or for taking mitigation actions. Some examples are the operation of water supply and water treatment plants, the development of flood early warning systems (FEWS), or for producing inputs to hydraulic models for the delineation of areas prone to flooding.

On the other hand, qualitative forecasting represents a classification problem consisting of classifying floods into distinct categories or river states according to their severity (i.e., runoff magnitude). The utility of categorizing runoff magnitudes is that they can be used for producing runoff susceptibility states in a semaphore-like FEWS (e.g., no-alert, pre-alert, and alert of flooding), which is easy to understand by non-hydrologists (decision-makers and the public). Another application is the mapping of compound flood vulnerability occasioned by the combined effects of hydrological, meteorological, oceanic, and anthropogenic processes (e.g., urbanization) [72], [73]. In general, the advantage of the classification over the regression approach is the possibility to account also for inputs not directly related to flash flood driving forces.

However, regardless of the forecasting approach (regression or classification), the use of ML for flash flood forecasting represents a scientific challenge. This is the selection of the optimal ML technique for developing robust models able to provide accurate forecasts with a sufficient lead time for decision-making. To date, the problem has received scant attention, and as far as our knowledge no previous work has examined the potential and efficacy of ML techniques for flash flood forecasting in complex hydrological systems (i.e., systems that face data scarcity issues regarding the highly variable driving forces behind peak runoffs).

Partially based on the publication of Contreras, P., Orellana-Alvear, J., Muñoz, P., Bendix, J., & Céleri, R. (2021). Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment. Atmosphere, 12(2), 238. <https://doi.org/10.3390/atmos12020238>.

2.1 Aim and objectives

The aim of this chapter is to propose a methodological framework for developing and evaluating peak runoff and flash flood forecasting models with ML techniques.

Objectives:

- To propose a methodological framework for developing and evaluating qualitative (ML classification) peak runoff and flash flood forecasting models.
- To propose a methodological framework for developing and evaluating quantitative (ML regression) peak runoff and flash flood forecasting models.

This chapter is organized into three sections. The first section explores state-of-the-art on peak runoff including flash flood forecasting with ML techniques. In this section, we describe the learning mechanisms of the most-employed ML techniques. With this background, the second section describes the overall process for developing ML forecasting models, starting with the composition of the input feature space, the hyperparameterization of ML models, and finally feature reduction for meeting parsimony criteria. And finally, the third section proposes an evaluation framework for both quantitative and qualitative peak runoff forecasting.

2.2 Review of Machine Learning (ML) techniques

ML techniques can be grouped according to their functionality. For peak runoff and flash flood forecasting, the five most employed groups worldwide are [37]:

- i. Regression techniques to model relations between input-output variables. For instance, linear regression, logistic regression, multivariate adaptive regression splines, etc.).
- ii. Instance-based techniques relying on memory-based learning. This represents a decision problem, some examples are the K-nearest neighbors' algorithm, locally weighted learning, learning vector quantification, etc.).
- iii. Bayesian algorithms using Bayes' theorem on conditional probability, some examples are Naive Bayes, Gaussian Naïve Bayes, Bayesian network, etc.

- iv. Decision tree algorithms, whose idea is to progressively divide the input feature space into data subsets according to feature values or scenarios. For instance, the Random Forest algorithm, regression tree, M5, etc.
- v. Neural Network-based algorithm inspired by the functioning of biological neural networks. The idea is to transform input to outputs through specified transient states which enables the model to learn in a sophisticated way. For instance, perceptron, multi-layer perceptron, long short-term memory networks, radial basis function networks, convolutional networks, etc.

Below, we describe five ML techniques, one from each group. These are logistic regression (LR), K-nearest neighbors (KNN), naive Bayes (NB), random forest (RF), and Multi-layer perceptron (MLP). LR, KNN, and NB are classification techniques, whereas RF and MLP can be employed for both classification and regression applications.

2.1.1 Logistic Regression

Logistic Regression (LR) is a discriminative classification algorithm. LR focuses on the decision boundary between classes. For this, existent relationships between input features are obtained via linear regressions. Then, the conditional probability of belonging to a class is obtained with a logistic (Sigmoid) function useful for outliers (binary classification).

From the obtained probabilities, the LR is used to classify, with regularization, the dependent variables into the created classes. The extension for multiclass problems takes into account all binary classification possibilities. In the end, the classification decision is based on the maximum probability (multinomial LR) calculated with the *softmax* function [74]. The calculated probability for each class is positive with the logistic function and normalized across all classes. The *softmax* function is calculated as follows.

$$softmax(z)_i = \frac{e^{z_i}}{\sum_{l=1}^k e^{z_l}} \quad \text{Equation 1}$$

Where z_i is the i th input of the softmax function, corresponding to class i from the k number of classes.

2.1.2 K-Nearest Neighbors

K-nearest neighbors (KNN) is a non-parametric algorithm for statistical pattern recognition. The classification concept of KNN is based on memory-based learning (intuitive statistical procedure) rather than on a theoretical or analytical background. The classification decision is performed based on a distance function (e.g., Euclidean, Manhattan, Chebyshev, Hamming, etc.). Moreover, the use of multiple neighbors is recommended to deal with noisy features misleading the classification task. In that case, the majority vote of the nearest neighbors determines the classification decision (see formulation in the study of Bishop [74]).

The number of neighbors can be optimized to achieve a global minimum, avoid longer computation times, and reduce the influence of class size. Although the greatest advantage of KNN is its simplicity, a major drawback is that KNN is memory intensive. In practice, the entire training dataset must be stored and computed for the evaluation of new information.

2.1.3 Naïve Bayes

Naïve Bayes (NB) is a classification algorithm that relies on Bayes' theorem, and with the "naive" assumption of independence between features in a class, even when there is dependence [75]. Bayes' theorem can be expressed as follows:

$$P(y|X) = \frac{P(X|y) P(y)}{P(X)} \quad \text{Equation 2}$$

where $P(y|X)$ is the conditional probability of y (hypothesis) given the occurrence of X (features), and X can be defined as $X = x_1, x_2, \dots, x_n$. According to Bayes' theorem equation 2 can be rewritten as:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(x_1|y) P(x_2|y) \dots P(x_n|y) P(y)}{P(x_1) P(x_2) \dots P(x_n)} \quad \text{Equation 3}$$

Moreover, depending on the assumption of the distribution of $P(X|y)$, different NB classifiers can be used. In this regard, the study of [75] proved the optimality of NB under the Gaussian distribution even when there is feature dependence (real application cases). The extension for multiclass problems outputs the class with the maximum probability. For the Gaussian NB algorithm, there are no parameters to be tuned.

2.1.4 Random Forest

Random Forest (RF) is a supervised algorithm where an input feature space (i.e., set of predictors) is related to output through a forest of decision-tree regression models [76]. The RF algorithm mines an ensemble of a multitude of decorrelated decision trees (DTs), where a single DT is a particular model obtained by hierarchically applying a set of conditions. Decorrelation between DTs is assured by applying a bagging technique aimed at growing DTs from different randomly resampled training subsets. The mean prediction of the individual trees is the solution for regression applications, whereas, for classification problems, the result is the class with the majority of votes.

In summary, the highest-level node of a DT is split into two self-similar lower-level nodes according to simple conditions related to the input data, and until stopping criteria are met. This process is repeated to obtain purer nodes than their precedent ones. The split of each node is performed by randomly selecting several features from the total number of features. The random component is used to both resample the data and to determine the optimal successive features (directions) for splitting the data. Every terminal node represents a regression or classification model applying in that very node only. A complete description of the RF algorithm can be found in the studies of Breiman [77], [78].

Among RF hyperparameters, some control the structure of decision trees (e.g., depth of the tree minimum number of samples in the leaf nodes, the maximum number of leaf nodes in the trees, etc.), some others control diversity in the forest (e.g., number of trees, number of features for the splitting, percentage of dataset employed for build trees), and more advanced hyperparameters define the internal divisions on each tree (e.g., decision quality, and minimum samples for dividing an internal node, etc.). Now, considering the efficiency and popularity of the RF, we conducted a sensibility analysis experiment for determining the most relevant RF hyperparameters for runoff forecasting. From this analysis, it is now well established that considerably higher accuracies and reduction of equifinality are obtained for an optimal number of trees and an adequate combination of the depth of the tree and the number of features [79]. Moreover, for lead times exceeding the concentration time of the catchment, more effort must be put into the hyperparameterization since forecasts' efficiency depends more on an appropriate hyperparameterization.

2.1.5 Multi-layer Perceptron

The multi-layer perceptron (MLP) is a type of fully-connected feedforward artificial neural networks (ANNs) that can be used for both regression and classification applications. A perceptron is a linear classifier used for separating inputs into two categories for producing a single output. The architecture of the MLP is multiple neurons allocated in fully-connected multiple layers. The first layer of MLP corresponds to the input feature space, and all other nodes are employed for relating inputs to outputs through the use of linear combinations with weights and bias terms together with an activation function.

The advantage of MLP when compared to the single-layer case is that MLP can reproduce non-linear functions by the addition of several so-called hidden layers. For the classification case, the probabilities of belonging to a class are calculated with the *softmax* function (equation 1). The efficiency of MLP can be evaluated with a logistic loss function based on the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS). A more detailed and comprehensive description of MLP can be found in [80].

2.3 Methodology for developing ML peak runoff forecasting models

The methodology for developing ML peak runoff including flash flood forecasting models was based on the study of Muñoz et al. [47], and it is summarized in Figure 2.1. In short, the first step of the methodology contemplates the composition of an input feature space from which ML forecasting models can learn. An input feature space is composed of three components: i) features coming from the available timeseries of precipitation and runoff, ii) past features (lags) of the information in the first component, according to statistical analyses, and iii) additional features derived from the application of FE strategies (Chapters Four and Five of this thesis).

The second step is referred to the model construction process, where it is initially required to split the input feature space into training and testing subsets. For timeseries modeling, the data splitting contemplates continuous hydrological periods for training/testing rather than randomly selected samples. Whereas for event-based modeling, a fraction of the events can be selected for training (around 70 %) and the remaining events are left for testing purposes. For the cases when the number of events is reduced, a more exhaustive evaluation consists of using the leave-one-out cross-validation (LOOCV) algorithm [81].

Moreover, an hyperparameterization task is performed for the specific hyperparameters of each ML technique. This task is carried out on the training subset and can be done either by using a full or random grid-search (RGS) procedure for finding the optimal combination of hyperparameters for a selected efficient metric. Then, a feature selection algorithm is applied for retaining only relevant features and trimming off noisy features for the forecasting process.

The last step encompasses model evaluation on the testing subset according to a combination of performance metrics (comparison between forecasts and observations) and graphical analyses to account for flash floods. The metrics selected depend on the ML modeling approach, i.e., classification or regression. Below we detailed some of the steps of the methodology.

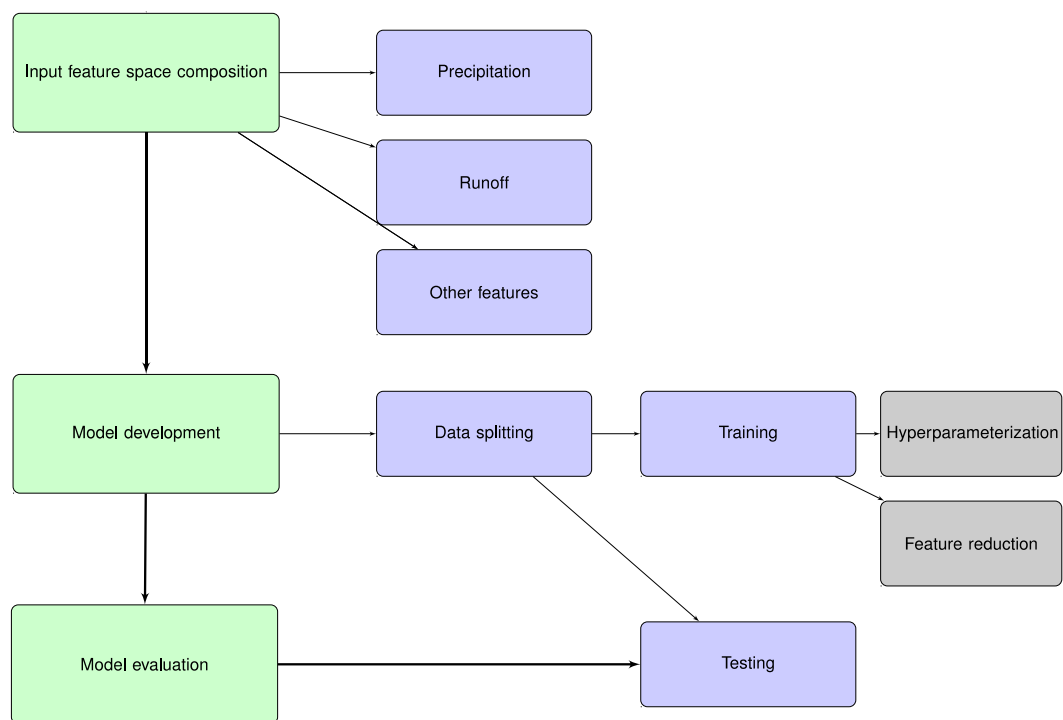


Figure 2.1. Step-wise methodology scheme for developing ML peak runoff forecasting models.

2.3.1 Statistical lag analyses

For ML peak runoff forecasting, the information on endogenous (i.e., runoff) and exogenous variables (e.g., precipitation) at the current time is not sufficient for describing the inner relations in the runoff generation process. Additional information can be derived from past endogenous and exogenous data (lagged information).

The utility of lagged information of exogenous variables is the addition of physically relevant processes to the model. For instance, in the case of precipitation, this information can be added as additional features to the input feature space to mimic the soil moisture state of the system. Thus, for the cases when the soil is non-saturated (dry periods), lagged precipitation information informs ML models that an initial precipitation water volume is used for saturating the soil before becoming overland flow. Conversely, for already saturated conditions in the system (wet periods), lagged precipitation indicates that most of the precipitation volume becomes runoff. Thus, the lack of consideration of antecedent soil moisture conditions in runoff models has been reported to lead to runoff underestimation and overestimation issues during dry and wet periods, respectively [36].

The statistical analyses contemplate a qualitative method proposed by [82] for determining the adequate number of lags from endogenous and exogenous variables. The application of this method avoids complex and computationally-intensive trial-and-error procedures for determining the optimal number of lags. For exogenous variables, the optimal number of lags is determined through cross-correlation analyses between each exogenous variable and the endogenous variable (i.e., runoff). In addition, a correlation threshold has to be set for removing the neglectable influence of certain lags on runoff. The threshold value depends on the correlation level between the station and runoff timeseries. However, Muñoz et al. [47] have suggested a threshold value of 0.2. For runoff, the number of lags is determined by the autocorrelation function (ACF), and the partial autocorrelation function (PACF). The ACF and PACF are applied with 95 % confidence levels.

2.3.2 Model hyperparameterization

Once the input feature space is composed, the optimal architecture or combination of hyperparameters (for a given ML technique) has to be defined during the training stage of the modeling process. The optimal hyperparameter combination is intended in this research to maximize accuracy. For this, we selected performance metrics of accuracy for both ML problems; these are the Nash-Sutcliffe efficiency (NSE) for regression, and the f1-macro score for classification. The NSE and f1-macro scores are described below in section 2.4.

In terms of ML computational cost, both the full or RGS procedures can be applied together with a k-fold cross-validation scheme to reduce computation times and overfitting. This means that the full searching procedure develops models with all possible combinations of hyperparameters, and

determines the optimal combination based on maximum accuracy between observations and forecasts. Whereas for the RGS procedure, the evaluation focuses on a discretized continuous hyperparameters' domain. In any case, the k-fold cross-validation scheme means that the training dataset is split into k folds (subsets); the models were iteratively fitted on the k-1 folds, and accuracy is evaluated on the remaining one.

The LR, KNN, NB, RF, and MLP techniques as well as the hyperparameter searching procedures were implemented through the scikit-learn package for ML in Python® (Pedregosa et al., 2011). Table 2.1 presents the hyperparameters for each ML technique and their search space for tuning.

Table 2.1. Model hyperparameters of the most-employed ML techniques for flash flood forecasting.

ML technique	Hyperparameters				
LR	<i>C</i>	<i>penalty</i>			
KNN	<i>neighbor's</i>	<i>weights</i>	<i>metric</i>	<i>algorithm</i>	
RF*	<i>Number of trees</i>	<i>max_features</i>	<i>max_depth</i>	<i>min_samples_leaf</i>	<i>min_samples_split</i>
MLP	<i>solver</i>	<i>max_iter</i>	<i>alpha</i>	<i>hidden_layers</i>	

* Most relevant hyperparameters for runoff forecasting according to the study of [79]

2.3.3 Feature space reduction

The development of ML models deals with the assimilation of high-dimension and complex input feature spaces. Assimilation is referred to the ability of ML models to exploit input information to find their relations with the output variable (i.e., runoff) during the learning process (training). High dimensionality results from the use of a large number of features necessary for learning spatial and temporal relations between inputs and outputs. On one hand, the use of high-dimension spaces demands substantial amounts of memory and computational costs, and on the other hand, the use of high-dimension feature spaces might include information that might not be relevant to the model. In other words, although the inclusion of a certain feature might be conceptually correct, the internal relations between that feature and the target, and the interactions between features might be noisy rather than useful. Thus, the ML efficiency is reduced.

For these reasons, a feature space reduction procedure is recommended for including only the features relevant to the model and trimming off the noisy ones. Apart from shortening computation times, in some cases, feature space reduction even improves the model's accuracy [83].

Feature space reduction can be done in several ways. One intuitive approach is the application of a principal component analysis (PCA). The PCA is aimed at finding the dimension of maximum variance to exclude correlated features that do not add information to the model. However, since each ML technique assimilates data differently, instead of defining a fixed threshold of variance explanation (e.g., 80-90%), the optimal number of components can be treated as an additional ML hyperparameter. The threshold selected ultimately depends on the specific problem, i.e., on the required trade-off between accuracy, computational cost, and model complexity to address parsimony criteria. Another alternative for feature reduction is the application of a process known as feature selection. Feature selection can be done based on e.g., a variance sensitivity analysis, univariate statistical tests, or recursive elimination, among other methods. Here, we rely on the sensitivity analysis proposed by [84].

The selected sensibility analysis measures the output's variance produced by a single feature without the influence of the feature's interaction. As a result, the isolated impact of each feature can be calculated to keep only the features accounting for a certain total relative importance, for instance, 80 %. Thus, the remaining features can be considered unimportant, and removed from the input feature space.

The variance (V_k) and its relative importance (R_k) can be calculated with equations 4 and 5, respectively.

$$V_k = \frac{\sum_{j=1}^L [\hat{y}_{t-k}(j) - \overline{\hat{y}_{t-k}(j)}]^2}{L-1} \quad \text{Equation 4}$$

$$R_k = \frac{V_k}{\sum_{i=1}^m V_i} \times 100 \% \quad \text{Equation 5}$$

Where $\hat{y}_{t-k}(j)$ is the model output when all m features are held at their average values except \hat{y}_{t-k} , which can vary through its entire range with $j \in \{1, \dots, L\}$ levels.

2.4 Evaluation of machine learning peak runoff forecasting models

Model efficiency or performance can be determined by a direct comparison between model outputs (forecasts) and observations. Outputs and observations can be either quantities (regression problems) or labels (classification). Thus, considering the nature of the problem we develop an evaluation framework for both cases.

2.4.1 Evaluation of ML qualitative models

The evaluation of classification models for the cases of extreme values (e.g., peak flows) turns into an imbalanced classification problem. This is because peak runoffs (minority class) are rare events that occur with a lower frequency when compared to normal conditions runoff magnitudes (majority class). Moreover, a classification problem aimed at forecasting more than two (non-binary) flood warning labels such as no-alert, pre-alert, and alert of flash floods becomes a multi-class problem. The imbalance problem for this case is that ML classification algorithms focus on the minimization of the overall error rate, i.e., the minimization focuses on the majority class (no-alert) which leads to high errors in the minority (pre-alert and alert classes) [85].

The imbalance problem can be treated by resampling the class distribution of the data to obtain an equal number of samples per class. However, a more accepted approach relies on training ML models with the assumption of imbalanced data. Training imbalanced models contemplates the penalization of errors in samples belonging to the minority classes rather than under-sampling or over-sampling data. In practice, this means that for a given efficiency metric, its overall score is the average metric for all classes after being multiplied by a weight factor according to class distribution. The weight factors for each class can be calculated using equation 11.

$$w_i = \frac{N}{C n_j} \quad \text{Equation 11}$$

where w_i is the weight of class i , N is the total number of observations, C is the number of classes, and n_j the number of observations in class i . This implies that higher weights will be obtained for minority classes.

2.4.1.1 Efficiency metrics

The efficiency metrics for imbalanced datasets can be derived from the well-known confusion matrix. These are the *f1 score*, the geometric mean (*G-mean*), and the logistic regression loss (*Log loss*) score. For a proper model evaluation, it is suggested to use the *f1 score*, *G-mean* and *Log loss* together since they complement each other [85]–[90].

F1 score

The *f score* is a metric that relies on precision and recall, which is an effective metric for imbalanced problems. When the *f score* as a weighted harmonic mean, we name this score *f1 score*. The latter score can be calculated with equation 12.

$$f1\ score = \frac{2 * Precision * Recall}{(Precision + Recall)} \quad \text{Equation 12}$$

Where precision and recall are defined with the following equations:

$$Precision = \frac{TP}{TP + FP} \quad \text{Equation 13}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{Equation 14}$$

Where *TP* stands for True Positives, *FP* for False Positives, and *FN* for False Negatives.

The *f1 score* ranges from 0 to 1, indicating perfect precision and recall. The advantage of using the *f1 score* compared to the arithmetic or *G-mean* is that it penalizes models most when either the precision or recall is low. However, classifying a No-Alert label as Alert might have a different impact on the decision-making than when the opposite occurs. This limitation scales up when there is an additional state, e.g., Pre-alert. Thus, the interpretation of the *f1 score* must be taken with care. For multiclass problems, the *f1 score* is commonly averaged across all classes and is called the *f1 – macro score* to indicate the overall model performance.

Geometric-mean

The geometric-mean (*G-mean*) measures simultaneously the balanced performance of TP and True Negative (TN) rates. This metric gives equal importance to the classification task of both the

majority (No-alert) and minority (Pre-alert and Alert) classes. The G-mean is an evaluation measure that can be used to maximize accuracy to balance TP and TN examples at the same time with a good trade-off [87]. The G-mean can be calculated using equation 15.

$$G\text{-mean} = \sqrt{(TP_{rate} * TN_{rate})} \quad \text{Equation 15}$$

Where TP_{rate} and TN_{rate} are defined by:

$$TP_{rate} = Recall \quad \text{Equation 16}$$

$$TN_{rate} = \frac{TN}{TN+FP} \quad \text{Equation 17}$$

The value of the *G-mean* metric ranges from 0 to 1, where low values indicate deficient performance in the classification of the majority class even if the minority classes are correctly classified.

Logistic regression loss

The metric logistic regression loss (*Log loss*) measures the performance of a classification model when the input is a probability value between 0 and 1. It accounts for the uncertainty of the forecast based on how much it varies from the actual label. For multiclass classification, a separate *Log loss* is calculated for each class label (per observation), and the results are summed up. The *Log loss* score for multi-class problems is defined as:

$$Log\ loss = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij}) \quad \text{Equation 18}$$

where N is the number of samples, M the number of classes, y_{ij} equal to 1 when the observation belongs to class j ; else 0, and p_{ij} is the predicted probability that the observation belongs to class j . Starting from 0 (best score), the *Log loss* magnitudes increase as the probability diverges from the actual label. The *Log loss* penalizes worse errors more harshly to promote conservative predictions. For probabilities close to 1, the *Log loss* slowly decreases. However, as the predicted probability decreases, the *Log loss* increases rapidly.

Although we can directly compare performance metrics of ML alternatives and claim to have found the best one based on the score, it is not certain whether the difference in metrics is real or the

result of statistical chance. Different statistical frameworks are available allowing us to compare the performance of classification models (e.g., a difference of proportions, paired comparison, binomial test, etc.).

Among them, Raschka et al. [91] recommend using the chi-square test to quantify the likelihood of the samples of skill scores, being observed under the assumption that they have the same distributions (null hypothesis). The assumption states, therefore, that the results (error rates) of the two ML models are equal. If the null hypothesis is rejected, it can be concluded that any observed difference in performance metrics is due to a difference in the models and not due to statistical chance. In practice, the chi-square test applied to ML flash flood forecasting models can be used to assess whether the difference in the observed proportions of the contingency tables of a pair of ML algorithms (for a given lead time) is significant. For this, a significance value of 0.05 is often agreed for proving the statistical significance of model improvements/degradations.

2.4.2 Evaluation of ML quantitative models

For the evaluation of regression models, previous research has established the need of using a combination between goodness-of-fit metrics and graphical analyses [36], [92]. This is because a single efficiency metric represents the mean performance of a model without consideration of the unbalanced influence of peak runoffs such as flash floods. Moreover, graphical interpretation techniques serve to further identify model strengths and weaknesses that might be hidden in a single value measure of efficiency.

2.4.2.1 *Efficiency metrics*

In terms of goodness-of-fit (efficiency) metrics, we used a collection of four indices following the guidelines of Moriassi et al. [92]. Among them, the Nash-Sutcliffe Efficiency (*NSE*) [93] can be set as the reference metric for measuring and comparing the overall fit of model forecasts and observations. The complementary metrics are the Kling-Gupta Efficiency (*KGE*) [94] to account for extreme value underestimations/overestimations, the Percent Bias (*PBIAS*), and the Root Mean Square Error (*RMSE*). The corresponding equations are as follows:

$$NSE = 1 - \frac{\sum_{i=1}^n (Q_S(i) - Q_O(i))^2}{\sum_{i=1}^n (Q_O(i) - \bar{Q}_O)^2} \quad \text{Equation 19}$$

$$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad \text{Equation 20}$$

$$PBIAS = \frac{\sum_{i=1}^n (Q_o - Q_s)}{\sum_{i=1}^n Q_o} \quad \text{Equation 21}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Q_s - Q_o)^2} \quad \text{Equation 22}$$

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, being $NSE = 1$ 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 $KGE = 1$ as 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.

Moreover, for event-based forecasting, the LOOCV algorithm treats each event as an independent testing dataset while the remaining events are used for training purposes. In the end, the overall model efficiency corresponds to the average NSE obtained for all scenarios (each event used as a testing dataset).

2.4.2.2 Graphical techniques focused on flash-floods

For the evaluation of peak runoffs and flash floods, we complemented the goodness-of-fit metrics with graphical techniques including the peak values frequency distribution, and the Box-Cox transformation for runoff. For the first case, the behavior of the distribution towards the tail for both observations and forecasts determines whether the performances of the models are acceptable for peak conditions. Whereas, for the second case, the Box-Cox transformation aims at dealing with the oversensitivity of model residuals for peak values. The Box-Cox transformation is calculated with the following equation:

$$Box - Cox (Q) = \frac{Q^{\lambda-1}}{\lambda} \quad \text{Equation 23}$$

Where Q is runoff, and λ is the parameter graphically calibrated until reaching homoscedasticity in the residuals (i.e., constant standard deviation). For runoff, a reference value of $\lambda = 0.25$ is suggested in the study of [95].

An additional consideration for peak flow evaluation is the serial dependence of runoff magnitudes to the timescale employed (hourly scale for flash floods). This means that the graphical evaluation of flows suffers from a higher representation of low flows when compared to the reduced number of peak flows in the timeseries. Moreover, the serial dependence for extreme peak flows is stronger for shorter timesteps. To overcome this issue, nearly independent observations must be selected by splitting the runoff timeseries in events and using one value per event. This can be done using the Peak-over-threshold approach [95].

Chapter three: exploration of quantitative and qualitative machine learning flash flood forecasting using ground-based precipitation data.

Related publications:

- ❖ **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: Djalante, R., Bisri, M.B.F., Shaw, R. (eds) *Integrated Research on Disaster Risks. Disaster Risk Reduction.* Springer, Cham. https://doi.org/10.1007/978-3-030-55563-4_2.
 - ❖ **Muñoz, P., Orellana-Alvear, J., Bendix, J., Feyen, J., & Célleri, R. (2021).** *Flood Early Warning Systems Using Machine Learning Techniques: The Case of the Tomebamba Catchment at the Southern Andes of Ecuador.* *Hydrology*, 8(4), 183. <https://doi.org/10.3390/hydrology8040183>.
-

On this chapter, we present two machine learning (ML) case studies, one for qualitative and one for quantitative flash flood forecasting. Each case study has a different aim. For the qualitative (classification) forecasting application, we aimed at developing a semaphore-like flood early warning system (FEWS) with three warnings (river states), no-alert, pre-alert, and alert of flash flooding. Moreover, considering that very little attention has been paid to the selection of the optimal ML qualitative technique, we evaluated and compared the forecasting efficiencies of FEWSs powered by the most-employed ML techniques for flash flood forecasting.

On the other hand, for the quantitative (regression) forecasting application, we investigated the ability of one single ML technique, the random forest (RF), for developing operational runoff forecasting models with special attention to flash floods. We selected the RF for three main reasons. The first one is that extensive research for quantitative hydrological forecasting has already demonstrated that the RF algorithm is a suitable ensembled tool for obtaining accurate forecasts, producing promising results in comparison to more advanced ML techniques such as support vector machines (SVMs) and artificial neural networks (ANNs). The second reason is

certain RF advantages for its implementation in real-time applications, for instance, fewer parameters to calibrate and higher accuracies when compared to other ML techniques, model robustness, overfitting reduction, and the possibility to interpret results through calculation of estimator importance [47], [48], [55], [77], [79], [85], [96]–[98]. These advantages have some implications, for instance, the construction process of RF models is shorter when compared to other ML techniques. Another example is that, contrary to ANN models, RF models do not require input normalization for the training stage of models. And the third reason is the opportunity to exploit the ability of the RF algorithm to deal with small-size samples and complex data structures as encountered in complex systems such as the Andes (extreme heterogeneity and temporal and spatial data scarcity, issues [54], [99]–[102]).

Both the ML classification and regression approaches are applied in a meso-scale hydrological system, the Tomebamba catchment in southern Ecuador. For both case studies, we used the existing ground-based precipitation data.

3.1 Aim and objectives

To explore the use of ML for flash flood forecasting using ground-based precipitation data.

