Are LSTM and conceptual rainfall-runoff models able to cope with limited training datasets under diverse hydrometeorological conditions?

Fadil Boodoo, Renaud Hostache, Nadia Skifa, Joris Guerin, Carole Delenne

Are LSTM and conceptual rainfall-runoff models able to cope with limited training datasets under diverse hydrometeorological conditions ?

Fadil Boodoo[1,2*], Renaud Hostache[4], Nadia Skifa[1], Joris Guerin[4], Carole Delenne[1,2,3]

[1*]HSM, Univ. Montpellier, CNRS, IRD, France.
[2]Lemon, Inria, Montpellier, France.
[3]currently at IUSTI, Aix Marseille Univ., CNRS, Marseilles, France.
[4]EspaceDev, UMR Espace-Dev, Univ. Montpellier, IRD, Univ. Réunion, Univ. Guyane, Univ. Nouvelle Calédonie, Univ. Antilles, UPVD, Montpellier, France.

*Corresponding author(s). E-mail(s): fadil.boodoo@gmail.com;
Contributing authors: renaud.hostache@ird.fr; N.Skifa2@newcastle.ac.uk; Joris.guerin@ird.fr; carole.delenne@univ-amu.fr;

Abstract

As climate change exacerbates variability and non-stationarity in rainfall patterns, it is crucial to assess the predictive capabilities of forecasting models. Previous researches on rainfall-runoff modeling have focused on the impact of training dataset size on Artificial Neural Networks (ANNs) results, with limited consideration of hydrometeorological diversity. This study first evaluates the influence of the training dataset length (1 to 15 years) on performance of a Long Short-Term Memory (LSTM) and a traditional conceptual model, Superflex, across 10 validation years. Next, training years are categorized based on hydrometeorological diversity (wetter, standard, drier). This clustering allows for experiments where models are trained on data from similar or different clusters, enhancing understanding of how data diversity, and therefore climate change, can affect model performance.

Results indicate that the LSTM model is highly sensitive to training length, showing poor performance with short datasets (below three years), reaches similar performance to Superflex around six training years on average, and overperforms with 15 years of training. Conversely, Superflex maintains rather constant performance levels regardless of the dataset length. LSTM model benefits from diverse

training data, achieving higher accuracy and reliability when trained on years with diverse hydrological typology. Despite their potential to outperform traditional models (with six or more training years on average), LSTM models are highly dependent on the quality and diversity of training data. In climate change scenarios, caution is needed when applying LSTM models to unfamiliar conditions, as their predictive accuracy may decline more rapidly than that of more traditional hydrological models.

Keywords: Machine Learning in Hydrology, Hydrometeorological Typopolgy, Artificial Intelligence, Univariate Analysis, Long Short-Term Memory (LSTM), Training Dataset, Rainfall-Runoff Modelling, Conceptual Models

1 Introduction

The history of rainfall-runoff modelling goes back 170 years, with the first attempts to predict discharge based on precipitation events using regression-type approaches [1, 2]. Over time, models have evolved by gradually integrating physically-based processes and concepts into the mathematical model formulations.

Physically-based models [3] are built on fundamental equations of fluid mechanics (mass and momentum balance) and aim to simulate the hydrological processes that occur within a catchment, such as infiltration, surface runoff and groundwater flow [4, 5]. These models require a detailed understanding of the catchment characteristics such as size, shape and topography, as well as the land use, soil type, vegetation cover and drainage network [6]. Although these models does not theoretically require heavy calibration work, their high computation demand make them not suitable yet for real-time applications at a large scale [7].

In 1933, Horton [8] proposed a simplified representation of the hydrologic cycle that focused on the processes of infiltration, evaporation and runoff, which is considered the basis for the development of conceptual models in hydrology. These were next further developed to address the lack of precise information on catchment topography, soils and geology. They provide a simplified representation of the catchment response to precipitation events, using a limited number of parameters to capture the essential processes happening in a catchment, such as storage and runoff processes. They are more suitable for real-time applications than the physically-based one while generally providing satisfying performances [9]. This makes them easier to calibrate and faster to use but they often require more data than physically-based models.

Since the 1990s, alternative approaches have been explored for rainfall-runoff modelling, such as data-based mechanistic modelling [10], as well as fully data-driven methods, including regression, fuzzy-based approaches or artificial neural networks (ANNs) [11, 12].

ANNs belong to machine learning (ML) models as they can learn the underlying relationships between precipitation and runoff from historical data. ML models can handle nonlinear relationships and high-dimensional data, and do not require a detailed prior knowledge of the catchment characteristics. ML models can exploit temporal patterns of atmospheric variables like precipitation, atmospheric pressure,

temperature, humidity, shortwave radiation or wind speed, to accurately predict river streamflows [13].

Among ANNs, Convolutional Neural Networks (CNNs) are the most commonly used in rainfall-runoff modelling [14]. RNNs were introduced in the mid-1980s [15] to address the limitations of traditional neural networks in processing sequential data, such as time-series data. They can handle inputs that have temporal dependencies by introducing feedback loops. This ability to retain information from previous steps makes RNNs particularly suitable for tasks such as speech recognition [16], language modelling [17] and time-series prediction [18].

One of the key challenges in RNN applications was the problem of vanishing gradients, which occurs when the gradients used in backpropagation become very small and make it difficult for the network to learn long-term dependencies [19]. This problem was partially addressed with the development of Long Short-Term Memory (LSTM) networks in 1997, which introduced a memory cell and gating mechanisms to control the flow of information through the network [20]. Despite these advances, RNNs and LSTM were still limited by their sequential nature and difficulty in modelling complex, long-term dependencies [21]. This changed with the introduction of the Attention Mechanism in 2017 [22], which allowed the so-called Transformers to focus on specific parts of the input sequence and improve their performance on tasks such as machine translation [23]. Although the use of Transformers for rainfall-runoff modelling recently shows first promising results [24], LSTMs are still more widely used [14] as they demonstrates very good performances especially for short-term predictions and even in areas with complex catchment behaviours, e.g. annual precipitation with high variance and strong gradient [25].

In the context of conceptual hydrological modelling, some studies [26–28] have investigated the minimum data length necessary to robustly calibrate conceptual models. It has been shown that satisfactory performance, as evidenced by a Nash-Sutcliffe Efficiency (NSE) metric surpassing 0.65 out of 1, can be achieved with less than one year of data using parsimonious model calibration [26, 28]. Moreover, a calibration period of at least five years was recommended [27] to capture most of the temporal hydrological variability whereas using more than 15 years did not result in significant improvement. For ANNs models, some papers have investigated the influence of data length on rainfall-runoff ML models. For example in [29], twenty catchments with different hydrological behaviours from the CAMELS-US dataset [30] were used to train an LSTM and a Feedforward Neural Network (FFNN) with a length of training data ranging from 3 to 15 years. Both models produced acceptable predictions (NSE greater than 0.65) with a minimum of 9 years of training data. Prediction carried out using both models improved with 12 years of data, and no significant improvement was observed beyond 15 years (NSE equal to 0.83). Remarkably, the LSTM model showed acceptable efficiency (NSE greater than 0.65) with a small 3-years dataset, outperforming the FFNN model, that required 9 years for comparable results.

In light of the current and anticipated effects of climate change, it is crucial to assess the robustness of rainfall-runoff models under increasingly variable climatic conditions [31]. The performance of LSTM and other machine learning models is highly dependent on the quality and representativeness of training data, often requiring large

datasets for accurate predictions [25]. However, historical data is often sparse and difficult to obtain for many catchments worldwide [32]. Climate change is expected to intensify the non-stationarity of environmental systems, creating conditions that significantly diverge from historical norms. This poses a challenge for models trained on past data, especially artificial neural networks (ANNs), which are sensitive to data quality and characteristics. Traditional models may show greater resilience to such shifts. Therefore, it is essential to evaluate model performance across diverse climate datasets, including training on years with varying hydrometeorological typology, to measure robustness and ensure reliable predictions in the face of changing climate scenarios. In this paper, the primary objective is to assess how an ANN data-driven model, such as LSTM, performs under different training regimes relatively to a more traditional and conceptual approach, like Superflex. This approach allows for a more comprehensive evaluation of each model's robustness and adaptability to evolving environmental conditions.

To enable more general conclusions, we selected three catchments from diverse regions in Europe and the USA, each with distinct sizes and characteristics. These catchments were chosen based on model's relative strong performance using 15 years of training. Specifically, the selected catchments are the Severn River at Saxons Lode (UK), the Fish River near Fort Kent (Maine, USA) and the Narraguagus River at Cherryfield (Maine, USA).

Although LSTM networks have been evaluated within the context of a single training basin. We acknowledge that our approach results in a limited dataset and does not align with the multi-basin methodology in Kratzert & al[33]. However, this was intentional, as our goal is to evaluate LSTM performance with limited datasets and specific training scenarios. While the multi-basin approach increases dataset size and captures patterns from extreme events, it can sometimes worsen local-scale predictions, suggesting single-basin training may be more effective in some cases[33]. Pre-training on many basins expands the model's training envelope, which is generally beneficial but conflicts with our objective to assess the model's ability to predict events outside its envelope. We hypothesize that future events, under climate change, will often lie outside any model's training envelope, even for models trained on large datasets. Thus, limiting the training dataset size is more appropriate for validating the model's extrapolation capabilities. Additionally, multi-basin training increases complexity and computational costs. Our analysis of hydrologic typology scenarios (wetter, drier, standard years) would not be feasible in a multi-basin context. For these reasons, we argue a mono-basin approach is methodologically and practically relevant for our study.

The paper is organized as follows: Section 2 (Materials and Methods) describes the study areas, data sources, the LSTM and Superflex models, as well as the experimental design set up to assess the impact of both the size and the hydrometeorological diversity of the training data set on model performance. Section 3 (Results and Discussion) presents the findings on how these factors affect model accuracy. Finally, Section 4 (Conclusion) provides an overview of the pros and cons of each model, discusses the broader impact of the results and highlights areas for future investigation.

2 Materials and Methods

2.1 Study areas

To generalize our findings across different basin sizes and climatic regions, we included basins with varied geographical and hydrological characteristics. The Severn Basin, situated in the temperate climate of the UK, spans 6,865 km^2, with data sourced from CAMELS-GB [32]. Meanwhile, the Fish River near Fort Kent and the Narraguagus River at Cherryfield, both in continental climate zones of the USA, cover drainage areas of 2,252.7 km^2 and 573.6 km^2, respectively, with data provided by CAMELS-US [30].

2.1.1 Severn catchement

The Severn River is the longest in Great Britain and runs from its source in the Welsh hills to the Bristol channel. The basin is mainly rural, with some urban settlements like Worcester, Tewkesbury and Evesham. Our study focuses on the Severn basin at Saxons Lode, located near the village of Ripple (Fig. 1). The water level at Saxons Lode typically range from 0.42 meters to 4.00 meters above mean sea level. Throughout the monitoring period, the river has stayed within this range approximately 90% of the time. However, on Sunday, July 22, 2007, at 3:15 pm, the river reached its highest recorded level of 5.93 meters since monitoring began [34].

The river can be divided into four sections: the Upper Severn in the Welsh Mountains, the Shropshire-Cheshire plain section between Welshpool and Ironbridge, the middle Severn from Ironbridge to Worcester, and the lower Severn in the valley of Gloucester. The Cambrian mountains, reaching up to 1,000 meters in altitude, experience annual precipitation of up to 2,500 mm. The Valley of Powis, which follows a southwest to northeast trough, contains a narrow floodplain. The channel slope is relatively steep in this area, with gravel and frequent braided sections. The part of the valley experience regularly floodings [35].

2.1.2 Fish River near Fort Kent catchment

The Saint John River, with the Fish River as one of its tributaries, is the eastern Canada's longest at 673 kilometers, flows from the Notre Dame Mountains near the Maine-Quebec border through western New Brunswick to the northwest shore of the Bay of Fundy, boasting a vast drainage basin of about 55,000 km2 [36]. Originating in the New England/Acadian forests of Maine and Quebec, it includes branches like the Southwest, Northwest and Baker branches, along with the Allagash River [37]. Noteworthy tributaries in the middle section include the Meduxnekeag River. Moving downstream, the river landscape transforms into a mix of lakes, islands, wetlands and a tidal estuary, with significant tributaries like the Nashwaak and Nerepis rivers. The Kennebecasis River forms a fjord near its mouth, leading to the Reversing Falls. Spring flooding is a recurrent issue, exacerbated by heavy rainfall, ice jams, high tides and rapid snowmelt, increasing sixfold compared to the average rate [38]. Floods, documented for over 300 years [39], affect areas such as Edmundston, Grand Falls, Perth-Andover, Hartland, Woodstock and particularly Fredericton and Saint John.

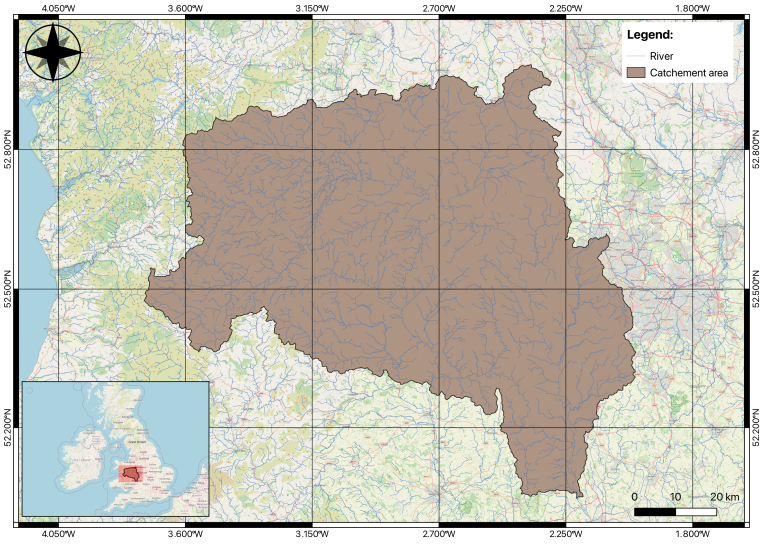

Fig. 1 Saxons Lode Gauging station, Riverstream, Severn Basin

Major flooding occurred in 1923, reaching 8 meters above normal winter levels. In 1936, high temperatures accelerated snowmelt, raising water levels to 8.9 meters, about 7.6 meters above summer levels. Similar conditions led to the same water level during the 1973 flood, and major flooding recurred in 2018 and 2019. Even if flooding has been less severe since 2019, it is expected to worsen due to climate change, with New Brunswick's average temperature predicted to rise by 5°C by 2100, accompanied by increased precipitation [40].

The Fish River, spanning 112.5 km [41], is situated in northern Maine, United States (Fig. 2). It serves as a tributary to the Saint John River, ultimately flowing into the Bay of Fundy in New Brunswick, Canada. Originating at the merging point of Fox Brook and Carr Pond Stream in Maine the river flows northward to Fish River Lake, then eastward to Portage Lake. Continuing its course, it proceeds northward through St. Froid Lake and Eagle Lake before reaching the Saint John River at Fort Kent. This latter stretch roughly parallels Maine State Route 11.

The watershed covers an area of 527 km^2, with an elevation of 705 m and a slope of 47.86 $m.km^{-1}$. The proportion of forested area within the watershed is less than 1% [42].

From 1980 to 2014, the watershed experiences an annual average runoff of 2.794 mm/day, with precipitation measuring 5.281 mm/day. Potential evapotranspiration during this period is recorded at 2.804 mm/day, with an average temperature of 12.779°C [42].

