How can we benefit from regime information to make use of LSTM runoff models more effectively?

Reyhaneh Hashemi[1], Pierre Brigode[2,3], Pierre-André Garambois[1], and Pierre Javelle[1]

[1]Aix-Marseille Université, INRAE, UR RECOVER, Aix-en-Provence, France
[2]Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, IRD, Géoazur, Sophia-Antipolis, France
[3]Université Paris-Saclay, INRAE, UR HYCAR, Antony, France

Correspondence: Reyhaneh Hashemi (reyhaneh.hashemi@inrae.fr)

Abstract. Long Short Term Memory (LSTM) networks have been so far successfully applied to a key problem in hydrology — prediction of runoff. Very contrary to traditional conceptual models, LSTM models are built on concepts that basically avoid the formal need for encoding our knowledge of hydrology into them. This brings the strong interest of benefiting from domain knowledge and traditional practices, not for building LSTM models, as we do for conceptual models, but for using them in a more effective way. In this paper, we take this perspective and investigate how we can use information related to hydrologic characteristics of catchments for LSTM runoff models. As the first application of LSTM to the French context, we use 361 gauged catchments with very diverse hydrologic conditions from all over France. The catchments have long time series of at least 30 years. Our main axes of investigation include a) the relationship between LSTM performance and the length of the input sequence of LSTM within different hydrologic regimes, b) the importance of hydrologic homogeneity of catchments when training LSTMs on a group of catchments, and c) the interconnected influence of local tuning of the two important LSTM hyperparameters, namely length of input sequence and hidden unit size, on performances of group trained LSTMs. We present a classification, which is built on three indices coming from the regime of runoff, precipitation, and temperature. We use this classification as our measure of homogeneity — catchments within the same regime are assumed to be hydrologically homogeneous. We train LSTMs on individual catchments (local level training), on catchments within the same regime (regime level training), and on the entire sample (national level training). We benchmark the local LSTM with the conceptual GR4J model — which is able to represent water gains/losses in a catchment. We show that in the two Uniform and Nival regimes where the dominant hydrologic process of the regime has a clear hysteretic pattern, LSTM performances have the highest sensitivity with respect to the length of the input sequence and large lengths should be chosen. In other regimes, this level of sensitivity is not found, in some of them an almost no-sensitivity level is observed — the size of the input sequence in these regimes does not need to be large, therefore. Overall, our homogeneous regime level training slightly outperforms our heterogeneous national level training. This shows that in both trainings the same level of data adequacy with respect to complexity of to-be-learned representation(s) is achieved. We, however, do not exclude the potential role of the "regime-informed" property of our national LSTMs — they use the classification variables as static attributes. Last but not least, we show that how a local selection of the two important hyperparameters of LSTM combined with a national level training can bring the best runoff prediction performances.

1 Introduction

Surface runoff (in short, runoff) is the response of catchment to its intakes and yields. A reliable prediction of runoff is key information for many water related hazards and management of water resources and has been the focus of numerous studies in hydrology over the past several decades. Nevertheless, an accurate prediction of runoff has since remained a challenge due to non linearity of the several involved surface and subsurface processes (Kachroo and Natale, 1992; Phillips, 2003). Promising continuous runoff models based on Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) were first introduced by Kratzert et al. in 2018. Its very successful first application has since then encouraged many researchers to explore more widely the predictive capability of LSTM based runoff models. Some examples include Kratzert et al. (2019a, b); Gao et al. (2020); O et al. (2020); Feng et al. (2020); Frame et al. (2021); Gauch et al. (2021a, b); Lees et al. (2021); Nearing et al. (2021). Unlike traditional conceptual rainfall-runoff models where hydrological rules are hardwired into them, LSTM based models borrow their principles from the fields that are not native to hydrology. A central interest is thus whether and how we can benefit from domain knowledge and traditional practices in hydrology when using LSTM models for prediction of runoff. This paper is about some paths towards this goal.

PATH 1 – Conforming to the daily runoff model from Kratzert et al. (2018), the LSTM takes a “sequence” of past forcing variables to predict runoff. Its sequence type input reflects the distinct property of LSTM — capturing time dependencies. In the previous studies by Kratzert et al. (2018) and Lees et al. (2021), the length of this sequence, hereafter called lookback, was set to 365 $[\mathrm{day}]$ so that the dynamics of a full annual cycle could be captured. Kratzert et al. (2019b) tested four lookbacks (90, 180, 270 and, 365 $[\mathrm{day}]$) and reported that a lookback of 270 $[\mathrm{day}]$ gave the best results in their study. However, Gauch et al. (2021b) systematically reduced the size of the data and showed that the choice of lookback should be made taking into account the amount of data — when the available data were limited, a too long lookback could impair LSTM performance. From the point of view of pure Deep Learning, lookback is a hyperparameter, just like batch size, learning rate, and so forth. However, we can find some compelling reason to separate lookback from usual hyperparameters. The catchment response is known to depend on the current soil-water state of the catchment, which is itself a result of antecedent conditions and forcing history — succession of dry/wet, cold/hot periods, and so on. However, this dependence is time limited; what has happened in the past is progressively forgotten by the catchment, and, over time, it will have no (or very limited) influence on the current conditions. It is also known that each catchment has its own degree of hysteresis (memory length), which depends on the dominant hydrologic process of the catchment as well as the catchment size. For instance, large catchments connected to major ground aquifers can have a very long memory up to several years (de Lavenne et al., 2021). Or, small catchments located on the surface of an impermeable bedrock with no infiltration can have a very short memory of only a few days. We thus naturally expect the choice of lookback to depend, not only on the length of training data as shown by Gauch et al. (2021b), but also on the hydrologic characteristics of the catchment. We can thus define our first research question, which is largely unaddressed in the existing literature, as follows: *“Q1- does the LSTM performance-lookback pattern depend on the regime of catchment?”*

Deep Learning context of PATHs 2 & 3 – We can decompose the error associated to any Deep Learning (including LSTM) network to the following three components (Beck et al., 2022): 1) approximation error, 2) generalization error, 3) optimization error. The approximation error is the error of the network in approximating the true underlying mapping function. This error is controlled by model representational capacity (which model architecture and family: LSTMs? CNNs? ANNs?), the choice, and the number of input features (Goodfellow et al., 2016). The generalization error is the error of the network on unseen data. The optimization error is the error of the optimization algorithm in finding the global minimum of the loss function. This error results from the optimization algorithm. The training and validation errors that the learning algorithm visits during training are a reflection of approximation + optimization and generalization errors, respectively. But the training and validation errors are only "expectations" or "estimates" of the true errors — since they are computed only on "a finite number" of samples drawn from the distribution of inputs the system is expected to visit in practice. As the number of training examples increases, the network's learning can be refined given the more accurate losses. This provides intellectual justification to account for data size as a model independent factor controlling the performance of the model. Let's assume that the model family and architecture, and the optimization algorithm are fixed to some given choices — all errors associated to theme are thus considered to remain unchanged. In this paper, we want to alter other error controlling variables (i.e. model's feature, data size, data homogeneity) — in ways conforming to traditional hydrologic practices — and study how LSTM performances change. PATH 2 allows to investigate the influence of model's feature and data size. Through PATH 3, we intend to see the influence of data homogeneity.

PATH 2 – Conforming to classical regionalisation (Oudin et al., 2008), we pass from individually trained (local) to group trained (regional) LSTMs. In this passage, we incorporate static feature (into regional LSTMs) — thus both data size and model capacity increase. Bigger data and a higher capacity improve the training error and model precision, but, is it without losing some generalization? This path allows us to form our second question: *"Q2- how well does the LSTM trade generalization for precision from local to regional training?"*

Local and regional LSTMs have been already investigated and compared against multiple conceptual models in several studies. See Kratzert et al. (2018) for examples of local LSTMs (compared against the conceptual SAC-SMA+Snow-17 model). For examples of regional LSTMs see Kratzert et al. (2019b), or Lees et al. (2021), and Gauch et al. (2021a). Kratzert et al. (2018) discuss that, in regional training, not only the training data are significantly augmented, but also different contributing catchments would bring different complementary information about rainfall-runoff transformation under more general hydrological conditions and, consequently, learning would improve. Kratzert et al. (2019a, b) showed that their regional LSTMs that used both dynamic (e.g. forcing data) and static (e.g. catchment attributes) features outperformed the regional LSTMs that did not take any static features, as well as, all tested local conceptual benchmark models. Later, Lees et al. (2021) observed also an outperformance of regional LSTM models over their four conceptual benchmark models in the climatic context of Great Britain and on a sample of 518 catchments. However, in any of the previous studies, local LSTMs have not been compared with regional LSTMs with static attributes.

PATH 3 – Looking for benefiting from the traditional practice of hydrologic classification (Haines et al., 1988; Chiverton et al., 2015), we here investigate hydrologically homogeneous versus hydrologically heterogeneous training, at the regional scale. Classification of catchments according to their hydrologic behavior naturally conveys the idea that all catchments in the same class are hydrologically similar to each other and thus have the same behavior; or same "representation" in the language of Deep Learning. Does this also confer the learning advantage for the LSTM that the representation of the shared behavior can be captured by a single training on the data of the class? This is the main motivation of this path where we compare regional LSTMs under two conditions: a) when the training examples are more but collected from distributions that are very different in horologic statistics (heterogeneous national training set) and b) when the training examples are much less but drawn from hydrologically similar distributions (homogeneous regime training set). In this comparison, the model capacity/complexity remains the same, the size of the training data increases, and the complexity of the latent rules to be learned varies due to the difference in heterogeneity/homogeneity. More specifically, we are interested in answering to the following question: *"Q3- is there a performance gain for regional LSTMs in the passage from hydrologically heterogeneous to homogeneous training, and vice versa?"*

To identify hydrologic similarity, we present a purely hydrologic classification built on three indices obtained from analysis of regime of runoff, precipitation, and temperature. To the time of this writing, only one other study has investigated, in very parallel to this paper, the data homogeneity component in training LSTMs (Fang et al., 2022). However, there are a number of important differences between this paper and the study by Fang et al. Their study is conducted for the U.S. context. They use the "ecoregion based" classification of Omernik and Griffith (2014), which is built on geological, land form, soil, vegetation, climatic, land use, wildlife, and hydrologic compositions (Fang et al., 2022). The measure of homogeneity that is used in their experimental design is the "proximity" of ecoregions — the farther the two regions are, the more dissimilar they are. However, this hypothesis has not been always found to be true as it is shown in Oudin et al. (2008). Likewise, this hypothesis is largely contradicted in our classification — we can have very close but totally dissimilar catchments, and vice versa. Besides that their LSTM model is also different in many aspects (a different architecture, number of hidden layers, activation function, loss function), Fang et al. have performed no hyperparameter tuning for lookback (it is fixed to 365 [day]). Also, their number of epochs is predefined and similar for all experiments, which is not either the case in this paper.

PATH 4 – The last investigation path of this paper — inspired by the fine tuning experiment of Kratzert et al. (2018) — is about improving LSTM performances by a way other than increasing the size of the data or model capacity or changing the homogeneity/heterogeneity of the data. Here, we study the change in only our approach to selection of the best value for the two important LSTM hyperparameters — lookback and hidden unit size. Following this path, our last research question is defined as follows: *"Q4- what is the most effective way of using LSTM for making runoff predictions?"*

In following these paths, we apply LSTM to a sample consisting of 361 gauged catchments with very diverse hydrologic conditions from all over France — this paper is the first application of the LSTM to the French context. The discharge time series of the catchments is at least 30 years long ($30 \leq \leq 60$ [year]). In all experiments, the LSTM is tuned with respect to lookback, hidden unit size, as well as dropout rate, and three disjoint subsets (training, validation, and test) are used. We use also the non mass conservative conceptual GR4J model to benchmark the LSTM.

The remainder of this paper is organized as follows. The following section presents the available data and our hydrologic catchment classification. Section 3 details the methods used in this paper and describes the experimental design. Results are provided in Section 4. Research questions of the paper are discussed in Section 5. The paper is concluded in Section 6, which also outlines some future perspectives based on the findings of this study.

2 Data

2.1 Hydro meteorological data

The data set used in this study contains time series of hydro-meteorological variables and time invariant catchment attributes. It is a subset of a larger dataset of 4190 French catchments processed by the HYCAR research unit of INRAE (Delaigue et al., 2020). The meteorological forcing data are produced by the daily SAFRAN (Système d'Analyse Fournissant des Renseignements Atmosphériques à la Neige) reanalysis that is run by Météo France at a resolution of $8{\times}8$ [km^2] (Quintana-Segui et al., 2008; Vidal et al., 2010). For each catchment, spatially averaged forcing data consisting of daily total precipitation, mean, minimum, and maximum air temperature, wind speed, air moisture, atmospheric radiation, and visible radiation are available for the common period from 1958-08-01 through 2019-07-31. Hydrometric data consist of daily time series of discharge and are collected by the French Ministry of Environment covering the period of the forcing data.

The catchment sample of this paper includes 361 catchments from all over France with discharge time series ranging from 30 to 60 [year]. These catchments range in size from 5 to 13806 [km^2] with a median size of 219 [km^2]. Their annual runoff ranges from 47 to 2312 [$\mathrm{mm\ per\ year}$], with a median value of 466 [$\mathrm{mm\ per\ year}$] and their annual total precipitation varies between 621 and 2128 [$\mathrm{mm\ per\ year}$], with a median value of 1053 [$\mathrm{mm\ per\ year}$]. The mean daily temperature of the catchments varies between -1.8 and 14.8 [°C] and has a median value of 9.8 [°C].

