Learning Landscape Features from Streamflow with Autoencoders

Alberto Bassi[1,2], Marvin Höge[2], Antonietta Mira[3,4], Fabrizio Fenicia[2], and Carlo Albert[2]

[1]Department of Physics, ETH Zurich, Switzerland
[2]Swiss Federal Institute for Aquatic Science and Technology (EAWAG), Dübendorf, Switzerland
[3]Euler Institute, Università della Svizzera italiana, Lugano, Switzerland
[4]Insubria University, Como, Italy

Correspondence: Alberto Bassi (abassi@ethz.ch)

Abstract.

~~Understanding the number and types of signatures that best describe streamflow time series is a crucial objective in hydrological science, serving applications such as catchment classification, hydrological model development and calibration. With the main objective of learning a minimal number of streamflow features, we employ an explicit noise conditional autoencoder (ENCA), which, together with meteorological forcings, allows for~~ Recent successes with Machine Learning (ML) models in catchment hydrology have highlighted their ability to extract crucial information from catchment properties pertinent to the rainfall-runoff relationship. In this study, we aim to identify a minimal set of catchment signatures in streamflow that, when combined with meteorological drivers, enable an accurate reconstruction of the ~~whole~~ entire streamflow time series. To achieve this, we utilize an Explicit Noise Conditional Autoencoder (ENCA), which effectively separates the influences of meteorological drivers and catchment properties on streamflow. The ENCA architecture feeds ~~the meteorological forcing~~ meteorological forcing and climate attributes to the decoder in order to incentivize the encoder to only learn features that are related to landscape properties. By isolating the effect of meteorology, these features can thus be interpreted as landscape fingerprints. The optimal number of features is found by means of an intrinsic dimension estimator. We train our model on the hydro-meteorologic time series data of 568 catchments of the continental United States from the CAMELS dataset. We compare the reconstruction accuracy with ~~state-of-the-art~~ models that take as input a subset of static catchment attributes (both climate and landscape attributes) along with the meteorological forcing variables. Our results suggest that available ~~static catchment~~ landscape attributes compiled by experts ~~account for almost all the relevant information about the rainfall-runoff relationship. Yet, these catchment attributes~~ can be summarized by only two relevant learnt features (or signatures), while at least a third one is needed for about a dozen difficult to predict catchments in the central US, mainly characterized by high aridity index~~and intermittent flow~~. The principal components of the learnt features strongly correlate with the baseflow index and aridity indicators, which is consistent with the idea that these indicators capture the variability of catchment hydrological response~~and relate to needed model complexity~~. The correlation analysis further indicates that soil-related and vegetation attributes are of ~~high importance. Finally, in the attempt to interpret the learnt catchment features, we relate them to typical hydrological model components, with specific reference to the parameters of the GR4J model and their function on the hydrograph.~~ importance.

1 Introduction

~~*Hydrological signatures*~~ Hydrological signatures encompass descriptive statistics derived from meteorological and streamflow time series. They serve various purposes in hydrology, such as hydrological model calibration or evaluation (Fenicia et al., 2018; Kiraz et al., 2023), process identification (McMillan, 2020a), and ecological characterization (Olden and Poff, 2003). Alongside with catchment attributes (distinguished here in landscape and climate attributes), they are also used for catchment classification and regionalization studies (Wagener et al., 2007).

~~*Streamflow signatures*~~Streamflow signatures, i.e. hydrological signatures solely based on streamflow, hold significant importance as they relate to the variable one aims to predict and understand (Gnann et al., 2021). Hydrologists have developed diverse signatures reflecting different aspects of streamflow dynamics. Examples include those linked to the flow duration curve (e.g., slope of various segments), the baseflow index, or the flashiness index. Numerous other such signatures exist. For instance, Olden and Poff (2003) compiled a list of 171 indices from prior work, reflecting aspects associated to magnitude, frequency, duration, timing, and rate of change of flow events. As streamflow depends on meteorological forcing and landscape attributes, streamflow signatures generally contain information from both sources. In particular for predictions in ungauged basins, it is vital to be able to disentangle them.

One way of doing so is through ~~*hydrological models*~~hydrological models, which condense catchment attributes into ~~*model parameters*~~ model parameters (Wagener et al., 2003). Previous research indicates that observed hydrographs can be represented by a handful of model parameters (Jakeman and Hornberger, 1993). For instance, the ~~*GR4J model*~~ GR4J model (Perrin et al., 2003), resulting from a continuous refinement process aimed at balancing model complexity and performance, has only four parameters. However, these analyses are based on predefined model assumptions.

Model parameters can, in principle, directly be estimated from streamflow signatures. The Approximate Bayesian Computation (ABC) technique (Albert et al., 2015) has recently been used to infer model parameters from streamflow signatures - which in this context are called ~~*summary statistics*~~ summary statistics - bypassing the need to directly compare the complete time series (Fenicia et al., 2018). If these summary statistics contained all the information necessary to estimate model parameters they would be termed as ~~*sufficient*~~sufficient. Sufficiency is therefore not an inherent property of the summary statistics but depends on the specific hydrological model and on the parameters that need to be inferred. For ABC to converge efficiently, we also want the summary statistics to be ~~*minimal*~~minimal, i.e., while they should ideally encode all the parameter related information available in the streamflow, they should encode no other information, neither from the forcing nor from the noise that is used for the simulations (Albert et al., 2022). Such minimally sufficient summary statistics could thus be considered the relevant fingerprints of landscape features on the streamflow. Of course, this holds true only if the model is capable of encoding all the information in such features that is relevant for the input-output relationship. However, recent studies show that purely data-driven models ~~might~~ outperform process-based models in prediction accuracy (Kratzert et al., 2019; Mohammadi, 2021), ~~becuase~~ because they suggest information in catchment attributes previously not utilized for streamflow prediction.

Our goal is to employ Machine Learning (ML) techniques to identify a minimal set of streamflow features enabling accurate streamflow predictions when combined with meteorological forcing. Thus, our aim is to eliminate all forcing-related informa-

tion from the streamflow, distilling features solely from the catchments themselves. We approach this objective from a purely data-driven perspective~~, aiming to reduce assumptions relative to traditional process-based modeling~~.

To identify minimal sets of streamflow features, we employ ~~a novel ML architecture recently proposed for extracting minimally sufficient summary statistics from noisy outputs of stochastic models. The architecture is~~ an Explicit Noise Conditional Autoencoder (ENCA) (Albert et al., 2022), where the bare noise utilized by the stochastic model simulator is fed into the decoder together with the learnt summary statistics. This way, the encoder is encouraged to encode only those features containing information on the model parameters while disregarding the noise. Albert et al. (2022) applied ENCA to infer parameters of simple one-dimensional stochastic maps, showing that the learnt features allow for an excellent approximation of the true posterior. In our case, instead of noise, we input meteorological forcing into the decoder~~and~~. By feeding also climate attributes to the decoder, we encourage the encoder to exclusively encode landscape-related information within the streamflow. Moreover, since we make use of uni-directional LSTM (see Appendix C), conditioning ENCA also on climate attributes could help the decoder to obtain future information about the climate that it would not be normally able to retrieve.

In order to reduce the computational costs and learn a minimal set of catchment features, the dimension of the latent space is chosen according to the estimation of its Intrinsic Dimension (ID) (Facco et al., 2017; Allegra et al., 2020; Denti et al., 2022). In particular, we employ the ID estimator GRIDE (Denti et al., 2022), which is robust to noise. Learnt features will then be compared with known catchment attributes (both from the landscape and the climate) and hydrological signatures to provide a hydrological interpretation and guide knowledge domain experts towards the pertinent information necessary for streamflow prediction.

We apply our approach to the US-CAMELS dataset (Newman et al., 2015), covering several hundred catchments over the continental US. ~~In order to benchmark our results, we used previous modelling work on the same dataset.~~ LSTMs (Long ~~Short Term~~ Short-Term Memory networks) have emerged as state-of-the-art models for streamflow data-driven predictions~~in ungauged basins~~. In the study of Kratzert et al. (2019), LSTMs validated on unseen catchments, enriched with static landscape and climate attributes from Addor et al. (2017), outperformed conceptual models. First investigations towards mechanistic interpretation of the LSTM states, e.g. linking hidden states to the dynamics of soil moisture, demonstrated the potential of eliciting physics from data-driven models (Lees et al., 2022). Here, by linking learnt features to known catchment attributes, we explore a further aspect in this broader field of explainable AI or interpretable ML (Molnar, 2024; Molnar et al., 2020).

Our specific objectives are: (i) find the minimal number of dominant streamflow-features stemming from the landscape; (ii) relate them to known landscape and climate attributes as well as established hydrological signatures. This will not only allow us to determine how many features are required for streamflow prediction, but also to answer the question whether there is missing information in known catchment attributes.

A similar attempt of learning signatures has recently been made by Botterill and McMillan (2023). In pursuit of an interpretable latent space on a continental scale, they employed a convolutional encoder to compress high-dimensional information derived from meteorological forcing and streamflow data. This approach was aimed at learning hydrological signatures (McMillan, 2020b) within the US-CAMELS dataset. Their approach differs from ours in three aspects: (i) they used a a traditional conceptual model as a decoder whereas we use an LSTM architecture which has been shown to be superior to conceptual

models when provided with catchment properties; (ii) they fed both streamflow and meteorological forcing into the encoder whereas we feed in only streamflow data in an attempt to separate landscape- from forcing-information; (iii) they did not attempt to find a minimal number of signatures sufficient for streamflow prediction, whereas this is a primary objective of our work.

It is important to note that our main objective is not to beat state-of-the-art models regarding their predictive performance ~~in ungauged basins~~(Kratzert et al., 2021; Klotz et al., 2022). Our goal is rather to investigate the information content in streamflow. However, we believe our research will provide valuable insights into the most critical features for streamflow prediction. ~~Ultimately, this knowledge may be utilized for prediction in ungauged catchments in the future.~~

2 Models and Methods

2.1 Data

We employ the Catchments Attributes and Meteorology for Large-sample Studies (CAMELS) (Newman et al., 2015), which consists of 671 catchments in the contiguous United States (CONUS), ranging in size from 4 to $25 \cdot 10^3$ km^2. For this study, we select those 568 catchments out of 671 whose data span continuously on a daily basis the time period from 1 October 1980 to 30 September 2010, corresponding to 30 hydrological years. The first 15 years are used for ~~calibration~~ training and the last 15 for ~~validation~~testing. Along with the streamflow time series and the meteorological forcing variables, US-CAMELS also provides information about catchments static attributes (Addor et al., 2017), encompassing both landscape (vegetation, soil, topography and geology) and climate. Streamflow data is retrieved from the U.S. Geological Survey gauges, while the meteorological forcing comes from the extended North America Land Data Assimilation Systems (NLDAS) (Kratzert, 2019) and includes maximum and minimum daily temperature, solar radiation, vapour pressure and precipitation.