2.1.3 Narraguagus River at Cherryfield catchement

The Narraguagus River, flowing through Cherryfield, Maine, in the US, is a vital hydrological feature within the region. Originating from Eagle Lake at an elevation of 124 m, the river traverses a predominantly north-to-south course, draining an area

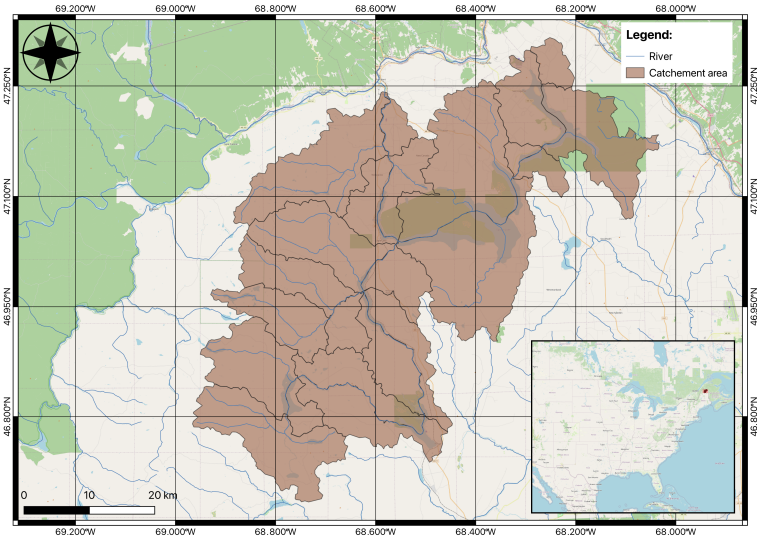

Fig. 2 Fish River near Fort Kent catchement

of approximately 2.6 km^2 [43]. The catchment area is characterized by a diverse array of hydrological and ecological features, including varied streamflow patterns and significant seasonal fluctuations. The river's discharge and gauge height have been meticulously monitored since 1948 [44], providing a comprehensive understanding of its long-term hydrological behavior. The river's flow regime is influenced by both natural and anthropogenic factors, including precipitation patterns, land use changes and water management practices. Monthly median streamflows, derived from extensive historical records, reveal distinct seasonal variations, with higher flows typically observed during spring snowmelt and lower flows during late summer and early autumn [43]. The Narraguagus River supports a rich biodiversity, serving as a critical habitat for various aquatic species, including the endangered Atlantic salmon [45]. Additionally, the river plays a significant role in the local economy, supporting recreational activities such as fishing and kayaking, and providing water resources for agricultural and domestic use.

A key feature of the river is the Cherryfield Dam, a rock shale timber dam built in 1965 to reduce flood risk [46]. Standing 25 feet high and spanning 500 feet in length [46], the dam prevents ice chunks from floating downriver and causing floods in the town [47]. However, it has also posed a significant barrier to the migration of endangered Atlantic salmon and other sea-run fish, preventing them from reaching their spawning grounds. In 2024, the Downeast Salmon Federation has announced plans to replace the existing ice retention dam with a nature-like fishway. This new design aims to maintain the current elevation and pond behind the dam while allowing fish to travel freely upstream [47].

The watershed (Fig. 3) encompasses an area of 573.6 km^2, characterized by an elevation of 93 m and a slope of 18 m.km^{-1} [42]. Between 1980 and 2014, the watershed exhibits an average annual runoff of 2.145 mm/day, with precipitation measuring

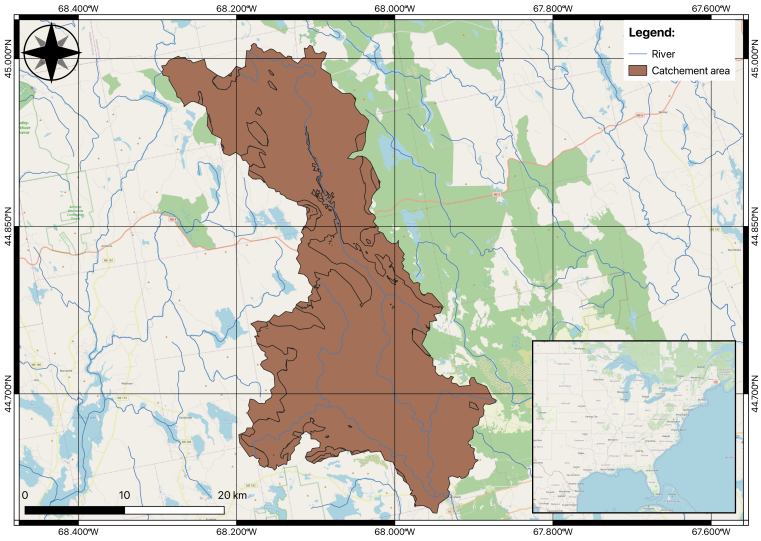

Fig. 3 Narraguagus River at Cherryfield, Maine catchement

3.6 mm/day. Potential evapotranspiration during this period averages 2.1 mm/day, alongside a mean temperature of 6.5°C [42].

2.1.4 Comparative analysis of basin climate and flow regimes

Based on a detailed 35-year analysis (1980–2014), we compared the hydroclimatic regimes of the Severn, Fish and Narraguagus basins.

Temperature analyses reveals significant spatial heterogeneity among the basins (Fig. 5). The Fish and Narraguagus basins exhibit a higher temperature variability, with summer temperatures reaching approximately 20°C and winter temperatures dropping to around -10°C. In contrast, Severn basin displays more moderate temperature ranges, with winter temperatures around 0°C and summer temperatures reaching approximately 15°C. The mean annual temperatures (Fig. 4) show that Fish basin has the highest mean temperature, followed by Narraguagus, while Severn exhibits the lowest mean temperature. All three basins demonstrated clear seasonal patterns, with peak temperatures occurring during summer months (July-August) as shown in the example of year 1997 (Fig. 5).

Precipitation dynamics reveal distinct characteristics across the basins (Fig. 6). The Narraguagus basin exhibits the highest mean daily precipitation with a variability around 7.5 mm/day), while Severn and Fish basins show moderately lower values with a variability around 4.1 and 5.2 mm/day respectively. All three basins demonstrate day-to-day variability in precipitation to some extent depending on the basin, as indicated by the large standard deviation for Narraguagus, and the smaller one for Severn. This variability suggests complex and different atmospheric processes affecting precipitation amounts between these watersheds.

PET patterns exhibit strong seasonal variability across all basins, closely following temperature trends (Fig. 8). Peak PET values occur during the summer months

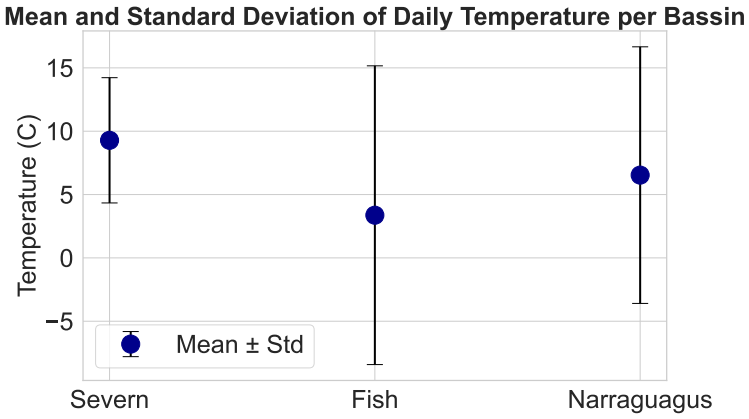

Fig. 4 Mean temperature and standard deviation for Severn, Fish and Narraguagus basins.

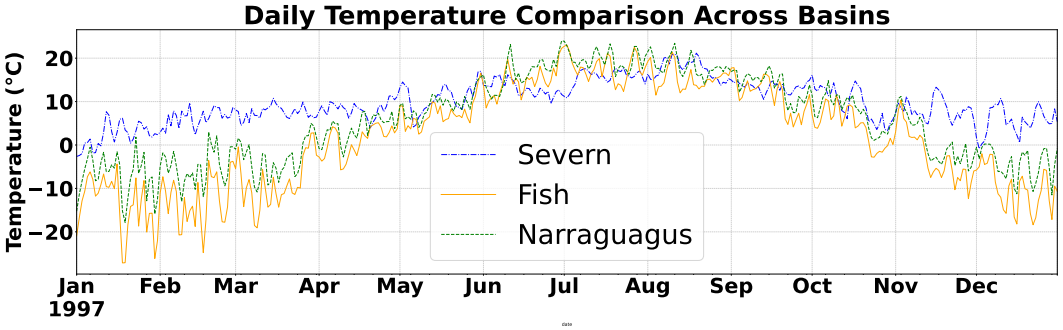

Fig. 5 Seasonal variation of temperature across Severn, Fish and Narraguagus basins during 1997.

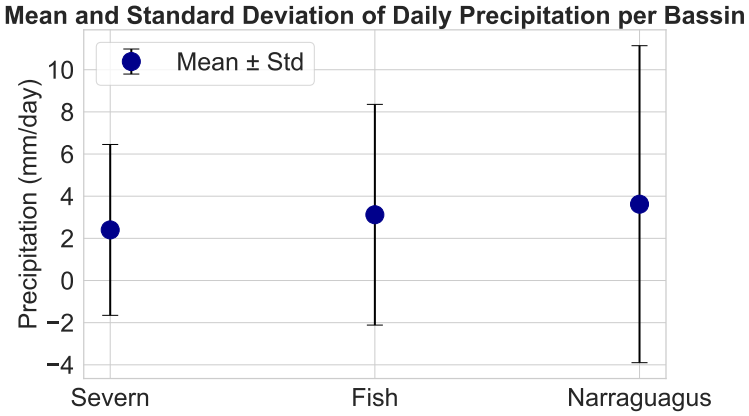

Fig. 6 Mean precipitation and standard deviation for Severn, Fish, and Narraguagus basins.

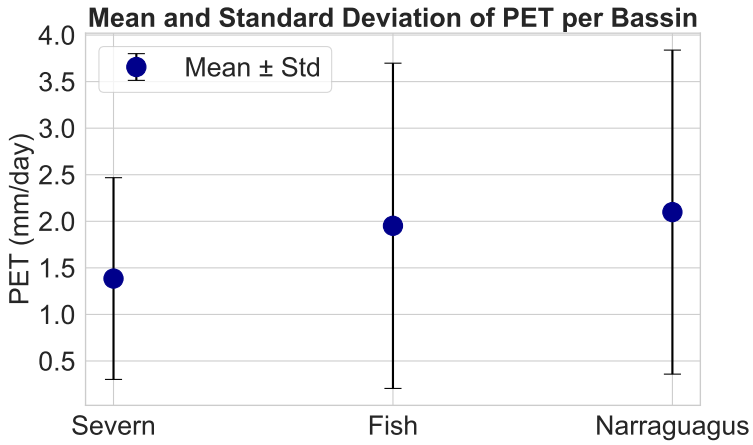

Fig. 7 Mean potential evapotranspiration and standard deviation for Severn, Fish, and Narraguagus basins.

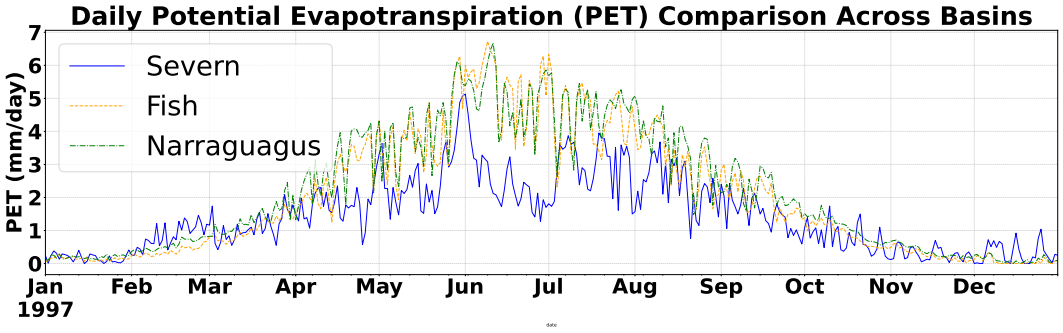

Fig. 8 Seasonal variation of potential evapotranspiration across Severn, Fish, and Narraguagus basins during 1997

(June–August), reaching 6–7 mm/day. The Severn basin records the lowest mean PET (≈2 mm/day). In contrast, the Fish and Narraguagus basins display slightly higher mean PET values (Fig. 7), consistent with their climates and differing hydrological regimes.

Streamflow analysis reveals notable differences in hydrological responses among the basins. The Severn basin exhibits the highest mean streamflow (90 m^3/s), far exceeding the Fish basin (45 m^3/s) and the Narraguagus basin (15 m^3/s) (Fig.9). The Fish basin's mean streamflow is approximately three times greater than that of the Narraguagus basin, highlighting a significant disparity between these smaller basins. These differences can be attributed to variations in drainage area size and runoff generation processes. Despite these contrasts, all basins show considerable streamflow variability, with the Severn displaying the largest standard deviation.

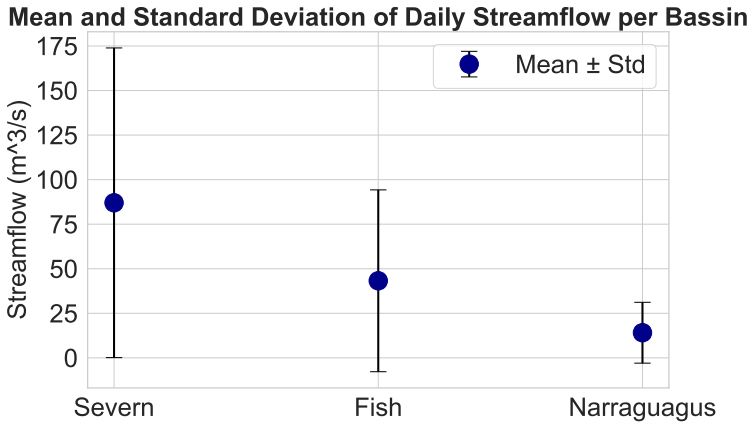

Fig. 9 Mean streamflow and standard deviation for Severn, Fish, and Narraguagus basins

2.1.5 The CAMELS-GB dataset

CAMELS-GB (Catchment Attributes and MEteorology for Large-sample Studies) [32] builds upon existing data for Great Britain catchments to create a comprehensive, unified, well-documented and up-to-date dataset. CAMELS-GB provides daily time series data of meteorological and hydrological variables in each gauged catchment. These include streamflow, rainfall, potential evapotranspiration, temperature, short-wave radiation, long-wave radiation, specific humidity and wind speed. The dataset covers a 45-year period, from October 1, 1970 to September 30, 2015, including major hydroclimatic events such as the 1976 drought and the 2007 floods events that affected the River Severn [48].

In CAMELS-GB, daily meteorological time series are derived from the Climate Hydrology and Ecology research Support System (CHESS) meteorology dataset. The dataset encompasses daily 1 km^2 gridded estimates covering Great Britain from January 1st, 1961, to December 31st, 2015. It features various meteorological variables originating from observational data. CHESS-met stems from the MORECS (Meteorological Office Rainfall and Evaporation Calculation System) which is derived from observational data and has a 40 km resolution. This dataset is the result of interpolating daily station data. To achieve the 1 km resolution in CHESS-met, MORECS variables are downscaled and adjusted for local topography considering factors like lapse rates, modeled wind speeds, and established relationships (see [32] for detail). Directly from MORECS, CHESS-met's air temperature and wind speed are downscaled. Specific humidity is determined from MORECS vapour pressure, downward short-wave radiation from its sunshine hours, and long-wave radiation using the downscaled temperature, vapour pressure and sunshine duration. Daily streamflow data are collected by various measuring authorities such as the Environment Agency (EA), Natural Resources Wales (NRW), Scottish Environmental Protection Agency (SEPA) or the UK National River Flow Archive (NRFA). The CAMELS-GB dataset indicates, for the catchment area upstream of Saxons Lode, a mean daily precipitation of 2.40 $mm.day^{-1}$, a mean daily potential evapotranspiration

(Penman-Monteith equation without interception correction) of 1.39 mm.day^{-1}, a dryness (calculated as the ratio of the mean daily potential evapotranspiration to the mean daily precipitation) of 0.58 in the area, and a mean daily outflow discharge of 1.10 mm.day^{-1}.

2.1.6 The CAMELS-US dataset

The CAMELS-US [30] data set provides attributes for 671 catchments within the contiguous United States (CONUS), specifically chosen as they have limited human infrastructure. These attributes includes six main categories: topography, climate, streamflow, land cover, soil and geology. Complementing daily meteorological forcing and streamflow time series from [49], the data set integrates diverse sources to provide a comprehensive understanding of each catchment. Meteorological forcing data are drawn from three datasets – Daymet, Maurer and NLDAS – with Daymet data specifically used for computing climatic indices. Streamflow measurements, sourced from the United States Geological Survey (USGS), cover a continuous 20-year period from 1990 to 2009. To accommodate variations in spatial resolution across data sources, upscaling is performed using arithmetic means. Noteworthy geological classes in the CAMELS-US catchments include siliciclastic sedimentary rocks, unconsolidated sediments, metamorphic rocks and carbonate sedimentary rocks. Temporal coverage for climate indices and hydrological signatures extends from 1 October 1989 to 30 September 2009. The CAMELS-US data set, an extension of the N15 data set [50], consolidates both hydrometeorological time series and catchment attributes, enabling comprehensive analyses. This dataset serves as a valuable resource for large-sample studies in hydrology and related disciplines, facilitating nuanced research into catchment behavior and environmental processes across the CONUS.