Objectives:

- To develop and evaluate qualitative ML flash flood forecasting models using ground-based precipitation data.
- To develop and evaluate quantitative ML flash flood forecasting models using ground-based precipitation data.

3.2 Qualitative ML flash flood forecasting

3.2.1 Introduction

Flood Early Warning Systems (FEWSs) have proved to be cost-efficient solutions for life preservation, damage mitigation, and resilience enhancement [57], [103]. To date, there is no report of any operational FEWS in the Andean region for scales other than continental [56], [57], [104]. An alternative attempt in Peru was targeted to derive daily maps of potential floods based on the spatial cumulated precipitation in the past days [105]. Other endeavors in Ecuador and Bolivia focused on the monitoring of the runoff in the upper parts of the catchment to predict the

likelihood of flood events in the downstream basin area [56], [106]. However, such studies are unsatisfactory as countermeasures against floods and especially flash floods, where it is required to have reliable and accurate forecasts with lead times shorter than the response time between the farthest precipitation station and the runoff control point.

In this context, this classification application aims at developing FEWSs for the Tomebamba catchment. For this, the most-employed ML-based classification models were implemented and a comparison and ranking of the efficiency of flash flood forecasting was performed. The ML models were evaluated concerning their capacity to forecast three flood warning stages or river states (No-alert, Pre-alert, and Alert of flash floods) for varying lead times of 1, 4, and 6 hours (flash-floods), but also 8 and 12 hours to further test whether the lead time can be satisfactorily extended with sufficient accuracy.

3.2.2 Dataset and processing

Data comprises 4 years of hourly time series of precipitation and runoff for the Tomebamba catchment (see Figure 3.1). The study period runs from January 2015 to January 2019. Precipitation data were derived from 3 tipping-bucket rain gauges, respectively Toredora (3955 m a.s.l.), Virgen (3626 m a.s.l.), and Chirimachay (3298 m a.s.l.), installed along the altitudinal gradient of the catchment. Whereas runoff measurements were obtained from the Matadero-Sayausí hydrological station (2693 m a.s.l.).

For the labeling of flash flood warnings or river states, we rely on the definitions of the Empresa Pública Municipal de Telecomunicaciones, Agua Potable, Alcantarillado y Saneamiento de Cuenca (ETAPA-EP). ETAPA-EP is the local water company for the city of Cuenca and defined three flood alert levels at the Matadero-Sayausí station in the Tomebamba catchment. These are: i) No-alert when runoff at the outlet of the catchment is less than $30 \text{ m}^3/\text{s}$, ii) Pre-alert when runoff varies between 30 and $50 \text{ m}^3/\text{s}$, and iii) Alert warning when runoff exceeds $50 \text{ m}^3/\text{s}$. With these definitions, it is clear that the No-alert warning stands for the majority of the data, while the Pre-alert and Alert warnings comprise the minority yet the most dangerous classes (Figure 3.1). Moreover, for ML training and testing, we split the available dataset into training (from 2015 to 2017), and testing (2018) subsets.

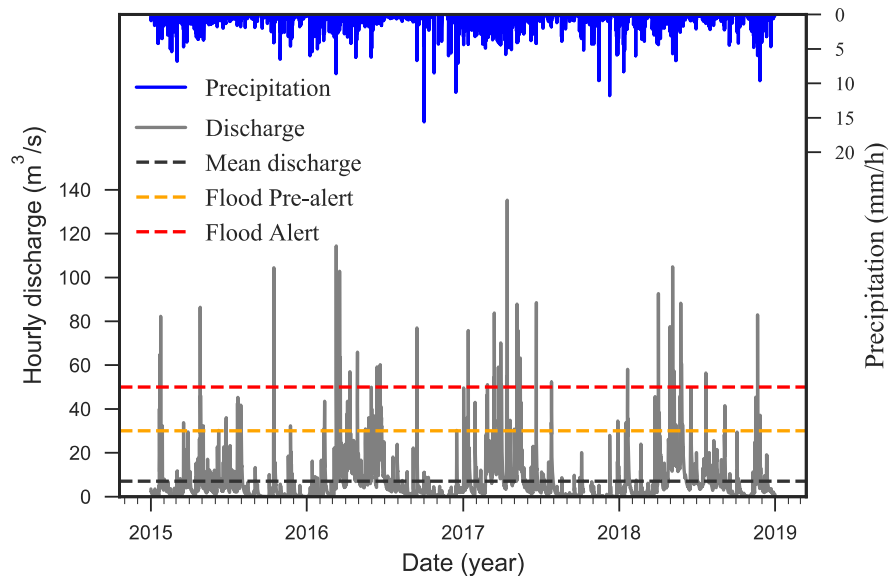


Figure 3.1. Time series of precipitation (Toreadora) and discharge (Matadero-Sayausi). Horizontal dashed lines indicate the mean runoff and the currently employed flood alert levels for labeling the Pre-alert and Alert flood warnings classes.

3.2.3 Methodology

Figure 3.2 depicts the methodology employed for the development of ML classification forecasting models. The following is based on the methodology described in Chapter Two for qualitative ML modeling.

In summary, the available dataset (precipitation and labeled runoff) is split into training and testing subsets. Then, the corresponding input feature spaces are composed according to statistical analyses and the lead times selected. For a given lead time, the selected ML techniques (LR, KNN, RF, NB, and MLP) are used to construct, hyperparameter, and evaluate the corresponding forecasting models. Finally, the intercomparison and ranking of ML models across lead times were supported by a statistical test aimed at proving significance in improving/deterioration of the efficiency.

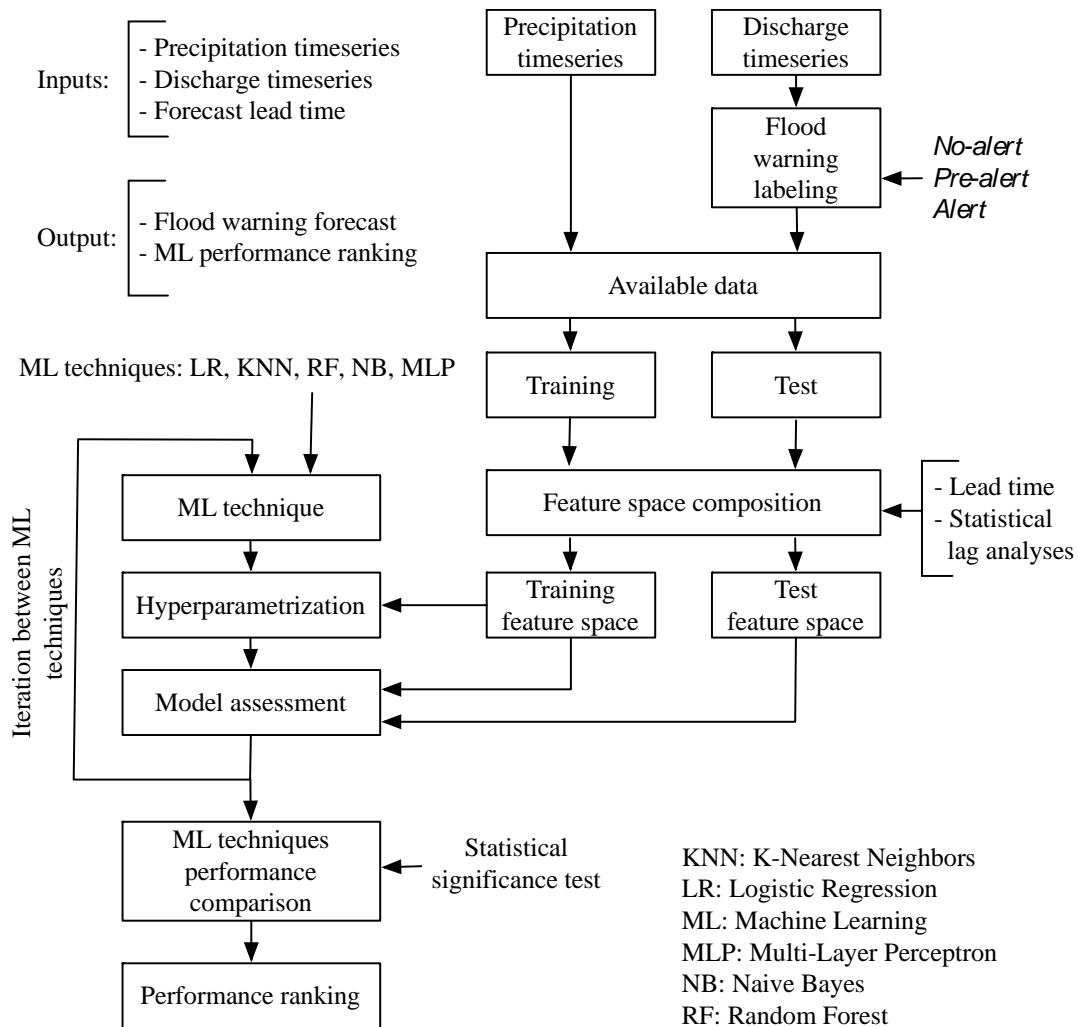


Figure 3.2. Methodologic scheme for the development and testing of ML flash flood forecasting models.

3.2.3.1 Feature space composition

Concerning input feature space composition, we defined specific training and test feature spaces. Feature spaces were composed of features (predictors) coming from two variables: precipitation and runoff. The amount of precipitation and runoff features (current time and lagged instances) was determined according to statistical analyses on precipitation and runoff timeseries.

For precipitation, the number of lags from each station was selected by setting up a Pearson correlation threshold of 0.2 according to the recommendations of [47]. Whereas for runoff, we

relied on correlations from partial and auto-correlation analyses applied to the runoff timeseries with the consideration that the number of runoff features triples since we replace each runoff feature with 3 features (one per flood warning class). This is a process known as one-hot encoding or binary encoding. In practice, each feature denotes 0 or 1 when the correspondent alarm stage is false or true, respectively. Moreover, features in the input spaces were transformed using a standardization process before the computation stage of the KNN, LR, NB, and NN algorithms. To this end, we subtracted the mean and scale it to unit variance, resulting in a distribution with a standard deviation equal to 1 and a mean equal to 0.

3.2.3.2 Model hyperparameterization

All ML techniques and the random-grid search (RGS) hyperparameterization procedure were implemented through the scikit-learn package for ML in Python® [107]. For the hyperparameterization, we selected the *f1 score* as the objective function for finding the optimal hyperparameter combination for each forecasting model. Table 3.1 presents the relevant hyperparameters for each ML technique and the search space selected for tuning. It is worth noting that feature reduction was applied by adding a hyperparameter for controlling the number of components for the PCA.

Table 3.1. Model hyperparameters and their ranges/possibilities for tuning.

ML technique	Hyperparameters				
LR	<i>C</i>	<i>penalty</i>			
	0.001 - 1000	{ 'l1', 'l2' }			
KNN	<i>neighbor's</i>	<i>weights</i>	<i>metric</i>	<i>algorithm</i>	
	3 - 75	{ 'uniform', 'distance' }	{ 'euclidean', 'manhattan', 'minkowski' }	{ 'auto', 'ball_tree', 'kd_tree', 'brute' }	
RF	<i>Number of trees</i>	<i>max_features</i>	<i>max_depth</i>	<i>min_samples_leaf</i>	<i>min_samples_split</i>
	50 - 1000	{ 'auto', 'sqrt', 'log2' }	50 - 1000	1 - 500	1 - 500
MLP	<i>solver</i>	<i>max_iter</i>	<i>alpha</i>	<i>hidden_layers</i>	
	{ 'lbfgs' }	10 - 5000	1 E-9 - 0.1	1 - 16	

3.2.3.3 Model performance evaluation

For model evaluation, we employed a compendium of metrics accounting for imbalanced and multiclass problems (see Chapter Two). For the imbalance problem, we employed weighting factors according to the frequency of samples in each warning class (No-alert, Pre-alert, and Alert of a flash flood). The selected efficiency metrics are the *f1 score*, the *G-mean*, and the *Log loss*. Moreover, for the comparison and ranking of ML techniques, we used the chi-squared test to assess whether the efficiency difference of a pair of ML algorithms is significant under a value of 0.05. In all cases, the MLP model was used as the base model to which the other models were compared. This was done since MLP models depicted the highest efficiencies when compared to the remaining ML techniques (see next section).

3.2.4 Results and discussion

This section presents the results of the flood forecasting models developed with the LR, KNN, RF, NB, and MLP techniques, and for lead times of 1, 4, 6, 8, and 12 hours. For each model, we addressed the forecast of three flood warnings, *No-alert*, *Pre-alert*, and *Alert*. First, we present the results of the feature space composition process, taking the 1-hour lead time case as an example. Then, we show the results of the hyperparameterization for all models, followed by an evaluation and ranking of the performance of the ML techniques.

3.2.4.1 Feature space composition

Figure 3.3 shows the results of the runoff lag analyses for the 1-hour flood forecasting model. Figure 3.3a plots the ACF and its corresponding 95% confidence interval from lag 1 up to 600 (hours). We found a significant correlation up to a lag of 280 h (maximum correlation at the first lag), and thereafter, the correlation fell within the confidence band. To complement the ACF, Figure 3.3b presents the runoff PACF and its 95% confidence band from lag 1 to 30 h. Here, we found a significant correlation up to lag 8 h (first lags outside the confidence band). As a result, based on the interpretation of the ACF and PACF analyses, and according to [47], we decided to include 8 runoff lags (hours) for the case of 1-hour flood forecasting models.

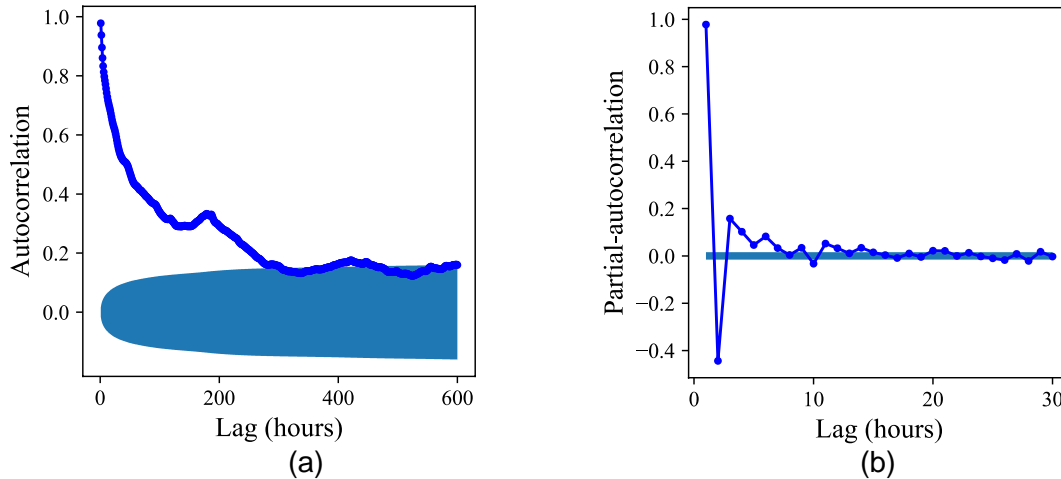


Figure 3.3. (a) Autocorrelation function (ACF) and (b) Partial-autocorrelation function (PACF) of the Matadero-Sayausí (Tomebamba catchment) discharge series. The blue hatch indicates in each case the correspondent 95% confidence interval.

On the other hand, Figure 3.4 plots Pearson’s cross-correlation between precipitation at each location and runoff at the Matadero-Sayausí station. For all precipitation locations, we found a maximum correlation at lag 4 (maximum 0.32 for Chirimachay). With the fixed correlation threshold of 0.2, we included 11, 14, and 15 lags for Virgen, Chirimachay, and Toreadora stations, respectively.

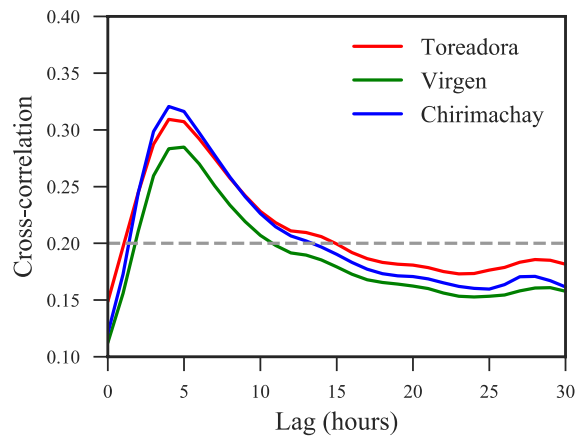


Figure 3.4. Pearson’s cross-correlation comparison between the Toreadora (3955 m a.s.l), Virgen (3626 m a.s.l.), and Chirimachay (3298 m a.s.l.) precipitation stations and the Matadero-Sayausí discharge series. Note the grey horizontal line at a fixed correlation of 0.2 for determining the number of lags.

Similarly, the same procedure was applied for the remaining lead times (i.e, 4, 6, 8, and 12 hours). In Table 3.2, we present the input feature space composition and the resulting total number of features obtained from the lag analyses for each forecasting model. For instance, for the 1-hour case, the total number of features in the feature space equals 67, from which 43 are derived from precipitation (40 past lags and one feature from present time for each station), and 24 from discharge (one-hot-encoding).

Table 3.2. Input feature space composition (number of features) for all ML models of the Tomebamba catchment.

Lead time (hours)	Discharge lags* (hours)	Precipitation lags (hours)			Number of features
	Matadero-Sayausí	Toreador a	Chirimachay	Virgen	
1	8	15	14	11	67
4	12	18	17	14	88
6	14	20	19	16	100
8	16	22	21	18	112
12	20	26	25	22	136

* Note that each discharge feature triples (three flood warning classes) after a one-hot-encoding process.

3.2.4.2 Model hyperparameterization

The results of the hyperparameterization including the number of PCA components employed for achieving the best model efficiencies are presented in Table 3.3. No evident relation between the number of principal components and the ML technique nor the lead time was found. In fact, for some models, we found differences in the $f1 - macro$ score lower than 0.01 for a low and high number of principal components. See for instance the case of the KNN models where the optimal number of components significantly decayed for lead times greater than 4 hours. For the 1-hour lead time, 96% of the components were used, whereas for the rest of the lead times only less than 8%.

If we turn to the evolution of models' complexity with lead time (Table 3.3) more complex ML architectures are needed to forecast greater lead times. This is underpinned by the fact that the corresponding optimal models require for greater lead times a stronger regularization (lower

values of C) for LR, a greater number of neighbors ($n_neighbors$) for KNN, more specific trees (lower values of $min_samples_split$) for RF and more hidden layers ($hidden_layers$) for MLP.

Table 3.3. Model hyperparameters and the number of principal components used for each specific model (ML technique and lead time).

ML technique	Hyperparameter	Lead time				
		1h	4h	6h	8h	12h
LR	C	0.01	0.00001	0.0001	0.0001	0.001
	$penalty$	'l2'	'l2'	'l2'	'l2'	'l2'
	$PCA_components$ *	58	62	78	75	51
KNN	$n_neighbors$	15	15	23	33	55
	$weights$	'uniform'	'uniform'	'uniform'	'uniform'	'uniform'
	$metric$	'minkowski'	'minkowski'	'minkowski'	'minkowski'	'minkowski'
	$Algorithm$	'auto'	'auto'	'auto'	'auto'	'auto'
	$PCA_components$ *	64	6	6	6	4
RF	$n_estimators$	700	700	700	700	800
	$max_features$	'sqrt'	'auto'	auto	'log2'	'auto'
	max_depth	350	350	350	350	300
	$min_samples_leaf$	450	450	480	480	450
	$min_samples_split$	10	5	5	2	4
	$PCA_components$ *	66	79	90	45	78
NB	$PCA_components$ *	63	64	87	89	15
MLP	$solver$	'lbfgs'	'lbfgs'	'lbfgs'	'lbfgs'	'lbfgs'
	max_iter	2000	2000	2000	2000	2000
	$alpha$	0.0001	0.0001	0.0001	0.0001	0.0001
	$hidden_layers$	2	3	2	2	4
	$PCA_components$ *	63	51	64	76	4

* From the total number of features: 1h=67, 4h=88, 6h=100, 8h=112, 12h=136 features

3.2.4.3 Model performance evaluation

Model performances were calculated with the $f1 - score$, $G - mean$, and the $Log\ loss$ metrics. The overall performances across all classes (warnings) were obtained by weighting factors according to class frequencies. Table 3.4 presents the frequency distribution for the complete

dataset respectively for the training and test subsets. Here, the dominance of the No-alert flood class is evident, with more than 95% of the samples in both subsets. With this information, the class weights for the training period were calculated as $w_{\text{No-alert}} = 0.01$, $w_{\text{Pre-alert}} = 0.55$ and $w_{\text{Alert}} = 0.51$.

Table 3.4. The number of samples and relative percentage for the entire dataset and the training and test subsets.

Class (Warning)	Complete	Training	Test
<i>No-alert</i>	32596 (96.1%)	24890 (96.2%)	7706 (95.7%)
<i>Pre-alert</i>	720 (2.1%)	473 (1.8%)	247 (3.1%)
<i>Alert</i>	609 (1.8%)	509 (2.0%)	100 (1.2%)

The results of the model performance evaluation for all ML models and lead times (test subset) are summarized in Table 3.5. We proved for all models that the differences in performance metrics for a given lead time were due to the difference in the ML techniques rather than to the statistical chance. As expected, ML models' ability to forecast floods decreased for a longer lead time. For instance, for the case of 1-hour forecasting, we found a maximum $f1 - macro$ score of 0.88 (MLP) for the training and 0.82 (LR) for the test subset. Whereas, for the 12-hour case, the maximum $f1 - macro$ score was 0.71 (MLP) for the training and 0.46 (MLP) for the test subset.

The extensive hyperparameterization (RGS scheme) powered by 10-fold cross-validation served to assure robustness in all ML models and reduced overfitting. We found only a small difference between the performance values by using the training and the test subsets. For all models, maximum differences in performances were lower than 0.27 for the $f1 - macro$ score and 0.19 for the $G - mean$.

Table 3.5. Models' performance evaluation on the test subset. Bold fonts indicate the best performance for a given lead time.

Lead time (hours)	RF	KNN	LR	NB	MLP
<i>f1 – macro score</i>					
1	0.59	0.73	0.82	0.57	0.78
4	0.47	0.57	0.59	0.46	0.62
6	0.47	0.45	0.50	0.41	0.51
8	0.44	0.41	0.44	0.45	0.51
12	0.42	0.36	0.44	0.43	0.46
<i>G – mean</i>					
1	0.86	0.77	0.88	0.81	0.83
4	0.75	0.63	0.76	0.73	0.71
6	0.70	0.56	0.72	0.68	0.62
8	0.73	0.53	0.67	0.62	0.62
12	0.69	0.50	0.69	0.64	0.56
<i>Log – loss score</i>					
1	0.28	0.38	1.09	3.14	0.09
4	0.38	0.46	0.74	4.10	0.11
6	0.45	0.58	0.47	4.71	0.14
8	0.50	0.65	0.53	0.59	0.16
12	0.59	0.70	0.57	2.17	0.20

Note: All improvements and degradations are statistically significant

In general, for all lead times, the MLP model obtained the highest *f1 – macro score*, followed by the LR model. This performance dominance was confirmed by the ranking of the models according to the *Log loss score*. The ranking of the remaining models was highly variable and therefore not conclusive. For instance, the results of the KNN models obtained the second-highest score for the training subset, but the lowest for the test subset, especially for longer lead times. This is because the KNN is a memory-based algorithm and therefore more sensitive to the inclusion of information different from the training subset in comparison to the remaining ML techniques. This can be noted in Table 3.4, where the training and test frequency distributions are different for the Pre-alert and Alert classes.

On the other hand, for the *G – mean*, we obtained a different ranking of the methods. We found the highest scores for the LR model, followed by the RF and MLP models. Despite this behavior, the values of the g-mean were superior to the *f1 – macro scores* for all lead times and subsets. This is because the *f1 score* relies on the harmonic mean. Therefore, the *f1 score* penalizes a low precision or recall in comparison with a metric based on a geometric or arithmetic mean. Results of the *G – mean* served to identify that the LR is the most stable method in terms of

correctly classifying both the majority (No-alert) and the minority (Pre-alert and Alert) flood warning classes, while the MLP model could be used to focus on the minority (flood alert) classes.

To extend the last idea, we analyzed the individual $f1$ scores of each flood warning class. This unveils the ability of the model to forecast the main classes of interest, i.e., Pre-alert and Alert. Figure 3.5 presents the evolution of the $f1 - score$ of each ML algorithm at the corresponding lead time. We found that for all ML techniques, the Alert class is the most difficult to forecast when the $f1 - macro$ score was selected as the metric for the hyperparameterization task. An additional exercise consisted in choosing the individual $f1 - score$ for the Alert class as the target for hyperparameterization of all models. However, although we obtained comparable results for the Alert class, the scores of the Pre-alert class were highly deteriorated, even reaching scores near zero.

The most interesting aspect of Figure 3.5 is that the most efficient and stable models across lead times (test subset) were the models based on MLP and LR techniques. It is also evident that for all forecasting models, we found a lack of robustness for the Pre-alert warning class, this means major differences between the $f1 - scores$ for the training and test subsets. An explanation for this might be that the Alert class implies a Pre-Alert warning class, but not the opposite. Consequently, this might mislead the learning process causing overfitting during training and leading to poor performances when assessing unseen data during the test phase.

Moreover, although we added a notion of classes' frequency distribution (weights) to the performance evaluation task, it can be noted that for all models, the majority class is most perfectly classified. This is because the No-alert class arises from low-to-medium discharge magnitudes. This helps and simplifies the learning process of the ML techniques since these magnitudes can be related to normal conditions (present time and past lags) of precipitation and discharge.

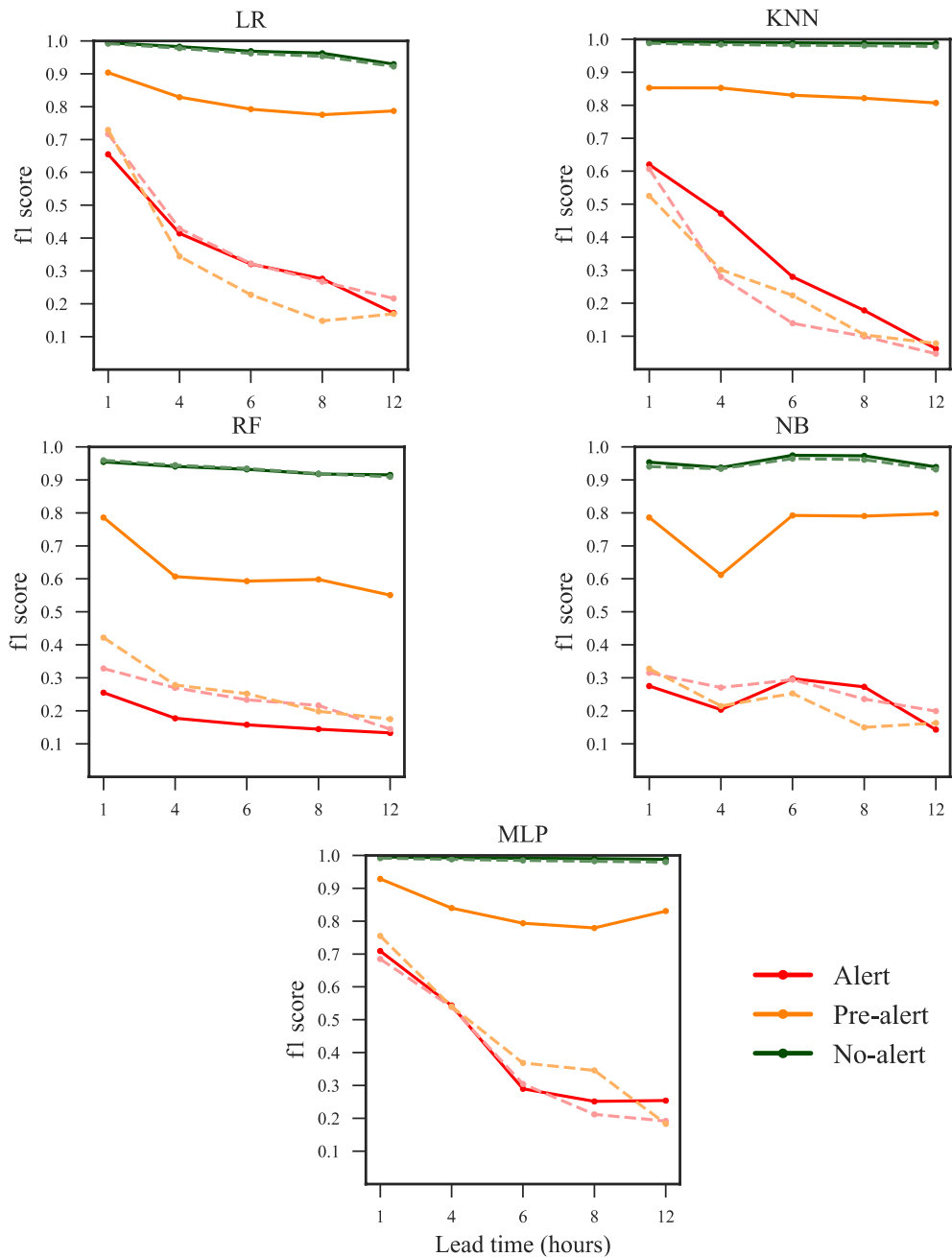


Figure 3.5. *F1 scores* per flood warning state (No-alert, Pre-alert, and Alert) for all combinations of ML techniques and lead times. The brightest and dashed lines in each case (color coding) represent the scores for the test subset.

3.2.4.4 Discussion

In this application, we developed and evaluated five different FEWSs relying on the most common ML techniques for flood forecasting, short-term lead times of 1, 4, and 6 hours for flash floods, and 8 and 12 hours to assess models' operational value for longer lead times. Historical runoff data were used to define and label the three flood warning scenarios to be forecasted (No-alert, Pre-alert, and Alert). We constructed the feature space for the models according to the statistical analyses of precipitation and discharge data followed by a PCA analysis embedded in the hyperparameterization. This was aimed at better exploiting the learning algorithm of each ML technique. In terms of model assessment, we proposed an integral scheme based on the $f1$ – score , G – mean , and the Log loss score to deal with data imbalance and multiclass characteristics. Finally, the assessment was complemented with statistical analysis to provide a performance ranking between ML techniques. For all lead times, we obtained the best forecasts for both, the majority and minority classes from the models based on the LR, RF, and MLP techniques (G – mean). The two most suitable models for the dangerous warning classes (Pre-Alert and Alert) were the MLP and LR ($f1$ and Log loss scores). This finding has important implications for developing FEWSs since real-time applications must be capable to deal with both the majority and minority classes. It can therefore be suggested that the most appropriate forecasting models are based on the MLP technique.