2.2 Explicit Noise Conditional Autoencoder

We use the Explicit Noise Conditional Autoencoders (ENCA) (Albert et al., 2022), where the streamflow is fed to a convolutional encoder. ENCA has been designed to distill sufficient summary statistics which contain minimal noise information from the output of stochastic models. Here (Figure 1 - for the detailed architectures the reader is referred to Appendix C), the noise is substituted by all the variables we are not interested in, namely the meteorological forcing. The convolutional encoder is thus followed by a LSTM decoder that takes as input 15 hydrological years of meteorological forcing, i.e. 5478 ~~values (- for the detailed architectures the reader is referred to~~ time points, and nine climate attributes (reported in Table 1). The ~~memory cells of the LSTM are~~ LSTM capacity is limited by the dimension of the input layer. In order to enlarge the ~~memory available~~ available capacity and capture more complex patterns from the meteorological forcing, the meteorological time series are first fed to a single linear layer with 1350 output units. The output of this linear layer is then concatenated with the output of the encoder and fed to the LSTM decoder. This way the decoder sees tensors of size ~~$(B, 5478, 1350 + S)$, where B~~ $(BS, 5478, 1350 + N)$,

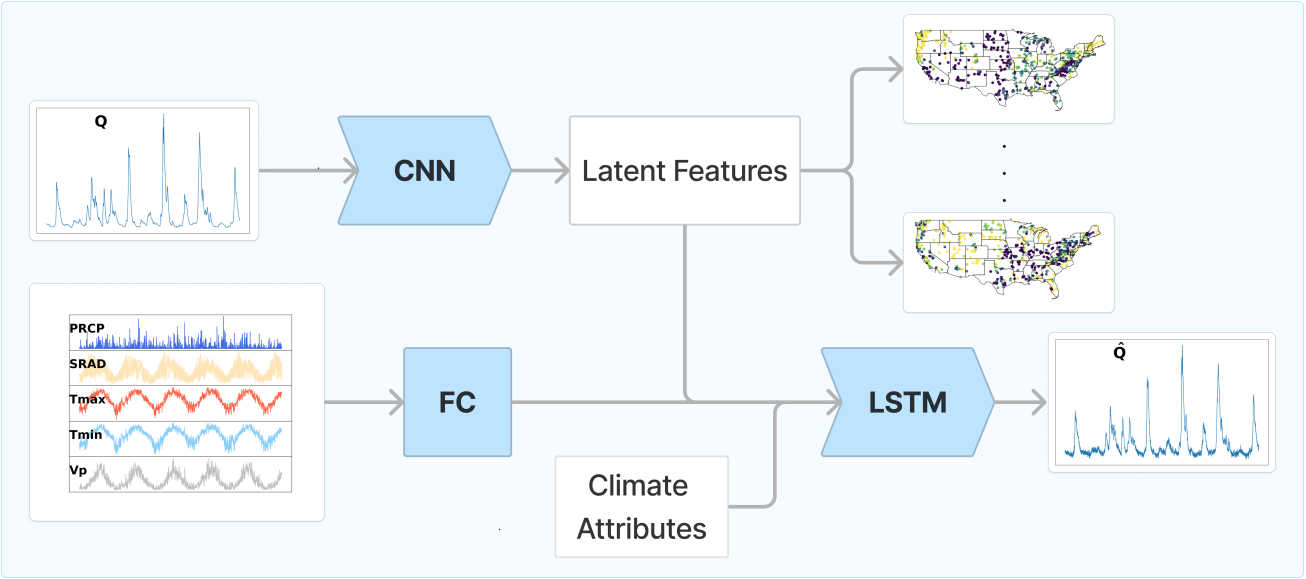

Figure 1. Explicit Noise Conditional Autoencoder used in this study. For the hyper-parameters and the implementation details, the reader is referred to Appendix C. The neural network architectures employed are Convolutional Neural Networks (CNN), a Fully Connected (FC) layer and LSTM. The observed and simulated streamflows are denoted with Q and $\hat{Q}$, respectively. The meteorological forcing variables are denoted with PRCP (precipitation), SRAD (solar radiation), Tmax (maximum temperature), Tmin (minimum temperature) and Vp (vapour pressure).

where BS is the batch size (batches are selected across different catchments) and ~~S~~ N is the latent space dimension. We opted for such an architecture in order to extract as much static information related to the streamflow as possible.

~~Obviously, setting the latent dimension equal to the length of the entire time-series would allow for a perfect streamflow reconstruction. However, we~~ We expect to be able to compress almost all streamflow information not already contained in the forcing into a ~~much smaller number of features~~low (N)-dimensional feature vector. Because they should be largely devoid of forcing information, we call ~~them~~ these features the *relevant landscape features* and explain in ~~the next section~~ subsection 2.3 how we fix their number. We refer to the ENCA model with latent space dimension ~~equal to~~ N as ENCA-N. Comparing relevant landscape features with known static catchment attributes in terms of their capacity for streamflow reconstruction will allow us to find out whether static catchment attributes lack information that is crucial for streamflow prediction. For this comparison, we use an LSTM model augmented with catchment attributes (Addor et al., 2017) in the input, both steaming from the landscape and the climate. We refer to this model as Catchment Attributes Augmented Model (CAAM). This model differs from Figure 1 solely by the fact that the latent features are substituted by known ~~catchment~~ landscape attributes. Following Kratzert et al. (2019), CAAM is fed with 27 catchment attributes (reported in Table 1~~and denoted with a *~~), which are representative of climate, ~~topology~~topography, geology, soil and vegetation. ~~As a control case, we also report the results obtained with ENCA-0, which is given neither catchment attributes nor learnt features as inputs.~~

In order to mitigate numerical instability, it is crucial to standardize the catchment attributes or latent features before feeding the LSTM. In CAAM, we standardize the catchment attributes ~~once and for all~~ with the mean and standard deviation computed over all the considered catchments. This is not possible for ENCA, since the mean and standard deviation of the latent features are not known a priori. Therefore, we standardize the latent features by ~~mean~~ means of a batch normalization layer. This way, we ensure that the magnitude of the LSTM input is comparable between CAAM and ENCA.

~~In order to estimate the ID, we apply the GRIDE estimator (Denti et al., 2022). Given sample points, $\mathbf{x}_i \in \mathbb{R}^D$, for $i = 1, \ldots, M$, and a distance measure, $r : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^+$, GRIDE assumes that points in a given neighbourhood are counted with a Poisson point process with intensity ρ, which is constant at least up to the scale of the diameter of the considered neighbourhood. We define $r_{i,l}$ as the distance between the point $\mathbf{x}_i$ and its l-th nearest neighbour. Let us define $\mu_{i,n_2,n_1} = \frac{r_{i,n_2}}{r_{i,n_1}}$, where n_1~~ Training and ~~n_2 (with $0 \leq n_1 \leq n_2 \leq M$) are nearest neighbours of generic order. The distribution of μ_{i,n_2,n_1} can be computed in closed form and depends only on the ID of the data while, crucially, does not depend on ρ, as long as ρ is constant in the considered neighbourhood of i whose diameter is set by the distance between i and its n_2-th nearest neighbour (Denti et al., 2022).~~

~~GRIDE evolution plot for the ENCA-27 (left) and the ENCA-5 (right) latent spaces for one random restart. The other restarts show a similar pattern.~~

~~In order to correctly identify the ID of a dataset, a scale-independent analysis is essential. We therefore make use of *GRIDE paths*, the evolution of the ID as a function of n_2, which can be interpreted as the scale at which we look at the data. We set $n_1 = n_2/2$, as is usually done in the literature. As a function of n_2, the ID is first expected to increase, due to the noise present at small distance scales, and then to reach a plateau corresponding to the correct ID. shows the GRIDE path for increasing values of n_1 from 1 to 270. The left panel shows that the ID estimate of the latent space of ENCA-27 decreases after showing a plateau around five, then reaches a minimum around three, then increases again and finally collapses to low values at larger distance scale. The plateau at five motivates us to train an ENCA-5 and study its ID. The right panel of shows that the local minimum of the GRIDE path of the latent space of ENCA-5 is consistent with an ID of three. We can deduce that, for most of the catchments, the ID is three, while for some it can be higher. However, the fact that the GRIDE path of the latent space of ENCA-27 shows two plateaus around five and three can be an indicator of the existence of two or more manifolds with different IDs. We will see below that, indeed, three features seem to capture most of the relevant information, which is in line with the GRIDE path for ENCA-5 (right plot in).~~

~~Note that, at increasingly large scales, the GRIDE estimator is not reliable anymore since the assumption of locally homogeneous intensity of the Poisson process - on which it relies - may fail to hold. With real data it is usually difficult to ascertain the scale at which the local homogeneity assumption is valid. We use the ID as a guide to reduce the dimension of the latent space of the autoencoder. However, in the end we train ENCA for several dimensions of the latent space and evaluate the information content of the learnt features in terms of their ability to reconstruct streamflows.~~

2.3 ~~Training and Validation~~

We use the first Testing]The Intrinsic DimensionTo identify the dimension of the latent space we proceed with the following methodology. First, we train an ENCA-N with a relatively large number of latent features N. Since we fed 27 catchment attributes to the reference model (CAAM), we used a 27-dimensional latent space in order to have a fair comparison in terms of model capacity. We refer to this model as ENCA-27. The dimension of the latent space does not matter so much, as long as it is larger than the expected number of relevant landscape features. Then, we estimate the ID of the latent space (see below), and train another ENCA with the number of latent features equal to the estimated ID and, in turn, estimate its ID to check if the dimension of the latent space can be further reduced.

In order to estimate the ID, we apply the GRIDE estimator (Denti et al., 2022). Given sample points, $\mathbf{x}_i \in \mathbb{R}^D$, for $i = 1, \ldots, M$, and a distance measure, $r : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^+$, GRIDE assumes that points in a given neighbourhood are counted with a Poisson point process with intensity ρ, which is constant at least up to the scale of the diameter of the considered neighbourhood. We define $r_{i,l}$ as the distance between the point $\mathbf{x}_i$ and its l -th nearest neighbour. Let us define $\mu_{i,n_2,n_1} = \frac{r_{i,n_2}}{r_{i,n_1}}$, where n_1 Training and n_2 (with $0 \leq n_1 \leq n_2 \leq M$) are nearest neighbours of generic order. The distribution of μ_{i,n_2,n_1} can be computed in closed form and depends only on the ID of the data while, crucially, does not depend on ρ, as long as ρ is constant in the considered neighbourhood of i whose diameter is set by the distance between i and its n_2-th nearest neighbour (Denti et al., 2022).

GRIDE evolution plot for the ENCA-27 (left) and the ENCA-5 (right) latent spaces for one random restart. The other restarts show a similar pattern.

In order to correctly identify the ID of a dataset, a scale-independent analysis is essential. We therefore make use of *GRIDE paths*, the evolution of the ID as a function of n_2, which can be interpreted as the scale at which we look at the data. We set $n_1 = n_2/2$, as is usually done in the literature. As a function of n_2, the ID is first expected to increase, due to the noise present at small distance scales, and then to reach a plateau corresponding to the correct ID. shows the GRIDE path for increasing values of n_1 from 1 to 270. The left panel shows that the ID estimate of the latent space of ENCA-27 decreases after showing a plateau around five, then reaches a minimum around three, then increases again and finally collapses to low values at larger distance scale. The plateau at five motivates us to train an ENCA-5 and study its ID. The right panel of shows that the local minimum of the GRIDE path of the latent space of ENCA-5 is consistent with an ID of three. We can deduce that, for most of the catchments, the ID is three, while for some it can be higher. However, the fact that the GRIDE path of the latent space of ENCA-27 shows two plateaus around five and three can be an indicator of the existence of two or more manifolds with different IDs. We will see below that, indeed, three features seem to capture most of the relevant information, which is in line with the GRIDE path for ENCA-5 (right plot in).

Note that, at increasingly large scales, the GRIDE estimator is not reliable anymore since the assumption of locally homogeneous intensity of the Poisson process - on which it relies - may fail to hold. With real data it is usually difficult to ascertain the scale at which the local homogeneity assumption is valid. We use the ID as a guide to reduce the dimension of the latent space of the autoencoder. However, in the end we train ENCA for several dimensions of the latent space and evaluate the information content of the learnt features in terms of their ability to reconstruct streamflows.

2.3 Training and Validation

~~We use the first~~ Testing We use the first 15 years of data for ~~calibration~~ training and the last 15 for ~~validation. Calibration~~ testing. training is performed by maximizing the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970), defined as:

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{\mathrm{T}} \left(q_{\mathrm{sim},t} - q_{\mathrm{obs},t}\right)^2}{\sum_{t=1}^{\mathrm{T}} \left(q_{\mathrm{sim},t} - \mu_{obs}\right)^2} \frac{\sum_{t=1}^{\mathrm{T}} \left(q_{\mathrm{sim},t} - q_{\mathrm{obs},t}\right)^2}{\sum_{t=1}^{\mathrm{T}} \left(q_{\mathrm{obs},t} - \mu_{obs}\right)^2}, \tag{1}$$

where $q_{obs,t}$ and $q_{sim,t}$ are, respectively, the observed and predicted streamflow expressed in mm/day at day t, and μ_{obs} is the average of the observed streamflow. We notice that maximizing the NSE is equivalent to minimizing the Mean Square Error (MSE) between data and prediction. Each model is trained with the Adam optimizer (Kingma and Ba, 2015), with learning rate equal to 10^{-5} for ~~a maximum of $20 \cdot 10^3$ epochsand early stopping with patience of $2 \cdot 10^3$ epochs. The models with best validation NSE is then chosen. The~~ 10,000 epochs. The batch size is set to 64 and the first 270 days of the predicted streamflow are excluded when computing the loss.