2.2 The Long Short-Term Memory (LSTM) model

Artificial Neural Networks (ANNs) represent a cornerstone in the realm of machine learning, serving as computational frameworks mimiking the neural configurations found in biological brains. Comprised of interconnected nodes, cells or neurons, these networks are organized into distinct layers: the input layer for receiving features, one or more hidden layers for computation and the output layer for delivering predictions. Each connection between neurons is associated with a numerical weight, which is adaptively updated during the training phase through optimization algorithms such as Gradient Descent [51]. ANNs employ activation functions like ReLU, Tanh or Sigmoid to introduce non-linearity into the system, enhancing their capacity to model complex relationships. The learning process in ANNs involves a cycle of forward propagation [52] to make predictions, loss calculation [53] to measure deviations from the actual target, and backpropagation [54] to adjust the network's weights. Recurrent Neural Networks (RNNs) occupy a specialised niche within the broader landscape of Artificial Neural Networks, designed explicitly for handling sequential and time-series data. Unlike traditional ANNs (feedforward networks for example), RNNs include loops that allow information to persist, thereby enabling the network to maintain a "memory" of previous inputs as shown in Fig. 10. This unique architecture makes

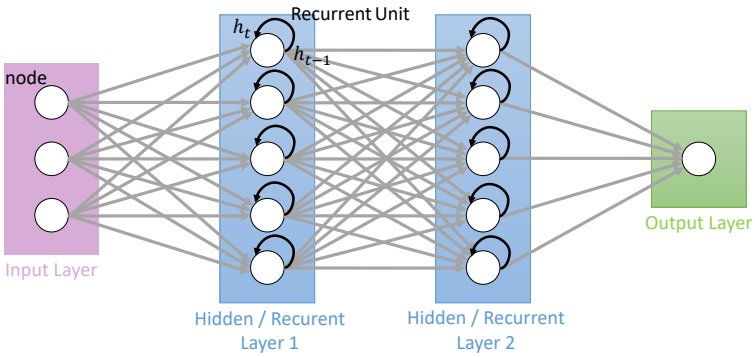

Fig. 10 Structure of a Recurrent Neural Network (RNN). The input layer is on the left, and the output layer is on the right. Nodes in the recurrent layers are connected not only to the next layer but also to themselves.

RNNs particularly well-suited for tasks that require the understanding of temporal dynamics, such as catchment modelling.

We adopt here a LSTM model for rainfall-runoff modelling, which is a variant of recurrent neural networks (RNNs). In a traditional RNN cell, only one internal state $\mathbf{h}_t$ is considered. It is computed at each time step t using equation (1) involving activation function, parameterisable weight matrices ($\mathbf{W}$ and $\mathbf{U}$) and a parameterisable bias vector ($\mathbf{b}$ and $\mathbf{c}$). The time step t represents an individual moment within a designated sequence that serves as an input. In an RNN model with a sequence length of n, the model considers the previous n time steps to make predictions for the $n+1$ time step or beyond. The initial hidden state is typically set to the null vector and its length is a user-defined hyperparameter of the network.

$$\begin{aligned} \mathbf{a}_t &= \mathbf{b} + \mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{x}_t \\ \mathbf{h}_t &= \tanh(\mathbf{a}_t) \\ \mathbf{o}_t &= \mathbf{c} + \mathbf{V}\mathbf{h}_t \\ \mathbf{y}_t &= \mathrm{softmax}(\mathbf{o}_t) \end{aligned} \tag{1}$$

where $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ are the weight matrices, $\mathbf{b}$ and $\mathbf{c}$ the bias vectors, $\mathbf{x}_t$ is the input data, $\mathbf{y}_t$ the predicted output value and softmax is the Softmax function [55].

The LSTM architecture is very similar to traditional RNNs, with the exception of the internal operations within the recurrent cell. The LSTM cell operations are presented in Fig. 11. Instead of a single internal state (as in traditionnal RNNs), LSTMs include several internal states that allow for the retention and utilization of information over long sequences. These memory cells are equipped with specialized gates: the input, forget and output gates. These are crucial for regulating the flow of information within the memory cell.

The LSTM model operates in the following manner: in the first step, it integrates new input data $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$ (which is initialized to 0 at t=0) and introduces them into the first gate of the cell, namely the forget gate. The

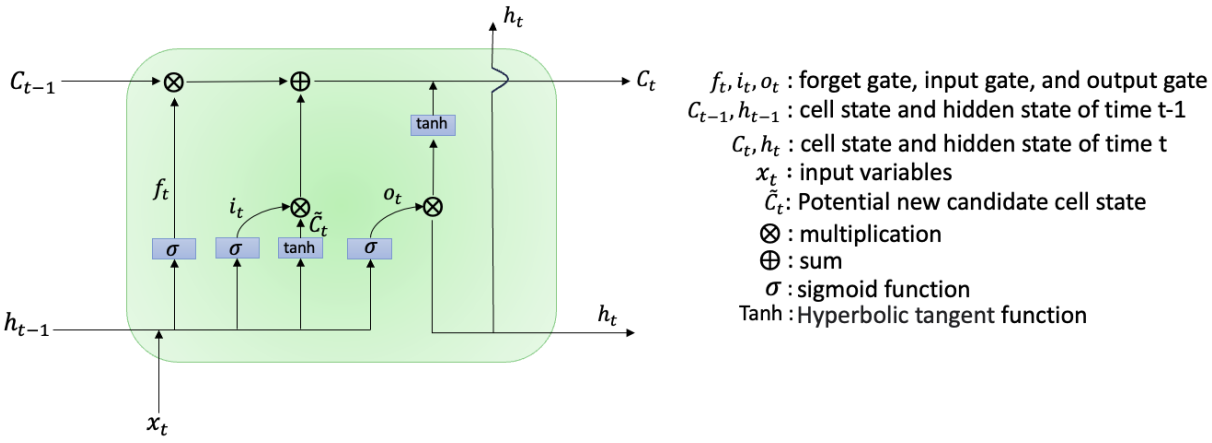

Fig. 11 Structure of an LSTM Cell

forget gate determines how much information from the previous time step is forgotten or kept according to Eq. (2) where the sigmoid activation function σ returns values between 0 and 1. The closer the value is to 0, the more the information is forgotten. Conversely, the closer the value is to 1, the more the information is retained.

$$f_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \tag{2}$$

where f_t is the forget gate function at time t, defined as the logistic sigmoid function σ, that takes as input the dot matrix product of the weight matrix $\mathbf{W}_f$ with the new input data $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$ plus the bias vector $\mathbf{b}_f$.

In the second step of the cell operation, the input gate $\mathbf{i}_t$ controls the flow of new information coming into the memory cell at time t. It takes as input $\mathbf{x}_t$ and $\mathbf{h}_{t-1}$ and decides how much new information is allowed into the memory cell according to Eq. (3). The input gate also uses the sigmoid activation function σ to indicate the importance of the incoming information.

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \tag{3}$$

where $\mathbf{i}_t$ is the input gate at time step t, $\mathbf{W}_i$ represents the weight matrices, $\mathbf{b}_i$ is the bias vector.

Moreover, the memory cell has also an internal cell state $\mathbf{C}_t$ that allows to retain information over time and helps address the vanishing gradient problem [56]. The cell state $\mathbf{C}_t$ is updated using the forget gate $\mathbf{f}_t$ and the input gate $\mathbf{i}_t$. The forget gate determines how much information of the previous cell state $\mathbf{C}_{t-1}$ is forgotten, while the input gate determines how much information from the potential new candidate cell state $\widetilde{\mathbf{C}}_t$ is updated with new information according to Eq. (4).

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \widetilde{\mathbf{C}}_t \tag{4}$$

where $\odot$ denotes element-wise multiplication, and $\widetilde{\mathbf{C}}_t$ represents a potential new candidate cell state which is calculated as follows:

$$\widetilde{\mathbf{C}}_t = \tanh\left(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c\right) \tag{5}$$

Finally, the output gate $\mathbf{O}_t$ controls the amount of information that should be outputted from the memory cell. It takes the current input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$ as inputs and uses the σ function to determine how much of the memory cell's content should be passed to the next time step according to Eq. (6):

$$\mathbf{O}_t = \sigma(\mathbf{W}_O[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_O) \tag{6}$$

The output from the memory cell is the combination of the output gate's activation $\mathbf{O}_t$ and the tanh activation of the cell state $\mathbf{C}_t$ as indicated in Eq. (7). This output is passed as the hidden state to the next time step $\mathbf{h}_t$ and is also used to make predictions or further processing in the network.

$$\mathbf{h}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t) \tag{7}$$

Similar to other neural networks, the LSTM architecture allows for flexibility in choosing the number of cell in a layer, that is generally defined as a hyperparameter. Furthermore, multiple LSTM layers can be added on top of each other. The final output prediction of the discharge is computed by connecting the output of the last LSTM layer at the last time step $\mathbf{h}_n$ to a single output cell through a traditional dense layer using the following equation:

$$\mathbf{y} = \mathbf{W}_d\mathbf{h}_n + \mathbf{b}_d \tag{8}$$

where $\mathbf{y}$ represents the discharge prediction $Q(t)$, $\mathbf{h}_n$ is calculated using the output of the last LSTM layer at the last time step, as derived from equation Eq.(7). This output is multiplied by the weight matrix $\mathbf{W}_d$ of the dense layer and added to the bias term $\mathbf{b}_d$.

2.3 LSTM model training

In ML, the dataset is commonly divided into features and labels. The features are the attributes of the dataset that are used as input of the ML model. In our study, these are the daily averaged meteorological time series: precipitation [mm.d^{-1}], temperature (°C), wind speed [m.s^{-1}], specific humidity [g.kg^{-1}], downward short-wave radiation [W.m^{-2}], long-wave radiation [W.m^{-2}]. The labels represent the output of the ML model. In our study, these correspond to the streamflow time-series [m^{3}s^{-1}]. In our study, as usual, we split our dataset in two parts: a training set and a validation set. The former is used to train the model by optimising its weights. The latter is used to validate the previously optimized model by comparing the streamflow predicted by the model with the observation.

During the training of LSTMs, one iteration step for the optimization typically operates on a subset of the training data, referred to as a batch or mini-batch. Each sample within a batch consists of the discharge value of a specific day and the meteorological input from a given number of previous days (i.e. the sequence length, measured in days, that can be adjusted as a hyperparameter). Since the streamflow at a particular time step only depends on the meteorological inputs from the past days, the

samples within a batch can be randomly selected and do not need to be arranged chronologically. Indeed, as shown in [57], random samples within a batch can facilitate faster convergence. This approach differs from traditional hydrological model calibration, where typically all calibration data is used at each iteration, considering all pairs of simulated and observed streamflow for model evaluation.

At each training iteration, the loss function is computed as the average mean-squared error [58] (MSE) between the simulated and observed streamflow of these samples. Since the neural network equations are differentiable, the gradients of the loss function (MSE) can be explicitly calculated as a function of the parameters. This property, known as the backpropagation step, contributes to minimizing the overall loss and allows parameters (i.e., weights and biases) to be updated. These training iterations are referred to as an epoch. An epoch is one complete forward and backward pass of all training samples to update the model parameters. Over multiple training epochs, the LSTM gradually learns the complete rainfall-discharge relationship from scratch, resulting in improved representation of the streamflow dynamics at each epoch. To optimize LSTM performance, multiple models are often trained [59] with varying weight initializations based on different random seeds. Predictions are then combined by selecting the most recurring output from these models. This approach bolsters model stability and diminishes uncertainty.

2.4 Calibration of the hyperparameters

The LSTM model requires the configuration of six hyperparameters: output dropout, learning rate, hidden size, epoch, batch size and sequential length. In our study, the first four are set according to literature as explained hereafter, while the last two are calibrated.

The dropout is a regularization technique commonly used in neural networks, to prevent overfitting. It involves randomly dropping out (setting to zero) a certain proportion of neurons in a layer during each training iteration. The learning rate determines the step size at which the model's parameters are updated during the training process. It controls the magnitude of adjustments made to the model's weights and biases based on the computed gradients. The output dropout and learning rate are set empirically [20]. In our study, the output dropout was set to 0.4 as suggested in [25] and the learning rate was set to: 1e-2 from 0 to 30 epoch, 5e-3 from 30 to 40 epoch and 1e-3 from 40 to 50 epoch.

The batch size (i.e. the number of training samples processed in one iteration) was set via trials and errors to 300. To stop the model training process, we set the number of epochs to 50 as the MSE reaches a plateau around 50 epochs in all tests we carried out.

The hidden size corresponds to the number of cells in each hidden layers whereas a sequence length describes the number of time steps that are considered in a single input sequence. For example, with a sequence length of 10, the model considers 10 previous time steps of data to generate a prediction for the 11th time step or beyond. To tune these two hyperparameters, we propose to carry out a series of simulations with varying hidden size and sequence length. To reduce uncertainty and increase the robustness of the results, we run 50 simulations for each parameter set, with

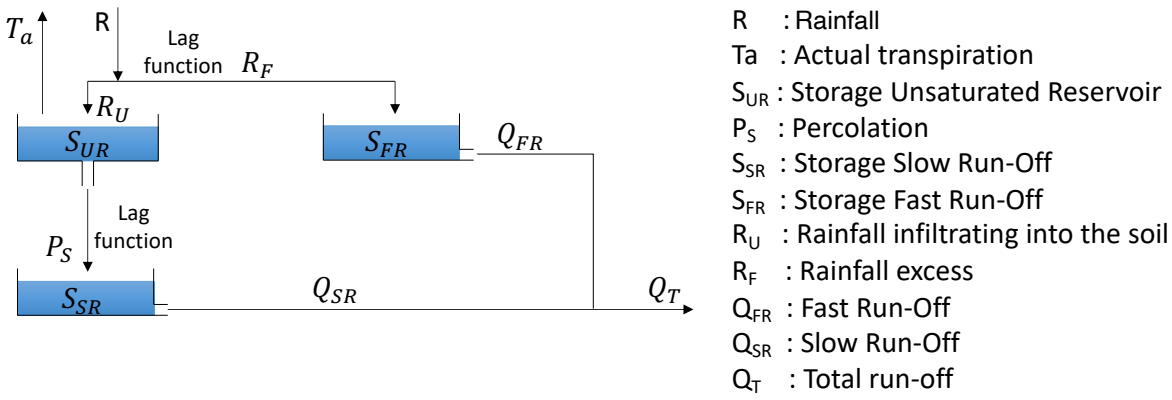

Fig. 12 Superflex scheme

each simulation involving the initialization of weights with different random seeds. To aggregate the findings from the 50 simulations, we selected the mode of the NSE values as a representative indicator of overall performance. We then compiled a table documenting the NSE associated with this mode for each hyperparameter set.

2.5 Superflex model

In this study, we used as a conceptual hydrological model a lumped version of Superflex hydrological model [60].

As illustrated in Fig. 12, Superflex is structurally compartmentalised into three reservoirs:

- An unsaturated soil reservoir, refered to as S_{UR}, models the unsaturated root zone. It collects part of the rainfall R denoted as R_u and calculated using Eq. (9a):

$$R_u = (1 - C_r)R \tag{9a}$$

$$C_r = \frac{1}{1 + \exp\left(\frac{\frac{S_{\mathrm{UR}}}{S_c} + 0.5}{\beta}\right)} \tag{9b}$$

  where C_r represents the runoff coefficient, $S_{\max}$ and β represent respectively the maximum S_{UR} storage parameter and the runoff shape parameter.
- A fast response reservoir, S_{FR}, simulates the fast component of water propagation through the basin that often dominates during flood events. It collects the remaining part of the rainfall R denoted as R_F, and calculated using Eq. (10):

$$R_f = C_r R \tag{10}$$

- A slow response reservoir S_{SR} that simulates the slow component of water propagation through the basin that controls the baseflow. It is fed by the percolation of

water P_S from the root zone reservoir, calculated using Eq. (11)

$$P_S = P_{\max} \frac{S_{\mathrm{UR}}}{S_{\max}} \tag{11}$$

where $P_{\max}$ represents the limit for percolation.