2.2 Catchment classification

The classification proposed in this paper uses readily available data and is inspired by Pardé (1933) and Sauquet (2006). It is built on three hydroclimatic indices, namely IQ [-], IP [-], and T_{min} [°C], concluded from the analysis of the interannual monthly signals of runoff (Q [$\mathrm{mm\ per\ month}$]), total precipitation (P [$\mathrm{mm\ per\ month}$]), and temperature (T [°C]). These indices are defined as follows:

$$IQ = \frac{Q_{\mathrm{max}} - Q_{\mathrm{min}}}{Q_{\mathrm{mean}}} \tag{1}$$

$$IP = \frac{P_{\mathrm{max}} - P_{\mathrm{min}}}{P_{\mathrm{mean}}} \tag{2}$$

$$T_{\mathrm{min}} = \min(T_1, ..., T_i) \quad i \in 1, 2, ..., 12 \tag{3}$$

where T_i is the mean annual temperature of month i. $Q_{\max}$ and $Q_{\min}$ are maximum and minimum interannual monthly runoff $[\mathrm{mm\ per\ month}]$, respectively. $P_{\max}$ and $P_{\min}$ are maximum and minimum interannual monthly precipitation $[\mathrm{mm\ per\ month}]$, respectively.

In this definition, the IQ and IP indices give information on runoff variability and precipitation variability throughout the year, respectively. A low value of IQ and IP indicates a uniform distribution of them over the year while a high value reflects the presence of contrasted dry and humid seasons. A low IQ can also imply the presence of ground water effects or reservoirs (natural or artificial) tending to attenuate runoff fluctuations at the catchment outlet. The $T_{\min}$ index is a proxy to determine whether or not precipitation is received as snow during winter. Figure 1 shows the spatial variation of the three indices across France. High IQ levels are fragmented in patches in the west and south east of France. The areas with high IP levels are found on the Mediterranean coast in the south of France and Corsica. Low $T_{\min}$ values occur in the mountainous areas — the Alps in the east, the Pyrenees in the south west, and the Massif Central in the center of France.

Using the defined indices, the following classification criteria are defined and applied to each catchment in the sample to determine its hydrologic regime (Fig. 2):

Nival:	$T_{\min} \leq -2$
Nival-Pluvial:	$-2 < T_{\min} < 0$
Mediterranean:	$T_{\min} \geq 0$ and $IP > 1$
Uniform:	$T_{\min} \geq 0$ and $IP \leq 1$ and $IQ < 1$
Oceanic:	$T_{\min} \geq 0$ and $IP \leq 1$ and $IQ \geq 1$

The location of the catchments within each regime is shown in Fig. 3. It is observed that the regimes are geographically plausible and compatible with the geographical characteristics of the region. For example, the Nival and Nival Pluvial regimes occur in the mountainous ranges, and the catchments of the Mediterranean regime are found along the French Mediterranean coastline and in the Mediterranean island of Corsica. The oceanic catchments are spread across other parts of France, except areas known to have important aquifers which are held by the catchments of the Uniform regime — e.g. the Paris Basin region in the north of France.

For each regime, variations of interannual monthly runoff, total precipitation, and mean temperature is presented in Fig. 4. In the Uniform regime, runoff remains in the range between 4% and 13% of the annual discharge all over the year and no wet or dry period is observed. This is while the other regimes clearly exhibit periods of low and high flows. The Oceanic regime is characterized by low flows during the summer and high flows during the winter. This is due to higher evaporation in summer relative to winter. Total precipitation displays a rather uniform pattern in this regime. For catchments in the Mediterranean regime, high flows have a wider period but are less pronounced compared to the Oceanic regime. However, low flows occur at lower levels as a result of extremely dry summers. Autumn precipitation is abundant in this regime, making autumn a period prone to thunderstorms which could in turn induce sudden flash floods. Pattern of runoff in the Nival class is also recognizable with its snowmelt induced peak in the late spring/early summer where there is a rise in temperature. The Nival Pluvial regime appears to be a combined regime of the Oceanic and Nival regimes with two high flow periods, in autumn and spring.

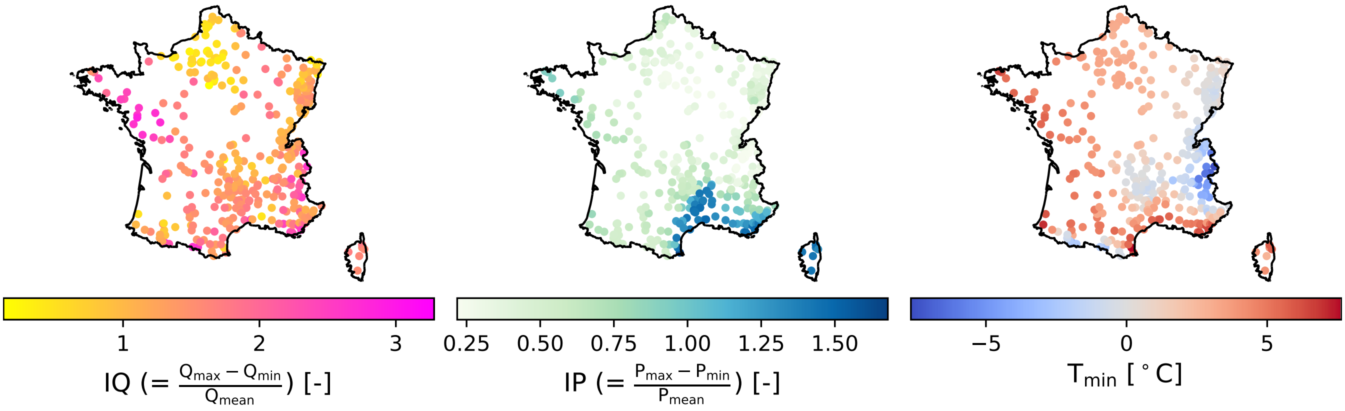
1
2
3
IQ (= (Qmax − Qmin)/Qmean) [-]
0.25
0.50
0.75
1.00
1.25
1.50
IP (= (Pmax − Pmin)/Pmean) [-]
−5
0
5
Tmin [° C]

Figure 1. Spatial variation of IQ, IP, and $T_{\min}$ — the three indices used for hydrologic classification of the catchments. Each catchment in the sample is shown as a point.

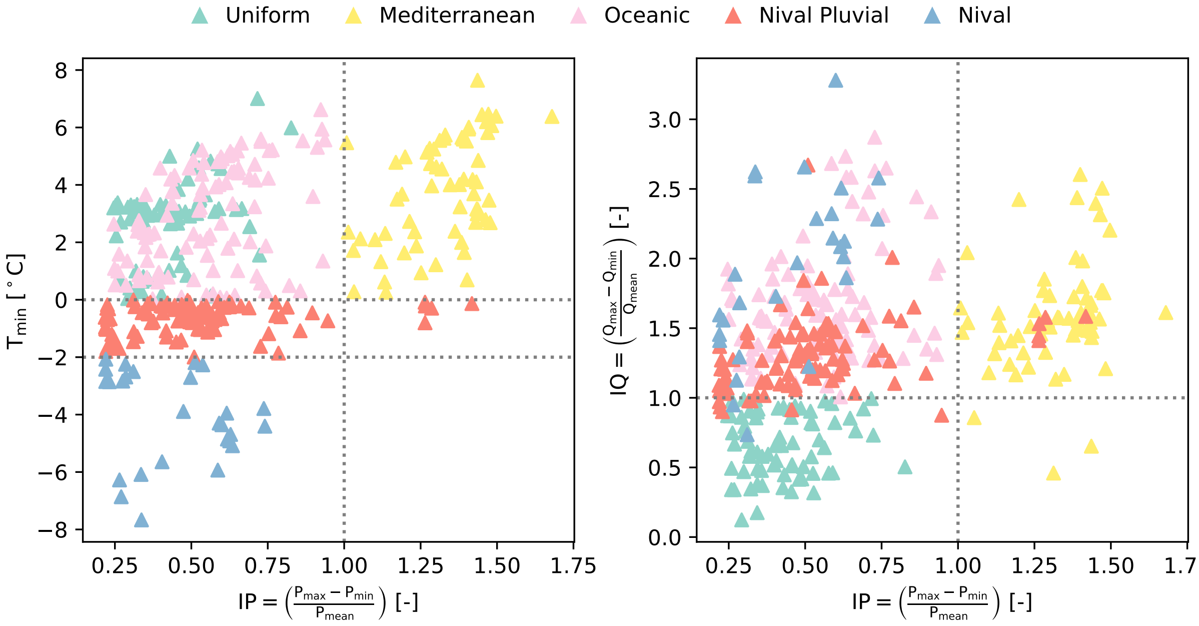

Figure 2. Classification of the catchments into five hydrologic regimes based on five conditions built on $T_{\min}$, IP, and IQ. In order of priority, we first evaluate two $T_{\min}$ conditions — whether $T_{\min} < -2$, if not, whether $T_{\min} < 0$. For catchments not satisfying any of these two conditions, we then check whether $IP > 1$. If this condition is not either satisfied, the $IQ > 1$ condition will be evaluated. Each point represents a catchment and its colour indicates its regime.

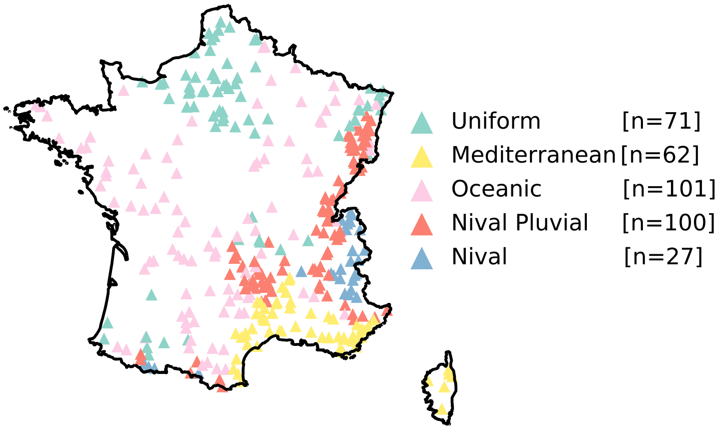

Figure 3. Location of the catchments within each of the five Uniform, Mediterranean, Oceanic, Nival Pluvial, and Nival regimes across France. Each point represents one catchment and is coloured according to its regime.

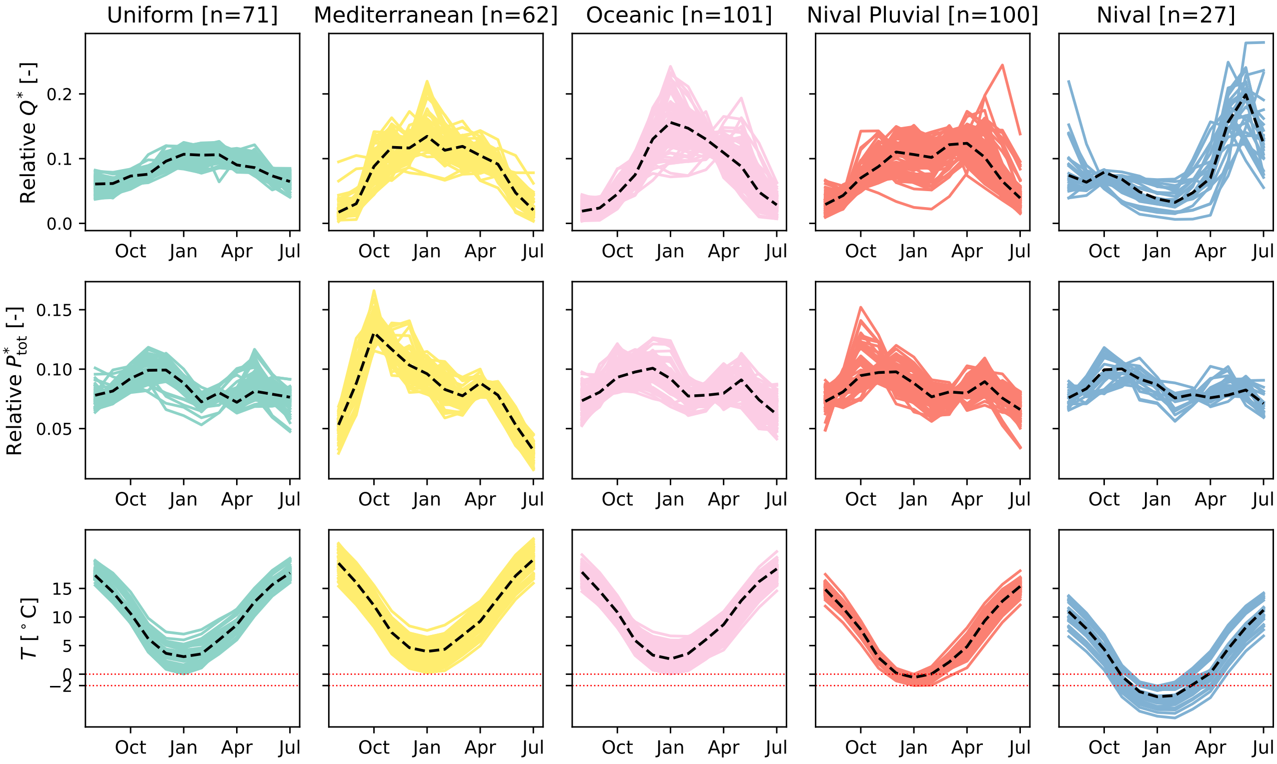

Figure 4. Interannual monthly (≡ Regime of) runoff (Q^*) [−], total precipitation (P^*_{tot}) [−], and temperature (T) for the catchments within different hydrologic classes. The (*) symbol in Q^* and P^*_{tot} indicates that values of these two variables are relative to the total annual amount. Each solid line reflects one catchment. The black dashed line in each panel represents the panel's median regime.

2.3 Catchment physical and climatic attributes

In this paper, we use four physical attributes — surface area $[\mathrm{km}^2]$, median slope $[\%]$, median drainage density $[\%]$, and median altitude $[\mathrm{m}]$ — as well as six climatic attributes — IP, IQ, $T_{\min}$, mean daily liquid precipitation (P_{liq}) $[\mathrm{mm\ per\ day}]$, mean daily solid precipitation (P_{sol}) $[\mathrm{mm\ per\ day}]$, and mean daily potential evapotranspiration (PET) $[\mathrm{mm\ per\ day}]$. The quartile distribution of the physical attributes and P_{sol}, P_{sol}, and PET is shown in Fig. 5 and Fig. 6. We note that surface areas in all regimes are spread between the four quartiles. That is, all regimes have catchments from almost all four quartiles. This is not, however, the case for other attributes. For instance, catchments having the highest 25% of values of altitude or slope are more likely to belong to the Nival or Nival Pluvial regimes. Similarly, it is more probable that catchments with the lowest 25% of drainage densities belong to the Uniform regime than to the Nival or Mediterranean regimes. Compatible with the features of the regime, Nival catchments have significant snow days. Nival Pluvial catchments have both major snow and rain days. Mediterranean catchments are of high evapotranspiration and rainfall rates.