~~Since the NSE overweight flow peaks due to the square values, it is not well suited to asses the performance on low flow regimes. Therefore, following Kratzert et al. (2019) we also report the percentage bias , defined as~~

$$\mathrm{BIAS} = \frac{\mu_{\mathrm{sim}} - \mu_{\mathrm{obs}}}{\mu_{\mathrm{obs}}}.$$

~~In addition, to assess the performance in streamflow variability, we report the standard deviation ratio of the streamflow logarithm, defined as~~ We also report the three components in which NSE can be decomposed, see Eq. 4 in the main text of Gupta et al. (2009). These components are the linear correlation coefficient (R), the bias normalized by the observed streamflow standard deviation (BIAS) and standard deviation ratio (STDEV). The linear correlation coefficient is related to timing, whereas STDEV measures the streamflow variability and it is defined as

$$\mathrm{LOG-STDEV}\mathrm{STDEV} = \frac{\sigma_{\log(\mathrm{sim})}}{\sigma_{\log(\mathrm{obs})}} \frac{\sigma_{\mathrm{sim}}}{\sigma_{\mathrm{obs}}}, \tag{2}$$

where ~~$\sigma_{\log(\mathrm{sim})}$ and $\sigma_{\log(\mathrm{obs})}$~~ σ_{sim} and σ_{obs} are the standard deviations of the ~~logarithm~~ of the simulated and observed streamflows, respectively. Finally, the BIAS is related to volume errors and it is defined as

$$\mathrm{BIAS} = \frac{\mu_{\mathrm{sim}} - \mu_{\mathrm{obs}}}{\sigma_{\mathrm{obs}}}. \tag{3}$$

Each algorithm is affected by noise, due to the random initialization of the neural network parameters. To minimize this effect we run each model with four random restarts, each one providing the streamflow prediction in the whole ~~validation~~ testing period. We compute the evaluation metrics on the predicted streamflow after averaging the streamflow over these four random restarts.

2.3 ~~Latent Space Interpretation~~Intrinsic Dimension Estimation

The ~~relevant features~~ selection of the encoder latent space dimension, specifically the number of relevant features, is informed by the ID estimator GRIDE (Denti et al., 2022). We utilize the GRIDE paths, which involve estimating the ID at several

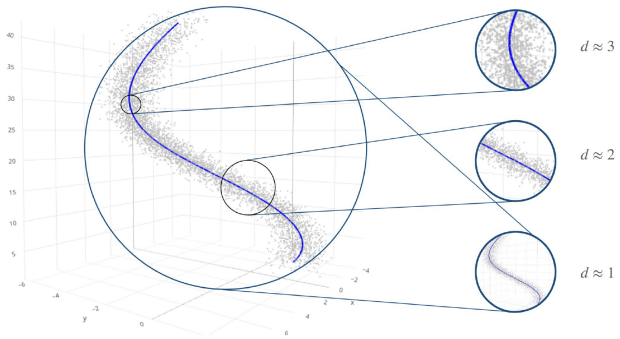

Figure 2. Unidimensional line embedded in three dimensions. The estimated intrinsic dimension depends on the distance scale.

distance scales at which the data is analyzed (for an in-depth discussion on ID, refer to Appendix A). The ID intuitively measures the dimension of the manifold where the data resides, which may be lower than the dimension of the embedding space. Most ID estimators depend on calculating the distance scale between data points, and the estimated ID itself can vary with this distance scale. Figure 2, derived from Denti et al. (2022), illustrates a one-dimensional spiral dataset embedded in a three-dimensional space with added noise. When the distance scale is too small, the data points appear to fill the space uniformly, making the manifold seem three-dimensional. However, as the distance scale increases and the noise is bypassed, the estimated ID decreases until the correct value of one is achieved.

To identify the dimension of the latent space of ENCA we proceed with the following methodology. First, we train an ENCA-N with a relatively large number of latent features N. Since we fed 27 catchment attributes to the reference model (CAAM), we use a 27-dimensional latent space in order to have a fair comparison in terms of model capacity. We refer to this model as ENCA-27. The exact dimension of the latent space we start with does not matter much, as long as it is larger than the expected number of relevant landscape features. Then, we estimate the ID of the latent space, and train another ENCA with the number of latent features equal to the estimated ID and, in turn, estimate its ID to check if the dimension of the latent space can be further reduced. We thus use the ID as a guide to progressively diminish the dimension of the latent space of the autoencoder. However, in the end we train ENCA for several dimensions of the latent space and evaluate the information content of the learnt features in terms of their ability to reconstruct streamflows. In Figure A1 we report GRIDE paths for different models trained in this work.

3 Latent Space Interpretation

The relevant features are first projected using a Principal Component Analysis (PCA), since in general the autoencoder latent representation is in arbitrary coordinates. ~~By doing this~~Doing so, we ensure a fair comparison ~~between~~ among different random restarts, since we change the basis of each latent space by ordering the new coordinates according to the explained variance.

Finally, in order to interpret the relevant landscape features, we report the absolute Spearman correlation (Zar, 2005) matrix between among the learnt features and , static catchment attributes and hydrological signatures, which are reported in Table 1.

Meteorological forcing variables, climate and landscape (topological, geological, soil and vegetation) attributes and hydrological signatures compared in this study. The attributes fed to CAAM are denoted with a *.

Meteorological Forcing Variables	
PRCP	Average daily precipitation (mm/day)
SRAD	Surface incident solar radiation (W/m^2)
Tmax	Maximum daily atmosphere temperature (°C)
Tmin	Minimum daily atmosphere temperature (°C)
Vp	Nearly surface daily vapour pressure average (Pa)
Climate Attributes	
Prec Mean	Mean daily precipitation.
PET Mean	Mean daily potential evapotranspiration.
Prec Seasonality	Seasonality of precipitation estimated by using sinusoidal waves.
Fraction Snow	Fraction of precipitation falling on days with T < 0 °C.
Aridity Index	Ratio between the mean PET and mean precipitation.
High Prec Frequency	Frequency of days with $\leq$ 5x mean daily precipitation.
High Prec Duration	Mean duration of high precipitation events.
Low Prec Frequency	Frequency of days with $\leq$ 1 mm/day of precipitation.
Low Prec Duration	Mean duration of dry periods.
Hydrological Signatures	
Q Mean	Mean daily streamflow (mm/day).
Streamflow Ratio	Ratio between the mean daily streamflow and mean daily precipitation.
Slope FDC	Slope of the flow duration curve.
Baseflow Index	Ratio between the average daily baseflow and streamflow.
Stream ELAS	Streamflow precipitation elasticity.
Q5	5 % flow quantile (mm/day).

Landscape Attributes	
Topological Attributes	
Elevation Mean	Mean elevation of the catchment.
Slope Mean	Mean slope of the catchment.
Area Catchment	Area of the catchment.
Geological Attributes	
Carbonate Rocks Fraction	Carbonate sedimentary rocks fraction area in the catchment.
Geological Permeability	Surface permeability (in log10 scale).
Soil Attributes	
Soil Depth (Pelletier)	Depth to bedrock (maximum 50 m).
Soil Depth (STATSGO)	Soil depth (maximum 1.5 m).
Soil Porosity	Volumetric porosity.
Soil Conductivity	Saturated hydraulic conductivity.
Max Water Content	Maximum water content of the soil.
Sand Fraction	Fraction of sand in the soil.
Silt Fraction	Fraction of silt in the soil.
Clay Fraction	Fraction of clay in the soil.
Vegetation Attributes	
Fraction Forest	Fraction of the catchment covered by forest.
LAI Max	Maximum monthly mean of leaf area index.
LAI Diff	Difference between the max. and the min. monthly mean of the leaf area index.
GVF Max	Maximum monthly mean of the green vegetation fraction.
GVF Diff	Difference between the max. and min. monthly mean of the green vegetation fraction.

Table 1. Meteorological forcing variables, climate and landscape (topological, geological, soil and vegetation) attributes and hydrological signatures compared in this study. Both climate and landscape attributes are fed to CAAM. ENCA models are conditioned on climate

4 Results and Discussion

4.1 The Number of Relevant Landscape Features

Validation NSE, BIAS and LOG-STDEV values for the considered models. The NSE and BIAS distributions of the attributes augmented model (CAAM) lie between ENCA with 2 and 3 latent features. ENCA tends to underestimate the flow variability, which approaches one by increasing the number of latent features.

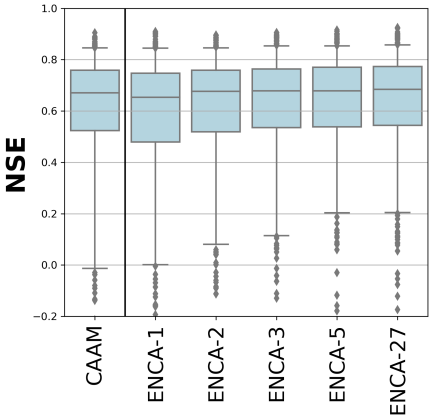

Figure 3. NSE values for the considered models in the test period. The boxes are delimited by the 25 % and the 75 % quantiles, while the whiskers indicate to the 5 % and 95 % quantiles. CAAM performance is similar to ENCA-2.

Figure 3 depicts the ~~boxplot of~~ boxplot of the test NSE values for the ~~validation NSE, BIAS and LOG-STDEV values for the considered models. The associated statistics are reported in~~ considered models. We report and discuss the associated statistics, the boxplots of the NSE components (R, BIAS, STDEV) and the correlations among them in Appendix B. ~~In all the metrics considered, the control model ENCA-0 achieves the worst results, which is consistent with what Kratzert et al. (2019) found. In particular, the big performance difference between ENCA-0 and CAAM is an indicator of the utility of the information contained in the 27 selected catchmentattributes in terms of streamflow prediction.~~

~~In general, and the related statistics () show also that increasing the number of latent features improves the prediction accuracy of the considered metrics. All the ENCA-N models (for $N > 0$) perform better (with respect all the metrics considered) than the control case ENCA-0.~~

In terms of NSE, we observe a performance improvement from ENCA-1 to ENCA-2 in the bulk of the distribution, and a further minor improvement from ENCA-2 to ENCA-3~~which produces a lower number of NSE outliers~~. The NSE ~~improvements~~ improvement between ENCA-3 and ENCA-5 is minor and is mainly related to outliers while ENCA-5 and ENCA-27 ~~are minor both in the outliers and in the bulk distribution. In terms of the BIAS, we observe a similar pattern. By~~ distributions are almost identical.

In general, Figure 3 and the related statistics (Figure B1) show that increasing the number of latent features ~~, the bulk distribution improves significantly. However, we observe the biggest gap between ENCA-2 and ENCA-3 and this gap is mostly related to high BIAS outliers.~~ improves the prediction accuracy of the considered metrics. Even though it is difficult to set a cut-off dimension, we can state that: i) with more than five latent features we do not observe a performance improvement anymore, meaning that 5 features are a sufficient set of summary statistics of the streamflow (which, however, can still depend on the chosen encoder architecture). ii) Overall, CAAM performance is most similar to ENCA-2~~in terms of BIAS and NSE~~. We therefore argue that ~~catchment attributes collected by experts~~ known catchment attributes (selected in this study) account for two relevant landscape features that appear to be sufficient for most catchments, while at least a third one is needed to resolve specific catchments.

