https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Better data or better architecture? Improving deep-learning-based prediction in ungauged basins

Benedikt Heudorfer[1], Hoshin Gupta[2], Alexander Dolich[3], Ralf Loritz[3]

[1]Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research – Atmospheric Trace Gases and Remote Sensing, Karlsruhe, Germany
[2]Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ, USA
[3]Karlsruhe Institute of Technology (KIT), Institute for Water and Environment, Karlsruhe, Germany

Correspondence to: Benedikt Heudorfer (benedikt.heudorfer@kit.edu)

Abstract.

Large-sample hydrology has recently been driven by two key developments. First, the introduction of hydrological benchmark datasets such as CAMELS-US and CARAVAN, and second, the emergence of deep-learning modelling frameworks, particularly LSTM-based regional models, which have demonstrated performance on par with, and in some cases exceeding, that of process-based models for streamflow prediction in gauged and ungauged settings. Building on these developments, we investigate whether (i) further enhanced LSTM architectures, (ii) new sets of static features, or (iii) a combination of both enable us to significantly improve Predictions in Ungauged Basins (PUB). In this study, we evaluate a state-of-the-art regional LSTM model (base LSTM) against embedded (EMB-LSTM) and cross-attention enhanced (CA-LSTM) variants, in combination with a suite of newly applied static features, namely MODIS surface reflectance bands, ALPHAEARTH embeddings, DEM-, meteorology- and catchment coordinate-derived auxiliary aggregates, and conventional CAMELS attributes. We tested these model-and-data combinations in pseudo-ungauged 5-fold cross-validation across the 531 CAMELS-US catchments. Model performance was quantified by the Nash-Sutcliffe Efficiency (NSE), while latent-space complexity was assessed via the Shannon effective rank (erank). Results show that the quality of static features is more important than architectural improvements. ALPHAEARTH embeddings attained the highest median NSE, but only in combination with auxiliary static feature data ($\text{ALPHAEARTH}_{\text{plus}}$). Architectural refinements yielded only modest improvements. Thereby the relatively simple EMB-LSTM, which allowed the LSTM layer to better ingest $\text{ALPHAEAERTH}_{\text{plus}}$ static features, outperformed the other architectures. With this combination, we achieved a median performance of NSE=0.726, significantly improving the state-of-the-art PUB performance (NSE = 0.69) for the CAMELS-US dataset. Auxiliary analysis indicates that further improvement is possible when adding MODIS bands as additional dynamic features to the model. In conclusion, our study indicates that, broadly speaking, (a) better data is more important than better architecture, (b) better architecture is necessary only to accommodate better data, (c) the single layer LSTM remains the most suitable core model as of now, and (d) the Shannon effective rank complexity of the latent space is a useful diagnostic for linking improved PUB performance to improved quality of latent hydrological representation inside the model. Overall, this highlights the need for improved measurement-derived descriptor datasets, especially for soil and geology.

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

1 Introduction

In hydrology, one of the catalytic developments of the past 10 years is the advent of hydrological benchmark datasets. Kicked off by the CAMELS-US dataset as a follow-up dataset of the MOPEX dataset across the contiguous United States (Addor et al., 2017; Newman et al., 2015), numerous national datasets followed (e.g. Alvarez-Garreton et al., 2018; Coxon et al., 2020; Loritz et al., 2024), culminating in the worldwide CARAVAN dataset (Kratzert et al., 2023), integrating most of the national CAMELS datasets. These datasets triggered an unprecedented boom in large sample studies within and beyond hydrology (Kratzert et al., 2024), significantly shaping the community of large-sample hydrology (Gupta et al., 2014).

The existence of these datasets kicked off a second catalytic development in hydrology, namely the advent of deep-learning models. Two initial methodological developments made this happen. First, the new deep-learning models were set up as "regional" models (sometimes called "global" models) across larger sets of catchments (usually the entire dataset), enabling neural network architectures such as Long-Short Term Memory Networks (LSTM, Hochreiter & Schmidhuber, 1997) to synergistically construct representations of hydrological functioning from multiple samples in a meaningful and performative way (Kratzert et al., 2018). Second, the fusion of information about catchment characteristics (hereafter called static features) with information pertaining to meteorological drivers (hereafter called dynamic features) as model input data enabled the LSTM-based models to significantly outperform process-based hydrological models at the continental scale (Kratzert et al., 2019a). Although experiments with neural networks (see literature review provided in Kratzert et al., 2018; Mount et al., 2016) long pre-date those works, only now was a real breakthrough achieved, placing the new deep learning methods at the centre of attention of significant parts of the hydrological community.

These developments assigned to static features a critical role in neural network-based lumped hydrological modelling. The rational explanation to the results of Kratzert et al. (2019a) is that, by exploiting the information encoded by the static features (which characterize each entities' hydrological catchment functioning), the fusion of static and dynamic features allows the model to discern entities (catchments) within a larger population of entities in a physiographical meaningful way. Heudorfer et al. (2025) coined the term "entity aware" (EA) modelling for this general approach, following Ghosh et al. (2023), who introduced the term in the context of the wider deep learning community.

However, static features can, by definition, only be useful when they represent new information provided externally to the model. This is because the dynamical input-output data taken together *already* encapsule all of the necessary information required to make inferences about the input-output transformation function of a model – assuming of course that the model represents a complete input-output relation (Willems, 2007), which we assume to be the case for the deep learning models used today. Thus, static features are only useful for characterizing the input-output relationship (model) when they serve as a proxy for the information otherwise contained in the dynamic input-output data. This happens to be the case for "Predictions in Ungauged Basins" (PUB, Hrachowitz et al., 2013, Sivapalan et al., 2003) when this relation cannot be established due to the output part of the input-output data being unavailable, or when tasked with making spatial out-of-sample predictions. Note that this is also the explanation for why simple embeddings (Nolte et al., 2024; Yang et al., 2026; Yang & Chui, 2024) or even

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

random values (Heudorfer et al., 2024; Li et al., 2022) have been shown to be equally sensible alternatives for physiographically meaningful static features in the gauged prediction case (or spatial in-sample prediction), thereby enabling models to achieve state-of-the-art performance without any static feature "support points".

In this context, Heudorfer et al. (2025) showed that, despite their demonstrated predictive skill, static features are underutilized by LSTM models in ungauged prediction. By analysing the CAMELS catchment attributes typically used as static inputs, they found that most of these features can be replaced by simple feature-engineered variants derived from the model's dynamic inputs. This finding raises questions about the actual informational value of the CAMELS static features and motivates further work to better exploit available data for improved ungauged prediction. At the same time, recall that accurate ungauged prediction fundamentally depends on a robust hydrological representation that transforms inputs to outputs without relying on an entities' output data for conditioning, i.e. a model that generalizes well. Therefore, progress in this area would mark a significant step forward in enhancing the models' generalization capabilities.

Against this backdrop, we propose that generalization capabilities in DL models are limited either by a) data constraints or b) inadequate learning techniques and/or model architectures. These two pathways mirror the two catalytic events for deep-learning hydrology described at the beginning of this introduction (emergence of benchmark datasets, and static feature-informed regional LSTM models), and align with the two principal types of approaches possible in machine learning-based hydrologic studies, namely data-centric and model-centric approaches (Roscher et al., 2024). Coincidentally, the LSTM currently faces a lot of hot contenders for the task of streamflow prediction, especially with the advent of attention-based architectures (Liu et al., 2024, 2025). Likewise, promising new datasets (Brown et al., 2025) are now available that may be suitable extensions to, or replacements for, the conventional sets of static features. Consequently, to improve the performance of ungauged predictions via improved generalization capability, in this paper we address two research questions:

1) To what extent can (i) deeper architectures, (ii) improved static feature sets, or (iii) a combination of both enhance the models' ability to extract hydrologically relevant information and improve performance in ungauged settings?
2) To what extent do deeper architectures and improved static features shape latent representations of hydrological functioning, and what do these representations reveal about model performance and limitations in ungauged settings?

To answer these questions, we test three different LSTM-based model architectures in combination with a range of additional data products for static features (catchment attributes). All data used for this task is described in Chapter 2. Chapter 3 provides an overview of the experimental setup, as well as the models and explainability methodology used. Chapter 4 then presents the study results and their discussion, while Chapter 5 provides the concluding remarks of this study.

2 Data

In this chapter, we introduce all data used for our models. First of all, streamflow time series as the target (output) feature (Chapter 2.1). As input, we use a combination of meteorological forcing time series as dynamic features (Chapter 2.2), and environmental catchment descriptors as static features (Chapter 2.3). The former remain the same throughout all experiments,

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

whereat the focus of this paper are the latter, i.e. static features. Consequently, particular care is given in describing the selection logic (Chapter 2.3.1) and preprocessing (Chapter 2.3.2) applied to arrive at the specific sets of static features used for this task.

2.1 Target feature: streamflow

As target data we use streamflow from the CAMELS-US dataset (Addor et al., 2017). Specifically, following Newman et al. (2017), we use a sub-selection of 531 catchments from the 671 catchments originally included in the dataset. We use this subset because the studies we benchmark against (Heudorfer et al., 2025; Kratzert et al., 2021) used this selection as well. This allows comparability.

