Applied Computing and Geosciences 23 (2024) 100181

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences

# Modeling river flow for flood forecasting: A case study on the Ter river

Fabián Serrano-López [a], Sergi Ger-Roca [a], Maria Salamó [a], Jerónimo Hernández-González [b,*]

[a] Departament de Matemàtiques i Informàtica, Universitat de Barcelona (UB), Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain
[b] Departament d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona (UdG), Campus Montilivi, 17003, Girona, Spain

ARTICLE INFO

Dataset link: https://jhernandezgonzalez.github.io/supp_ter.html

*Keywords:*
Real-time flood forecasting
Spatio-temporal calibration
Machine learning
Ter river

ABSTRACT

Floods affect chronically many communities around the world. Their socioeconomic impact increases year-by-year, boosted by global warming and climate change. Combined with long-term preemptive measures, preparatory actions are crucial when floods are imminent. In the last decade, machine learning models have been used to anticipate these hazards. In this work, we model the Ter river (NE Spain), which has historically suffered from floods, using traditional ML models such as K-nearest neighbors, Random forests, XGBoost and Linear regressors. Publicly available river flow and precipitation data was collected from year 2009 to 2021. Our analysis measures the time elapsed between observing a flow rise event at different stations (or heavy rain, for rainfall stations), and use the measured time lags to align the data from the different stations. This information provides increased interpretability to our river flow models and flood forecasters. A thorough evaluation reveals that ML techniques can be used to make short-term predictions of the river flow, even during heavy rain and large flow rise events. Moreover, our flood forecasting system provides valuable interpretable information for setting up timely preparatory actions.

## 1. Introduction

Faced with catastrophic events such as floods, droughts or violent storms, responsible administrations need to plan prevention and rapid-response actions to mitigate the impact of these events. Floods, which chronically affect many riversides and coastal areas every year, are among the most devastating natural hazards; the World Meteorological Organization (2011) estimates that one third of all losses due to natural phenomena can be attributed to floods. Their impact involves damage or destruction of crops, infrastructure and buildings along riverbanks, but also can cause personal harm and, in the most severe cases, the loss of human lives. In Europe, floods account for 43% of the economic losses caused by climate-related extremes (European Environment Agency, 2023). In Spain, and particularly in Catalonia, floods are the main cause of loss of human lives and property damage due to natural disasters (Barnolas and Llasat, 2007).

Climate change, natural or human-induced, can affect the frequency or strength of extreme weather events such as extreme precipitation (Cubasch et al., 2013). Floods follow extreme precipitation events, and the frequency of these is strongly correlated with temperature increase (Drobinski et al., 2018). Recent studies by Mallakpour and Villarini (2015) or by Blöschl et al. (2020) evidence that the frequency of flooding has, in general, increased over the last decades following the global warming, even though in specific areas the trend might be the opposite. Direct human intervention through land use changes has been associated with an increase in frequency and severity of floods (Apollonio et al., 2016; Hounkpè et al., 2019). Being able to anticipate them is crucial for planning mitigation actions and diminishing their potentially severe consequences.

Analyzing meteorological hazards and using predictive models to implement preventive measures is paramount in the research community, addressing, e.g., wildfire detection (Sayad et al., 2019), snowfalls prediction (Kulie et al., 2020; Panegrossi et al., 2022), drought detection (Chang et al., 2023), or river flow modeling (Norsyuhada et al., 2022). We are interested in river flow modeling to enable the prediction of floods using machine learning (ML). Previous works on river flow analysis range from flow forecasting (Dibike and Solomatine, 2001; Saint-Fleur et al., 2023) to flood prediction or detection (Ha et al., 2021; Norsyuhada et al., 2022; Chebii et al., 2022), most of them based on traditional statistical methods, including multivariate time series analysis tools. In the related literature, hydrological forecasting has been addressed with ML models like support vector machines (Bürger et al., 2007; Hamitouche and Ribalta, 2023; Kumar et al., 2021), tree-based ensembles (Chakraborty et al., 2021; Hamitouche and Ribalta, 2023; Chang et al., 2024) or artificial neural networks (ANNs) (Bafitlhile and Li, 2019; Jimeno-Sáez et al., 2018; Tayfur et al., 2018; Chebii et al., 2022). ANNs are a popular choice since seminal works like those by Hsu et al. (1995), modeling the rainfall-runoff relationship in

* Corresponding author.
*E-mail address:* jeronimo.hernandez@udg.edu (J. Hernández-González).

https://doi.org/10.1016/j.acags.2024.100181
Received 21 March 2024; Received in revised form 15 July 2024; Accepted 22 July 2024
Available online 24 July 2024

the Leaf river watershed (USA), or by Dibike and Solomatine (2001), forecasting the streamflow in the Apure river watershed (Venezuela). Modern *deep* neural networks (DNN) have also been used for hydrological modeling and flood forecasting (Gao et al., 2020; Lin et al., 2020; Kratzert et al., 2019; Chang et al., 2024; Bhasme and Bhatia, 2024). Xu et al. (2021) used temporal DNN to model the rainfall-runoff relationship on two Chinese rivers, and Ghimire et al. (2021) integrated two DNN structures to make short-term hourly flow predictions on two Australian rivers. Recently, Jiang et al. (2024) used dynamic temporal graph convolutional networks for flood forecasting, capturing dynamic spatiotemporal features of flood data. These ML models have traditionally used flow, rainfall and other meteorological data (Bürger et al., 2007; Zhang et al., 2021; Jimeno-Sáez et al., 2018; Saint-Fleur et al., 2023; Hamitouche and Ribalta, 2023; Xu et al., 2021; Ghimire et al., 2021), and more recently remote sensing/satellite data (Kumar et al., 2021; Liu et al., 2023; Kratzert et al., 2019; Jiang et al., 2024). Although ML has had a transformative impact on different scientific areas, mainly after the emergence of deep learning (DL) methods, this has not yet taken place in hydrology (Nearing et al., 2021). The low interpretability of this type of model has slowed down the adoption of DL techniques in hydrology. However, several studies have recently shed light on this issue. Kratzert et al. (2019) designed a general physics-informed model that allows them to improve flow predictions in a vast amount of river basins and find a hidden representation which, compared to observable catchment characteristics, reveals that vegetation type and seasonality are relevant factors. Similarly, Bhasme and Bhatia (2024) and Saint-Fleur et al. (2023) show that physics-informed ML can be applied to enhance and make more interpretable results in both general and specific contexts. In this direction, theory-guided machine learning aims to produce scientifically interpretable models, simplifying model search and enhancing model generalizability due to their grounding in scientific knowledge (Karpatne et al., 2017). The use of attention mechanisms in the DL models, which can be easily visualized, was exploited by Chang et al. (2024) to understand the relationship between flood sensors and spatiotemporal variables. Salience maps over convolutional DNNs have also been used to understand how sea temperature surface influences Amazon and Congo rivers flow in relation with the El Niño Southern Oscillation (Liu et al., 2023). On the contrary, most of the traditional ML models are directly explainable or can take advantage of off-the-self techniques like SHAP or LIME (Molnar, 2022) to produce interpretable yet competitive predictions on structured tabular hydroclimatic data (Chakraborty et al., 2021).

In this work, we focus on river flow modeling to enable the prediction of floods in the Ter river, which flows from the eastern Pyrenees into the Mediterranean Sea in NE Spain. It flows through the city of Girona and supplies water to the metropolitan area of Barcelona (about 5 million people). Floods have been documented in the Ter watershed since 1193 (Ribas Palom, 2007). Several meteorological studies have analyzed and categorized the historical floods of the river, identifying those that can be considered serious or catastrophic (Barnolas and Llasat, 2007; Llasat et al., 2005). Although other rivers in the west-Mediterranean region have been studied with ML techniques (Bürger et al., 2007; Hamitouche and Ribalta, 2023; Jimeno-Sáez et al., 2018; Saint-Fleur et al., 2023; Tayfur et al., 2018), this is, up to our knowledge, the first study on ML-based streamflow modeling of the Ter river. We consider standard ML models due to (i) their simplicity and efficiency, (ii) our use of structured tabular data, where they are competitive with DNNs (Chakraborty et al., 2021) and, more importantly, (iii) their straightforward interpretability, which fulfills our requirement of an explainable and actionable predictive system. Using meteorological and river flow data from 2009 to 2021, we calibrate our models with an estimation of the time elapsed between the observation of an event (heavy rainfall, flow rise) in different meteorological and river flow stations. This enhances the interpretability of our models in line with the current trend towards interpretable ML (Molnar, 2022, Ch. 3.1). We present a solution combining two models fitted to calm and flow rise periods.

The contributions of this work can be summarized as follows:

- A dataset collected with 12-year rainfall and river-flow data for the Ter basin, curated after imputing missing values and frequency homogenization, publicly available on the website associated with this study[1];
- The estimation of the time elapsed (lag) between the observation of the same event (heavy rainfall, flow rise) between meteorological and river flow stations;
- A set of interpretable regressors for river flow modeling at two points of the river able to provide short-term predictions during both flow rises and clam periods.

For the rest of the paper, we first describe the case study, the Ter river, the data sources and the preprocessing techniques. Section 3 presents the descriptive data analysis where the time elapsed between stations is estimated, followed by our river-flow predictive modeling and its validation. Finally, in Section 5, our findings are discussed and future research lines are drawn.

## 2. Case study: Ter river

In the NE Iberian peninsula (Catalonia, Spain), the Ter river is born at 2400 masl in an ancient glacial cirque of the Eastern Pyrenees and flows into the Mediterranean Sea. With a length of 208 km and a watershed area of 3010 $km^2$, it is the longest river in the Pyrenees-Mediterranean hydrographic area. The river suffers from numerous flow diversions: It greatly influences the local agricultural sector and its water is exported to other areas such as the Barcelona area. It holds large hydroelectric production facilities such as those in Sau (capacity: 151.3 $hm^3$), Susqueda (233 $hm^3$), and Pasteral (2 $hm^3$) reservoirs. These reservoirs, located approximately halfway through the course of the river, are used to regulate the flow downstream along the second half of the river. Thus, the flow at the beginning of the lower course is completely regulated. For this reason, we perform *separate studies* for the upper and the lower courses.