~~Finally, regarding LOG-STDEV, it is evident that both CAAM and ENCA models tend to underestimate the streamflow variability. The results of ENCA-27 are most similar to CAAM while all other ENCA variants result in lower LOG-STDEV values. We hypothesize that ENCA models perform worse than CAAM in LOG-STDEV because of the different standardization procedure employed. However, the general underestimation of streamflow variability was also described in other LSTM-based hydrological modelling investigations (e.g. Kratzert et al., 2019). It can partly be attributed to using NSE as objective function for training which puts more weight on matching high flow. Also, using daily averaged data means covering faster dynamics and variability in the system. In addition, LSTM models that are trained on these data might not capture all the dynamics of even this daily data using the static attributes in CAMELS - that themselves are averages over entire catchments. The ENCA results show lower but increasing LOG-STDEV values indicating that using more latent features might help to better match the hydrograph variability. At the same time, ENCA-27, that would be "free" to learn whatever feature it deems necessary to be encoded to match the hydrograph variability, does now exceed the CAAM performance in this respect and this might confirm the existence of an upper bound to the LSTMs performance.~~

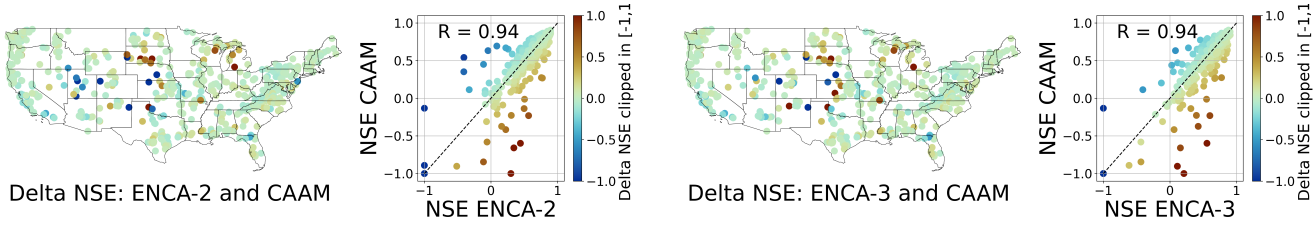

Figure 4. ~~Validation~~ Test NSE of ENCA-2 (left panel) and ENCA-3 (right panel) versus CAAM, color-coded with the NSE difference per catchment clipped in [-1,1]. Red means ENCA performs better, blue means CAAM performs better.

To study which catchments are most affected when using the latent features of the ENCA models in place of the known catchment attributes in CAAM, we report (Figure 4) the NSE difference between CAAM and ENCA-2 (left panels) or ENCA-3 (right panels), respectively. While, for most catchments, switching from ENCA-2 to ENCA-3 does not result in a high performance gain, we see a clear improvement on about a dozen or so catchments, mostly located in the central CONUS. This corroborates the hypothesis that the ~~collected~~ known catchment attributes account for two relevant landscape features and the

improvement due to the third one is related to only few catchments that are particularly difficult to predict. It is interesting to count the number of catchments for which CAAM~~fails (negative NSE values, ~~'s NSE is negative (i.e. predictions that are worse than average streamflow) but ENCA~~succeeds (positive NSE values)~~'s NSE is positive. This number ~~increases from 13 (~~is 17 for ENCA-2 ~~) to 21 (~~and 15 for ENCA-3~~)~~. On the other hand, ~~there are only 9 (3) catchments ,~~ the number of catchments for which CAAM~~succeeds but ~~'s NSE is positive but ENCA's negative decreases from eight (ENCA-2~~(-~~)~~ to only two (ENCA-3)~~fails~~.

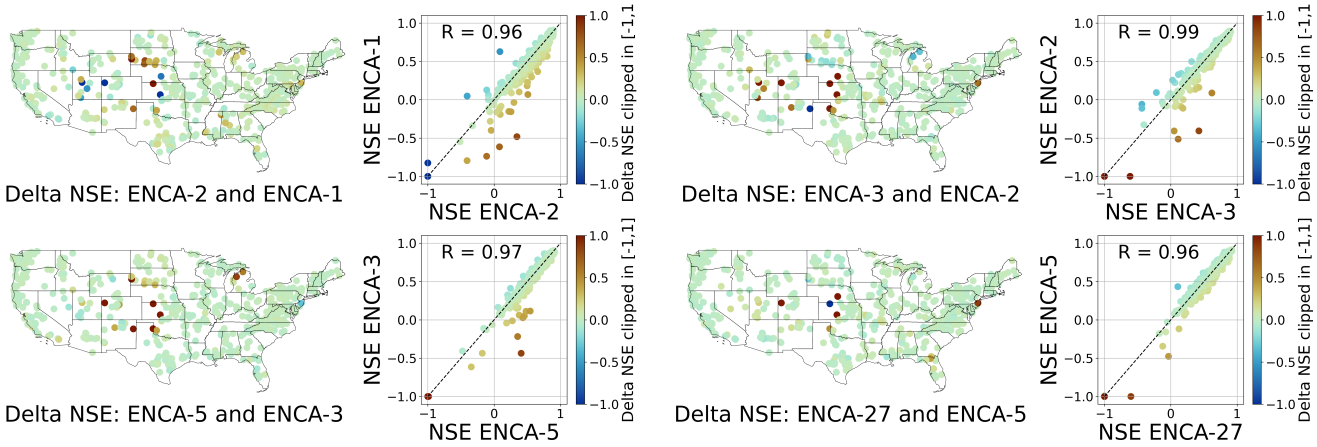

Figure 5. ~~Validation~~ Test NSE of ENCA models with different number of latent features, color-coded according to the NSE difference per catchment clipped in [-1,1]. The improvement of increasing the number of latent features is significant from ENCA-2 to ENCA-3 and marginal for more complex models. The improved catchments are mainly located in the central CONUS, dominated by arid climate conditions.

In order to evaluate the impact of additional features, we compare the performances of ENCA models differing in their number of latent variables (Figure 5). The results corroborate our earlier findings that two features are sufficient to cover most of the catchments, and additional features provide information about relatively few, difficult to predict, ~~mostly catchments~~ catchments mostly located in the central CONUS and dominated by arid climate conditions. While the number of such catchments informed by the third feature is relatively high, additional features only have a minor effect. Indeed, adding a third feature turns ~~15 catchments from failures (negative NSE) to successes (positive NSE)~~ seven catchments from negative NSE to positive NSE and only leads to marginal deterioration ~~on a few~~ of three other catchments. Additional features have much less dramatic effects. ~~In order to understand the characteristic of the improved catchments, in we report the performance metrics and some attributes of the 15 catchments failing under ENCA-2, but succeeding under ENCA-3. They are generally characterized by high aridity indexes and intermittent flows, i.e. time-windows in the streamflow time series with low to zero flow, and are for this reason very difficult to predict.~~

~~Note, in we computed the GRIDE paths of~~

Note that test NSE obtained in this work are good, but still far from state-of-the-art approaches on the same dataset. For comparison, the ~~ENCA-27 and ENCA-5 latent spaces and concluded that most catchments can be characterized by only three~~

relevant landscape features, while more are only needed for a few special cases. This discrepancy may arise global model of Kratzert et al. (2019) (augmented with the same catchment attributes reported in Table 1) achieves a median NSE of 0.74, while in this work the best model achieves a median NSE of 0.68. Later approaches (with multiple input forcings) achieve even higher performance of 0.82 (Kratzert et al., 2021). One might be tempted to attribute this to over-fitting due to the presence of two or more manifolds with different IDs in the latent space of ENCA. An interesting direction of investigation would be to study this latent space with the ID estimator HIDALGO (Allegra et al., 2020), which allows consideration of multiple manifolds with different IDsvery large number of parameters of our architecture (about 3 million). However, the test MSE loss curves (Figure C1) do not support this hypothesis. Also the very long sequences fed to the LSTMs might affect their performance, as they are known to suffer from vanishing/exploding gradients. While we use time-series of length 5478, state-of-the-art approaches use lengths of 270 (Kratzert et al., 2019) and 365 (Kratzert et al., 2021).

4.2 Interpretation of the Relevant Feature Principal Components

Figure 6 shows the absolute Spearman correlation matrix between the principal components of the identified three relevant featuresand , the known streamflow signatures and catchment attributes across different random restarts of the model. The relevant features share information with catchment attributes and hydrological signatures. For instance, feature one carries information about basic hydrological attributes like baseflow index and low flow frequency. Moreover, feature one is (weakly) correlated with soil-related attributes like soil porosity and conductivity, sand, silt and clay fraction. Feature two is correlated with climatic indicators, such as the aridity index, the mean precipitation, high and low precipitation frequency, but also with hydrological signatures like mean discharge streamflow and the 95% quantile of the flow duration curve. We point out that even though the encoder is explicitly designed to learn non-climate landscape features, we can still observe a correlation between latent features and climate attributes. This correlation can be due to collinearities between landscape and climate attributes. In this case, the collinear attributes are those related to vegetation, like the fraction of forests and the maximum GVF (Green Vegetation Fraction), which are obviously correlated with climate. For instance, from Figure D1 we can observe that the aridity index is highly correlated with the mean precipitation (0.88) and the fraction of forest (0.74), while these last two attributes are fairly correlated between each other (0.67).

Finally, feature three is mostly correlated with high and low flow duration and frequency, signatures relating to the extremes of streamflow. Interestingly, this principal component does not hold much information about neither landscape nor climate attributes, indicating that it encodes catchment information that has not yet been considered or that is not related to any discernible catchment feature. Since the third feature mainly conveys information about certain dry and hard to predict catchments, the latter might very well be the case.

The discussed principal components, however, do not share the same amount of explained variance: feature one accounts for about 60%, feature two for about 30% and feature three for about 10%. In Appendix E we report the correlation matrices for the other ENCA models, where we can verify that the principal components carry the same information for streamflow prediction consistently across different models. A specific analysis of the correlation between NSE improvements and certain static attributes is provided in .

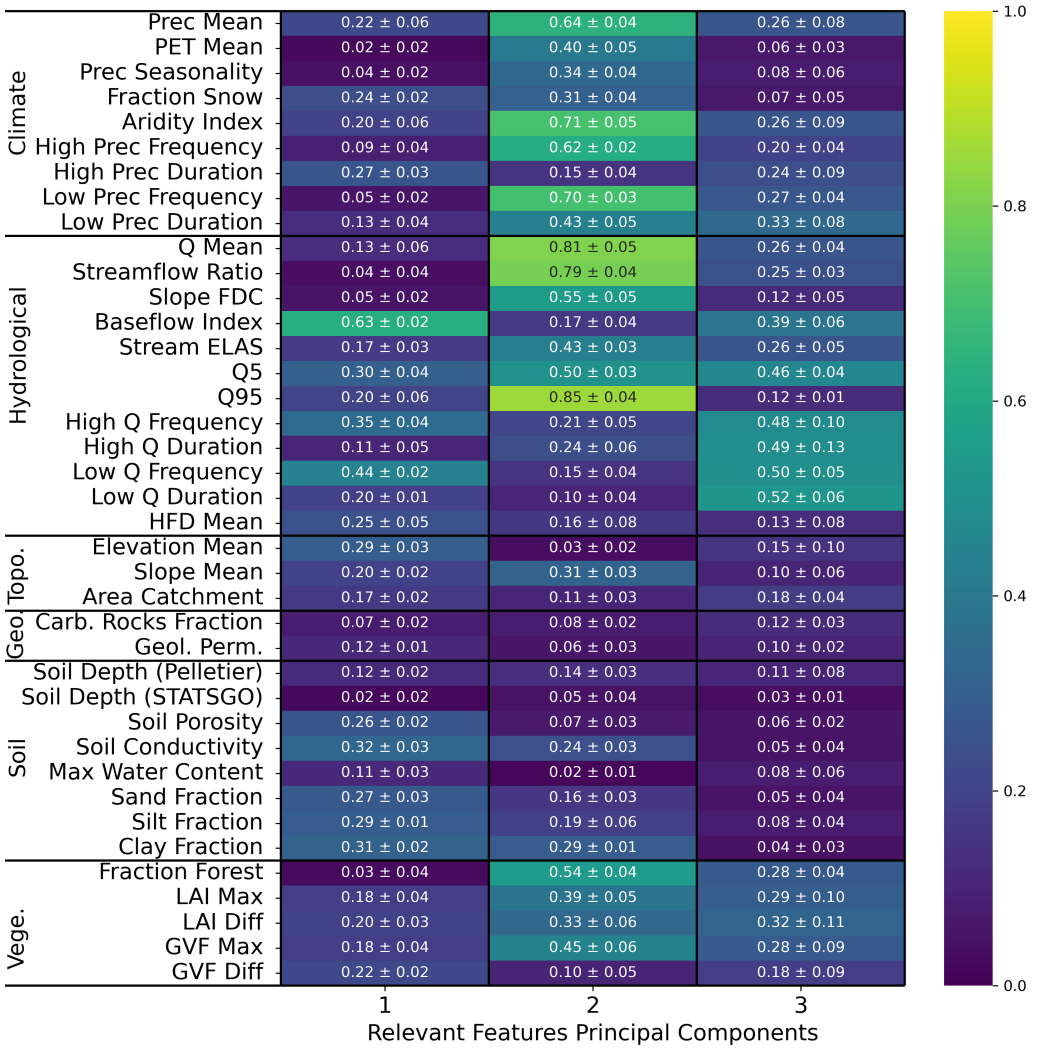

Figure 6. Average (plus minus the standard deviation) of the absolute Spearman correlation ~~matrix~~ of ~~the~~ relevant features principal components of ENCA-3 with respect catchment attributes and hydrological signatures across four different random model restarts. The colors refer to the average.