2.2 Dynamic features (meteorological forcing)

All of our experiments leverage the same set of dynamic input features, consisting of several meteorologic variables – namely three different sources of precipitation, solar radiation, minimum/maximum temperature, and vapor pressure from the Daymet (Thornton et al., 1997), Nldas (Xia et al., 2012) and Maurer (Maurer et al., 2002) datasets. These data accompany the CAMELS-US streamflow dataset and are extensively described in the CAMELS paper (Addor et al., 2017).

In total, this makes 15 dynamic meteorologic input features, which is seemingly redundant information. However, Kratzert et al. (2021) showed that the LSTM can reliably disregard redundancy and is instead able to leverage small additional information advantages between the three meteorological datasets, leading to significantly elevated performance when predicting CAMELS-US streamflow time series. Accordingly, we follow this rationale and use all 15 meteorological variables as dynamic features.

2.3 Static features (catchment attributes)

2.3.1 Rationale for dataset selection

In the experiments reported here, we compare model performance with different combinations of datasets used as static features. The selection of these combinations was guided by the question of which datasets (taken together) might represent richer data structures than provided by the original CAMELS-US static features. Table 1 juxtaposes the different subgroups of CAMELS static features with the suitable replacements we chose.

In the selection process, we operated under the premise that a dataset that is closer to real observations is one that is better exploitable by data-driven models. Accordingly, we assume that, for a given training task, these models are most effective when they can directly extract relevant information from minimally processed data. In other words, if feature engineering is important for the task, a sufficiently expressive model architecture may learn equivalent transformations through its internal representations. This means that, given advanced enough models, we assume that manually or semi-manually classified data, or feature-engineered variants of observation data, do not necessarily help the model to better generalize and produce good performance. On the contrary, data manipulation can potentially, and unwittingly, veil the actual information content and

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

thereby obstruct the ability of the model to construct meaningful semantic relationships between the input-output data. So, under this premise, we only provided our models with replacement datasets that are as close to observations as possible.

Table 1: List of CAMELS static features used in our base LSTM model, and the corresponding underlying datasets from which these features were derived from. Next to those are the direct measurement dataset we chose as replacement here, and the spatial aggregation we used to derive static features from the catchment maps of these datasets, which are quantiles taken from the long-term average feature maps. CAMELS static features are all derived as spatial averages. References: [1] Thornton et al. (1997), [2] Thornton et al. (2022), [3] Falcone (2011), [4] Jarvis et al. (2008), [5] Newman et al. (2015), [6] Vermote et al. (2002), [7] Brown et al. (2025), [8] Miller & White (1998), [9] Hartmann & Moosdorf (2012), [10] Gleeson et al. (2014).

static feature	underlying dataset in CAMELS	replacement chosen here	spatial aggregation chosen here
Precipitation mean PET mean Aridity index Precipitation seasonality Snow fraction High precipitation frequency High precipitation duration Low precipitation frequency Low precipitation duration	daymet[1]	daymet[2] prcp, srad, vp, tmin, tmax (labeled **METEO**)	temporal: q0,q10,q25,q50, q75,q90,q100 (precip: no q0-q25) spatial: average
Elevation Slope	GAGES-II[3]	**DEM**[4] elevation	spatial: q0,q10,q20,q30, q40,q50, q60,q70, q80,q90,q100
Area		GAGES-II[3]	
Forest fraction	USGS data[5]	**MODIS**[6] bands 1-7 **ALPHAEARTH**[7] bands 1-64	MODIS: q0,q10,q25, q50,q75,q90,q100 ALPHAEARTH: q50
LAI max LAI difference GVF max GVF difference	MODIS[6]		
Soil depth (Pelletier) Soil depth (STATSGO) Soil Porosity Soil conductivity Max water content Sand fraction Silt fraction Clay fraction	STATSGO[8]	---	---
Carbonate rocks fraction	GLiM[9]	---	---
Geological permeability	GLHYMPS[10]		
Coordinates	---	WGS84 (labeled **XY**)	min/max

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

This reasoning confronts us with the problem that not all datasets of environmental attributes are purely observation based. Specifically, the soil and geology datasets used in the CAMELS dataset (see Table 1) are essentially manual classifications (e.g. Gleeson et al., 2014; Hartmann & Moosdorf, 2012; Miller & White, 1998). Broadly speaking, the classification process is the following. As a basis, it does to some degree involve local (point or cross-section) measurements of physical or chemical soil and geologic properties, as well as interpretations, static and dynamic metadata etc. But this data then only indirectly serves experts as a foundation to their judgement of soil and geology classes, which they derive qualitatively from whatever data is available for a given landscape unit (which may be plenty or scarce, depending on the location). So, while the resulting datasets have tremendous value for a vast array of decisions, we argue that they are more suited for evaluation via human cognition, while data-driven models better rely on *actual* data (i.e. measurements) instead of qualitative assessments. This perspective is supported by the findings of previous studies, in which soil and geology attributes rank notably low in feature importance analyses of data-driven models (e.g. Kratzert et al., 2019b). Consequently, we chose not to include any soil and geology datasets, there being essentially no observation-based datasets available for these domains.

Regarding vegetative attributes, finding suitable replacement data was straightforward, as the basis for all CAMELS vegetative attributes is MODIS (Vermote et al., 2002), which is readily available via the Google Earth Engine (Gorelick et al., 2017). We used the raw MODIS surface reflectance bands directly, instead of using variables such as the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR, Monteith, 1972), Normalized Difference Vegetation Index (NDVI, Rouse, 1973), or Leaf Area Index (LAI, Watson, 1947) that are derived from those bands and are also readily available . This is because the derived variables must typically contain less information than the raw data, since they are feature-engineered re-combinations of raw surface reflectance bands, constructed with the aim of assigning semantic meaning to the data, achieved by processing data through physical, rule-based formulas that can be considered lossy compressions. As lossy compressions, their use can be considered detrimental to data-driven modeling (see arguments above). We still tested this in pre-experiments (not shown) by comparing model performance when adding raw surface reflectance bands vs. adding feature engineered re-combinations of these bands (e.g. FAPAR, NDVI, LAI) and found that using raw MODIS bands resulted in significantly improved performance, compared to no improvement in performance when engineered features were used. Thus, we rely on raw MODIS surface reflectance bands.

On the other hand, ALPHAEARTH also represents a feature engineered data product readily available via Google Earth Engine. However, it merges multiple reflectance products such as Landsat (Wulder et al., 2019) and Sentinel (Spoto et al., 2012), as well as auxiliary data such as geo-located text, via an Autoencoder that is essentially used as a compression tool. The final product consists of the latent embeddings of the Autoencoder, exported and published as a globally available feature set (Brown et al., 2025). Thus, it can be considered to be a purely data based product derived from a data-driven compression tool. Furthermore, it can be assumed to contain more rather than less information, even if autoencoding is considered lossy compression as well, because multiple input data sources are ingested, expanding the informational coverage rather than limiting it. Consequently, recent studies have demonstrated its considerable value (Qu et al., 2026), and we decided to test the ALPHAEARTH embeddings product alongside MODIS as a second satellite-based dataset.

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

For the meteorology-based static features, we used the same database as in the CAMELS dataset, namely the Daymet dataset of meteorological input features, but ingested as a gridded product (Thornton et al., 2022). Choosing a replacement dataset for the topographic static features was also straightforward: the basis in CAMELS was a DEM (Newman et al., 2015), and we used a DEM here as well, namely the SRTM DEM (Jarvis et al., 2008). Lastly, we added the catchments' coordinates as additional static features, because they have been shown to add some value when used as static input features (e.g. in groundwater, Ohmer et al., 2025), but were not included in previous deep learning-based streamflow prediction studies.

2.3.2 Preprocessing

All static feature datasets used in this study (METEO, DEM, MODIS, ALPHAEARTH, XY, see Table 1 and Chapter 2.3.1) were preprocessed in the following manner. First, all datasets were downloaded from Google Earth Engine (Gorelick et al., 2017) as raster data. These raster were subsequently resampled to a spatial resolution of 500m. This resolution was chosen based on preliminary experiments comparing 500 and 90m resolutions, which found no performance improvement at higher resolution, but a substantial increase in computational cost. Following resampling, the rasters were clipped to the respective catchment extents, resulting in gridded spatial feature maps for each catchment and dataset channel. For the METEO, MODIS and ALPHAEARTH datasets, these maps constitute spatial feature map *time series*.

From these feature maps, we computed quantiles of the spatiotemporal value distributions. Quantile-based summarization was chosen as the simplest approach to preserve distributional information in spatial aggregation, providing greater informational richness compared to the spatial averaging used for CAMELS static features (simple averaging). We extensively tested more advanced compression methods, including Principal Component compression and Masked Variational Autoencoder-based latent feature extraction (following Ehret et al., 2025). However, these methods did not yield performance improvements over the quantile-based approach and were therefore not adopted.

The quantiling scheme as well as the number of quantiles used for resampling varied depending on the dataset (see Table 1). For spatial coordinates (XY), only the minimum and maximum values were derived. For the DEM, we chose a dense range of quantiles (q0, q10, q20, q30, q40, q50, q60, q70, q80, q90, q100) for spatial resampling, because we assume high information content of elevation data, and because it is only one channel, allowing finer resampling. Catchment area was additionally included as a feature.

