Geophysical Research Letters

RESEARCH LETTER
10.1029/2024GL113036

Are Deep Learning Models in Hydrology Entity Aware?

Benedikt Heudorfer[1], **Hoshin V. Gupta**[2], **and Ralf Loritz**[3]

[1]Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research - Atmospheric Trace Gases and Remote Sensing (IMK-ASF), 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

Key Points:

- Meteorological dynamic features are the main driver of the models' generalizability capability, more than physiographic static features
- Results reveal limitations regarding the suitability of current deep learning (DL) models when applied to changing environmental conditions
- DL models need better methods to exploit information encoded in physiographic static features, and more data for training

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

B. Heudorfer,
benedikt.heudorfer@kit.edu

Citation:

Heudorfer, B., Gupta, H. V., & Loritz, R. (2025). Are deep learning models in hydrology entity aware? *Geophysical Research Letters*, *52*, e2024GL113036. https://doi.org/10.1029/2024GL113036

Received 15 OCT 2024
Accepted 25 FEB 2025

Author Contributions:

Conceptualization: Benedikt Heudorfer, Hoshin V. Gupta, Ralf Loritz
Data curation: Benedikt Heudorfer
Formal analysis: Benedikt Heudorfer
Funding acquisition: Ralf Loritz
Investigation: Benedikt Heudorfer
Methodology: Benedikt Heudorfer
Project administration: Ralf Loritz
Resources: Benedikt Heudorfer
Software: Benedikt Heudorfer
Supervision: Hoshin V. Gupta, Ralf Loritz
Validation: Benedikt Heudorfer
Visualization: Benedikt Heudorfer
Writing – original draft: Benedikt Heudorfer
Writing – review & editing: Hoshin V. Gupta, Ralf Loritz

Abstract

Hydrology is experiencing a shift from process-based toward deep learning (DL) models. Entity-aware (EA) DL models with static features (predominantly physiographic proxies) merged to dynamic forcing features show significant performance improvements. However, recent studies challenge the notion that combining dynamic forcings with static attributes make such models entity aware, suggesting that static features are not effectively leveraged for generalization. We examine entity awareness using state-of-the-art Long-Short Term Memory (LSTM) networks and the CAMELS-US data set. We compare EA models provided with physiographic static features to ablated variants not provided with static inputs. Findings suggest that the superior performance of EA models is primarily driven by information provided by meteorological data, with limited contributions from physiographic static features, particularly when tested out-of-sample. These results challenge previously held assumptions regarding how physiographic proxies contribute to generalization ability in EA Models, highlighting the need for new approaches for robust generalization in DL models.

Plain Language Summary

We investigate whether deep learning (DL) models used in streamflow prediction can make good use of catchment characteristics to improve predictions. Catchment characteristics (e.g., elevation, soil type), describe physical catchment properties. Models that combine catchment characteristics with meteorological data (e.g., rainfall and temperature) to predict streamflow are called entity-aware (EA) models. While EA models have shown superior performance in predicting streamflow, recent studies suggest catchment characteristics may not be as useful as previously thought, especially when making predictions at new locations that the model has never see before (out-of-sample). To examine this, we use the CAMELS-US streamflow data set to build models that predict streamflow with meteorological data. Our results show that providing the models with catchment characteristics enhances the models understanding of catchments only marginally, and leads to only slightly improved performance. We conclude that meteorological data rather than catchment characteristics play the key role in the models' predictive success. This raises questions about how DL models exploit the information encoded within catchment characteristics and suggests the need for new methods to better integrate such information into our models for robust generalization. Also, we probably need more data to successfully solve this issue, but we need the right kind of data.

1. Introduction

The hydrologic domain is undergoing a major shift from the use of process-based models toward deep learning (DL) approaches. In the past few years, a series of studies on streamflow prediction have compared process-based (PB) models and DL models and consistently found superior performance of the latter (Acuña Espinoza et al., 2024; Feng et al., 2020; Kratzert et al., 2019a, 2019b, 2021; Lees et al., 2021; Ma et al., 2021; Nearing et al., 2024).