The geospatial distribution of feature importance is shown in Figure 7, where a non-trivial distribution of the features appears, highlighting that the different features have different information content for different regions: feature one dominates in the less to non-arid eastern CONUS, while feature two is mainly dominant in the western part. Feature three does not show such a clear spatial representation. Overall, the potential of delineating geospatial relations is another indicator that the encoder has learnt from the landscape signal in the data.

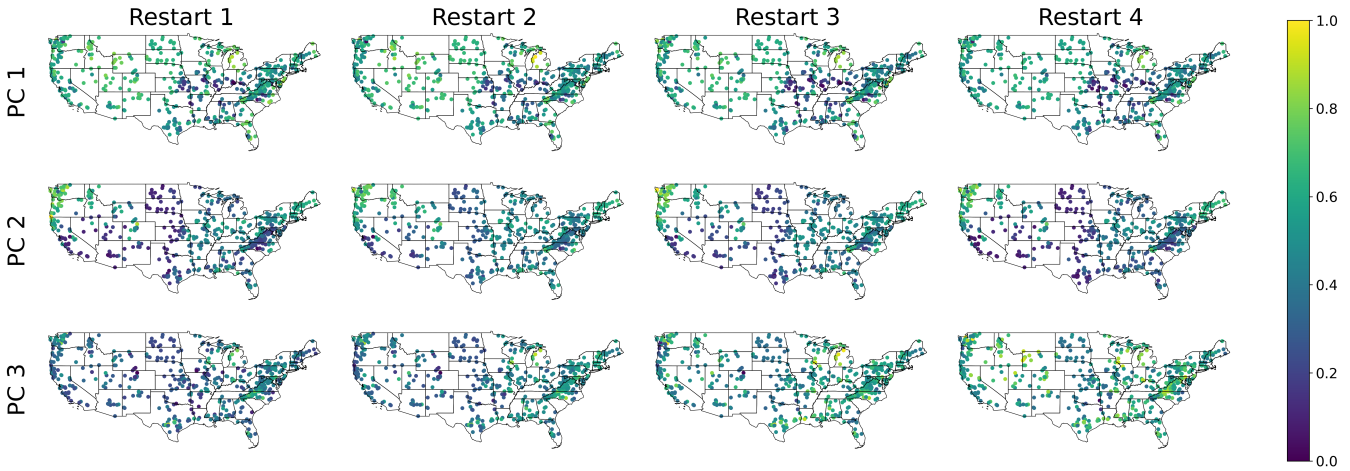

Figure 7. Principal ~~components~~ Components (PCs) of the relevant features of ENCA-3 in the CONUS. For a better comparison across different restarts, PCs are normalized in the interval $[0, 1]$ and their sign is adjusted such that the PCs of the first catchment in the dataset are always positive.

~~The elicitation of principal feature components further allows us to pinpoint a subset of the particularly bad performing catchments. These catchments show intermittent flows and they are characterized by relatively high aridity indices (see for the catchment characteristic and the hydrographs). depicts the learnt features of ENCA-3, color-coded according to baseflow index (a), aridity (b) and the high flow frequency (c). These are the attributes that show the highest correlation with the learnt features of ENCA-3. The red diamonds represent those 15 catchments failing under ENCA-2, but succeeding under ENCA-3 (see also). They mainly lie in a sub-region of the latent space characterized by high aridity and low baseflow.~~

~~Relevant features of ENCA-3 for a random restart, colored-coded according some standardized catchment attributes which correlate strongly with the first three principal components. Red diamonds are those catchments whose prediction has improved from failure (ENCA-2) to success (ENCA-3). These catchments are characterized by intermittent flows.~~

4.3 ~~Relationship between Relevant Features and Parameters in Conceptual Models~~

~~In the literature, a great number of conceptual models are available. It is well known that only a handful of parameters can be reliably estimated from the rainfall-streamflow data (Jakeman and Hornberger, 1993), indicating that models with a few parameters and states, like the GR3J (Edijanto et al., 1999) and its successor GR4J (Perrin et al., 2003), must capture the main features of a hydrograph through their structure and parameters. As a consequence, in makes sense to try to relate the parameters of such models to the learnt features identified in this work.~~

~~The GR4J model has only four parameters that can be related to specific characteristics of a hydrological system: (i) the maximum capacity of the production store; (ii) the groundwater exchange coefficient that influences the catchment mass~~

balance; (iii) the maximum capacity of the routing store; (iv) the time base of the unit hydrograph that controls the time lag between rainfall and streamflow.

It is conceivable that the number of parameters that need to be calibrated to a specific catchment should be similar to the number of non-meteorological features in the runoff and thus to the minimal number of features needed by ENCA to make good runoff reconstructions. Although a one-to-one mapping may not be possible, we can try to connect these parameters to the relevant learnt features: The first GR4J parameter (maximum capacity of production store) could be related to the second feature (vegetation attributes) since it relates to how much water actually ends up in storage or streamflow and how much is evapotranspired. This threshold parameter in the rainfall-runoff relationship has been related to the root zone and to vegetation indicators (obviously also related to climate) by previous work (Gao et al., 2014). The first relevant feature can be related with the third GR4J parameter that relates to the routing store capacity. In particular, the routing store capacity in GR4J mainly affects streamflow recessions. The first relevant feature is related to baseflow, therefore arguably related to a similar fingerprint the hydrograph. Soil-related attributes have been traditionally related to baseflow (Gnann et al., 2021) which underpins the relation to subsurface storage.

Including the third ENCA-found feature improved performance notably. Meanwhile, it could only weakly be related to static catchment attributes showing some correlation to high and low flow hydrological indices regarding frequency and duration, and to vegetation attributes in Figure 6. Further, including this feature generally helped minimizing the gap between baseflow predictions and observations as it is shown in Figure **??**. While the former correlation relates to timing in the hydrograph, the latter two points refer to the water-balance or hydrograph magnitude.

Hence, we conjecture that feature three relates to the two other parameters in GR4J: time lag and groundwater exchange coefficient. First, the third feature results in a smoothing of hydrographs and buffering of discharge spikes (as indicated in). In GR4J, the time lag parameter has a similar effect, in other hydrological models this is sometimes referred to as a routing parameter. Hence, the third feature appears to play a role in regulating the timing of increase and decrease of the memory states of the LSTM. It is noteworthy that these hidden and cell states of the LSTM essentially resemble the role traditionally held by reservoirs in conceptual models and that any routing routine or time lag parameter in conceptual hydrological models acts as a convolutional filter to match the reservoir output to the observed hydrograph. Second, we conjecture that it accounts for offsets in the water balance (see) potentially due to water exchange with other catchments. Such a relaxation of a strictly enforced water balance for a modelled catchment further improves the model performance. Interestingly, it was shown that when using LSTMs for streamflow prediction a strictly enforced water balance deteriorates model performance (Frame et al., 2022). From a hydrological perspective this makes sense: even if the surface delineation of a catchment might be well known, the subsurface delineation might not be identical and unknown exchange fluxes may occur. Further, the observed water inputs and outputs of the system are per se subject to uncertainty and therefore a full closure of the water balance is not guaranteed. LSTMs appear to account for resulting offsets which also holds for our LSTM-ENCA.

Overall, these potential resemblances illustrate how the learning mechanism encapsulates distinctive hydrological characteristics embedded in the model's parameters, making them at least partially interpretable.

5 Conclusions and Outlook

We employed a ~~new kind of~~ conditional autoencoder to distill a minimal set of streamflow features (signatures) necessary for streamflow reconstruction in conjunction with meteorological data. Thus, these features can be interpreted as landscape fingerprints on the streamflow. We compared these features with known catchment attributes in terms of their capacity for streamflow reconstruction. The primary conclusions we highlight in this study are:

- For all the metrics considered, ENCAs (Explicit Noise Conditional Autoencoders) perform better ~~in terms of NSE and BIAS~~ than the reference attributes enhanced model (CAAM) when the number of latent features is greater than two. In fact, two features seem to be sufficient for most catchments, while a relatively small number of catchments, mostly located in the central CONUS, require a third one. Including more than three features, however, only leads to marginal improvements. We therefore conjecture that most of the information contained in the static attributes used for CAAM, insofar as it is relevant for streamflow prediction, can be reduced to two independent features. The third latent feature, however, seems to encode information that is not fully contained in those static attributes.

- The correlation between attributes and importance of the relevant features (see Figure 6) suggests an ordering of the information contained in the features for accurately predicting discharge: ~~fFirst~~first, basic hydrological attributes like baseflow and soil-related attributes, followed by the average streamflow and the 95% flow quantile (correlated to climate due to collinearities with vegetation-related attributes) and, third, specifics on the high and low flow, i.e. the extremes of the hydrograph. Looking back at Figure 4, this last feature appears to encode the information that is needed to exceed the model performance that is only based on the 27 static attributes (CAAM).

In summary, our research reveals a significant reduction in the dimensionality of the streamflow time series. Despite the plethora of hydrological signatures and catchment attributes at our disposal, only a small subset proves essential for accurate streamflow prediction. This finding echoes established results from prior studies (Jakeman and Hornberger, 1993; Edijanto et al., 1999; Perrin et al., 2003), suggesting that hydrological systems might be effectively modelled using only a limited set of parameters. The low dimensionality of the relevant catchment information opens up the opportunity for a better ~~*understanding*~~ understanding of its nature, suggesting some future research directions:

- A promising approach could be the adoption of NeuralODEs (Höge et al., 2022), which offer a high level of interpretability due to their low number of states. This combination of a few states and a few features may help to decipher not only the nature of the relevant catchment information but also how it influences streamflow.

- Preliminary analysis (not shown in this paper) has revealed that the known static catchment attributes live on a low-dimensional manifold, which is in line with our finding that only two independent features seem to capture most of the information that is relevant for streamflow. While the correlation-based analysis presented in this paper gives some clues as to how these features can be interpreted, more sophisticated types of analysis like those based on Information Imbalance (Glielmo et al., 2022) might allow for a more precise understanding of their physical nature.

Code and data availability. The US-CAMELS dataset, as well as the catchment attributes, is available at the site https://ral.ucar.edu/solutions/products/camels. The extended NLDAS forcing dataset is available at https://doi.org/10.4211/hs.0a68bfd7ddf642a8be9041d60f40868c. All the code used for this work is publicly available at https://doi.org/10.5281/zenodo.13132951.