The MODIS and ALPHAEARTH datasets were first temporally aggregated by taking the long-term average per raster cell over the full record period until December 2025. Subsequently, for MODIS we chose a moderately dense quantile range (q0, q10, q25, q50, q75, q90, q100) for spatial resampling, because it is a multi-channel feature map (7 channels) and we did not want to unnecessarily inflate the static feature set. For ALPHAEARTH, we took only the median (q50) because of its excessive number of channels (64), and the fact that these channels already represent compressed embeddings.

Note that the record period of MODIS (24.02.2002 – 31.12.2025) differs from ALPHAEARTH (01.01.2017 – 31.12.2025) as well as from the CAMELS availability period of streamflow (01.01.1980 – 31.12.2014). However, since we use long-term

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

averages as basis for computing static features that are only indicative of time-invariant overall catchment properties, we argue the non-overlap is acceptable.

For the METEO dataset, we reversed the aggregation order: quantiles (q0, q10, q25, q50, q75, q90, q100) were derived first as per-channel, long-term temporal aggregates, and then averaged spatially. We did this because Heudorfer et al. (2025) demonstrated the value of using multiple temporally aggregates (mean and standard deviation) for static feature derivation from meteorological time series. Finally, we had to omit the q0, q10 and q25 quantiles for precipitation, because those were all-zero which broke model training when included.

The resulting number of static features was 4 for XY (min/max for lon/lat), 11 for DEM (10 quantiles for elevation plus catchment area), 32 for METEO (7 quantiles for srap, vp, tmin, tmax and 4 quantiles for prcp), 49 for MODIS (7 quantiles for 7 surface reflectance bands) and 64 for ALPHAEARTH (median values for 64 embedding channels).

3 Methods

3.1 Experimental setup

The paper encompasses two *modeling* experiments and two *explainability* applications, whose results are presented in Chapter 4.1 and Chapter 4.2, respectively. Since we test a lot of different combinations of static feature data and model architecture, an overview which data-architecture combination is used in which figure or table is given in Appendix A.

The first modeling experiment is a stepwise addition experiment, wherein different static feature datasets are added to the base LSTM model. The experiment begins with the MODIS and ALPHAEARTH datasets, to which auxiliary datasets are added (DEM, METEO, XY, see Chapter 2.3.1) in a stepwise manner. The order of addition was guided by the absolute performance of the base LSTM model when using the individual auxiliary datasets separately (Appendix B). Note that the dynamic features always remain the same throughout the paper, and only the static features are changed.

The second modeling experiment compares model architectures (Figure 1) having different levels of data fusion complexity. For these, only the two best-performing dataset combinations (one based on the MODIS dataset, the other on the ALPHAEARTH dataset) from the first experiment are used when running the different models. The baseline model is the base LSTM model with simple concatenation-style data fusion of static and dynamic features (Chapter 3.3.1), against which we compare a mid-complexity architecture with simple embeddings preceding the LSTM layer (EMB-LSTM, Chapter 3.3.2), and a high-complexity architecture with an additional cross-attention layer preceding the LSTM layer (CA-LSTM, Chapter 3.3.3). The best-performing model from this second experiment integrates performance advantages from both data and model improvements.

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

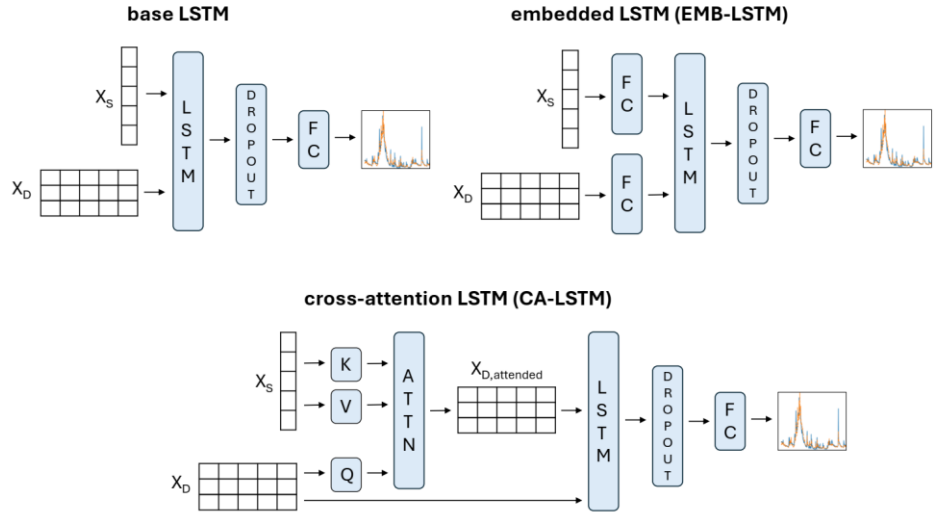

Figure 1: Overview of the streamflow prediction models used in this study. X_S denotes static features, X_D dynamic features. LSTM stands for Long Short-Term Memory layers, FC for fully connected layers, ATTN for attention operation on the key (K), value (V) and query (Q) matrices.

In the subsequent explainability section (Chapter 4.2), the overall best-performing models are compared in terms of the Shannon effective rank (erank, further explained in Chapter 3.4) applied to the latent activations of the cell state and hidden state of the LSTM layer of the EMB-LSTM. Here, erank serves as a measure of the compressibility of the models' latent space, which has the unit of number of ranks or dimensions. It can be interpreted as the overall number of LSTM cells that are processing significant amounts of information during inference, i.e. the number of cells in a trained model that are effectively active. Therefore, depending on how compressible the latent space of the LSTM is, we can draw conclusions regarding the complexity of the hydrologic representation inside the model. We interpret the erank measure for itself, as well as in correlation to conventional static features.

3.2 Miscellaneous experimental details

All models are 1D (input) to 1D (output) dynamical systems models, where the inputs are daily meteorological time series data spatially aggregated over the catchment, fused with static feature values repeated at every time step. The output is a daily streamflow time series at the corresponding gauge location. Models are trained using data from all 531 catchments simultaneously, and are set up as sequence-to-one, i.e. they infer streamflow at day t from the input features over the previous 365 days (with sequence length 365). To allow comparability with previous studies, we used the same train / val / test periods as in Heudorfer et al. (2025) and Kratzert et al. (2021). The training period was 01.10.1999 – 30.09.2008, the validation period was 01.10.1980 – 30.09.1989, and the testing period was 01.10.1989 – 30.10.1999 for all experiments in this study, except one supplementary experiment (see Appendix C).

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

All results shown in this study represent performance in a pseudo-ungauged prediction setup. That means the models were run temporal as well as spatial out-of-sample, practically implemented as a 5-fold cross validation, where in each fold, training happened on the train period of 80% (N) of the 531 catchments, and prediction on the test period of the other 20% (N) of catchments.

The loss function was the basin-averaged Nash-Sutcliffe Efficiency (NSE, Nash & Sutcliffe, 1970) defined in Kratzert et al. (2019b). The NSE was also used as a metric to evaluate performance in a bootstrapping procedure as defined by Heudorfer et al. (2025) to account for uncertainty due to random sampling variability. For this, the model was first run with 5 seed initializations in general, but 8 seeds for all models in Table 3 (to support higher confidence), and each catchments' predicted test period from all seed realizations were bagged into a single sample bag. From this bag, we drew an 80% sample 100 times with replacement, and calculated the NSE for each sample, resulting in 100 metric realizations. The final performance measure (NSE_{median}) and uncertainty bands ($U_{5\text{-}95}$) used in this paper are the median and the average Q5-Q95 range of these 100 metric realizations. This more rigorous metric calculation leads to somewhat diminished NSE values compared to the benchmark we compare our results against Kratzert et al. (2021), as was discussed in Heudorfer et al. (2025). Finally, to interrogate the significance of model differences, we compare the score distributions between models using the two-sided Kolmogorov-Smirnov test (KS) assuming a 5% significance level.

3.3 Models

3.3.1 Base LSTM

All models used in this study are based on the LSTM as the central layer. Descriptions of the LSTM has been re-iterated many times in hydrology by now. Instead of reformulating it, we provide here a direct quotation from the description provided in Kratzert et al. (2019b), which in our opinion represents the most concise and clear description of the LSTM in a hydrological paper.

> 'An LSTM works as follows: given an input sequence $x = [x[1], \dots, x[T]]$ with T time steps, where each element $x[t]$ is a vector containing input features (model inputs) at time step $t (1 \leq t \leq T)$, the following equations describe the forward pass through the LSTM:
>
> $$i[t] = \sigma(W_i x[t] + U_i h[t-1] + b_i) \quad (1)$$
> $$f[t] = \sigma(W_f x[t] + U_f h[t-1] + b_f) \quad (2)$$
> $$g[t] = tanh(W_g x[t] + U_g h[t-1] + b_g) \quad (3)$$
> $$o[t] = \sigma(W_o x[t] + U_o h[t-1] + b_o) \quad (4)$$
> $$c[t] = f[t] \odot c[t-1] + i[t] \odot g[t] \quad (5)$$
> $$h[t] = o[t] \odot \tanh(c[t]) \quad (6)$$
>
> where $i[t]$, $f[t]$, and $o[t]$ are the *input gate*, *forget gate*, and *output gate*, respectively, $g[t]$ is the *cell input* and $x[t]$ is the *network input* at time step $t (1 \leq t \leq T)$, and $h[t-1]$ is the *recurrent input*, $c[t-1]$ the *cell*

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

state from the previous time step. At the first time step, the hidden and cell states are initialized as a vector of zeros. $\boldsymbol{W}$, $\boldsymbol{U}$, and $\boldsymbol{b}$ are learnable parameters for each gate, where subscripts indicate which gate the particular weight matrix/vector is used for, $\sigma(\cdot)$ is the sigmoid function, $\tanh(\cdot)$ is the hyperbolic tangent function, and $\odot$ is element-wise multiplication. The intuition behind this network is that the cell states ($c[t]$) characterize the memory of the system. The cell states can get modified by the forget gate ($f[t]$), which can delete states, and the input gate ($i[t]$) and cell update ($g[t]$), which can add new information. In the latter case, the cell update is seen as the information that is added and the input gate controls into which cells new information is added. Finally, the output gate ($o[t]$) controls which information, stored in the cell states, is outputted.'