3 Method

3.1 A primer in Long Short Term Memory (LSTM)

LSTM networks are a family of RNNs that address both issues of vanishing and exploding gradients (Hochreiter, 1998). They have proven well suited to modeling a time dependent system where there can be “unknown lags” between the response of the system to a continuous input to it. This is the case for the transformation of rainfall into runoff in a catchment. In the language of LSTM, capturing time dependencies can be translated as sharing important information between time steps of a time sequence (Goodfellow et al., 2016). Information sharing of RNNs is supposed to be deep — i.e. between the time steps that are distant. However, in practice, this occurs only at a shallow level due to the vanishing gradient problem. The LSTM is designed, in turn, to allow for both shallow and deep information sharing across a sequence. In the following paragraphs, we give a quick review of the forward propagation equations of a standard LSTM cell for time step t. For a comprehensive description of LSTM networks, we refer the reader to Chapter 10 of the textbook of Goodfellow et al. (2016). Figure 7 illustrates an unfolded computational LSTM cell corresponding to the last time step (t) of a sequence of length T (so including time steps $t-T+1$ to t). This sequence reflects one sample in a (mini) batch.

The standard LSTM involves two feedback connections operating at different time scales: the shallow level hidden state ($\boldsymbol{h_t}$), for capturing short term dependency details, and the deep level cell state ($\boldsymbol{C_t}$), for transferring information from the distant past to the present in a more effective way than the hidden state — thanks to its self loop structure. The equation of this self loop is the core equation of the LSTM and is as follows:

$$\boldsymbol{C}_t = \boldsymbol{f}_{\mathrm{t}} \odot \boldsymbol{C}_{\mathrm{t}-1} + \boldsymbol{i}_{\mathrm{t}} \odot \tanh(\mathbf{W}_{\mathrm{xc}}^{\intercal} \boldsymbol{X}_{\mathrm{t}} + \mathbf{W}_{\mathrm{hc}}^{\intercal} \boldsymbol{h}_{\mathrm{t}-1} + \boldsymbol{b}_{\mathrm{c}}) \tag{4}$$

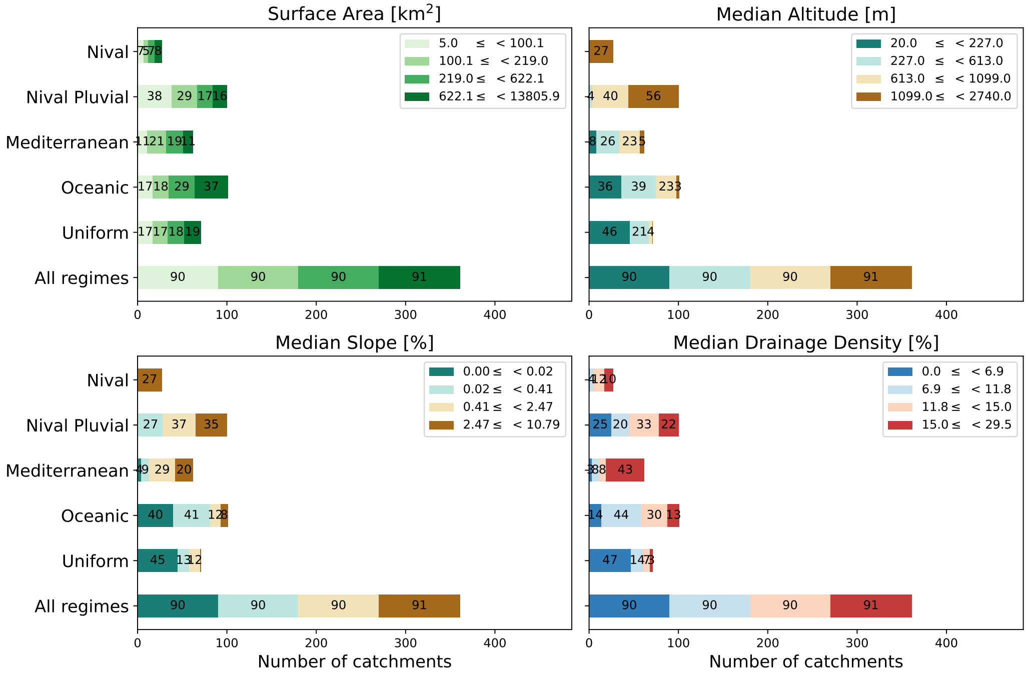

Figure 5. Stacked bar charts showing the variation of the four physical attributes used in this paper within each regimes and the entire sample. The end to end segments of each bar correspond to the intervals delimited by the quartiles of the physical attribute of interest. The quartiles are computed taking all 361 catchments into account. The number inside each segment is its length.

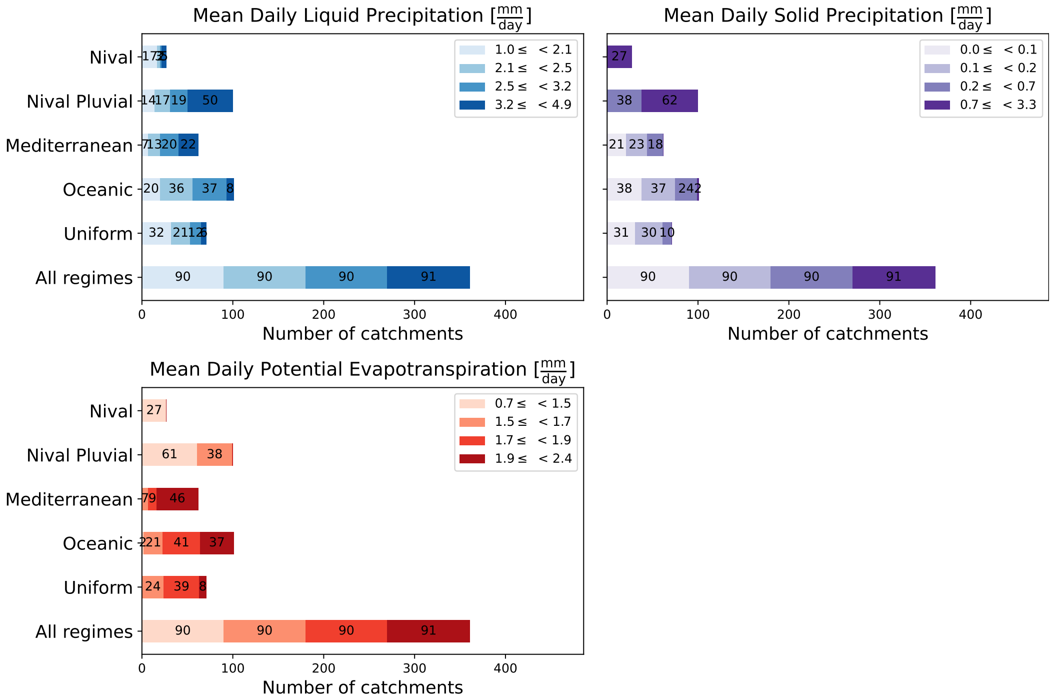

Figure 6. Stacked bar charts showing the variation of the three climatic attributes used in this paper within the regimes and the sample. The end to end segments of each bar correspond to the intervals delimited by the quartiles of the climatic attribute of interest. The quartiles are computed taking all 361 catchments into account. The number inside each segment is its length.

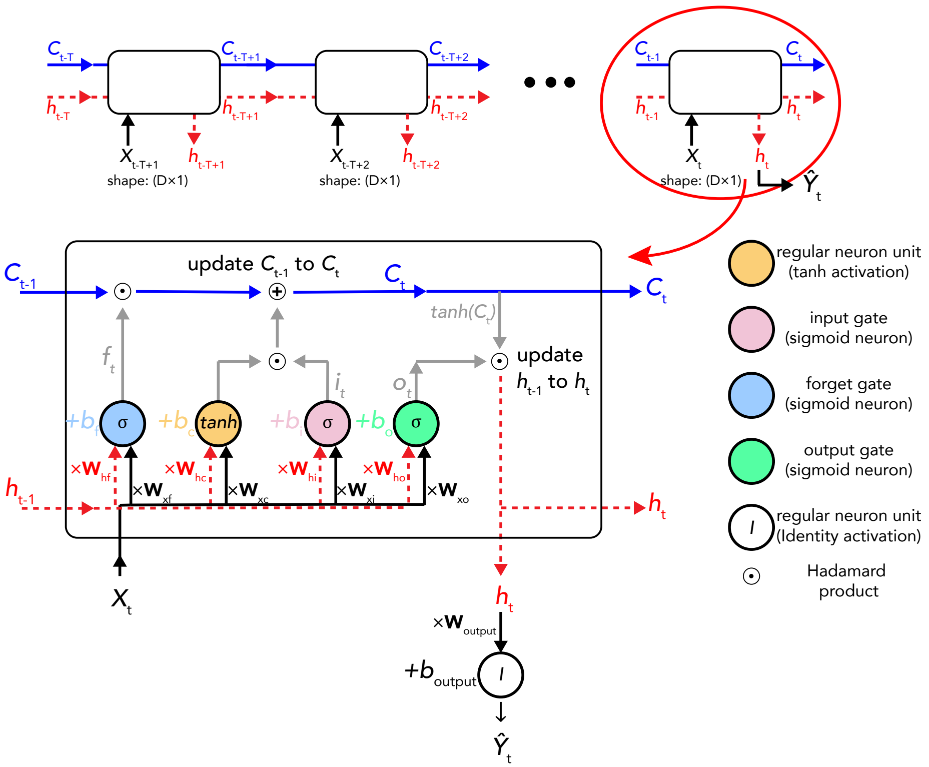

Figure 7. Time unfolded schematic representation of data processing of one time step of the sequence in 1 sample through an LSTM cell. $\mathbf{X}_{\mathrm{t}}$ is the input of time step t. $\boldsymbol{h}_{\mathrm{t}}$ is the hidden state (dashed red line) and $\boldsymbol{C}_{\mathrm{t}}$ represents the cell state (solid blue line). σ and $\tanh$ are the sigmoid and hyperbolic tangent activation functions.

It describes that the cell state is a linear self loop of form $\boldsymbol{C}_t := A\ \boldsymbol{C}_{\mathrm{t}-1} + B$ with $A := \boldsymbol{f}_{\mathrm{t}}$ and $B := \boldsymbol{i}_{\mathrm{t}} \odot \tanh(\mathbf{W}_{\mathrm{xc}}^{\mathsf{T}} \boldsymbol{X}_{\mathrm{t}} + \mathbf{W}_{\mathrm{hc}}^{\mathsf{T}} \boldsymbol{h}_{\mathrm{t}-1} + \boldsymbol{b}_{\mathrm{c}})$. $\boldsymbol{f}_{\mathrm{t}}$ is called forget gate and has the following definition and properties:

$$\boldsymbol{f}_{\mathrm{t}} = \sigma(\mathbf{W}_{\mathrm{xf}}^{\mathsf{T}} \boldsymbol{X}_{\mathrm{t}} + \mathbf{W}_{\mathrm{hf}}^{\mathsf{T}} \boldsymbol{h}_{\mathrm{t}-1} + \boldsymbol{b}_{\mathrm{f}}) \tag{5}$$

1. It is a unit analogous to a neuron in nature: it takes 1) a weighted sum of its inputs ($\boldsymbol{X}, \boldsymbol{h}$) and a vector of bias ($\boldsymbol{b}$); 2) it applies an element wise non linearity ($\equiv$ activation function) to their sum.

2. Its non linear function is the sigmoid function (σ) and outputs values between 0 and 1 — $\boldsymbol{f}_{\mathrm{t}} \in (0, 1)$. Its "gate" functionality comes from this property, with 0 to completely throw away information and 1 to fully retain information.

3. The presence of the term $\mathbf{W}\boldsymbol{X}_{\mathrm{t}}$ reflects a conditioning on the inputs of the current time step ($\boldsymbol{X}_{\mathrm{t}}$). Therefore, $\boldsymbol{f}_{\mathrm{t}}$ is a function of $\boldsymbol{X}_{\mathrm{t}}$ and different for different time steps. The weights $\mathbf{W}$ and bias $\boldsymbol{b}$ are independent of the inputs and are shared between the time steps of the sequence.

$\boldsymbol{h}_{\mathrm{t}-1}$ in Eq. (5) is the hidden state of the previous time step $(\mathrm{t}-1)$ and is defined as follows:

$$\boldsymbol{h}_{\mathrm{t}} = \boldsymbol{o}_{\mathrm{t}} \odot \tanh\left(\boldsymbol{C}_{\mathrm{t}}\right) \tag{6}$$

where $\boldsymbol{o}_{\mathrm{t}}$ is called output gate and has the following definition:

$$\boldsymbol{o}_{\mathrm{t}} = \sigma\left(\mathbf{W}_{\mathrm{xo}}^{\intercal} \boldsymbol{X}_{\mathrm{t}} + \mathbf{W}_{\mathrm{ho}}^{\intercal} \boldsymbol{h}_{\mathrm{t}-1} + \boldsymbol{b}_{\mathrm{o}}\right) \tag{7}$$

$\boldsymbol{o}_{\mathrm{t}}$ has the exact same properties as of $\boldsymbol{f}_{\mathrm{t}}$.
So far, we have provided the definition of all terms in Eq. (4), except for $\boldsymbol{i}_{\mathrm{t}}$. It is called input gate and is given by:

$$\boldsymbol{i}_{\mathrm{t}} = \sigma\left(\mathbf{W}_{\mathrm{xi}}^{\intercal} \boldsymbol{X}_{\mathrm{t}} + \mathbf{W}_{\mathrm{hi}}^{\intercal} \boldsymbol{h}_{\mathrm{t}-1} + \boldsymbol{b}_{\mathrm{i}}\right) \tag{8}$$

Similar to the other gates and as Eq. (8) suggests, it shares all of the properties mentioned above for $\boldsymbol{f}_{\mathrm{t}}$. The current output — which depends on the T last hidden states — is then computed as follows:

$$\widehat{\boldsymbol{Y}}_{\mathrm{t}} = \mathbf{W}_{\mathrm{output}}^{\intercal} \boldsymbol{h}_{\boldsymbol{t}} + \boldsymbol{b}_{\mathrm{output}} \tag{9}$$

The notation, shape, and definition of the different variables in the LSTM's forward pass equations are given in Table 1.