Ter is the Catalan river with the largest number of *historical floods* (121 events from year 1322 to 2000) (Llasat et al., 2005). Floods mainly occur in the late summer and fall months (Barnolas and Llasat, 2007). Catalan Water Agency (2019, Map 01H02D, Annex I) identified the highest flood risk areas.

Recently, several high-impact flood episodes have hit the Ter basin. For example, the *Gloria* storm brought rainfall ranging from 200 to 500 mm in three days of January 2020. This event filled the reservoirs in the middle course of the river, forcing river managers to open sluice gates. Thus, for several hours, the lower course of the river was unregulated. In consequence, the Ter river and its tributaries overflowed, causing significant floods that forced the evacuation or confinement of several populations. In October 2018, the *Leslie* storm, with more than 200 mm (20–30 mm in 30 min), caused a similar flood risk event. Although the Ter did not overflow, the warning height threshold was exceeded. Both events are recorded in our data and are used as key events for model evaluation.

For this study, river flow and rainfall data were collected from open data sources at the Catalan Water Agency[2] (ACA) and Catalan Meteorological Service[3] (SMC) websites, respectively. The supplementary material available on the website associated with this study[1] includes a spreadsheet and a map of the Ter watershed with the list of the river flow and rainfall stations considered and their location.

[1] https://jhernandezgonzalez.github.io/supp_ter.html

[2] https://aplicacions.aca.gencat.cat/sdim21/ (accessed July 15th, 2024).

[3] https://www.meteo.cat/wpweb/serveis/cataleg-de-serveis/serveis-oberts/dades-obertes/ (accessed July 15th, 2024).

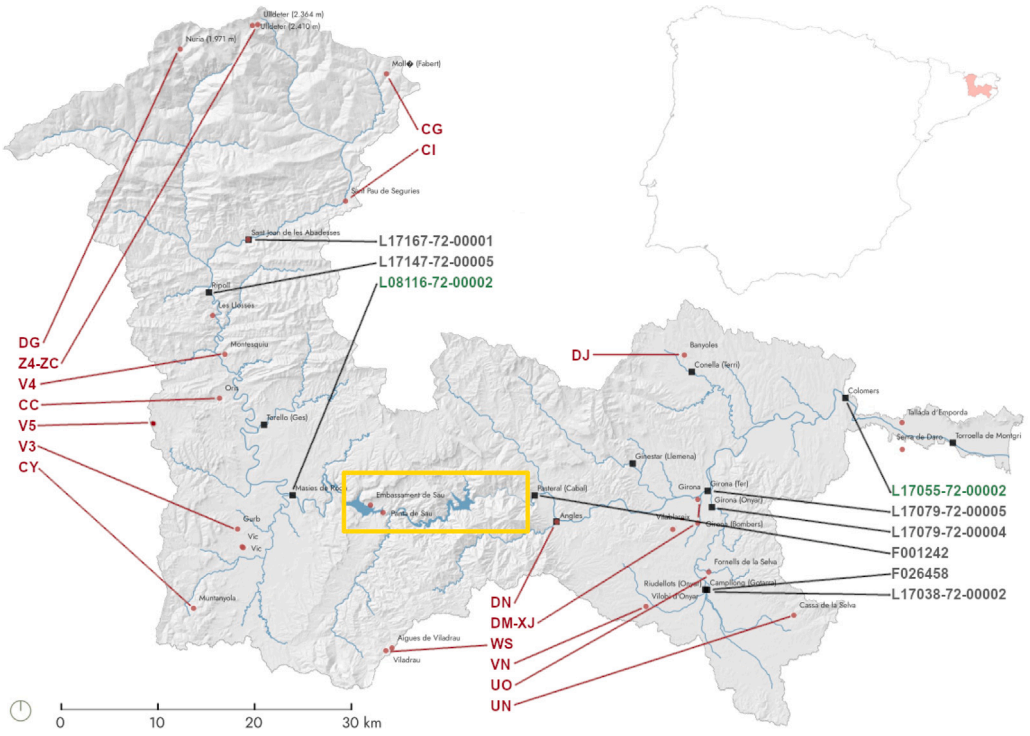


**Fig. 1.** Ter river watershed and the location of the stations considered in this study. Codes of river flow stations are given in gray, reference stations in green, and meteorological stations in red. Codes are not given for discarded stations. The yellow box covers the three reservoirs that divide the river into two differentiated courses.

### 2.1. River flow data

ACA's open data website[2] allows for collecting measurements of river flow and height in 14 different stations throughout the Ter river and its main tributaries. River flow is the quantity of water that runs through a section of the river in a period of time. Height measures the distance from the bottom of the river up to the water surface. These measurements are provided in cubic meters per second ($m^3/s$) and meters (m), respectively, every 5 min.

**Data pre-processing.** We first homogenized the frequency (30 min) of the time series using the mean value. Four stations initially with more than 50% of missing values (ranging from 57% to 95%) were eliminated from the study. Defective measurements were easily identified due to the generally stable flow-height relationship: too-large or too-small height values attached to regular flow values (or the other way around) were discarded. Moreover, we applied a moving average individually to each river flow and height time series to detect and *remove outliers*. With a window size of 25, we removed all the detected outliers: values above/below the mean plus/minus two times the standard deviation. After this step, the percentage of missing values ranged from 5% to 60% among the river flow time series and from 3% to 62% among the height time series.

To fill *missing values*, we took advantage again of the flow-height relationship. For each station, we fitted to this relationship a $K$-nearest neighbor ($K$NN) regressor (Cover and Hart, 1967) with $K = 7$. When a river height value was available and the corresponding flow measurement missing (and the other way around), the model estimation was used to impute it. We also used simple interpolation to fill in small gaps (up to 30 consecutive missing points) in both river flow and height series taking advantage of their usual smoothness. After these imputation steps, the percentage of missing values ranged from 0.1% to 49% among the river flow time series. Finally, combining the flow time series of several stations, we applied multivariate $K$NN-based imputation assuming that the river behavior is stable across time. We performed it twice, separately for the upper course series and for those of the lower course. After this, all the river flow time series are complete.

Two stations are used as reference points to model the river flow, one from each part of the river: *Masies de Roda* (Upper course, *L08116-72-00002*) and *Colomers* (Lower course, *L17055-72-00002*). Our decision was motivated by their location (both are near the final section of the upper and lower courses) and their proximity to potentially flood-affected areas according to the Catalan Water Agency (2019). *Pasteral* (*F001242*) is a key river flow station in the lower course of the river because, due to its location at the last reservoir's gates, it contains information on the impact of reservoir management on downstream flow. One station was discarded as it is physically located downstream from the reference station of the lower course. As a result, the final number of river flow stations is 9.

### 2.2. Rainfall data

SMC open data website[3] allows for collecting measurements of different meteorological features throughout the whole Catalan territory. We are interested in rainfall measurements from stations located inside the Ter watershed: we found 21 well-functioning stations in the period of interest (2009–2021). Rainfall is the amount of rain that falls on an area in a particular period of time. Rainfall measurements are provided in millimeters (mm) and their frequency varies (half-hourly or hourly).

**Data pre-processing.** We first homogenized the frequency (30 min) of the time series. The collected time series had, in part, hourly rainfall measurements. To simulate half-hourly measurements in these parts, we made the simplifying assumption of a constant rainfall rate throughout the hour. Thus, from each point in the time series, two new points were created with half the measured rainfall each.

Discontinued rainfall stations were disregarded, with only two exceptions: DM and Z4. Right after disassembling them, two new stations were operational in their vicinity (XJ and ZC, resp.). Each pair of time series was combined into a single one (DM-XJ and Z4-ZC). The time gap between the original series (merely 29 and 405 points, resp., 0.6