The results on the evolution of model performances across lead times suggest that the models are acceptable for lead times up to 6 hours, i.e., the models are suitable for flash-flood applications in the Tomebamba catchment. For lead times greater than 6 hours, we found a strong decay in model performance. In other words, the utility of the 8 and 12-hour forecasting models is limited by the models' operational value. This is because, in the absence of precipitation forecasts, the assumption of future rain is solely based on runoff measurements at past and present times. This generates forecasts that are not accurate enough for horizons greater than the concentration time of the catchment. The concentration time of the Tomebamba catchment was estimated between 2 and 6 hours according to the equations of Kirpich, Giandotti, Ven Te Chow, and Temez, respectively. A summary of the equations can be found in [108]. This results in an additional performance decay for the 8 and 12-hour cases in addition to the error in modeling.

The study of Furquim et al. [67] is comparable since they analyzed the performance of different ML classification algorithms for flash-flood nowcasting (3 hours) in a river located in an urban area

of Brazil. They found that models based on neural networks and decision trees outperformed the ones based on the Naive Bayes technique. However, this study only evaluated the percentage of correctly classified instances which is a simplistic evaluation. Thus, we recommend a more integral assessment of model performances, like the one in the current study, which allows for better decision-making support. Other studies related to quantitative forecasting such as Aichouri et al. [66], Khosravi et al. [68], and Solomatine and Xue [69] revealed that neural network-based models usually outperform the remaining techniques proposed in our study. Nevertheless, in certain cases, the use of less expensive techniques regarding the computational costs produces comparable results as in Solomatine and Xue [69]; this is also the case in our short-rain and flash-flood classification problem. As a further step, we propose the development of ensemble models for improving the performance results of individual models. This can be accomplished by combining the outcomes of the ML models with weights obtained, for instance, from the log-log scores. Another alternative that is becoming popular is the construction of hybrid models as a combination of ML algorithms for more accurate and efficient models [37], [68], [69]. As stated by Solomatine and Xue [69], inaccuracies in forecasting floods are mainly due to data-related problems. In this regard, Muñoz et al. [20] reported a deficiency in precipitation-driven models due to rainfall heterogeneity in mountainous areas, where orographic rainfall formation occurs. In most cases, rainfall events are only partially captured by punctual measurement, and even the entire storm coverage can be missing.

In general precipitation-runoff models will reach at a certain point an effectiveness threshold that cannot be exceeded without incorporating new types of data such as soil moisture [109], [110]. In humid areas, the precipitation-runoff relation also depends on other variables such as evapotranspiration, soil moisture, and land use, which leads to significant spatial variations of water storage. However, these variables are difficult to measure or estimate.

3.3 Quantitative flash flood forecasting

3.3.1 Introduction

The necessity of flash flood qualitative forecasting can be understood by analyzing the report of the Andean community for the period 1970–2007 (<http://www.comunidadandina.org>). In this report, it is revealed that in the Andes of Ecuador, 263 floods and 357 landslides (as a side effect, mostly in the city of Cuenca) caused 429 human deaths as well as the destruction of 2149 houses.

Moreover, according to the Empresa Pública Municipal de Telecomunicaciones, Agua potable, Alcantarillado y Saneamiento de Cuenca (ETAPA-EP), city of Cuenca is annually affected by flood events from which local media have reported human losses, destruction of infrastructure (e.g., bridges), and interruption of the water supply for the city and surrounding rural areas.

As a countermeasure against these flash flood impacts, it seems crucial for ETAPA-EP to count with quantitative runoff and flash flood forecasts for proper water management in aspects related to risk communication and mitigation, and for ensuring water production. Therefore, ETAPA-EP launched in 2014 a flash flood monitoring program that merely consists of monitoring (in real-time) the main currents at specific locations with the purpose to inspect the hydrograph transit [106]. The limitation of this monitoring program is the dependence on instrumentation which could be damaged during extreme events. Additionally, the time in advance in which an alert can be emitted is in the order of one or two hours, insufficient for taking mitigation actions.

In this context, the objective of this application is to develop flash flood forecasting models for the Tomebamba catchment. We produced forecasting models based on the RF algorithm and the forecasting ability was tested for lead times of 4, 8, 12, and 24 hours.

3.3.2 Dataset and processing

Data comprises precipitation and runoff hourly timeseries for the periods Jan/2015 to Sep/2018. We used the information of 3 rain gauges installed within the Tomebamba catchment and along its altitudinal gradients, Toreadora, Chirimachay, and Virgen, at elevations of 3955, 3626, and 3298 m asl, respectively. For model development purposes, we split the length of the data for training and testing. Training runs from Jan/2015 to May/2017 and testing from Feb/2017 to Jan/2019.

3.3.3 Methodology

Similar to the classification application, the development of a quantitative flash flood forecasting model begins with the composition of the input feature space, followed by model hyperparameterization, feature selection, and finally model performance evaluation. Figure 3.6 summarizes the methodology employed (see Chapter Two).

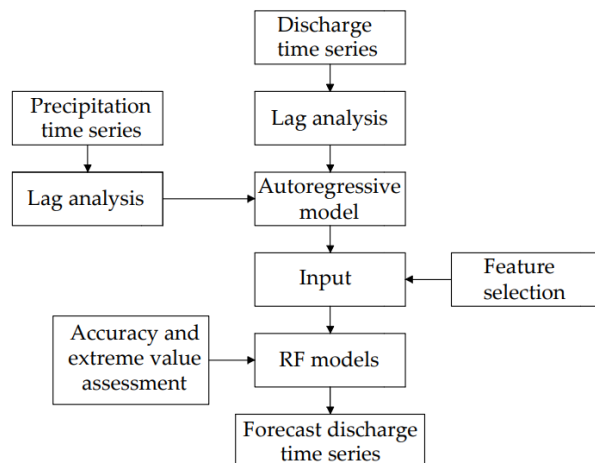


Figure 3.6. Methodology scheme for parsimonious model development.

3.3.3.1 Feature space composition

The feature space composition phase consists of correlation analyses to determine the necessary number of previous timesteps (lags) of precipitation and runoff that have a major influence on runoff forecasting. Here, we expect similar results when compared to the feature space composition for the classification application. In practice, this means similarity in the results of the ACF and PACF analyses for runoff and the Persian cross-correlation analysis for precipitation.

3.3.3.2 Model hyperparameterization

For the RF algorithm, the structure of the trees in the forest and their level of randomness can be controlled by RF hyperparameters [111]. And although the algorithm can be run with default hyperparameters, the study of Contreras et al. [79] showed that higher accuracies can be obtained by tuning the most relevant hyperparameters to the algorithm (see Table 3.6). For the hyperparameterization task, we employed a RGS procedure aimed to find the best combination (lower model residual) of hyperparameters from a previously defined grid of parameter ranges (Table 3.6). To avoid overfitting during the RGS process, a 3-fold cross-validation scheme was selected.

Table 3.6. Random Forest most-relevant model hyper-parameters and their search domain for tuning.

Hyperparameter	Value
n_estimators*	50-700
max_features	'auto', 'sqrt', and 'log2'
min_samples_split	2, 5 and 10
min_samples_leaf	1, 2 and 4
max_depth*	10-700

* Increment of 10 units

3.3.3.3 Model performance evaluation and feature selection

For a proper comparison between forecasts and observations, we employed a goodness-of-fit statistic, the NSE, which gives a measure of agreement focused on mean runoff values. To complement this analysis and evaluate the forecasting ability for flash floods, we employed a Box-Cox transformation to the discharge timeseries, and employ only nearly-independent peak runoff events.

In the pursuit of model parsimony, we develop alternative parsimonious forecasting models based on the feature selection process. For this, for the full input models, we calculated the relative importance of each feature to the model's output with the purpose to keep only features accounting for 80 % of the total relative importance. This means that the remaining features we trimmed off from the model's input of the parsimonious models. The idea was to contrast the forecasting efficiency between the full input and the reduced models (parsimonious).

3.3.4 Results and Discussion

3.3.4.1 Feature space composition

We obtained similar results than for the classification application in section 3.2. These are the inclusion of 8 runoff lags (hours) according to ACF and PACF analyses. For precipitation, we found a maximum cross-correlation with runoff at lag 4 (maximum 0.33 for Chirimachay). Based on this result, we decided to use 24, 10, and 15 lags for Toreadora, Virgen, and Chirimachay

precipitation stations, respectively. For the Pearson cross-correlation, we employed a correlation threshold of 0.2.

3.3.4.2 Model hyperparameterization and feature selection

The previous process to select lags provides a starting point for constructing RF forecasting models, however, in the pursuit of model parsimony, we applied a feature reduction process. For the 4-hour forecasting model, we found that including 9 lags from each precipitation station and 8 discharge lags would be enough to achieve 80.36% of the total relative importance. The percentage of reduction of model features was 58%. Table 3.7 summarizes the input feature space composition and the total number of features utilized for the RF forecasting models for all lead times.

Table 3.7. Input feature space composition of the RF models and their parsimonious versions (4, 8, 12, and 24-hour lead time) for the Tomebamba catchment.

Lead time [hours]	Discharge lags	Toreadora lags	Chirimachay lags	Virgen lags	Total Features
4	8	24	15	10	60
4*	8	9	9	9	38
8	8	32	23	19	85
8*	8	15	15	15	56
12	8	36	27	23	98
12*	8	18	18	18	65
24	15	48	39	35	140
24*	15	21	21	21	81

* Parsimonious version

3.3.4.3 Model performance evaluation

Table 3.8 presents the obtained model performances in terms of the NSE. Notice that the NSE was calculated for the whole spectrum of flows. In this table, we also contrast the NSE coefficients of the full-input and their parsimonious model obtained through a feature selection process. Results prove that NSE coefficients obtained from the parsimonious models do not differ significantly from the correspondent full-input version of the model (maximum difference in

calibration and validation of 0.01). In some cases, parsimonious models even outperform their correspondent full-input models.

Table 3.8. Model performance of the RF models and their parsimonious versions (4, 8, 12, and 24-hour lead time).

Lead time [hours]	Total Features [#]	NSE	
		Training	Test
4	60	0.9193	0.8604
4*	38	0.9211	0.8682
8	85	0.8486	0.7494
8*	56	0.8441	0.7523
12	98	0.8131	0.6759
12*	65	0.8074	0.6799
24	140	0.7541	0.538
24*	81	0.7483	0.5454

* Parsimonios version

Regarding RF model overfitting, we found maximum differences between the NSE coefficients of the calibration versus the validation period of 0.20.

For the evaluation focused on flash floods, Figure 3.7 shows the empirical extreme peak value distributions for all forecast horizon models (4, 8, 12, and 24 hours). For this, we employed the simulations obtained from the so-called parsimonious models. Overall results for both catchments, revealed that the underestimation of peak flows towards the upper tail of the distribution becomes stronger as the lead time increases. We found maximum underestimations of 48, 53, 57, and 66% for the 4, 8, 12, and 24-hour forecasting models, respectively.

On the other hand, Figure 3.8 presents a scatter plot of forecasts (vertical axis) and observed discharge (horizontal axis) for peak flows. Here, model residuals are represented by the horizontal and vertical differences between each point and the bisector line. The dependence of the standard deviation on the flow magnitude was disrupted (constant standard deviation) with a λ -value of 0.25. Results confirm the observed in Figure 3.8 where higher scatters and biases were found for longer forecast horizons.

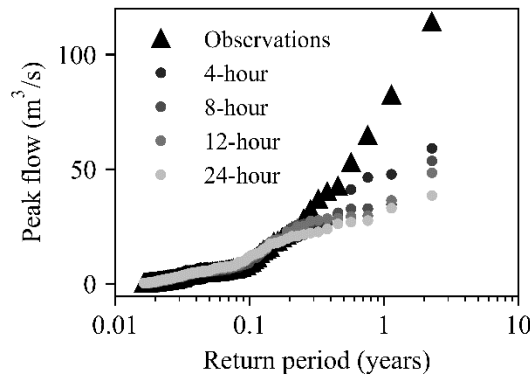


Figure 3.7. Empirical extreme value distribution of peak flows (flash floods).

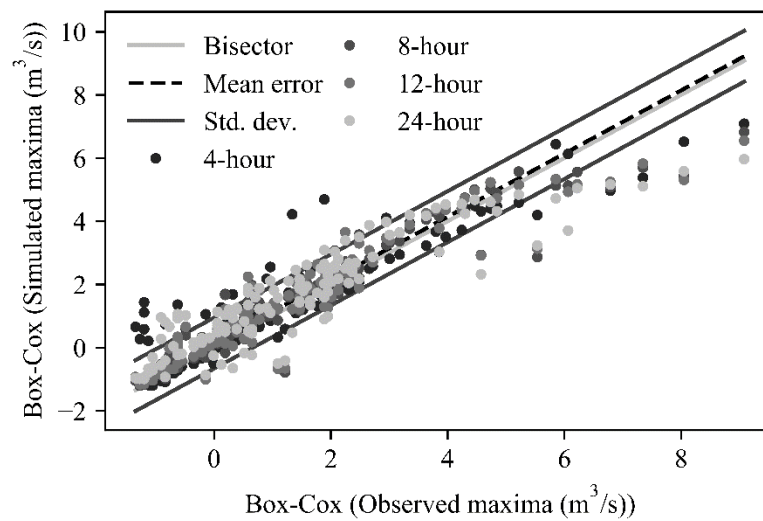


Figure 3.8. Comparison of nearly independent peak flow maxima

3.4 Summary and conclusions

In this chapter, we presented two case studies (qualitative and quantitative) peak runoff and flash flood forecasting in a meso-scale hydrological system representative of the tropical Andes of Ecuador. For this, we exploited precipitation data obtained from available ground-based measurements.

For the qualitative forecasting case study, we developed FEWSs using ML techniques, and special attention was taken to the selection of the most appropriate technique among the universe of ML techniques. We assessed FEWSs with three warning or river states, No-alert, Pre-alert,

and Alert for flooding, and for lead times between 1 to 12 hours. In terms of modeling, we used the most-employed ML techniques that belong to five different ML strategies, which are the Multi-Layer Perceptron (MLP), Logistic Regression (LR), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Random Forest (RF). For all lead times, the MLP models achieved the highest performance followed by LR models, with $f1 - macro$ (*Log loss*) scores of 0.82 (0.09) and 0.46 (0.20) for the 1- and 12-hour cases, respectively. The ranking was highly variable for the remaining ML techniques. According to the $G - mean$, the LR models correctly forecast and show more stability at all states, while the MLP models perform better for the Pre-alert and Alert warnings.

For the qualitative application, the following main conclusions can be drawn:

- In contradiction to other studies, our results related to model comparison are statistically significant and validate our results regarding model performance comparison and ranking.
- For all lead times, the most suitable models for flood forecasting were based on the MLP models followed by the LR techniques. Based on the performance metrics, we believe that the LR models are the most efficient and stable option for the classification of both the majority (No-alert) and minority (Pre-alert and Alert) classes. While we recommend the MLP models when the interest lies in the minority classes.
- The forecasting models developed in this study were robust. Differences in the averaged $f1$ scores, $G - mean$, and *Log loss scores* between training and test were consistent for all models. The utility of the models for flash-flood applications (lead times up to 6 hours) is limited. For longer lead times, we recommend improvement in precipitation representation, and even forecasting this variable for lead times longer than the concentration time of the catchment.
- A more detailed model assessment (individual $f1$ scores) unveiled the difficulties to forecast the *Pre-alert* and *Alert* flood warnings. This was evidenced when the hyperparameterization was driven for the optimization of the forecast for the alert class and this, however, did not improve the model performance of this specific class.

For the quantitative forecasting case study, we examined the feasibility to develop precipitation-runoff forecasting models and their ability to forecast flash floods according to a multi-criteria evaluation framework. We used a methodological framework to develop flash flood forecasting

models for several lead times (4, 8, 12, and 24 hours). We found that derived models can reach maximum validation performances (NSE) from 0.860 (4-h) to 0.545 (24-h) for optimal inputs composed only by features accounting for 80% of the model's outcome variance.

For the regression the following main conclusions can be drawn:

- As expected, the ability of the RF models to forecast flash floods decreased with increasing lead time.
- The use of a feature selection technique (based on the output's variance) proved to be successful not only in reducing models' complexity due to the dimension of the input feature space but also to keep forecasting efficiencies.
- It was demonstrated the difficulty to forecast flash floods rather than mean runoffs. From a data-driven perspective, this is occasioned by imbalance data problems (i.e., the number of independent events for peak flows is scarce. The solution would be to collect more data on flash floods or to develop specialized models (event-based modeling).

Now, based on both applications we can conclude that overall, the forecasting of flash floods is challenging mainly due to a lack of or insufficient resolution of relevant data (driving forces), and insufficient extreme events from which ML models can learn and forecast. For instance, for the meso-scale Tomebamba catchment (dominated by a paramo ecosystem), it is well-known that soils govern flow processes, and therefore, lack of direct measurements limits the forecasting of extreme flows. Moreover, the extreme spatial and temporal variability of precipitation in this catchment representative of the Ecuadorian Andes is hardly collected by a few rain gauge stations within the catchment.

The logical solution would be then to expand the precipitation monitoring network to improve the representativeness of precipitation. This will improve, to a certain degree, model performances. However, it must be taken into account that the major shortcoming of the use of rain gauges in the Andean region is the occurrence of focalized precipitation events due to complex topography. Thus, an adequate representation of the spatial variability of precipitation is rarely available for forecasting applications. Moreover, budget constraints in the Andean region and particularly in Ecuador often limit its viability.

To overcome this issue, there is the opportunity to use spatial precipitation estimations from remote sensing products such SPPs, or even combine this spatial information with available ground-based data. Other future directions for flash flood forecasting with ML techniques can encompass a deep exploration of the effect of input feature space composition, and the FE strategies aimed at improving data representation, and thus, forecasting efficiencies.

These ideas will be explored in the following Chapters of this thesis for the case of regression. This is because in this way we can test the quantitative accuracy of forecasting models and their ability to forecast peak flow magnitudes in a more detailed way. Another reason for selecting a regression approach is due to the superior number of applications or uses that can be derived from quantitative flash flood forecasting when compared to the classification approach. Some applications are for instance the management of water plants or hydropower dams where peak runoff forecasts can also be used to close entrance gates and avoid the adverse effects of high levels of water contamination and sediments coming from erosion processes produced by flash floods.

Finally, the utility of these two forecasting case studies for the Tomebamba catchment, and other comparable systems, is conclusive. Although further recommendations for improving model performance have been identified, the models and the methodology followed in both applications can be immediately used and the results interpreted by decision-makers and politicians.

Chapter four: a feature engineering strategy for exploiting of satellite-based precipitation data in machine learning models.

Related publication:

- ❖ **Muñoz, P.,** Perez, G. A. C., Solomatine, D., Feyen, J., & Céleri, R. (2022). *Use of near-real-time satellite precipitation data and machine learning to improve extreme runoff modeling*. AGU books. Accepted. Authorea. doi: 10.1002/essoar.10508861.1
-

The lack of sufficient and relevant ground information in the tropical Andes limits the applicability of physically-based models for runoff and—by extension—peak runoffs. In contrast, the use of machine learning (ML) techniques allows data-driven models to exploit the available data in order to provide adequate simulations. Beyond its capacity to provide accurate simulations, ML has also received critical attention due to its potential to infer hydrological processes in a particular system [112]–[116].

It is argued, for instance, that conceptualizing and understanding forcing processes in a system would enable ML models to simulate the hydrological response beyond the range of training data, or even to transfer models to similar ungauged systems [1]. The process of adding hydrological knowledge to ML models is known as feature engineering (FE). Several FE strategies have proven to be successful in hydrological modeling, for instance, runoff separation into subflow components [50], [95], [117], exploitation of topographic characteristics [118], the addition of stream network information [118], [119], the addition of mass and energy balance equations for the training task [120], or various ways of employing hydrological knowledge in choosing input attributes [114].

On the other hand, to deal with ground data scarcity, the continuous development of SPPs has dramatically enhanced the quantity and quality of areal precipitation observations. Yet, SPPs obtained from a single satellite hardly provide accurate estimations [121]. This has stimulated the development of multi-satellite SPPs such as the NASA Global Precipitation Measurement (GPM)

Integrated Multi-satellite Retrievals for GPM (IMERG) [122], and the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) [23]. These two SPPs exhibit the highest spatial (< 11 km x 11 km) and temporal resolutions (< 1 hour), and latency times (< 5 hour), which are useful for flash flood forecasting and/or real-time applications. However, a major problem in the use of the IMERG and PERSIANN products is the fact that the accuracy of these SPPs have only been validated with ground information in certain regions, leading to precipitation uncertainties in unvalidated regions such as the Andes.

The validation of SPPs with ground information is mandatory in the cases when the interest lies in providing accurate precipitation estimations (see for instance precipitation validation studies of Laverde-Barajas et al. [123] and Li et al. [124]) or for providing precipitation inputs for traditional physically-based hydrological models. Nonetheless, one of the greatest opportunities about the use of ML for hydrology is the freedom to employ unvalidated SPP estimates for complex systems. This is because ML exploits not quantitative precipitation but rather precipitation differences within the system. The hypothesis is that precipitation uncertainties at local scales remain more or less constant. Investigating FE strategies for improving SPPs assimilation in ML is a continuous concern within the field of hydrological modeling [125]–[127]. A successful strategy is the use of object-based methods for deriving precipitation attributes from satellite imagery [48], [123], [124], [126], [128], [129]. With these attributes, there is the potential to build specialized runoff models able to discriminate between different precipitation event types. This is based on the concept that different precipitation events produce different runoff responses as a result of different runoff generation processes, mainly infiltration and saturation excess [130].

In summary, it seems that a robust solution to the difficulty to forecast peak runoffs is the development of ML models able to exploit and digest spatial SPPs data, and account for key hydrological concepts about the functioning of the system. Both concerns can be faced through the application of a combination of FE strategies aimed at assisting ML forecasting. But since the FE application attempts to add hydrometeorological knowledge of the system, the hypothesis employed and their implementation must first demonstrate success in runoff and flash flood modeling (current-time) before forecasting.

Thus, in this chapter, we propose and implement FE strategies for improving flash flood modeling in a complex mountain systems in terms of spatial and temporal data scarcity. Among FE strategies, we focused on: i) an object-based methodology for processing SPPs imagery to derive

additional features meaningful for the ML learning process, and ii) classification of precipitation events leading to peak runoffs. The ML technique selected is the Random Forest (RF) algorithm for regression given its demonstrated flexibility, accuracy, and competent computation times for operational hydrology (Chapter Three).

This chapter is organized as follows, first we defined the aforementioned two FE strategies for assisting the assimilation of SPPs to ML models. Second, we test the CCA in a study for the Jubones basin, and discuss the advantages and disadvantages in developing specialized ML models according to precipitation conditions triggering peak runoffs.

4.1 Aim and objectives

To implement a FE strategy for assisting ML flash flood modeling through the exploitation of satellite-based precipitation data.

Objectives:

- To propose a FE strategy, the connected component analysis (CCA), for exploiting satellite-based precipitation data and for deriving precipitation attributes.
- To demonstrate the utility of the CCA for flash flood modeling.

4.2 Feature engineering strategies

4.2.1 Object-based Connected Component Analysis (CCA)

The opportunity to employ SPPs is essential for characterizing the spatial distribution of precipitation events to produce accurate flash flood forecasts. However, characterizing small-scale precipitation dynamics with SPPs is still a major concern within the remote sensing field [131]. This is because most studies have evaluated SPPs in terms of efficiencies and correlations but have failed to address the description of key attributes to flash flood forecasting, these are for instance total water volume, storm location, and type of precipitation event, among others [123].

In this regard, object-based methods are an alternative for analyzing spatial precipitation. A simple yet effective object-based method is the Connected Component Analysis (CCA) proposed by Laverde-Barajas et al. [123]. The CCA extracts precipitation attributes from SPP data through a multidimensional connected component labeling algorithm. The extracted attributes provide a

physical description of precipitation events (localization, centroids, area), and meteorological features (duration, volume, maximum intensity, etc.). For exemplifying and summarizing the main CCA steps, we used a precipitation image covering the Jubones basin. The main CCA steps are as follows:

- i. Precipitation retrieval for selected flash flood events, and imagery clipping to the Jubones basin (Figure 4.1a).
- ii. Detection and localization (latitude, longitude, see Figure 4.1b) of precipitation objects. For this, a detection sensitivity threshold is defined to remove noise and keep only clear precipitation objects in the precipitation imagery (Figure 4.1c). The detection sensitivity was calibrated on a trial-and-error basis with a precipitation threshold volume of 0.5 mm. This means that precipitation objects associated with a depth of less than 0.5 mm were trimmed-off.
- iii. Precipitation object filtering according to size criteria. We defined a minimum object area corresponding to two pixels of the finest-resolution product (~39 km²).
- iv. Morphologically closure of precipitation objects found in step (iii). For this, a dilation-and-erosion algorithm was used to refine precipitation objects (Figure 4.1d); dilation expands objects while erosion removes the boundaries of the expansion.
- v. Extraction of physical (centroid and extension area) and meteorological attributes (volume of precipitation, maximum intensity, precipitation duration) from the objects refined in step (iv). 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 3 h. This threshold was also calibrated on a trial-and-error basis. These characteristics are then used for classifying precipitation events which can be paired with their associated runoff responses (see the next section).

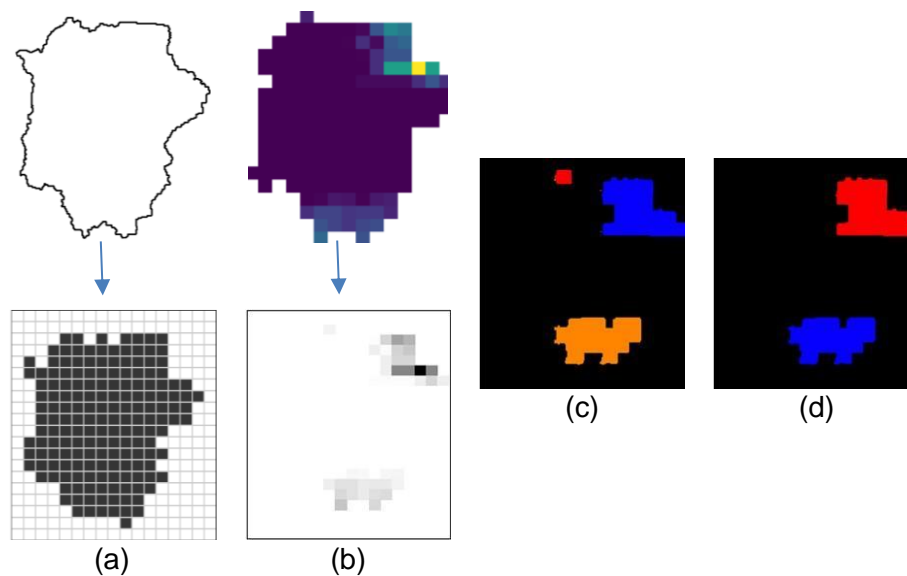


Figure 4.1. Precipitation identification with an object-based Connected Component Analysis (CCA) Illustration of the PERSIAN-CCS 2021-12-25 05:00 UTC image. (a) Jubones basin boundary, (b) Precipitation identification in mm from the PERSIANN-CCS product, (c) Identification of three precipitation objects with the CCA, and (d) Final identification of two precipitation objects after object size filtering and morphological closing.

4.2.2 Classification of precipitation events leading to peak runoffs

Precipitation events can be distinguished (classified) by applying object-based methods to SPP data [123], [124], [128], [129], [132], [133]. From the CCA, precipitation events triggering peak runoffs can be classified by focusing on two precipitation attributes, the extension of the precipitation objects (local and spatial extensive) and the duration of the events (short and long). As a result, four precipitation event classes can be defined: i) Local and short-duration 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) [123].

Once classified, there is the potential to produce specialized flash flood forecasting models for each precipitation class (LSE, LLE, SEE, and SLE). Here the idea is to reduce noise during the learning process of ML models. Let's think of a single ML model without precipitation characterization. This single model has to learn different precipitation runoff mechanisms, and it is well known that flash flood responses can be the result of more than one mechanism, for instance, infiltration- and saturation-excess. The arising problem is that the learning of multiple

mechanisms will be biased toward the occurrence of the main mechanism and associated precipitation attributes, thus, failing to model the response of other flash flood events. As a solution, we rely on the premise that the development of specialized models will produce more accurate forecasts than a single model for all precipitation classes.

4.3 Implementation of FE strategies for peak runoff modeling

In this study, we develop specialized ML peak runoff models for the Jubones basin, a 3393-km² basin in Ecuador. We used a combination of FE strategies, the CCA to improve the areal representation of precipitation, and specialized runoff modeling to maximize model efficiencies by identifying and classifying precipitation events associated with flash flood responses.

4.3.1 Dataset and processing

The dataset for this application comprises hourly satellite-derived precipitation covering the Jubones basin, and hourly runoff data collected at a hydrological station situated in the outlet of the basin, consisting of the Minas-San Francisco hydropower dam. Since the dam was completed in 2018, the study period ran from November 2018 to April 2021 (~2.5 years). As mentioned in Chapter One, precipitation data were retrieved from two near-real-time databases (considering the absence of ground-based precipitation stations), the IMERG-Early Run (ER), and the PERSIANN-Cloud Classification System (CCS) products. Data were extracted at the finest temporal resolution and then aggregated to the hourly time step. Mean (maximum) annual precipitation depths are 729 (1167) and 1532 (2759) mm, for the PERSIANN-CCS and IMERG-ER, respectively.

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 4.2a depicts the runoff information for the study period, whereas Figure 4.2b the corresponding probability of runoff exceedance.