Table 1. Notation, shape, and definition of the terms and operators of Eq. (4) to Eq. (9) for the forward pass of a standard LSTM cell involving forget, input, and output gates.

Notation	Shape	Definition
$\boldsymbol{X}_{\mathrm{t}}$	$D \times 1$	input for time step t for a given sample
$\widehat{\boldsymbol{Y}}_{\mathrm{t}}$	1×1	output vector for time step t for a given sample
$\mathbf{W}_{\mathrm{xf}}, \mathbf{W}_{\mathrm{xi}}, \mathbf{W}_{\mathrm{xo}}, \mathbf{W}_{\mathrm{xc}}$	$D \times M$	inputs to forget, input, output gate weights, inputs to regular neuron unit weights
$\mathbf{W}_{\mathrm{hf}}, \mathbf{W}_{\mathrm{hi}}, \mathbf{W}_{\mathrm{ho}}, \mathbf{W}_{\mathrm{hc}}$	$M \times M$	hidden state to forget, input, and output gate weights, hidden state to regular neuron unit weights
$\mathbf{W}_{\mathrm{output}}$	$M \times 1$	hidden state to output weights
$\boldsymbol{b}_{\mathrm{f}}, \boldsymbol{b}_{\mathrm{i}}, \boldsymbol{b}_{\mathrm{o}}, \boldsymbol{b}_{\mathrm{c}}$	$M \times 1$	forget, input, output gate biases, and regular neuron unit bias
$\boldsymbol{b}_{\mathrm{output}}$	1×1	output bias
$\boldsymbol{f}, \boldsymbol{i}, \boldsymbol{o}$	$M \times 1$	forget, input, output gates
$\boldsymbol{h}, \boldsymbol{C}$	$M \times 1$	hidden state, cell state
$\sigma()$		sigmoid function
$\tanh()$		hyperbolic tangent function
$\odot$		linear algebra element wise (Hadamard) product
$\intercal$		linear algebra transpose operator

T = sequence length of each sample (= lookback)
D = number of features — dynamic+static — for each sample
M = number of hidden units of the LSTM layer

3.2 Training, validation, and test data sets

The period of full record discharge differs between the catchments in the sample. To obtain training, validation, and test data sets, we split the data of each catchment according to its own full record period. We obtain the three sets as follows. The first period from the end containing 10 years of full record discharge is set as the test period. The next subsequent period that contains 10 years of full record discharge is set as the validation period. What remains constitutes the training period, the length of which varies between 10 to 40 $[\mathrm{year}]$ in the sample.

Since the features and the target vary in wide ranges of values, we perform a feature wise standardization for the features and the target. The standardization is performed using the mean and the standard deviation of the training data. This form of standardization — centering the input data around 0 and scaling them by the standard deviation — is also used by Kratzert et al. (2018) and is appropriate for runoff simulation using LSTM. LeCun et al. (2012) explain why this form of standardization works generally well by making the gradient descent converge faster. Also, since the temperature feature can take negative values, we can not in principle benefit from a [0, 1] scaling. Last but not least, the useful area of the LSTM's activation functions (sigmoid and hyperbolic tangent functions) — i.e. where their derivatives are most dynamic — is an area centered around 0. This form of standardization could thus help to fall in this area where the weights can get updated more effectively.

3.3 Criteria for performance evaluation

In this paper, to evaluate runoff prediction performances, we use the Kling–Gupta Efficiency (KGE) score (Gupta et al., 2009) since it combines the three fundamental diagnostic properties of a predictive hydrologic model, i.e. variability (α), bias (β), and linear correlation (r).

$$\mathrm{KGE} = 1 - \sqrt{(1-\alpha)^2 + (1-\beta)^2 + (1-r)^2} \tag{10}$$

$$\alpha = \frac{\mathrm{std}(\widehat{\mathbf{Y}})}{\mathrm{std}(\mathbf{Y})} \tag{11}$$

$$\beta = \frac{\overline{\widehat{\mathbf{Y}}}}{\overline{\mathbf{Y}}} \tag{12}$$

$$r = \frac{\sum_{\mathrm{n}=1}^{N_P} \left(\mathbf{Y}_{\mathrm{n}} - \overline{\mathbf{Y}}\right)\left(\widehat{\mathbf{Y}}_{\mathrm{n}} - \overline{\widehat{\mathbf{Y}}}\right)}{\mathrm{std}(\widehat{\mathbf{Y}}) \times \mathrm{std}(\mathbf{Y})} \tag{13}$$

Where $\widehat{\mathbf{Y}}$ and $\mathbf{Y}$ are predicted and true values, respectively. $\overline{\widehat{\mathbf{Y}}}$ and $\overline{\mathbf{Y}}$ are the mean values of $\widehat{\mathbf{Y}}$ and $\mathbf{Y}$, respectively. std is the standard deviation function and N_p is the number of time steps in the period for which we want to calculate the KGE. For instance, if we are interested in calculating the KGE on the training data set, N_p will be the number of time steps the training data contain. The calculation of the KGE score is catchment wise throughout the paper.

3.4 Hyperparameter tuning

When addressing a research question using a DL model, it is important to limit, as much as possible, potential conclusion biases due to using a not hyperparameter tuned model. LSTM has, in particular, two interconnected hyperparameters that need to be tuned together — lookback and hidden unit size. For this purpose, for each of the LSTMs of the paper, we have tested all combinations of all variations of the hyperparameters listed in Table 2 — 6 (lookback variations)$\times 3$ (hidden unit variations) $\times 3$ (dropout variations) = 54 tuning cases. In all of these cases, the batch size, the number of LSTM layers, and the learning rate are kept the same — 128, 1, and 10^{-4}, respectively.

Table 2. List and variations of the hyperparameters tested for all LSTMs of the paper

Hyperparameter	**Lookback** $[\mathrm{day}]$	**Hidden Unit**	**Dropout Rate**	**Batch Size**	**Number of LSTM Layers**	**Learning Rate**
Variations	30, 60, 90, 180, 365, 730	64, 128, 256	0.0, 0.2, 0.4	128	1	10^{-4}

3.5 Model training and selection of the best hyperparameter set

Here we want to train an LSTM that takes the past T time steps of $\boldsymbol{X}_{\mathrm{t-T+1}},\ldots,\boldsymbol{X}_{\mathrm{t}}$ as inputs ($\mathbf{X}$) to output $\widehat{\boldsymbol{Y}}_{\mathrm{t}}$ — i.e. runoff at time step t $[\mathrm{mm\ per\ day}]$. The input thus necessarily contains sequences of length T of a number of time varying forcing variables (dynamic features). In some cases, we wish to use also time invariant variables (static features), such as physical or climatic catchment attributes. Kratzert et al. (2019b) proposed a variant of LSTM — EA LSTM — that is able to treat static features separately from the dynamic ones. We here use the vanilla LSTM and the simplest way to integrate static features, i.e. to repeat each static feature T times to get its corresponding sequence and then concatenate the obtained sequences with $\mathbf{X}$. Doing this and assuming that D is the total number of features, we will have $\mathbf{X}_{\mathrm{T\times D}}$. The complete list of dynamic and static feature used in the paper is presented in Table 3. Given $\mathbf{X}_{\mathrm{T\times D}}$ and the set of equations presented in Subsection 3.1, the LSTM can therefore output $\widehat{\boldsymbol{Y}}_{\mathrm{t}}$. If we need to have runoff prediction for more than one time step ($\equiv$ sample), the exact same task can be performed for all N time steps to get N runoff predictions — $\widehat{\mathbf{Y}}_{\mathrm{N\times 1}}$. Note that here N reflects the number of samples in the (mini) batch, or the batch size. Now, the goal to accomplish is to find the best set of weights $\mathbf{W}$ and biases $\boldsymbol{b}$ that map $\mathbf{X}_{\mathrm{N\times T\times D}}$ to $\widehat{\mathbf{Y}}_{\mathrm{N\times 1}}$. By the the best set, we mean the weighs and biases that make the overall difference between the LSTM's runoff predictions and runoff true values minimum. This overall difference can be measured by a loss function $l(\widehat{\mathbf{Y}}_{\mathrm{N\times 1}},\mathbf{Y}_{\mathrm{N\times 1}})$, where $\mathbf{Y}$ represents runoff true values. In other words, the goal is to learn the optimal $(\mathbf{W},\boldsymbol{b})_{\mathrm{opt}}$ so that the loss function is globally minimized: $\{\theta_{\mathrm{opt}}=(\mathbf{W},\boldsymbol{b})_{\mathrm{opt}}\}=\underset{\theta=(\mathbf{W},\boldsymbol{b})}{\operatorname{argmin}}\, l(\widehat{\mathbf{Y}}_{\mathrm{N\times 1}},\mathbf{Y}_{\mathrm{N\times 1}})$, or less formally, $\{\theta_{\mathrm{opt}}\}=\underset{\theta}{\operatorname{argmin}}\, l(\widehat{\mathbf{Y}}(\theta),\mathbf{Y})$. Depending on whether we train the LSTM on only an individual catchment or on a group of catchments, we use the mean squared error (MSE, Equation 14) or the NSE^* (Equation 15, Kratzert et al. (2019b)) as loss function, respectively. The NSE^* is catchment specific and is in particular useful when the input data come from different catchments; thus the discharge variance could vary in a wide range. The NSE^* is normalized with respect to the variance of discharge in each catchment. This will

Table 3. List of the dynamic and static features used in different LSTM models of the paper.

Feature	Nature	Time step	Unit	Notation	Comment
total precipitation	Dynamic	Daily	[mm per day]	P_{tot}	SAFRAN output
wind speed	Dynamic	Daily	[m per second]	WS	SAFRAN output
specific air humidity	Dynamic	Daily	[g per kg]	HU	SAFRAN output
atmospheric radiation	Dynamic	Daily	[joule per cm^2]	AR	SAFRAN output
visible radiation	Dynamic	Daily	[joule per cm^2]	VR	SAFRAN output
minimum air temperature	Dynamic	Daily	[°C]	TN	SAFRAN output
maximum air temperature	Dynamic	Daily	[°C]	TX	SAFRAN output
runoff index	Static	-	[-]	IP	$\frac{P_{\text{max}}-P_{\text{min}}}{P_{\text{mean}}}$
total precipitation index	Static	-	[-]	IQ	$\frac{Q_{\text{max}}-Q_{\text{min}}}{Q_{\text{mean}}}$
minimum monthly temperature	Static	-	[°C]	T_{min}	$\min(T_1,...,T_{12})$
mean daily liquid precipitation	Static	-	[mm per day]	P_{liq}	(1 - solid fraction) $\times$ P_{tot}
mean daily solid precipitation	Static	-	[mm per day]	P_{sol}	solid fraction $\times$ P_{tot}
mean daily potential evapotranspiration	Static	-	[mm per day]	PET	Oudin et al.'s formula
surface area	Static	-	[km^2]	A	-
median altitude	Static	-	[m]	$Z50$	-
median slope	Static	-	[%]	S	-
median drainage density	Static	-	[%]	DD	-

prevent giving smaller or larger weights to catchments with a lower or higher variance.

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N}(\widehat{\mathbf{Y}}_{\text{n}} - \mathbf{Y}_{\text{n}})^2 \tag{14}$$

$$\text{NSE}^* = \frac{1}{B}\sum_{\text{b}=1}^{B}\sum_{\text{n}=1}^{N}\frac{(\widehat{\mathbf{Y}}_{\text{n}} - \mathbf{Y}_{\text{n}})^2}{(s_{\text{b}}+\epsilon)^2} \tag{15}$$

Where B is the number of catchments and s_{b} is the standard deviation of discharge for catchment b, which is computed using discharges of the training data. ϵ in equation NSE* is added to the denominator to prevent division by a value very close to 0 in catchments with a very small discharge variance — i.e. when $s_b \to 0$.

We used the Keras library (Chollet et al., 2015) written in Python 3.8 (Van Rossum and Drake, 2009) to build and train all LSTM models of the paper. The gradient based Adam algorithm (Kingma and Ba, 2017) with a learning rate of 0.0001 (Kratzert et al., 2018; Lees et al., 2021) is used as optimization algorithm in all experiments. All other arguments in the Adam optimization module, including β_1, and β_2 (L^1 and L^2 norms) are kept at their default values. To control overfitting, we use the early stopping algorithm of Keras. An early stopping algorithm does not impose to all simulations the same predefined non traversable number of training epochs. It allows the model to continue to learn as long as its performance (on the validation data) is improving.

We train the LSTM both locally using the data from “individual catchments” and regionally using the data from “a group of catchments”. In local training, the loss function is the MSE and only the dynamic features of Table 3 are used. We call the LSTMs trained on individual catchments SINGLEs; since the data from only a single catchment is used in their training. In regional training, the loss function is the NSE^* and both dynamic and static features of Table 3 are used. Further, in regional training, we once train all catchments together at a national level and once at a regime level using only catchments belonging to the same regime (See Subsection 2.2). We call the national level trained LSTMs “REGIONAL NATIONAL”s and the regime level trained LSTMs “REGIONAL REGIME”s. For each of the SINGLEs, REGIONAL REGIMEs, and REGIONAL NATIONALs, we perform the 54 hyperparameter tuning cases, that is:

- 361×54 individual trainings → for SINGLEs
- 54 group trainings on the 71 catchments of the Uniform regime → for REGIONAL REGIME Uniform
- 54 group trainings on the 62 catchments of the Mediterranean regime → for REGIONAL REGIME Mediterranean
- 54 group trainings on the 101 catchments of the Oceanic regime → for REGIONAL REGIME Oceanic
- 54 group trainings on the 100 catchments of the Nival Pluvial regime → for REGIONAL REGIME Nival Pluvial
- 54 group trainings on the 27 catchments of the Nival regime → for REGIONAL REGIME Nival
- 54 group trainings on the 361 catchments of the sample → for REGIONAL NATIONAL

This gives in total 19818 $(= 361 \times 54 \; + \; 6 \times 54)$ trainings.