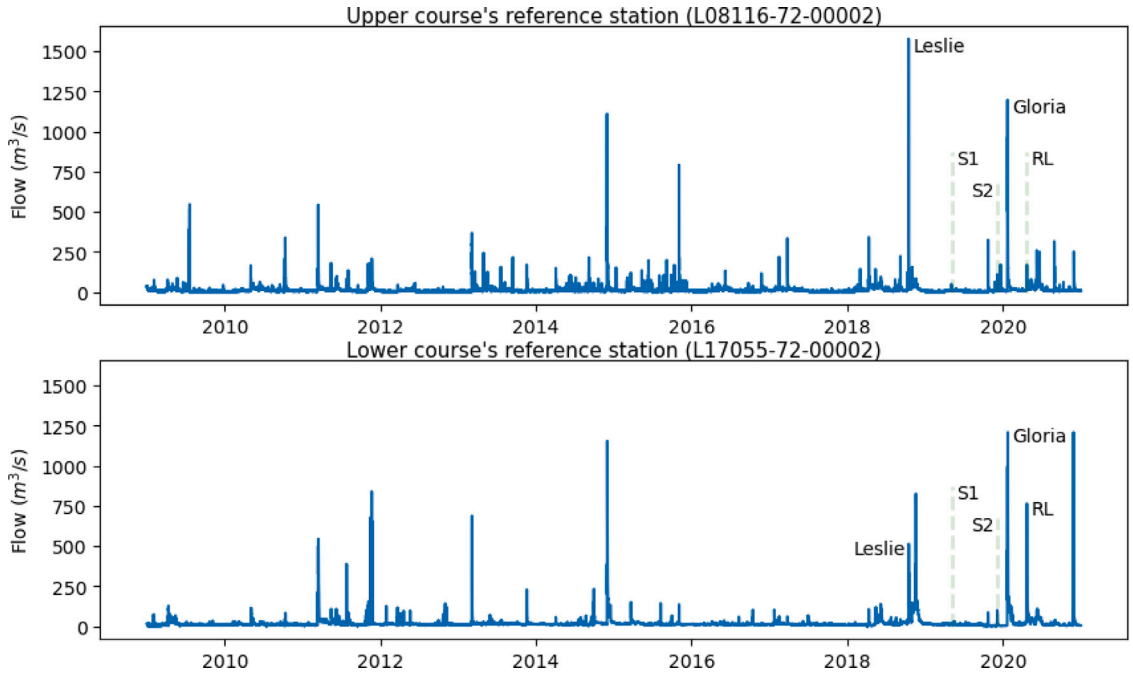


**Fig. 2.** Time series of the Ter river flow at the upper and lower reference stations.

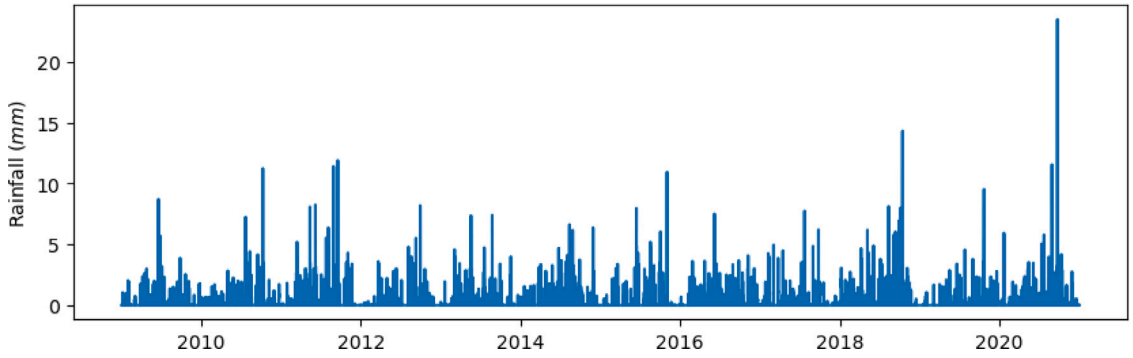


**Fig. 3.** Illustrative example of a rainfall time series (CI station).

and 8.4 days) was filled with data from the closest stations (WF and DG, resp., although only DG measured some rain in 11 out of the 405 points). No treatment of missing values or outliers was required. Two other rainfall stations were disregarded as they are physically located downstream from the reference stations. All in all, the number of rainfall stations is finally 16.

## 3. Descriptive data analysis

The river flow in Ter follows a relatively constant pattern (see Fig. 2), with a mean streamflow of 14.78 $m^3/s$ in the upper course and 19.53 $m^3/s$ in the lower course. The median values are, respectively, 9.98 and 13.34 $m^3/s$. The river flow increases considerably at specific time points usually due to large rainfall events. The number of these events changes yearly, but as a general pattern, they are more common during the fall and winter months. Counterintuitively, the flow peaks are consistently smaller in the lower course of the river. This can be explained by the presence of the reservoirs that split the upper and lower courses into two streams with different characteristics. These reservoirs can catch most of the water from flow rise events in the upper course, as they are the origin of large flow diversions. Only on rare and dangerous occasions when the reservoirs are full and sluice gates need to be opened, like during the *Gloria* event, the flow rises in the upper and lower courses with similar magnitudes.

The rainfall time series (see Fig. 3) shows a general periodic pattern: the first months of the year are drier, and the rest of them have similar rainfall, with a clear peak in the late-summer/fall months, when heavy rain events traditionally happen in this area.

Given the rather flat or periodic characteristics of these time series, focusing on heavy-rain or river flow rise events can help us understand the dynamics of the river. In this study, we are interested in measuring the lag or time elapsed between the signal of these events registered on each of the upstream rain and river-flow stations and the corresponding signal registered at the reference station. We can take advantage of this information to calibrate or align the time series of the different stations.

The source code of the analysis presented hereunder is available on the website associated with this study.[1] Written in Python (v3.7.13) and run in Jupyter notebooks (*notebook*, v6.4.12), it uses standard ML libraries: *numpy* (v1.21.5), *pandas* (v1.3.5), *matplotlib* (v3.5.2), *scipy* (v1.7.3), *scikit-learn* (v1.0.2), *smogn* (v0.1.2), *xgboost* (v1.6.2). The default parametrization of all methods is used unless otherwise explained.

### 3.1. Finding rare events

We are interested in finding events that stand out due to their high streamflow automatically. In this work, for the sake of simplicity, a

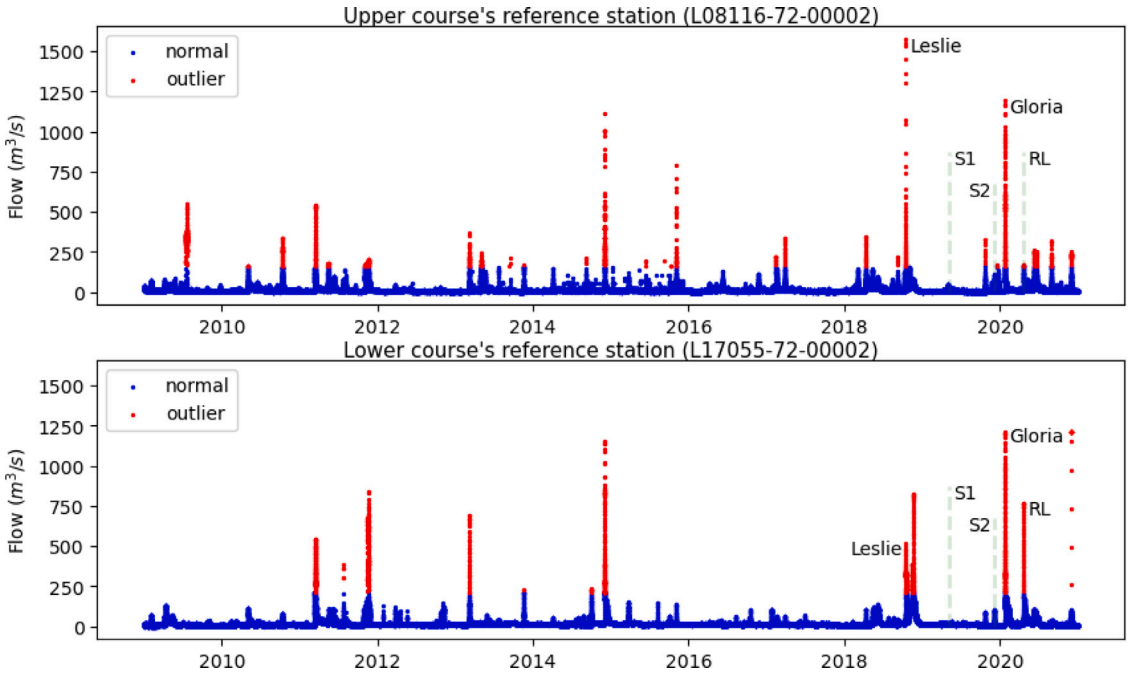


**Fig. 4.** Rare events found on the time series of both reference stations.

measurement above the 99.5th percentile is considered to stand out. Consequently, a rare event is a (short) period including one or more outstanding measurements.

Fig. 4 shows the rare events found in the time series of the two reference stations. 32 and 14 such events are found in the upper and lower course, respectively. The smaller number of rare events in the lower course is in line with the above description of the river as a two-course river. The initial flow in this lower course is regulated. Among the events detected as anomalies are the storms *Gloria* and *Leslie* (see Section 2).

### 3.2. Lag estimation

One of the main objectives of the descriptive analysis is to understand and infer from data the time elapsed between the observation of the same event at two different stations. The aim is to estimate this time for every station, both rainfall and river flow, with respect to the corresponding reference station. Remember that we have identified two stations, one from the upper and another one from the lower course, as references.

The signal we aim to find might be too weak in most of the analyzed period as the river flow is rather constant (see Fig. 2) and heavy rain events are scarce (see Fig. 3). Therefore, we focus on the rare events found in the previous section. We assume that it is possible to estimate from data the time elapsed between a streamflow rise at two different river stations. Similarly, we expect to understand how a heavy rain event is later observed in the river as a streamflow raise and how long this process takes.

For each pair of time series to be aligned (any station with its reference station in the corresponding part of the river), our process uses Pearson's cross-correlation to find a set of relevant time lapses/lags. Algorithm 1 shows its pseudocode. Given a rare event, the origin, `ts_orig`, and reference, `ts_ref`, time series are cross-correlated within the limits of that rare event for up to `M` lag values or shifts of `ts_ref`. Iteratively, we test increasingly larger lag values and employ Pearson coefficient to measure the linear correlation between a window of size `ws` of both series. Whenever the correlation is above `r > 0.5` and statistically significant (`p < 0.05`), we keep the lag value. This cross-correlation procedure is iteratively repeated (as many times as possible) by pushing back the origin of the rare event in both time series `step` points per iteration. For all the lags stored due to their significance, we keep the T-most frequent values. This function is applied to each pair of stations (`ts_orig`, `ts_ref`) for all the rare events detected in the previous section. For each pair of stations, we select a single lag among the T-most frequent values of all the events: the lag with the highest mean correlation that appears in the most-frequent list of at least half of the events.

**Algorithm 1** Pseudo-code of the cross-correlation procedure

**Require:** $ts\_orig, ts\_ref, ws > 0, step > 0, M \geq 0$
  $T_0 \leftarrow 0$
  $R \leftarrow \{\}$ ▷ Initially empty set
  **while** $T_0 < length(ts\_orig) - M$ **do**
    $lag \leftarrow 0$
    **while** $lag < M$ **do**
      $(r, p) \leftarrow pearson(ts\_orig[T_0 : ws + T_0],$
        $ts\_ref[lag + T_0 : ws + lag + T_0])$ ▷ Pearson correlation
      **if** $(r > 0.5)$ and $(p < 0.05)$ **then** ▷ (coeff., $r$, and p-value, $p$)
        $R \leftarrow R \cup \{(lag, r, p)\}$
      **end if**
      $lag \leftarrow lag + 1$
    **end while**
    $T_0 \leftarrow T_0 + step$
  **end while**
  **return** $R$

The final time-lapses estimated for each station are summarized in Table 1. The inferred lags between river flow stations are largely in line with the physical distance between them (see Fig. 1). Checking the validity of the measured time-elapse for the rainfall stations is not that straightforward. Future deployment of a monitoring and predictive system based on these methods would require validation from field experts.

**Table 1**
Lag (time elapsed, in minutes) for each station, together with the percentage of rare events (Perc.) in which the lag was found relevant and the mean Pearson correlation (Mean corr.) in these events. Symbols are used to differentiate river flow° and rainfall† stations (find in Fig. 1 their physical location).