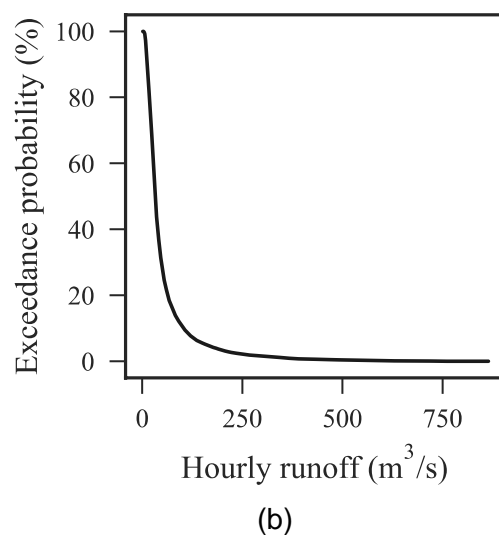
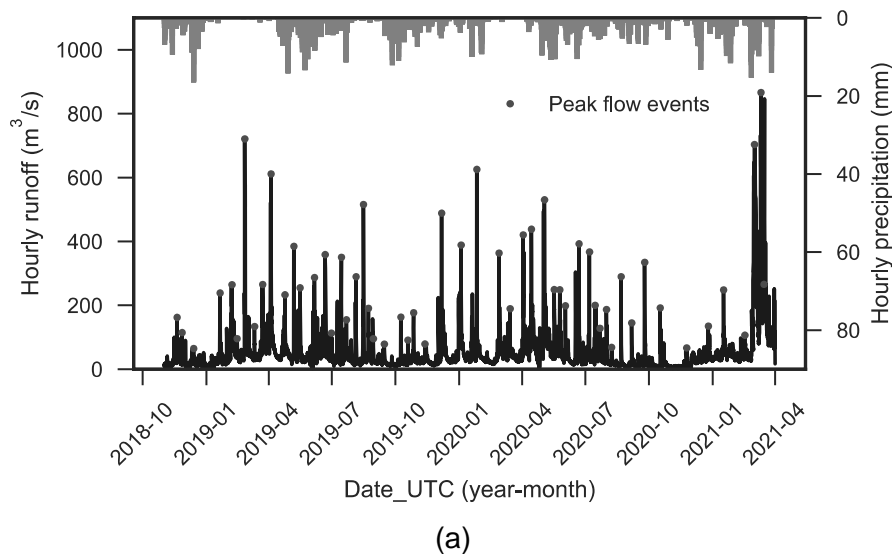


Figure 4.2. (a) Hourly runoff and precipitation (PERSIANN-CCS) time series at the outlet of the Jubones basin. Peak flow events are displayed as dots. (b) Exceedance probability for the study period (18/11/2018 to 01/04/2021).

4.3.2 Methodology

A summary of the methodology employed in this application is as follows. First, flash flood events must be selected from the runoff timeseries. Second, for the selected events, precipitation information from both SPPs is retrieved and the CCA (FE strategy) is applied to derive hydrometeorological attributes. Third, these attributes are used to classify precipitation events leading to a flash flood response. Fourth, differentiated modeling is performed for each class.

Fifth, to evaluate the performance and utility of specialized models, we contrast the performance of specialized models with base models developed with the same input information but without consideration of precipitation classes. For the comparison, we employ a combination of efficiency metrics focused on mean and extreme values. Below, we describe the most important aspects of the methodology.

4.3.2.1 Selection of nearly independent flash flood events

Flash flood events were derived from the runoff time series by applying the following two criteria: i) flash flood events must exceed 90% of runoff quartile values ($98.8 \text{ m}^3 \cdot \text{s}^{-1}$), and ii) flash flood events must be nearly independent. For meeting both criteria, we used the WETSPRO time series tool developed by Willems [95]. The WETSPRO splits the runoff series in nearly independent peak flow events following a peak-over-threshold approach. For this, the WETSPRO has two parameters to be calibrated, the inter-event time and peak height. In other words, flash flood events were selected from the runoff timeseries with a definition of independence controlled by the recession time and peak height difference of two consecutive runoff events.

4.3.2.2 Object-based CCA

To apply the CCA, we designed a modular approach for SPP data acquisition. This was aimed to deal with the cases when a SPP fails to observe precipitation yet there is a flash flood response. For instance, when no precipitation is observed by the PERSIANN-CCS product, 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 $0.1 \times 0.1^\circ$ was directly divided into ~ 6.4 cells with a resolution of $0.04 \times 0.04^\circ$, matching the resolution of the PERSIANN-CCS product. This modular approach assures that flash flood events are trained with an existent precipitation signal, reducing noise and improving the learning process.

Apart from processing SPP data, the CCA served to derive two meteorological attributes, the extension of the precipitation objects, and the duration of the precipitation events. The CCA was implemented through the scikit-image processing package in Python® version 3.7 [134].

4.3.2.3 Classification of precipitation events

Precipitation data passed through the CCA, and from this analysis, we defined extension and duration thresholds. According to this threshold, we established 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). The classified events were used to develop specialized flash flood models.

4.3.2.4 Model construction, hyperparameterization, and evaluation

The specialized LSE, LLE, SEE and SLE models, and the base models were built using the RF for regression. 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 functions for precipitation (see Chapter Two).

For the hyperparameterization task, we tuned the most-influencing RF hyperparameters, these 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 the maximum depth for pruning purposes (max_depth) [79]. For all models, we determined the optimal combinations of hyperparameters following a RGS 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 4.1 presents the domain of the selected hyperparameters which forms the search space for the optimization task.

For the evaluation of specialized models, and the comparison with the base model, we used four goodness-of-fit metrics for evaluating the efficiencies of the four runoff models. The NSE was set as the reference for measuring and comparing the overall model accuracy. To complement the analysis, we relied on KGE, the PBIAS, and the RMSE. All equations are described in Chapter Two.

Table 4.1. 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.

4.3.3 Results

4.3.3.1 Determination of nearly independent flash flood events

The WETSPRO tool for the Jubones basin was calibrated using the following parameters: inter-event time of 12 hours, and a maximum ratio of runoff drop down of 0.6. Moreover, we considered only events exceeding the 90% quartile values of the runoff time series ($98.8 \text{ m}^3 \cdot \text{s}^{-1}$). With these criteria, we obtained 55 nearly independent peak hydrological events (see Figure 4.2a).

4.3.3.2 Object-based CCA

For the 55 peak hydrological events, we first 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 4.1.

CCA results showed that, for 15 extreme hydrological events, there was nearly or even an inexistent precipitation signal from the PERSIANN-CCS product. For these 15 cases, we performed the CCA algorithm on the IMERG-ER 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 analysis. The utility of the precipitation modular approach can be seen in the events depicted in Figure 4.3. For the event from 2019 to 07-13 20:00 to 2019-07-14 20:00 UTC (Figure 4.3a), it seemed evident that the higher resolution of the PERSIANN-CCS product lead to a stronger precipitation-runoff relation

when compared to precipitation obtained from the IMERG-ER product. Thus, the precipitation data from the PERSIANN-CCS were used to feed the forecasting models. The opposite was true for the event from 2019 to 10-07 at 16:00 to 2019-10-08 at 16:00 UTC (Figure 4.3b), where the PERSIANN-CCS signal was nonexistent for almost 24 h before the runoff peak, whereas there is a significant amount of precipitation from the IMERG-ER product.

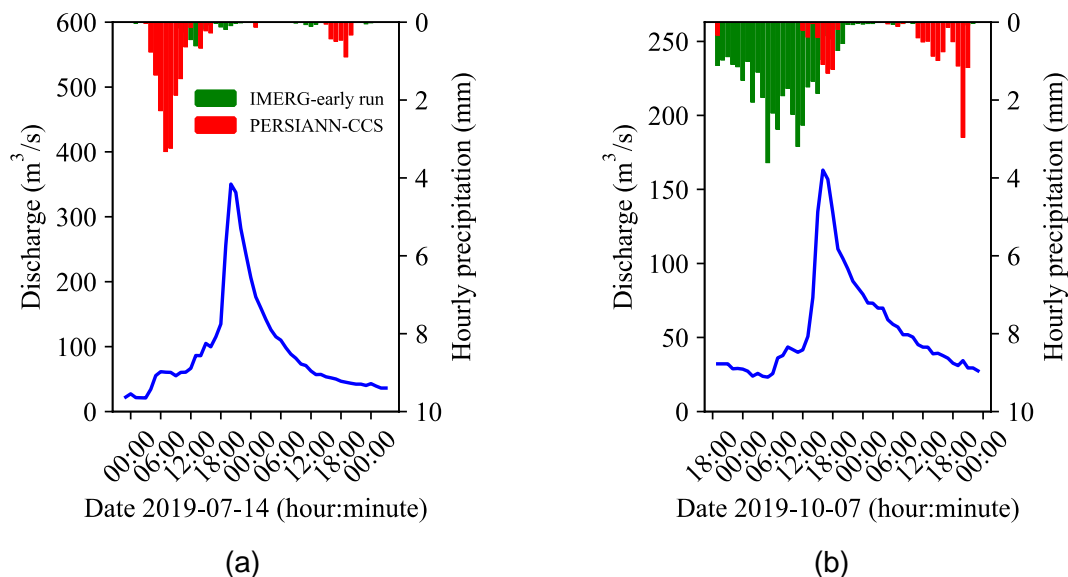


Figure 4.3. Illustration of the precipitation-retrieval modular approach using PERSIANN-CCS and IMERG-ER data sources, respectively for the events from (a) 2019-07-13 18:00 to 2019-07-14 18:00 UTC, and (b) from 2019-10-07 12:00 to 2019-10-08 12:00 UTC.

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 4.4 and served to infer duration and extension thresholds of 7 hours and 50 km², respectively. These thresholds were used in the following subsection to classify the precipitation events.

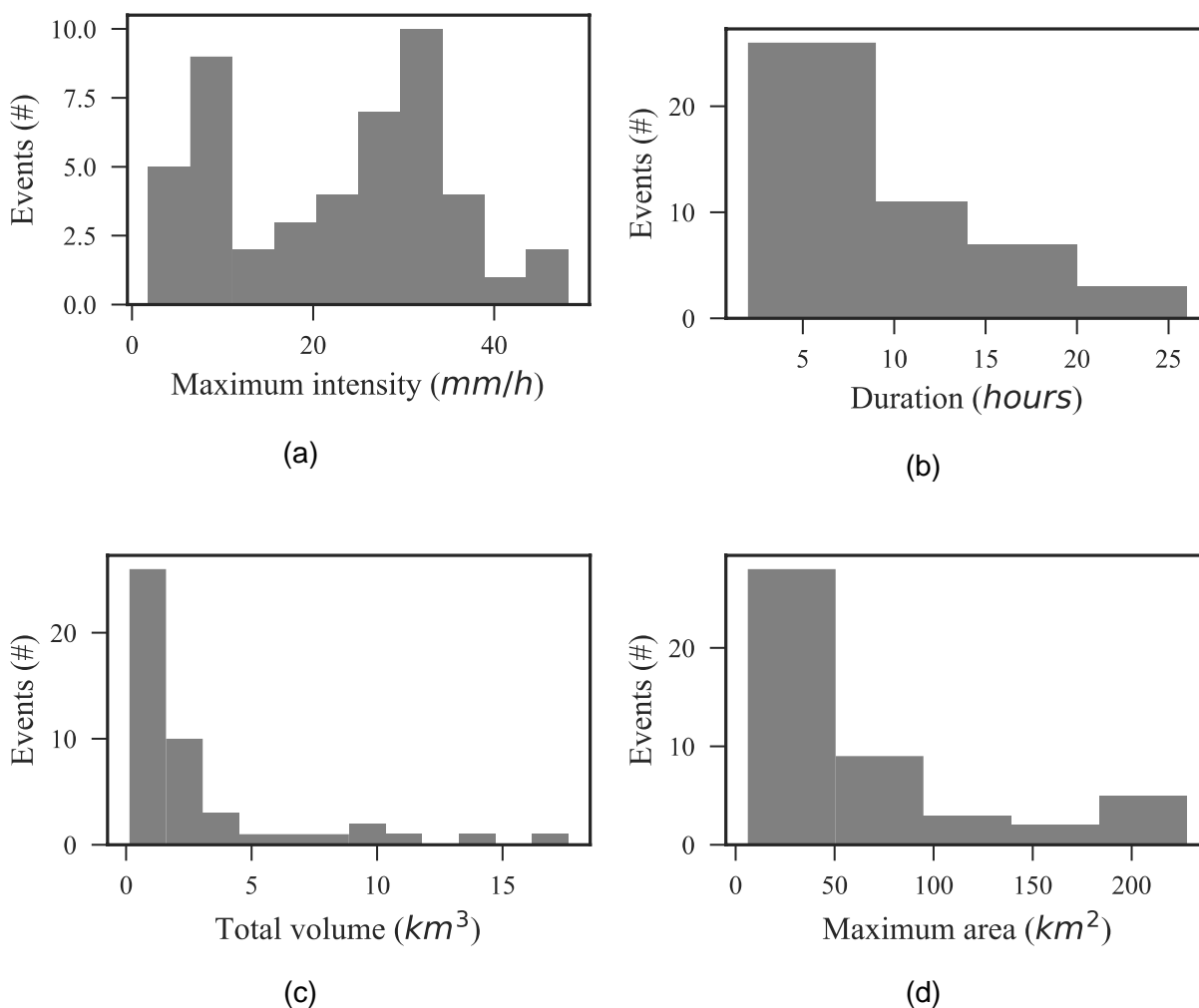


Figure 4.4. Meteorological precipitation information was retrieved from 46 extreme hydrological events: (a) maximum intensity, (b) duration, (c) total volume, and (d) maximum area.

Moreover, analysis of the centroid occurrence of the precipitation objects did not reveal any precipitation hotspots in the basin that could be associated with peak runoff events (see Figure 4.5). There was no evidence that centroid occurrence is driven or can be related to any physical attribute of the Jubones basin (e.g., soil type, land use, elevation, topography, etc.). This might

indicate the nonexistence of orographic precipitation enhancement (i.e., cloud formation due to orographic lifting of air masses).

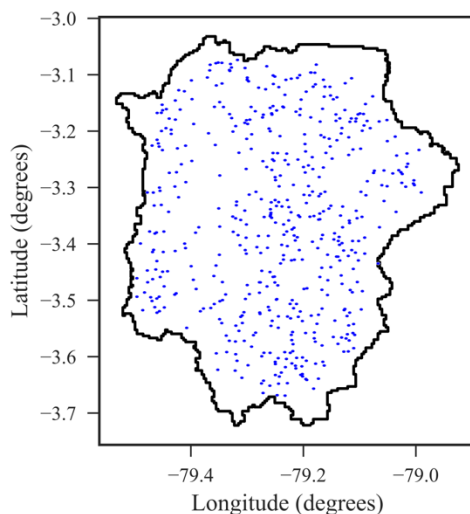


Figure 4.5. Localization of precipitation object centroids (blue dots) associated with extreme hydrological events in the Jubones basin.

4.3.3.3 Classification of precipitation events

The combination of duration and extension thresholds of 7 hours and 50 km² 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 4.6 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).

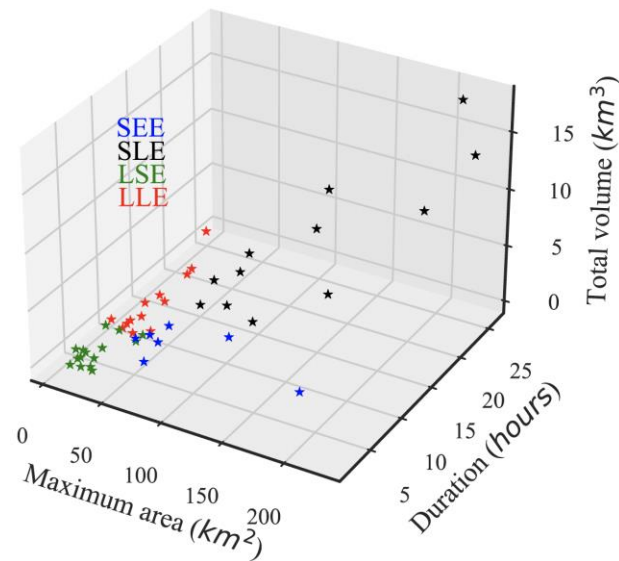


Figure 4.6. Precipitation classes associated with extreme hydrological events: Local and short extreme events (LSE), Local and long-duration extreme events (LLE), Spatially extensive extreme events (SEE), and Spatially extensive and long-duration extreme events (SLE).

4.3.3.4 Event-based flash flood modeling

First, we defined the dimension of the input feature space. Results from ACF and PACF for runoff suggested 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 [135]).

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 24 events available for the LSE precipitation class; therefore, we assigned 17 events for training and 7 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 4.2 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 4.2. 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)	n_features ^(1/2)	2100 ^(1/2)	log ₂ (2100)
max_depth*	200	200	150	180	200

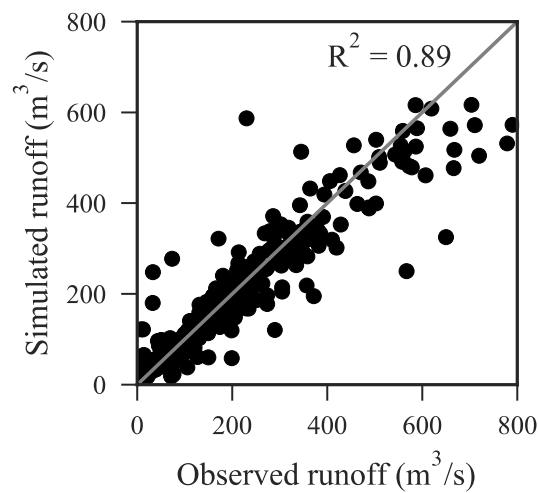
Table 4.3 summarizes the number of events used for developing extreme runoff models, and a comparison of the NSE coefficients obtained for each precipitation class and the base model. It is apparent from this table that LSE and especially SEE precipitation events are causing decay in the overall NSE-value of 0.83 (see also Figures 4.7b and 4.7d). 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 4.3. The number of events and efficiencies on test subsets of runoff models specifically developed for different precipitation events.

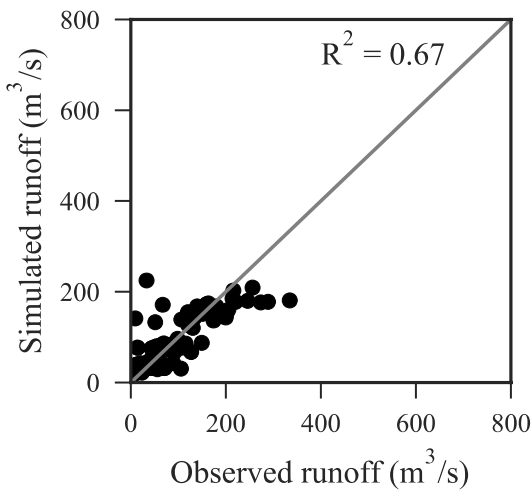
Precipitation class	# Total Events (Test)	NSE	KGE	PBIAS	RMSE
None	46 (14)	0.83	0.85	4.49	55.38
LSE	24 (7)	0.67	0.71	-1.45	35.00
LLE	5 (2)	0.72	0.74	-23.94	41.76
SEE	7 (3)	-1.93	-0.48	-61.44	60.44
SLE	10 (3)	0.90	0.94	-2.72	69.09

From the data in Figure 4.7, we can infer the spectrum of the runoff magnitudes modeled for each precipitation class. What is striking from the subfigures in Figure 4.7 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

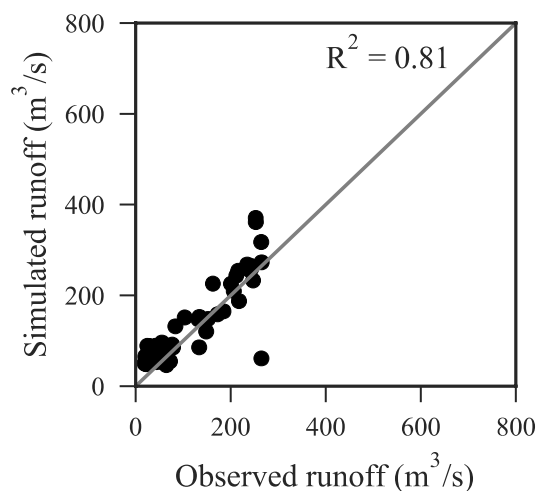
is evident that the lowest NSE coefficients for the LSE and SEE classes are found. Physically, this finding may be explained by the fact that the runoff response of short-duration events is somehow softened by the infiltration and saturation processes. This means that the volume of precipitation that 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 4.7a), 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 (Figure 4.7c and Figure 4.7e).



(a)



(b)



(c)

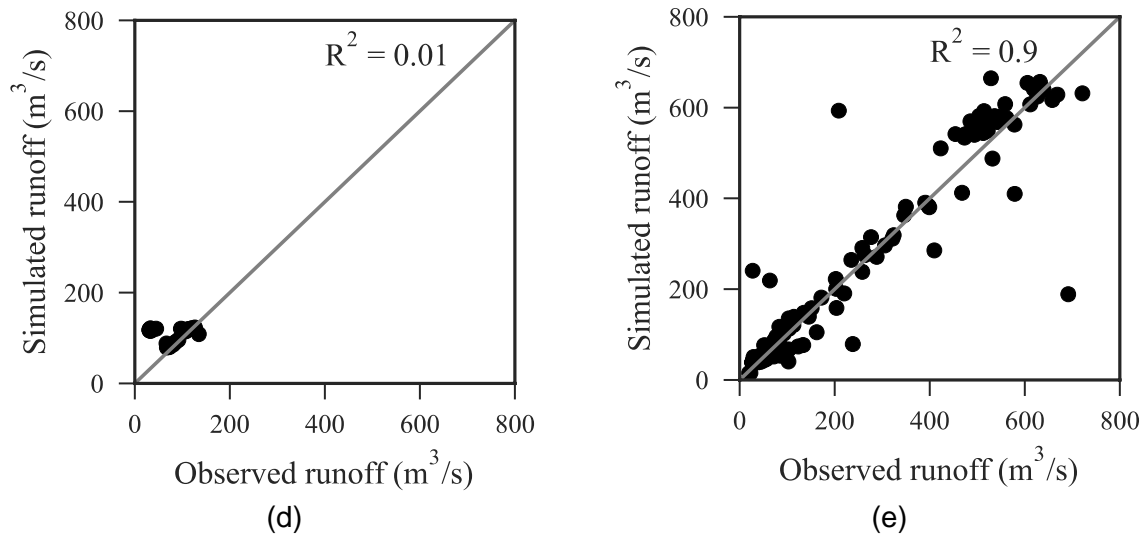


Figure 4.7. Scatter plot between extreme runoff observations and simulations for (a) No-precipitation event classification, (b) LSE events, (c) LLE events, (d) SEE events, and (e) SLE events.

4.3.4 Discussion

In this study, specialized flash flood models were developed for a 3391-km² 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. [29]), wflow-sbm (see Laverde-Barajas et al. [126]), and the hydrologic-hydraulic HiResFlood-UCI model (see Nguyen et al.[30]). 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 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. [29], 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 in 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=0.72). The SLE runoff model was 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. [136] 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, the exploration of additional ML algorithms for the modeling of LSE and SEE events, and come up with a superior model consisting of an ensemble of specialized runoff models.

4.3.5 Conclusions

This study exploits the possibility of using two near-real-time satellite precipitation sources (without ground validation) for the development of specialized flash flood models for a 3391-km²

basin. Specialized models are characterized by the use of a ML algorithm assisted by a FE strategy for assimilating SPPs data, and for deriving hydrometeorological attributes for the classification of precipitation conditions leading to peak runoffs. The major finding emerging from this study is that improvement of the representation of precipitation maximizes the efficiency of flash flood models. In addition, precipitation classification 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. 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 tested two near-real-time precipitation satellite sources, the PERSIAN-CCS and IMERG-ER products. We used a modular framework of precipitation data acquisition that reduced 40% of precipitation events with nearly- or even inexistent precipitation signals.

All in all, the knowledge gained from the functioning of the basin, the proposed feature engineering methodology, and the evaluation of near-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.

The success of the CCA for assisting the assimilation of SPPs by ML models opens the path for further applications focused on forecasting objectives and/or real-time applications. Moreover, an additional fruitful area for further development would be the combination of SPPs with ground-based precipitation for the development of more accurate flash flood forecasting models. Both the CCA and the FE strategies represents promising tools for hydrologists to develop forecasting models following hydrological concepts in regions otherwise limited by data scarcity issues.

Chapter five: feature engineering strategies for exploiting ground- and satellite-based precipitation data and for adding process-based hydrological knowledge.

Related publications:

- ❖ **Muñoz, P., Corzo, G., Solomatine, D., Feyen, J., & Céleri, R. (2023).** *Near-real-time satellite precipitation data ingestion into peak runoff forecasting models. Environmental Modelling & Software, 160, 105582.*
 - ❖ **Muñoz, P., Muñoz, D.F., Orellana-Alvear, J., & Céleri, R. (2023).** *Flash flood forecasting using readily-available satellite precipitation and machine learning feature engineering strategies. Hydrological Sciences Journal. In review.*
-

The application of FE strategies for improving the efficiency of ML flash flood forecasting models represents a promising opportunity for hydrologists. The opportunity is to understand the use of ML technique as a starting point for exploiting available information coming from ground- and satellite-based sources, and for adding hydrological knowledge by testing forecasting hypotheses on top of ML statistical/computational advantages.

In terms of precipitation data availability, two common case scenarios can be encountered when developing forecasting systems for mountain complex systems such as the Andes. These scenarios are: i) the absence of ground-based data, or ii) the existence of insufficient ground-based data for characterizing spatial precipitation patterns. The first scenario, i.e., precipitation ungauged hydrological systems, is commonly experienced in mountain macro-scale hydrological systems where climate variability and the complexity of the terrain makes it unfeasible to monitor precipitation. Here, the solution lies in exploiting precipitation from SPPs, and to use of FE strategies for deriving as much as process-based hydrological knowledge for improving ML forecasting efficiencies.

Whereas the second scenario is referred to the cases when there is an existing yet insufficient ground-based precipitation monitoring network for precipitation, which is more often found in meso-scale hydrological systems. For this case, the utility of SPPs is to complement its spatial representation with the accuracy of ground-based information in the pursuit of accurate flash flood forecasting models. In addition, we added process-based hydrological knowledge such as flow division (directflow and baseflow), and the corresponding water residence times in the soil layers producing these subflows. For both case scenarios, the use of FE strategies for assisting ML techniques allows the use of unvalidated SPPs due to lack or low density of precipitation gauges.

5.1 Aim and objectives

To develop ML flash flood forecasting models assisted by FE strategies for the exploitation of satellite-based precipitation data and the addition of process-based hydrological knowledge in meso- and macro-scale hydrological systems.

Objectives

- To develop ML flash flood forecasting models using FE strategies to exploit SPPs in a precipitation ungauged macro-scale hydrological system.
- To develop ML flash flood forecasting models assisted by FE strategies in a meso-scale hydrological system with an existent ground-based precipitation data.

5.2 Peak runoff forecasting in a precipitation ungauged macro-scale hydrological system

This case study addresses the knowledge gap in developing flash flood forecasting models for precipitation ungauged hydrological systems. For this, the solution is to employ non-validated near-real-time satellite products together with a combination of FE strategies for comprehending the functioning of a catchment for short-term lead times, and for improving the forecasting efficiency of RF models. The FE strategies selected were flow separation into baseflow and directflow, and precipitation-runoff event classification according to precipitation attributes derived from satellite imagery. This event-based forecasting approach was done for gaining an understanding of the hydrological functioning of the basin as well as for improving the forecasting efficiencies of peak runoffs. An additional objective was to unravel the influence of multiple precipitation-runoff responses through specialized runoff forecasting of the classified events. The

proposed methodology was applied to the Jubones basin located in the southern Andes of Ecuador, and for short-term lead times between 1 and 6 hours to account for peak runoffs.

5.2.1 Dataset processing

The dataset comprises ~3.5 years of hourly information on two variables, precipitation, and runoff for the period January 2019 to June 2022. Precipitation data were retrieved from two near-real-time databases, the IMERG-Early Run (ER), and the PERSIANN-Cloud Classification System (CCS) products. Data were extracted at the finest temporal resolution (30 minutes and 1 hour for the IMERG-ER and PERSIANN-CCS, respectively) and then aggregated to the hourly time step.

Figure 5.1 presents the mean annual precipitation measured by both satellite products in the Jubones basin, with mean (maximum) annual precipitation depths of 729 (1167) and 1532 (2759) mm, respectively. The mean annual precipitation differences of 803 and 1592 mm for the mean and the maximum precipitation are attributed to the resolution differences as well as the measuring principle of each product. It is also worth noting the difference in the number of pixels (timeseries) obtained with each satellite product, 174 and 30 pixels for the PERSIANN-CCS and the IMERG-ER, respectively.

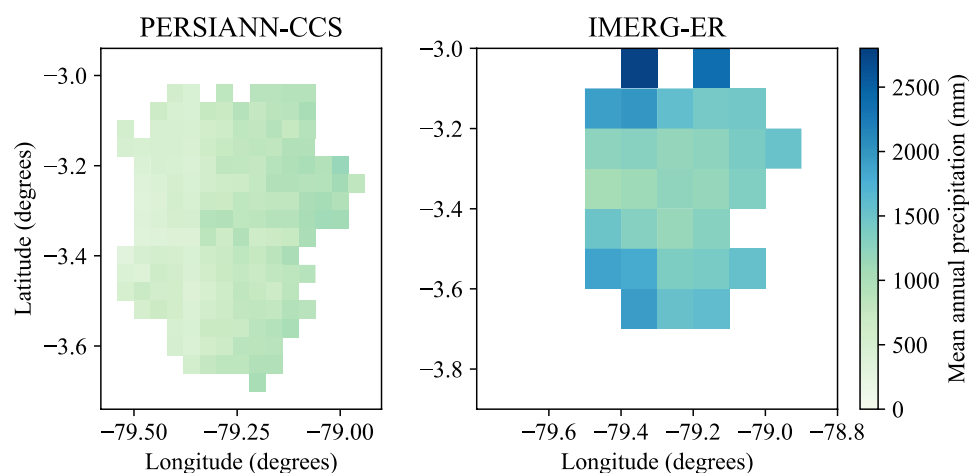


Figure 5.1. Mean annual precipitation measured by the PERSIANN-CCS and the IMERG-ER satellite products for the study period from January 2019 to June 2022 (Jubones basin, Ecuador).