In hydrology, DL models achieved a significant breakthrough through the work of Kratzert et al. (2018, 2019a, 2019b), who successfully applied Long-Short Term Memory (LSTM) neural networks (Hochreiter & Schmidhuber, 1997) to streamflow prediction, demonstrating their potential to outperform traditional PB approaches. These studies introduced two major innovations. First, the model was set up as a Global DL Model (referred to as a "regional" model in the aforementioned studies), meaning that a single LSTM was able to simultaneous make predictions at all locations constituting the data set at hand, which in their case was the CAMELS-US data set (Addor et al., 2017; Newman et al., 2015). In contrast, previous approaches were limited to predictions only at the level of individual catchments. Second, the LSTMs were designed as what is known as Entity-Aware models (hereafter EA Models), following Ghosh et al. (2023). Today, nearly all global DL models in hydrology incorporate the concept of entity awareness. But what exactly is entity awareness?

The concept of EA Models is neither unique nor original to the field of streamflow prediction (Ghosh et al., 2023). It refers in general to models that are able to deliver distinguishable predictions between entities situated in a larger data set. In hydrology, the first use of the term EA Model was in Kratzert et al. (2019a) in reference to their introduction of a modified LSTM architecture with a slightly different structure for internal data processing compared to the original "vanilla LSTM", leading to minor changes in the parameter space. However, they reported that this modified LSTM underperformed and, as a consequence, was rarely ever used again. The superior performing vanilla LSTM used a simple concatenation-type approach to data fusion. It was provided with static features (i.e., physiographic proxies, catchment characteristics) alongside dynamic features (meteorological forcings) by simply concatenating the static features to the dynamic features at every time step. These concatenated features were jointly fed into the LSTM, enabling it to discern between individual entities (catchments) when predicting the target feature (streamflow) in response to the dynamic input features.

All subsequent studies by those authors (e.g., Kratzert et al., 2019a, 2021, 2024) and in fact by most researchers in the field (e.g., Frame et al., 2022; Gauch et al., 2021; Klotz et al., 2022; Lees et al., 2022; Loritz et al., 2024) make use of this concatenation-type model setup. However, other data fusion methods and architectures exist (e.g., Arsenault et al., 2023; Heudorfer et al., 2024) and, in fact, any DL model (e.g., Transformers, Liu et al., 2024) that is fed static features along with dynamic features is today implicitly treated as being entity aware. In other words, the model is generally expected to be able to discern entities in a physiographically meaningful way by exploiting information encoded by the static features. The rationale is that, since the static features are assumed to represent meaningful physiographic proxies that are indicative of catchment response, the model will be able to generalize streamflow predictions to locations with similar values for static features.

Of course, this generalization capability can only be exploited when the model is run spatially out-of-sample. To clarify, any time series prediction model implicitly predicts temporally out-of-sample while remaining spatially in-sample; that is, testing happens on a test period unseen during training but on catchments seen during training, which we refer to as the in-sample (IS) mode. But a model run spatially out-of-sample is simultaneously tested on both unseen test periods and unseen locations, which we refer to as the out-of-sample (OOS) mode. Therefore, OOS is just another term for Prediction in Ungauged Basins (PUB, Hrachowitz et al., 2013; Sivapalan et al., 2003), or regionalization, which is the context under which EA Models could really be put to beneficial use. A model class capable of regionalization by making use of external observable data regarding catchment attributes (static features) would signify a major step forward in PUB. Further, understanding how DL models regionalize is key to understanding how they will behave under changing environments (e.g., changes in land-use or climate) or under extrapolation beyond previously seen behaviors, opening the possibility of scenario analysis (e.g., changing static features or using unseen combinations of static features).

To date, however, little proof has been provided that DL models developed in the fashion mentioned above—fed with information about static features concatenated along with the dynamic features—are actually able to generalize from static features in this way. Given that during IS runs, an entities' exact combination of static and dynamic features is always known during training and testing, the model could just as well leverage knowledge about this exact combination to consistently provide strong performance, without acquiring deeper generalization capabilities. Consequently, the implications for regionalization (PUB) can only be tested in a consistent OOS test setting.

In a previous study, Heudorfer et al. (2024) compared IS performance and OOS performance of EA models with various sets of static features in a groundwater data set in Germany. They found that the OOS-EA model could not make good use of static features, but relied mostlys on dynamic features for prediction. However, the data fusion method used in that study was not state-of-the-art; it uses a two-branched model architecture where static and dynamic features did not get processed jointly by the same LSTM layer, leading to implausible loss in out-of-sample runs. Furthermore, their results were generated with data from the groundwater domain, which has known and severe data scarcity problems due to the hidden conditions underground, rendering static features inherently more uncertain (Barthel, 2014; Barthel et al., 2021), aggravated by the fact that the data set was small and heavily biased toward climate predictability.