The model computes the total runoff by combining both the fast Q_{FR} and slow Q_{SR} runoff components, calculated by the following equation Eq. (12a):

$$Q_{\mathrm{FR}} = \frac{S_{\mathrm{FR}}}{K_{\mathrm{FR}}} \tag{12a}$$

$$Q_{\mathrm{SR}} = \frac{S_{\mathrm{SR}}}{K_{\mathrm{SR}}} \tag{12b}$$

where K_{FR} and K_{SR} represent the fast and slow runoff time scale respectively.

To better simulate the water transfer time across the basin, two triangular lag functions are applied on the fast and slow reservoirs. This ensures a more realistic representation of delayed hydrological responses in the system.

2.6 Superflex model calibration

To calibrate the superflex model, we also used CAMELS-GB dataset as for the LSTM model. More precisely, we restricted the input to two specific variables: precipitation [mm.d^{-1}] and potential evapotranspiration [mm.d^{-1}]. Superflex uses specific input variables, unlike the LSTM where various input variables can be used. We therefore assume that these two variables effectively capture the meteorological dynamics reflected by other variables that the LSTM leverages. The streamflow, model's output variable, remains identical to the LSTM model.

As for the LSTM model, the dataset is split into two parts: the calibration period and the validation period. The calibration period is used to determine the optimal parameters, while the validation period is used to evaluate the model performance. In our study, the calibration period for the Superflex model corresponds to the training period used for the LSTM model, and similarly, the validation period is the same for both models. It is important to note that Superflex model requires a two-year warm-up period prior to either the calibration or the validation phase.

The Superflex model is calibrated through a Monte Carlo method [61] involving all eight parameters. Within defined behavioral ranges for each parameter, random sets are generated iteratively. A Python script automates the generation of these random values at specified intervals, which are then input into the Superflex model to simulate discharge time series. These simulated outputs are compared to actual measurements from gauging stations on the same day, and the Nash-Sutcliffe Efficiency (NSE) is calculated accordingly (see equation 13 below). After running 10,000 simulations, the parameter set with the highest NSE is identified as optimal.

2.7 Evaluation metrics

To evaluate model performance, we use the Nash-Sutcliffe efficiency [62] (Eq. 13), a widely used metric in hydrology. NSE quantifies the proportion of variance in observed streamflow time series explained by a model. A higher NSE value (i.e. approaching 1) indicates a better agreement between model prediction and observation. Conversely, an NSE value of 0 indicates that the model's performance is equivalent to the mean observed discharge, while a negative NSE value suggests that the mean observed discharge is a better estimate than the model result. It is important to note that extreme values have a significant impact on the NSE, as the squared difference between predictions and measurements amplifies their influence on the metric.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(Q_{\mathrm{obs}_i} - Q_{\mathrm{pred}_i})^2}{\sum_{i=1}^{n}(Q_{\mathrm{obs}_i} - \overline{Q_{\mathrm{obs}}})^2} \tag{13}$$

where Q_{obs_i} and Q_{pred_i} are respectively the observed and predicted streamflow, n is the number of data available for comparison, and $\overline{Q_{\mathrm{obs}_i}}$ is the mean of observed streamflow.

We also used the Kling–Gupta efficiency (KGE) [63] (Eq. 14) that serves as a pivotal goodness-of-fit indicator in the hydrological sciences, essential for comparing simulations with observations. Developed by hydrologic scientists Harald Kling and Hoshin Vijai Gupta, its inception aimed to refine prevalent metrics like the coefficient of determination and the Nash–Sutcliffe model efficiency coefficient. By integrating multiple aspects of model performance, including correlation, bias, and variability, KGE provides a comprehensive evaluation tool, facilitating more nuanced and accurate assessments in hydrological modeling endeavors.

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{Q_{\mathrm{pred}}}}{\sigma_{Q_{\mathrm{obs}}}} - 1\right)^2 + \left(\frac{\overline{Q_{\mathrm{pred}}}}{\overline{Q_{\mathrm{obs}}}} - 1\right)^2} \tag{14}$$

where r is the is the Pearson correlation coefficient, $\overline{Q_{\mathrm{pred}}}$ and $\overline{Q_{\mathrm{obs}}}$ is the mean of the simulated and observed time series respectively, and $\sigma_{Q_{\mathrm{pred}}}$ and $\sigma_{Q_{\mathrm{obs}}}$ the standard deviation of simulated and observed data respectively.

The Mean Squared Error (MSE) [64] (Eq. 15) was also used as evaluation metric, it is a widely used metric in various fields to evaluate predictive models. It measures the average squared difference between predicted and actual values, providing a straightforward assessment of model accuracy. While lower MSE values indicate better fit, it does not offer insights into error direction or distribution. Despite its simplicity and popularity, MSE can be sensitive to outliers.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Q_{\mathrm{pred}_i} - Q_{\mathrm{obs}_i})^2 \tag{15}$$

where Q_{obs_i} and Q_{pred_i} are respectively the observed and predicted streamflow, n is the number of data available for comparison.

However, we primarily present the results using NSE, as the different metrics yield similar conclusions in the experiments.

2.8 Experimental design

We trained the models over various training duration, specifically 1, 3, 6, 9, 12 and 15 years, within the training period ranging from 1976 to 2004. We then evaluate the model performance for each individual year within the validation period, between 2005 and 2014, for each catchment separately.

To investigate the impact of hydrometeorological diversity and non-stationary factors on the performance of both LSTM and Superflex models, we initiated our investigation by classifying the years based on their respective hydrometeorological diversity for each catchment separately. For each year, we computed the annual mean and standard deviation of daily rainfall, and then applied a k-means clustering technique [65, 66] to these values. This three clusters are meant to categorize respectively drier, standard and wetter years. We then tested different scenarios by varying the number of training years and selecting these training years from the same cluster or different clusters:

- Scenario 1 (S1): we trained the models using three consecutive years (referred to as the training years) from the dataset, repeating the experiment for each set of three year from 1976 to 2004. We then evaluated the model's performance over each individual year (referred to as the validation year) within the dataset, from 2005 to 2014.
- Scenario 2 (S2): we trained the models in each catchment by randomly choosing 10 sets of three years from the same cluster for each cluster and evaluate their performance over each validation year.
- Scenario 3 (S3): we trained the models by randomly selecting 10 sets of three years from three different clusters with the same evaluation over each individual year within the validation phase of the dataset, and within the same catchment.

Throughout the experiments, to reduce uncertainty and increase the robustness of the experiments, we run 30 simulations for each experiments (same training dataset), with each simulation involving the initialization of weights with different random seeds. To combine the results of the 30 model run, we select the mode of the NSE values as a representative indicator of overall performance.

Finally, we create a heatmap where each cell's color and value indicate the NSE attained by both models for every pair of training and validation years. This visual representation facilitates the observation of fluctuations and trends in NSE values across varying combinations of training and validation years. To examine the influence of hydrometeorological diversity on model performance, the years listed on the heatmap's axes are reordered in an ascending fashion, based on their corresponding annual mean precipitation measurements.

3 Results and discussion

3.1 Software

Python 3.7 is the selected programming language for our research [67]. The LSTM model is provided by the Neuralhydrology library [68]. For data postprocessing

Sequence length \ Hidden size	32	64	128	256	512
30	0.849	0.841	0.854	0.856	0.867
90	0.882	0.882	0.883	0.885	0.900
180	0.892	0.903	0.898	0.899	0.916
270	0.897	0.895	0.896	0.906	0.908
360	0.901	0.903	0.896	0.898	0.914

Table 1 NSE obtained for the validation period for 2 hyperparameters: i) Sequence length representing the number of past data used to predict the next outcome and ii) hidden size represents the number of cells in each hidden layer.

and general data management tasks, we relied on Numpy [69], Pandas [70] and Scikit-Learn [71] libraries. Additionally, all figures presented here are created using Matplotlib [72] and Seaborn [73]. The code was run on two high performance computing (HPC) systems: Grid5000 [74] and MESO@LR-Platform at the University of Montpellier.

3.2 Hyperparameters: hidden size and sequence length

The LSTM model requires the adjustment of various hyperparameters, usually determined by empirical methods. However, for hidden size and sequence length, we carried out several model runs by systematically varying the values of these two hyperparameters to find the best fit. For the hidden size, we tested a range of values that are commonly used from 32 to 512 cells. For the sequence length, we test a range of values from 30 to 360 days. The training period was established from August 1, 1995, to July 31, 2005. The validation period extended from August 1, 2005, to December 31, 2010, and the results for this period are presented in Table 1. This shows that all sets of hyperparameters (sequential length, hidden size) yield good result with NSE greater than 0.8. We also looked at the computation time using some of these set of hyperparameters, we found for example that, while keeping the same number of cells at 64, the computation time is 5 minutes for 90 days, 8 minutes for 180 days and 13.5 minutes for 360 days, on a node with dual Intel Xeon E5-2680 v4 CPUs (28 cores, 2.4 GHz, 120 GB RAM). We therefore decided to set the sequential length at 90 days and hidden size at 64 cells, as this provides good results while maintaining a reasonable computation time.

3.3 Impact of training data size on model performance

To investigate the effect of training data size on model performance, both the LSTM and Superflex models were initially trained on each year of the training phase and then evaluated for each year of the validation phase, for each catchment. Fig. 13 and Fig. 14, shows the heatmap of the NSEs obtained from this experiment using LSTM and Superflex respectively for the Severn catchment. The x-axis of this heatmap specifies the training years, while the y-axis represents the validation years. The colour scheme in the heatmap varies from green, indicating a high NSE approaching 1 and

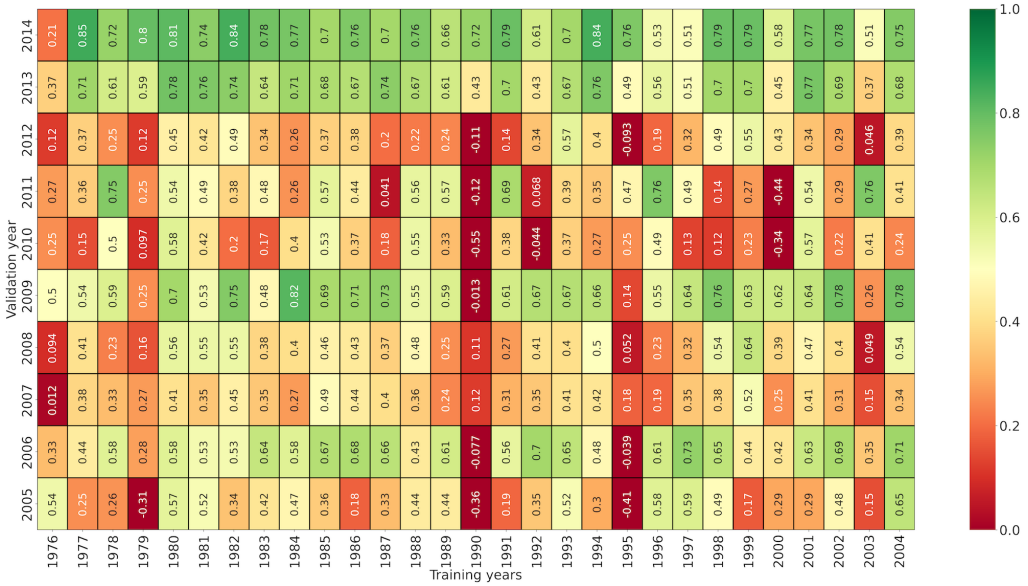

Fig. 13 Heatmap of NSE values obtained with one year of train/calibration data for LSTM model for the Severn catchment. Training years are shown in abscissa and validation years in ordinate.

representing perfect model performance, through yellow, indicating an NSE approaching 0.5 and reflecting moderate performance), to red, representing an NSE approching 0 and indicating poor model performance.

At first glance, the Superflex heatmap exhibits significantly greener pixels than that of the LSTM model. This shows a systematically better performance of the Superflex model compared to the LSTM model when trained using only one year of data. Additionally, the performance of the LSTM model exhibits more variability across the validation years, as indicated by the wider colour range in the heatmap, when compared to the Superflex model.

Next, both models were trained using the same approach but with varying numbers of training years – 3, 6, 9, 12, and 15 years – selected from the training period ranging from 1976 to 2004. The LSTM model experiments were repeated 30 times. Fig. 15 shows the average NSE over the training years for each validation year obtained by Superflex (dotted lines) and LSTM (continuous lines) models for the different numbers of training years. Each individual graph corresponds to a fixed number of training years.

Fig. 15 not only reinforces the observations made with the heatmaps in Fig. 13 and Fig. 14 for a one-year training, but also provides insight into how both models evolve in terms of overall performance and consistency across the various validation years as the number of training years increases for each basins. For each catchment, the performance of both models generally improves with an increase in the number of training years, suggesting that longer training periods lead to enhanced model fitting and improved prediction accuracy. The Severn catchment demonstrates a more

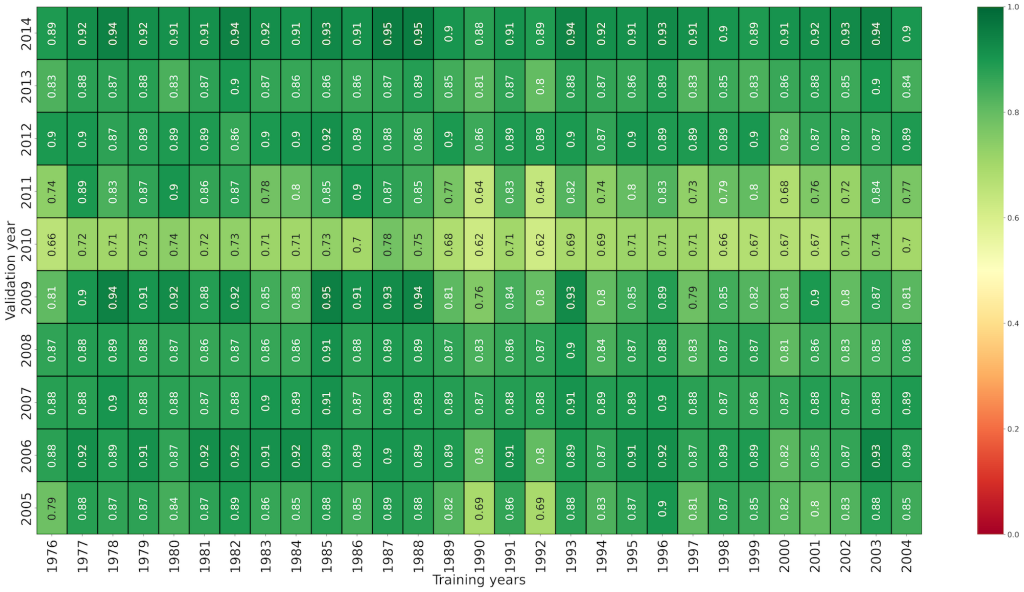

Fig. 14 Heatmap of NSE values obtained with one year of train/calibration data for Superflex model for the Severn catchment. Training years are shown in abscissa and validation years in ordinate.

consistent enhancement in NSE values as training years increase for both models, in contrast to the Fish River and Narraguagus River catchments, which exhibit greater variability in model performance. Notably, the Narraguagus River consistently shows lower NSE values compared to the Severn and Fish River, particularly with extended training periods. This reduced performance may be influenced by the presence of a dam in the catchment, which can significantly modify the natural flow regime, introduce regulated discharge, and increase the complexity of hydrological processes, thus challenging the models' ability to accurately capture the system's dynamics. Furthermore, the LSTM model exhibits a more pronounced improvement in performance with additional training data, indicating a greater sensitivity to the length of the training period compared to the Superflex model. This suggests that while both models benefit from increased training data, LSTM's data-driven approach may leverage extended training more effectively to refine its predictive capabilities, making it more sensitive to the quality and quantity of training data.

As a further analysis, Fig. 16 shows the evolution of the NSE across the validation years and the number of training years across all basins on average. For a one-year training, Superflex outperforms the LSTM model for each validation year. However, as the number of training years extends, the NSE of Superflex remains relatively stable across the validation years, while the LSTM model NSE increases substantially for every validation years. When the number of training years extends to six years, the LSTM model begins to catch up with Superflex and overperform it for nearly every validation years until fifteen years of training. This behavior is consistent across each catchments.