Moreover, although there is lack of ground precipitation gauges operating in the basin, a comparison with the study of Hasan and Wyseure [63] was presented in Chapter One. As a result, it was concluded that for the period 1982-1998, mean annual ground-based precipitation better agreed with the PERSIANN-CCS product. Nevertheless, the PERSIANN-CCS validation with ground measurements is not a limiting issue since precipitation is merely an estimator of peak runoffs when ML techniques are employed. Thus, we rather exploited the spatiotemporal variability of both precipitation products under the assumption that the overall bias of each of them remains constant for the study area.

On the other hand, hourly runoff data was collected for a hydrological station in the outlet of the basin, i.e., the entrance MSF hydropower dam. The runoff data were facilitated by the Corporación Eléctrica del Ecuador (CELEC EP, <https://www.celec.gob.ec/>), the company that manages the MSF hydropower dam.

5.2.2 Methodology

Figure 5.2 summarizes the methodology of this study. First, nearly-independent peak runoff events were selected, and the hourly runoff time series (total flow) was separated into baseflow and directflow series (Figure 5.2a). The purpose of separating total flow (runoff) is to characterize the different orders of magnitude of hydrological processes [95]. Here, the assumption is that differentiated subflow modeling, followed by the summation of both subflows will produce more efficient total flows than the modeling of total flow directly. This will also allow building ML models of different complexity for each subflow. For the subflow separation task, the generalized Chapman filter technique was selected following the recommendations of Willems [95], and Corzo and Solomatine [117]. The subflow filtering principle is based on a numerical digital filter implemented through a linear reservoir modeling concept. The subflow separation method is available within the WETSPRO tool [95].

Second, the precipitation imagery associated with peak events was processed using an object-based Connected Component Analysis (CCA) to extract key precipitation object attributes (Figure 5.2b). The CCA is applied to the precipitation dataset following a modular approach. This means that the CCA is preferably applied to the finest spatial-resolution product (PERSIANN-CCS), and for the cases when no precipitation is detected by the PERSIANN-CCS, the CCA is applied to the supplementary IMERG-ER database. The precipitation attributes extracted from the CCA serve

to classify multiple extreme precipitation-runoff events. Third, for the development of forecasting models, we employed two internal ML sub-models, one for baseflow and the other one for directflow, which were summed up to provide the total flow (Figure 5.2c). Finally, we contrasted the performances of the developed forecasting models developed for increasing lead times and considered specialized models according to the classification of extreme events. A step-by-step explanation of the proposed methodology is presented in the following subsections.

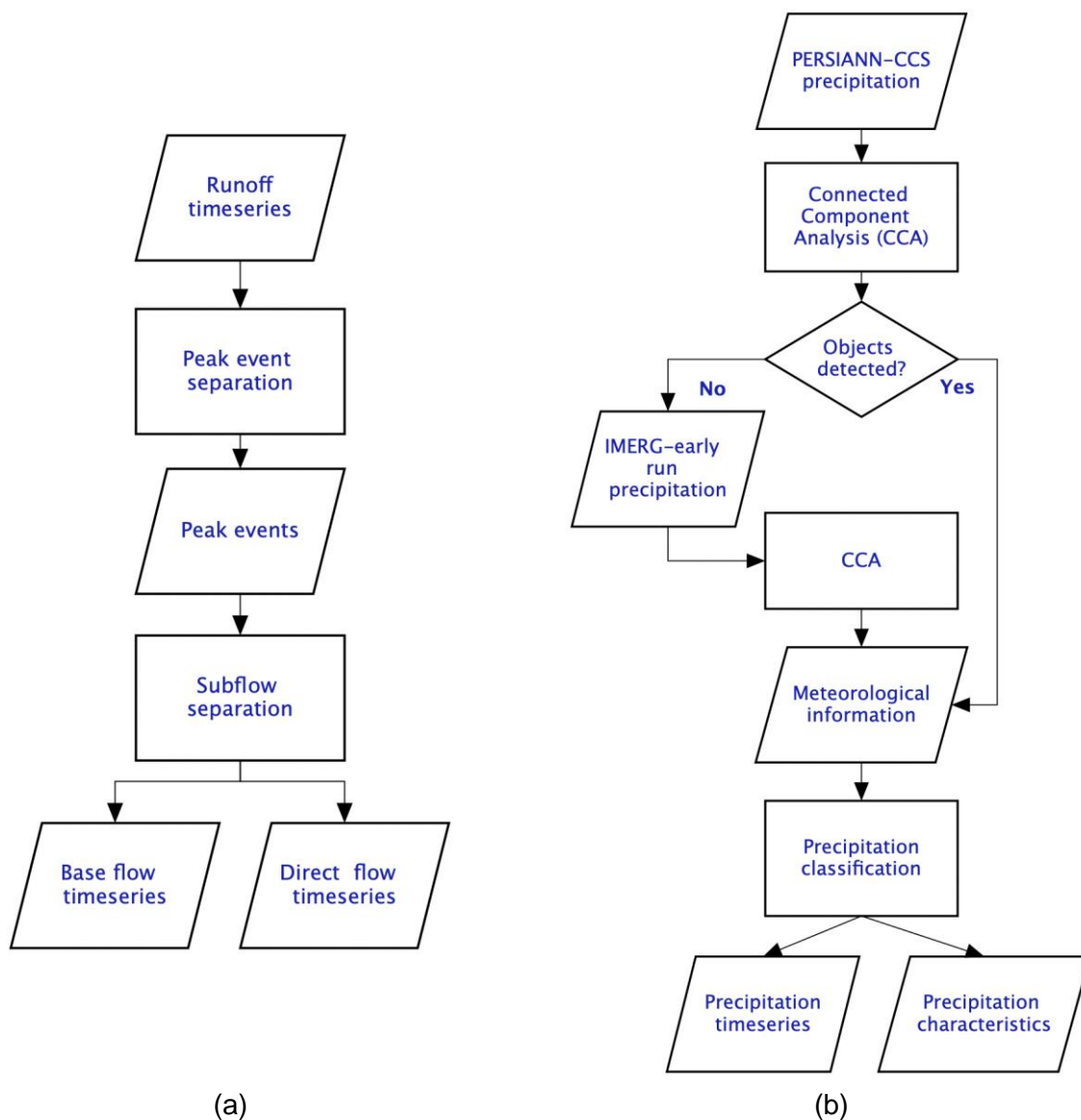
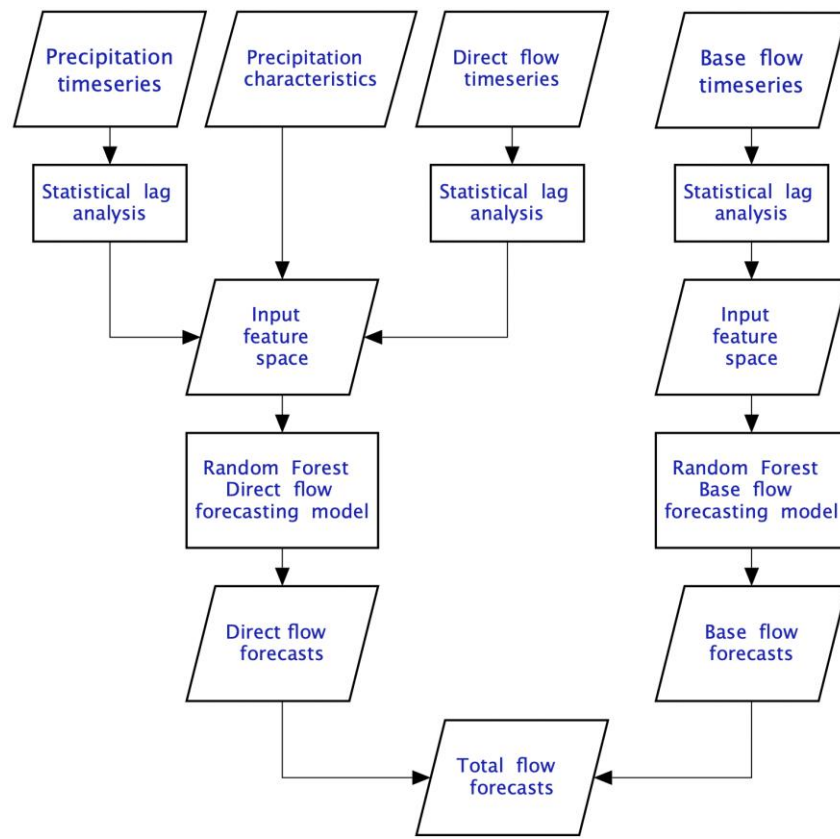


Figure 5.2. Scheme of the methodology for developing peak runoff forecasting models, (a) extreme peak runoff selection and subflow separation, (b) satellite precipitation processing, and (c) forecast modeling approach.



(c)

Figure 5.2 (continuation). Scheme of the methodology for developing peak runoff forecasting models, (a) extreme peak runoff selection and subflow separation, (b) satellite precipitation processing, and (c) forecast modeling approach.

5.2.2.1 Determination of nearly-independent flash flood events and subflow separation

Flash events were selected from the complete runoff time series by using the WETSPRO time series tool [95]. For the subflow separation task, the generalized Chapman filter technique was selected following the recommendations of Willems [95] and Corzo and Solomatine [117]. The subflow separation method is also available within the WETSPRO tool.

The calibration of the WETSPRO tool was done with the following parameter values. First, an inter-event time of 12 hours, i.e., two consecutive events are considered nearly independent when

separated by a period of at least 5 days. Secondly, a runoff maximum drop-down ratio of 0.6, which means that runoff, q , drops down in between two consecutive events to a ratio $\frac{q_{min}}{q_{max}} < 0.6$. Based on the calibration, 81 nearly-independent peak flow events could be delineated. Figure 5.3a shows the obtained hourly baseflow and directflow time series together with the 81 flash flood events depicted as blue dots, while Figure 5.3b plots the exceedance probability of total flow.

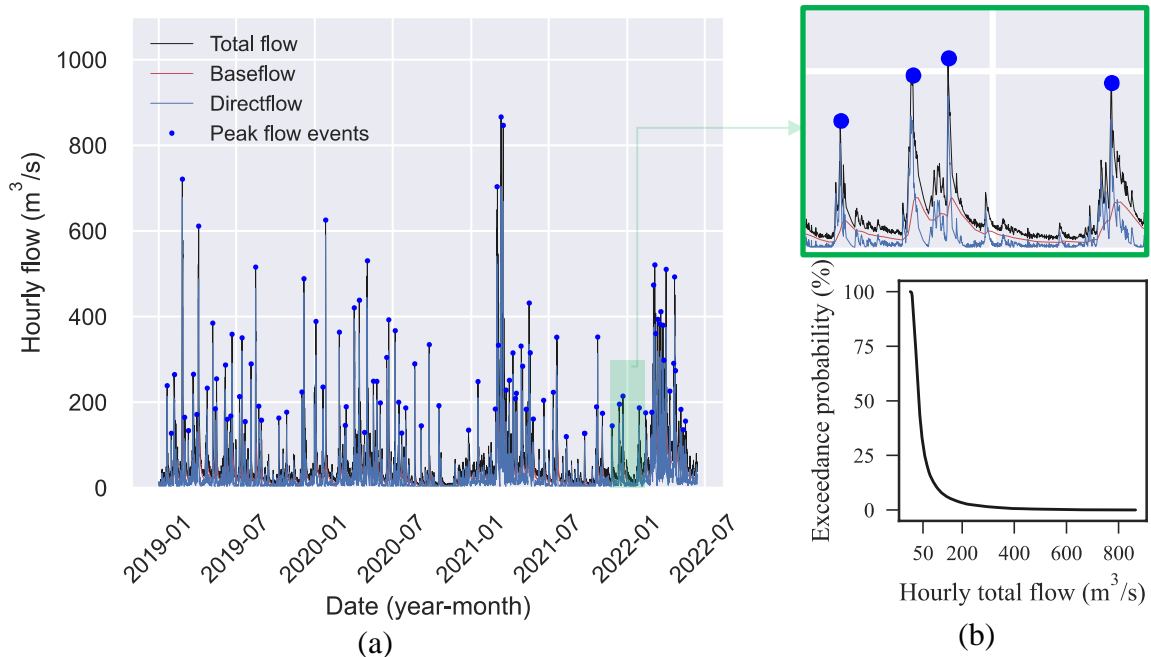


Figure 5.3. (a) Directflow and baseflow separation from the total flow time series at the outlet of the Jubones basin. Peak flow events selected with the WETSPRO tool are displayed as blue dots. (b) Exceedance probability of total flow for the study period (01/01/2019 to 13/06/2022).

5.2.2.2 Object-based CCA and precipitation-event classification

The precipitation imagery corresponding to the selected flash events was processed using the object-based CCA developed by Laverde-Barajas [123] (see Chapter Four). Moreover, according to the modular approach for precipitation data acquisition, the IMERG-ER was used as a supplement dataset to the finest spatial resolution, the PERSIANN-CCS. For this, we applied a simple under-sampling technique. It consisted of dispersing the information contained in a pixel into several subdivided pixels, i.e., the IMERG-ER cell of size $0.1 \times 0.1^\circ$ was converted into ~ 6.4

cells with a resolution of $0.04 \times 0.04^\circ$. The CCA was implemented through the scikit-image processing package in Python® version 3.7 [134].

From the CCA, two attributes of each precipitation event were derived: the extension of the precipitation objects (local and spatial extensive) and the duration of the events (short and long). As a result, four precipitation event classes could be defined: i) Local and short-duration 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) [123].

5.2.2.3 Development of peak runoff forecasting models

The forecasting of peak runoffs was obtained by summing up the forecasts of two internal models, one model for baseflow and one model for directflow. The purpose of separating total flow into baseflow and directflow was to characterize the different orders of magnitude of hydrological processes, i.e., subflow responses to precipitation [95]. The subflow separation was done for the base model (considering all flash flood events) as well as for each precipitation event class. All the models were developed using the RF algorithm for regression (see Chapter Two).

For the baseflow model, we assumed a slow (neglectable) response of this subflow to precipitation. As a result, baseflow is assumed to be solely affected by gradual changes in the past baseflow, i.e., fully autoregressive. On the contrary, the quick response of directflow to precipitation was assumed to be influenced by changes in precipitation and directflow.

The input feature space construction for the RF models was conducted following the methodology used by [47]. In summary, the input feature space was formed by three elements. First, hourly precipitation (for each pixel) and runoff timeseries (i.e., baseflow and directflow). Second, two precipitation characteristics from the CCA: total volume and total area of the precipitation objects. And third, past lag information of precipitation (for each pixel) and runoff. The number of precipitation and runoff lags were determined according to statistical correlation analyses: partial- and auto-correlation functions for runoff, and cross-correlation functions for precipitation.

Moreover, we performed a feature selection analysis to reduce the input dimension of the RF models. For this, we used a sensitivity analysis aimed at calculating the relative importance of each feature to the output [84]. This is done by measuring the variance of the output produced by a single feature without considering the interaction between features. The purpose was to retain

only features accounting for at least 80 % of the total relative importance. The variance produced can be calculated using the equations depicted in Chapter Two.

Moreover, the implementation of the RF algorithm demands the tuning of several hyperparameters. For hydrological applications, the most influencing hyperparameters are the number of trees (n_trees), the maximum depth for pruning (max_depth), and the maximum number of features to perform the splits ($max_features$) [79]. We obtained the optimal combination of these three hyperparameters based on a RGS implemented under a 10-fold cross-validation algorithm to prevent overfitting. The NSE (defined in Chapter Two) between simulations and observations was used as a measure of agreement for the training subsets of each model. Table 5.1 shows the search space (domain) of the selected RF hyperparameters for the optimization task.

Table 5.1. Search space of the RF hyperparameters.

Hyperparameter	Domain
n_trees^*	20;1000;10
$max_features$	$n_features^\psi$, $n_features^{(1/2)}$, $\log_2(n_features)$
max_depth^*	40;800;10

* Domain defined by min, max, and increment. ψ $n_features$ refers to the number of estimators (features) in the input feature space

5.2.2.4 Model evaluation

For model evaluation, instead of selecting a fraction of the total number of peak flow events for training/testing purposes, we employed the leave-one-out cross-validation (LOOCV) algorithm. This means that each event was treated as an independent testing subset while the remaining events were used for training purposes. In the end, the overall performance of a model (NSE) was calculated by averaging the NSE on the testing subsets when all events were tested separately. This was done since only a few events might be available after the classification task.

For each event, we simulated the peak runoff inside a 24-hour window for capturing the entire hydrograph. To quantify model performance, we used a collection of four metrics following the guidelines proposed by Moriasi et al. [92]. The NSE coefficient [93] was set as the reference metric for measuring and comparing the overall fit of model simulations to observations. The

evaluation was complemented with the KGE [94] to account for peak flow underestimations and low flow overestimations, the percent bias, and the RMSE. Moreover, we contrasted the average NSE coefficients of each model with the corresponding OOB errors (also calculated as NSE coefficients). The corresponding equations are listed in Chapter Two.

5.2.3 Results

First, CCA results for the 81 nearly-independent peak flow events showed that for 15 events (19 %) there was no clear precipitation signal from the PERSIANN-CCS product. For these 15 cases, we applied the CCA to the IMERG-ER dataset following the modular approach described in Chapter Four. From the CCA, we derived duration and extension thresholds of 7 hours and 50 km², respectively (Figure 5.4). These thresholds served to classify peak runoff events into four precipitation classes, 23 events for the LSE class, 24 for the LLE, 25 for the SEE, and 9 for the SLE. Moreover, analysis of the centroid occurrence of the precipitation objects did not reveal any precipitation hotspots in the basin that could be associated with peak runoff events. There was no evidence that centroid occurrence is driven or can be related to any physical attribute of the Jubones basin (e.g., soil type, land use, elevation, topography, etc.). This might indicate the nonexistence of orographic precipitation enhancement (i.e., cloud formation due to orographic lifting of air masses).

On the other hand, the input feature space to each model with a certain lead time was partly formed with lagged information on precipitation and runoff. For runoff, results of the ACF and PACF suggested using 12 lags (hours), with a 95% confidence level for both correlation values. Similarly, for precipitation, the Pearson cross-correlation function determined correlations higher than 0.2 for 13 lags (hours). Cross-correlation results are consistent with the estimated concentration time of the Jubones basin of 11 hours (average time using the equations of the U.S. Army Corps of Engineers, Johnstone, and Giandotti, being the equations recommended for the basin extension [108]). For the RF hyperparameterization of each model, we obtained averaged NSE coefficients between simulations and observations always higher than 0.98. Table 5.2 exemplifies the optimized combination of hyperparameters found for the forecasting models of 1-hour lead time.

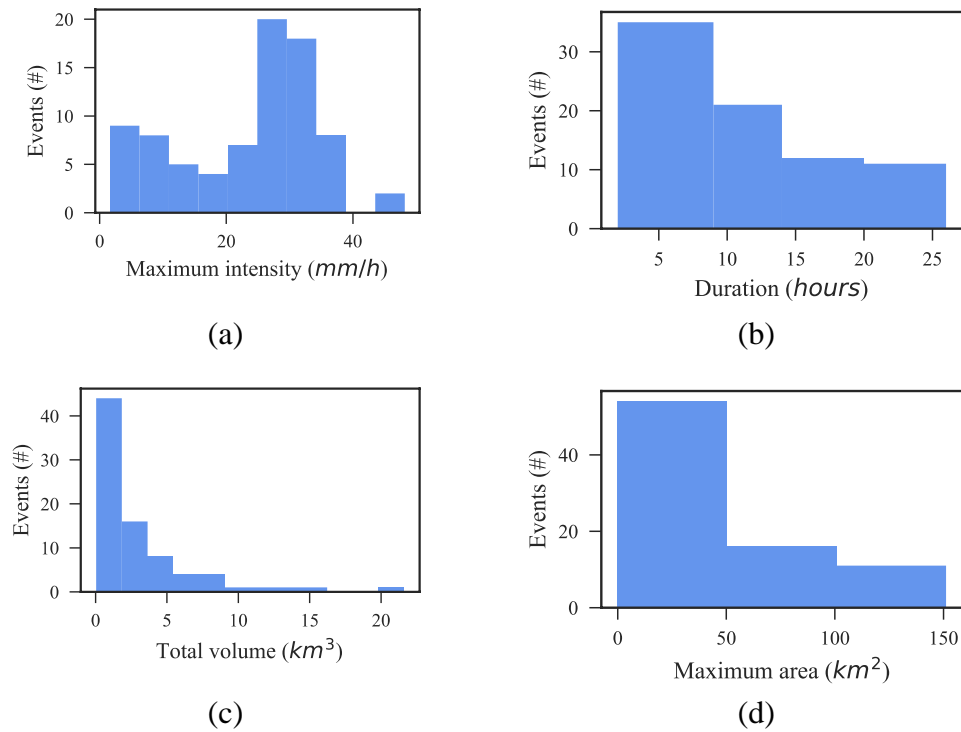


Figure 5.4. Meteorological precipitation information retrieved from 81 extreme hydrological events: (a) maximum intensity, (b) event duration, (c) total volume, and (d) maximum area

Table 5.2. RF hyperparameterization of the forecasting models for the 1-hour lead time

	Random Forest hyperparameters			
	Events [#]	n_trees	max_features	max_depth
Base model	81	300	21	200
LSE	23	280	9	220
LLE	24	250	21	190
SLE	25	300	21	160
SEE	9	300	9	180

Concerning model efficiencies, Table 5.3 presents the averaged performance metrics for the base models, and for the specialized peak runoff forecasting models. In all cases, we present separately the performances for the baseflow, directflow, and the resulting total flow. The color mapping of this table was done on a column-by-column basis. This allows comparing the performances of the models (for a given metric) across lead times, and between the base and the specialized models. The darkest colors represent the best performances.

The first striking result visible in this table is that the base models proved to be satisfactory, with NSE coefficients for total flow varying from 0.86 to 0.59, for lead times between 1 to 6 hours, respectively. We contrasted these values with the NSE coefficients obtained for the validation subset (OOB errors), and we found an overall pattern for the OOB errors to be higher than NSE values using the LOOCV. We found a maximum difference of 0.14 for the 4-hour directflow model, whereas for baseflow models the differences were lower (maximum 0.06 for the 1-h models). The higher OOB errors can be attributed to the fact that the validation is performed on a randomly selected one-third of the training subset; thus, considering the 24-h window of each event, it might be possible that most of the scrutinized runoff does not correspond to peak values. As a result, the calculation of NSE coefficients using the LOOCV provides a more severe evaluation of the forecasting models.

Moreover, according to the criterion of Singh et al. [137], the obtained RMSE values for the base models were also satisfactory for all lead times since their magnitudes were lower than half the standard deviation of measured total flow, $126.2 \text{ m}^3 \cdot \text{s}^{-1}$. The evolution of model performance with lead time is explained by previous thoughts and follows a logical path: the forecast ability of RF decreases with increasing lead time. Moreover, it was also clear that the modeling difficulty came from the modeling of directflow where NSE-values for the base models decayed to 0.36 (6-hour lead time). Nevertheless, the satisfactory performance for total flow was a remarkable outcome since the input feature space was derived from non-validated near-real-time satellite estimates using only the object-based CCA as the processing tool.

Once the base models were evaluated, further analysis focused on the specialized peak runoff forecasting models. First, for total flow, it is apparent from Table 5.3 that spatially-extensive and short-duration events (SEE) produced the lowest performances (NSE values across lead times) when compared to the remaining event classes. These results were confirmed by the OOB errors, which followed a similar pattern across even types, where the lowest values were obtained for SEE. Therefore, it is apparent that SEE are the major source of error for the base models. Here, the hypothesis is that for SEE, precipitation over a mosaic of land uses and soil types produces complex directflow responses that are difficult to be learned by RF regressors. The reason is that small precipitation volumes over extensive areas might be lost before converting into runoff, especially in non-saturated conditions. Although this issue is strongly linked to land uses, soil types, and the saturation state of the basin, such biophysical information was neither available

(soil type and soil moisture) nor updated (land use) for the basin, and could therefore not be used as additional inputs to the forecasting models.

Table 5.3. Model efficiencies (LOOCV evaluation framework) for the base and specialized forecasting models across lead times.

Lead time	Class	Baseflow					Directflow					Totalflow					
		NSE	KEG	RMSE	PBIAS	NSE	KEG	RMSE	PBIAS	NSE	KEG	RMSE	PBIAS	NSE	KEG	RMSE	PBIAS
1h	Base model	0.86	0.70	20.50	15.40	0.88	0.85	28.60	4.50	0.86	0.81	33.55	4.85	0.86	0.81	33.55	4.85
	LLE	0.89	0.86	5.06	5.10	0.81	0.83	19.58	2.53	0.87	0.84	38.89	5.69	0.87	0.84	38.89	5.69
	LSE	0.87	0.79	5.64	11.12	0.81	0.79	15.68	10.56	0.78	0.71	29.48	9.87	0.78	0.71	29.48	9.87
	SLE	0.85	0.80	24.56	8.79	0.70	0.72	57.46	15.40	0.70	0.74	63.27	19.80	0.70	0.74	63.27	19.80
	SEE	0.78	0.68	27.86	18.89	0.30	0.33	93.65	101.50	0.29	0.32	89.63	51.22	0.29	0.32	89.63	51.22
2h	Base model	0.82	0.75	16.58	6.52	0.75	0.76	46.28	4.50	0.82	0.75	65.50	5.87	0.82	0.75	65.50	5.87
	LLE	0.87	0.88	6.22	7.50	0.71	0.75	28.80	3.69	0.70	0.72	42.77	8.70	0.70	0.72	42.77	8.70
	LSE	0.79	0.77	6.98	5.44	0.65	0.65	33.58	8.06	0.65	0.64	55.16	10.45	0.65	0.64	55.16	10.45
	SLE	0.68	0.77	35.56	9.02	0.58	0.59	70.05	16.06	0.49	0.53	84.50	20.50	0.49	0.53	84.50	20.50
	SEE	0.48	0.50	37.81	23.06	0.37	0.37	123.40	121.33	0.22	0.15	131.84	83.01	0.22	0.15	131.84	83.01
4h	Base model	0.78	0.69	35.40	7.80	0.63	0.70	50.40	0.48	0.69	0.70	81.41	9.87	0.69	0.70	81.41	9.87
	LLE	0.79	0.75	9.33	7.80	0.38	0.43	40.22	5.68	0.63	0.66	62.50	10.44	0.63	0.66	62.50	10.44
	LSE	0.60	0.48	23.87	11.56	0.35	0.31	70.80	12.50	0.50	0.52	66.90	15.62	0.50	0.52	66.90	15.62
	SLE	0.49	0.46	40.52	13.25	0.19	0.22	65.80	7.95	0.26	0.27	62.50	21.60	0.26	0.27	62.50	21.60
	SEE	0.32	0.28	38.96	27.13	0.06	0.10	140.52	129.12	0.09	0.13	146.58	88.56	0.09	0.13	146.58	88.56
6h	Base model	0.67	0.71	26.50	3.56	0.40	0.38	68.50	6.89	0.59	0.55	75.34	5.11	0.59	0.55	75.34	5.11
	LLE	0.70	0.66	14.23	8.08	0.32	0.38	50.24	6.06	0.40	0.39	75.48	8.97	0.40	0.39	75.48	8.97
	LSE	0.59	0.62	26.38	15.60	0.18	0.21	85.67	4.03	0.35	0.41	125.40	16.58	0.35	0.41	125.40	16.58
	SLE	0.45	0.42	45.60	17.69	0.14	0.15	89.96	17.36	0.29	0.28	133.68	28.60	0.29	0.28	133.68	28.60
	SEE	0.25	0.26	55.73	29.86	-3.68	-1.67	119.50	145.10	-0.19	0.02	158.80	112.50	-0.19	0.02	158.80	112.50



This can be also seen in Figure 5.5, where it stands out that overall LSE and LLE perform better than the base model, whereas SLE models perform almost similarly but have the advantage of forecasting the highest peak runoffs. For local events, irrespective of their duration (LLE and LSE), the runoff response strongly depends on the land use and soil characteristics where the precipitation occurs. In these cases, the error seems to be absorbed by the RF algorithm by relying on more specific trees. This can be seen in the higher values for the `max_depth` hyperparameter. On the contrary, RF models with lower `max_depth` values are less specific and complex, thus relying more on input data rather than in the complexity of the models.

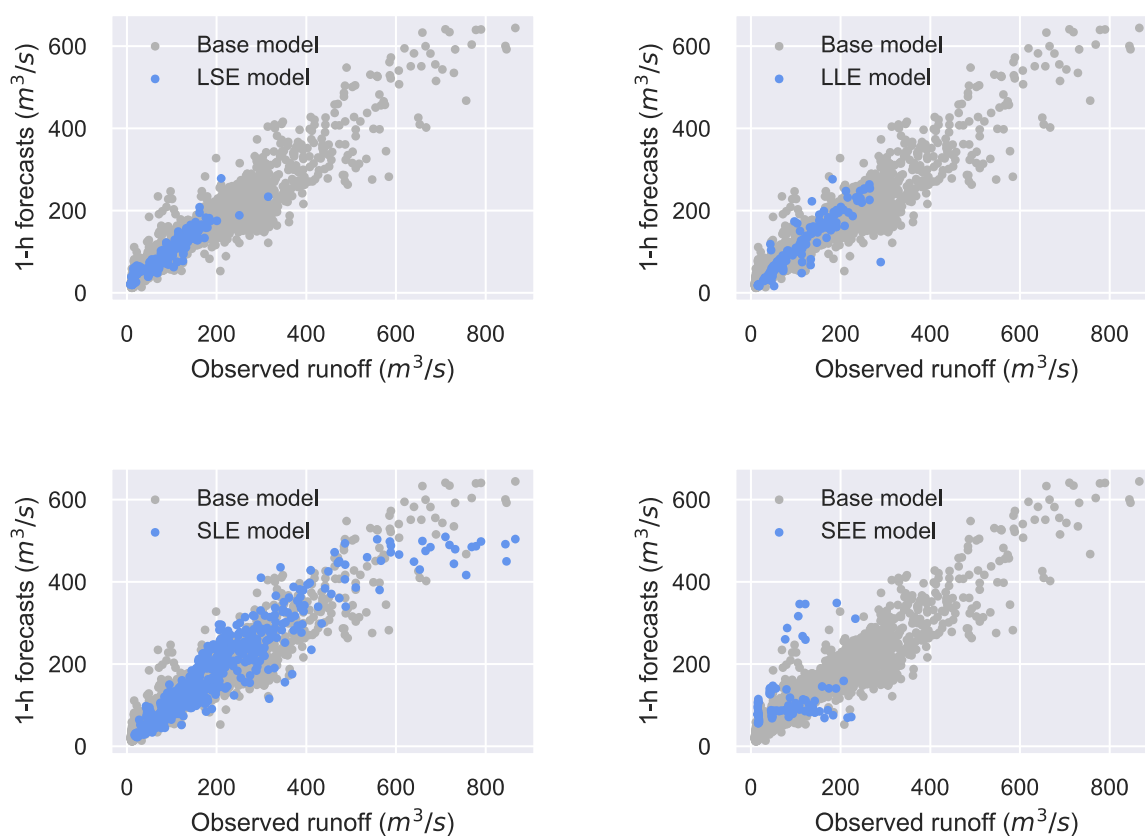


Figure 5.5. Comparison of the scatter plots of the observed and forecasted runoff for the base model and the specialized models for the 1-hour lead time.