Appendix A: ~~Validation Metrics~~The Intrinsic Dimension

In order to estimate the ID, we apply the GRIDE estimator (Denti et al., 2022). Given sample points, $\mathbf{x}_i \in \mathbb{R}^D$, for $i = 1, \ldots, M$, and a distance measure, $r : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^+$, GRIDE assumes that points in a given neighbourhood are counted with a Poisson point process with intensity ρ, which is constant at least up to the scale of the diameter of the considered neighbourhood. We define $r_{i,l}$ as the distance between the point $\mathbf{x}_i$ and its l -th nearest neighbour. Let us define $\mu_{i,n_2,n_1} = \frac{r_{i,n_2}}{r_{i,n_1}}$, where n_1 and n_2 (with $0 \leq n_1 \leq n_2 \leq M$) are nearest neighbours of generic order. The distribution of μ_{i,n_2,n_1} can be computed in closed form and depends only on the ID of the data while, crucially, does not depend on ρ, as long as ρ is constant in the considered neighbourhood of i whose diameter is set by the distance between i and its n_2-th nearest neighbour (Denti et al., 2022).

In order to correctly identify the ID of a dataset, a scale-independent analysis is essential. We therefore make use of GRIDE paths, the evolution of the ID estimate as a function of n_2, which can be interpreted as the scale at which we look at the data. We set $n_1 = n_2/2$, as usually done in the literature. As a function of n_2, the ID is first expected to increase, due to the noise present at small distance scales, and then to reach a plateau corresponding to the correct ID.

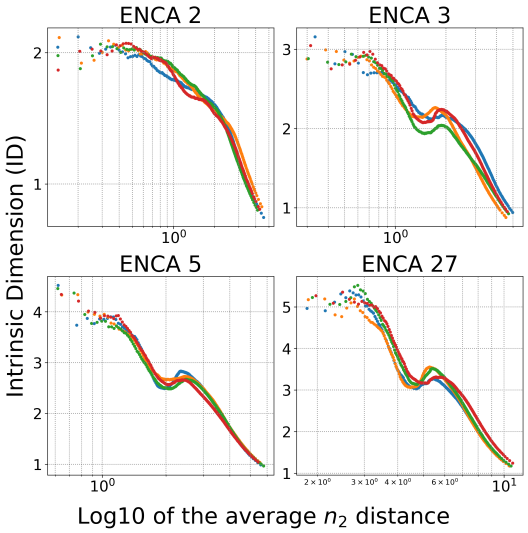

Figure A1. GRIDE evolution plot for the different ENCA models employed for the four random restarts of the models.

Figure A1 shows the GRIDE path for different ENCA models trained in this work. In particular, the ID estimate of the latent space of ENCA-27 decreases after showing a plateau around five, then reaches a minimum around three, then increases again and finally collapses to low values at larger distance scale. The plateau at five motivates us to train an ENCA-5 and study its ID. The local minimum of the GRIDE path of the latent space of ENCA-5 is consistent with an ID of three. We can deduce that, for most of the catchments, the ID is three, while for some it can be higher. However, the fact that the GRIDE path of the latent space of ENCA-27 shows two plateaus around five and three can be an indicator of the existence of two or more manifolds with different IDs. From Figure 3 we see that, indeed, three features seem to capture most of the relevant information, which

is in line with the GRIDE path for ENCA-5.

Appendix B: Metric Comparisons

We report some summary statistics of the NSE, ~~BIAS and LOG-STDEV~~ R, BIAS and STDEV values across the 568 catchments considered in this study.

In terms of of the linear correlation coefficient R and the BIAS (left and central panels of Figure B1), we observe a similar pattern to what we observed for the NSE distribution (Figure 3). By increasing the number of latent features to three, the bulk distribution improves significantly. At the same time, further increasing the number of latent features improves only the BIAS outliers.

Regarding the STDEV (right panel of Figure B1), it is clear that both CAAM and ENCA models tend to underestimate streamflow variability, a known issue associated with using NSE as the objective function (Gupta et al., 2009). However, the differences between CAAM and ENCA models are less pronounced, and unlike the observations for NSE, R, and BIAS, a clear performance hierarchy cannot be established.

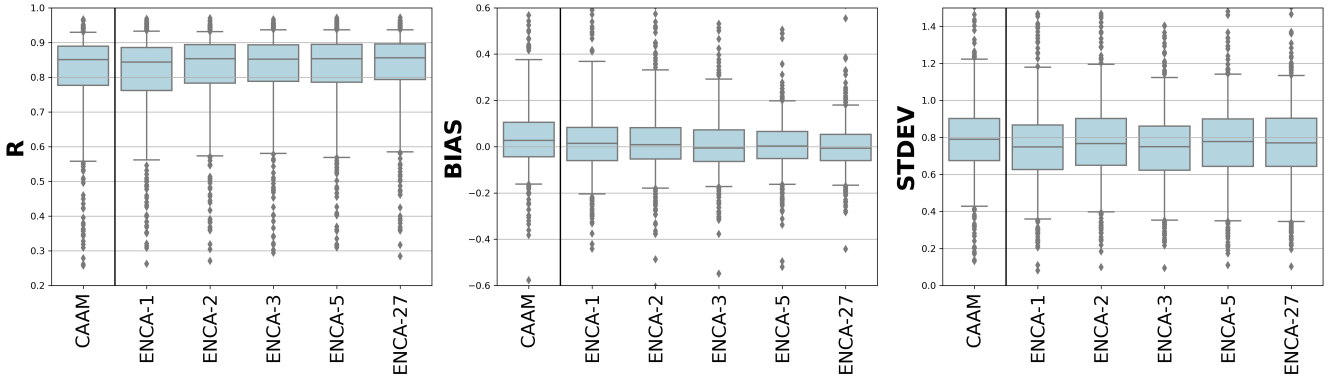

Figure B1. R, BIAS and STDEV values for the considered models in the test period. The boxes are delimit by the 25 % and the 75 % quantiles, while the whiskers indicate to the 5 % and 95 % quantiles. For these metrics as well, CAAM performance is similar to ENCA-2.

Table B1 ~~shows that the greatest performance gap in terms of the NSE is obtained for the mean, the minimum and the 5% quantile~~ demonstrates that increasing the number of latent features in ENCA generally enhances performance. However, the improvement is modest between ENCA-~~2~~ 3 and ENCA-~~3, indicating~~ 5 and nearly negligible between ENCA-5 and ENCA-27. The performance of CAAM is more comparable to ENCA-2 overall, suggesting that the third latent feature is ~~needed to improve prediction in particularly difficult catchments. A similar pattern is shown by the BIAS, whereas the biggest gap is found in the mean, maximum and 95% quantile. Instead, we observe a general tendency of ENCA models to underestimate the flow variability.~~ crucial for better predictions in certain catchments, while five features appear to be the maximum number our encoder can effectively learn. Furthermore, as shown in Figure 5, it is clear that almost all catchments that improved from ENCA-3 to ENCA-5 still exhibit very poor performance (NSE below -1.0).

Appendix C: ~~Analysis of Improved Catchments from ENCA-2 to ENCA-3~~

From Table B1, we can observe that the NSE is strongly correlated with the linear coefficient R and fairly correlated with STDEV and the correlation increases with the performance (see Figure 3). At the same time, the NSE is anti-correlated with the BIAS, but this correlation is low. While the correlation between R and BIAS and between BIAS and STDEV is weak, the correlation between R and STDEV is intermediate and increases with performance. These results agree with the findings of (Gupta et al., 2009), who showed that with optimal BIAS values, optimal NSE values are reached when R is correlated with STDEV.

~~In we report the hydrographs of the catchments failing under ENCA-2 (i. e. with NSE values below zero), but succeeding under ENCA-3 (i. e. with NSE values above zero). We report the corresponding validation metrics (NSE,~

		~~CAAM~~ CAAM	~~ENCA-0~~ ENCA-1	~~ENCA-1~~ENCA-2	~~ENCA-2~~ ENCA-3	~~ENCA-3~~ ENCA-5
	Mean	~~0.48~~ 0.52	~~0.06~~ 0.47	~~0.43~~ 0.50	~~0.43~~ 0.53	~~0.54~~ 0.58
	Min	~~-23.78~~ -17.66	~~-71.46~~ -15.08	~~-18.62~~ -17.74	~~-22.45~~ -14.77	~~-10.66~~ -11.03
	Q5	~~-0.05~~ -0.02	~~-0.61~~ -0.00	~~-0.03~~ 0.07	~~-0.10~~ 0.11	~~0.13~~ 0.19
	Q25	~~0.54~~ 0.52	~~0.42~~ 0.48	~~0.51~~ 0.52	0.53	0.54
NSE	Median	~~0.69~~ 0.67	~~0.60~~ 0.65	~~0.67~~ 0.68	0.68	0.68
	Q75	~~0.77~~ 0.76	~~0.73~~ 0.75	~~0.74~~ 0.76	0.76	0.77
	Q95	0.85	~~0.83~~ 0.85	0.85	0.86	0.85
	Max	0.90	0.91	0.90	0.91	0.92
	~~0.92~~ # < 0.0	~~0.92~~ 31	~~0.93~~ 29	22	18	14
	Mean	~~0.22~~ 0.81	~~0.37~~ 0.80	~~0.29~~ 0.82	~~0.32~~ 0.82	~~0.17~~ 0.82
	Min	0.11	0.12	0.14	~~0.09~~ 0.15	0.13
	~~Min~~ Q5	~~-0.55~~ 0.55	~~-0.66~~ 0.55	~~-0.82~~ 0.57	~~-0.67~~ 0.58	~~-0.69~~ 0.57
	~~-0.61~~ Q25	0.78	0.76	0.78	0.79	0.79
R	~~Q5~~ Median	~~-0.22~~ 0.85	~~-0.23~~ 0.84	~~-0.27~~ 0.85	~~-0.24~~ 0.85	~~-0.26~~ 0.85
	~~-0.25~~ Q75	0.89	0.89	0.89	0.89	0.89
	Q95	0.93	0.93	0.93	0.94	0.94
	Max	0.97	0.97	0.97	0.97	0.97
	Mean	0.05	0.05	0.05	0.02	0.02
	Min	-1.13	-0.86	-0.86	-1.05	-0.52
	Q5	-0.16	-0.21	-0.18	-0.17	-0.16
	Q25	~~-0.05~~ -0.04	~~-0.02~~ -0.06	-0.05	-0.06	~~-0.08~~ -0.05
BIAS	Median	~~0.07~~ 0.03	~~0.12~~ 0.01	~~0.09 0.05 0.04~~ 0.01	-0.01	0.00
	Q75	~~0.24~~ 0.11	~~0.38~~ 0.08	~~0.31~~ 0.08	~~0.26~~ 0.07	~~0.21~~ 0.07
	Q95	~~1.16~~ 0.40	~~1.81~~ 0.40	~~1.52~~ 0.34	~~1.66~~ 0.30	~~0.87~~ 0.20
	Max	~~6.03~~ 1.70	~~15.26~~ 2.87	~~11.12~~ 2.53	~~15.63~~ 2.03	~~8.64~~ 1.70
	Mean	0.82	~~0.78~~ 0.77	0.80	0.76	~~0.75~~ 0.78
	Min	~~0.09~~ 0.13	0.08	~~0.07~~ 0.10	~~0.06~~ 0.09	~~0.09~~ 0.11
	Q5	~~0.23~~ 0.42	~~0.19~~ 0.35	~~0.23~~ 0.40	~~0.18~~ 0.35	~~0.27~~ 0.35
	Q25	~~0.60~~ 0.67	~~0.51~~ 0.63	~~0.56~~ 0.65	~~0.56~~ 0.62	~~0.60~~ 0.64
~~LOG-STDEV~~STDEV	Median	~~0.78~~ 0.79	~~0.70~~ 0.75	~~0.72~~ 0.77	~~0.73~~ 0.75	~~0.75~~ 0.78
	Q75	~~0.95~~ 0.90	0.87	~~0.88~~ 0.90	~~0.88~~ 0.86	~~0.91~~ 0.90
	Q95	~~1.44~~ 1.23	~~1.47~~ 1.18	~~1.36~~ 1.20	~~1.33~~ 1.13	~~1.33~~ 1.14
	Max	~~7.23~~ 4.26	~~10.95~~ 3.64	~~3.81~~ 4.12	~~4.03~~ 3.88	~~4.26~~ 3.55

Table B1. Metrics comparison for different models computed in the test period. We report the mean, the minimum, the 5% quantile, the 25% quantile, the median, the 75% quantile, the 95% quantile and the maximum values~~of~~ . Additionally, we report the ~~distribution~~ number of ~~validation~~ catchments whose predicted NSE values ~~for the three metrics considered in this work~~are lower than zero.