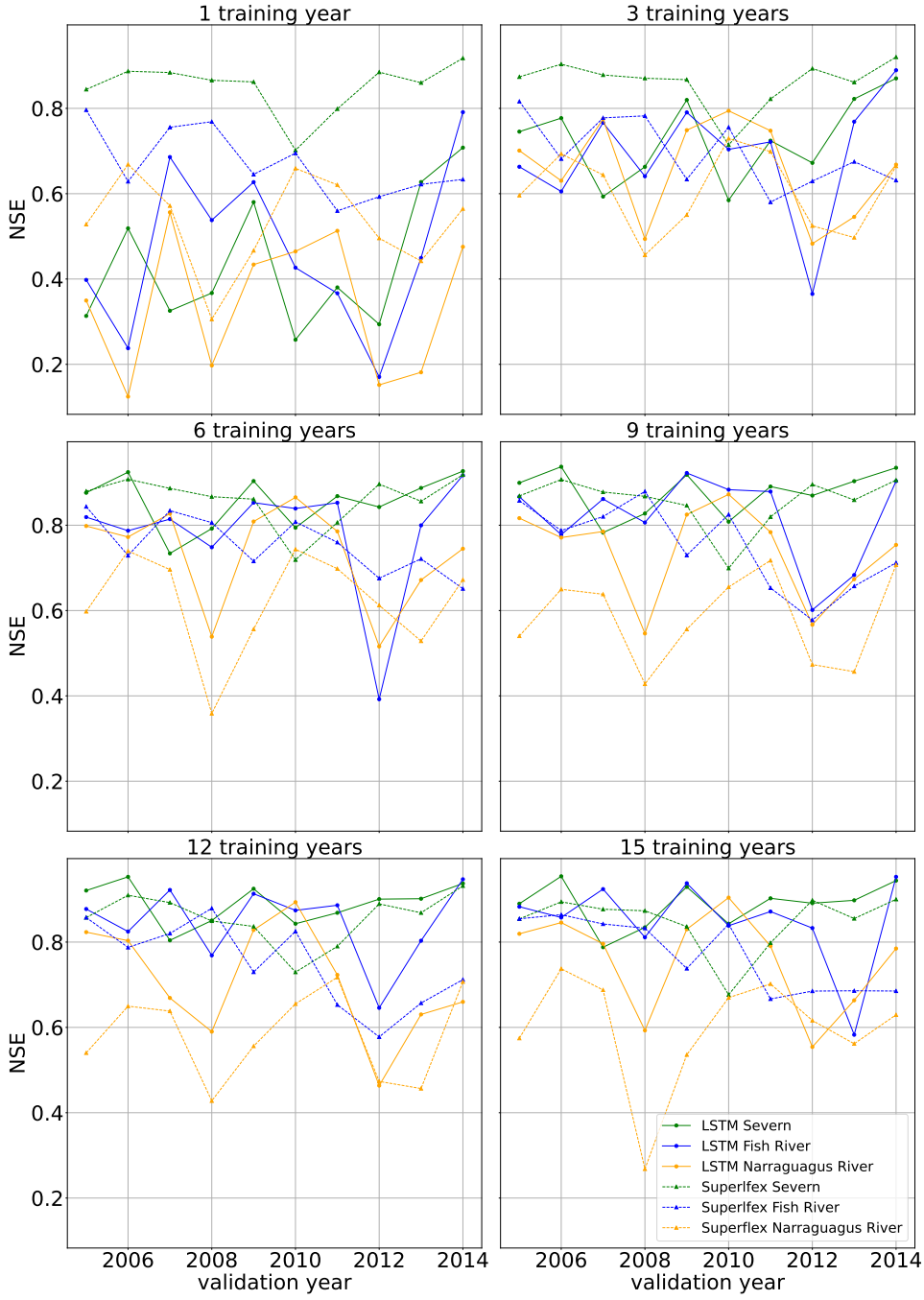

Fig. 15 Average NSE according to validation year obtained by both models and each catchment. Each graph represents a fixed number of training years from 1 to 15.

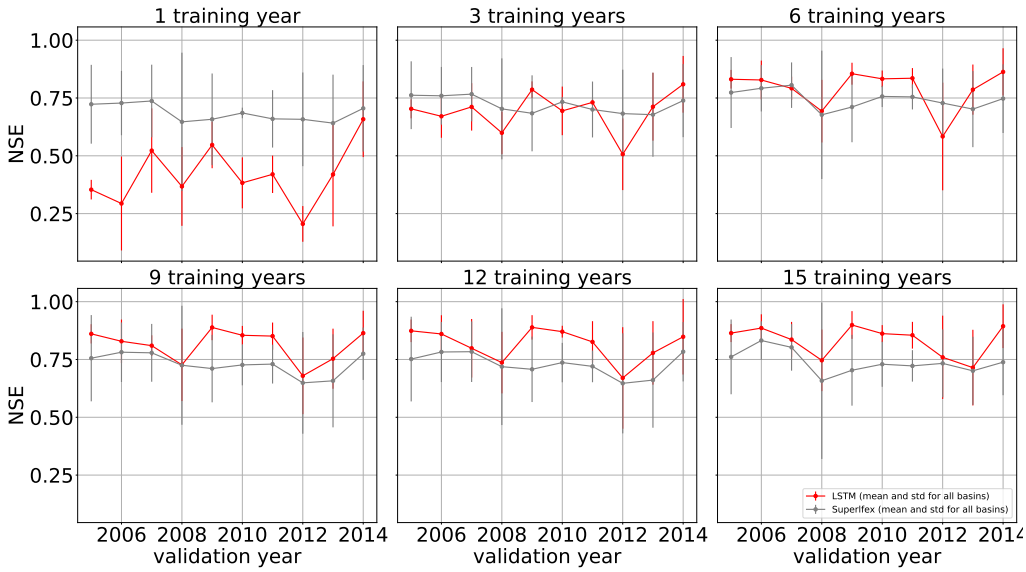

Fig. 16 Average NSE and standard deviation across validation years for both models, averaged over the three catchments. Each graph corresponds to a fixed number of training years, ranging from 1 to 15.

The graphs in Fig. 17 illustrate the median and standard deviation of NSEs values obtained in Fig. 15 as a function of the number of training years (1, 3, 6, 9, 12, and 15) for the LSTM and Superflex models across the Severn, Fish River, and Narraguagus River catchments. The plots at the top illustrate the median NSE in relation to the number of training years for all catchments and models. The top-left plot shows this data for each catchment individually, while the top-right plot presents the mean NSE and its standard deviation aggregated across all basins. Conversely, the plots at the bottom display the standard deviation of NSE with respect to the number of training years for all catchments and models. The bottom-left plot represents each catchment separately, while the bottom-right plot shows the aggregated mean standard deviation of NSE across all basins.

The upper panel of Fig. 17 supports the previous analysis while offering additional insights. The median NSE generally increases with the number of training years across all catchments and models, with the LSTM model for the Severn catchment achieving the highest median values. For the LSTM model, a pronounced increase in median NSE is observed as the training period extends from 1 to 9 years, after which it stabilizes around 0.9. This trend underscores the significant sensitivity of the LSTM model's performance to the training period, particularly for training years shorter than 9 years. In contrast, the Superflex model maintains consistently high median NSE values, hovering around 0.75 (slightly lower than the LSTM model during the 9 to 15-year range) across various training years duration. This consistent performance indicates that the Superflex model is considerably less influenced by the length of the training period. This behavior is generally consistent across all catchments, although the degree of

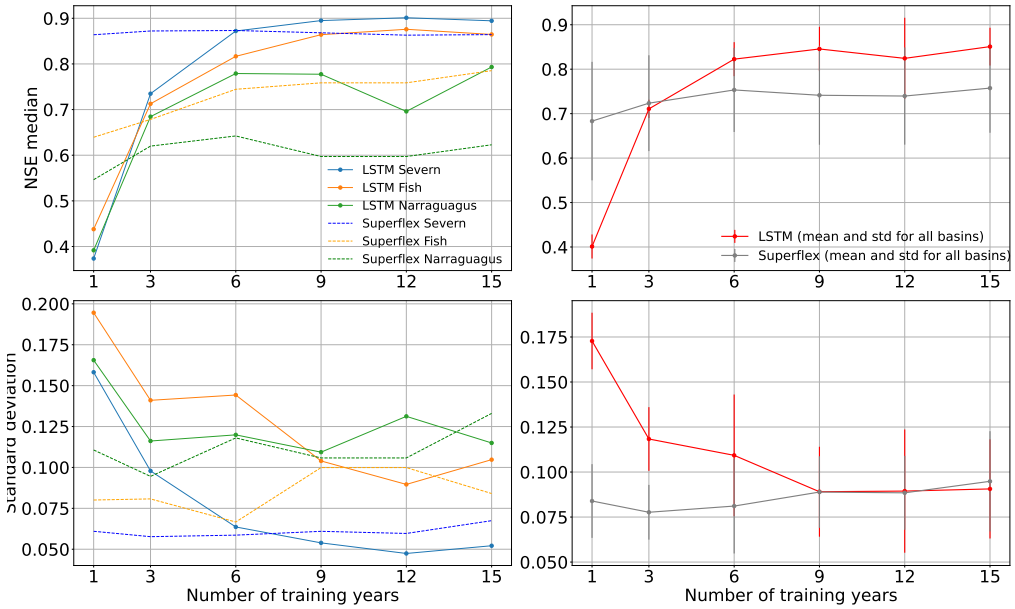

Fig. 17 Comparison between LSTM and Superflex models across two metrics: Median NSE (top) and its standard deviation (bottom). The left side of the figure displays results for each individual catchment, while the right side shows the mean and the standard deviation values across all catchments.

variation differs depending on the specific characteristics of each basin. For instance, the Narraguagus River catchment shows a lower median NSE and a higher standard deviation compared to the Severn and Fish River catchments, indicating that, on average, the models perform less effectively and their predictions are more variable in this catchment. This diminished performance and increased variability could be associated with the presence of a dam in the catchment, which may disrupt the natural flow regime and present challenges for the models in accurately capturing the hydrological dynamics.

In the bottom panel of Fig. 17, a notable decrease in the standard deviation of the LSTM model is observed from 1 to 6 years of training. Beyond 9 years, the standard deviation stabilizes, indicating that the LSTM model's performance initially exhibits considerable variability with respect to validation years when trained over shorter periods, but this variability diminishes rapidly as the training duration increases. In contrast, the Superflex model consistently maintains low variability values, approximately 0.085 (slightly lower than the LSTM model during the 9 to 15-year range), regardless of the number of training years. This stability confirms that the Superflex model is significantly less affected by the duration of the training period. This consistent behavior is observed across different catchments and evaluation metrics (KGE and MSE), though with varying degrees of prominence.

In summary, LSTM model's median NSE improves significantly with training up to 9 years, then stabilizes, while the Superflex model remains consistently stable around

0.75 regardless of training duration. LSTM model's variability decreases rapidly with more training, while Superflex maintains low, stable variability across all training periods.

The markedly different behaviours of the two models can be explained by the difference in the number of parameters, or degrees of freedom. The Superflex model is designed with a set of 8 parameters, whereas the LSTM model operates with hundreds of parameters. This difference in one order of magnitude is a key reason why the LSTM model requires more data or constraints to robustly calibrate its parameters. Once the LSTM model's parameters are well calibrated, it tends to outperform the Superflex model. This superior performance is likely due to the LSTM's greater number of degrees of freedom, which allows it to more precisely adapt to the unique characteristics of the studied catchment. For the Superflex model, its lower number of parameters, combined with its foundation in hydrological processes, enables quicker adjustment relative to the amount of data available. However, this limited parameter set may constrain the model's ability to precisely adapt to the specific characteristics of the studied catchment. These results also indicate that with 6 or more training years, the LSTM model generally outperforms the Superflex model, making it a preferable choice. However, the question remains: How does this performance hold up when faced with changing hydrometeorological typologies? This is what we aim to explore in the following section.

3.4 Impact of the hydrometeorological diversity of the training dataset on model performance

In the previous section, we highlighted the LSTM model's greater sensitivity to the length of training data compared to the Superflex model. We can now shift our focus to examining how hydrometeorological diversity within the training dataset affect model performance as introduced in section 2.8. To achieve this, we classified the years in the dataset for each catchment according to hydrological typology using a K-means clustering approach. Default parameters from scikit-learn [66] were employed, with the number of clusters set to three. The clustering was based on annual mean and standard deviation of daily rainfall, effectively categorizing the years into distinct hydrological types. Fig. 18 illustrates the clustering results for the Severn catchment, clearly distinguishing between drier, average, and wetter years, represented by green, blue, and red dots, respectively (clustering results for other catchments are detailed in A). This classification provides a clearer understanding of the variability in hydrological conditions, laying the groundwork for a more in-depth analysis of how different training data characteristics influence model performance.

Three distinct clusters emerge:

1. A green cluster comprising years characterised by low mean and standard deviation of precipitation. These years correspond to drier years.
2. A blue cluster consists of years that exhibit average mean and standard deviation of precipitation, and can thus be categorised as more typical or standard years.
3. A red cluster encompassing years with higher mean precipitation and larger standard deviation. These are indicative of wetter years.

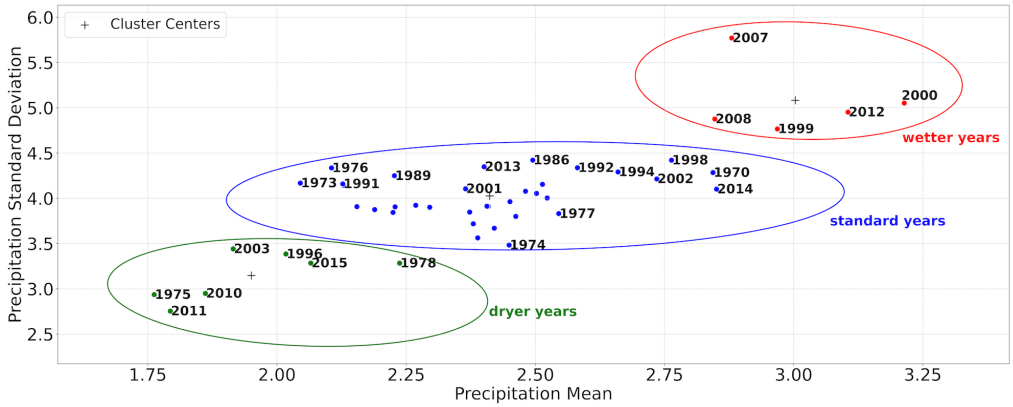

Fig. 18 Clusters on annual mean and standard deviation of daily rainfall of the Severn catchment. Three distinct yearly precipitation patterns: dryer years in green, standard years in blue, and wetter years in red

Fig. 19 shows an example of yearly time series of precipitation and streamflow for years representative of each of the three clusters:

- The top panel of this figure depicts a relatively dry year (1996), during which the number of flood events is noticeably low.
- In contrast, the middle panel represents a normal year (2009), marked by a higher frequency of flood events compared to the dry year.
- The bottom panel of the figure illustrates a wet year (1990), which is characterized by more frequent and longer-lasting flood events.

In light of this clustering analysis (drier years, standard years, wetter years), we carried out training experiments with various scenarios (see section 2.8), including trainings with single years, trainings with three consecutive years (S1), trainings with three training years belonging to the same cluster (S2), and three training years belonging to distinct clusters (S3).

3.4.1 Training with single clustered training year

Fig. 20 and Fig. 21 display heatmaps similar to those depicted in Fig. 13 and Fig. 14 respectively. The key distinction is the arrangement of the x-axis, which represents the training years, and the y-axis, which represents the validation years, organised based on the clustering and sorted by annual mean daily precipitation values.

What becomes evident is a notable difference in the general tendencies between the LSTM and Superflex models for the Severn catchment. The LSTM model exhibits higher variability, ranging from a minimum of -0.5 to a maximum of 0.85, in contrast to the Superflex model, which seems more consistent around 0.88 except for validation years 2010 and 2011, with NSE$\approx$ 0.75. This slight decrease in performance for these two drier years may suggest a challenge for Superflex in accurately predicting dry years.