So far, we have trained different local and regional LSTMs for 54 hyperparameter sets. Now, we need to choose the best hyperparameter set for the trained LSTMs. For SINGLEs, the only possible approach is to select, for each catchment, its own best set — the hyperparameter set that gives the best KGE on the validation data. But, for REGIONALs, either NATIONALs or REGIMEs, two possibilities exist. We can either identify one best set for each of the training catchments or one best overall set for the entire model. In this paper, we investigate both approaches. Crossing the two local and regional training approaches with the two approaches to selection of the best hyperparameter set as shown in Fig. 8, we obtain five LSTM models. SINGLEs are trained locally and have locally tuned hyperparameters. REGIONALs are trained regionally and their best hyperparameter is regional, as well. HYBRIDs, as their name suggest, are the LSTMs that are trained regionally but their best hyperparameter set is chosen locally. Table 4 gives a summary of the important features of these models.

3.6 Conceptual benchmark model: GR4J

We select the daily lumped GR4J model (Perrin et al., 2003, Génie Rural à 4 paramètres Journalier) and its snowmelt routine CemaNeige (Valéry et al., 2014) to benchmark the LSTM. We opt for GR4J since it is able to account for groundwater

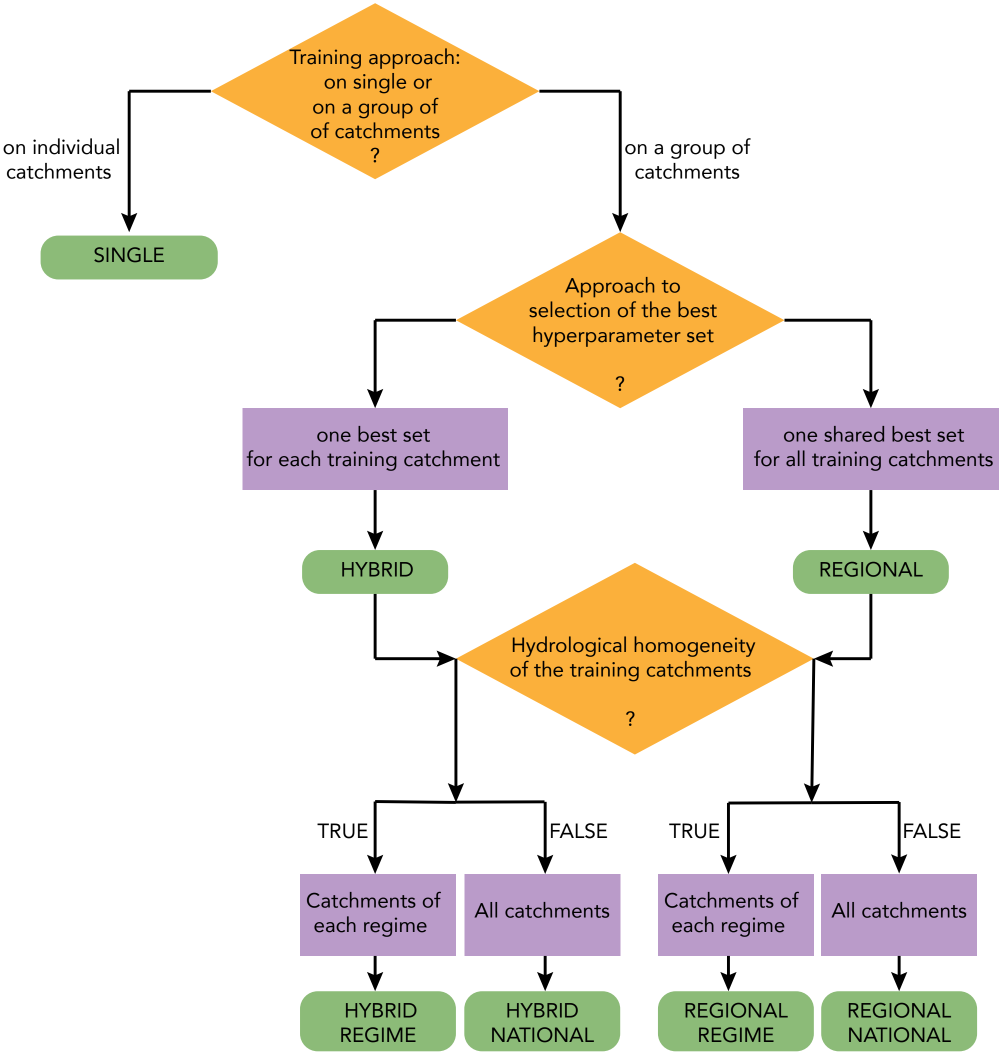

Figure 8. Conceptual flowchart of how SINGLE, REGIONAL, and HYBRID LSTM models and their sub models (green rounded rectangles) are built based on three decision criteria (orange rhombuses) — training approach, approach to selection of the best hyperparameters, and training catchments.

exchanges with aquifers and/or adjoining catchments thanks to its gain/loss function. This is a distinctive feature of GR4J compared against the benchmark conceptual models used in previous studies (Kratzert et al., 2018; Lees et al., 2021).

GR4J is a parsimonious model incorporating only four free parameters. CemaNeige has two parameters and computes snow accumulation and snow melt as outputs (Valéry et al., 2014). We use GR4J coupled with CemaNeige to perform one simulation for each catchment in the sample. It consists of calibrating the coupled model on the training+validation data sets and evaluating it on the test data set. For model calibration, the optimization algorithm Michel (Michel, 1989) is used. For the sake of compar-

Table 4. Name, training catchments, approach to select the best hyperparameter set, and features used for the five LSTM models of the paper

Model	Training catchments	Approach to selection of the best hyperparameter set	Features	Loss
SINGLE	individual catchments	one set for each catchment	All dynamic features of Table 3	MSE
REGIONAL REGIME	catchments in each regime	one shared set for all catchments within the same regime	All dynamic + All static features of Table 3	NSE*
REGIONAL NATIONAL	all catchments together	one shared set for all catchments	All dynamic + All static features of Table 3	NSE*
HYBRID REGIME	catchments in each regime	one set for each catchment	All dynamic + All static features of Table 3	NSE*
HYBRID NATIONAL	all catchments together	one set for each catchment	All dynamic + All static features of Table 3	NSE*

ison with LSTM, the NSE is selected as the objective function for the optimization algorithm. For GR4J, it is recommended to consider a period of warm up to provide the model with an initial state rather than starting with an arbitrary state (Perrin and Littlewood, 2000). Therefore, in all simulations, we set the first 2 years of data as the warm up period when calibrating or evaluating the coupled model. The length of the warm up period corresponds to the longest lookback tested for the LSTM. All GR4J simulations are performed using the airGR package (Coron et al., 2017, 2020) in the R interface (R Core Team, 2019). Compulsory inputs to the GR4J model consist of daily total precipitation $[\mathrm{mm\ per\ day}]$, potential evapotranspiration $[\mathrm{mm\ per\ day}]$ computed using Oudin et al. (2005)'s formula, and runoff $[\mathrm{mm\ per\ day}]$ — where runoff is only used for model calibration. Compulsory inputs to the CemaNeige snowmelt routine are daily total precipitation $[\mathrm{mm\ per\ day}]$ and mean air temperature [°C]. We also use the hypsometric data of each catchment as an optional input for the CemaNeige model. It uses this information to account for orographic gradients (Valéry et al., 2014).

4 Results

The dropout rates $0.2, 0.4$ did not bring performance improvements with respect to the dropout rate 0. All results presented herein correspond to the dropout rate 0.

4.1 Variation of LSTM performances with respect to the length of its input sequence (lookback)

We show in Fig. 9 three curves — the median KGE scores calculated on the training and validation data sets and their average, for SINGLE (panels on the left) and REGIONAL REGIME (panels on the right) LSTMs within the five regimes. For each lookback, the median KGE score corresponds to the best hyperparameter set of that lookback. For instance, for lookback 30 $[\mathrm{day}]$, the hyperparameter sets to select from are the following three sets: (Lookback=30, Dropout=0, Hidden Unit=64), (Lookback=30, Dropout=0, Hidden Unit=128), and (Lookback=30, Dropout=0, Hidden Unit=256). We conjecture that the true underlying performance-lookback pattern lies in between the pattern represented by the training and validation curves. Because the former has the advantage of being used for model training and the latter for hyperparameter selection. For this reason, we choose to look at the average of these two curves.

For both models, the curves tend to show a consistent pattern in different regimes. The median KGE first increases at a certain slope and then from a specific lookback on the KGE remains largely unchanged or even decreases. Both the slope and the

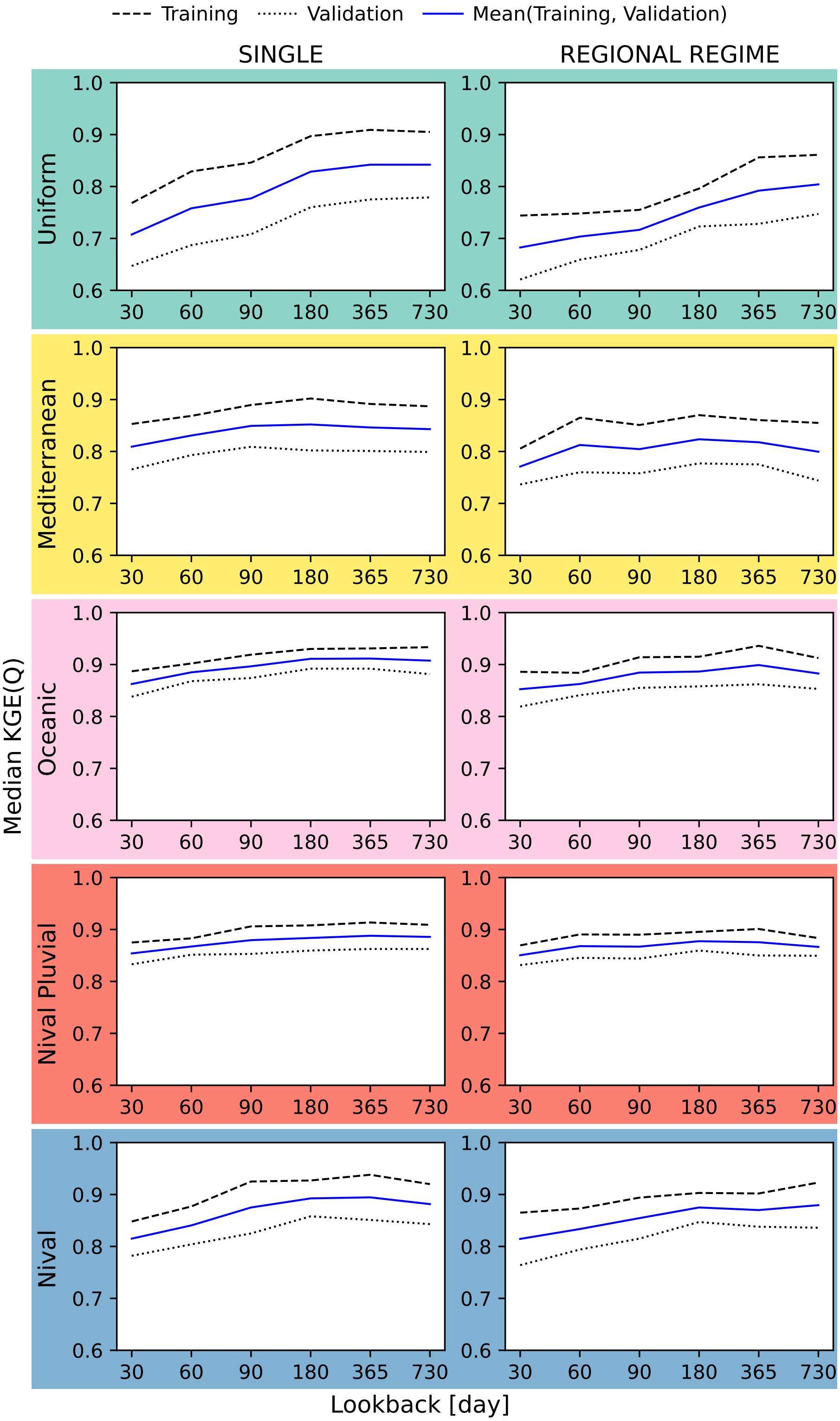

Figure 9. LSTM performance variations with respect to the length of its input sequence within different regimes for the SINGLE and REGIONAL REGIME models. In each panel, the solid, dashed, and dotted lines correspond respectively to the training, validation, and test data. Each line plots the median KGE scores (on the y axis) for different lookback sizes (on the x axis). The median KGE score for a given lookback in a given panel, is the median of KGE scores from the panel's catchments.

lookback appear to be regime dependent. In the Uniform and Nival regimes, the slope is distinctively pronounced for both models — we find the highest sensitivity within these two regimes. In the Mediterranean regime the median KGE varies between 0.81 and 0.85 and between 0.77 and 0.82 for the SINGLE and REGIONAL REGIME models, respectively. The initial slope is more important in this regime than the Oceanic regime and the point after which the KGE stalls occurs earlier. In both regimes, the global sensitivity of performance to lookback size is low. In the Nival Pluvial regime, the starting slope is small — making the pattern almost flat reflecting also low global sensitivity with respect to lookback variations. The range of variation of the median KGE is 0.85-0.89 and 0.85-0.88 for the SINGLE and REGIONAL REGIME models, respectively.

The tendency for performance improvement with lookbacks greater than a year within the Uniform regime, compared against the multi month scale in other regimes, is consistent with the multi year and multi month catchment memory scales concluded by de Lavenne et al. (2021) for the Uniform and non Uniform catchments in the French context.