(a) Upper course.

| Station | Lag | Perc. | Mean corr. |
|---|---|---|---|
| L17147-72-00005° | 300 | 53% | 0.886 |
| L17167-72-00001° | 300 | 53% | 0.849 |
| CC† | 390 | 47% | 0.740 |
| CG† | 420 | 29% | 0.747 |
| CI† | 480 | 35% | 0.747 |
| CY† | 480 | 24% | 0.717 |
| DG† | 420 | 24% | 0.735 |
| V3† | 450 | 35% | 0.711 |
| V4† | 420 | 47% | 0.745 |
| V5† | 390 | 35% | 0.733 |
| WS† | 540 | 35% | 0.775 |
| Z4-ZC† | 510 | 24% | 0.731 |

(b) Lower course.

| Station | Lag | Perc. | Mean corr. |
|---|---|---|---|
| L17038-72-00002° | 120 | 67% | 0.894 |
| L17079-72-00004° | 210 | 67% | 0.897 |
| L17079-72-00005° | 210 | 67% | 0.876 |
| F001242° | 420 | 44% | 0.897 |
| F026458° | 120 | 67% | 0.852 |
| DJ† | 540 | 56% | 0.759 |
| DN† | 690 | 44% | 0.771 |
| UN† | 780 | 33% | 0.822 |
| UO† | 480 | 44% | 0.758 |
| VN† | 810 | 33% | 0.817 |
| WS† | 720 | 56% | 0.768 |
| DM-XJ† | 600 | 44% | 0.756 |

## 4. Predictive modeling

In this section, the lag estimations are used to calibrate a set of models that compose our interpretable and actionable predictive system.

### 4.1. Set up

We prepare two standard ML datasets, one for each reference station. Each dataset is formed by the time series of the upper or lower course, correspondingly, aligned using the lags of Table 1. For each dataset, the time series of the corresponding reference station becomes the outcome vector, $y$, that is, what we want to predict. The time series of the rest of the stations form the columns of the input matrix, $\boldsymbol{X}$, that is, the predictive or independent variables. Each row $(\boldsymbol{X}_t, y_t)$ encodes the situation of the reference station in time '$t$', $y_t = TS_t^{ref}$, and the situation of each station $j$ in time '$t - \text{lag}_j$', $X_{t,j} = TS^j_{t-\text{lag}_j}$. This alignment of the columns (stations) will provide actionable information to decision-makers: "to anticipate what is going to happen at the reference station, observe what happened $\text{lag}_j$ minutes before at station $j$".

**Regressors:** Four different types of regressors are considered: $K$NN (Cover and Hart, 1967), Linear regression (LR) (Hastie et al., 2008), Random Forest (RF) (Breiman, 2001) and XGBoost (XGB) (Chen and Guestrin, 2016). We use halving grid search for hyper-parameter tuning of $K$NN, RF and XGB, with $R^2$ metric as scoring function. It uses $\{1, 3, 5, 10, 20\}$ neighbors, leaf sizes of $\{20, 30, 50\}$, and uniform or distance-based weights for $K$NN. RF and XGB optimization considers $\{50, 100, 200\}$ estimators, $\{1, \text{sqrt}, \log 2\}$ features per split, and $\{8, 10, 25, \text{None}\}$ depth. The rest of the hyper-parameters are set to default values. The source code is available on the website associated with this study[1].

**Data subsets.** Noting the potential difficulty of fitting heavy-rain or extreme flow rise measurements, we perform *additional* experiments using specific data subsets:

- A subset with *no-precipitation* data, i.e., it only includes data from periods without precipitation. We hypothesize that when the training data includes heavy rainfall episodes (rare) the model tries to adjust to them, losing the ability to fit well the calm periods (the vast majority).
- A subset with *precipitation-only* data, i.e., it only includes data from periods with precipitation. As before, attempting to model the curves of river flow when the training signal is mostly plain (no precipitation) is presumably inadequate. We hypothesize that learning only from data with signal (at least a minimum amount of precipitation) can boost the performance of a model fitted to this curve.

To split the dataset into rainy/non-rainy periods, we use the average rainfall across all the rainfall stations over/under 0.1 mm. Dunkerley (2021) recently reviewed the thresholds used in the related literature and found that light or low-intensity rainfalls are usually characterized to be approximately under 0.2 mm/h (or 0.1 mm in our 30-minute frequency data). With this threshold, the rain volume left in the so-called non-rainy periods accounts for $2.3 - 4.2\%$ of the total volume. Using precipitation to decide which model to use is an easily actionable criterion that can be easily implemented in the hypothetical deployment of this prediction system.

Due to the small proportion of rainy periods, we perform synthetic oversampling in the precipitation-only data subset to help the models gain robustness. We use Synthetic Minority Oversampling technique for regression with Gaussian Noise (SMOGN) (Branco et al., 2017), which adapts the classical SMOTE (Chawla et al., 2002) to generate synthetic samples for regression. Each new sample is generated by interpolating its values from two real neighbor samples and adjusted by a linear model to guarantee generalization.

**Validation.** Standard cross-validation (CV) is not valid for time series. In the usual alternative, time series split validation (TSS), the training data slice grows with the CV iterations and the validation fold is the next fixed-size slice of the series (Fig. 5a). In the common validation set (CVS) strategy, the training data slice also grows with the validation iterations but the validation fold is held constant (the last end of the series) (Fig. 5b). Whereas the increased robustness of a model learned from a larger dataset can be hidden with TSS, CVS cannot measure the model's generalization to different validation sets. We used both TSS and CVS in these experiments.

Performance is measured in terms of root mean squared error, $\text{RMSE} = \sqrt{N^{-1} \cdot \sum_{j=1}^{N} (y_j - \hat{y}_j)^2}$ where $\boldsymbol{y} = \{y_1, \ldots, y_N\}$ are the true values and $\hat{\boldsymbol{y}} = \{\hat{y}_1, \ldots, \hat{y}_N\}$ are the predictions, and Nash–Sutcliffe efficiency coefficient, $\text{NSE} = 1 - \left(\sum_{j=1}^{N} (y_j - \hat{y}_j)^2\right) / \left(\sum_{j=1}^{N} (y_j - \bar{y})^2\right)$, where $\bar{y}$ represents the mean true value through the evaluation period $\left(\bar{y} = N^{-1} \cdot \sum_{j=1}^{N} y_j\right)$. Both are standard metrics to evaluate hydrological models. Whereas NSE gathers insights about the increased performance of the model regarding a simplistic system that always predicts the mean value, RMSE provides information about the divergence between the real and predicted values while maintaining the same scale of the data, which makes it more interpretable.

Reporting model performance exclusively relying on aggregation-based metrics such as RMSE or NSE can mask the performance in specific relevant situations that might interest domain experts (Burnell et al., 2023). In these cases, local inspection is valuable. In this study, we inspect five different events: two standard situations where no particularly high rise is observed (**S1** and **S2**) and three events of heavy rain and large flow rise: one specifically relevant in the lower course (**RL**), **Leslie** storm, with a bigger impact in the upper course, and **Gloria** storm, which impacted both parts.

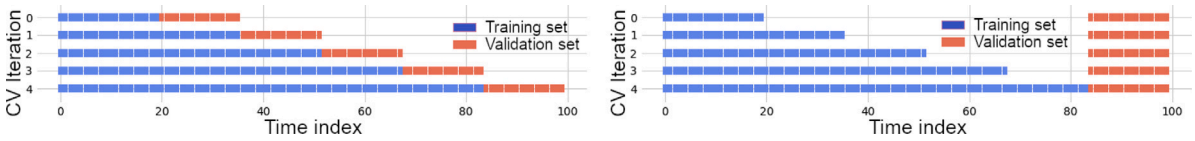


**Fig. 5.** Graphical representation of TSS and CVS validation strategies.

**Table 2**
Results in terms of averaged RMSE and NSE (and associated standard deviation) following TSS (left) and CVS (right) validation strategies for 4 regressors in the upper and lower courses. Results are shown with the complete dataset, and with the subsets of no-precipitation and precipitation-only data. Bold type indicates the best model per dataset.

(a) Results with TSS validation strategy.

| Model | Upper course | | Lower course | |
|---|---|---|---|---|
| | RMSE | NSE | RMSE | NSE |
| KNN | 15.34 ± 4.63 | 0.64 ± 0.14 | 21.31 ± 18.08 | 0.54 ± 0.18 |
| LR | **13.05 ± 4.39** | **0.75 ± 0.07** | 19.31 ± 13.15 | 0.51 ± 0.30 |
| RF | 13.53 ± 5.43 | 0.74 ± 0.05 | **17.92 ± 14.43** | **0.68 ± 0.10** |
| XGB | 15.02 ± 5.28 | 0.67 ± 0.10 | 18.83 ± 15.70 | 0.65 ± 0.11 |
| Complete dataset | | | | |
| KNN | 8.17 ± 1.66 | 0.52 ± 0.29 | 14.33 ± 12.84 | 0.50 ± 0.17 |
| LR | 6.08 ± 1.68 | 0.74 ± 0.18 | 12.61 ± 8.50 | 0.36 ± 0.55 |
| RF | **6.06 ± 1.04** | **0.77 ± 0.07** | 14.16 ± 12.59 | 0.56 ± 0.09 |
| XGB | 6.05 ± 1.43 | 0.75 ± 0.15 | **11.88 ± 11.31** | **0.65 ± 0.16** |
| No-precipitation subset | | | | |
| KNN | 63.82 ± 46.56 | 0.69 ± 0.26 | 107.03 ± 107.87 | 0.65 ± 0.36 |
| LR | **51.69 ± 18.80** | **0.78 ± 0.06** | 104.36 ± 73.62 | 0.67 ± 0.16 |
| RF | 56.71 ± 57.71 | 0.73 ± 0.35 | **103.38 ± 117.68** | **0.68 ± 0.42** |
| XGB | 57.72 ± 54.91 | 0.72 ± 0.33 | 126.01 ± 95.17 | 0.44 ± 0.43 |
| Precipitation-only subset (oversampled) | | | | |

(b) Results with CVS validation strategy.