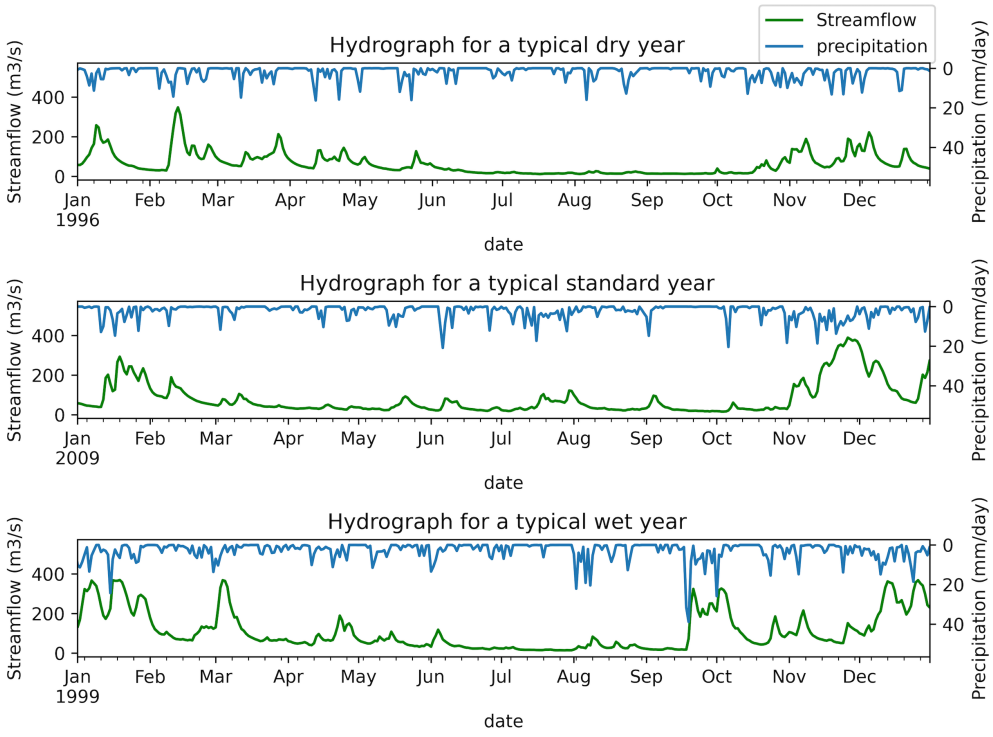

Fig. 19 Precipitation and streamflow time series of typical years belonging to each cluster for the Severn catchement.

This disparity underscores several discernible patterns: within the middle region, there appears to be a prevalence of greener hues, while the top-left and bottom-right corners lean toward redder shades. This observation implies that the performance of the LSTM model appears to be sensitive to the hydrometeorological typology of the training and validating year. Specifically, a standard training year tends to yield better predictive performance for another standard year, while a drier training year results in poorer model performance when predicting wetter years, and vice versa. In contrast, the Superflex model displays lower variability, suggesting a reduced sensitivity to the hydrometeorological typology of the training and validating years. Similar conclusions can also be drawn for the other catchments.

3.4.2 Training with three training years belonging to the same cluster

Fig. 22 and Fig. 23 shows the heatmap of NSE obtained for the LSTM and Superflex models respectively for Severn catchment, using various three-year training data subsets from the same cluster. The x-axis identifies the sets of three-year training data, grouped by clusters. On the y-axis, different validation years are represented along

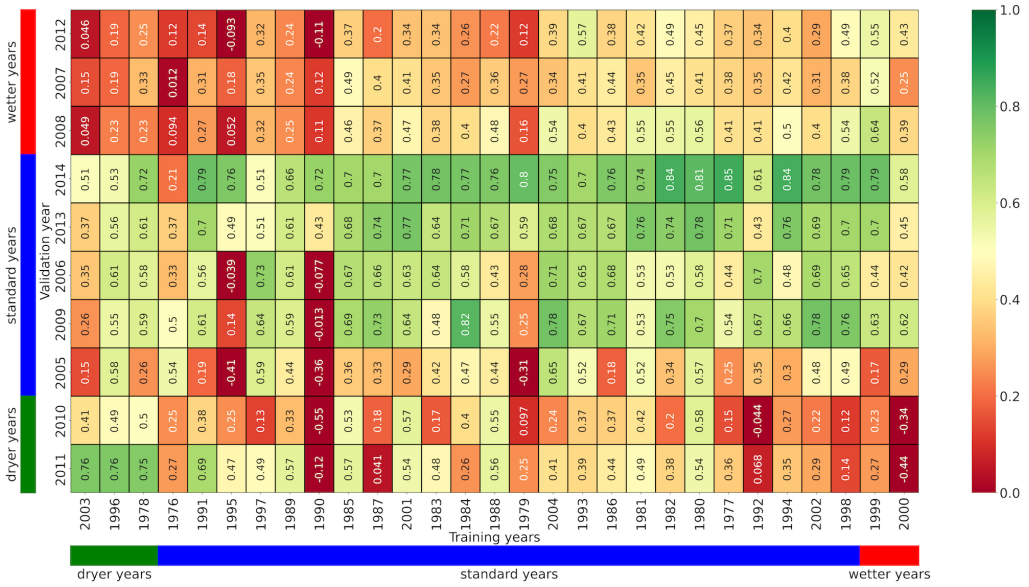

Fig. 20 Heatmap of NSE values obtained with one year of training/calibration data for LSTM model within the Severn catchment area. The x-axis represents the training years, while the y-axis represents the validation years. Both the training and validation years are arranged based on the annual mean daily precipitation.

with their cluster, enabling an analysis of how each model performs when assessed against diverse years. Due to the limited number of dryer and wetter years in our dataset, we had to include some of the validation years in the training set. However, we systematically didn't evaluate the models' performance on those validation years when also included in the training set (depicted in grey in the figure).

Superflex exhibits a more consistent color intensity pattern across different three-year intervals within the heatmap. This uniformity implies that the Superflex model performance remains relatively stable, regardless of the specific set of training data under consideration. In contrast, the LSTM model displays variations in color intensity within the heatmap, indicating its sensitivity to the particular three-years combination of training data. As previously observed, this difference highlights discernible patterns: a prevalence of greener hues in the middle region, with darker green tones in the top-right and bottom-left corners. This suggests that the LSTM model performance is influenced by the hydrometeorological diversity of the training and validation years. More specifically, a set of three dryer, standard or wetter training years tends to yield improved predictive performance for a corresponding dryer, standard or wetter year in validation, while other combinations may yield less favorable results, even if wetter training years enhance predictive performance for corresponding standard years, whereas standard training years do not improve the model predictive performance for wetter years. In comparison to Fig. 20, it is evident that when utilizing a set of three

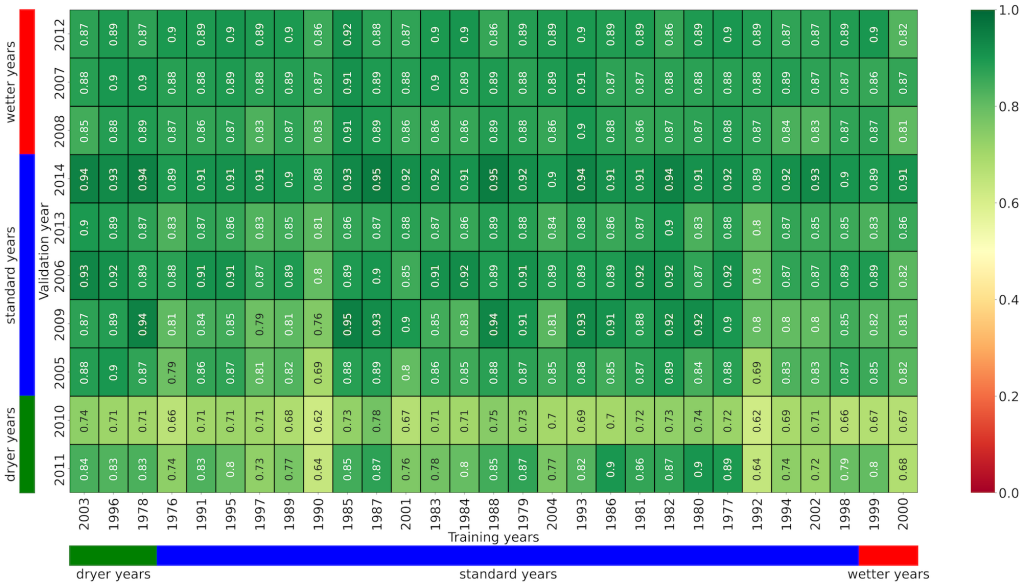

Fig. 21 Heatmap of NSE values obtained with one year of training/calibration data for superflex model within the Severn catchment area. The x-axis represents the training years, while the y-axis represents the validation years. Both the training and validation years are arranged based on the annual mean daily precipitation.

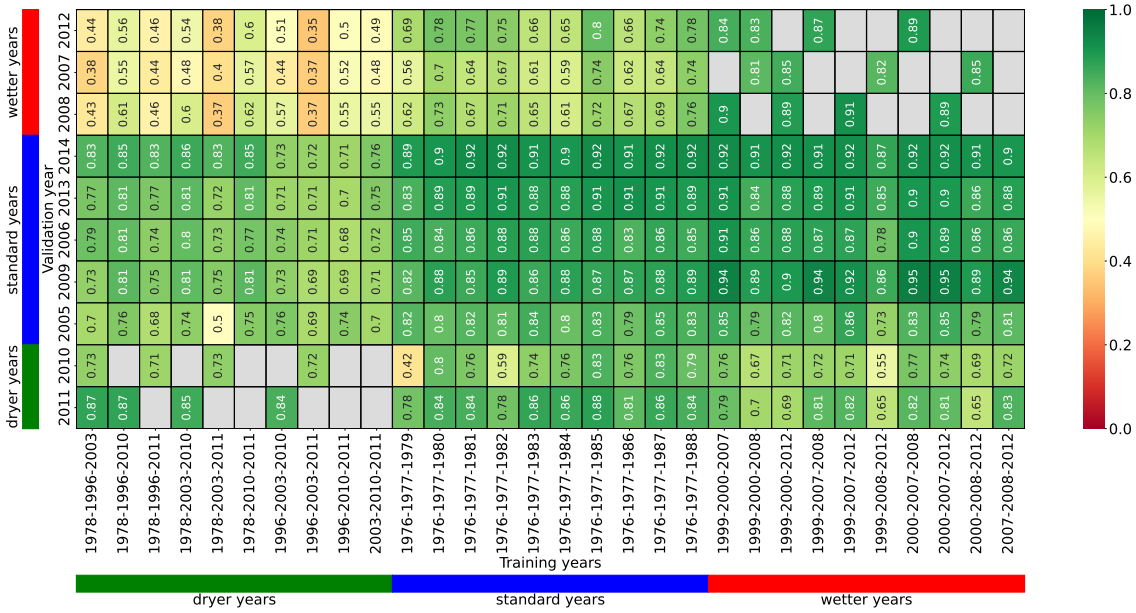

Fig. 22 Heatmap of NSE values for S2 (set of three-year training data from the same cluster) organised by cluster for LSTM model for the Severn catchement.

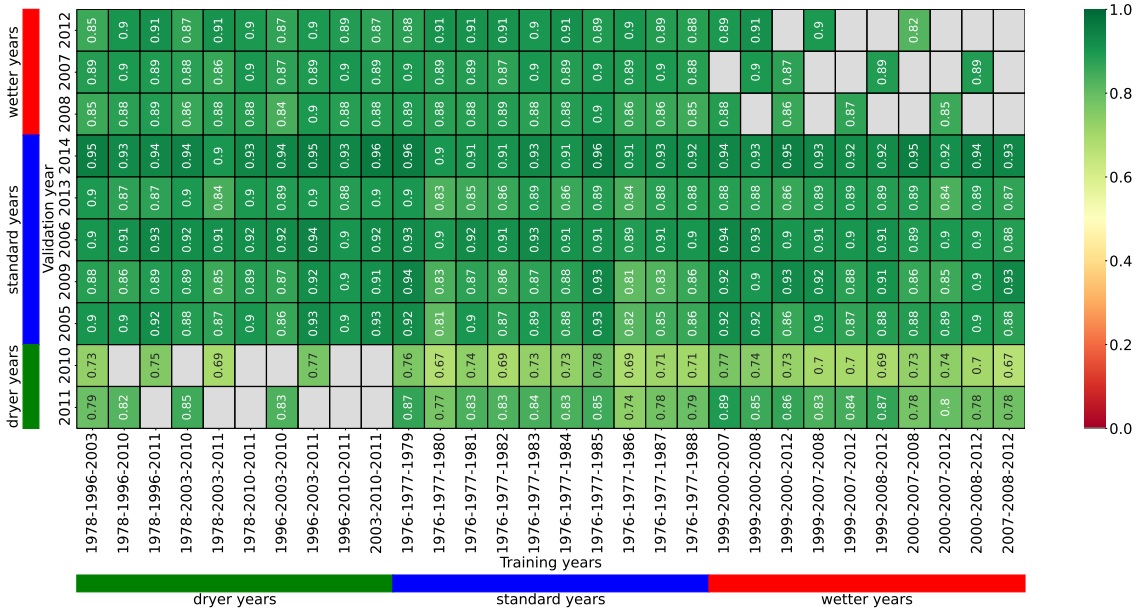

Fig. 23 Heatmap of NSE values for S2 (set of three-year training data from the same cluster), organised by cluster for Superflex model for the Severn catchement.

training years from the same cluster, the LSTM model exhibits improved overall performance compared to using just one training year. However, the disparities in relative performance across different combinations of training and validating years persist. We can witness identical phenomena occurring within the other two catchments as well).

3.4.3 Training with three training years belonging to distinct clusters

Fig. 24 and Fig. 25 illustrate the NSE heatmap for the LSTM and Superflex models respectively using distinct three-year training datasets, each year originating from a different cluster from Severn catchment. The various sets of three-year training data are on the x-axis, while the y-axis represents different validation years, along with their respective clusters.

In the context of the LSTM model for the Severn catchment, Fig. 24 exhibits a high level of color uniformity across the heatmap in comparison to a set of three training years from the same cluster. This observation indicates an enhanced reliability of model performance across various combinations of training and validating years. It further supports our hypothesis that the hydrometeorological diversity of the training years significantly influence model performance. This pattern is more prominent in the LSTM model compared to the Superflex model for the Severn catchement as illustrated in Fig. 25.

To enable further comparison, the boxplots in Fig. 26 depict the distribution of NSE values obtained with Superflex and LSTM models for 3 training years from the

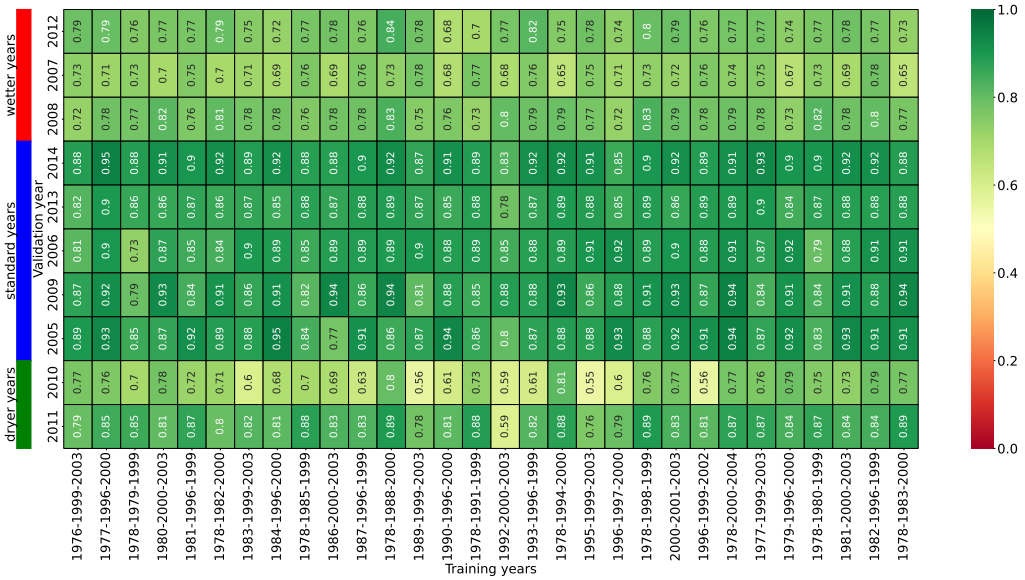

Fig. 24 Heatmap of NSE values for S3 (set of three-year training data from different clusters), organised by cluster for LSTM model