4.2 Variation of LSTM performances with respect to its training approach

Figure 10 compares the cumulative distribution function (CDF) of the KGE for the locally trained SINGLE, the REGIONAL REGIME, as well as, the REGIONAL NATIONAL LSTMs (see Fig. 8 and Table 4 for their description). Comparing first the median KGE for local training against regional (both regime and national) training, in almost all regimes, regional training outperforms local training. However, except for the Uniform regime, the difference in performance between the SINGLE model and the best REGIONAL model remains slight. Overall, if we take all catchments into account, we have a median KGE of 0.80 for the SINGLE model versus 0.82, and 0.81 for the REGIONAL REGIME and REGIONAL NATIONAL models.

Now, we compare specifically homogeneous group training — REGIONAL REGIME — with non homogeneous group training — REGIONAL NATIONAL. We see that, in the Mediterranean catchments, REGIME has a lower median KGE than NATIONAL while, in the Nival Pluvial regime, it shows a better score. In all other regimes, both trainings have almost the same median KGE. In the Nival Pluvial regime the CDF of the REGIONAL REGIME model is completely shifted towards higher KGE scores. In the Nival regime, although both models have the same median KGE, the CDF curve of the REGIONAL NATIONAL regime is shifted towards better KGEs. Overall, when all catchments are taken into account, the homogeneous group training slightly outperforms the group training with mixed regimes, in terms of the median KGE score. But, their CDFs superpose for high KGEs.

4.3 Variation of LSTM performances with respect to the approach to selection of its best hyperparameter set

Figure 11 compares the CDFs for the group trained REGIONAL and HYBRID LSTMs, which differ in the approach to selection of their best hyperparameter sets. The HYBRIDs thus benefits from the advantage of group training at the same time as using local hyperparameters.

We see that in almost all regimes and overall there is clearly a performance improvement from the REGIONAL NATIONAL model to the HYBRID NATIONAL model. This is in both terms of the median KGE score and the shift of the CDF curve towards better KGEs. However, from the REGIONAL REGIME model to the HYBRID REGIME model, such performance

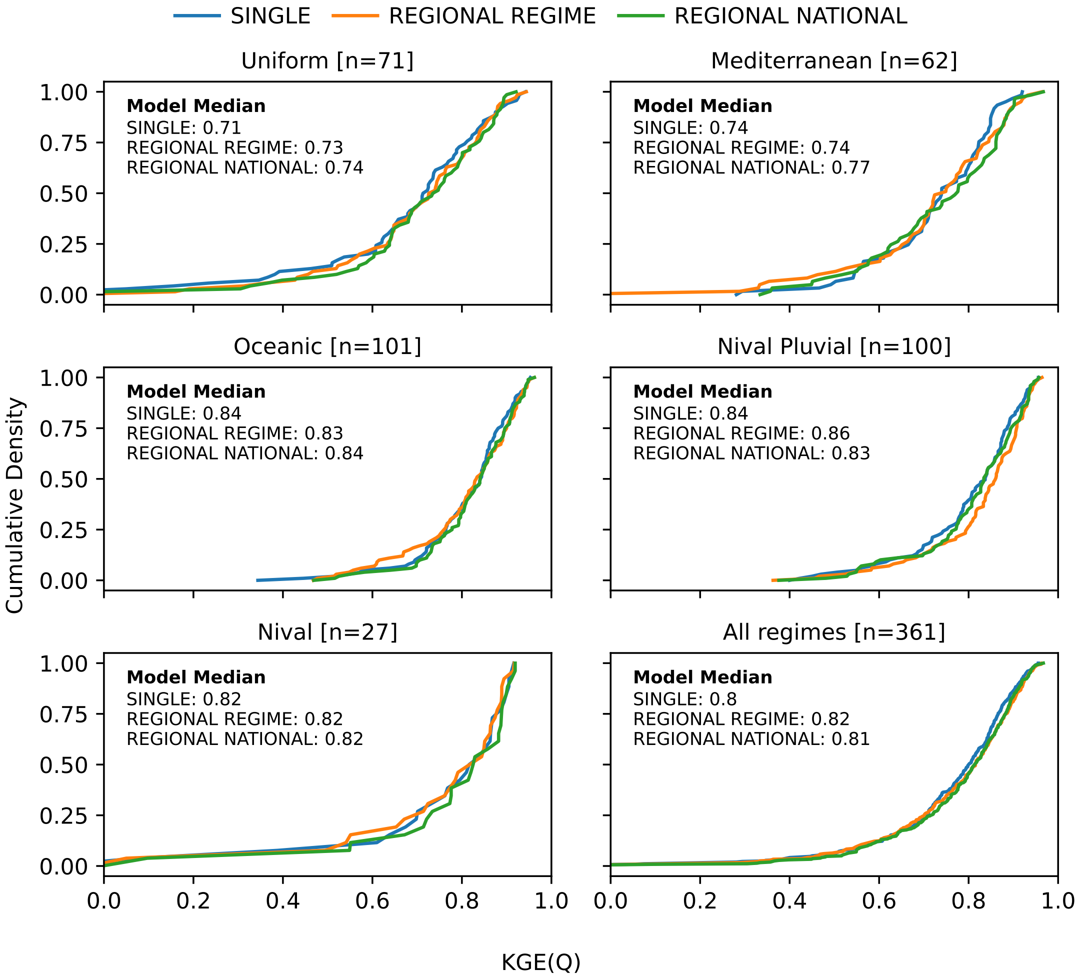

Figure 10. Cumulative distribution functions (CDFs) of the KGE score of the test data for three LSTM models — SINGLE (blue), REGIONAL REGIME (orange), and REGIONAL NATIONAL (green). From top to bottom, the first five panels indicate the CDFs of one of the five regimes — Uniform, Mediterranean, Oceanic, Nival Pluvial, and Nival. The last panel corresponds to the distributions of the entire sample.

improvement does not occur or occurs in a much less pronounced way, except for the Mediterranean regime. The HYBRID NATIONAL model performs best between all tested LSTMs.

4.4 Performance comparison between the LSTM and GR4J model

Table 5 compares the median KGE scores from the GR4J model against the LSTM models for the training+validation as well as test periods. We see in the table that, overall, GR4J is more robust than local and regional LSTMs. Looking at the median KGE score across different regimes for the test period, except for the Uniform and Mediterranean regimes, all LSTMs outperform

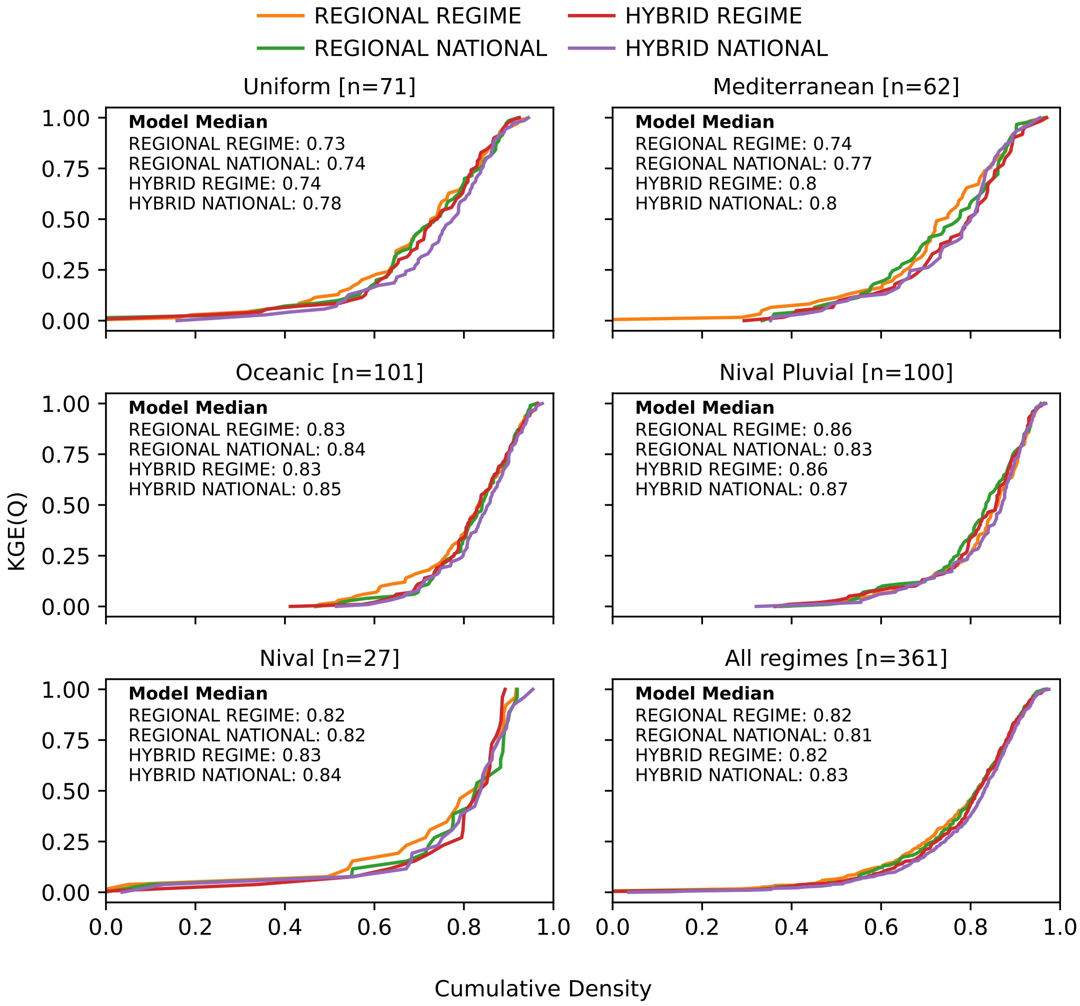

Figure 11. Cumulative distribution functions (CDFs) of the KGE score of the test data for the group trained LSTM models — REGIONAL REGIME (orange), REGIONAL NATIONAL (green), HYBRID REGIME (red), and HYBRID NATIONAL (purple). From top to bottom, the first five panels indicate the CDFs of one of the five regimes — Uniform, Mediterranean, Oceanic, Nival Pluvial, and Nival. The last panel corresponds to the distributions of the entire sample.

GR4J or have the same score — the latter happens in only two cases. In the Mediterranean regime, GR4J outperforms only the SINGLE LSTM. Overall and taking all catchments from different regimes into account, SINGLE and GR4J have the same score; the group trained LSTMs outperform the GR4J model — the performance difference is however small. Group trained LSTMs in previous studies (Kratzert et al., 2019b; Lees et al., 2021) had also better overall performances when compared against conceptual local models, although the LSTM's outperformance in these studies was much more pronounced. One explaining reason is the difference between GR4J and previous studies' conceptual models — including Sacramento Soil Moisture Accounting (SAC-SMA), FUSE, mHM, ARNOVIC, TOPMODEL and PRMS. These models are explicitly mass

Model	Dataset	Uniform	Mediterranean	Oceanic	Nival Pluvial	Nival	All regimes
GR4J	Training+Validation	0.84	0.84	0.89	0.83	0.86	0.85
	Test	0.77	0.75	0.83	0.82	0.75	0.8
SINGLE	Training+Validation	0.87	0.88	0.93	0.91	0.94	0.91
	Test	0.71	0.74	0.84	0.84	0.82	0.8
REGIONAL REGIME	Training+Validation	0.82	0.85	0.92	0.9	0.9	0.89
	Test	0.73	0.74	0.83	0.86	0.82	0.82
REGIONAL NATIONAL	Training+Validation	0.84	0.85	0.92	0.9	0.92	0.89
	Test	0.74	0.77	0.84	0.83	0.82	0.81
HYBRID REGIME	Training+Validation	0.84	0.85	0.93	0.9	0.9	0.89
	Test	0.74	0.80	0.83	0.86	0.83	0.82
HYBRID NATIONAL	Training+Validation	0.86	0.87	0.92	0.91	0.92	0.90
	Test	0.78	0.80	0.85	0.87	0.84	0.83

Table 5. Median KGE scores, within different regimes and overall, for the GR4J model compared against the LSTM models.

conservative — unlike GR4J that is explicitly designed to capture water losses and gains through an exchange parameter (Perrin et al., 2003). GR4J is thus able to simulate runoff in catchments with water balance disclosure. Lees et al. (2021) discuss that when the water balance closure is not satisfied, the LSTM performs better than conceptual models. Figure 12 shows a diagnostic plot of runoff coefficient (=$\frac{Q}{P_{\text{tot}}}$) versus the wetness index WI (=$\frac{P_{\text{tot}}}{PET}$) for the 361 catchments. The points — representing the catchments — are colored by the KGE score. Between the 361 catchments plotted in each panel of Fig. 12, 9 catchments occur in zone z_1 (above the horizontal water limit line). Given that in this zone $Q > P_{\text{tot}}$, there is a surplus in its water balance and it does not therefore close. The z_2 zone (located between the horizontal and curved lines) contains 255 catchments in which the water balance is satisfied. Finally, 97 catchments are found in the z_3 zone (located below the curved line) where the water balance does not close since $\frac{Q}{P_{\text{tot}}} < 1 - \frac{1}{WI}$ and therefore $Q < P_{\text{tot}} - PET$ indicating a potential water deficit. The mini plot inside each panel shows the KGE score of the catchments located in each of the z_1, z_2, and z_3 zones as well as their median value. In z_1 and z_3, where the water balance is not satisfied, the GR4J model has the same or a better median score than z_2 — where the water balance closes. This is in contrast with the corresponding finding of the previous study by Lees et al. (2021). Interestingly, we can make the same but more clear observation for LSTM — the median KGE score for all LSTMs is lower in z_2 than in z_1 and z_3. For catchments occurring in the z_1 zone, the regime LSTMs outperform clearly NATIONAL and SINGLE LSTMs. Within z_3, the group models produce similar scores, which are better than the corresponding scores of the SINGLE model. However, taking into account the 136 catchments having either a water surplus (9 catchments) or deficit (97 catchments) in their water balance, the median KGE scores of LSTM models is better than GR4J: 0.81 (SIMPLE), 0.84 (REGIONAL REGIME), 0.84 (HYBRID REGIME), 0.83 (REGIONAL NATIONAL), and 0.84 (HYBRID NATIONAL)