The base LSTM as used here (Figure 1) is a replica of the model used in Kratzert et al. (2021) as implemented by Acuña Espinoza et al. (2025) based on the neural hydrology repository (Kratzert et al., 2022). In this model, static features are simply repeated at every time step and concatenated with the dynamic features before funneling them into one single LSTM layer with 256 hidden states, followed by a 40% dropout layer and a linear output layer. The forget gate bias was set to 3 and the batch size was 256. The initial learning rate of the Adam optimizer was 0.001, and the learning rate was adapted to 0.0005 at epoch 10 and 0.0001 at epoch 20. The total number of epochs was 30, and the last epochs' model state was used for prediction.

3.3.2 Embedded LSTM (EMB-LSTM)

The embedding-enhanced LSTM model complements the base LSTM with an anterior dense layer that receives the static and dynamic features separately (Figure 1). Thus, the concatenation of static and dynamic features in the base LSTM model is replaced by the full neural connection between the embeddings and the LSTM layer. The dynamic feature embedding has a fixed size of 15, corresponding to the number of dynamic input features used in this study (see Chapter 2.2). The static feature embedding has a fixed size of 50 neurons. That means for smaller sets of static features (<50), the embedding gives the model more dimensions of freedom for combinatory input of static features into the LSTM layer, while for larger sets of static features (>50), it serves to reduce the dimensionality of the static features. All other hyperparameters are the same as in the base LSTM model.

3.3.3 Cross-attention LSTM (CA-LSTM)

In this architectural variant, the base LSTM model is complemented with an anterior cross-attention module module that makes use of the attention mechanism taken as defined for the Transformer architecture (Vaswani et al., 2017). Thereby, the dynamic features are cross-attended by the static features. This mechanism replaces the concatenation of static and dynamic features in the base LSTM (Figure 1). Subsequently, the cross-attended dynamic features are funneled into the LSTM layer parallel to the "raw" dynamic features in order to provide a skip connection. To come up with this cross-attention LSTM architecture, we tested many different variants, but this was the best-performing one. To keep the scope of this paper concise, these preliminary experiments are not shown, since they represented failures mostly.

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Formally, the following equations describe the forward pass through the multi-headed cross-attention module:

$$Q = W_Q x_d[t] + b_Q \quad (7)$$

$$K = W_K x_s[t] + b_K \quad (8)$$

$$V = W_V x_s[t] + b_V \quad (9)$$

$$Q_i = W_i^Q Q \quad (10)$$

$$K_i = W_i^K K \quad (11)$$

$$V_i = W_i^V V \quad (12)$$

$$a_i = softmax(Q_i K_i^T / \sqrt{d}) \quad (13)$$

$$c_i = a_i V_i \quad (14)$$

$$\tilde{X}_d = concat(a_1, \dots, a_n) W_o \quad (15)$$

Here, dynamic inputs $x_d[t]$ and static inputs $x_s[t]$ are first turned into linear projections named query (Q), key (K) and value (V) via learnable parameters W and b to establish uniform Q, K, V dimensions d (also called attention size). Then, for multiple heads $i = 1, \dots, n$ the attention weights a_i are calculated by taking the $softmax$ function over Q_i and K_i in formula (13). Through subsequent multiplication of a_i with V_i we arrive at contexts c_i, which, concatenated across all heads n, form the dynamic representation that has been informed by or modulated with static context. Thereby, we chose attention size $d = 64$ and number of heads $n = 8$. All multiplications in formulas (7)-(15) are matrix multiplications.

3.4 Shannon effective rank

We use the Shannon effective rank ($erank$) as a tool to analyze the internal latent space complexity of the LSTM layer in order to draw conclusions regarding the complexity of the hydrological representation inside the model. The Shannon effective rank was introduced by Roy & Vetterli (2007) and is defined as:

$$erank(M) = exp\big(H(M)\big) \quad (16)$$

which is the effectively the Shannon entropy $H(M)$ for a matrix M with $m * n$ dimensions, but unit-converted to *number of ranks* (dimensions) by taking its exponential. An important property of the $erank$ measure is that it is equal to the size m of matrix M at maximum entropy. For a LSTM layer size of 256, as used in this study, this translates to $erank_{max}$=256.

Finally, for the total amount r of positive (> 0) singular values σ of M, the entropy H is defined as

$$H(M) = -\sum_{i=1}^{r} p_i * log\, p_i \quad (17)$$

over the singular value probability distribution p_i:

$$p_i = {\sigma_i}/{\sum_j \sigma_j} \quad (18)$$

Accordingly, the Shannon effective rank is a measure of the spread of the distribution of singular values of a matrix. Thus, it approximates the information complexity of a matrix as the number of linearly independent dimensions it possesses. To arrive at its interpretability as the number of important dimensions, it effectively ignores the value of the weights of singular values

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

as being indicative measures of importance of these dimensions, and treats them as if they were equally important. Applied to neural network activations, the Shannon effective rank expresses the complexity of the learned representational or latent space, where higher $erank$ indicates a higher complexity latent space (i.e. more information-processing pathways).

Here, we used Shannon effective rank to measure the complexity of the latent space of the core LSTM layers of our models by calculating $erank$ from the cell state and hidden state activations. To do this, we forward-passed the test data through the model, and extracted the last cell state and hidden state activations of each sequence for all time steps in the test period. We did this for each of the 531 catchments in our dataset, arriving at 531 $erank$ values for cell state and hidden state, i.e. one for each catchment. Calculating the $erank$ of these activations then provides a measure of the approximate number of information-processing pathways needed for each catchment during inference. We then analyze the $erank$ distributions of different models, to gauge the representational complexity that a model is able to achieve when predicting across the entire dataset.

4 Results & Discussion

4.1 Improving ungauged performance

4.1.1 Experiment 1: better data

Table 2 shows the results of the stepwise static feature data addition experiment, with the aim to outperform the benchmark model provided with CAMELS static features. Model performance is expressed as NSE_{median} for different dataset combinations based on the MODIS and ALPHAEARTH data. The results indicate that:

- At first, the model provided with the MODIS static features (NSE_{median}=0.6668) underperforms the one provided with CAMELS static features (NSE_{median}=0.696).
- When progressively provided with additional static features (DEM+METEO+XY, hereafter called $MODIS_{plus}$), the model eventually performs on par with the CAMELS-based model, achieving NSE_{median}=0.694.
- When instead provided with ALPHAEARTH static features, the model significantly outperforms the CAMELS-based model right away (NSE_{median}=0.705), however without significance (KS=0.057, p=0.364).
- Finally, when provided with additional sets of static features (DEM+METEO+XY, hereafter called $ALPHAEARTH_{plus}$), the model is able to further improve significantly (KS=0.081, P=0.061) to NSE_{median}=0.718.

At first sight, these results seem to indicate clear superiority of ALPHEAEARTH as a static feature set, compared to conventional static features, or to "raw" surface reflectance data (MODIS). They are also in line with previous studies that reported large performance improvements when incorporating ALPHAEARTH into streamflow prediction models, however on a degraded performance level that is not state of the art (Qu et al., 2026, which report NSE=0.612 with ALPHAEARTH, and 0.553 with CAMELS static features). However, when examined at the 95% confidence level (KS=0.094, p=0.018) the $ALPHAEARTH_{plus}$ model is, strictly speaking, *not significantly better* than the $MODIS_{plus}$ model. So the advantage is not as large as it may appear to be, and not as groundbreaking as previous studies have suggested (Qu et al., 2026) when taking into

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

account the uncertainty of the model, which, although seemingly quite low in absolute terms, proves to be decisive in comparison to the relatively small overall performance improvements we observe. This result indicates, therefore, that the ALPHAEARTH embeddings are not perfect. Although good, and representing a significant advancement, there is further value to be found and researchers should not abandon the quest for additional sources of information simply because the ALPHAEARTH dataset has been published. One suggestion would be that ALPHAEARTH should in the future expand its inclusion of a larger observational database (beyond mainly Landsat and Sentinel datasets, e.g. MODIS) into their embedding framework. This would likely provide considerable added value to the product.

Table 2: Ungauged performance of the base LSTM model with different dataset combinations expressed as NSE calculated across the test period next to uncertainty (Unc). The ablated model ("None"), which is not provided with static features, and the model provided with conventional 27 catchment attributes as static features ("CAMELS") are taken as benchmarks. The static features were then replaced with new data based on the MODIS and ALPHAEARTH datasets, and stepwise addition of additional data (DEM, METEO, XY, see Chapter 2.3). The order of the stepwise addition was guided by their performative value based on Appendix B. Bold font indicates significant improvement over the CAMELS base LSTM model with respect to a 90% (*) and a 95% (**) confidence level.