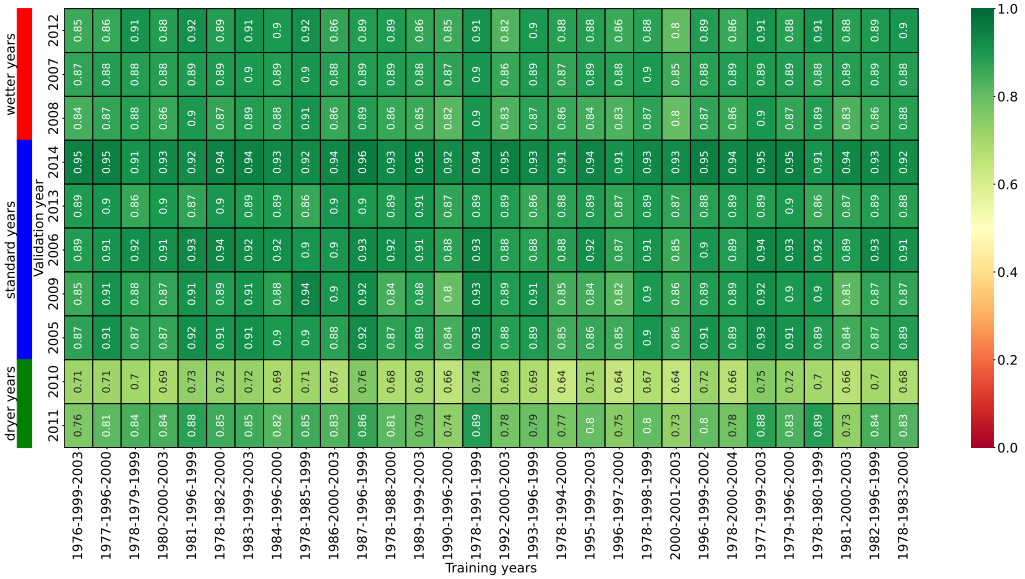

Fig. 25 Heatmap of NSE values for S3 (set of three-year training data from different clusters), organised by cluster for Superflex model

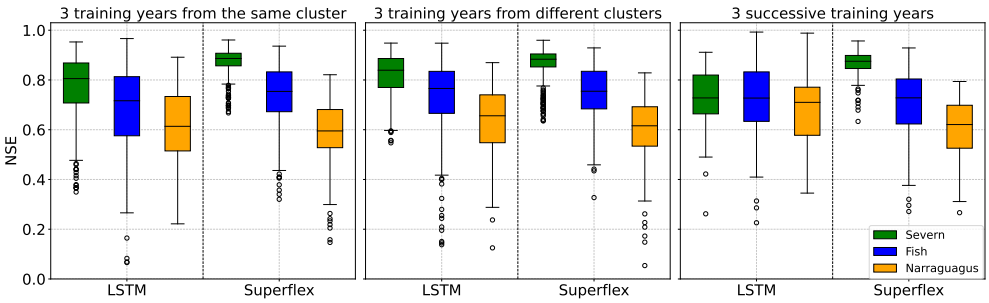

Fig. 26 Boxplots illustrating the distribution of the NSE performance metric for both the LSTM and Superflex models for each catchment, across three distinct scenarios: 3 training years from, left: the same cluster (S2) middle: different clusters (S3) and right: successive (S1).

same cluster S2 (in the right panel), 3 training years from different clusters S3 (in the middle panel), and 3 successive training years S1 (in the right panel) for each catchement. The boxplot visualizes data distribution through key components: the box itself spans from the 25% quartile (Q1) to the 75% quartile (Q3), with a line marking the median (Q2). The box width (Q3 - Q1) represents the interquartile range (IQR), indicating data spread. The upper whisker extends to 1.5 times the IQR above Q3, indicating the upper range. The lower whisker shows the lower range, ending at 1.5 times the IQR below the lower quartile (Q1). Points beyond the whiskers are considered outliers. We have truncated the boxplots at 0 to improve visibility.

In Fig. 26, we observe notable trends in the performance metrics for both the Superflex and LSTM models. Regarding the Superflex model, there is a prevailing uniformity in the median and interquartile range (IQR) across various scenarios for each catchment. This uniformity indicates a stable level of performance and reliability, underscoring its stability across different scenarios.

In contrast, the LSTM model displays variation in its median values. Specifically, for the Severn catchment (in green), the highest median value (0.84) is achieved for the different cluster's training years, followed by the same cluster's training years (0.81), and the lowest median is obtained for successive training years. This pattern suggests that the LSTM model generally performs better when trained on data from different clusters, followed by data from the same cluster, and performs relatively less favourably when trained on successive years.

Moreover, the range between the maximum and minimum values (max-min) decreases from successive training years (0.65) to different cluster's training years (0.40), with same cluster's training years falling in between at 0.60. This finding indicates that the model's performance is more consistent when trained using data from different clusters compared to data from the same cluster or successive training years. Similar conclusions can be drawn for the two other catchments.

These findings suggest that under non-stationary climate conditions, LSTM models, and potentially other data-driven approaches such as artificial neural networks (ANNs), may face challenges in maintaining precision and reliability compared to traditional models like empirical or physically-based models. This could be attributed

to several factors inherent to data-driven models. First, these models rely heavily on the quality and representativeness of the training data. In environments where climatic and hydrological conditions are rapidly changing, the training data may not adequately capture the variability and emerging patterns, leading to reduced model accuracy and generalization capabilities.

Moreover, data-driven models often require extensive historical data to discern complex relationships and patterns. In the absence of a sufficient volume of diverse and high-quality data, these models may overfit to specific patterns present in the training dataset, thereby failing to perform effectively when exposed to new or unseen conditions. This overfitting is particularly problematic in non-stationary contexts, where the statistical properties of the input data can change over time.

Traditional conceptual and physically-based models are grounded on established theoretical principles and can incorporate prior knowledge about the system dynamics. For example, these models can integrate hydrological processes such as precipitation-runoff relationships, groundwater flow, which are less sensitive to changes in input data characteristics. Consequently, these models are often more robust to variations in climate and land-use conditions, providing more reliable predictions under non-stationary scenarios.

Table 2 proposes various metrics for comparing the three scenarios to the reference of 15 training years. The three training years from different clusters scenario shows a similar or higher mean and median, as well as a lower standard deviation, compared to both the three successive and same cluster scenarios for each catchment respectively. This suggests that the scenario with training years from different clusters exhibits better overall performance and greater consistency than the scenarios with the same or three successive years. However, the different cluster scenario still performs less effectively, with lower mean and median values, and is less consistent (with a lower standard deviation) than the scenario with 15 successive training years. Furthermore, in the "different cluster" scenario, performance scores exceeding 0.7 are consistently achieved for each type of validation year when considered individually, with a minimum occurrence rate of 52%. In the case of dryer and wetter years, the "different cluster" scenario generally achieves scores above 0.7 but often faces challenges in reaching the 0.8 threshold.

When the LSTM model is trained with 15 successive years of data, it performs exceptionally well in validating wetter years, achieving an NSE greater than 0.7 for 100% of the years and an NSE greater than 0.8 with an average rate of 67% across the Severn catchment, even though the model was not explicitly trained on years with similar hydrological conditions. This high performance is also observed during dryer years, where the model reaches an NSE greater than 0.8 all the time. Similar patterns are evident in the other catchments, indicating the model's strong generalization capability and robustness across different environmental conditions, regardless of specific training clusters.

These results suggest that the LSTM model is capable of capturing the underlying hydrological dynamics across a wide range of conditions, showcasing its robust generalization ability across different environmental scenarios, irrespective of the specific training clusters.

NSE properties	(S2)			(S3)			(S1)			15 successive		
All years												
Mean NSE	0.76	0.69	0.62	0.82	0.72	0.64	0.73	0.69	0.66	0.89	0.85	0.76
Median NSE	0.81	0.72	0.62	0.84	0.77	0.66	0.73	0.72	0.68	0.89	0.86	0.79
Std NSE	0.13	0.23	0.15	0.08	0.18	0.13	0.12	0.19	0.13	0.05	0.10	0.11
Max-min NSE	0.60	1.76	0.77	0.40	1.38	0.74	0.65	1.35	0.52	0.17	0.37	0.35
All years [%]												
NSE > 0.7	69	55	36	91	69	36	61	55	47	100	90	70
NSE > 0.8	47	32	15	60	40	8	31	28	10	90	90	40
Dryer years [%]												
NSE > 0.7	67	-	-	77	-	-	56	-	-	100	-	-
NSE > 0.8	33	-	-	42	-	-	6	-	-	100	-	-
Standard years [%]												
NSE > 0.7	95	55	41	1.00	71	46	87	59	52	100	86	67
NSE > 0.8	71	31	18	97	44	09	60	36	19	100	86	0
Wetter years [%]												
NSE > 0.7	28	55	34	87	64	32	22	46	45	100	100	71
NSE > 0.8	14	34	14	10	30	08	0	8	06	67	100	57

Table 2 NSE performance metric for LSTM models across four distinct scenarios: 3 training years from the same cluster (S2 in the left), 3 training years from different cluster (S3 in the middle left), 3 successive training years (S1 in the middle right), and 15 successive training years (in the rigth). These scenarios are assessed for the Severn catchement (in green), the Fish river catchement (in blue), and the Narraguagus catchement (in orange). For "All validation years" and the properties associated with validation years listed beneath it, the values indicate a ratio.

This strong generalization capability is not limited to a single catchment but is consistently observed across other catchments as well. It indicates that the LSTM model can adapt effectively to variations in hydrometeorological diversity, thereby providing reliable performance even when exposed to diverse and previously unseen scenarios. Such adaptability is particularly valuable in hydrological modeling, where the ability to generalize beyond the training data is crucial for accurate forecasting under variable and changing climate conditions.

In contrast, the LSTM model also shows a degree of sensitivity to the diversity of hydrometeorological training conditions. This sensitivity can lead to variability in model performance when the training data does not adequately represent the full spectrum of climatic scenarios. However, this limitation can be mitigated in some extend by training the model on a diverse range of hydrometeorological diversity, as evidenced by its superior performance in the "different cluster" scenario. In this context, the LSTM model outperforms when trained on heterogeneous datasets that encompass a variety of hydrological regimes, demonstrating higher resilience and reliability.

Nonetheless, caution is necessary when applying data-driven models like LSTM in situations where the hydrometeorological diversity are significantly different from those present in the training data. The model's predictive accuracy may diminish under such circumstances, underscoring the importance of carefully accumulating a training dataset that represents a wide spectrum of possible conditions. While LSTM models, with sufficient training data, offer powerful tools for capturing complex, nonlinear relationships in hydrology - particularly when compared to traditional models like

Superflex, as demonstrated in the previous section - their effectiveness is contingent on the quality and diversity of the training data.

4 Conclusion

This study explored how the length and hydrometeorological diversity of training datasets affect the performance of LSTM models compared to traditional rainfall-runoff models like Superflex. The findings reveal that LSTM models are highly sensitive to the length of the training period. For instance, when trained with only one year of data, the LSTM model used showed poor predictive performance, with low median NSE values between 0.2 and 0.44. However, its performance improved significantly with longer training datasets. With six years of training, the model achieved median NSE values ranging from 0.78 to 0.87; it further improved to a range of 0.79 to 0.89 with 15 years, along with reduced values for standard deviation, indicating more consistent predictions.

In contrast, the Superflex model displayed less sensitivity to training data length, maintaining stable performance with median NSE values between 0.6 and 0.9, regardless of the amount of training data. This suggests that while traditional models like Superflex may not achieve the high performance of LSTM models with extensive training, they are more robust in scenarios with limited data availability.

The study also examined the impact of hydrometeorological diversity on LSTM performance. When the model was trained on datasets that included a variety of conditions – such as wetter, average, and drier years – it exhibited higher performance and reliability, achieving an average median NSE of 0.76 with a low standard deviation of 0.13. Conversely, training on more homogeneous datasets, such as data from the same cluster or successive years, resulted in lower performance and greater variability. This highlights the importance of diverse training data in improving the LSTM model's generalization capabilities.

Our findings show that with six or more training years, the LSTM model generally outperforms the Superflex model, making it a preferable choice. However, despite their potential, LSTM models have a critical limitation: they rely heavily on the training dataset and are highly sensitive to hydrometeorological variability. Unlike traditional models, their predictive accuracy may degrade significantly when faced with conditions outside the range of their training data.

To address these challenges, future research should focus on developing strategies to enhance model resilience, such as incorporating synthetic data or using transfer learning techniques. These approaches could help ensure reliable performance of LSTM models even in regions with limited historical data and unseen hydrometeorological typology.

Additionally, conducting a sensitivity analysis to understand the relationship between hydrometeorological diversity and LSTM performance could help identify mathematical thresholds at which the performance of LSTM and other ANN models begins to decline. Such an understanding would enable researchers to select the most appropriate modeling approach – either data-driven or traditional conceptual models

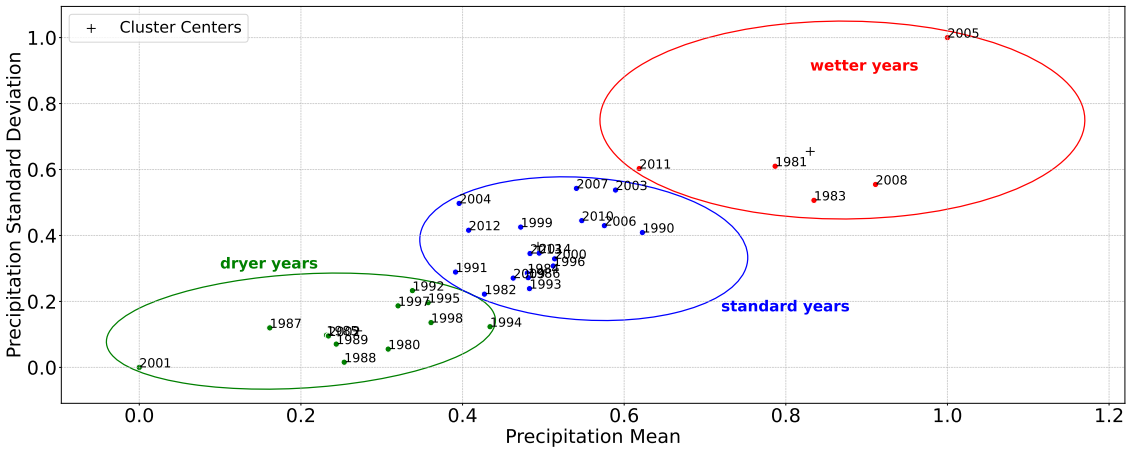

Fig. A1 Clusters on annual mean and standard deviation of daily rainfall for the Fish river catchement. Three Distinct Yearly Precipitation Patterns: dryer years in green, standard years in blue, and wetter years in red

– based on the specific conditions of a study area. This would optimize model performance and reliability across diverse hydrological contexts, providing more effective tools for managing water resources and adapting to the impacts of climate change.

Acknowledgements. This work has been partly supported by AXIAUM Project ANR-20-THIA-0005. The research reported herein was partly funded by the National French Research Agency through the SWIFT (ANR-23-CE56-0009-03) projects. This work has been realized with the support of MESO@LR-Platform at the University of Montpellier, and a part of the experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

Declarations

The authors did not receive support from any organization for the submitted work. The authors have no financial or proprietary interests in any material discussed in this article.

Appendix A Clusters on annual mean and standard deviation of daily rainfall for the Fish and Narraguagus catchement

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding

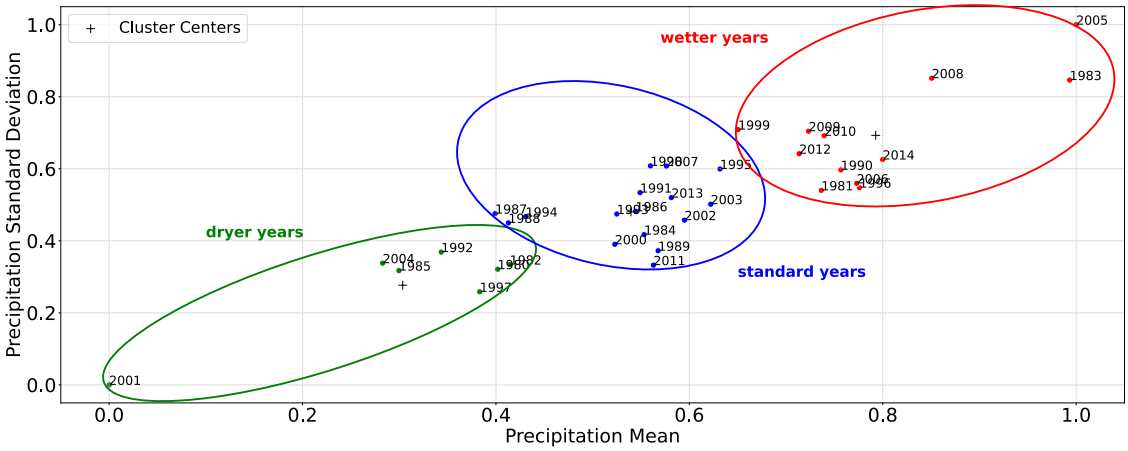

Fig. A2 Clusters on annual mean and standard deviation of daily rainfall for the Narraguagus catchement. Three Distinct Yearly Precipitation Patterns: dryer years in green, standard years in blue, and wetter years in red

of the research problem or it is information that is too cumbersome to be included in the body of the paper.