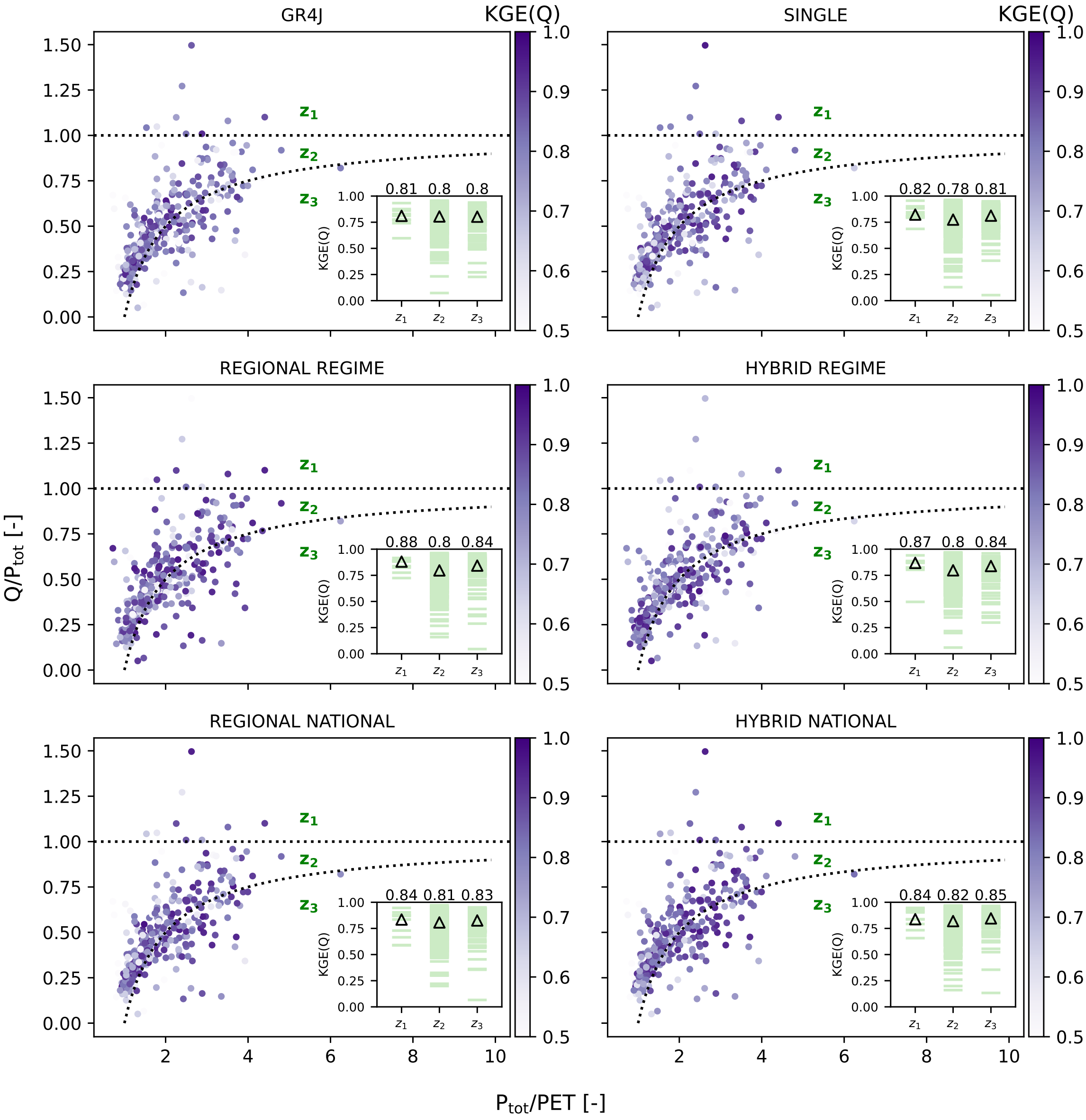

Figure 12. Variation of the KGE score with respect to runoff ratio ($\frac{Q}{P_{\text{tot}}}$) and the wetness index ($\frac{P_{\text{tot}}}{PET}$) for the GR4J and LSTM models. Scores lower than 0.5 are shown in the same color as the lower boundary of the color bar. The mini plot inside each panel shows the KGE score of the catchments located in each of the z_1 (above the horizontal water limit line), z_2 (between the horizontal and curved lines), and z_3 (below the curved line) zones. The △ symbol and numbers in the mini plot represent the median KGE score of the three zones. The KGE scores correspond to the test data excluding the first two years constituting the warm up period in GR4J for which it does not output any results.

versus 0.80 (GR4J). This agrees with the corresponding overall better performances of the LSTM over the four conceptual models in Lees et al. (2021).

5 Discussion

5.1 Does the LSTM performance-lookback pattern depend on the regime of catchment?

We distinguish the Uniform and Nival regimes as the two regimes with the most clean performance-lookback pattern — increasing performances with increasing lookback size. We relate this to the clear hysteretic pattern of their dominant hydrologic process: recharge and discharge of aquifer and thawing of accumulated snow, which are two prime examples of long term dynamics.

Uniform catchments occur mainly in areas known to be highly influenced by large aquifers, such as aquifers of the Seine or the Somme River basins in the north of France (Fig. 3). Such aquifers can significantly modify the temporal dynamics of the impacted catchments and largely hamper correlating runoff with hydro-climatic conditions in the present (Fig. 4). Runoff at the outlet of Uniform catchments can depend on precipitations of several years ago (de Lavenne et al., 2021). In snow dominated catchments, we have storage of precipitation as snow, which is later released (as snow melt) during the late spring/early summer.

In the Mediterranean regime, the performance-lookback pattern is characterized by a narrow spread in KGE scores for different lookbacks, whereas we expected a clear offset for small lookback values. In this regime, internal states (e.g. soil moisture) do not depend on long antecedent periods due to the flash flood generating nature of precipitations in this regime, which are particularly intense in the autumn (Fig. 4). Although, we see a mild tendency to lookback values 90 and 60 $[\mathrm{day}]$, for local and regional LSTMs, at both scales, the KGE scores vary in a narrow range regardless of which lookback is chosen. One explaining reason is that various levels of lookback sensitivity may exist for different catchments within this regimes due to inter-regime differences in non hydrologic characteristics — such as, soil type, bedrock geology, drainage class, and so forth. For examples of such a variability, see Fig13. After a further investigation, we note that many of Mediterranean catchments occur in a karstic region exerting potentially an influence, although very locally, on their degree of hysteresis. We have not further investigated this hypothesis in this paper. But, if that is the case, we can relate the unclear pattern of this regime to the absence of one single dominant process — a combination of dominant processes exists and the degree of combination varies between catchments.

In the Oceanic, and Nival Pluvial regimes, the performance-lookback pattern is very monotonous and there is a far less sensitivity — almost an insensitivity — to lookback when we look at the median of the KGE scores. We relate this to the low degree of hysteresis of the dominant hydrologic processes in these two regimes — they can be correlated relatively well with current precipitation (Fig. 4) and evapotranspiration.

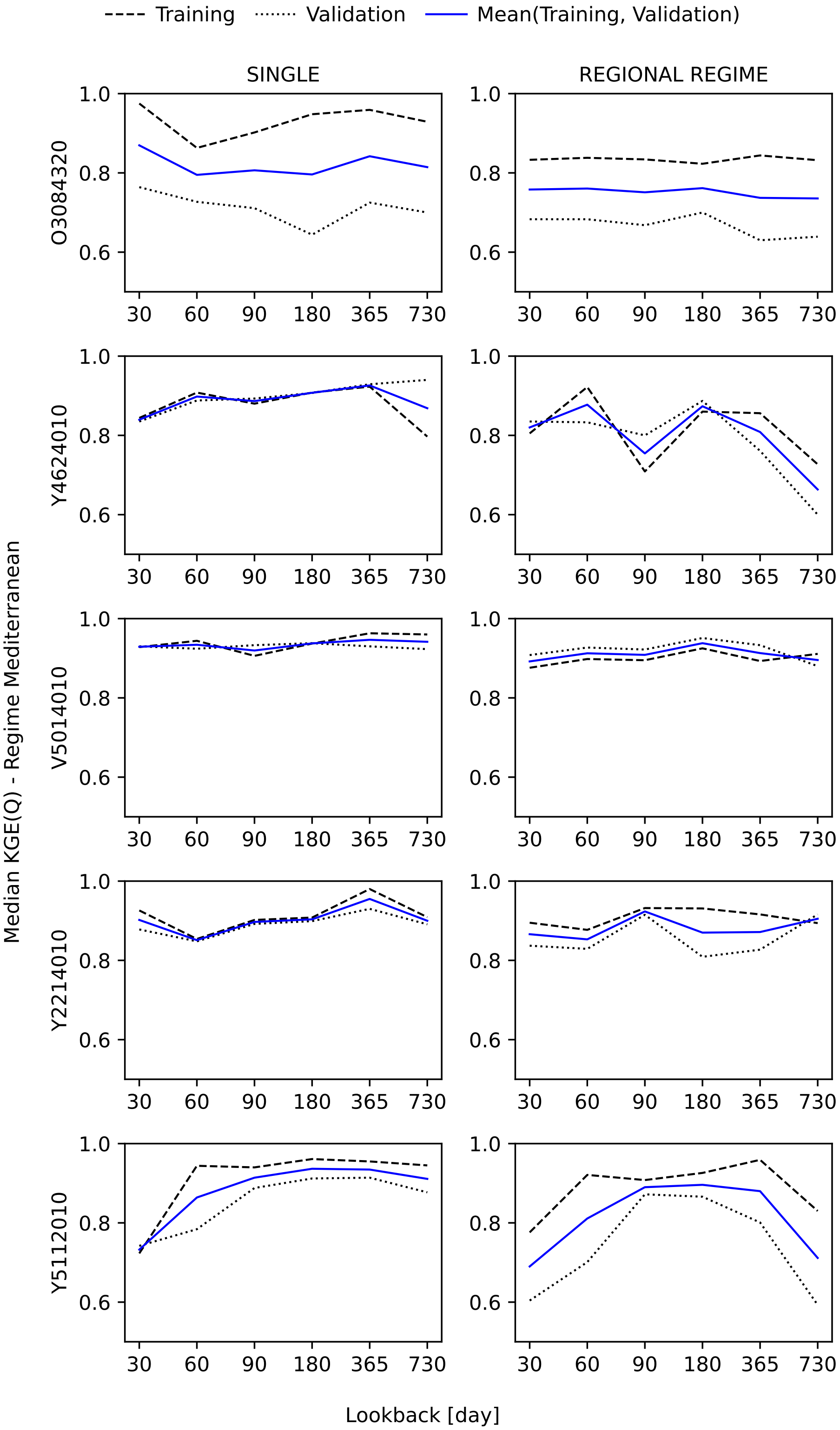

Figure 13. Five examples from the Mediterranean regime, each with a different lookback sensitivity pattern

5.2 How well does the LSTM trade generalization for precision in the passage from local training to regional training?

To answer this question, we need to take into account, SINGLE, REGIONAL REGIME, and REGIONAL NATIONAL LSTMs. From individual catchment (local) training to group (regional) training, we have increased the capacity of the model (by adding 10 static attributes) and the size of the data. In this passage, LSTM performances have improved in almost all regimes and overall. That is, in both local to homogeneous regional and local to heterogeneous regional passages, the precision that the LSTM gains is "almost" always more than the generalization it looses. For Uniform, Mediterranean, and to a lesser extent, Nival Pluvial catchments, the passage from local to at least one of the regional LSTMs, is a real gain. For the two other regimes, the trade is rather a trade-off and performance improvements do not turn out to be significant.

One explaining reason for the small performance difference between local and regional (homogeneous or heterogeneous) trainings is that the available data at the local level has been already large enough with respect to the complexity of catchments representations. The LSTM has thus already asymptoted to an error very close to the minimum possible error. At the regional level, although the amount of data has increased greatly, the resultant of the gained precision, lost generalization, and varied complexity is not remarkably positive to push the final error to a point closer to the minimum possible error. Besides, in local training, selection of the best hyperparameter set is also local (catchment wise), allowing each catchment to take its own best set.

5.3 Is there a performance gain for regional LSTMs in the passage from hydrologically heterogeneous to homogeneous training, and vice versa?

To answer this question, we need to compare the REGIONAL REGIME model against the REGIONAL NATIONAL model. For almost all regimes and overall, when hydrologically similar but less catchments are used, we have the median KGE scores as equally good as when we use much more training catchments from various regimes. This is interesting at least for two reasons.

First, both models are benefiting from group training and their data are already several times greater than local level data. But between them, it is not the one with greater amount of the training data that outperforms. To clarify this point, let's take into account the Nival regime. National (heterogeneous) model uses the data that are "13 times" larger than the data that regime (homogeneous) model uses in this regime. However, they both have the same median KGE score. The point to note here is that, in the "regime Nival to national" passage, we have not increased the data of this regime (representation) 13 times. We have added a remarkable amount (13 times of the regime size) of data of some "dissimilar representations". This is very different from adding a large amount of data of the "same representation", which happens in the local to regime passage. Therefore, for non homogeneous training there is a "varied", not necessarily an added, complexity with respect to the representations.

Second, for both trainings, the complexity (learning) capacity of the model is the same — we are using the exact same model with the exact same static attributes for both trainings. In regime (homogeneous) training each REGIME LSTM is supposed to learn a single representation while in national (non homogeneous) training, the LSTM is exposed to the representations from all regimes.