| Model | Upper course | | Lower course | |
|---|---|---|---|---|
| | RMSE | NSE | RMSE | NSE |
| KNN | 20.51 ± 0.38 | 0.75 ± 0.01 | 40.30 ± 0.65 | 0.52 ± 0.02 |
| LR | **16.90 ± 0.15** | **0.83 ± 0.00** | **33.47 ± 1.45** | **0.67 ± 0.03** |
| RF | 19.34 ± 2.01 | 0.78 ± 0.05 | 33.91 ± 1.10 | 0.66 ± 0.02 |
| XGB | 19.42 ± 2.10 | 0.78 ± 0.05 | 37.26 ± 0.94 | 0.59 ± 0.02 |
| Complete dataset | | | | |
| KNN | 8.06 ± 1.00 | 0.80 ± 0.05 | 28.67 ± 0.04 | 0.43 ± 0.00 |
| LR | **5.40 ± 0.09** | **0.91 ± 0.00** | **22.13 ± 0.57** | **0.66 ± 0.02** |
| RF | 7.22 ± 0.31 | 0.84 ± 0.01 | 27.06 ± 2.72 | 0.49 ± 0.10 |
| XGB | 5.85 ± 0.24 | 0.90 ± 0.01 | 24.45 ± 1.36 | 0.59 ± 0.05 |
| No-precipitation subset | | | | |
| KNN | 53.34 ± 28.63 | 0.72 ± 0.29 | 59.42 ± 38.68 | 0.75 ± 0.29 |
| LR | **46.20 ± 7.02** | **0.82 ± 0.06** | **63.79 ± 7.05** | **0.78 ± 0.05** |
| RF | 46.70 ± 38.77 | 0.74 ± 0.37 | 71.88 ± 36.30 | 0.67 ± 0.27 |
| XGB | 46.56 ± 37.67 | 0.74 ± 0.36 | 109.85 ± 79.85 | 0.11 ± 1.06 |
| Precipitation-only subset (oversampled) | | | | |

### 4.2. Results

The overall results of our predictive analysis using both TSS and CVS validation strategies are displayed in Table 2 for both the upper and lower courses. Holding constant the validation set (with CVS) shows consistently lower variance and larger error than using TSS. For example, in the upper course with the whole data, *K*NN shows a mean RMSE of 15.34 with TSS and 20.51 with CVS, whereas the associated standard deviation is 4.54 and 0.38 respectively. As shown in Fig. 4, in the last portion of the series a concentration of more and heavier flow rise events exists. These events are included in the last validation subset, the only one used by CVS. In contrast, TSS uses different validation subsets, some of them with long easy-to-predict valley periods. Thus, we can describe the different sources of variation: whereas the standard deviation measured with CVS is due to an increasingly-large training set (and the intrinsic variability of the model), TSS also registers the variability due to changing validation subsets. The behavior with the precipitation-only subset is slightly different because synthetic oversampling is applied to the training data, forcing the model to fit larger curves. Linear regression stands out as the best model, in general. Random forest and XGBoost, in this order, show competitive results. In the upper course, LR is the best model in the vast majority of experiments. In the lower course, its superiority is less clear. Per validation strategy, the best model is always LR with CVS and in the majority of cases RF with TSS. These results might indicate that simpler models, with a smaller number of parameters to fit, might be more appropriate for this problem: while holding top performance, the limited standard deviation suggests that a competitive model fit can be achieved with smaller data sizes. Reasonably, the lowest errors are achieved with the no-precipitations dataset, as large flow rise events are not present in the dataset and the models can fit the data better. These events are collected in the precipitation-only subset, and hence the associated large error values. Note that RMSE measures absolute error. Results in terms of NSE show that models for the no-rain periods might be relatively worse than those for the rainy periods. The lower course seems more complex to model, as all the evaluation measurements are worse than those of the upper course. This can be partially explained by the presence of the reservoirs right before this part of the river, the management of which depends on different aspects such as reservoir level, weather forecasts or water diversions.

Tables 3 and 4 show the results of the models learned with the different datasets when validated on 5 different events in the upper and lower courses, respectively: two standard situations of the river (S1 and S2), a larger flow rise in the upper course which was moderate in the lower course (Leslie storm) and the other way around (RL), as well as Gloria storm, which affected both parts of the river. The validation with these 5 events is similar to CVS (Fig. 5b): each experiment is repeated with training datasets of increasing size and, hence, the observed variability (standard deviation). In standard situation S1, models learned with the complete data or the no-precipitation subset show similar results in both courses. Regardless of the RMSE values, NSE results indicate that this type of small flow rise event is difficult to model, especially in the lower course. As expected, models learned with precipitation-only data are not competitive in this case. However, with standard situation S2, which covers a slightly heavier flow rise, some regressors (LR with the precipitation-only subset in the upper course, KNN with the no-precipitation dataset in the upper course) provide good results while the vast majority fails to model it properly. Interestingly, models learned only with data from rainy events start showing up as competitive. Results during RL, *Gloria* and *Leslie* events are consistently better when the models are learned from the precipitation-only subset. For example, for *Gloria* in the lower course, LR improves from a RMSE value of 296.32 with the complete dataset to 138.62 with the subset and, in terms of NSE, from 0.04 to 0.79. To better illustrate this gain, Fig. 6 shows the original river flow time series of these three last flow rise events, and the prediction of the LR and RF models learned with the complete data and the precipitation-only subset. It can be easily appreciated how the models learned with data of rainy periods fits better the real curve. By visual inspection, RF fits better the real time series. LR overestimates (or underestimates,

**Table 3**
Results in terms of averaged RMSE and NSE (and associated standard deviation) in 5 events in the upper course using 4 different types of regressors. Results are shown with the complete dataset, and with the subsets of no-precipitation and precipitation-only data. Bold type indicates the best model per event.

| Model | S1 | S2 | RL | Leslie | Gloria | Metric |
|---|---|---|---|---|---|---|
| KNN | 10.48 ± 0.59 | 14.57 ± 0.67 | 26.76 ± 0.94 | 141.68 ± 5.7 | 217.28 ± 3.76 | RMSE |
| | −0.08 ± 0.12 | 0.63 ± 0.03 | 0.45 ± 0.04 | 0.63 ± 0.03 | 0.49 ± 0.02 | NSE |
| LR | **8.27 ± 0.93** | 12.54 ± 1.63 | 22.26 ± 1.69 | 89.24 ± 5.3 | 213.28 ± 6.66 | RMSE |
| | **0.32 ± 0.15** | 0.72 ± 0.07 | 0.62 ± 0.06 | 0.85 ± 0.02 | 0.5 ± 0.03 | NSE |
| RF | 8.73 ± 0.15 | 13.17 ± 0.75 | 22.93 ± 1.29 | 145.58 ± 20.96 | 211.71 ± 25.55 | RMSE |
| | 0.25 ± 0.03 | 0.7 ± 0.03 | 0.6 ± 0.04 | 0.61 ± 0.11 | 0.51 ± 0.12 | NSE |
| XGB | 9.89 ± 0.62 | 13.52 ± 1.02 | 27.84 ± 2.4 | 144.61 ± 32.31 | 210.17 ± 17.7 | RMSE |
| | 0.04 ± 0.12 | 0.68 ± 0.05 | 0.41 ± 0.1 | 0.6 ± 0.16 | 0.52 ± 0.08 | NSE |
| Complete dataset | | | | | | |
| KNN | 10.41 ± 0.46 | 15.74 ± 0.03 | 27.77 ± 0.71 | 199.33 ± 3.01 | 320.19 ± 7.29 | RMSE |
| | −0.07 ± 0.09 | 0.57 ± 0.0 | 0.41 ± 0.03 | 0.27 ± 0.02 | −0.12 ± 0.05 | NSE |
| LR | 10.05 ± 1.31 | 18.84 ± 2.61 | 28.27 ± 1.99 | 109.52 ± 4.31 | 247.98 ± 9.97 | RMSE |
| | −0.01 ± 0.27 | 0.37 ± 0.18 | 0.39 ± 0.09 | 0.78 ± 0.02 | 0.33 ± 0.05 | NSE |
| RF | **8.29 ± 0.15** | 12.68 ± 0.59 | 22.65 ± 2.74 | 215.7 ± 0.17 | 359.52 ± 1.08 | RMSE |
| | **0.32 ± 0.02** | 0.72 ± 0.03 | 0.6 ± 0.09 | 0.15 ± 0.0 | −0.41 ± 0.01 | NSE |
| XGB | 9.51 ± 0.68 | 16.21 ± 2.45 | 27.68 ± 1.7 | 203.28 ± 10.55 | 329.2 ± 22.71 | RMSE |
| | 0.11 ± 0.13 | 0.54 ± 0.13 | 0.41 ± 0.07 | 0.24 ± 0.08 | −0.18 ± 0.16 | NSE |
| No-precipitation subset | | | | | | |
| KNN | 9.79 ± 5.82 | 12.7 ± 2.9 | 14.5 ± 2.34 | 131.91 ± 71.23 | 215.31 ± 120.75 | RMSE |
| | −0.16 ± 1.31 | 0.71 ± 0.14 | 0.84 ± 0.05 | 0.62 ± 0.4 | 0.39 ± 0.66 | NSE |
| LR | 11.11 ± 3.18 | **10.43 ± 2.39** | 18.32 ± 2.86 | **83.03 ± 15.51** | 192.45 ± 37.02 | RMSE |
| | −0.28 ± 0.64 | **0.8 ± 0.09** | 0.74 ± 0.08 | **0.87 ± 0.05** | **0.59 ± 0.16** | NSE |
| RF | 10.78 ± 7.08 | 11.34 ± 6.61 | **10.46 ± 0.37** | 106.98 ± 106.88 | 189.58 ± 160.81 | RMSE |
| | −0.47 ± 1.78 | 0.73 ± 0.3 | **0.92 ± 0.01** | 0.65 ± 0.54 | 0.42 ± 0.83 | NSE |
| XGB | 10.33 ± 4.86 | 10.88 ± 4.95 | 11.12 ± 0.67 | 111.83 ± 104.02 | **189.23 ± 156.64** | RMSE |
| | −0.2 ± 1.12 | 0.77 ± 0.21 | 0.91 ± 0.01 | 0.64 ± 0.54 | 0.43 ± 0.8 | NSE |
| Precipitation-only subset (oversampled) | | | | | | |