Regarding subflow separation for the specialized forecasting models, the NSE values for the 1-hour lead time for baseflow and directflow were comparable for local precipitation events irrespective of their duration (LLE and LSE). On the contrary, for spatially extensive events (SLE

and SEE), there was a clear reduction in the ability of RF models to forecast directflow. This issue becomes critical as the lead time increases, where NSE values for directflow tend to completely deteriorate for the 6-hour case (SEE). Overall, for a given lead time, SLE and SEE models depicted the lowest NSE values for directflow, and consequently total flow. Therefore, the considerably greater amount of precipitation input features is rather producing noise in the directflow models. In such cases, the forecasting ability tends to rely more on their autoregressive power and not on what the models can learn from the processed satellite precipitation.

Concerning the remaining performance metrics (KGE, RMSE, and PBIAS), we found patterns similar to NSE, between specialized runoff forecasting models, and across lead times. For instance, for any lead time, the PBIAS for SEE had the highest values between precipitation event classes, i.e., the highest propensity of forecasted values to be larger than the measured total flow. PBIAS for SEE were always higher than 51 %, with values up to 113 % for the 6-hour lead time. Similarly, the KGE metric, which exposes runoff variability to a greater extent than NSE, revealed the lowest efficiencies for SEE for all lead times. Physically, NSE- and KGE-values might be explained by the fact that the precipitation-runoff correspondence is clear (straightforward) for the cases when either soil saturation is reached or infiltration capacity is exceeded, SLE and LLE, respectively. The straightforward precipitation-runoff relations seem to be well detected by the RF models, especially LLE models, where the highest NSE- and KGE-values were obtained.

5.2.4 Discussion

In this case study, the RF algorithm was used to develop event-based flash flood forecasting models for a representative basin of the tropical Andes, where physically-based modeling is restricted by the lack of sufficient information. The methodology of this study proposes a solution for exploiting near-real-time satellite precipitation data even without validation with ground precipitation networks.

We developed general base models for lead times between 1 and 6 hours to account for peak runoffs in the Jubones basin, which although representing a particular solution for the MSF hydropower dam, is useful for planning the operation of other dams under peak extreme runoff conditions. In addition to the base models, we focused on characterizing extreme hydrological events by analyzing satellite precipitation through an object-based CCA. The development of

specialized models according to precipitation characteristics (duration and extension) served to identify the hidden strength and weaknesses of the already satisfactory base models.

The performances obtained for the base models (NSE=0.86) are comparable to the results obtained in other studies using traditional physically-based models, such as HEC-RAS (NSE=0.92) in the study of Belabid et al. [29], wflow-sbm (NSE=0.58) in the study of Laverde-Barajas [126], and the hydrologic-hydraulic HiResFlood-UCI model (NSE=0.94) in the study of Nguyen et al. [30]. And although this study did not aim at outperforming physically-based models, a clear advantage of the models hereby developed is the possibility to exploit raw near-real-time satellite precipitation. This possibility is, however, feasible under a modular approach for data acquisition, where a second satellite source is used to overcome detection issues of the primary satellite source. We are aware, however, that further analyses must be performed for choosing not only existing precipitation signals but the satellite source, or even data fusion such as the efforts of Chen et al. [138] and Xu et al. [139], presenting the highest correlations with observed runoff. For this, additional near-real-time data sources must be considered.

The framework for unveiling the strengths and weakness of the base models can be replicated to understand the reasons behind unacceptable low performances (e.g., negative NSE), see for instance the study of Belabid et al. [29]. The superiority in performance of the developed local models when compared to spatially extensive events can be explained by the straightforward infiltration- and saturation-excess runoff generation processes in reduced portions of the basin. Conversely, whenever precipitation is extensively distributed within the basin, the forecasting models lower their ability to characterize and learn the specificities of the multiple precipitation-runoff relations. For such cases, the forecasting ability is attributed to the autoregressive dependency in the flow time series. The performances of the specialized forecasting models revealed the need to include information describing the dynamics of antecedent soil saturation during extreme events. This is especially required for initializing the forecast of short-duration precipitation events (SEE and LSE). The antecedent soil saturation state will serve to explain why short-duration and non-extreme precipitation intensities can trigger extreme hydrological events. Given the previous assumption, a future direction would be to include satellite soil moisture observations to improve forecasting efficiencies in ungauged basins, as done in the study of Massari et al. [136].

We are also aware that the findings of this study were obtained with a relatively short-length database when compared to other ML studies; however, the use of the RF together with processing tools, and severe evaluation framework for reducing overfitting served to ensure the use of the models for this study and the daily operation of the MSF hydropower plant. Particularly, for the Jubones basin, we did not find a pattern or hotspot of the precipitation storms that triggered extreme runoff responses in the study period. This has direct implications for the ability of RF models to recognize patterns and demonstrates the necessity of developing specialized forecasting models according to other precipitation characteristics, and not only in distributed modeling (subbasins), which is commonly employed by traditional physically-based models. Finally, the findings and limitations encountered in this study open the path for future research on exploring additional ML techniques for the modeling of spatially-extensive events, or even model ensemble strategies.

5.3 Flash flood forecasting exploiting ground- and satellite-based precipitation data in a meso-scale hydrological system

This case study aims to develop flash flood forecasting models by leveraging ground-based and PERSIANN-CCS precipitation estimates for the Tomebamba catchment located in the tropical Andes of Ecuador. The hydrological model is based on the RF algorithm together with a combination of FE strategies including flow separation into baseflow and directflow, soil moisture modeling, and inclusion of precipitation attributes derived from satellite imagery. The accuracy and suitability of the PERSIANN-CCS are evaluated, and the forecasting ability is tested for lead times between 1 to 12 hours.

5.3.1 Dataset processing

The dataset comprises hourly information on two variables, runoff measured at the outlet of the catchment and precipitation within the catchment for the period Jan/2015 to May/2021. Runoff time series for the Sayausí station were obtained from the local drinking water facility of Cuenca, ETAPA-EP.

Precipitation data were retrieved from ground and satellite sources. Ground estimates were acquired from three rain gauges installed in the upper and middle parts of the catchment, Toreadora at 3395, Virgen del Cajas at 3626, and Chirimachay at 3298 m a.s.l. These rain gauges are located within microcatchment M1 of the Tomebamba. On the other hand, satellite estimates

of precipitation were retrieved from the PERSIANN-CCS database, resulting in 15 pixels-based information for the Tomebamba catchment. Figure 1.3 (Chapter One) shows the PERSIANN-CCS coverage over the study catchment as well as a comparison between the annual cumulated precipitation measured by the satellite- and ground-based products for the study period.

For training and testing purposes, the dataset was split up into training (Jan/2016 to May/2021) and testing (from Jan/2015 to Dec/2016) periods. The selection of these periods was aimed at capturing the highest runoff peaks in the training phase.

5.3.2 Methodology

5.3.2.1 *Evaluation framework and metrics*

The evaluation framework consisted firstly in assessing the accuracy of PERSIANN-CCS precipitation data at a microcatchment-wide scale for hourly, daily, and monthly timescales. Then, we evaluated the PERSIANN-CCS suitability for runoff forecasting with special attention to peak runoffs (flash floods).

The evaluation of quantitative precipitation estimates (QPEs) was performed by comparing the average PERSIANN-CCS and the available rain gauge measurements for the microcatchment M1, where the three rain gauges are installed. The accuracy of the PERSIANN-CCS estimates was evaluated through the coefficient of correlation (CC) and Bias. Whereas the precipitation detection ability was measured with the Probability of Detection (POD), the False Alarm Ratio (FAR), and the Critical Success Index (CSI). The corresponding metrics are presented in Table 5.4 below.

Table 5.4. Efficiency metrics for precipitation.

Metric	Equation	Ideal value
CC	$\frac{\sum_{i=1}^n [(P_i - \bar{P})(G_i - \bar{G})]}{\sqrt{\sum_{i=1}^n (P_i - \bar{P})^2} \sqrt{\sum_{i=1}^n (G_i - \bar{G})^2}}$	1
Bias	$\frac{\sum_{i=1}^n (X_s - X_o)}{\sum_{i=1}^n X_o}$	0
POD	$\frac{A}{A + B}$	1
FAR	$\frac{C}{A + C}$	0
CSI	$\frac{A}{A + B + C}$	1

Where n is the number of instances, P is satellite precipitation, G is ground precipitation, \bar{P} is the mean satellite precipitation, \bar{G} is the mean ground precipitation, $X = P$ for precipitation, $X = Q$ for runoff, A is the frequency of observed precipitation detected by satellite, B is the frequency of observed precipitation not detected by satellite, and C is the frequency of false precipitation detected by a satellite product.

On the other hand, the comparison between runoff observations and forecasts was done using the NSE [93], KGE [94], the RMSE, and the percent bias (see Chapter Two). Additionally, for a more specific evaluation focused on flash floods, we complemented the performance metrics with graphical techniques such as the frequency distribution for peak values and a Box-Cox transformation of total flow.

5.3.2.2 Satellite data processing using an object-based CCA

The satellite data retrieved from the PERSIANN-CCS were processed before their use in a hydrological model. For this, we employed the object-based CCA developed by Laverde-Barajas et al. [123] (see Chapter Four). The CCA serves to identify precipitation objects (storms) and allows extraction of key meteorological features such as maximum intensity, maximum areal extension, maximum volume, storm location, etc. Among them, the use of maximum intensity and maximum areal extension as additional input features was proved to improve RF runoff modeling and forecasting performances [48], [140]. The maximum intensity of each satellite imagery

corresponded to the maximum value encountered among all precipitation objects; whereas the maximum areal extension was calculated as the sum of the areas of the identified objects.

For the CCA detection and localization of precipitation objects, we defined a precipitation sensitivity threshold of 0.5 mm. The detected objects were then filtered according to a minimum threshold area corresponding to 2 pixels ($\sim 20 \text{ km}^2$). Both thresholds were calibrated on a trial-and-error basis to remove noise (e.g., isolated precipitation objects) in the satellite imagery. Then, the filtered objects were passed through a dilation-and-erosion algorithm for image refining. Lastly, the CCA extracted two precipitation attributes, maximum intensity, and maximum areal extension. The CCA was implemented through the scikit-image processing package in Python® version 3.7 [134].

5.3.2.3 Determination of nearly-independent flash flood events and subflow separation

The separation of runoff (total flow) into baseflow and directflow components was performed using the generalized Chapman filter technique, according to the recommendations of Willems [95] and Corzo and Solomatine [117]. The separation method is implemented in the WETSPRO time series tool [95]. Moreover, for the extraction of nearly-independent peak runoff events using a peak-over-threshold approach. Those events will be used for the evaluation framework focused on flash floods.

The calibration of the WETSPRO tool was achieved with the following parameter values. First, recession constants of 8 days and 12 hours for baseflow and directflow, respectively. Second, an inter-event time of 10 hours, i.e., two consecutive events are considered nearly independent when separated by a period of at least 10 hours. And third, a runoff maximum drop-down ratio of 0.1, means that the minimum runoff, q_{\min} , between two events is 0.1 of antecedent q_{\max} . Figure 5.6 shows the obtained hourly baseflow and directflow time series together with the 156 peak flows depicted as blue dots.

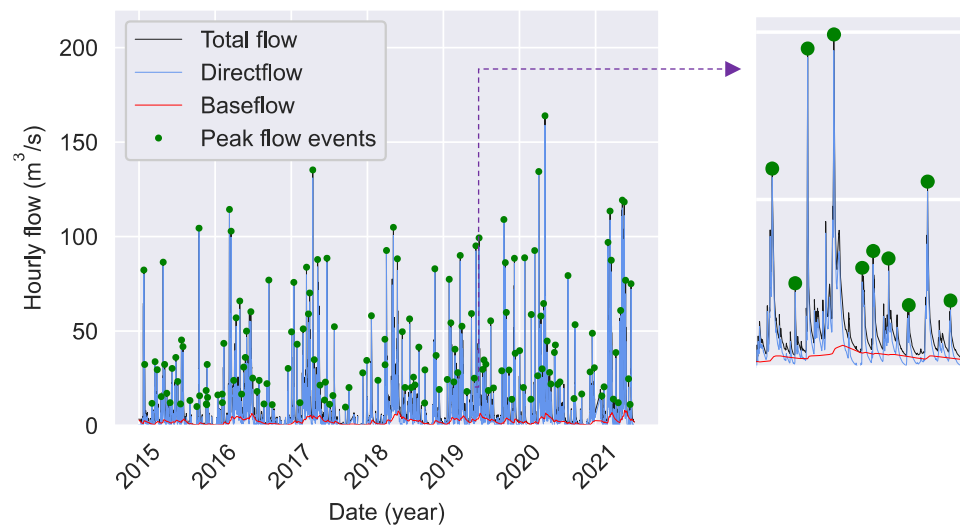


Figure 5.6. Hourly runoff (total flow) of the Sayausí station, and its subflow components (baseflow and directflow). 156 nearly-independent peak flows are displayed as green dots. Study period from Jan/2015 to May/2021.

5.3.2.4 Hydrological forecasting model based on the Random Forest (RF) algorithm

The RF hydrological model consisted of two internal sub-models, baseflow, and directflow, respectively. The forecasts were summed up to obtain the total flow. The subflow separation task was aimed to characterize the different orders of magnitude of hydrological processes [95]. Both models were based on the RF algorithm for regression, which is detailed in Chapter Two.

In terms of model structures, the baseflow model was built as completely autoregressive, i.e., the baseflow response is forecasted regardless of the precipitation input (Figure 5.7). This assumption was based on the study of Mosquera et al. [59] for the same catchment, where isotope analyses revealed that water resides in the deeper soil layer for around 4 weeks, and this layer remains near saturation through the year. Thus, considering the fast response of the catchment (4 hours), flash-flood responses are rather explained by the dynamic of directflow.

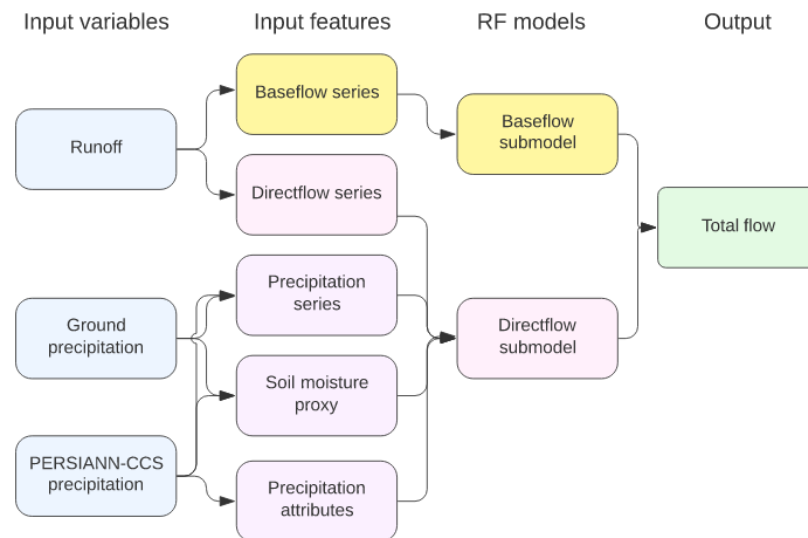


Figure 5.7. Model structures for baseflow and directflow RF forecasting models.

The directflow model considered the fast soil moisture reaction occurring in the rooted layer, with an approximate residence time of 2 weeks [59]. In this case, changes in both precipitation (ground- and satellite-based) and directflow are assumed to influence and/or control the directflow (Figure 5.7). Regarding precipitation features, in addition to the current time and lags of the precipitation series, we included the attributes derived from the CCA, and soil moisture dynamics in the directflow sub-model by adding a proxy feature that accounts for 2 weeks of cumulative precipitation. Moreover, the mean concentration time of each microcatchment served to weigh the precipitation input for the calculation of directflow. This served to account for different runoff responses according to the localization of precipitation storms. The concentration times were calculated by averaging the outputs of the equations of Johnstone, Corps Engineers, Williams, and Haktanir & Sezen, recommended for the extension of the microcatchments [141].

Model hyperparameterization

Several hyperparameters are involved in the construction process of the forest, yet the most influencing for hydrological applications are the number of trees (`n_trees`), the maximum depth for pruning (`max_depth`), and the maximum number of features to perform the splits (`max_features`) [79]. The optimal hyperparameter combinations were obtained through a RGS under a 10-fold cross-validation approach. For this, we used the NSE between forecasts and observations as the measure of agreement for the training subsets.

5.3.2.5 *Input feature space construction and feature selection*

The input feature space construction for the RF models was conducted following the methodology proposed by Muñoz et al., (2018). In summary, the input feature space was formed by three elements. First, hourly precipitation (for each pixel) and runoff timeseries, i.e., baseflow and directflow. Second, the two precipitation characteristics from the CCA: maximum intensity and maximum areal extension. Third, past lag information of precipitation (for each pixel) and runoff. The number of precipitation and runoff lags were determined according to statistical correlation analyses, the ACF and the PACF for runoff and Pearson cross-correlation functions for precipitation.

Moreover, we reduced the input feature dimension by implementing a feature selection procedure. For this, we calculated the relative importance of each feature to the output using sensitivity analysis [84] (see Chapter Two). The input feature space was only composed of features that summed up to 80 % of the total relative importance.

5.3.3 Results and discussion

5.3.3.1 *Satellite data validation at a microcatchment-wide scale*

Figure 5.8 presents the histograms of ground-based and PERSIANN-CCS hourly, nonzero precipitation. For both datasets, the occurrence of precipitation events according to their intensities follows an exponential distribution, where low precipitation intensities are far more frequent than higher values. What stands out in Figure 5.8 is that the PERSIANN-CCS underestimates the occurrence of very light precipitation ($< 1 \text{ mm}\cdot\text{hour}^{-1}$), while it tends to overestimate the occurrence of higher precipitation intensities ($> 5 \text{ mm}\cdot\text{hour}^{-1}$). The underestimation of very light events, which occurs $> 80 \%$ of the time, leads to a lower CC (0.04) and negative bias (-35 %) between satellite and ground-based measurements. This underestimation/overestimation is congruent with the findings of Hong et al. [142], who concluded that PERSIANN-CCS presents higher sensitivity to convective events at expense of missing light precipitation events.

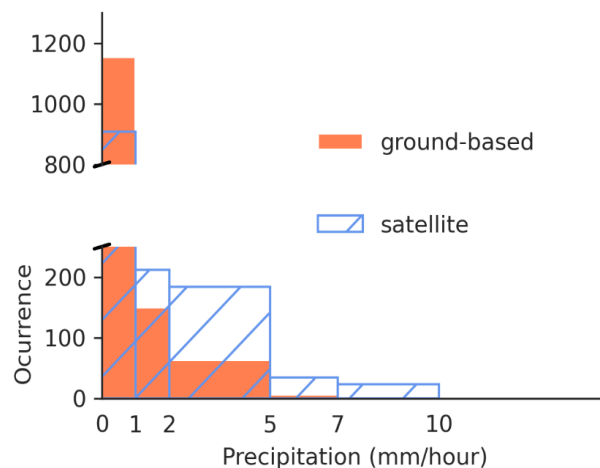


Figure 5.8. Comparison of PERSIANN-CCS and average ground-based (microcatchment M1) histograms of hourly precipitation. Considering the asymmetry of the data, the histograms were split up into different size class bins.

Figure 5.9 shows the scatter density plots from rain gauges and the PERSIANN-CCS at daily and monthly averaged series, respectively. As shown in these figures, the highest correlation was found for the monthly ($CC=0.49$), with the lowest value for the daily scale (0.21). These correlations are comparable to those from the studies of Anjum et al. [143], Salehi et al [144], Moura Ramos Filho et al. [145], Eini et al. [146], and Nguyen et al. [24]. Overall, for both time scales, it can be noted a tendency to underestimate precipitation based on the slope of the data (gray continuous line) when compared to the bisector line (blue dotted line). This is attributed to the error accumulation of the aforementioned underestimation/overestimation issues at the hourly timescale.

On the other hand, the mean POD values are 0.07 and 0.24 for the daily and monthly timescales, respectively. The maximum POD values were obtained for intensities in the range 0-1 mm.hour⁻¹. The FAR and the CSI obtained were unsatisfactory for the hourly timescale, while for the daily scale, we obtained mean values of 0.5 and 0.2, respectively. Overall, the evaluation of QPEs seemed to be highly variable at sub-daily timescales, which matches the findings of J. Li et al. [147], Sadeghi et al. [148], and Zeweldi and Gebremichael [149]. Thus, we suggest that precipitation retrieved from PERSIANN-CCS could not be used with confidence for precipitation studies in the Tomebamba catchment without image correction/calibration/adjustment, thus limiting the use of traditional physically-based hydrological models. Nonetheless, although the

validation results were not very encouraging, the challenge of this study was rather to exploit the spatial characterization for runoff forecasting with the RF algorithm.

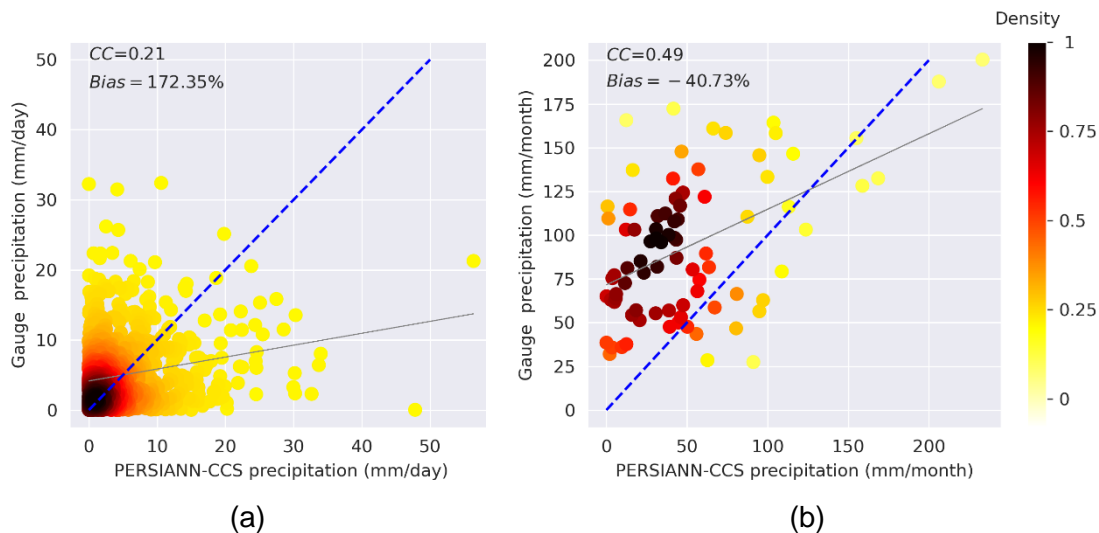


Figure 5.9. Scatter density of PERSIANN-CCS estimates and corresponding ground-based precipitation at (a) daily and (b) monthly scales. Period of analysis from Jan/2015 to May/2021.

5.3.3.2 Meteorological characteristics obtained from the object-based CCA

We found that the majority of precipitation events had a duration of fewer than 8 hours (Figure 5.10a), with predominant intensities below $5 \text{ mm}\cdot\text{hour}^{-1}$ (Figure 5.10b). If we now turn to areal extension, the majority of precipitation events can either cover most of the catchment area or less than 50 km^2 (local events, see Figure 5.10c). Moreover, the combination of intensities, duration, and areal extension results in precipitation volumes of less than 0.0125 km^3 or $38 \text{ l}\cdot\text{m}^{-2}$ per event (Figure 5.10d). Taken together, these results suggest that most extreme hydrological events in the catchment are rather driven by saturation excess and not infiltration excess processes, and their potential to trigger flash floods depends on areal extension and/or the proximity to the catchment outlet for local events. Similar to the previous section, we are aware that the derived precipitation features can be biased. However, the maximum intensity and maximum areal extension attributes were used to improve the learning process of the RF models.

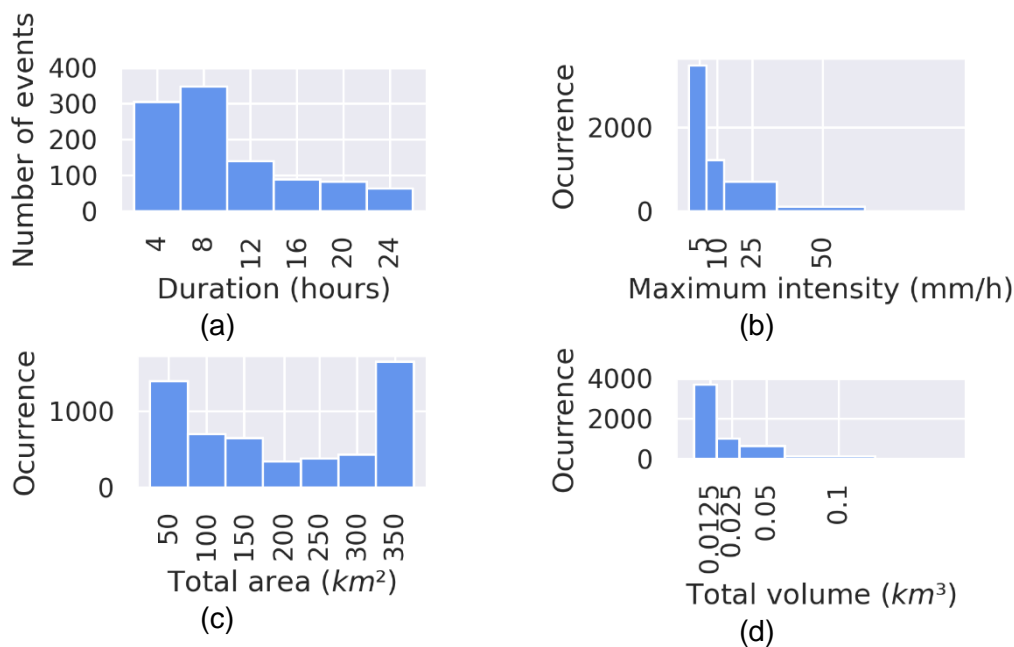


Figure 5.10. Precipitation-event characteristics derived from the CCA: (a) event duration, (b) maximum intensity plotted with different class widths, (c) areal extension, and (d) total volume plotted with different class widths.

5.3.3.3 Hydrological modeling and hyperparameterization

For the baseflow models, the input feature space for each lead time was constructed with baseflow data at the current time, 12 baseflow lags as determined by PAC and AC functions with 95 % of confidence, and cumulated runoff volume for the past 7 days. The cumulated volume was initially set to 4 weeks since is the average residence time in the deeper soil layer; however, 7 days seemed plausible for detecting rapid changes in water storage. This value was determined on a trial-and-error basis during the RF-hyperparameterization task, which is described at the end of this section.

On the other hand, the directflow models for each led time were constructed with both flow and precipitation information. Flow information contemplated directflow data at the current time, and 12 directflow lags (similarly to baseflow). Whereas the precipitation features included precipitation at the current time for each ground- and pixel-based in the catchment, and their corresponding 9 lags as determined by cross-correlations results with a threshold of 0.2. The precipitation series were weighted according to the mean concentration times calculated for each microcatchment (4.6, 3.6, 4.2, 4.9, 1.6, and 1.5 hours for M1-M6, respectively). The mean concentration time of

the entire catchment is 6.4 hours. In addition, we employed the maximum intensity and maximum areal extension attributes derived from the CCA. Moreover, we included 2 days of cumulated past precipitation as a soil moisture proxy variable. The late value was calibrated during the RF tuning task starting from 1 up to 14 days (to reach the average residence time in the rooted layer).

With the aforementioned considerations, each RF model was hyperparameterized with all features available in the input feature space. Then, for each model, up to 70 % of the features were trimmed-off according to their calculated relative importance. For instance, the input feature space for the 4-hour lead time model for directflow contained the information of 114 features, and after trimming off the space dimension was reduced to 35 features. The optimal hyperparameter combinations are presented in Table 5.5 below.

Table 5.5. Optimal combination of RF hyperparameters for the baseflow and directflow forecasting models across lead times.

Lead time [h]	Baseflow			Directflow		
	n_trees	max_features	max_depth	n_trees	max_features	max_depth
1	330	3 (22 %*)	30	300	6 (6 %*)	70
4	320	3 (22 %*)	40	320	6 (6 %*)	80
8	270	8 (58 %*)	80	270	40 (35 %*)	110
12	390	8 (58 %*)	100	420	60 (53 %*)	110

* Percentage of features from the total number of features employed

For all models (different subflows and lead times), the NSE values between observations and simulations during training were always above 0.98. The most interesting aspect of Table 5.5 is that for lead times exceeding the concentration time of the catchment (4 hours), the forecasting performance relies more on the hyperparameterization task than on the information contained in the corresponding input feature space. This can be noted by the higher values obtained for the max_depth, and the max_features hyperparameters, which means that more specific trees and stronger randomization are required for obtaining optimal model performances. Conversely, for the 1- and 4-hour lead times, the input feature space provides information that better describes the response of the catchment to precipitation events.

5.3.3.4 Model evaluation

The forecasting performances (NSE) on the testing subset for baseflow, directflow, and total flow, and across lead times are provided in Table 5.6. This table is quite revealing in several ways.