References

[1] Mulvaney, T.J.: On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and of flood discharges in a given catchment. Proceedings Institution of Civil Engineers **4**, 18–31 (1850)

[2] Beven, K.J.: Rainfall-runoff Modelling: The Primer, 2nd edn., p. 25. Wiley-Blackwell, Hoboken, NJ (2012)

[3] Islam, Z.: A Review on Physically Based Hydrologic Modeling. University of Alberta (2011). https://doi.org/10.13140/2.1.4544.5924 . http://rgdoi.net/10.13140/2.1.4544.5924

[4] Beckie, R.: Fundamental Hydrologic Equations. John Wiley & Sons, Ltd (2005). https://doi.org/10.1002/0470848944.hsa004 . https://doi.org/10.1002/0470848944.hsa004

[5] Shaw, S.B., Beslity, J.O., Colvin, M.E.: Working toward a more holistic set of hydrologic principles to teach non-hydrologists: Five simple concepts within catchment hydrology. Hydrological Processes (2019) https://doi.org/10.1002/hyp.13485

[6] Hostache, R., Rains, D., Mallick, K., Chini, M., Pelich, R., Lievens, H., Fenicia, F., Corato, G., Verhoest, N.E.C., Matgen, P.: Assimilation of soil moisture and ocean

salinity (SMOS) brightness temperature into a large-scale distributed conceptual hydrological model to improve soil moisture predictions: the Murray–Darling basin in australia as a test case. Hydrol. Earth Syst. Sci. **24**(10), 4793–4812 (2020)

[7] Clark, M.P., Bierkens, M.F.P., Samaniego, L., Woods, R.A., Uijlenhoet, R., Bennett, K.E., Pauwels, V.R.N., Cai, X., Wood, A.W., Peters-Lidard, C.D.: The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism. Hydrology and Earth System Sciences **21**(7), 3427–3440 (2017) https://doi.org/10.5194/hess-21-3427-2017

[8] Horton, R.E.: The role of infiltration in the hydrologic cycle. Transactions, American Geophysical Union **14**(1), 446 (1933) https://doi.org/10.1029/tr014i001p00446

[9] Jehanzaib, M., Ajmal, M., Achite, M., Kim, T.-W.: Comprehensive review: Advancements in rainfall-runoff modelling for flood mitigation. Climate **10**(10), 147 (2022) https://doi.org/10.3390/cli10100147 . Accessed 2023-08-24

[10] Young, P.C., Beven, K.J.: Data-based mechanistic modelling and the rainfall-flow non-linearity. Environmetrics **5**, 335–363 (1994)

[11] Remesan, R., Mathew, J.: Hydrological Data Driven Modelling. Springer, ??? (2015). https://doi.org/10.1007/978-3-319-09235-5 . https://doi.org/10.1007%2F978-3-319-09235-5

[12] Solomatine, D., See, L.M., Abrahart, R.J.: Data-driven modelling: Concepts, approaches and experiences. In: Abrahart, R.J., See, L.M., Solomatine, D.P. (eds.) Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, pp. 17–30. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79881-1_2

[13] Zounemat-Kermani, M., Matta, E., Cominola, A., Xia, X., Zhang, Q., Liang, Q., Hinkelmann, R.: Neurocomputing in surface water hydrology and hydraulics: A review of two decades retrospective, current status and future prospects. Journal of Hydrology **588**, 125085 (2020) https://doi.org/10.1016/j.jhydrol.2020.125085

[14] Sit, M., Demiray, B.Z., Xiang, Z., Ewing, G.J., Sermet, Y., Demir, I.: A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources. arXiv (2020). https://doi.org/10.48550/ARXIV.2007.12269 . https://arxiv.org/abs/2007.12269

[15] Elman, J.L.: Finding structure in time. Cognitive Science **14**(2), 179–211 (1990) https://doi.org/10.1016/0364-0213(90)90002-E

[16] Graves, A., Mohamed, A., Hinton, G.E.: Speech recognition with deep recurrent neural networks. arXiv: Neural and Evolutionary Computing (2013) https://doi.org/10.1109/icassp.2013.6638947

[17] Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative Study of CNN and RNN for Natural Language Processing (2017)

[18] Dixon, M., London, J.: Financial forecasting with α-rnns: A time series modeling approach. Frontiers in Applied Mathematics and Statistics (2021) https://doi.org/10.3389/fams.2020.551138

[19] Noh, S.: Analysis of gradient vanishing of rnns and performance comparison. Information-an International Interdisciplinary Journal (2021) https://doi.org/10.3390/info12110442

[20] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation (1997) https://doi.org/10.1162/neco.1997.9.8.1735

[21] Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with lstm recurrent networks. Journal of machine learning research **3**(Aug), 115–143 (2002)

[22] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. NIPS (2017) https://doi.org/10.48550/arXiv.1706.03762

[23] Gao, P., Geng, S., Qiao, Y., Wang, X., Dai, J., Li, H.: Scalable Transformers for Neural Machine Translation (2021)

[24] Yin, H., Guo, Z., Zhang, X., Chen, J., Zhang, Y.: Rr-former: Rainfall-runoff modeling based on transformer. Journal of Hydrology **609**, 127781 (2022) https://doi.org/10.1016/j.jhydrol.2022.127781

[25] Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrology and Earth System Sciences **22**(11), 6005–6022 (2018) https://doi.org/10.5194/hess-22-6005-2018 . Accessed 2023-03-25

[26] Brath, A., Montanari, A., Toth, E.: Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model. J. Hydrol. (Amst.) **291**(3-4), 232–253 (2004)

[27] Merz, R., Parajka, J., Blöschl, G.: Scale effects in conceptual hydrological modeling: SCALE EFFECTS IN CONCEPTUAL HYDROLOGICAL MODELING. Water Resour. Res. **45**(9) (2009)

[28] Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., Mathevet, T.: Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. Hydrol. Sci. J. **52**(1), 131–151 (2007)

[29] Boulmaiz, T., Guermoui, M., Boutaghane, H.: Impact of training data size on the

LSTM performances for rainfall–runoff modeling. Model. Earth Syst. Environ. **6**(4), 2153–2164 (2020)

[30] Newman, Andrew: A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. UCAR/NCAR - GDEX (2014). https://doi.org/10.5065/D6MW2F4D . https://gdex.ucar.edu/dataset/id/fbc54ccc-5184-4f54-b306-f58112a34700.html

[31] Adler, C., Wester, P., Bhatt, I., Huggel, C., Insarov, G.E., Morecroft, M.D., Muccione, V., Prakash, A.: Cross-chapter paper 5: Mountains. In: Pörtner, H.O., Roberts, D.C., Tignor, M., Poloczanska, E.S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., Rama, B. (eds.) Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 2273–2318. Cambridge University Press, Cambridge, UK and New York, USA (2022). https://doi.org/10.1017/9781009325844.022.2273

[32] Coxon, G., Addor, N., Bloomfield, J.P., Freer, J., Fry, M., Hannaford, J., Howden, N.J.K., Lane, R., Lewis, M., Robinson, E.L., Wagener, T., Woods, R.: Catchment attributes and hydro-meteorological timeseries for 671 catchments across Great Britain (CAMELS-GB). NERC EDS Environmental Information Data Centre (2020)

[33] Kratzert, F., Gauch, M., Klotz, D., Nearing, G.: Hess opinions: Never train a long short-term memory (lstm) network on a single basin. Hydrology and Earth System Sciences **28**(17), 4187–4201 (2024) https://doi.org/10.5194/hess-28-4187-2024

[34] River Severn at Saxons Lode: River level and flood alerts. https://riverlevels.uk/river-severn-ripple-saxons-lode. Accessed: 2023-5-17 (2023)

[35] Newson, M.D., Newson, M., Newson, M.D., Harrison, J.G.: Channel studies in the plynlimon experimental catchments. null (1978)

[36] Marsh, J.H.: Saint John River. https://www.thecanadianencyclopedia.ca/en/article/saint-john-river. Accessed: 2024-2-17 (2006)

[37] Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R.: Terrestrial ecoregions of the world: A new map of life on earth. BioScience **51**(11), 933 (2001) https://doi.org/10.1641/0006-3568(2001)051[0933:teotwa]2.0.co;2

[38] Flooding in New Brunswick. https://flooding-inondations-geonb.hub.arcgis.com. Accessed: 2024-2-17 (2019)

[39] Worst floods in New Brunswick history: how 2018 compares. https://www.

cbc.ca/news/canada/new-brunswick/st-john-river-flooding-history-1.4641969. Accessed: 2024-2-17 (2018)

[40] New Brunswick's Flood Risk Reduction Strategy. https://www2.gnb.ca/content/dam/gnb/Departments/env/pdf/Flooding-Inondations/NBFloodRiskReductionStrategy.pdf. Accessed: 2024-2-17 (2019)

[41] Geological Survey (U.S.): National Hydrography Dataset. [Reston, Va.] : U.S. Dept. of the Interior, U.S. Geological Survey, 2004- (2004). https://search.library.wisc.edu/catalog/9910061259502121

[42] Addor, N., Newman, A.J., Mizukami, N., Clark, M.P.: Catchment attributes for large-sample studies. UCAR/NCAR (2017). https://doi.org/10.5065/D6G73C3Q . https://gdex.ucar.edu/dataset/camels.html

[43] Dudley, R.W., Dudley, R.W., Nielsen, J.P., Nielsen, J.P.: Streamflow statistics for the narraguagus river at cherryfield, maine. null (2000) https://doi.org/10.3133/ofr200095

[44] Narraguagus River at Cherryfield, Maine — waterdata.usgs.gov. https://waterdata.usgs.gov/monitoring-location/01022500. [Accessed 07-09-2024] (n.d.)

[45] Whiting, M.C., Whiting, M.C., Otto, W.D., Otto, W., Federation, D.S., Federation, D.S.: Spatial and temporal patterns in water chemistry of the narraguagus river: A summary of available data from the maine dep salmon rivers program. null (2008) https://doi.org/null

[46] Cherryfield Dam, Maine — All You Need To Know — damsoftheworld.com. https://damsoftheworld.com/usa/maine/cherryfield-dam/. [Accessed 07-09-2024] (n.d.)

[47] Cherryfield community takes steps to replace ice dam with natural fishway — newscentermaine.com. https://www.newscentermaine.com/article/tech/science/environment/cherryfield-residents-conservationists-replace-dam-natural-fishway/97-aeabff7e-fd1c-4f07-a262-fc089f0d8287. [Accessed 07-09-2024] (n.d.)

[48] Folland, C.K., Hannaford, J., Bloomfield, J.P., Kendon, M., Svensson, C., Marchant, B.P., Prior, J., Wallace, E.: Multi-annual droughts in the english lowlands: a review of their characteristics and climate drivers in the winter half-year. Hydrol. Earth Syst. Sci. **19**(5), 2353–2375 (2015)

[49] Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous usa: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences **19**(1), 209–223 (2015)

https://doi.org/10.5194/hess-19-209-2015

[50] Newman, Andrew: A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. UCAR/NCAR - GDEX (2014). https://doi.org/10.5065/D6MW2F4D . https://gdex.ucar.edu/dataset/id/fbc54ccc-5184-4f54-b306-f58112a34700.html

[51] Ruder, S.: An overview of gradient descent optimization algorithms (2017). https://arxiv.org/abs/1609.04747

[52] Kohan, A.A., Rietman, E.A., Siegelmann, H.T.: Error forward-propagation: Reusing feedforward connections to propagate errors in deep learning. CoRR **abs/1808.03357** (2018) 1808.03357

[53] Terven, J., Cordova-Esparza, D.M., Ramirez-Pedraza, A., Chavez-Urbiola, E.A., Romero-Gonzalez, J.A.: Loss Functions and Metrics in Deep Learning (2024). https://arxiv.org/abs/2307.02694

[54] Linnainmaa, S.: Taylor expansion of the accumulated rounding error. BIT **16**(2), 146–160 (1976) https://doi.org/10.1007/bf01931367

[55] Cardarilli, G.C., Re, M., Di Nunzio, L.: A pseudo-softmax function for hardware-based high speed image classification. Scientific Reports (2021) https://doi.org/10.1038/s41598-021-94691-7

[56] Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **06**(02), 107–116 (1998) https://doi.org/10.1142/S0218488598000094 https://doi.org/10.1142/S0218488598000094

[57] Orr, G.B., Muller, K.R. (eds.): Neural Networks Work. Springer, Berlin, Germany (1998)

[58] Gauss, C.F.: Theoria Motvs Corporvm Coelestivm in Sectionibvs Conicis Solem Ambientivm [Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections. Boston, Little, Brown and company, ??? (1809)

[59] Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: Neural Networks, pp. 437–478 (2012). https://api.semanticscholar.org/CorpusID:10808461

[60] Di Mauro, C., Hostache, R., Matgen, P., Pelich, R., Chini, M., Leeuwen, P.J., Nichols, N., Blöschl, G.: A tempered particle filter to enhance the assimilation of sar-derived flood extent maps into flood forecasting models. Water Resources Research **58**(8), 2022–031940 (2022) https://doi.org/10.1029/2022WR031940 https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022WR031940. e2022WR031940 2022WR031940

[61] Metropolis, N., Ulam, S.: The monte carlo method. Journal of the American Statistical Association **44**(247), 335–341 (1949) https://doi.org/10.1080/01621459.1949.10483310 https://www.tandfonline.com/doi/pdf/10.1080/01621459.1949.10483310. PMID: 18139350

[62] Nash, J.E., Sutcliffe, J.V.: River flow forecasting through conceptual models part i-a discussion of principles. J. Hydrol **10**, 282–290 (1970)

[63] Gupta, H.V., Kling, H.: On typical range, sensitivity, and normalization of mean squared error and nash-sutcliffe efficiency type metrics. Water Resources Research **47**(10) (2011) https://doi.org/10.1029/2011WR010962 https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011WR010962

[64] Sammut, C., Webb, G.I. (eds.): Mean Squared Error, p. 653. Springer, Boston, MA (2010). Chap. Mean Squared Error. https://doi.org/10.1007/978-0-387-30164-8_528 . https://doi.org/10.1007/978-0-387-30164-8_528

[65] Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory **28**(2), 129–137 (1982) https://doi.org/10.1109/tit.1982.1056489

[66] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[67] Van Rossum, G., Drake, F.L.: Python 3 Reference Manual. CreateSpace, Scotts Valley, CA (2009)

[68] Kratzert, F., Gauch, M., Nearing, G., Klotz, D.: Neuralhydrology — a python library for deep learning research in hydrology. Journal of Open Source Software **7**(71), 4050 (2022) https://doi.org/10.21105/joss.04050

[69] Walt, S., Colbert, S.C., Varoquaux, G.: The NumPy array: A structure for efficient numerical computation. Comput. Sci. Eng. **13**(2), 22–30 (2011)

[70] McKinney, W.: Data Structures for Statistical Computing in Python. In: Walt, Millman (eds.) Proceedings of the 9th Python in Science Conference, pp. 56–61 (2010). https://doi.org/10.25080/Majora-92bf1922-00a

[71] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**(85), 2825–2830 (2011)

[72] Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in Science &

Engineering **9**(3), 90–95 (2007) https://doi.org/10.1109/MCSE.2007.55

[73] Waskom, M.L.: seaborn: statistical data visualization. Journal of Open Source Software **6**(60), 3021 (2021) https://doi.org/10.21105/joss.03021

[74] Balouek, D., Carpen Amarie, ., Charrier, G., Desprez, F., Jeannot, E., Jeanvoine, E., Lèbre, A., Margery, D., Niclausse, N., Nussbaum, L., Richard, O., Pérez, C., Quesnel, F., Rohr, C., Sarzyniec, L.: Adding virtualization capabilities to the Grid'5000 testbed. In: Ivanov, I.I., Sinderen, M., Leymann, F., Shan, T. (eds.) Cloud Computing and Services Science. Communications in Computer and Information Science, vol. 367, pp. 3–20. Springer, ??? (2013). https://doi.org/10.1007/978-3-319-04519-1_1