Here, we transfer the experimental setup to the hydrologic domain, using a much larger data set that is better balanced (CAMELS-US, see Section 2.1), using state-of-the-art EA model architecture to establish an extended experimental setup (Section 2.2) and test against an established benchmark (Kratzert et al., 2021). We compare the IS and OOS performance of a LSTM-based EA model when using (a) physiographic static features, (b) static features derived purely from the associated meteorological dynamic features and (c) dynamic features only. With this approach, the study aims to address a central question: to what degree do current state-of-the-art EA models actually possess generalization ability (GA) arising from entity awareness? After reporting on the results (Section 3) and discussing model theoretic implications (Section 4), we crystallize the implications for the field going forward (Section 5).

2. Data and Methods

2.1. Data

We used the CAMELS-US streamflow data set (Addor et al., 2017) as target data. Of the 671 catchments, we used a sub-selection of 531 catchments (Newman et al., 2017) to allow benchmark comparison, since this subset was used by the most important DL benchmark studies (Kratzert et al., 2019a, 2019b, 2021). For dynamic input data we used 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) data sets accompanying the CAMELS-US data set, totaling 15 dynamic features. These were fed into all three model variants used in this study. Two of these three variants were additionally fed static features. In the $\mathrm{EA}_{\mathrm{CAMELS}}$ model, we used a set of 27 physiographic proxies from the CAMELS-US data set as also used by for example, Acuña Espinoza et al. (2024), Kratzert et al. (2019a, 2021). Due to exhaustive description of these features provided elsewhere (Addor et al., 2017; Kratzert et al., 2019a), we refrain from further discussing them here. In the $\mathrm{EA}_{\mathrm{meteo}}$ model, we used means and standard deviations of the 15 meteorological dynamic features defined above, totaling 30 static features. Because this number is close to the 27 physiographic static features used in $\mathrm{EA}_{\mathrm{CAMELS}}$, influence of the stabilizing effect that a multiplicity of static features can have (Li et al., 2022) is avoided.

2.2. Model and Experimental Setup

We used a replica of the model used in Kratzert et al. (2021) as implemented by Acuña Espinoza et al. (2024) based on the neural hydrology repository (Kratzert et al., 2022) in Pytorch (Paszke et al., 2019). The model is a 1D (input) to 1D (output) dynamical systems model, where the inputs are meteorological time series data spatially aggregated over the catchment, and the output is a streamflow time series at the corresponding gauge location. The model is trained using data from multiple catchments simultaneously. It consists of a 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. The batch size was 256, the initial learning rate of the Adam optimizer was 0.001, and, in slight deviation to Kratzert et al. (2021) and Acuña Espinoza et al. (2024), the learning rate was adapted every 5 epochs to 80% of its previous value, over a total of 20 epochs to allow faster training while maintaining performance. For prediction, the last epochs' model state was used. Loss function was the basin-averaged Nash-Sutcliffe Efficiency (NSE, Nash & Sutcliffe, 1970) defined in Kratzert et al. (2019b). Training period was October 1999–September 2008, testing period was October 1989–October 1999, sequence length was 365 days.

We used an EA Model (rationale described in the introduction) and, following Heudorfer et al. (2024), ran three different variants thereof.

- **$\mathrm{EA}_{\mathrm{CAMELS}}$**: Using 27 physiographic static features from the CAMELS data set concatenated to 15 meteorologic forcings used as dynamic features.
- **$\mathrm{EA}_{\mathrm{meteo}}$**: Using 30 static features comprised of mean and standard deviation of the same 15 meteorological dynamic features that are used as forcing to the model. This entails no additional external data products as input to the model other than meteorological dynamic features.
- **$\mathrm{EA}_{\mathrm{ablated}}$**: Ablated variant without static features, relying solely on the 15 meteorological dynamic features for prediction.