First, the developed forecasting models proved to be satisfactory for lead times up to 3 times the concentration time of the catchment, i.e., 12 hours, with NSE varying from 0.95 to 0.61. The validity of the methodology proposed in this study can be demonstrated by the fact that we obtained efficiencies comparable to and even higher than in other studies using the PERSIANN-CCS but under physically-based approaches [144], [146], [150]. Therefore, the success of this study has direct implications for the development of flash flood early-warning systems (FEWSs) not only for the study area but also for other Andean catchments where lack of data and non-validated satellite precipitation has been the limiting reason for FEWSs implementation.

Table 5.6. Forecasting performances for the baseflow, directflow, and total flow models across increasing lead times.

Baseflow				
Lead time [h]	NSE	PBIAS	RMSE	KGE
1	0.85	-6.99	0.53	0.89
4	0.86	-6.76	0.52	0.90
8	0.86	-6.54	0.50	0.90
12	0.87	-6.30	0.49	0.90
Directflow				
Lead time [h]	NSE	PBIAS	RMSE	KGE
1	0.94	0.47	1.89	0.95
4	0.81	-3.75	3.24	0.88
8	0.65	-9.19	4.49	0.77
12	0.54	-12.64	5.10	0.68
Total flow				
Lead time [h]	NSE	PBIAS	RMSE	KGE
1	0.95	-1.73	1.84	0.96
4	0.85	-4.64	3.22	0.90
8	0.70	-8.41	4.49	0.81
12	0.61	-10.77	5.13	0.74

Second, it is clear that the forecasting difficulty comes from the modeling of directflow, with NSE varying from 0.94 to 0.54 as the lead time increases. The fact that baseflows are well forecasted (NSE-values are always higher than 0.85) is due to the nearly year-round saturation of the deeper soil layer of the catchment [59]. Thus, the development of fully autoregressive baseflow models seemed to be sufficiently complex, and also fulfills the model parsimony criteria. On the other

hand, the higher complexity of the models for directflow depicted NSE values higher than 0.80 for lead times up to the concentration time of the catchment. For longer lead times, the decay in NSE (minimum 0.54) is attributed to an insufficient description of the governing forces of this subflow component (e.g., topsoil layers and land use information), but mainly due to the lack of relevant precipitation information considering the latency of the satellite imagery and the fact that precipitation is not being forecasted. Thus, precipitation events occurring between the most recent precipitation instance and the forecast horizon are not available for the forecasting task. All these results were corroborated by the remaining performance metrics (RMSE, bias, and KGE). Furthermore, consideration of weights according to the concentration times of each microcatchment served to characterize time differences in precipitation-runoff responses, especially for saturation excess processes where some of the precipitation water volume is used for saturating the topsoil layers and only a fraction reaches the outlet of the catchment. Although there were no substantial differences between the mean concentration times of M1-M6, this approach has the potential to substantially improve model efficiencies in larger catchments, as in the study of P. C. Huang and Lee [118] for an extension greater than 500 km².

Third, even though the PERSIANN-CCS database presented lower correlations with hourly ground-based measurements, the use of longer time scale estimates (as in the soil moisture proxy accounting for 2 days of precipitation) served to better relate the spatial precipitation patterns to directflow, and consequently total flow. Similarly, the addition of available ground-based satellite estimates was crucial for achieving a good trade-off between precipitation accuracy and spatial characterization. This was demonstrated since the efficiency of the forecasting models whose precipitation information was solely derived from ground-based data was inferior when compared to the models exploiting both satellite- and ground-based precipitation (maximum difference of 0.15 in NSE). Another comparison can be done with the previous study of Muñoz et al. [47] in the Tomebamba catchment, where the efficiencies of forecasting models using ground-based information alone, yet for a shorter study period (2.5 years), achieved efficiencies lower than the ones found in this study (maximum difference of 0.12 in NSE).

In addition, a comparison between total flow forecasts and observations across increasing lead times is presented in Figure 5.11. The most striking result is the substantial difference in the degree of correlation for lead times shorter (Figures 5.11a-b), and longer (Figures 5.11c-d) than

the concentration time of the catchment. For the 8- and 12-hour lead times there is a trend of underestimating hourly runoff, thus negative biases of -8.4 and -10.8, respectively.

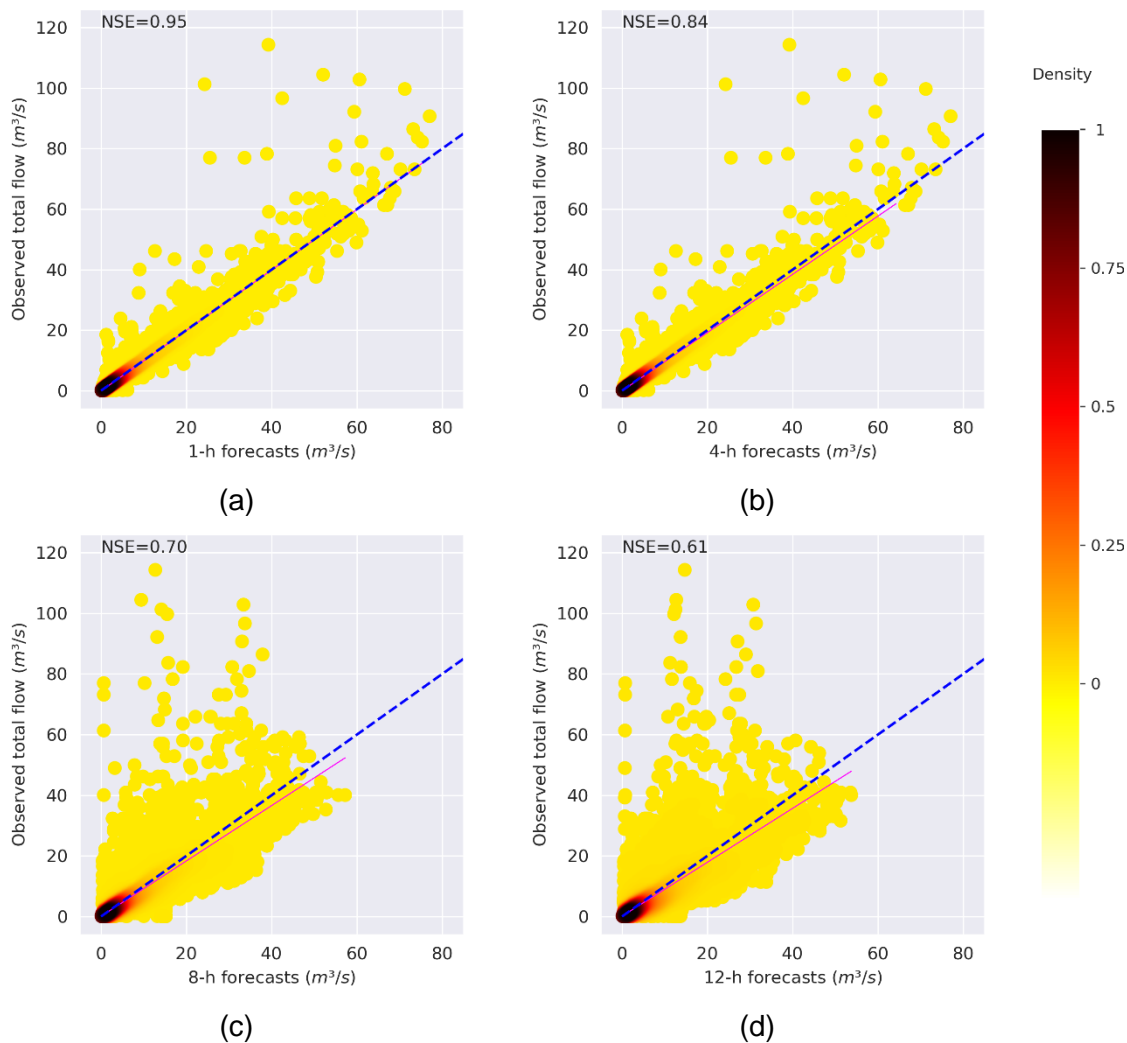


Figure 5.11. Scatter density plot of timeseries of forecasted total flow for the testing periods: (a-d) plots for lead times of 1, 4, 8, and 12 hours, respectively.

To complement the previous assessment, the empirical peak value distributions for both forecasts and total flow observations are provided in Figure 5.12a. Overall total flows up to $40 m^3 \cdot s^{-1}$, there is a good match between observations and forecasts for lead times as long as the concentration time of the catchment (4 hours). On the other hand, for flows higher than $40 m^3 \cdot s^{-1}$, which occur less than 2% of the time, there is a systematic underestimation of peak values towards the upper tail of the distribution. The underestimation becomes critical as the lead time increases.

Additionally, Figure 5.12b shows the correlation between forecasted and observed peak flows, where the mean error and the standard deviation correspond to the 1-hour forecasts. In this figure, model residuals are represented by the horizontal and vertical differences between each point and the bisector line, and the dependence of the standard deviation on the total flow magnitude was disrupted with a λ -value of 0.25 (Box-Cox transformation). What stands out in this figure is the higher scatter (higher standard deviation of peak flows from the bisector lines), and higher bias (systematically lower mean peak flows) for increasing lead times.

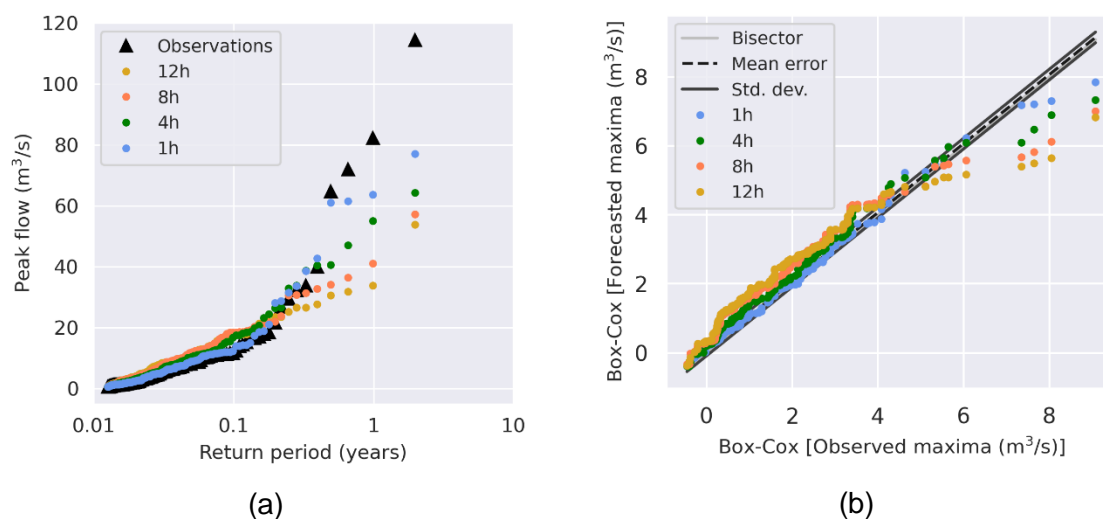


Figure 5.12. (a) Empirical peak value distribution, (b) Comparison of nearly independent peak flow maxima.

5.4 Summary and conclusions

In this chapter, we addressed the two common case scenarios that can be encountered when developing peak runoff forecasting models for mountain complex systems. These are the cases of precipitation ungauged macro-scale hydrological systems, and meso-scale systems where there is an existence yet insufficiently-dense ground-based monitoring network for precipitation. In both cases, we developed forecasting models using FE strategies for exploiting SPPs data by ML models, and for adding process-based hydrological knowledge and concepts for improving forecasting efficiencies.

First, for the case of precipitation ungauged hydrological systems, we developed event-based forecasting models for the Jubones basin. For this application, the following main conclusions can be drawn:

- The use of near-real-time SPPs data assisted by the CCA served to improve precipitation representation over the basin, and consequently, enhanced the performance of flash floods forecasting models.
- FE strategies applied to precipitation and runoff datasets allowed the development of specialized models for precipitation events and for subflow components. This application also showed that for the Jubones basin, spatially extensive events are the most difficult precipitation scenarios to model without deep characterization of the study area (land use, soil moisture, and topography, among other features).
- The description of soil saturation conditions might also enhance runoff forecasting associated with local precipitation events, yet their higher efficiencies are attributed to the straightforward infiltration- and saturation-excess runoff generation relations.
- The best forecasting performances were obtained for peak runoffs triggered by short-extension precipitation events (<50 km²) where infiltration- or saturation-excess runoff responses are well learned by the RF models. Conversely, the forecasting difficulty is associated with extensive precipitation events. For such conditions, a deeper characterization of the biophysical characteristics of the basin is encouraged for capturing the dynamic of directflow across multiple runoff responses.

Second, for the case of systems with existent ground-based monitoring network, we developed continuous time-series flash flood forecasting models for the Tomebamba catchment. In this application, SPPs data served to complement the existing ground-based monitoring network. For this case study, the following main conclusions can be drawn:

- The PERSIANN-CCS product was validated at a microcatchment-wide scale, and even though the quality of the hourly satellite data might be questionable, the development of forecasting models proved to be satisfactory. We attribute the forecasting success of this study to both the merging of ground- and satellite-based precipitation, and to the FE strategies applied for adding physical knowledge of the system to the forecasting models. For instance, for precipitation, the CCA served to derive key precipitation attributes that enriched the input feature space of the RF models. We are aware, of course, that the

uncertainties involved (occasioned by lack of precipitation accuracy) might be misleading when using physically-based models; however, instead of discarding this high-resolution information, this represented a great opportunity for the RF algorithm (or other ML techniques).

- It was crucial to separate the total flow signal into baseflow and directflow components and model each subflow with different levels of complexity. This approach served also to efficiently forecast peak flows with anticipation times up to 12 hours (i.e., 3 times the concentration time of the catchment). Taken together, the findings of this research have significant implications for operational applications such as the development of FEWSs, but also for gaining an understanding of precipitation-runoff responses in catchments otherwise limited by insufficiently dense monitoring networks
- This application has also shown differences in the hyperparameterization of the RF models across lead times. Overall, the inclusion of physical knowledge regarding the functioning of the catchment, and the reasoning behind the hyperparameterization task served to enlighten the always-questioned veracity of black-box, data-driven models. In this way, we attempt to endorse the use of ML hydrological models as a blank page, where hydrological forecasting hypothesis can be tested on top of statistical/computational advantages.

Based on both case studies, it can be concluded that the application of FE strategies for assisting ML flash flood forecasting is a promising approach not only for modeling but for forecasting flash floods in macro- and meso-scale complex systems. Moreover, the proposed methodology for processing non-validated SPPs, and the modular approach for data acquisition have direct implications for operational hydrology where the lack of precipitation data has been a limiting issue.

A natural extension of this work would be to better represent the physical conditions of the basin before and during a precipitation event, not necessarily extreme, which might cause a peak runoff response. It would be also advisable to explore additional ML algorithms, hybridization, and/or model assembling aimed at maximizing the encountered forecasting efficiencies.

Chapter six: summary, conclusions and feature work.

This doctoral thesis presents original research that advances the field of machine learning (ML) peak runoff including flash flood forecasting. The research focuses on developing innovative methodologies for constructing efficient forecasting models in complex systems at both macro and meso scales considering that traditional physically-based hydrological models are not feasible due to the highly-variable and poorly-monitored flash flood driving forces. Therefore, the thesis aims to enhance ML peak runoff forecasting models through the implementation of feature engineering (FE) strategies. These FE strategies will allow for the effective utilization of ground- and satellite-based precipitation data and process-based hydrological knowledge in macro- and meso-scale complex systems. The ultimate goal of this research is to improve the accuracy and reliability of ML-based forecasting models.

In our research, we focused on improving the accuracy of peak runoff forecasting models by implementing feature engineering (FE) strategies in both macro- and meso-scale systems. For the macro-scale system, we utilized satellite precipitation products (SPPs) data to not only obtain spatiotemporal information but also to derive process-based hydrological knowledge related to precipitation events that cause peak runoffs. This was made possible due to the areal extension of the system, which enabled us to extract features from the satellite precipitation imagery.

On the other hand, for the meso-scale system, where local land use and topography significantly impact the occurrence of flash floods, we focused on exploiting the accuracy of ground-based precipitation data in conjunction with the spatial representation from SPPs using FE. Additionally, we incorporated process-based hydrological knowledge related to subflow division into directflow and baseflow, as well as the corresponding residence times in the soil layers producing these subflows. By implementing these FE strategies, we were able to improve the effectiveness of flash flood forecasting models in both macro- and meso-scale systems.

In order to achieve the goal of our thesis, we conducted research on two primary aspects. The first one focused on utilizing the latest machine learning (ML) techniques to develop peak runoff forecasting models. The second aspect involved the implementation of a feature engineering (FE) strategy, specifically the connected component analysis (CCA), to leverage readily-available satellite precipitation products (SPPs) information and overcome issues related to spatial and temporal data scarcity. We explored the use of SPPs and FE in two distinct hydrological systems.

Firstly, for a macro-scale system where ground-based data is lacking, we utilized SPPs and FE to address this data gap and improve the accuracy of the peak runoff forecasting model. Secondly, for a meso-scale hydrological system where ground-based data was available but insufficient for characterizing spatial precipitation patterns, we employed SPPs and FE to complement the existing data and further enhance the flash flood forecasting models. By integrating these advanced ML techniques and FE strategies, we were able to significantly improve the accuracy of flash flood forecasting models in both macro- and meso-scale hydrological systems.

To address the first aspect of our research, we focused on developing machine learning (ML) models for both qualitative and quantitative peak runoff forecasting, which are detailed in Chapters Two and Three of the thesis. Our approach utilized precipitation data sourced solely from ground-based stations, and we evaluated the forecasting performance of various commonly-used ML techniques at varying lead times. Through our experimentation, we arrived at the conclusion that the random forest (RF) algorithm displayed the most potential for further development, given its high levels of accuracy, robustness, and ability to effectively handle complex, short datasets. As such, we identified RF as the most promising ML technique for operational hydrology and real-time applications.

For the second aspect, we proposed FE strategies to exploit readily-available SPPs, as well as to inform ML models with process-based hydrological knowledge. We developed techniques to derive precipitation storm attributes, model runoff dynamics (including subflow modeling), and create specialized models for various precipitation conditions that trigger peak runoffs. Initially, we validated our approach in runoff and peak runoff modeling in current time, before transitioning to a forecasting problem (as discussed in Chapter Four). Our findings demonstrated that improving the spatial representation of precipitation using near-real-time SPPs data led to more accurate peak runoff forecasts. Furthermore, by developing specialized models based on precipitation attributes such as duration and areal extension, we were able to improve the efficiency of our models and identify precipitation-runoff scenarios that were previously difficult for the RF forecasting models to learn.

Given the success of FE strategies for flash flood modeling, we employed the developed methodology for two case studies representative of macro- and meso-scale systems, i) a macro-scale precipitation ungauged system, and ii) a meso-scale system with an existing yet insufficient ground-based precipitation data (Chapter Five). The first case study focused on event-based

forecasting for a macro-scale ungauged precipitation system. The use of FE not only improved the accuracy of peak runoff forecasting but also provided valuable insights into the hydrological behavior of the system. The second case study aimed to address operational hydrology problems for a meso-scale system with insufficient ground-based precipitation data. Continuous timeseries forecasting using the developed RF models was used to forecast both runoff and flash floods, demonstrating the versatility and effectiveness of the FE approach in various hydrological applications.

In the case of the precipitation-ungauged macro-scale system, we found that the best forecasting results were obtained for peak runoff events triggered by short-duration precipitation events, which have simple infiltration- or saturation-excess runoff responses that can be effectively learned by the RF models (Chapter Five). Conversely, forecasting for extensive precipitation events with complex runoff responses was more challenging, as these events involve multiple soil types and land uses. For such macro-scale systems, we recommend using the developed methodology to create general or base models, while specialized forecasting would require a more detailed characterization of the system's biophysical characteristics to capture the dynamics of runoff across multiple response types.

Whereas, in the case of the meso-scale system, our findings showed that the ML forecasting methodology, assisted by FE strategies, was effective in assimilating SPPs information, despite concerns about the quality of SPPs for short timescales (hourly). A key factor in achieving efficiency improvements was the separation of the total flow into baseflow and directflow components, as the developed ML models for baseflow and directflow had different levels of complexity according to soil dynamics. This demonstrates the importance of considering the hydrological processes underlying the data and tailoring the modeling approach accordingly, rather than relying solely on raw data inputs.

The integration of ML techniques, SPPs, and FE strategies in peak runoff forecasting presents a significant advancement in the field of hydrology and has the potential to provide valuable information for decision-makers and hydrologists in complex mountain systems. These methodologies overcome the challenges associated with sparse monitoring networks and the limitations of traditional physically-based models. Our findings demonstrate that the proposed methodologies can be applied to other macro- and meso-scale systems with some modifications based on available data and system-specific characteristics. The development and evaluation of

peak runoff forecasting models in other systems can be facilitated using the methodologies presented in this study, which provides a significant contribution towards mitigating the devastating impacts of flash floods.

One potential future direction of this work is to explore the derivation of new features to enhance the input space of ML models or to construct more complex models with specific physical conditions or restrictions. Terrain-based attributes, such as slope, soil types, land use, and the presence of depressions and floodplains, can be considered to calculate the water retention potential. Moreover, whenever sufficient data is available, physically-based hydrological modeling can be also a promising approach in complex hydrological systems (mountainous areas) [151]. Various studies adopt the Ensemble Kalman Filter and wavelet analysis to forecast precipitation, temperature, evaporation, and runoff [152], while others concentrate on generating multi-system, multi-member seasonal predictions of mountain snow depth and resources [153]. Additionally, hydro-meteorological ensemble prediction systems are employed for short, medium, and long-term forecasting [154], [155], also exploiting both ground-based and remote sensing imagery of climatological and hydrological variables in these regions [156]. Physical models prove to be effective in flat terrains, however, in areas with sudden climate changes such as oceans and mountains, data-driven models, such as the LSTM model, exhibit higher accuracy in flat areas but lesser accuracy in mountainous or oceanic regions [157].

Another potential direction is to explore more advanced ML techniques, such as deep learning (DL). DL techniques, such as long short-term memory networks and convolutional neural networks, have shown superior forecasting performances compared to RF. Moreover, they have the potential to transfer parameters between systems, which can help to address one of the biggest issues of RF - the extrapolation of peak runoffs. However, the application of DL models requires considerable datasets (i.e., estimators and target variables), which may not be available for recently monitored systems. Nevertheless, DL models could be the future of forecasting models, and their development warrants further investigation.

Finally, in addition to the development of flash flood forecasting models, it is crucial to consider the speed and effectiveness of communication to the public once a flood warning is triggered. Therefore, we recommend the development of a web portal and/or mobile applications to disseminate peak runoff and flash flood forecasts in a timely and effective manner. Such tools can enhance preparedness among the population and help authorities evaluate hazard risk.

Additionally, the development of an integrated action plan from a local and regional perspective can further mitigate the impact of flash floods. It is imperative to prioritize these developments alongside the advancement of forecasting models to ensure a comprehensive and effective approach to peak runoff management.

References

- [1] R. P. Trends and C. Andes, "Recent Precipitation Trends and Floods in the Colombian Andes," pp. 1–22, 2019, doi: 10.3390/w11020379.
- [2] M. M. Q. Mirza, "Climate change, flooding in South Asia and implications," *Reg Environ Change*, vol. 11, no. 1, pp. 95–107, 2011.
- [3] D. Paprotny, A. Sebastian, O. Morales-Nápoles, and S. N. Jonkman, "Trends in flood losses in Europe over the past 150 years," *Nat Commun*, vol. 9, no. 1, p. 1985, 2018.
- [4] S. Stefanidis and D. Stathis, "Assessment of flood hazard based on natural and anthropogenic factors using analytic hierarchy process (AHP)," *Natural Hazards*, vol. 68, no. 2, pp. 569–585, 2013, doi: 10.1007/s11069-013-0639-5.
- [5] R. Vos, M. Velasco, and E. Labastida, "Economic and social effects of" El Nino" in Ecuador, 1997-8," *ISS Working Paper Series/General Series*, vol. 292, pp. 1–55, 1999.
- [6] L.-C. Chang *et al.*, "Building an Intelligent Hydroinformatics Integration Platform for Regional Flood Inundation Warning Systems." Multidisciplinary Digital Publishing Institute, 2019.
- [7] S. K. Min, X. Zhang, F. W. Zwiers, and G. C. Hegerl, "Human contribution to more-intense precipitation extremes," *Nature*, vol. 470, no. 7334, pp. 378–381, 2011, doi: 10.1038/nature09763.
- [8] G. Sofia, G. Roder, G. Dalla Fontana, and P. Tarolli, "Flood dynamics in urbanised landscapes: 100 years of climate and humans' interaction," *Sci Rep*, vol. 7, no. July 2016, pp. 1–12, 2017, doi: 10.1038/srep40527.
- [9] F. J. Chang and Y. Y. Hwang, "A self-organization algorithm for real-time flood forecast," *Hydrol Process*, vol. 13, no. 2, pp. 123–138, 1999, doi: 10.1002/(SICI)1099-1085(19990215)13:2<123::AID-HYP701>3.0.CO;2-2.
- [10] R. Brouwer and R. van Ek, "Integrated ecological, economic and social impact assessment of alternative flood control policies in the Netherlands," *Ecological Economics*, vol. 50, no. 1–2, pp. 1–21, 2004, doi: 10.1016/j.ecolecon.2004.01.020.
- [11] Y. Hundecha, J. Parajka, and A. Viglione, "Flood type classification and assessment of their past changes across Europe," *Hydrology and Earth System Sciences Discussions*, pp. 1–29, 2017.
- [12] T. Turkington, K. Breinl, J. Ettema, D. Alkema, and V. Jetten, "A new flood type classification method for use in climate change impact studies," *Weather Clim Extrem*, vol. 14, pp. 1–16, 2016.
- [13] M. Borga, E. Gaume, J. D. Creutin, and L. Marchi, "Surveying flash floods: gauging the ungauged extremes," *Hydrol Process*, vol. 2274, no. 4, pp. 2267–2274, 2008, doi: 10.1002/hyp.7111.

- [14] E. T. Knocke and K. N. Kolivras, "Flash flood awareness in southwest Virginia," *Risk Analysis: An International Journal*, vol. 27, no. 1, pp. 155–169, 2007.
- [15] I. Braud *et al.*, "Advances in flash floods understanding and modelling derived from the FloodScale project in South-East France," *FLOODrisk 2016 - 3rd European Conference on Flood Risk Management DOI:*, vol. 04005, 2016, doi: 10.1051/e3sconf/20160704005.
- [16] I. Ruin, J. D. Creutin, S. Anquetin, and C. Lutoff, "Human exposure to flash floods - Relation between flood parameters and human vulnerability during a storm of September 2002 in Southern France," *J Hydrol (Amst)*, vol. 361, no. 1–2, pp. 199–213, 2008, doi: 10.1016/j.jhydrol.2008.07.044.
- [17] B. F. Ochoa-Tocachi *et al.*, "Impacts of land use on the hydrological response of tropical Andean catchments," *Hydrol Process*, vol. 30, no. 22, pp. 4074–4089, Oct. 2016, doi: 10.1002/HYP.10980.
- [18] R. Rollenbeck and J. Bendix, "Rainfall distribution in the Andes of southern Ecuador derived from blending weather radar data and meteorological field observations," *Atmos Res*, vol. 99, no. 2, pp. 277–289, Feb. 2011, doi: 10.1016/J.ATMOSRES.2010.10.018.
- [19] D. Ballari, R. Giraldo, L. Campozano, and E. Samaniego, "Spatial functional data analysis for regionalizing precipitation seasonality and intensity in a sparsely monitored region: Unveiling the spatio-temporal dependencies of precipitation in Ecuador," *International Journal of Climatology*, vol. 38, no. 8, pp. 3337–3354, Jun. 2018, doi: 10.1002/JOC.5504.
- [20] P. Muñoz, R. Célleri, and J. Feyen, "Effect of the Resolution of Tipping-Bucket Rain Gauge and Calculation Method on Rainfall Intensities in an Andean Mountain Gradient," *Water (Basel)*, vol. 8, no. 11, p. 534, 2016.
- [21] R. Padrón, J. Feyen, M. Córdova, P. Crespo, and R. Célleri, "Rain Gauge Inter-Comparison Quantifies Differences in Precipitation Monitoring," *LA GRANJA. Revista de Ciencias de la Vida*, vol. 31, no. 1, pp. 7–20, 2020.
- [22] G. J. Huffman *et al.*, "NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG)," *Algorithm Theoretical Basis Document (ATBD) Version*, vol. 4, p. 26, 2015.
- [23] K. Hsu, X. Gao, S. Sorooshian, and H. v Gupta, "Precipitation estimation from remotely sensed information using artificial neural networks," *Journal of applied meteorology*, vol. 36, no. 9, pp. 1176–1190, 1997.
- [24] P. Nguyen *et al.*, "The PERSIANN family of global satellite precipitation data: A review and evaluation of products," *Hydrol Earth Syst Sci*, vol. 22, no. 11, pp. 5801–5816, Nov. 2018, doi: 10.5194/HESS-22-5801-2018.
- [25] G. Tang, D. Long, and Y. Hong, "Systematic anomalies over inland water bodies of High Mountain Asia in TRMM precipitation estimates: No longer a problem for the GPM era?," *IEEE Geoscience and remote sensing letters*, vol. 13, no. 12, pp. 1762–1766, 2016.