Static feature data	NSE_{median}	Unc_{5-95}
None (dynamic only)	0.634	±0.073
CAMELS	0.696	±0.054
MODIS	0.668	±0.071
MODIS+DEM	0.690	±0.055
MODIS+DEM+METEO	0.690	±0.056
MODIS+DEM+METEO+XY (=$MODIS_{plus}$)	0.694	±0.053
ALPHAEARTH	0.705	±0.055
ALPHAEARTH+DEM	**0.717***	±0.055
ALPHAEARTH+DEM+METEO	**0.715***	±0.050
ALPHAEARTH+DEM+METEO+XY (=$ALPHAEARTH_{plus}$)	**0.718***	±0.048

4.1.2 Experiment 2: better architectures

Table 3 shows the results of the improved data fusion experiment. The starting point is the two dataset combinations $MODIS_{plus}$ and $ALPHAEARTH_{plus}$. Based on those, the base LSTM architecture was compared to two new architectures: the cross-attention LSTM (CA-LSTM) and the embedded LSTM (EMB-LSTM). The results indicate that:

- The first model, CA-LSTM, shows some insignificant (KS=0.042, p=0.752) improvement when used in combination with the CAMELS (NSE_{median}: 0.696→0.703). In combination with the $MODIS_{plus}$ dataset the performance improvement is substantial (NSE_{median}: 0.696→0.715), however also surprisingly insignificant (KS=0.07, p=0.151).

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Further, when used in combination with the ALPHAEARTH$_{plus}$ dataset, CA-LSTM is not able to improve performance at all, instead slightly deteriorates it.

- The second model, EMB-LSTM, does not work well with the CAMELS dataset, but when used with the MODIS$_{plus}$ dataset, we see some performance improvement, however smaller than when using the CA-LSTM with this dataset. Used in combination with the ALPHAEARTH$_{plus}$ model, the EMB-LSTM improves performance insignificantly from NSE$_{median}$=0.718→0.721.
- However, the EMB-LSTM + ALPHAEARTH$_{plus}$ combination is, strictly speaking, the best-performing one, with a total absolute improvement over the initial (benchmark) model – the base LSTM with CAMELS static features – of 0.026, which is both a respectable improvement for ungauged prediction, and statistically significant (KS=0.079, p=0.072).

Table 3: Ungauged performance of the two winning datasets of the stepwise experiment (see Table 2) used in three different LSTM architectures: base LSTM with direct fusion of static features to the LSTM, cross-attention LSTM (CA-LSTM) where static features cross-attend dynamic feature prior to fusion to the LSTM, and an LSTM where features are fed to an embedding before fusing to the LSTM (EMB-LSTM). See Chapter 3.3 and Figure 1 for more information about these architectures. Performance reported in terms of NSE, next to uncertainty (Unc). Bold font indicates significant improvement over the CAMELS base LSTM model with respect to a 90% (*) and a 95% (**) confidence level.

	base LSTM		CA-LSTM		EMB-LSTM	
Static feature data	(NSE$_{median}$)	Unc$_{5-95}$	(NSE$_{median}$)	Unc$_{5-95}$	(NSE$_{median}$)	Unc$_{5-95}$
CAMELS	0.696	±0.054	0.703	±0.047	0.687	±0.045
MODIS$_{plus}$	0.694	±0.053	0.715	±0.046	0.709	±0.043
ALPHAEARTH$_{plus}$	**0.718***	±0.048	**0.714****	±0.042	**0.721***	±0.037

First of all, the insignificance in the 0.019 improvement in NSE of the CA-LSTM + MODIS$_{plus}$ combination over the benchmark (base LSTM with CAMELS) model is surprising. However, this but can likely be attributed to different distribution shapes, as the KS test does not solely compare medians, but the entire cumulative distribution function, i.e. also its shape. Despite the attested insignificant, we deem the CA-LSTM + MODIS$_{plus}$ combination on par with the the CA-LSTM + ALPHAEARTH$_{plus}$ combination, since they are equally performative.

Finally, interpreting the results from Tables 2 and 3 together, we conclude that while increasing the capability of the model – here, the CA-LSTM – can help to improve ungauged prediction, the higher complexity of the model does not pay off significantly when compared to simpler model setups (EMB-LSTM). This is in line with Liesch & Ohmer (2025), who tested different architectures fusing static and dynamic features in multiple ways in groundwater prediction, but found no significant advantage of attention-based architectures (or others) over the standard concatenation-style data fusion LSTM going back to (Kratzert et al., 2019a, 2019b). This is further in line with Liu et al. (2024, 2025), who showed that the LSTM consistently outperforms various Transformer-based architectures in streamflow prediction.

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

4.1.3 Synopsis: better data or better architecture?

In our study, the more important knob to adjust arguably proved to be the makeup of the input data (here, ALPHAEARTH$_{plus}$). Regarding architectural changes, the embeddings of the EMB-LSTM do not perform well with lower-complexity data like the CAMELS static feature set. Conversely, the CA-LSTM mostly helps with lower-complexity data like CAMELS and MODIS, but not with higher-complexity data like ALPHAEARTH$_{plus}$. Since CA-LSTM and EMB-LSTM perform on par with the ALPHAEARTH$_{plus}$ dataset, we conclude that architectural changes improve performance only insofar they helped the otherwise capable core model (the LSTM layer) to absorb lower-complexity data. Thereby, the attention mechanism did not prove to be significantly better than the simpler embedding-based architecture. This is in line with (Liu et al., 2024, 2025), who show that Transformers have not (yet) proved to be hugely beneficial. The LSTM is (so far) still one of the most capable architectures, having the best-matching inductive bias due to being a Markovian representation with dissipative states. We believe that this might change down the road, if/when we challenge our models with larger amounts and more complex data, or when applications other than catchment-based modeling are pursued, for example with other prediction tasks (Liu et al., 2025), in the emerging field of deep learning-based distributed (Eddin et al., 2025; Kraft et al., 2025) or semi-distributed (Gauch et al., 2025) hydrological modeling.

4.2 Investigating internal model functioning

4.2.1 Latent space complexity as indication of hydrological representation

Figure 2 shows the distributions of $erank$ values across all basins, providing an overview of the complexity of the latent activation space in the LSTM layer of the EMB-LSTM model. Overall, the range of erank is 108 to 165 latent dimensions (relative to n_{max}=256, dictated by the size of the LSTM layer), which translates to a latent space compression rate of 42-65%. This suggests that (overall) only ~42-65% of the LSTM dimensions actually process a significant amount of information. The lower end of this (42%) is represented by the ablated model without static features ("*None*" in Figure 2), which is given access only to dynamic meteorological features.

Compared to the ablated model, all models *with* access to static feature information show lower complexity of the cell state and hidden state activation spaces. For both cell state and hidden state, the erank for ALPHAEARTH$_{plus}$ and MODIS$_{plus}$ are similar, consistent with the statistical insignificance in performance advantage of ALPHAEARTH$_{plus}$ over MODIS$_{plus}$ (Chapter 4.1, Table 3). In terms of cell state activations, both models (erank=114) have slightly increased latent complexity compared to the baseline CAMELS model (erank=108), but slightly decreased complexity in terms of hidden state activations for (ALPHAEARTH$_{plus}$-erank=125 and MODIS$_{plus}$-erank=128 compared to CAMELS-erank=133).

This result has an interesting interpretation. Recalling LSTM model theory, we know that the LSTM mediates input-output dynamics via the cell state, which is able to store information in a manner analogous to state variables in conceptual hydrological models. On the other hand, hidden states are the multiplication of the cell state (modulated by a hyperbolic tangent) with the input and the previous hidden state (modulated by weights and biases, see Chapter 3.3.1). Accordingly, the

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

hidden state represents a *skip-connection-mediated cell state*. Putting this together, our finding of a higher *erank* number in cell states (i.e. higher complexity, more information processed via state space) coinciding with lower *erank* number in hidden states (i.e. lower complexity, less information processed via skip connections) can be interpreted as the model relying more on state-variable-related information processing, and less on the LSTMs' internal skip connection.

Our interpretation is that the models with improved PUB performance are able to more elaborately represent the overall hydrological system via the cell state, which thus reflects better memorization of the hydrological functioning of the system, allowing the model to ease reliance on the skip connections. In other words, the model better knows the systems' hydrological functioning, having learned a higher-quality hydrological representation. This allows us to robustly conclude that, when improving the static feature data base and the models' ability to absorb this data, the improved performance (Chapter 4.1) is attributable to improved hydrologic system representation by the model. This highlights that gains in performance are closely linked to how efficiently the model organizes information within its latent state space.