**Table 4**
Results in terms of averaged RMSE and NSE (and associated standard deviation) in 5 events in the lower course using 4 different types of regressors. Results are shown with the complete dataset, and with the subsets of no-precipitation and precipitation-only data. Bold type indicates the best model per event.

| Model | S1 | S2 | RL | Leslie | Gloria | Metric |
|---|---|---|---|---|---|---|
| KNN | 5.96 ± 0.49 | 45.77 ± 14.41 | 138.35 ± 41.93 | 71.94 ± 0.72 | 207.58 ± 35.38 | RMSE |
| | −6.76 ± 1.23 | −2.43 ± 1.84 | 0.34 ± 0.34 | 0.6 ± 0.01 | 0.52 ± 0.17 | NSE |
| LR | 5.9 ± 0.65 | 42.91 ± 2.2 | 97.85 ± 29.83 | **39.38 ± 10.54** | 296.32 ± 15.53 | RMSE |
| | −6.61 ± 1.7 | −1.83 ± 0.29 | 0.67 ± 0.21 | **0.87 ± 0.07** | 0.04 ± 0.1 | NSE |
| RF | **2.43 ± 0.53** | 44.0 ± 8.11 | 92.57 ± 6.3 | 50.26 ± 5.96 | 206.35 ± 34.41 | RMSE |
| | **−0.33 ± 0.59** | −2.04 ± 1.12 | 0.72 ± 0.04 | 0.8 ± 0.05 | 0.53 ± 0.16 | NSE |
| XGB | 3.58 ± 0.22 | 66.02 ± 24.64 | 139.2 ± 8.44 | 58.35 ± 10.83 | 216.04 ± 36.32 | RMSE |
| | −1.79 ± 0.35 | −6.31 ± 5.46 | 0.37 ± 0.08 | 0.73 ± 0.1 | 0.48 ± 0.18 | NSE |
| Complete dataset | | | | | | |
| KNN | 5.39 ± 0.02 | **14.94 ± 0.23** | 147.71 ± 0.25 | 77.62 ± 1.5 | 288.51 ± 16.68 | RMSE |
| | −5.31 ± 0.05 | **0.66 ± 0.01** | 0.29 ± 0.0 | 0.53 ± 0.02 | 0.09 ± 0.11 | NSE |
| LR | 6.95 ± 1.93 | 102.87 ± 24.26 | 293.86 ± 75.56 | 48.94 ± 14.16 | 592.97 ± 113.62 | RMSE |
| | −10.03 ± 6.27 | −15.83 ± 8.13 | −1.92 ± 1.48 | 0.8 ± 0.12 | −2.94 ± 1.51 | NSE |
| RF | 4.32 ± 1.35 | 38.04 ± 5.88 | 124.68 ± 20.57 | 81.51 ± 22.13 | 310.2 ± 51.2 | RMSE |
| | −3.32 ± 2.69 | −1.26 ± 0.71 | 0.49 ± 0.16 | 0.46 ± 0.27 | −0.07 ± 0.33 | NSE |
| XGB | 3.51 ± 0.12 | 25.23 ± 9.71 | 109.76 ± 7.83 | 49.8 ± 5.81 | 232.18 ± 24.08 | RMSE |
| | −1.68 ± 0.19 | −0.07 ± 0.8 | 0.61 ± 0.06 | 0.81 ± 0.05 | 0.41 ± 0.13 | NSE |
| No-precipitation subset | | | | | | |
| KNN | 23.41 ± 33.31 | 32.05 ± 12.66 | **59.63 ± 63.32** | **39.06 ± 19.45** | 188.87 ± 146.03 | RMSE |
| | −278.52 ± 477.58 | −0.74 ± 1.39 | 0.8 ± 0.32 | **0.86 ± 0.13** | 0.45 ± 0.74 | NSE |
| LR | 9.49 ± 8.41 | 24.97 ± 0.41 | 65.96 ± 20.36 | **42.82 ± 4.45** | **138.62 ± 6.17** | RMSE |
| | −28.79 ± 41.59 | 0.04 ± 0.03 | **0.85 ± 0.09** | **0.86 ± 0.03** | **0.79 ± 0.02** | NSE |
| RF | 23.09 ± 20.31 | 24.38 ± 14.98 | 68.0 ± 60.23 | 49.24 ± 44.96 | 182.48 ± 181.74 | RMSE |
| | −174.43 ± 222.7 | −0.14 ± 1.32 | 0.77 ± 0.33 | 0.71 ± 0.43 | 0.39 ± 0.93 | NSE |
| XGB | 13.73 ± 13.76 | 24.22 ± 10.43 | 61.62 ± 59.21 | 47.7 ± 28.33 | 198.06 ± 150.05 | RMSE |
| | −67.33 ± 104.2 | −0.01 ± 0.86 | 0.8 ± 0.3 | 0.78 ± 0.24 | 0.41 ± 0.79 | NSE |
| Precipitation-only subset (oversampled) | | | | | | |

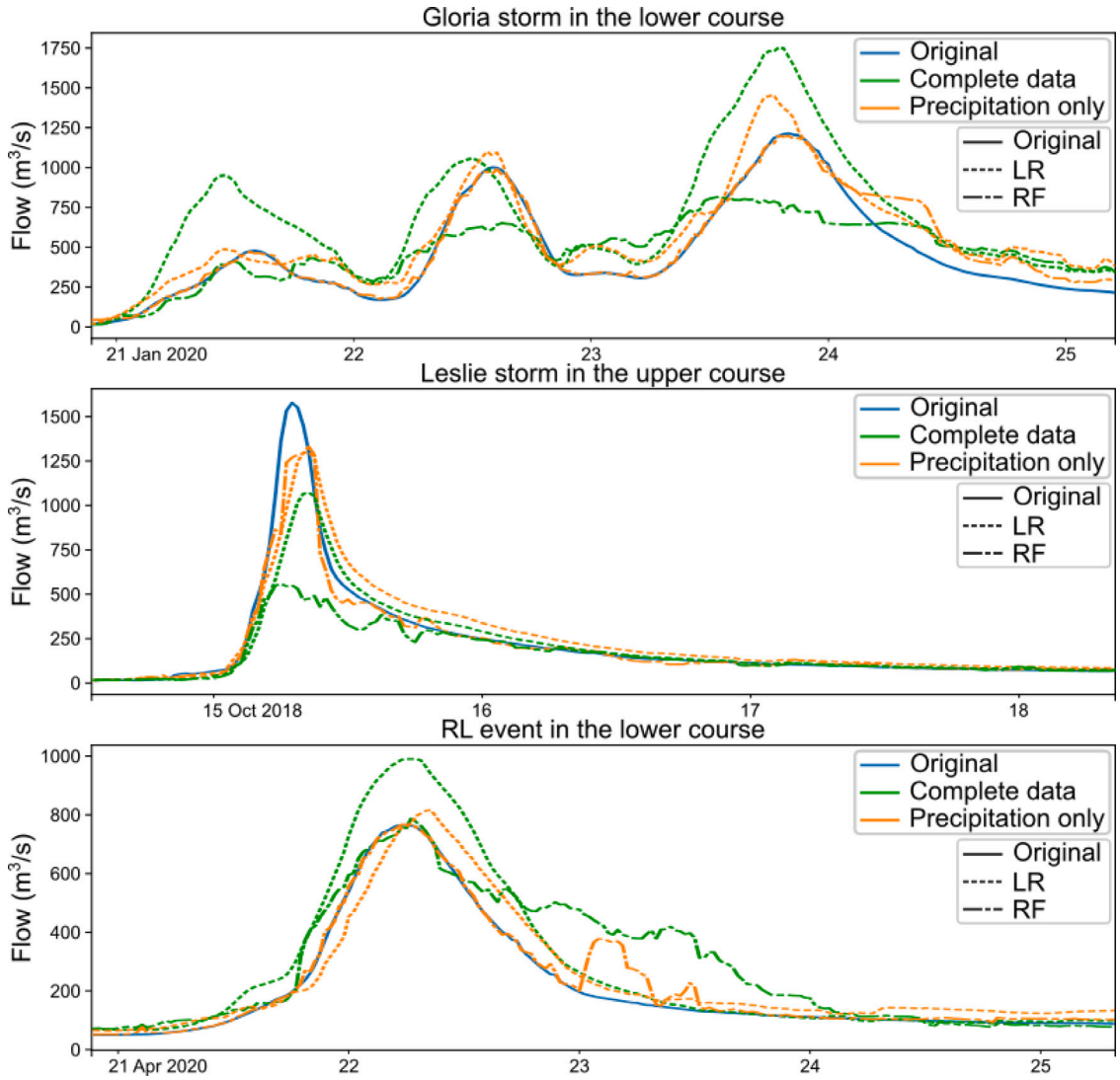


**Fig. 6.** Each plot shows, for a specific event (see Section 2), the real time series of the river flow and the predicted time series of LR and RF models learned with the complete data and models learned with (oversampled) precipitation-only data.

in *Leslie* event) and shifts the high flow peaks. However, LR produces smoother time series, closer to reality. RF introduces artifacts and produces shaky time series. Similar figures for other models, events and parts of the river can be found as supplementary material on the associated. website[1]

### 4.3. Measuring the impact of the alignment of stations

The alignment of the time series of the different rainfall and river flow stations regarding the corresponding reference station is, to our knowledge, novel in this type of machine learning based prediction of river flow. Measuring the impact of this alignment step can help us understand the extra information that this provides to our models. The following experiment aims to measure this impact.

The models presented above have been compared with a set of models learned from the same data without alignment. For these new models, we fix the prediction window in six hours. That is, they answer the question: given the current situation in the different stations, what will be the river flow in the reference station in six hours? Mean relative performance improvement (calibrated regarding uncalibrated models) is summarized in Table 5. Note that for the complete dataset, there is no practical *mean* improvement; only a subset of models of the river show improvement. For the precipitation-only subset, there is a clear mean improvement in terms of RMSE: the error of calibrated models is on average a 25.85% lower. In terms of NSE, the mean relative improvement (increase) is 18.97%. After a detailed inspection of the results, we observe improvement in almost every model from both parts of the river, although the improvement rates of the lower-course models are the largest.