NSE - R
NSE - BIAS ~~and LOG-STDEV) and some catchment attributes in . Observed and predicted (CAAM, ENCA-2 and ENCA-3) hydrographs for those 15~~
NSE - STDEV
R - BIAS
R - STDEV
BIAS - STDEV

Table B1. ~~Catchments failing under ENCA-2, but succeeding under ENCA-3~~Linear correlation coefficient between different metric in the models analyzed in this work. ~~These catchments~~ Catchments whose NSE prediction is lower than zero are ~~characterized by high aridity indexes and intermittent flows~~excluded.

Appendix C: Neural Networks Details and Training Losses

We report the architecture details of the encoder (Table C1) ~~and the LSTM decoder ()~~ of the ENCA models used in this work. For the encoder, we chose a single layer uni-directional LSTM, followed by a dropout layer (with probabiliy 0.4) and a fully connected layer that maps the LSTM output layer to the predicted output. For the LSTM, we chose a hidden size of 256 (number of memory cells), an initial forget bias of 5.

~~The Convolutional encoder architecture used in this study. Batch Normalization (BN) and Dropout (DR) with probability 0.4 are added between layers. A last BN layer is applied to the decoder output in order to standardize the latent features. N is the number of latent features.~~

Input	Layer name	Hyper-parameters	Output
	streamflow	input	(BS, 5478, 1)
streamflow	Conv 1	7, 8, BN, Leakyrelu, DR(O.4)	(BS, 5472, 8)
Conv 1	Avgpool 1	4 (BS, 1368, 8)	
Avgpool 1	Conv 2	5, 16, BN, Leakyrelu, DR(0.4)	(BS, 1364, 16)
Conv 2	Avgpool 2	4	(BS, 341, 16)
Avgpool 2	Conv 3	2, 32, BN, Leakyrelu, DR(0.4)	(BS,340, 32)
Conv 3	Avgpool 3	4	(BS, 85, 32)
Avgpool 3	Flatten	N/A	(BS, 2720)
Flatten	Linear	BN, Leakyrelu, DR(0.4)	(BS, 512)
Linear	Output	BN	(BS, N)

Table C1. The Convolutional encoder architecture used in this study. Batch Normalization (BN) and Dropout (DR) with probability 0.4 are added among layers. A last BN layer is applied to the decoder output in order to standardize the latent features. N is the number of latent features. The batch size is indicated with BS.

~~Hidden size~~ In Figure C1 we report the training and test losses of some models employed. We observe that all the models employed are about to reach a plateau, where the test loss does not decrease anymore. Though convergence is not perfectly reached due to computational limitations, the fact that the test loss is almost at the reachable minimum is an indicator that the models are not overfitting the dataset.

Additionally, we report the mean and standard deviations of the latent features of ENCA-5 (Table C2). We can appreciate a small amount of bias, even if the encoder succeeds in preserving the standard deviation of the latent features close to one. We found a similar behaviour in the latent features of other ENCA models (not reported).

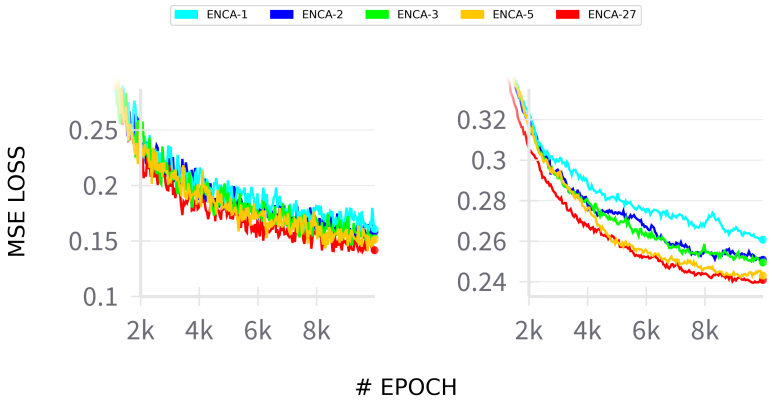

Figure C1. Training MSE loss (left panel) and test loss (right panel) during training for different ENCA models. The curves shown are obtained by averaging the losses across different random restarts. Loss variability across random restarts (not shown) is negligible.

Mean	~~Initial forget bias~~ 0.23	~~LSTM layers~~ 0.27	~~Dropout~~ 0.07	~~Bi-directional~~ 0.25	0.28
~~256~~ heightStandard Deviation	~~5.0~~ 1.15	~~1~~ 1.1	~~0.4~~ 0.92	~~False~~ 0.67	0.99

Table C2. ~~LSTM hyper-parameters. We choose~~ Mean and standard ~~hyper-parameters used in~~ deviation of the ~~literature (Kratzert et al., 2019)~~ latent features of ENCA-5.

Appendix D: ~~Performance Correlation with Known Attributes~~

~~We also report the NSE difference between ENCA-3 and ENCA-2 versus the values of some chosen known static attributes (). We can clearly appreciate a correlation between the performance improvement of ENCA-3 with respect CAAM for those catchments with baseflow index greater than 0.25, and for aridity indexes greater than 1.0, indicating that the learnt features are particularly important for improving the prediction accuracy in more arid catchments and for those basins with greater amount of underground water. A similar pattern is present for the Q95, which can be explained by the fact that autoencoders tend to improve the prediction of high flow peaks.~~

Appendix D: Known Attributes and Signatures Correlation

~~NSE difference between ENCA-3 and ENCA-2, clipped in -1,1, vs some static attributes and signatures.~~

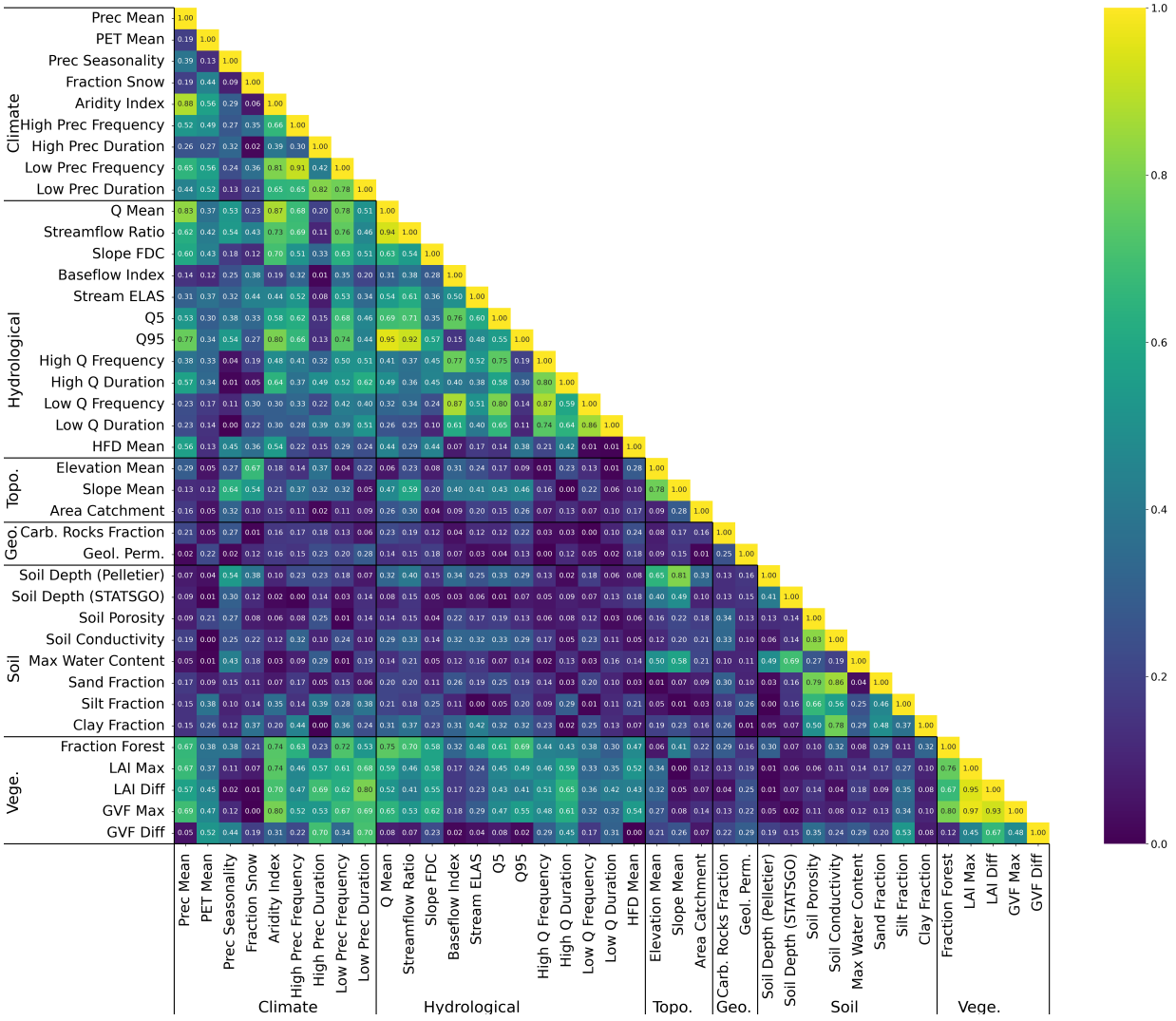

Figure D1. Absolute Spearman correlation matrix of the selected catchment attributes and hydrological signatures between each other.

Appendix E: Effect of Random Restart

We ~~clearly~~ ascertain that the random restart does not affect much the prediction accuracy (Figure E1). Apart from some catchments, most of them show a consistent behaviour across different random seeds.

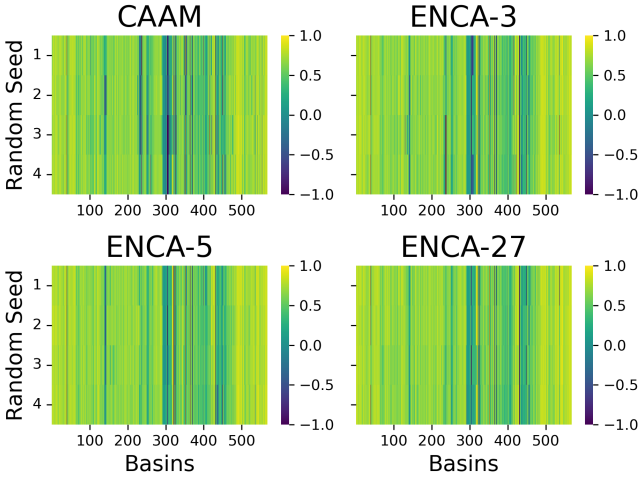

Figure E1. ~~Validation~~ Test NSE values of ~~the~~ four random restarts for CAAM (upper left), ENCA-~~2~~ 3 (upper right), ENCA-~~3~~ 5 (lower left) and ENCA-27 (lower right) for the 568 catchments considered ~~across four random restarts of~~ in this study. NSE values are clipped in the ~~model~~interval $[-1, 1]$. We do not observe much performance variability across different random restarts~~of the models~~.

We report the correlation matrix between the principal components of the learnt features of ENCA-~~4 (), ENCA-~~5 (Figure E2) and ENCA-27 (Figure E3) for different random restarts. We notice a consistency across random restarts and different models. Moreover, the correlation becomes weaker and weaker with the fourth component, indicating that ~~3~~ three features carry most of the information related to streamflow prediction.