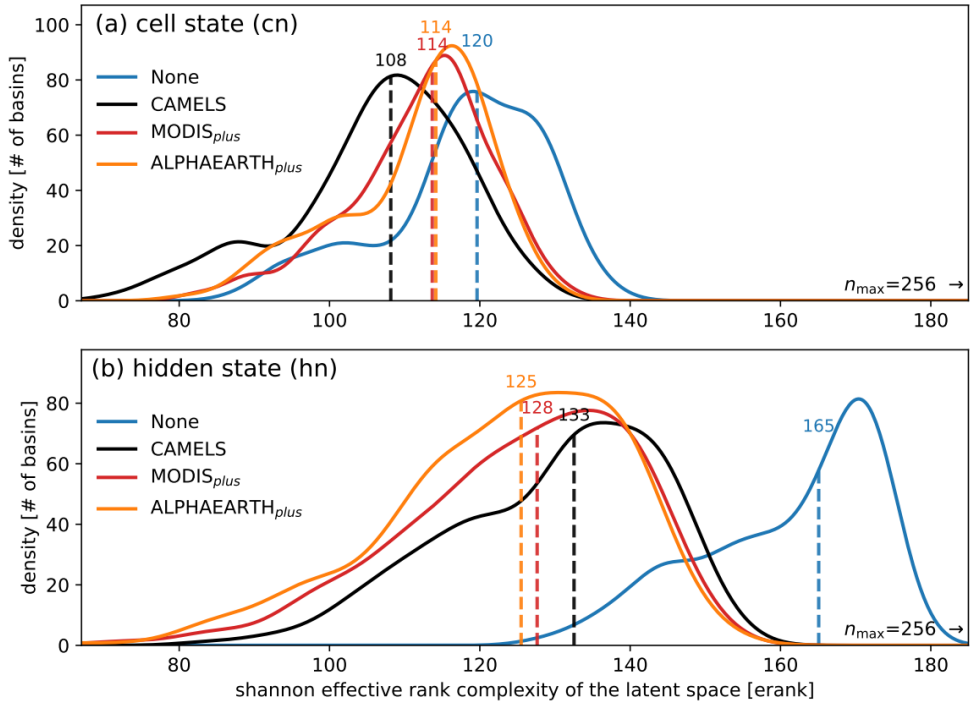

Figure 2: Density plot showing distributions of Shannon effective rank complexity of the latent space (erank), where (a) shows the erank of the cell state activations of the LSTM layer, and (b) the erank of the hidden state activations. Higher eranks indicate higher complexity latent space, i.e. more information-processing pathways. The different models are all run with the best-performing (see Table 3) EMB-LSTM architectures, but with different datasets used as static features: CAMELS uses the 27 CAMELS static features, MODIS$_{plus}$ and ALPHAEARTH$_{plus}$ use the best-performing MODIS and ALPHAEARTH-based dataset combination (see Table 3) and "None" uses no static features (ablated baseline). The maximum number of erank is n_{max}=256, which is the size of the LSTM layer. The MODIS$_{plus}$ and ALPHAEARTH$_{plus}$ erank distributions are significantly similar to each other (p=0.198 for hn, p=0.451 for cn), and significantly different from the CAMELS model (p<0.00008 for all).

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

4.2.2 Interpreting latent space complexity with catchment attributes

To analyze the information overlap between the main sets of static features used in this study, Figure 3 reports the correlation between the erank of the CAMELS, $MODIS_{plus}$ and $ALPHAEARTH_{plus}$ latent spaces (all run with the EMB-LSTM model) with the CAMELS static features. Overall, all models show the same overall correlation pattern. There are only small differences between the models, and differences are only significant where correlation is low. For example, the overall highest maximum absolute change in correlation strength was 0.24 (gvf_diff from r=0.08 down to -0.16). But for higher correlations (e.g. defined arbitrarily as r>0.5), the maximum change is only 0.1 (low_prec_dur from -0.51 to -0.61). This indicates that the additional data available to the $MODIS_{plus}$ and $ALPHAEARTH_{plus}$ models encompass the same information. It corroborates that the performance improvements observed in Tables 2 and 3 are due to higher information density of the data, which suggests indirectly that the CAMELS static features are not wrong per se, but simply represent a lesser degree of compression of the same information content.

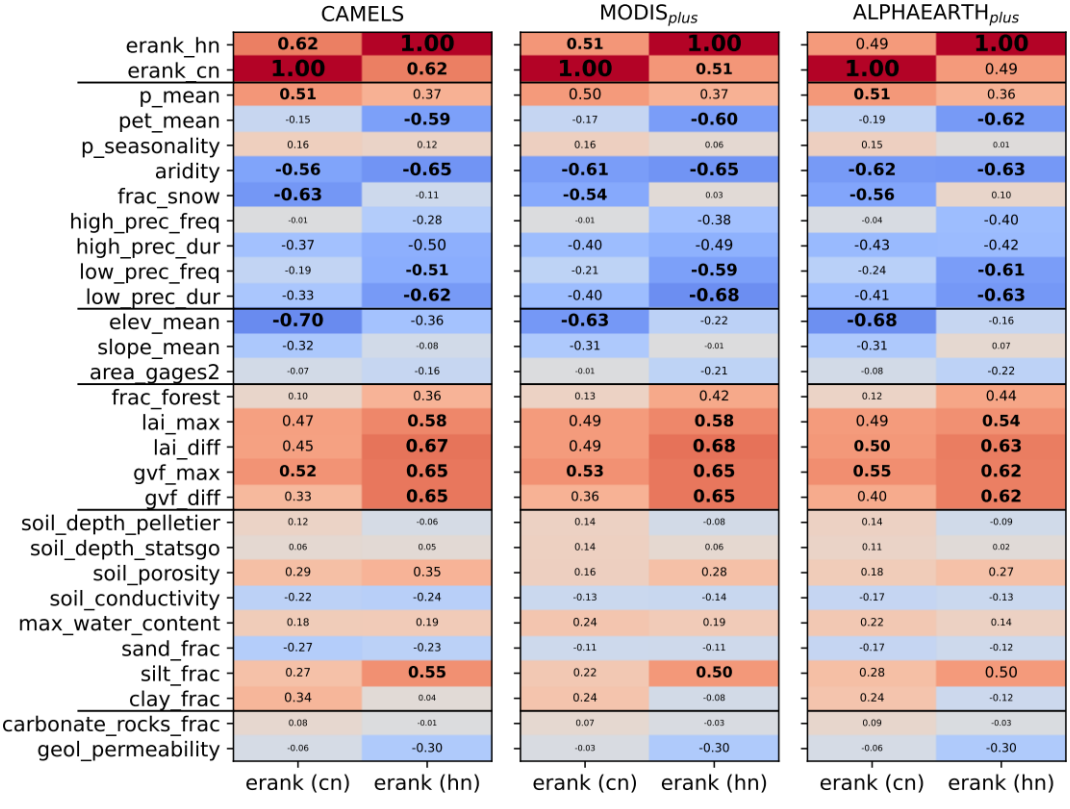

Figure 3: Pearson r correlation between CAMELS static features and the erank latent space complexity of the cell state (cn) and hidden state (hn) for the best-performing EMB-LSTM models with CAMELS, $MODIS_{plus}$ and $ALPHAEARTH_{plus}$ data as static features.

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Analyzing the specific correlation patterns (Figure 3), we observe that meteorological, topographic and vegetative attributes are highest correlated, whereby topography and meteorology are negatively correlated to latent space complexity, and attributes of vegetative state are positively correlated. Conversely, none of the soil and geological properties except silt fraction are significantly correlated to latent space complexity. This last fact is due to the lack of inclusion of soil- and geology-related datasets in the study for conceptual reasons (see Chapter 2.3.1), and further highlights that there is a significant gap for measurement-derived datasets covering soil and geology. Currently, no such dataset, purely based on measurement data exists. Regarding topography, it is interesting that catchment area and to some extent catchment slope do not seem to condition latent space complexity. This contradicts previous results based on feature importance tests (e.g. Kratzert et al., 2019a) that showed high relevance of catchment area within the same model framework. Only catchment elevation is important in conditioning the latent space significantly. It does so negatively, indicating that higher-altitude catchments are simpler to represent by the LSTM.

Regarding meteorology, there is an interesting inverse relationship between mean precipitation, which is positively correlated to both eranks (especially of the hidden state), whereas all other meteorologically attributes are negatively correlated. The model thus allocates *more* internal dimensions to represent humid systems than arid systems. This could be interpreted such that arid systems are simpler systems, but since arid catchments are known to perform worse, we instead suggest this is due to the fact that the data is insufficiently informative to properly characterize input-state-output behaviors of arid systems, hence leading to oversimplified internal characterization.

Finally, all vegetative attributes are positively correlated with latent space complexity. Thereby, all vegetative attributes indicate density of vegetative cover. Thus, the positive correlation with latent space means that the more dense the vegetive cover is, the more internal dimensions the model needs to correctly represent it, i.e. that catchments with dense growth have higher hydrological complexity. This is in line with conventional ecohydrological understanding that vegetation introduces additional hydrological dynamics to a hydrological system, increasing pathways and stores of the system.

4.2.3 Notes on interpretability

Our results suggest that we have only just scratched the surface of what can be understood by analyzing the latent space complexity. It gives an impression that the latent space of deep-learning models can serve as a useful diagnostic tool for investigating hydrological system functioning. Similarly, the study by Álvarez Chaves et al. (2026) demonstrates that latent space representations can further be leveraged to uncover underlying physical processes and system dynamics.