These results are partially determined by our selection of six hours for the prediction window of uncalibrated models. It was selected as a large enough, but still short, time to set up preparatory actions. Another value for this prediction window would definitively lead to different results. However, this fact highlights the need for some sort of lag estimation step in flow prediction studies so that the prediction window is not fixed arbitrarily and, mainly, to understand the dynamics of the river flow in large flow rise situations. The difference in improvement rates between experiments using the complete data or the precipitation-only subset can also be explained by our choice of rare events to estimate the lags. That is, by design, the calibration method focuses on adjusting the large-rise events and, thus, the large improvement observed when only using data from rainy (and flow rise) periods seems reasonable. In stable periods the river flow is rather constant, the flow values do not change considerably from hour to hour. Thus, for the complete dataset, where these stable periods are predominant, calibration is irrelevant. Nevertheless, we must emphasize that the

**Table 5**
Mean relative improvement in RMSE and NSE over all the models and both parts of the river before and after performing the calibration (aligning the time series according to the estimated lags). Results are shown with the complete dataset, and with the precipitation-only subsets.

| Complete dataset | | Precipitation-only subset | |
|---|---|---|---|
| RMSE | NSE | RMSE | NSE |
| 0.2% | −0.4% | −25.85% | 18.97% |

final objective of flow modeling is to predict large flow rises (potential floods), which highlights the importance of the results with the precipitation-only subset.

## 5. Discussion

Our models, which fit different scenarios of the river, show promising results. They not only meet the objective of helping to anticipate floods but also provide useful and interpretable information about their functioning.

**Data collection.** The first contribution of this work is the collection and curation of rainfall and river flow data for the study. In the future, more data or information could also be provided to the models. For example, floods are known to be associated with saturated soils that cannot absorb more water. To account for this behavior, soil moisture data could be gathered (Kumar et al., 2021) or, by feature engineering our current data, additional information on the number of rainy days in a certain past period or cumulative rainfall measurements (Saint-Fleur et al., 2023) over longer periods could be fed to the model.

**Modeling decisions.** Due to its physical traits, we decided to model the upper and lower courses of the river separately. We chose (and modeled the river flow at) one station from each part of the river, as stations close to flood-risk areas. The results of this work depend on the selection of these two stations as reference points; results might be different if other stations were chosen as references. Specifically, trying to model a station close to the source of the river could make the model depend exclusively on rainfall measurements. This type of analysis could be interesting to understand the contribution of rainfall to the river flow.

**Estimation of the time lag between stations.** The measurement of the lag or time elapsed between the observation of an event across different stations of the river is our result with the largest explanatory potential. This information was used to calibrate the ML models. This **calibration** is key to the *interpretability* of the produced models as it can guide domain managers not only to pay attention to the most relevant stations but also to indicate them *when* to look at these relevant sites. In the future, calibration could also be carried out within a multivariate correlation analysis, instead of the current bivariate study, to account for the interactions between stations. Recent studies using DNNs have been able to provide information similar to that of our calibration study taking advantage of spatial–temporal attention mechanisms (Chang et al., 2024) or physics-informed models (Saint-Fleur et al., 2023). It would be interesting to study whether these techniques can be used to estimate the lags for our alignment process.

The calibration also determines the largest achievable prediction window. As found also by Saint-Fleur et al. (2023), if our method finds that river flow rises with a difference of two hours between the reference and its closest station, our models will not be able to provide a prediction for a window longer than two hours. Eventually, a resulting short prediction window might not enable a feasible response to potential floods (not enough time for setting up preparatory actions). If this were the case, the station that determines the prediction window might be removed from the calibration step or, directly, from the predictive models at the cost of potential performance loss.

**Standard machine learning techniques.** We decided to avoid complex and black-box models such as ANNs and modern DNNs. Standard ML models provide competitive results and a good fit to flow rise events, as shown in Section 4. Moreover, thanks to the use of standard ML models, off-the-shelf model inspection techniques (Molnar, 2022) could be used to gain extra *interpretation*, such as feature importance metrics to understand which are the most relevant stations for determining river flow at each reference station. There might be room for improvement if the different components of our solution could be further optimized (using other types of models, enlarged hyperparameter tuning process, etc.).

**Adapting to extreme situations.** We have analyzed different models with various configurations of the data due to the difficulty of ML models to fit the extremes, that is, low and high flow levels (Kumar et al., 2021). In line with Saint-Fleur et al. (2023), we show that a model fitted to the heavy-rain/flow-rise events (precipitation-only subset) is useful. It shows a reduced predictive error regarding models learned with the complete dataset which registers, most of the time, a flat signal (no rain - no flow change). The improvement of models learned with the subset of no-precipitation data (only these predominantly flat periods) regarding those models learned with the complete data is more modest. In practice, a combination of models (precipitation-only/no-precipitation) might be more appropriate to build a predictive system that can anticipate high-flow episodes. The trigger of the decision to use one or another model in such a system, the mean precipitation, is easily measured and therefore actionable.

## 6. Conclusion

This study uses traditional machine learning to model the flow of the Ter river (NE Spain), which has historically suffered floods with a vast social impact. The goal is to anticipate flood episodes and enable the preparation of mitigation actions by the responsible institutions. We propose a predictive system composed of a set of models that fit different weather scenarios. Rainfall and river flow stations feed the models after calibrating them according to their estimated time lag with the modeled stations. The calibration step increases the interpretability of the models, providing actionable information about which station to pay attention to (and when) to anticipate overflow risk. Our predictive system shows a promising ability to model river flow. Empirical results suggest that a combination of two models is worthwhile: a model learned from data of heavy-rain periods which offers robust estimations during flow-rise events, and another one that models the more common calm periods.

River flow modeling is an open problem that can lead to several lines of future research, some of which are identified in our discussion above. Recovering or imputing data for the time series of the stations discarded due to their large missing rate is a potentially simple direction to enhance our models. Alternative approaches could consider pure time series analyses, possibly using current deep recurrent models. The present study could be extended by learning similar models using rainfall forecasts. This could enable an extensive analysis to understand how rainfall affects the flow in every river flow station. Domain experts should validate these data-driven results and our alignment of stations. We will explore the possibility of deploying a machine learning based predictive system that assists the Catalan Water Agency's decision-makers.

## CRediT authorship contribution statement

**Fabián Serrano-López:** Software, Investigation, Validation, Writing – original draft. **Sergi Ger-Roca:** Data curation, Investigation, Software, Writing – original draft. **Maria Salamó:** Methodology, Supervision, Writing – review & editing. **Jerónimo Hernández-González:** Conceptualization, Methodology, Supervision, Validation, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Source code and curated data are available on the website: https://jhernandezgonzalez.github.io/supp_ter.html. Raw data is available upon request.

## Acknowledgments

MS is partially funded by Generalitat de Catalunya, Spain (2021-SGR-00313). We thank ACA and SMC experts who provided useful feedback.

## References

Apollonio, C., Balacco, G., Novelli, A., Tarantino, E., Piccinni, A.F., 2016. Land use change impact on flooding areas: the case study of Cervaro basin (Italy). Sustainability 8, 996. http://dx.doi.org/10.3390/su8100996.

Bafitlhile, T.M., Li, Z., 2019. Applicability of $\epsilon$-support vector machine and artificial neural network for flood forecasting in humid, semi-humid and semi-arid basins in china. Water 11, 85. http://dx.doi.org/10.3390/w11010085.

Barnolas, M., Llasat, M.C., 2007. A flood geodatabase and its climatological applications: the case of catalonia for the last century. Nat. Hazards Earth Syst. Sci. 7, 271–281. http://dx.doi.org/10.5194/nhess-7-271-2007.

Bhasme, P., Bhatia, U., 2024. Improving the interpretability and predictive power of hydrological models: Applications for daily streamflow in managed and unmanaged catchments. J. Hydrol. 628, 130421. http://dx.doi.org/10.1016/j.jhydrol.2023.130421.

Blöschl, G., Kiss, A., Viglione, A., Barriendos, M., Böhm, O., Brázdil, R., Coeur, D., Demarée, G., Llasat, M.C., Macdonald, N., Retsö, D., Roald, L., Schmocker-Fackel, P., Amorim, I., Belínová, M., Benito, G., Bertolin, C., Camuffo, D., Cornel, D., Doktor, R., Elleder, L., Enzi, S., Garcia, J.a. C., Glaser, R., Hall, J., Haslinger, K., Hofstätter, M., Komma, J., Limanówka, D., Lun, D., Panin, A., Parajka, J., Petric, H., Rodrigo, F.S., Rohr, C., Schönbein, J., Schulte, L., Silva, L.P., Toonen, W.H.J., Valent, P., Waser, J., Wetter, O., 2020. Current european flood-rich period exceptional compared with past 500 years. Nature 583, 560–566. http://dx.doi.org/10.1038/s41586-020-2478-3.

Branco, P., Torgo, L., Ribeiro, R.P., 2017. Smogn: a pre-processing approach for imbalanced regression. In: Proc. of the 1st International Workshop on Learning with Imbalanced Domains: Theory and Applications. PMLR, pp. 36–50.

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

Bürger, C., Kolditz, O., Fowler, H., Blenkinsop, S., 2007. Future climate scenarios and rainfall–runoff modelling in the upper gallego catchment (Spain). Environ. Pollut. 148, 842–854. http://dx.doi.org/10.1016/j.envpol.2007.02.002, aquaTerra: Pollutant behavior in the soil, sediment, ground, and surface water system.

Burnell, R., Schellaert, W., Burden, J., Ullman, T.D., Martinez-Plumed, F., Tenenbaum, J.B., Rutar, D., Cheke, L.G., Sohl-Dickstein, J., Mitchell, M., Kiela, D., Shanahan, M., Voorhees, E.M., Cohn, A.G., Leibo, J.Z., Hernandez-Orallo, J., 2023. Rethink reporting of evaluation results in ai. Science 380, 136–138. http://dx.doi.org/10.1126/science.adf6369.