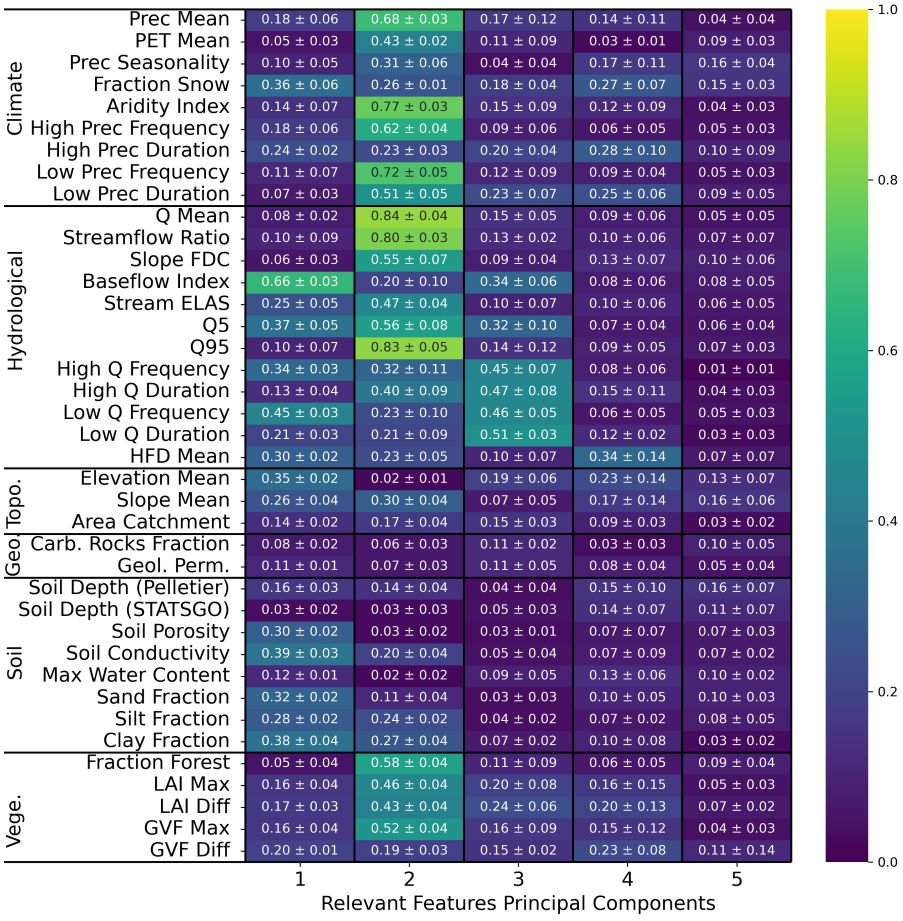

Figure E2. Average (plus minus the standard deviation) of the absolute Spearman correlation ~~matrix~~ of ~~the~~ relevant features principal components of ENCA-~~4~~ 5 with respect ~~to the selected~~ catchment attributes and hydrological signatures across four different random model restarts~~of~~. The colors refer to the ~~model~~average.

~~Spearman correlation matrix of the relevant features principal components of ENCA-5 with respect to the selected catchment attributes and hydrological signatures across four different random restarts of the model.~~

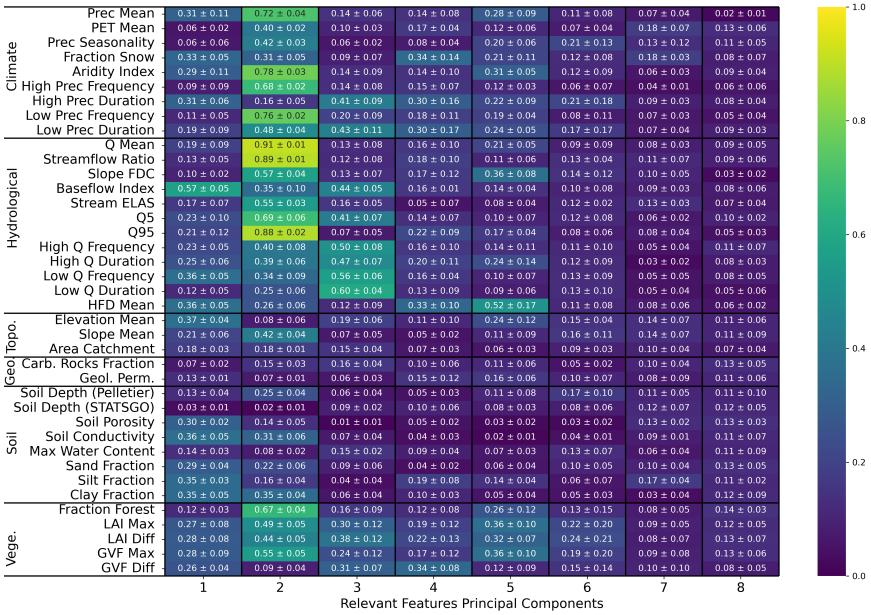

Figure E3. Average (plus minus the standard deviation) of the absolute Spearman correlation ~~matrix~~ of ~~the~~ relevant features principal components of ENCA-27 with respect ~~to the selected~~ catchment attributes and hydrological signatures across four different random model restarts~~of~~. The colors refer to the ~~model~~average. For a better visualization, we report only the first eight Principal Components.

Author contributions. CA had the original idea and AB and CA developed the conceptualization and methodology of the study. The idea of using intrinsic dimension is of AM. AB developed the software and conducted all model simulations and their formal analysis. Results were discussed and interpreted between MH, CA, AM, FF and AB. The visualizations and the original draft of the manuscript were prepared by AB, and reviewing and editing were provided by MH, CA, AM and FF. Funding was acquired by AM and CA. All authors have read and agreed to the current version of the paper.

Competing interests. At least one of the (co-)authors is a member of the editorial board of Hydrology and Earth System Sciences. The authors also have no other competing interests to declare.

Acknowledgements. This research has ~~partly been~~ been partly founded by the SNSF (Swiss National Science Foundation) grant 200021_208249. We would like to thank Antonio di Noia (Università della Svizzera italiana, ETH Zurich), Fernando Perez Cruz (ETH Zurich), Andreas Scheidegger (Eawag)~~and Marco Baity-Jesi~~ , Marco Baity-Jesi (Eawag) and Dmitri Kavetski (The University of Adelaide) for the insightful discussions.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Albert, C., Künsch, H., and Scheidegger, A.: A Simulated Annealing Approach to Approximate Bayes Computations, Statistics and Computing, 25, 1217–1232, https://doi.org/https://doi.org/10.1007/s11222-014-9507-8, 2015.

Albert, C., Ulzega, S., Ozdemir, F., Perez-Cruz, F., and Mira, A.: Learning Summary Statistics for Bayesian Inference with Autoencoders, SciPost Phys. Core, 5, 043, https://doi.org/10.21468/SciPostPhysCore.5.3.043, 2022.

Allegra, M., Facco, E., Denti, F., Laio, A., and Mira, A.: Data segmentation based on the local intrinsic dimension, Scientific Reports, 10, 16 449, https://doi.org/10.1038/s41598-020-72222-0, 2020.

Botterill, T. E. and McMillan, H. K.: Using Machine Learning to Identify Hydrologic Signatures With an Encoder–Decoder Framework, Water Resources Research, 59, e2022WR033 091, https://doi.org/https://doi.org/10.1029/2022WR033091, 2023.

Denti, F., Doimo, D., Laio, A., and Mira, A.: The generalized ratios intrinsic dimension estimator, Scientific Reports, 12, 20 005, https://doi.org/10.1038/s41598-022-20991-1, 2022.

Edijanto, N. D. O. N., Yang, X., Makhlouf, Z., and Michel, C.: GR3J: a daily watershed model with three free parameters, Hydrological Sciences Journal, 44, 263–277, https://doi.org/10.1080/02626669909492221, 1999.

Facco, E., d'Errico, M., Rodriguez, A., and Laio, A.: Estimating the intrinsic dimension of datasets by a minimal neighborhood information, Scientific Reports, 7, 12 140, https://doi.org/10.1038/s41598-017-11873-y, 2017.

Fenicia, F., Kavetski, D., Reichert, P., and Albert, C.: Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Empirical Analysis of Fundamental Properties, Water Resources Research, 54, 3958–3987, https://doi.org/https://doi.org/10.1002/2017WR021616, 2018.

Frame, J., Kratzert, F., Gupta, H. V., Ullrich, P., and Nearing, G. S.: On Strictly Enforced Mass Conservation Constraints for Modeling the Rainfall-Runoff Process, Hydrological Processes, in review, 2022.

Gao, H., Hrachowitz, M., Schymanski, S. J., Fenicia, F., Sriwongsitanon, N., and Savenije, H. H. G.: Climate controls how ecosystems size the root zone storage capacity at catchment scale, Geophysical Research Letters, 41, 7916–7923, https://doi.org/https://doi.org/10.1002/2014GL061668, 2014.

Glielmo, A., Zeni, C., Cheng, B., Csányi, G., and Laio, A.: Ranking the information content of distance measures, PNAS Nexus, 1, pgac039, https://doi.org/10.1093/pnasnexus/pgac039, 2022.

Gnann, S. J., McMillan, H. K., Woods, R. A., and Howden, N. J. K.: Including Regional Knowledge Improves Baseflow Signature Predictions in Large Sample Hydrology, Water Resources Research, 57, e2020WR028 354, https://doi.org/https://doi.org/10.1029/2020WR028354, e2020WR028354 10.1029/2020WR028354, 2021.

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, https://doi.org/https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., and Fenicia, F.: Improving hydrologic models for predictions and process understanding using neural ODEs, Hydrology and Earth System Sciences, 26, 5085–5102, https://doi.org/10.5194/hess-26-5085-2022, 2022.

Jakeman, A. J. and Hornberger, G. M.: How much complexity is warranted in a rainfall-runoff model?, Water Resources Research, 29, 2637–2649, https://doi.org/https://doi.org/10.1029/93WR00877, 1993.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, edited by Bengio, Y. and LeCun, Y., http://arxiv.org/abs/1412.6980, 2015.

Kiraz, M., Coxon, G., and Wagener, T.: A Signature-Based Hydrologic Efficiency Metric for Model Calibration and Evaluation in Gauged and Ungauged Catchments, Water Resources Research, 59, e2023WR035 321, https://doi.org/https://doi.org/10.1029/2023WR035321, e2023WR035321 2023WR035321, 2023.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 26, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.

Kratzert, F.: CAMELS Extended NLDAS Forcing Data, https://doi.org/10.4211/hs.0a68bfd7ddf642a8be9041d60f40868c, 2019.

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, https://doi.org/https://doi.org/10.1029/2019WR026065, 2019.

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 25, 2685–2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, 2022.

McMillan, H.: Linking hydrologic signatures to hydrologic processes: A review, Hydrological Processes, 34, 1393–1409, https://doi.org/https://doi.org/10.1002/hyp.13632, 2020a.

McMillan, H.: A review of hydrologic signatures and their applications, Wiley Interdisciplinary Reviews: Water, https://doi.org/10.1002/wat2.1499, 2020b.

Mohammadi, B.: A review on the applications of machine learning for runoff modeling, Sustainable Water Resources Management, 7, 98, https://doi.org/10.1007/s40899-021-00584-y, 2021.

Molnar, C.: Interpretable Machine Learning, open source online book, 2 edn., https://christophm.github.io/interpretable-ml-book, 2024.

Molnar, C., Casalicchio, G., and Bischl, B.: Interpretable machine learning–a brief history, state-of-the-art and challenges, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 417–431, Springer, https://doi.org/10.1007/978-3-030-65965-3_28, 2020.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

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., and 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, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

Olden, J. D. and Poff, N. L.: Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, River Research and Applications, 19, 101–121, https://doi.org/https://doi.org/10.1002/rra.700, 2003.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, https://doi.org/https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, Hydrological Processes, 17, 455–476, https://doi.org/https://doi.org/10.1002/hyp.1135, 2003.

Wagener, T., Sivapalan, M., Troch, P., and Woods, R.: Catchment Classification and Hydrologic Similarity, Geography Compass, 1, 901–931, https://doi.org/https://doi.org/10.1111/j.1749-8198.2007.00039.x, 2007.

Zar, J. H.: Spearman Rank Correlation, John Wiley & Sons, Ltd, https://doi.org/https://doi.org/10.1002/0470011815.b2a15150, 2005.