Finally, we want to highlight how central the correlation to human perception-derived catchment attributes is to our analysis. It shows that these attributes serve an important function in allowing us to access and interpret the models. In other words, only with these attributes were we able to reason concretely based on hydrologic understanding. This will allow us to, in turn, refine and conceptualize model application via improved theories of the systems' complexities, especially if taking this methodology further in future studies. We therefore do not connect the improved performance achieved by non-interpretable sets of static features like ALPHAEARTH with a call to fully abandon conventional static features. Instead, we call for renewed

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

efforts to improve upon the existing set of static features, by means of using these new tools and datasets to improve our ability to interpret and reason, instead of reducing it.

5 Conclusion

We tested a number of datasets and model architectures and examined how they affect performance on prediction in ungauged basins. To answer the title question, in our study, the more important knob to adjust arguably proved to be the makeup of the input data. Architectural changes improved performance only when they helped the otherwise capable core model (the LSTM layer) to absorb bigger and more information-complete data. In that, simpler architectures (EMB-LSTM) proved to be equally advantageous as more complex ones (CA-LSTM).

Detailing these conclusions, concerning data, ALPHAEARTH embeddings were technically most effective in improving ungauged prediction performance. But the addition of ALPHAEARTH embeddings led to peak performance only after addition of auxiliary data products such as DEM-, coordinate-, and meteorology-based static feature sets. Thus, our results imply that there is room for improvement of the ALPHAEARTH product. By testing these data products in combination with deeper network architectures, we were able to show that attention-based model architectures can substantially help to improve ungauged prediction by exploiting more of the existing data, but only if the data is under-representative (i.e. with the conventional CAMELS static features, or with MODIS-based ones). When more informationally-complete data was made available to the model (i.e. the ALPHAEARTH static features), the most effective architectural improvement was a relatively simple one that merely helped the LSTM layer to better absorb this data via one upstream embedding layer (EMB-LSTM), reaching a new improved PUB performance of NSE_{median}=0.721 compared to the previous benchmark of 0.696. It is, however, not significantly better than the attention-enhanced LSTM (CA-LSTM) in combination with the ALPHAEARTH-based static features (NSE_{median}=0.714) *as well as* the CA-LSTM in combination with MODIS-based static features (NSE_{median}=0.715).

Our study applies a new explainability tool in hydrology, the *Shannon effective rank complexity of the latent space* of the LSTM layer. We use this measure primarily as a way to empirically support the hypothesis that improved ungauged prediction performance can be attributed to higher-quality latent hydrological representation within the model. In addition, we show that the Shannon effective rank can serve as a useful tool for advancing hydrological understanding. Importantly, gaps were revealed regarding the representation of soil and geology datasets in data-driven models, which remained largely unconditional to the latent representation in our models. This highlights the need for improved attribute datasets, that are purely measurement-derived, for the soil and geology domain. In any case, the overall representational complexity of the LSTM built on the CAMELS dataset varies strongly depending on which catchment was passed through the model. This indicates that the method provides great potential to further advance our understanding of hydrological system functioning – turning deep learning based methods into data mining machines can hopefully allow us to challenge existing knowledge in future studies.

Finally, we reiterate that the correlation to human perception-derived catchment attributes was central to our explainability analysis. These attributes allowed us to access and interpret the models, and to reason based on hydrologic understanding.

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

565 Reasoning with models is central for further model development, as it will allow us to (for example) refine future model application. This underpins the importance of extending our ability to interpret deep-learning models. Accordingly, we call for more efforts to strengthen interpretation and reasoning, and not to focus exclusively on improving performance.

Appendices

Appendix A: Overview of all models used in this study, showing which combination of static feature data and model architecture are presented in which figure or table, next to a short description.

Static feature data	Model architecture	Tab. 2	Tab. 3	Fig. 1	Fig. 2	Description
None	base LSTM	x				base LSTM with no static features
None	EMB-LSTM			x		embedding LSTM with no static features
CAMELS	base LSTM	x	x			base LSTM with 27 CAMELS static features
CAMELS	CA-LSTM		x			cross-attention LSTM with 27 CAMELS static features
CAMELS	EMB-LSTM		x	x	x	embedding LSTM with 27 CAMELS static features
MODIS	base LSTM	x				base LSTM with MODIS static features
MODIS+DEM	base LSTM	x				base LSTM with MODIS and DEM static features
MODIS+DEM+METEO	base LSTM	x				base LSTM with MODIS, DEM and METEO static features
MODIS+DEM+METEO+XY(=$MODIS_{plus}$)	base LSTM	x	x			base LSTM with MODIS, DEM, METEO and XY static features
MODIS+DEM+METEO+XY(=$MODIS_{plus}$)	CA-LSTM		x			cross-attention LSTM with MODIS, DEM, METEO and XY static features
MODIS+DEM+METEO+XY(=$MODIS_{plus}$)	EMB-LSTM		x	x	x	embedding LSTM with MODIS, DEM, METEO and XY static features
ALPHAEARTH	base LSTM	x				base LSTM with ALPHAEARTH static features
ALPHAEARTH+DEM	base LSTM	x				base LSTM with ALPHAEARTH and DEM static features
ALPHAEARTH+DEM+METEO	base LSTM	x				base LSTM with ALPHAEARTH, DEM and METEO static features
ALPHAEARTH+DEM+METEO+XY($ALPHAEARTH_{plus}$)	base LSTM	x	x			base LSTM with ALPHAEARTH, DEM, METEO and XY static features
ALPHAEARTH+DEM+METEO+XY($ALPHAEARTH_{plus}$)	CA-LSTM		x			cross-attention LSTM
ALPHAEARTH+DEM+METEO+XY($ALPHAEARTH_{plus}$)	EMB-LSTM		x	x	x	embedding LSTM

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Appendix B: Performance of the base LSTM model with different datasets expressed as NSE calculated across the test period. CAMELS stands for the 27 static features from the CAMELS dataset, XY stands for WGS4 coordinates, METEO stands for static features derived from the dynamic meteorologic forcings, DEM stands for static features derived from a Digital Elevation Model, MODIS stands for static features based on MODIS-based surface reflectance bands, and ALPHAEARTH for a model with AlphaEarth embeddings as static features. More details on these datasets can be found in Table 1 and Chapter 2.3.

	Base LSTM	
Static feature data	NSE_{median}	$Unc_{5\text{-}95}$
None	0.634	±0.073
CAMELS	0.696	±0.054
XY	0.664	±0.067
METEO	0.673	±0.071
DEM	0.678	±0.063
MODIS	0.668	±0.071
ALPHAEARTH	0.705	±0.055

Code and data availability

All data and code will be made public via a github (code) and zenodo (data) repository upon paper publication. All coding was done in Python and facilitated by the use of various LLMs. Modeling was done in Pytorch (Paszke et al., 2019).

Author contributions

Conceptualization: Heudorfer, Loritz & Gupta / Data curation: Heudorfer & Dolich / Formal analysis: Heudorfer, Dolich / Funding acquisition: Loritz / Investigation: Heudorfer / Methodology: Heudorfer / Project administration: Loritz / Resources: Loritz, Gupta / Software: Heudorfer / Supervision: Gupta, Loritz / Validation: Heudorfer / Visualization: Heudorfer / Writing original draft: Heudorfer / Review and editing: Heudorfer, Gupta, Dolich, Loritz.

References

Acuña Espinoza, E., Chaves, M. Á., Dolich, A., & J, A. M. (2025). *Hy2DL: Hybrid Hydrological modeling using Deep Learning methods* [Software]. Zenodo. https://doi.org/10.5281/zenodo.17251944

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrol. Earth Syst. Sci.*

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Álvarez Chaves, M., Acuña Espinoza, E., Ehret, U., & Guthke, A. (2026). When physics gets in the way: An entropy-based evaluation of conceptual constraints in hybrid hydrological models. *Hydrology and Earth System Sciences*, *30*(3), 629–658. https://doi.org/10.5194/hess-30-629-2026

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., & others. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies–Chile dataset. *Hydrology and Earth System Sciences*, *22*(11), 5817–5846.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). *xLSTM: Extended Long Short-Term Memory*. NeurIPS. https://doi.org/10.52202/079017-3417

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., & Ilyushchenko, S. (2025). Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*. https://doi.org/10.48550/arXiv.2507.22291

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., & Kohli, P. (2025). *AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data* (arXiv:2507.22291). arXiv. https://doi.org/10.48550/arXiv.2507.22291

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J., Lane, R., Lewis, M., Robinson, E. L., & others. (2020). CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data*, *12*(4), 2459–2483.

Eddin, M. H. S., Zhang, Y., Kollet, S., & Gall, J. (2025). *RiverMamba: A State Space Model for Global River Discharge and Flood Forecasting* (arXiv:2505.22535). arXiv. https://doi.org/10.48550/arXiv.2505.22535

Ehret, U., Chen, J., & Lerch, S. (2025). *A comparative study of algorithms for lossy compression of 2-d meteorological gridded fields*. EGU General Assembly. https://doi.org/10.5194/egusphere-egu25-5977

Falcone, J. A. (2011). GAGES-II: Geospatial attributes of gages for evaluating streamflow. *USGS Report*, 41. https://doi.org/10.3133/70046617

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Gauch, M., Kratzert, F., Klotz, D., Shalev, G., Cohen, D., & Gilon, O. (2025). Towards Deep Learning River Network Models. *EGU General Assembly Conference Abstracts*, EGU25-9768.