What appears to count for both models is whether the varied complexity is shifted towards a simpler or a more difficult learning representation. And if the latter, whether there is an adequacy with respect to the amount of data. The variation of the complexity of representation(s) appears to be different from regime to regime. Given our results, we can identify three levels. 1) Regimes with a “self-sufficient” representation for which homogeneous training clearly outperforms heterogeneous training. The only instance of this level is found in the Nival Pluvial regime. In this regime, the new complexity appears to be shifted towards a “more complex” representation. 2) Regimes with a “self-insufficient” representation, which absolutely need contrasting/dissimilar representations for being learned by the LSTM. The only instance of this level is the Mediterranean regime. 3) Regimes with a “neutral” representation for which adding/removing contrasting representations does not (almost) change the complexity of the task for LSTM. The Uniform, Oceanic, and Nival regimes exhibit this level of representation. However, if we look at the overall performances, it turns out that almost the same level of data-complexity adequacy is achieved for both regime and national trainings.

One other important point to notice is that our non homogeneous (NATIONAL) LSTMs are “regime-informed”. That is, although their data come from all regimes, the exact same variables that we used to classify the regimes are given to the NATIONAL LSTMs as static attributes. They are not therefore absolutely naive with respect to the non homogeneity of data. Given their “regime-informed” property, we conjecture that, to some unknown but good extent, NATIONAL LSTMs have been already able to extract the classification. A systematic investigation is needed to prove this. But, if that happens, it is very useful because NATIONAL LSTMs are classification free — we will not need to encode, separately, the thresholds and conditions of the classification. Nevertheless, we still need a national data set to train them.

We did not observe in our results the performance improvement that Fang et al. (2022) obtained when they passed from LSTMs trained on single spatial ecoregions to the LSTM trained on all ecoregions. We find a number of explaining reasons for this difference. The measure of similarity between the two studies are very different. We have used purely hydrologic measures to classify catchments whereas in Fang et al.’s experiments, the measure of similarity is “spatial proximity”. The climatic context and data sets and their size are also very different in the two studies.

5.4 What is the most effective way of using LSTM for making runoff predictions?

Our results suggest that performances of an LSTM based runoff model is controlled by two factors: a) its training approach, b) its {lookback, hidden unit size} tuning. According to the results of this paper, maximizing the number of training catchments (national scale training) + selecting the {lookback, hidden unit size} set locally gives the best results within the regimes and overall. The interesting point to note is that it is only the “combination” of the the two components of this setting that gives the best results. Any of them separately does not appear to be a major winning factor — local LSTMs with local {lookback, hidden unit size} sets did not outperform regional LSTMs and NATIONAL LSTMs did not outperform REGIME LSTMs. We should also remember that the NATIONAL LSTMs that we tested are “regime-informed”. We might thus account for this property as the third component of this setting as well.

We have previously discussed the importance of lookback as a hyperparameter for LSTM. Here, we note the importance of having lookback tuned along with hidden unit size at a local scale so that the LSTM could better capture the hysteretic dynamics of each catchment separately. The strong relationship between these two hyperparameters has been previously recognized by Kratzert et al. (2019a).

6 Conclusions

In this study, we have used a sample of 361 gauged catchments in the hydrologically diverse French context. Our goal has been to exploit catchment hydrologic information when using LSTM based runoff models. We have thus proposed a regime classification built on three hydrologic indices to identify catchments with similar hydrologic behavior (representation). We have then trained the LSTM once locally — on individual catchments — and once regionally — on a group of catchments. We have performed the regional training at two scales: 1) at the scale of each hydrologic regime, i.e. only catchments from the same regime have been trained together 2) at the national scale, i.e. all 361 catchments have been trained together. For all trainings, we have performed 54 hyperparameter tunings on three hyperparameters — dropout rate (3 variations) as well as the two important LSTM's hyperparameters, namely sequence length (6 variations) and dropout rate (3 variations). We have investigated the relationship between the size of LSTM's input sequence and LSTM performance within different regimes. We have tested a new approach to selection of the best hyperparameter set and we have examined how different training and hyperparameter selection approaches change the performance of LSTM. For training and evaluation of all local and regional LSTMs, we have used three completely independent and long data sets — training $(10 \leq \leq 40$ [year]), validation $(= 10$ [year]), and test $(= 10$ [year]). In both local and regional trainings we have implemented the early stopping algorithm with no predefined number of epochs, allowing the LSTM to learn as long as its performances improve on the validation data. The results of our paper suggest the following main conclusions.

1. In the Uniform and Nival regimes where there is a clean hysteretic dominant process, we found a clear performance-lookback pattern — performances increased with increasing lookback up to an effective value, which depended on the time scaling of the dominant process. In the Mediterranean regime with a flash flood generating nature, we hypothetically expected a similar distinct pattern but with a much shorter effective lookback. What we found was a narrow spread in performance scores for different lookbacks. We related this to an underlying varying degree of hysteresis in this regime. Since several of catchments in this regime might be affected by localized controls from karstic geological features. In the Oceanic and Nival Pluvial regimes, we found a very monotonous performance-lookback pattern, reflecting an insensitivity of performances to different values of lookback. This indicates that in these regimes we could have adequate performances without using large lookbacks.

2. Whether or not the LSTM benefits from local to regional passage depends on a) the amount of the data at the local scale, b) on how it can negotiate the trade-off between the varied complexity of the representation(s) to be learned and the augmented data at the regional scale. If in this passage the model complexity increases as well, by, for instance, including many attributes in the regional model, this trade-off could become harder since the LSTM would need to

further trade generalization for precision (brought by the the more complex model). In local to regime passage, we saw a slightly better performance improvement compared to the performance improvement in the local to national passage.

3. At the local scale of one catchment, if the representation to be learned is “smooth” enough to elicit or the catchment’s data are so abundant that allow to elicit whatever complex representation, the LSTM would be already very close to the minimum possible error. In that case, there would be “less room” to improve performances by passing to regional LSTMs.

4. At the regional scale, from regime (hydrologically homogeneous) level to national (hydrologically heterogeneous) level, the model capacity is the same. An important amount of dissimilar data is added, and therefore the complexity of the new representations to be learned is varied. What appears to count for both models is whether the varied complexity is shifted towards a simpler or a more difficult learning representation. And if the latter, whether there is an adequacy with respect to the amount of data. Our results showed an overall outperformance of regime training but it was very slight and we can consider both regional trainings equivalent. It means that for both regime and national training levels, the amount of data has been adequate and appropriate with respect to the complexity of the representation(s) of that level. Nevertheless, we do not exclude the potential role of the “regime-informed” property of our national LSTMs in simplifying the representations in the heterogeneous space.

5. Given the almost equivalent performance of REGIME and “regime-informed” NATIONAL LSTMs, to choose between them, we may consider that the former needs less data but it requires an external classification — a precise encoding of our knowledge to the right classification. The latter needs a national data base but is classification free.

6. For improving the performance of an LSTM model, we found two important axes: its training approach and its {lookback, hidden unit size} tuning. Our best performances are given by the “regime-informed” HYBRID NATIONAL LSTMs — when we mix national training with local tuning of the two {lookback, hidden unit size} hyperparameters, along with providing regime information through attributes.

We can identify a number of directions that follow from our findings, which would benefit from further research:

1. The conclusions we made here are under one condition. In this paper, whenever we talked about an “increase in data size” at the regional scale, we were talking about increasing the data of dissimilar representations. And for this, we had been always within these bands: $361/101 \approx 4$ times (regime Oceanic) to $361/27 \approx 13$ times (regime Nival). We encourage to investigate the case in which we systematically change the degree of dissimilarity and size of data under a controlled environment.

2. If we intend to improve homogeneous training, one good step is to refine the current classification so that the number of regimes with a “self-sufficient” representation quality is maximized.

3. Our hydrologically heterogeneous LSTMs were “regime-informed”. We encourage to verify the conjecture that the LSTM is able to learn the classification if we provide it with regime information (through classification attributes). One

simple way to do this is to once include and once exclude the classification indices in and from static features of regional LSTMs and compare the results. This paper involves the former setting but not the latter.

4. A future research direction could be to explore in a catchment wise manner the relationship between LSTM's optimal lookback and memory related metrics such as Catchment Forgetting Curve (de Lavenne et al., 2021). This would allow us to predict for each catchment its optimal lookback without performing hyperparameter tunings.

5. The methods presented in this paper are developed for gauged catchments. A further step would be to extend them to approaches that are applicable to ungauged catchments — catchments not used in training.

Author contributions. RH designed all experiments with advice from PJ. RH conducted all experiments, visualized the results, and wrote the manuscript. RH analyzed the results with advice from PJ. PAG assisted in the execution of experiments and sample selection. PB contributed to the idea of a regime based analysis and gave guidance on the data and GR4J simulations. PJ supervised the work and was in charge of the overall direction. Revision of the manuscript was done collectively.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was granted access to the GPU resources of IDRIS under the allocation number 2022-AD011013339 made by GENCI. We gratefully acknowledge data provision from Météo France, the Banque HYDRO, and the HYCAR research team of INRAE. We would like to sincerely thank Jérémy Verrier for his constant support and help for using HPC resources of INRAE and GENCI. We would also like to thank Handling Editor Efrat Morin and appreciate the comments and reviews provided by Referee John Quilty and Anonymous Referee 2 — which have substantially improved this paper.

References

Beck, C., Jentzen, A., and Kuckuck, B.: Full error analysis for the training of deep neural networks, Infinite Dimensional Analysis, Quantum Probability and Related Topics, p. 2150020, 2022.

Chiverton, A., Hannaford, J., Holman, I., Corstanje, R., Prudhomme, C., Bloomfield, J., and Hess, T. M.: Which catchment characteristics control the temporal dependence structure of daily river flows?, Hydrological Processes, 29, 1353–1369, 2015.

Chollet, F. et al.: Keras, https://github.com/fchollet/keras, 2015.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., and Andréassian, V.: The Suite of Lumped GR Hydrological Models in an R package, Environmental Modelling and Software, 94, 166–171, 2017.

Coron, L., Delaigue, O., Thirel, G., Perrin, C., and Michel, C.: airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling, r package version 1.4.3.65, 2020.

de Lavenne, A., Andréassian, V., Crochemore, L., Lindström, G., and Arheimer, B.: Quantifying pluriannual hydrological memory with Catchment Forgetting Curves, Hydrology and Earth System Sciences Discussions, 2021, 1–27, https://doi.org/10.5194/hess-2021-331, 2021.

Delaigue, O., Génot, B., Lebecherel, L., Brigode, P., and Bourgin, P.-Y.: Database of watershed-scale hydroclimatic observations in France, https://webgr.inrae.fr/base-de-donnees, 2020.

Fang, K., Kifer, D., Lawson, K., Feng, D., and Shen, C.: The Data Synergy Effects of Time-Series Deep Learning Models in Hydrology, Water Resources Research, 58, e2021WR029 583, 2022.

Feng, D., Fang, K., and Shen, C.: Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales, Water Resources Research, 56, e2019WR026 793, 2020.

Frame, J., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall-runoff predictions of extreme events, Hydrology and Earth System Sciences Discussions, 2021, 1–20, 2021.

Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., and Lin, Q.: Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation, Journal of Hydrology, 589, 125 188, 2020.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network, Hydrology and Earth System Sciences, 25, 2045–2062, 2021a.

Gauch, M., Mai, J., and Lin, J.: The proper care and feeding of CAMELS: How limited training data affects streamflow prediction, Environmental Modelling & Software, 135, 104 926, 2021b.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, 2016.

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

Haines, A., Finlayson, B., and McMahon, T.: A global classification of river regimes, Applied Geography, 8, 255 – 272, 1988.

Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6, 107–116, 1998.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, 1997.

Kachroo, R. and Natale, L.: Non-linear modelling of the rainfall-runoff transformation, Journal of Hydrology, 135, 341 – 369, 1992.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, 2017.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall-runoff modelling using long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci, 22, 6005–6022, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resources Research, 55, 11 344–11 354, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, 2019b.

LeCun, Y.-A., Bottou, L., Orr, G.-B., and Müller, K.-R.: Efficient backprop, in: Neural networks: Tricks of the trade, pp. 9–48, Springer, 2012.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, Hydrology and Earth System Sciences, 25, 5517–5534, https://doi.org/10.5194/hess-25-5517-2021, 2021.

Michel, C.: Hydrologie appliquée aux petits bassins versants ruraux (Applied hydrology for small catchments), Cemagref, Antony, France, internal Report, 1989.

Nearing, G. S., Klotz, D., Sampson, A. K., Kratzert, F., Gauch, M., Frame, J. M., Shalev, G., and Nevo, S.: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks, Hydrology and Earth System Sciences Discussions, pp. 1–25, 2021.

O, S., Dutra, E., and Orth, R.: Robustness of Process-Based versus Data-Driven Modeling in Changing Climatic Conditions, Journal of Hydrometeorology, 21, 1929–1944, 2020.

Omernik, J. M. and Griffith, G. E.: Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework, Environmental management, 54, 1249–1266, 2014.

Oudin, L., Michel, C., and Anctil, F.: Which potential evapotranspiration input for a lumped rainfall-runoff model?: Part 1—can rainfall-runoff models effectively handle detailed potential evapotranspiration inputs?, Journal of Hydrology, 303, 275–289, 2005.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., and Le Moine, N.: Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, Water Resources Research, 44, 2008.

Pardé, M.: Fleuves et rivières, Collection Armand Colin. Section de Géographie (France) fre no. 155, 1933.

Perrin, C. and Littlewood, I.: A comparative assessment of two rainfall-runoff modelling approaches: GR4J and IHACRES, in: Proceedings of the Liblice Conference (22-24 September 1998), V. Elias and IG Littlewood (Eds.), IHP-V, Technical Documents in Hydrology n, vol. 37, pp. 191–201, 2000.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of hydrology, 279, 275–289, 2003.

Phillips, J. D.: Sources of nonlinearity and complexity in geomorphic systems, Progress in Physical Geography: Earth and Environment, 27, 1–23, 2003.

Quintana-Segui, P., Moigne, P. L., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S.: Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France, Journal of applied meteorology and climatology, 47, 92–107, 2008.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2019.

Sauquet, E.: Mapping mean annual river discharges: geostatistical developments for incorporating river network dependencies, Journal of Hydrology, 331, 300–314, 2006.

Valéry, A., Andréassian, V., and Perrin, C.: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, Journal of Hydrology, 517, 1176 – 1187, 2014.

Van Rossum, G. and Drake, F. L.: Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, 2010.