Each of these model variants were run in-sample (IS, i.e. temporal out-of-sample but spatial in-sample) and out-of-sample (OOS, i.e. temporal as well as spatial out-of-sample). Practically, IS was implemented by using the train period of all 531 catchments for training and the test period of all 531 for prediction. OOS was implemented by a 5-fold cross validation, where in each fold, training happened on the train period of 80% (N) of the catchments, and prediction on the test period of the other 20% (N) of catchments.

Two test scores were calculated in the test period, the NSE (Nash & Sutcliffe, 1970) and the Kling-Gupta Efficiency (KGE, Gupta et al., 2009). To account for uncertainty due to model weight initialization, each model was run with 5 different seed initializations. Calculation of NSE and KGE was bootstrapped to obtain the final test score. For bootstrapping, each catchments' test periods from all 5 seed realizations were bagged into a single set. Then, from the bag an 80% sample was drawn 100 times with replacement. For each sample, NSE and KGE scores were calculated, resulting in 100 score realizations for each catchment. From these, the 50% (median), 5%, and 95% quantiles were calculated as median, lower, and upper uncertainty bounds (Figure 1).

To compare the score distributions of the different models, we ran two-sided Kolmogorov-Smirnov (KS) tests with the null hypothesis that distributions are identical (Hodges Jr, 1958). To test whether predicted mean streamflow is different between prediction scenarios (see below, and Text S4 in Supporting Information S1), we apply the Student's *t*-test for related samples (Efron & Hastie, 2021) with the null hypothesis that both samples' means are identical. For both, the Scipy (Virtanen et al., 2020) implementation was used. Also, to analyze the degree of shared information between the static features from the CAMELS data set and the static features derived as summary statistics from dynamic features, we calculated the pairwise mutual information score from the scikit-learn package (Pedregosa et al., 2011).

Furthermore, we compared scores to the benchmark of Kratzert et al. (2021), which remains the currently valid benchmark for CAMELS-US (Liu et al., 2024). Note that our score calculation is different from that in Kratzert et al. (2021), who first averaged the hydrographs of their 10 seed realizations and then calculated the score only once. To allow comparability with our bootstrapped scores, we re-calculated each catchments' final score of the results in (Kratzert et al., 2021) to be the median of 10 distinct scores calculated individually for each hydrograph in their seed ensemble. This results in a lowered median NSE of 0.797 compared to the median NSE of 0.821 (mean NSE 0.783) reported in Kratzert et al. (2021). Similar for KGE (see Figure S1 in Supporting Information S1).

Finally, to understand how far the models' generalization capabilities carry in real-world hydrological problem settings, we conducted a scenario-based sensitivity analysis, investigating how streamflow is affected by deforestation. We did metamorphic test runs (Reichert et al., 2024; Yang & Chui, 2021) with the OOS-$\text{EA}_{\text{CAMELS}}$, where attribute values (i.e., static features) are manipulated in inference mode (no retraining taking place) to test whether the changes made based on a deforestation scenario leads to the physically expected (Zhang et al., 2017) increase in discharge predicted during inference. Specifically, for a selection of 63 catchments with high likelihood of increased discharge after deforestation, we reduced the "forest fraction" feature by 50%, and the "LAI max" and the "GVF max" feature by 18%, according to the reasoning outlined in Text S4 in Supporting Information S1, based on empirical meta-studies of field experiments (Breuer et al., 2003; Zhang et al., 2017).

3. Results

Figure 1 highlights the NSE score performance differences between IS and OOS runs for the three EA model variants $\text{EA}_{\text{CAMELS}}$, EA_{meteo} and $\text{EA}_{\text{ablated}}$. Note first that the current benchmark for the CAMELS-US data set by Kratzert et al. (2021) is reproduced by our architectural counterpart IS-$\text{EA}_{\text{CAMELS}}$ with no significant difference in performance (Kolmogorov-Smirnov test statistic KS = 0.047, p-value = 0.599). Further, we can make three main observations based on this figure.

First, in both IS and OOS the order of performance of the three models is the same. The $\text{EA}_{\text{CAMELS}}$ model consistently achieves the best performance, indicating that, in principle, establishing entity awareness via hydrological static features leads to the best outcome. The EA_{meteo} model follows closely, with a small but significant (KS = 0.104, p-value = 6.6e−250) difference in performance. And, compared to these two models, the $\text{EA}_{\text{ablated}}$ model without any static features underperforms, as expected.