Ghosh, R., Yang, H., Khandelwal, A., He, E., Renganathan, A., Sharma, S., Jia, X., & Kumar, V. (2023). *Entity Aware Modelling: A Survey* (arXiv:2302.08406). arXiv. http://arxiv.org/abs/2302.08406

Gleeson, T., Moosdorf, N., Hartmann, J., & van Beek, L. P. H. (2014). A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity. *Geophysical Research Letters*, *41*(11), 3891–3898. https://doi.org/10.1002/2014GL059856

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2017.06.031

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

Hartmann, J., & Moosdorf, N. (2012). The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochemistry, Geophysics, Geosystems*, *13*(12). https://doi.org/10.1029/2012GC004370

Heudorfer, B., Gupta, H. V., & Loritz, R. (2025). Are Deep Learning Models in Hydrology Entity Aware? *Geophysical Research Letters*, *52*(6), e2024GL113036. https://doi.org/10.1029/2024GL113036

Heudorfer, B., Liesch, T., & Broda, S. (2024). On the challenges of global entity-aware deep learning models for groundwater level prediction. *Hydrology and Earth System Sciences*, *28*(3), 525–543. https://doi.org/10.5194/hess-28-525-2024

Hochreiter, S., & Schmidhuber, J. (1997). LONG-SHORT-TERM MEMORY. *Neural Computation*, *9*(8), 1735–1780.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., … Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Jarvis, A., Reuter, H. I., Nelson, A., & Guevara, E. (2008). Hole-filled SRTM for the globe Version 4. *Hole-filled SRTM for the globe Version 4*, *15*(25–54), 5.

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Kraft, B., Kauzlaric, M., Aeberhard, W. H., Zappa, M., & Gudmundsson, L. (2025). *DROP: A scalable deep learning approach for runoff simulation and river routing*.

Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. *Hydrology and Earth System Sciences*, *28*(17), 4187–4201. https://doi.org/10.5194/hess-28-4187-2024

Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. (2022). NeuralHydrology. *Journal of Open Source Software*, *7*(71), 4050. https://doi.org/10.21105/joss.04050

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019a). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, *25*(5), 2685–2703. https://doi.org/10.5194/hess-25-2685-2021

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019b). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., & Matias, Y. (2023). Caravan—A global community dataset for large-sample hydrology. *Scientific Data*, *10*(1), 61. https://doi.org/10.1038/s41597-023-01975-w

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., Xu, S., Tayal, K., Nieber, J., Duffy, C., Steinbach, M., & Kumar, V. (2022). Regionalization in a Global Hydrologic Deep Learning Model: From Physical Descriptors to Random Vectors. *Water Resources Research*, *58*(8), e2021WR031794. https://doi.org/10.1029/2021WR031794

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Liesch, T., & Ohmer, M. (2025). *Strategies for Incorporating Static Features into Global Deep Learning Models*. Groundwater hydrology/Modelling approaches. https://doi.org/10.5194/egusphere-2025-4048

Liu, J., Bian, Y., Lawson, K., & Shen, C. (2024). Probing the limit of hydrologic predictability with the Transformer network. *Journal of Hydrology*, *637*, 131389. https://doi.org/10.1016/j.jhydrol.2024.131389

Liu, J., Shen, C., O'Donncha, F., Song, Y., Zhi, W., Beck, H. E., Bindas, T., Kraabel, N., & Lawson, K. (2025). From RNNs to Transformers: Benchmarking deep learning architectures for hydrologic prediction. *Hydrology and Earth System Sciences*, *29*(23), 6811–6828. https://doi.org/10.5194/hess-29-6811-2025

Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S. K., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., & Tarasova, L. (2024). CAMELS-DE: Hydro-meteorological time series and attributes for 1582 catchments in Germany. *Earth System Science Data*, *16*(12), 5625–5642. https://doi.org/10.5194/essd-16-5625-2024

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., & Nijssen, B. (2002). A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States*. *Journal of Climate*, *15*(22), 3237–3251. https://doi.org/10.1175/1520-0442(2002)015%253C3237:ALTHBD%253E2.0.CO;2

Miller, D. A., & White, R. A. (1998). A Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling. *Earth Interactions*, *2*(2), 1–26. https://doi.org/10.1175/1087-3562(1998)002%253C0001:ACUSMS%253E2.3.CO;2

Monteith, J. L. (1972). Solar Radiation and Productivity in Tropical Ecosystems. *Journal of Applied Ecology*, *9*(3), 747–766.

Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.-J., & Abrahart, R. (2016). Data-driven modelling approaches for socio-hydrology: Opportunities and challenges within the Panta Rhei Science Plan. *Hydrological Sciences Journal*, *61*(7), 1192–1208.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology*, *10*(3), 282–290.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., & Duan, Q. (2015). Development of a large-sample watershed-scale hydrometeorological data set for

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, *18*(8), 2215–2225.

Nolte, A., Haaf, E., Heudorfer, B., Bender, S., & Hartmann, J. (2024). Disentangling coastal groundwater level dynamics in a global dataset. *Hydrology and Earth System Sciences*, *28*(5), 1215–1249. https://doi.org/10.5194/hess-28-1215-2024

Ohmer, M., Doll, F., & Liesch, T. (2025). Incorporating Spatial Information for Regionalization of Environmental Parameters in Machine Learning Models. *Mathematical Geosciences*, *57*(2), 251–273. https://doi.org/10.1007/s11004-024-10163-4

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., & others. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Qu, P., Ouyang, W., Zhang, C., Chai, Y., Xu, S., Ye, L., Piao, Y., Zhang, M., & Lu, H. (2026). Utilizing Earth Foundation Models to Enhance the Simulation Performance of Hydrological Models with AlphaEarth Embeddings. *arXiv preprint arXiv:2601.01558*.

Roscher, R., Russwurm, M., Gevaert, C., Kampffmeyer, M., Dos Santos, J. A., Vakalopoulou, M., Hänsch, R., Hansen, S., Nogueira, K., Prexl, J., & Tuia, D. (2024). Better, not just more: Data-centric machine learning for Earth observation. *IEEE Geoscience and Remote Sensing Magazine*, *12*(4), 335–355. https://doi.org/10.1109/MGRS.2024.3470986

Rouse, W. (1973). Monitoring vegetation system in the great plain with ERTS. *3rd ERTS symposium, NASA Washington DC, 1973*, *1*, 309–317.

Roy, O., & Vetterli, M. (2007). *The Effective Rank: A Measure of Effective Dimensionality*. 15th European Signal Processing Conference (EUSIPCO), Poznan, Poland.

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDONNELL, J. J., Mendiondo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., & Zehe, E. (2003).

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421

Spoto, F., Sy, O., Laberinti, P., Martimort, P., Fernandez, V., Colin, O., Hoersch, B., & Meygret, A. (2012). Overview Of Sentinel-2. *2012 IEEE International Geoscience and Remote Sensing Symposium*, 1707–1710. https://doi.org/10.1109/IGARSS.2012.6351195

Thornton, M., Shrestha, R., Wei, Y., Thornton, P., & Kao, S.-C. (2022). Daymet. *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1 (Version 4.1)*. https://doi.org/doi.org/10.3334/ORNLDAAC/2129

Thornton, P. E., Running, S. W., & White, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology*, *190*(3–4), 214–251. https://doi.org/10.1016/S0022-1694(96)03128-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vermote, E. F., Saleous, N. Z. E., & Justice, C. O. (2002). Atmospheric correction of MODIS data in the visible to middle infrared: First results. *Remote Sensing of Environment*, *83*(1), 97–111. https://doi.org/10.1016/S0034-4257(02)00089-5

Watson, D. J. (1947). Comparative Physiological Studies on the Growth of Field Crops: I. Variation in Net Assimilation Rate and Leaf Area between Species and Varieties, and within and between Years. *Annals of Botany*, *11*(1), 41–76. https://doi.org/10.1093/oxfordjournals.aob.a083148

Willems, J. C. (2007). The Behavioral Approach to Open and Interconnected Systems. *IEEE Control Systems Magazine*, *27*(6), 46–99. https://doi.org/10.1109/MCS.2007.906923

Wulder, M. A., Loveland, T. R., Roy, D. P., Crawford, C. J., Masek, J. G., Woodcock, C. E., Allen, R. G., Anderson, M. C., Belward, A. S., Cohen, W. B., Dwyer, J., Erb, A., Gao, F., Griffiths, P., Helder, D., Hermosilla, T., Hipple, J. D., Hostert, P., Hughes, M. J., … Zhu, Z. (2019). Current status of Landsat program, science, and applications. *Remote Sensing of Environment*, *225*, 127–147. https://doi.org/10.1016/j.rse.2019.02.015

https://doi.org/10.5194/egusphere-2026-1965
Preprint. Discussion started: 20 April 2026

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., & Mocko, D. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres*, *117*(D3), 2011JD016048. https://doi.org/10.1029/2011JD016048

Yang, Y., & Chui, T. F. M. (2024). *Learning Generative Models for Lumped Rainfall-Runoff Modeling* (arXiv:2309.09904). arXiv. https://doi.org/10.48550/arXiv.2309.09904

Yang, Y., Janssen, J., Gupta, H., & Chui, T. F. M. (2026). *On the Adversarial Robustness of Hydrological Models* (arXiv:2602.05237). arXiv. https://doi.org/10.48550/arXiv.2602.05237