- [26] P. Nguyen *et al.*, “Satellites track precipitation of super typhoon Haiyan,” *Eos, Transactions American Geophysical Union*, vol. 95, no. 16, pp. 133–135, 2014.
- [27] S. Sakib, D. Ghebreyesus, and H. O. Sharif, “Performance Evaluation of IMERG GPM Products during Tropical Storm Imelda,” *Atmosphere (Basel)*, vol. 12, no. 6, p. 687, 2021.
- [28] S. Sorooshian, P. Nguyen, S. Sellars, D. Braithwaite, A. AghaKouchak, and K. Hsu, “Satellite-based remote sensing estimation of precipitation for early warning systems,” *Extreme natural hazards, disaster risks and societal implications*, vol. 1, p. 99, 2014.
- [29] N. Belabid, F. Zhao, L. Brocca, Y. Huang, and Y. Tan, “Near-real-time flood forecasting based on satellite precipitation products,” *Remote Sens (Basel)*, vol. 11, no. 3, p. 252, 2019.
- [30] P. Nguyen, A. Thorstensen, S. Sorooshian, K. Hsu, and A. AghaKouchak, “Flood forecasting and inundation mapping using HiResFlood-UCI and near-real-time satellite precipitation data: The 2008 Iowa flood,” *J Hydrometeorol*, vol. 16, no. 3, pp. 1171–1183, 2015.
- [31] M. P. Clark *et al.*, “The evolution of process-based hydrologic models : historical challenges and the collective quest for physical realism,” no. 1969, pp. 3427–3440, 2017.
- [32] A. Brath, A. Montanari, and E. Toth, “Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model,” *J Hydrol (Amst)*, vol. 291, no. 3–4, pp. 232–253, 2004, doi: 10.1016/j.jhydrol.2003.12.044.
- [33] C. W. Dawson and R. L. Wilby, “Hydrological modelling using artificial neural networks,” *Prog Phys Geogr*, vol. 25, no. 1, pp. 80–108, 2001, doi: 10.1191/030913301674775671.
- [34] S. Galelli and A. Castelletti, “Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling,” *Hydrol Earth Syst Sci*, vol. 17, no. 7, pp. 2669–2684, 2013, doi: 10.5194/hess-17-2669-2013.
- [35] H. v. Gupta, T. Wagener, and Y. Liu, “Reconciling theory with observations: Elements of a diagnostic approach to model evaluation,” *Hydrol Process*, 2008, doi: 10.1002/hyp.6989.
- [36] P. Willems, “Parsimonious rainfall-runoff model construction supported by time series processing and validation of hydrological extremes - Part 1: Step-wise model-structure identification and calibration approach,” *J Hydrol (Amst)*, vol. 510, pp. 578–590, 2014, doi: 10.1016/j.jhydrol.2014.01.017.
- [37] A. Mosavi, P. Ozturk, and K. W. Chau, “Flood prediction using machine learning models: Literature review,” *Water (Switzerland)*, vol. 10, no. 11, pp. 1–40, 2018, doi: 10.3390/w10111536.
- [38] P. C. Young, “Advances in real-time flood forecasting,” *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 360, no. 1796, pp. 1433–1450, 2002.

- [39] M. Valipour, M. E. Banihabib, and S. M. R. Behbahani, "Parameters estimate of autoregressive moving average and autoregressive integrated moving average models and compare their ability for inflow forecasting," *J Math Stat*, vol. 8, no. 3, pp. 330–338, 2012.
- [40] M. Valipour, M. E. Banihabib, and S. M. R. Behbahani, "Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir," *J Hydrol (Amst)*, vol. 476, pp. 433–441, 2013, doi: 10.1016/j.jhydrol.2012.11.017.
- [41] J. Adamowski, H. Fung Chan, S. O. Prasher, B. Ozga-Zielinski, and A. Sliusarieva, "Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada," *Water Resour Res*, vol. 48, no. 1, Jan. 2012, doi: 10.1029/2010WR009945.
- [42] G. Bontempi, S. ben Taieb, and Y.-A. le Borgne, "Machine Learning Strategies for Time Series Forecasting.," in *eBIS*, 2012, pp. 62–77.
- [43] A. Elshorbagy, G. Corzo, S. Srinivasulu, and D. P. Solomatine, "Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology," *Hydrol Earth Syst Sci*, vol. 14, no. 10, 2010, doi: 10.5194/hess-14-1931-2010.
- [44] G. Corzo and D. Solomatine, "Knowledge-based modularization and global optimization of artificial neural network models in hydrological forecasting," *Neural networks*, vol. 20, no. 4, pp. 528–536, 2007.
- [45] D. Solomatine, L. M. See, and R. J. Abraham, "Data-driven modelling: concepts, approaches and experiences," *Practical hydroinformatics*, pp. 17–30, 2009.
- [46] D. P. Solomatine and M. B. Siek, "Modular learning models in forecasting natural phenomena," *Neural networks*, vol. 19, no. 2, pp. 215–224, 2006.
- [47] P. Muñoz, J. Orellana-Alvear, P. Willems, and R. Céleri, "Flash-flood forecasting in an andean mountain catchment-development of a step-wise methodology based on the random forest algorithm," *Water (Switzerland)*, vol. 10, no. 11, 2018, doi: 10.3390/w10111519.
- [48] P. Muñoz, G. A. Corzo Perez, D. Solomatine, J. Feyen, and R. Céleri, "Use of near-real-time satellite precipitation data and machine learning to improve extreme runoff modeling," *Earth and Space Science Open Archive*, p. 28, 2021, doi: 10.1002/essoar.10508861.1.
- [49] S. M. Hosseini and N. Mahjouri, "Integrating support vector regression and a geomorphologic artificial neural network for daily rainfall-runoff modeling," *Appl Soft Comput*, vol. 38, pp. 329–345, 2016.

- [50] H. Tongal and M. J. Booij, "Simulation and forecasting of streamflows using machine learning models coupled with base flow separation," *J Hydrol (Amst)*, vol. 564, pp. 266–282, 2018.
- [51] C.-C. Young, W.-C. Liu, and M.-C. Wu, "A physically based and machine learning hybrid approach for accurate rainfall-runoff modeling during extreme typhoon events," *Appl Soft Comput*, vol. 53, pp. 205–216, 2017.
- [52] B. Bhattacharya and D. P. Solomatine, "Neural networks and M5 model trees in modelling water level–discharge relationship," *Neurocomputing*, vol. 63, pp. 381–396, 2005.
- [53] Y. B. Dibike, S. Velickov, D. Solomatine, and M. B. Abbott, "Model induction with support vector machines: introduction and applications," *Journal of Computing in Civil Engineering*, vol. 15, no. 3, pp. 208–216, 2001.
- [54] D. P. Solomatine and K. N. Dulal, "Model trees as an alternative to neural networks in rainfall–runoff modelling," *Hydrological Sciences Journal*, vol. 48, no. 3, pp. 399–411, 2003.
- [55] J. Orellana-Alvear, R. Céleri, R. Rollenbeck, P. Muñoz, P. Contreras, and J. Bendix, "Assessment of Native Radar Reflectivity and Radar Rainfall Estimates for Discharge Forecasting in Mountain Catchments with a Random Forest Model," *Remote Sens (Basel)*, vol. 12, no. 12, p. 1986, 2020.
- [56] D. Dávila, "21 experiencias de sistemas de alerta temprana en América Latina." Soluciones Prácticas, 2016.
- [57] S. del Granado, A. Stewart, M. Borbor, C. Franco, E. Tauzer, and M. Romero, "Sistemas de Alerta Temprana para Inundaciones: Análisis Comparativo de Tres Países Latinoamericanos," 2016.
- [58] W. Buytaert, J. Deckers, and G. Wyseure, "Description and classification of nonallophanic Andosols in south Ecuadorian alpine grasslands (páramo)," *Geomorphology*, vol. 73, no. 3–4, pp. 207–221, Feb. 2006, doi: 10.1016/J.GEOMORPH.2005.06.012.
- [59] G. M. Mosquera, P. Crespo, L. Breuer, J. Feyen, and D. Windhorst, "Water transport and tracer mixing in volcanic ash soils at a tropical hillslope: A wet layered sloping sponge," *Hydrol Process*, vol. 34, no. 9, pp. 2032–2047, Apr. 2020, doi: 10.1002/HYP.13733.
- [60] G. Esquivel-Hernández *et al.*, "Moisture transport and seasonal variations in the stable isotopic composition of rainfall in Central American and Andean Páramo during El Niño conditions (2015–2016)," *Hydrol Process*, vol. 33, no. 13, pp. 1802–1817, Jun. 2019, doi: 10.1002/HYP.13438.
- [61] R. S. Padrón, B. P. Wilcox, P. Crespo, and R. Céleri, "Rainfall in the Andean Páramo: New Insights from High-Resolution Monitoring in Southern Ecuador," *J Hydrometeorol*, vol. 16, no. 3, pp. 985–996, 2015, doi: 10.1175/JHM-D-14-0135.1.

- [62] M. C. Peel, B. L. Finlayson, and T. A. McMahon, "Updated world map of the Köppen-Geiger climate classification," *Hydrol Earth Syst Sci*, vol. 11, no. 5, pp. 1633–1644, 2007, doi: 10.5194/HESS-11-1633-2007.
- [63] M. M. Hasan and G. Wyseure, "Impact of climate change on hydropower generation in Rio Jubones Basin, Ecuador," *Water Science and Engineering*, vol. 11, no. 2, pp. 157–166, 2018.
- [64] Y. Hong, K.-L. Hsu, S. Sorooshian, and X. Gao, "Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system," *Journal of Applied Meteorology*, vol. 43, no. 12, pp. 1834–1853, 2004.
- [65] J. F. Adamowski, "Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis," *J Hydrol (Amst)*, vol. 353, no. 3–4, pp. 247–266, 2008.
- [66] I. Aichouri, A. Hani, N. Bougherira, L. Djabri, H. Chaffai, and S. Lallahem, "River flow model using artificial neural networks," *Energy Procedia*, vol. 74, pp. 1007–1014, 2015.
- [67] G. Furquim *et al.*, "Combining wireless sensor networks and machine learning for flash flood nowcasting," in *2014 28th International Conference on Advanced Information Networking and Applications Workshops*, 2014, pp. 67–72.
- [68] K. Khosravi, H. Shahabi, B. Thai, J. Adamowski, and A. Shirzadi, "A comparative assessment of flood susceptibility modeling using Multi- Criteria Decision-Making Analysis and Machine Learning Methods," *J Hydrol (Amst)*, vol. 573, no. March, pp. 311–323, 2019, doi: 10.1016/j.jhydrol.2019.03.073.
- [69] D. P. Solomatine and Y. Xue, "M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China," *J Hydrol Eng*, vol. 9, no. 6, pp. 491–501, 2004.
- [70] M. Toukourou, A. Johannet, G. Dreyfus, and P.-A. Ayrat, "Rainfall-runoff modeling of flash floods in the absence of rainfall forecasts: the case of 'Cévenol flash floods,'" *Applied Intelligence*, vol. 35, no. 2, pp. 178–189, 2011.
- [71] S. Chen, Z. Xue, and M. Li, "Variable Sets principle and method for flood classification," *Sci China Technol Sci*, vol. 56, no. 9, pp. 2343–2348, 2013.
- [72] D. F. Munoz, P. Munoz, A. Alipour, H. Moftakhari, H. Moradkhani, and B. Mortazavi, "Fusing Multisource Data to Estimate the Effects of Urbanization, Sea Level Rise, and Hurricane Impacts on Long-Term Wetland Change Dynamics," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 14, pp. 1768–1782, 2021, doi: 10.1109/JSTARS.2020.3048724.
- [73] D. F. Muñoz, P. Muñoz, H. Moftakhari, and H. Moradkhani, "From local to regional compound flood mapping with deep learning and data fusion techniques," *Science of The Total Environment*, vol. 782, p. 146927, Aug. 2021, doi: 10.1016/J.SCITOTENV.2021.146927.

- [74] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
- [75] H. Zhang, "The optimality of naive Bayes," *AA*, vol. 1, no. 2, p. 3, 2004.
- [76] C. J. Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, "Classification And Regression Trees (1st ed.). Routledge.," p. 368, 1984.
- [77] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.
- [78] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," *Classification and Regression Trees*, pp. 1–358, Jan. 2017, doi: 10.1201/9781315139470/CLASSIFICATION-REGRESSION-TREES-LEO-BREIMAN-JEROME-FRIEDMAN-RICHARD-OLSHEN-CHARLES-STONE.
- [79] P. Contreras, J. Orellana-Alvear, P. Muñoz, J. Bendix, and R. Céleri, "Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment," *Atmosphere (Basel)*, vol. 12, no. 2, p. 238, 2021.
- [80] H. R. Maier and G. C. Dandy, "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications," *Environmental modelling & software*, vol. 15, no. 1, pp. 101–124, 2000.
- [81] K. Haddad, A. Rahman, M. A Zaman, and S. Shrestha, "Applicability of Monte Carlo cross validation technique for model development and validation using generalised least squares regression," *J Hydrol (Amst)*, vol. 482, pp. 119–128, Mar. 2013, doi: 10.1016/J.JHYDROL.2012.12.041.
- [82] K. P. Sudheer, A. K. Gosain, and K. S. Ramasastri, "A data-driven algorithm for constructing artificial neural network rainfall-runoff models," *Hydrol Process*, 2002, doi: 10.1002/hyp.554.
- [83] Y. Tang, P. Reed, K. van Werkhoven, and T. Wagener, "Advancing the identification and evaluation of distributed rainfall‐runoff models using global sensitivity analysis," vol. 43, pp. 1–14, 2007, doi: 10.1029/2006WR005813.
- [84] P. Cortez, "Sensitivity Analysis for Time Lag Selection to Forecast Seasonal Time Series using Neural Networks and Support Vector Machines," *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1–8, 2010, doi: 10.1109/IJCNN.2010.5596890.
- [85] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *University of California, Berkeley*, vol. 110, pp. 1–12, 2004.
- [86] J. Akosa, "Predictive accuracy: A misleading performance measure for highly imbalanced data," in *Proceedings of the SAS Global Forum, 2017*.
- [87] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," *Communications in Computer and Information Science*, vol. 51, pp. 461–471, 2009, doi: 10.1007/978-3-642-04962-0_53.

- [88] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015.
- [89] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach Learn*, vol. 85, no. 3, p. 333, 2011.
- [90] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Intern J Pattern Recognit Artif Intell*, vol. 23, no. 04, pp. 687–719, 2009.
- [91] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *arXiv preprint arXiv:1811.12808*, 2018.
- [92] D. N. Moriasi, J. G. Arnold, M. W. van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations," *Trans ASABE*, vol. 50, no. 3, pp. 885–900, 2007.
- [93] J. E. Nash and J. v Sutcliffe, "River flow forecasting through conceptual models part I — A discussion of principles," *J Hydrol (Amst)*, vol. 10, no. 3, pp. 282–290, 1970, doi: [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6).
- [94] H. v Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling," *J Hydrol (Amst)*, vol. 377, no. 1–2, pp. 80–91, 2009.
- [95] P. Willems, "A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models," *Environmental Modelling & Software*, vol. 24, no. 3, pp. 311–321, 2009.
- [96] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.
- [97] Z. Wang, C. Lai, X. Chen, B. Yang, S. Zhao, and X. Bai, "Flood hazard risk assessment model based on random forest," *J Hydrol (Amst)*, vol. 527, pp. 1130–1141, 2015, doi: [10.1016/j.jhydrol.2015.06.008](https://doi.org/10.1016/j.jhydrol.2015.06.008).
- [98] H. Tyralis, G. Papacharalampous, and A. Langousis, "A brief review of random forests for water scientists and practitioners and their recent history in water resources," *Water (Switzerland)*. 2019. doi: [10.3390/w11050910](https://doi.org/10.3390/w11050910).
- [99] S. Galelli and A. Castelletti, "Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling," *Hydrol Earth Syst Sci*, vol. 17, no. 7, pp. 2669–2684, 2013.
- [100] G. A. Papacharalampous and H. Tyralis, "Evaluation of random forests and Prophet for daily streamflow forecasting.," *Advances in Geosciences*, vol. 45, 2018.
- [101] B. Li, G. Yang, R. Wan, X. Dai, and Y. Zhang, "Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake

- in China,” *Hydrology Research*, vol. 47, no. S1, pp. 69–83, 2016, doi: 10.2166/nh.2016.264.
- [102] Z. Abda, B. Zerouali, M. Chettih, C. A. Guimarães Santos, C. A. S. de Farias, and A. Elbeltagi, “Assessing machine learning models for streamflow estimation: A case study in Oued Sebaou watershed (Northern Algeria),” *Hydrological Sciences Journal*, pp. 1–14, 2022.
- [103] M. Borga, E. N. Anagnostou, G. Blöschl, and J. D. Creutin, “Flash flood forecasting, warning and risk management: The HYDRATE project,” *Environ Sci Policy*, vol. 14, no. 7, pp. 834–844, 2011, doi: 10.1016/j.envsci.2011.05.017.
- [104] N. Boers, B. Bookhagen, H. M. J. Barbosa, N. Marwan, J. Kurths, and J. A. Marengo, “Prediction of extreme floods in the eastern Central Andes based on a complex networks approach,” *Nat Commun*, vol. 5, no. 1, pp. 1–7, 2014.
- [105] C. Aybar *et al.*, “Uso del Producto Grillado ‘PISCO’ de precipitación en Estudios, Investigaciones y Sistemas Operacionales de Monitoreo y Pronóstico Hidrometeorológico,” *Nota Técnica*, vol. 1, 2017.
- [106] C. Fernández de Córdova Webster and Y. Javier Rodríguez López, “Primeros resultados de la red actual de monitoreo hidrometeorológico de Cuenca, Ecuador,” *Ingeniería Hidráulica y Ambiental*, vol. 37, no. 2, pp. 44–56, 2016.
- [107] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, 2011.
- [108] I. K. de Almeida, A. K. Almeida, J. A. A. Anache, J. L. Steffen, and T. Alves Sobrinho, “Estimation on time of concentration of overland flow in watersheds: A review,” *Geociencias*, vol. 33, no. 4, 2014.
- [109] Y. Li, S. Grimaldi, J. P. Walker, and V. Pauwels, “Application of remote sensing data to constrain operational rainfall-driven flood forecasting: a review,” *Remote Sens (Basel)*, vol. 8, no. 6, p. 456, 2016.
- [110] C. Loumagne *et al.*, “Integration of remote sensing data into hydrological models for reservoir management,” *Hydrological sciences journal*, vol. 46, no. 1, pp. 89–102, 2001.
- [111] P. Probst, M. Wright, and A.-L. Boulesteix, “Hyperparameters and Tuning Strategies for Random Forest,” *ArXiv e-prints*, 2018.
- [112] K. Beven, “Deep learning, hydrological processes and the uniqueness of place,” *Hydrol Process*, vol. 34, no. 16, pp. 3608–3613, Jul. 2020, doi: 10.1002/HYP.13805.
- [113] H. M. V. V. Herath, J. Chadalawada, and V. Babovic, “Hydrologically informed machine learning for rainfall-runoff modelling: Towards distributed modelling,” *Hydrol Earth Syst Sci*, vol. 25, no. 8, pp. 4373–4401, Aug. 2021, doi: 10.5194/HESS-25-4373-2021.

- [114] V. Moreido, B. Gartsman, D. P. Solomatine, and Z. Suchilina, "How Well Can Machine Learning Models Perform without Hydrologists? Application of Rational Feature Selection to Improve Hydrological Forecasting," *Water (Basel)*, vol. 13, no. 12, p. 1696, 2021.
- [115] G. S. Nearing *et al.*, "What Role Does Hydrological Science Play in the Age of Machine Learning?," *Water Resour Res*, vol. 57, no. 3, Mar. 2021, doi: 10.1029/2020WR028091.
- [116] D. P. Solomatine and A. Ostfeld, "Data-driven modelling: some past experiences and new approaches," *Journal of Hydroinformatics*, vol. 10, no. 1, pp. 3–22, Jan. 2008, doi: 10.2166/HYDRO.2008.015.
- [117] G. Corzo and D. Solomatine, "Baseflow separation techniques for modular artificial neural network modelling in flow forecasting," *Hydrological Sciences Journal*, vol. 52, no. 3, pp. 491–507, 2007.
- [118] P. C. Huang and K. T. Lee, "Influence of topographic features and stream network structure on the spatial distribution of hydrological response," *J Hydrol (Amst)*, vol. 603, p. 126856, Dec. 2021, doi: 10.1016/J.JHYDROL.2021.126856.
- [119] M. K. Akhtar, G. A. Corzo, S. J. van Andel, and A. Jonoski, "River flow forecasting with artificial neural networks using satellite observed precipitation pre-processed with flow length and travel time information: case study of the Ganges river basin," *Hydrol Earth Syst Sci*, vol. 13, no. 9, pp. 1607–1618, 2009.
- [120] N. Wang, D. Zhang, H. Chang, and H. Li, "Deep Learning of Subsurface Flow via Theory-guided Neural Network," *J Hydrol (Amst)*, vol. 584, Oct. 2019, doi: 10.1016/j.jhydrol.2020.124700.
- [121] Y. Hong *et al.*, "Remote sensing precipitation: Sensors, retrievals, validations, and applications," *Observation and Measurement; Li, X., Vereecken, H., Eds*, pp. 1–23, 2019.
- [122] G. Huffman, D. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, and P. Xie, "Integrated multi-satellite retrievals for GPM (IMERG), version 4.4. NASA's Precipitation Processing Center." 2014.
- [123] M. Laverde-Barajas, G. Corzo, B. Bhattacharya, R. Uijlenhoet, and D. P. Solomatine, "Spatiotemporal analysis of extreme rainfall events using an object-based approach," in *Spatiotemporal Analysis of Extreme Hydrological Events*, Elsevier, 2019, pp. 95–112.
- [124] J. Li, K.-L. Hsu, A. AghaKouchak, and S. Sorooshian, "Object-based assessment of satellite precipitation products," *Remote Sens (Basel)*, vol. 8, no. 7, p. 547, 2016.
- [125] M. Cisty and V. Soldanova, "Flow prediction versus flow simulation using machine learning algorithms," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10935 LNAI, pp. 369–382, 2018, doi: 10.1007/978-3-319-96133-0_28.
- [126] M. Laverde-Barajas, G. A. C. Perez, F. Chishtie, A. Poortinga, R. Uijlenhoet, and D. P. Solomatine, "Decomposing satellite-based rainfall errors in flood estimation: Hydrological

- responses using a spatiotemporal object-based verification method,” *J Hydrol (Amst)*, vol. 591, p. 125554, 2020.
- [127] Y. Yang and T. Chui, “Uncertainties of Machine Learning in Predicting the Hydrological Responses of LID Practices,” in *Proceedings of the 22nd IAHR-APD Congress*, Sapporo, Japan, 2020. Accessed: Apr. 16, 2022. [Online]. Available: <https://www.iahr.org/library/infor?pid=7598>
- [128] C. Davis, B. Brown, and R. Bullock, “Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas,” *Mon Weather Rev*, vol. 134, no. 7, pp. 1772–1784, 2006.
- [129] J. M. Peña-Barragán, M. K. Ngugi, R. E. Plant, and J. Six, “Object-based crop identification using multiple vegetation indices, textural features and crop phenology,” *Remote Sens Environ*, vol. 115, no. 6, pp. 1301–1316, 2011.
- [130] K. Y. Gutiérrez-Jurado, D. Partington, O. Batelaan, P. Cook, and M. Shanafield, “What triggers streamflow for intermittent rivers and ephemeral streams in low-gradient catchments in Mediterranean climates,” *Water Resour Res*, vol. 55, no. 11, pp. 9926–9946, 2019.
- [131] R. Grayson and G. Blöschl, *Spatial patterns in catchment hydrology: observations and modelling*. CUP Archive, 2001.
- [132] M. F. A. Vogels, S. M. de Jong, G. Sterk, N. Wanders, M. F. P. Bierkens, and E. A. Addink, “An object-based image analysis approach to assess irrigation-water consumption from MODIS products in Ethiopia,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 88, p. 102067, 2020.
- [133] H. Shi, T. Li, J. Wei, W. Fu, and G. Wang, “Spatial and temporal characteristics of precipitation over the Three-River Headwaters region during 1961–2014,” *J Hydrol Reg Stud*, vol. 6, pp. 52–65, 2016, doi: 10.1016/j.ejrh.2016.03.001.
- [134] S. der Walt *et al.*, “scikit-image: image processing in Python,” *PeerJ*, vol. 2, p. e453, 2014.
- [135] I. K. de ALMEIDA, A. K. ALMEIDA, J. A. A. ANACHE, J. L. STEFFEN, and T. A. SOBRINHO, “Estimation on time of concentration of overland flow in watersheds: a review,” *Geociências (São Paulo)*, vol. 33, no. 4, pp. 661–671, 2014.
- [136] C. Massari, S. Camici, L. Ciabatta, and L. Brocca, “Exploiting satellite-based surface soil moisture for flood forecasting in the Mediterranean area: State update versus rainfall correction,” *Remote Sens (Basel)*, vol. 10, no. 2, p. 292, 2018.
- [137] J. Singh, H. v Knapp, and M. Demissie, “Hydrologic modeling of the Iroquois River watershed using HSPF and SWAT. ISWS CR 2004-08,” *Champaign, Ill.: Illinois State Water Survey*, 2004.
- [138] S. Chen, L. Xiong, Q. Ma, J. Kim, J. Chen, and C. Xu, “Improving daily spatial precipitation estimates by merging gauge observation with multiple satellite-based precipitation

- products based on the geographically weighted ridge regression method,” *J Hydrol (Amst)*, vol. 589, no. June, p. 125156, 2020, doi: 10.1016/j.jhydrol.2020.125156.
- [139] L. Xu, N. Chen, H. Moradkhani, X. Zhang, and C. Hu, “Improving Global Monthly and Daily Precipitation Estimation by Fusing Gauge Observations, Remote Sensing, and Reanalysis Data Sets,” *Water Resour Res*, vol. 56, no. 3, p. e2019WR026444, 2020.
- [140] P. Muñoz, G. Corzo, D. Solomatine, J. Feyen, and R. Céleri, “Near-real-time satellite precipitation data ingestion into peak runoff forecasting models,” *Environmental Modelling & Software*, vol. 160, p. 105582, Feb. 2023, doi: 10.1016/J.ENVSOFT.2022.105582.
- [141] I. K. de Almeida, A. K. Almeida, J. A. A. Anache, J. L. Steffen, and T. Alves Sobrinho, “Estimation on time of concentration of overland flow in watersheds: A review,” *Geociencias*, vol. 33, no. 4, pp. 661–671, 2014.
- [142] Y. Hong, D. Gochis, J. T. Cheng, K. L. Hsu, and S. Sorooshian, “Evaluation of PERSIANN-CCS Rainfall Measurement Using the NAME Event Rain Gauge Network,” *J Hydrometeorol*, vol. 8, no. 3, pp. 469–482, Jun. 2007, doi: 10.1175/JHM574.1.
- [143] M. N. Anjum *et al.*, “Assessment of PERSIANN-CCS, PERSIANN-CDR, SM2RAIN-ASCAT, and CHIRPS-2.0 Rainfall Products over a Semi-Arid Subtropical Climatic Region,” *Water 2022*, Vol. 14, Page 147, vol. 14, no. 2, p. 147, Jan. 2022, doi: 10.3390/W14020147.
- [144] H. Salehi, M. Sadeghi, S. Golian, P. Nguyen, C. Murphy, and S. Sorooshian, “The Application of PERSIANN Family Datasets for Hydrological Modeling,” *Remote Sensing 2022*, Vol. 14, Page 3675, vol. 14, no. 15, p. 3675, Jul. 2022, doi: 10.3390/RS14153675.
- [145] G. Moura Ramos Filho, V. Hugo Rabelo Coelho, E. da Silva Freitas, Y. Xuan, L. Brocca, and C. das Neves Almeida, “Regional-scale evaluation of 14 satellite-based precipitation products in characterising extreme events and delineating rainfall thresholds for flood hazards,” *Atmos Res*, vol. 276, p. 106259, Oct. 2022, doi: 10.1016/J.ATMOSRES.2022.106259.
- [146] M. R. Eini, A. Rahmati, and M. Piniewski, “Hydrological application and accuracy evaluation of PERSIANN satellite-based precipitation estimates over a humid continental climate catchment,” *J Hydrol Reg Stud*, vol. 41, Jun. 2022, doi: 10.1016/J.EJRH.2022.101109.
- [147] J. Li *et al.*, “Predicting floods in a large karst river basin by coupling PERSIANN-CCS QPEs with a physically based distributed hydrological model,” *Hydrol Earth Syst Sci*, vol. 23, no. 3, pp. 1505–1532, Mar. 2019, doi: 10.5194/HESS-23-1505-2019.
- [148] M. Sadeghi *et al.*, “PERSIANN-CNN: Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks–Convolutional Neural Networks,” *J Hydrometeorol*, vol. 20, no. 12, pp. 2273–2289, Dec. 2019, doi: 10.1175/JHM-D-19-0110.1.
- [149] D. A. Zeweldi and M. Gebremichael, “Sub-daily scale validation of satellite-based high-resolution rainfall products,” *Atmos Res*, vol. 92, no. 4, pp. 427–433, Jun. 2009, doi: 10.1016/J.ATMOSRES.2009.01.001.

- [150] M. B. Gunathilake, T. Senerath, U. Rathnayake, M. B. Gunathilake, T. Senerath, and U. Rathnayake, "Artificial neural network based PERSIANN data sets in evaluation of hydrologic utility of precipitation estimations in a tropical watershed of Sri Lanka," *AIMS Geosciences* 2021 3:478, vol. 7, no. 3, pp. 478–489, 2021, doi: 10.3934/GEOSCI.2021027.
- [151] "Physically Based Mountain Hydrological Modeling Using Reanalysis Data in Patagonia on JSTOR." <https://www.jstor.org/stable/24914930> (accessed May 03, 2023).
- [152] G. He, Y. Chen, G. Fang, and Z. Li, "Hydrometeorological Forecast of a Typical Watershed in an Arid Area Using Ensemble Kalman Filter," *Water* 2022, Vol. 14, Page 3970, vol. 14, no. 23, p. 3970, Dec. 2022, doi: 10.3390/W14233970.
- [153] S. Terzago, G. Bongiovanni, and J. Von Hardenberg, "Seasonal forecasting of snow resources at Alpine sites," *Hydrol Earth Syst Sci*, vol. 27, no. 2, pp. 519–542, Jan. 2023, doi: 10.5194/HESS-27-519-2023.
- [154] D. I. V. Domeisen *et al.*, "Advances in the Subseasonal Prediction of Extreme Events: Relevant Case Studies across the Globe," *Bull Am Meteorol Soc*, vol. 103, no. 6, pp. E1473–E1501, Jun. 2022, doi: 10.1175/BAMS-D-20-0221.1.
- [155] G. Pegram *et al.*, "Present and Future Requirements for Using and Communicating Hydro-Meteorological Ensemble Prediction Systems for Short-, Medium-, and Long-Term Applications," *Handbook of Hydrometeorological Ensemble Forecasting*, pp. 1–46, 2015, doi: 10.1007/978-3-642-40457-3_39-1.
- [156] T. Condom *et al.*, "Climatological and Hydrological Observations for the South American Andes: In situ Stations, Satellite, and Reanalysis Data Sets," *Front Earth Sci (Lausanne)*, vol. 8, p. 92, Apr. 2020, doi: 10.3389/FEART.2020.00092/BIBTEX.