Second, both $\text{EA}_{\text{CAMELS}}$ and EA_{meteo} experience a significant drop in performance compared to the IS performance when run OOS, almost to the level of the OOS-$\text{EA}_{\text{ablated}}$ model. The marginal performance advantage of the OOS-$\text{EA}_{\text{CAMELS}}$ and OOS-EA_{meteo} over the OOS-$\text{EA}_{\text{ablated}}$ model suggests that the IS-$\text{EA}_{\text{CAMELS}}$ and IS-EA_{meteo} both mostly leverage knowledge about the exact combination of static and dynamic features known

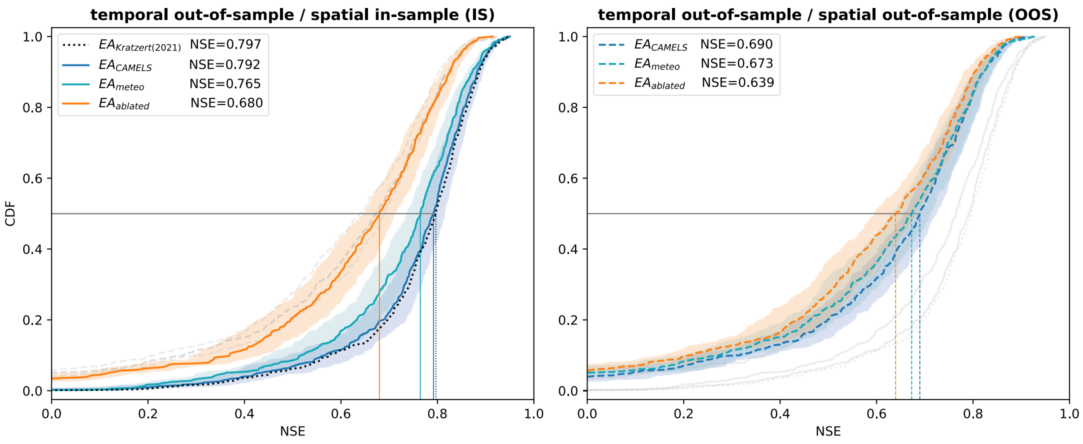

Figure 1. Cumulative distribution function (CDF) of the Nash-Sutcliffe Efficiency for three different models in spatial in-sample (IS) and out-of-sample (OOS) contexts. The models compared are (a): The EA model using the 27 static features from the CAMELS data set concatenated to the 15 meteorologic forcings as dynamic features ($\mathrm{EA}_{\mathrm{CAMELS}}$), (b): The EA model using 30 static features statistics derived from the mean and standard deviation of the 15 meteorological dynamic features used as input ($\mathrm{EA}_{\mathrm{meteo}}$), and (c): the ablated model without any static features, that is, only dynamic features as input ($\mathrm{EA}_{\mathrm{ablated}}$). Shades reflect the uncertainty estimates as specified in chapter *2.2*. Gray lines indicate the location of the CDFs in the respective other case (IS or OOS). Kling-Gupta Efficiency scores can be found in the supplements.

during training as well as testing to accomplish the impressive performance improvements over IS-$\mathrm{EA}_{\mathrm{ablated}}$, but derive only minimal generalization capability from the static features in OOS mode, when the exact combination is not known during testing. This means that both $\mathrm{EA}_{\mathrm{CAMELS}}$ and $\mathrm{EA}_{\mathrm{meteo}}$ are subject to overfitting to the training data, and that, as emphasized by the OOS results, the models' generalization capability derived from the static features is very limited. In this context, the only robust model proves to be the $\mathrm{EA}_{\mathrm{ablated}}$ model, which retains its overall level of performance on unseen data (Figure 1).

Third (and most importantly), the fact that the $\mathrm{EA}_{\mathrm{CAMELS}}$ model outperforms the other models, particularly in the IS case, seems to suggest that there is, in principle, value in providing EA Models with detailed hydrological static features. However, the use of simple summary statistics (i.e., the mean and standard deviation) from the models' own input forcing in the $\mathrm{EA}_{\mathrm{meteo}}$ model—which, notably, entails no additional external information as input other than the meteorological dynamic features—leads to similar but statistically different (KS = 0.048, *p*-value = 2.3e−53) performance. This indicates that the EA Models is able to compensate most of the physiographic static features provided by the CAMELS data set. This is further corrobated by additional analysis on the degree of mutual information between the CAMELS and METEO sets of static features (see Figure S2 in Supporting Information S1), which suggests that there is a fair share of mutual information between the topographic, vegetative and climatic static features from the CAMELS data set with the METEO static features derived from the input dynamic features. Put plainly, the information encoded in the form of physiographic proxies is to a certain extent redundant when compared with that provided by the meteorological data.