Catalan Water Agency, O.R.G., 2019. Revisió i actualització de l'avaluació preliminar del risc d'inundació del districte de conca fluvial de catalunya (2n cicle). URL: https://aca.gencat.cat/web/.content/30_Plans_i_programes/20_Gestio_del_risc_inundacions/2n-cicle-de-planificacio/APRI/00_Memoria_APRI_2018_CA.pdf.

Chakraborty, D., Basagaoglu, H., Winterle, J., 2021. Interpretable vs noninterpretable machine learning models for data-driven hydro-climatological process modeling. Expert Syst. Appl. 170, 114498. http://dx.doi.org/10.1016/j.eswa.2020.114498.

Chang, K.H., Chiu, Y.T., Su, W.R., Yu, Y.C., Chang, C.H., 2024. A spatial–temporal deep learning-based warning system against flooding hazards with an empirical study in taiwan. Int. J. Disaster Risk Reduct. 102, 104263. http://dx.doi.org/10.1016/j.ijdrr.2024.104263.

Chang, Q., Ficklin, D.L., Jiao, W., Denham, S.O., Wood, J.D., Brunsell, N.A., Matamala, R., Cook, D.R., Wang, L., Novick, K.A., 2023. Earlier ecological drought detection by involving the interaction of phenology and eco-physiological function. Earth's Future 11, e2022EF002667. http://dx.doi.org/10.1029/2022EF002667.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357. http://dx.doi.org/10.1613/jair.953.

Chebii, S.J., Mukolwe, M.M., Ong'or, B.I., 2022. River flow modelling for flood prediction using artificial neural network in ungauged Perkerra catchment, Baringo County, Kenya. Water Pract. Technol. 17, 914–929. http://dx.doi.org/10.2166/wpt.2022.034.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proc. of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory 13, 21–27. http://dx.doi.org/10.1109/TIT.1967.1053964.

Cubasch, U., Wuebbles, D., Chen, D., Facchini, M., Frame, D., Mahowald, N., Winther, J.G., 2013. Introduction. In: Stocker, T., Qin, D., Plattner, G.K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I To the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, pp. 119–158.

Dibike, Y., Solomatine, D., 2001. River flow forecasting using artificial neural networks. Phys. Chem. Earth B: Hydrol. Oceans Atmos. 26, 1–7. http://dx.doi.org/10.1016/S1464-1909(01)85005-X.

Drobinski, P., Da Silva, N., Panthou, G., Bastin, S., Muller, C., Ahrens, B., Borga, M., Conte, D., Fosser, G., Giorgi, F., Güttler, I., Kotroni, V., Li, L., Morin, E., Önol, B., Quintana-Segui, P., Romera, R., Torma, C.Z., 2018. Scaling precipitation extremes with temperature in the Mediterranean: past climate assessment and projection in anthropogenic scenarios. Clim. Dyn. 51, 1237–1257. http://dx.doi.org/10.1007/s00382-016-3083-x.

Dunkerley, D., 2021. Light and low-intensity rainfalls: A review of their classification, occurrence, and importance in landsurface, ecological and environmental processes. Earth-Sci. Rev. 214, 103529. http://dx.doi.org/10.1016/j.earscirev.2021.103529.

European Environment Agency, O.R.G., 2023. Economic losses from weather- and climate-related extremes in europe. URL: https://www.eea.europa.eu/en/analysis/indicators/economic-losses-from-climate-related. (Last access: 1 July 2024).

Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., Lin, Q., 2020. Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. J. Hydrol. 589, 125188. http://dx.doi.org/10.1016/j.jhydrol.2020.125188.

Ghimire, S., Yaseen, Z.M., Farooque, A.A., Deo, R.C., Zhang, J., Tao, X., 2021. Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. Sci. Rep. 11, 17497. http://dx.doi.org/10.1038/s41598-021-96751-4.

Ha, S., Liu, D., Mu, L., 2021. Prediction of Yangtze River streamflow based on deep learning neural network with El Niño–Southern Oscillation. Sci. Rep. 11, 11738. http://dx.doi.org/10.1038/s41598-021-90964-3.

Hamitouche, M., Ribalta, M., 2023. Daily streamflow modelling using ml based on discharge and rainfall time series in the besós river basin, spain. Environ. Sci. Proc. 25, http://dx.doi.org/10.3390/ECWS-7-14168.

Hastie, T., Tibshirani, R., Friedman, J., 2008. Linear Methods for Regression. Springer, pp. 43–100.

Hounkpè, J., Diekkrüger, B., Afouda, A.A., Sintondji, L.O.C., 2019. Land use change increases flood hazard: a multi-modelling approach to assess change in flood characteristics driven by socio-economic land use change scenarios. Nat. Hazards 98, 1021–1050. http://dx.doi.org/10.1007/s11069-018-3557-8.

Hsu, K.l., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modeling of the rainfall-runoff process. Water Resour. Res. 31, 2517–2530. http://dx.doi.org/10.1029/95WR01955.

Jiang, J., Chen, C., Zhou, Y., Berretti, S., Liu, L., Pei, Q., Zhou, J., Wan, S., 2024. Heterogeneous dynamic graph convolutional networks for enhanced spatiotemporal flood forecasting by remote sensing. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17, 3108–3122. http://dx.doi.org/10.1109/JSTARS.2023.3349162.

Jimeno-Sáez, P., Senent-Aparicio, J., Pérez-Sánchez, J., Pulido-Velazquez, D., 2018. A comparison of SWAT and ANN models for daily runoff simulation in different climatic zones of peninsular spain. Water 10, http://dx.doi.org/10.3390/w10020192.

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. 29, 2318–2331. http://dx.doi.org/10.1109/TKDE.2017.2720168.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resour. Res. 55, 11344–11354. http://dx.doi.org/10.1029/2019WR026065.

Kulie, M.S., Milani, L., Wood, N.B., L'Ecuyer, T.S., 2020. Global Snowfall Detection and Measurement. Springer, pp. 699–716. http://dx.doi.org/10.1007/978-3-030-35798-6_12.

Kumar, A., Ramsankaran, R., Brocca, L., noz Arriola, F.M., 2021. A simple machine learning approach to model real-time streamflow using satellite inputs: Demonstration in a data scarce catchment. J. Hydrol. 595, 126046. http://dx.doi.org/10.1016/j.jhydrol.2021.126046.

Lin, K., Sheng, S., Zhou, Y., Liu, F., Li, Z., Chen, H., Xu, C.Y., Chen, J., Guo, S., 2020. The exploration of a temporal convolutional network combined with encoder–decoder framework for runoff forecasting. Hydrol. Res. 51, 1136–1149. http://dx.doi.org/10.2166/nh.2020.100.

Liu, Y., Duffy, K., Dy, J.G., Ganguly, A.R., 2023. Explainable deep learning for insights in el niño and river flows. Nature Commun. 14, 339. http://dx.doi.org/10.1038/s41467-023-35968-5.

Llasat, M.C., Barriendos, M., Barrera, A., Rigo, T., 2005. Floods in Catalonia (NE Spain) since the 14th century. Climatological and meteorological aspects from historical documentary sources and old instrumental records. J. Hydrol. 313, 32–47. http://dx.doi.org/10.1016/j.jhydrol.2005.02.004.

Mallakpour, I., Villarini, G., 2015. The changing nature of flooding across the central United States. Nature Clim. Change 5, 250–254. http://dx.doi.org/10.1038/nclimate2516.

Molnar, C., 2022. Interpretable Machine Learning, second ed. URL: https://christophm.github.io/interpretable-ml-book.

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, H.V., 2021. What role does hydrological science play in the age of machine learning? Water Resour. Res. 57, e2020WR028091. http://dx.doi.org/10.1029/2020WR028091.

Norsyuhada, W., Abdul Malek, M., Reba, M., Zaini, N., Najah, A.M., Sherif, M., Elshafie, A., 2022. River flow prediction based on improved machine learning method: Cuckoo search-artificial neural network. Appl. Water Sci. 13, 28. http://dx.doi.org/10.1007/s13201-022-01830-0.

Panegrossi, G., Casella, D., Sanò, P., Camplani, A., Battaglia, A., 2022. Recent advances and challenges in satellite-based snowfall detection and estimation. In: Michaelides, S. (Ed.), Precipitation Science. Elsevier, pp. 333–376. http://dx.doi.org/10.1016/B978-0-12-822973-6.00015-9, (Chapter 12).

Ribas Palom, A., 2007. Les inundacions a Girona. Col.lecció Patrimoni Cultural. Ajuntament de Girona i Institut d'Estudis Gironins.

Saint-Fleur, B.E., Allier, S., Lassara, E., Rivet, A., Artigue, G., Pistre, S., Johannet, A., 2023. Towards a better consideration of rainfall and hydrological spatial features by a deep neural network model to improve flash floods forecasting: case study on the gardon basin, France. Model. Earth Syst. Environ. 9, 3693–3708. http://dx.doi.org/10.1007/s40808-022-01650-w.

Sayad, Y.O., Mousannif, H., Al Moatassime, H., 2019. Predictive modeling of wildfires: A new dataset and machine learning approach. Fire Saf. 104, 130–146. http://dx.doi.org/10.1016/j.firesaf.2019.01.006.

Tayfur, G., Singh, V.P., Moramarco, T., Barbetta, S., 2018. Flood hydrograph prediction using machine learning methods. Water 10, http://dx.doi.org/10.3390/w10080968.

World Meteorological Organization, O.R.G., 2011. Manual on Flood Forecasting and Warning. Technical Report 1072, World Meteorological Organization.

Xu, Y., Hu, C., Wu, Q., Li, Z., Jian, S., Chen, Y., 2021. Application of temporal convolutional network for flood forecasting. Hydrol. Res. 52, 1455–1468. http://dx.doi.org/10.2166/nh.2021.021.

Zhang, J., Chen, X., Khan, A., Zhang, Y.kuan., Kuang, X., Liang, X., Taccari, M.L., Nuttall, J., 2021. Daily runoff forecasting by deep recursive neural network. J. Hydrol. 596, 126067. http://dx.doi.org/10.1016/j.jhydrol.2021.126067.