These findings are reinforced by the distributions of a second score we looked at, the KGE (Figure S1 in Supporting Information S1), which reproduces the patterns described above at large. Furthermore, the drop in performance from $\mathrm{EA}_{\mathrm{CAMELS}}$ to $\mathrm{EA}_{\mathrm{ablated}}$ or $\mathrm{EA}_{\mathrm{meteo}}$ exhibits no clear regional pattern across the study region (Figure S3in Supporting Information S1). As expected with Figure 1 in mind, the maps in Figure S3 in Supporting Information S1 show a less pronounced drop in OOS performance compared to the drop in IS performance. Interestingly, at the individual catchment level, a significant number of catchments experienced a moderate improvement in performance when predicted using OOS-$\mathrm{EA}_{\mathrm{ablated}}$ or OOS-$\mathrm{EA}_{\mathrm{meteo}}$ instead of OOS-$\mathrm{EA}_{\mathrm{CAMELS}}$, though the spatial distribution remained random, with no clear regional pattern.

Finally, we ran a scenario-based experiment on a real-world hydrological problem setting to further test the limits of generalization capability found in above results. In field experiments investigating changes to streamflow when forest cover is significantly reduced, the expectation (based on findings from a large meta-study, Zhang et al., 2017) is that streamflow increases. Metamorphic test runs with manipulated static features in inference mode due to 50% forest cover reduction (see Section 2.2 and Text S4 in Supporting Information S1) show that for all seeds, a significant increase in predicted streamflow of 6.9% can be found (Table S5 in Supporting Information S1). However, in 43% of the catchments, an actual decrease in streamflow took place, which is counterintuitive to the physics of the experiment, because streamflow decrease after forest cover loss, while theoretically possible, was never empirically observed according to Zhang et al. (2017). Additionally, on the level of individual seeds, only 2 out of 5 seeds show significant increase in streamflow (Table S5 in Supporting Information S1), relativizing the significance indicated across all seeds. In concordance with the overall results, this indicates existing but limited generalization capability, and limited robustness thereof, especially since the expected response to such a drastic reduction in forest cover (50%) would be a much more fundamentally altered hydrological regime (Zhang et al., 2017).

4. Discussion

4.1. Evaluating the Models' Performance and Generalization Ability

Previous studies have shown high performance for out-of-sample experiments with entity aware models (e.g., Arsenault et al., 2023; Kratzert et al., 2019a; Nearing et al., 2024) and entity aware models have outmatched PB models both in-sample (Acuña Espinoza et al., 2024; Feng et al., 2020; Kratzert et al., 2019b; Lees et al., 2021) and out-of-sample (Kratzert et al., 2019a; Nearing et al., 2024). However, our results raise questions about the extent to which the current model "*works for the right reason*" and how it utilizes physiographic static features to achieve generalizability in a hydrological sense (see Section 4.2). Given our results, it seems reasonable to ask if the static features are mostly used in-sample as unique identifiers to enable the model to differentiate unique input-output relationships between catchments. In other words if the model really learns a hydrological meaningful relation between static and dynamic features. This is because models without static features, as well as those with static features tested out-of-sample, can compensate for missing physiographic information by relying solely on dynamic inputs or derivates thereof.

However, our results show that there is at least some degree of entity awareness, indicated by the small out-of-sample performance advantage that the $\mathrm{EA}_{\mathrm{CAMELS}}$ model has over the $\mathrm{EA}_{\mathrm{meteo}}$ and $\mathrm{EA}_{\mathrm{ablated}}$ models, as well as the ambiguous but significant results of the metamorphic forest cover reduction experiment. This is in contrast to Heudorfer et al. (2024), where out-of-sample performance generally drops below $\mathrm{EA}_{\mathrm{ablated}}$ performance. We ascribe the difference to (a) the suboptimal model architecture in Heudorfer et al. (2024), and (b) the fact that groundwater as a domain has known and severe data scarcity issues due to the hidden conditions underground, rendering static features inherently more uncertain (Barthel et al., 2021), whereat conditions in surface hydrology are more observable, that is, data is more abundant, and existing static features are more trusted to be meaningful physiographic proxies of catchment response (Barthel, 2014), as well as possibly c) the significantly larger database of the present study.

Despite the small degree of entity awareness found, we detect performance-oriented indications of the absence of larger GA. We say "performance-oriented" because we simply compared out-of-sample performance against their "benchmark", which we take to be in-sample performance. This is in contrast to, for example, evaluating the quality of the abstraction inside the model, which is a considerably more complicated task, and beyond the scope of our study. As such, we take a pragmatic approach because we simply set the goal that the model is capable of exploiting the information encoded in the relationships between dynamic and static features, so that it can (at best) achieve out-of-sample performance levels that approach in-sample performance levels. Of course, achieving this would not necessarily indicate actual "understanding" in the sense of true abstraction (see Section 4.2). But from a pragmatic point of view, this would correspond to an acceptable level of generalization.

In light of pragmatism, we maintain that the overall level of out-of-sample perfomance is remarkably high ($0.639 < \mathrm{NSE} < 0.690$), allowing a clear recommendation for practitioners to just use (for now) the full set of static features to enable best prediction performance in-sample as well as out-of-sample. There is no reason to believe that any subset of static features would provide better performance, since neural networks are known to not be impeded by redundant information in principle. That is, even if we now understand that static features are not as indicative of catchment functioning as we thought they are. Notably, this does not diminish the use of static features overall, for example, for assessments of hydrologic similarity (e.g., Jehn et al., 2020). The limitations revealed in the present study specifically concern the use of static features in state-of-the-art methods using physiographic static features and dynamic features together as the primary mechanism for developing entity-aware DL models. The important question, then, is how can DL models actually learn to generalize hydrologic similarity?

4.2. Theoretic Considerations and Hypotheses to Achieve Better Generalization in Hydrology

First, to set the context, we follow Maier et al. (2023) and define "Order-One" GA as the ability to interpolate within the range represented by the training data (i.e., in-sample conditions), while "Order-Two" GA refers to the ability to perform well in extrapolation (i.e., out-of-sample conditions, on data not yet seen). Extending upon this, we can think of "Order-Three" GA as the ability to extrapolate to isomorphic catchments, that is, to situations where the system has undergone structural change, whether due to natural causes (e.g., land-form evolution) or anthropogenic impacts (e.g., land-use/land-cover modifications). For further intuition on this see Gupta (2024).

The premise of our study is that we want our model to gain sufficient GA from the training data to be able to make good Order-Two prediction (out-of-sample) or higher. Theoretically, to be able to achieve Order-Two or higher GA, the model must learn a sufficiently comprehensive representation of all relevant fundamental principles that characterizes the underlying data generation process. In other words, the model must constitute a true (or close-to-true) representational abstraction of the system that is, therefore, able to make suitably good predictions when confronted with data generated by similar outside entities. Put plainly, true GA (entity awareness) in hydrology would entail the model to replicate catchment functioning in the form of an internal abstraction that allows the model to adequately predict isomorphic catchment behavior. Whether or not they actually achieve this, PB models are intended to be just that, founded as they are on principles of mass and energy conservation and other theoretical or semi-theoretical understanding of how physical systems actually function (e.g., Loritz et al., 2017).

But the results presented here suggest that today's state-of-the-art DL models struggle to achieve the aforementioned goal of Order-Two GA. To achieve this, machine learning models need to construct their own representations of system functioning by extracting information directly from the data, not being provided with preexisting domain knowledge (except modeler imposed assumptions regarding which features are actually relevant as input). So one likely explanation for our findings is that the data currently available represents a limited and/or biased sample that is not sufficiently representative of the underlying data generating process, making it difficult for the model to learn a sufficiently comprehensive representation. This necessarily reinforces the ever-present call for more data. However, it is not simply more data that is required—what we need is data that more comprehensively characterizes the full dynamical extent of the underlying data generating process, such that the true nature of that process can be decoded via the learning process and be suitably represented within the model.

Second, none of this means that physiographic static features have no beneficial role to play in addressing the PUB/regionalization problem. All that we show is that the available static features provide little information beyond what is already encoded by the meteorological input data and the output data (streamflow). In fact, the data-processing inequality of information theory (Cover & Thomas, 2012) indicates that this must be true—all information about the nature of the input-state-output transformation is already encoded within the dynamical input-output data, as long as we remain within the range of physical expressiveness of that data (i.e., in-sample). Especially if such data would be fully expressive of the entire data generating process.

The real value of any information encoded within physiographic static features is to serve as surrogates for information about the input-state-output transformation, which is not available when we do not have immediate access to the dynamical output data (i.e., out-of-sample). Then, without access to the information encoded in physiographic static features, the model would not be able to distinguish between catchments that have identical meteorological drivers but functionally different output responses. However, this applies only to the extent that machine learning techniques can be leveraged to establish meaningful functional relationships between differences in physiographic static features and corresponding differences in system output responses that are not due to differences in the dynamical input data. But DL models might not be able to make use of this information, as we showed in this study. Crucially, this does not necessarily mean that the physiographic proxies are useless. We might equally be missing adequate translation techniques to enable the model to decode the information expressed in physiographic proxies.

Given this, the goals of PUB would be within reach. In PB models, this is commonly pursued by attempting to construct robust mappings between such features and the parameters that mediate the functioning of processes represented within the model (Jiang et al., 2020). Within DL models, this would have to practically be realized either as alterations to the values of the model weights (analogous to the aforementioned parameters) or, analogously, as mechanisms that somehow alter the patterns of information flows within the directed graph networks that make up such models. Exactly how best to successfully realize this is a topic that warrants detailed investigation. However, some recent work (see e.g. De La Fuente et al., 2024) does suggest directions that may be worth pursuing.

5. Conclusions

This study investigated and discussed if current state-of-the-art entity aware DL models make the most of the information encoded in physiographic proxies given to it in the form of static features. Despite the model class' implicit assumption of spatial generalization capability drawn from static features, our model comparison shows that a similar model performance can be achieved if static features are replaced by features derived from the dynamic inputs. This indicates that when confronted with new data, the model is currently limited in its ability to generalize from static features, and instead relies mostly on meteorological data for prediction. Meanwhile, the excellent in-sample performance that is observed in state-of-the-art models is likely achieved through use of static features to serve as kind of unique catchment or catchment group identifiers.

In reflecting upon these results, we hypothesize that we encounter the dual problem of (a) insufficient data to allow the model to learn the true underlying nature of the entire data generating process, thus limiting its GA and, more crucially, that we are (b) missing appropriate decoding procedures to enable the model to extract relevant information that is encoded in static features, which are natively not comprehensible to the model.

In summary, our findings suggest that deeper investigation is necessary to understand how the information encoded within physiographic proxies can be exploited to achieve model GA, and underscore the need for more data as well as new approaches to better integrate physiographic proxies into DL architectures for more reliable predictions. We argue that while extensive research has examined the limitations and uncertainties of PB models (Loritz et al., 2017), similarly rigorous scrutiny of DL approaches in hydrology remains in its early stages (Baste et al., 2025). Addressing this gap is crucial to ensuring that DL models do more than memorize training data and instead leverage physiographic features in a hydrologically meaningful way. Ultimately, the ingenuity of contemporary machine learning models to potentially create abstract system representations from data, when provided with little or no access to pre-existing domain knowledge (beyond modeler imposed assumptions regarding which observed variables are actually relevant), reveals the crucial latent theoretical question: When does the model really embody a functionally comprehensive abstract representation of the system that is generating the data we have access to? Unfortunately, at this stage of investigation we can only open up the question to provoke deeper investigation instead of being able to definitively answer it. Clearly, to approach a definitive answer would require a whole body of investigation that may be beyond the ability of a single study or even a single research group to conduct. With this paper, we hope to help re-initiate a research field that is similar to the large interest in PUB that arose in previous decades (see Hrachowitz et al., 2013).

Data Availability Statement

The Python code to reproduce this study is publicly accessible (Heudorfer & Acuña Espinoza, 2025) as well as the underlying data and model results (Heudorfer, 2025).

Acknowledgments
The study was funded by the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments. Article processing charges were covered by the Karlsruhe Institute of Technology (KIT). Models were run on the bwHPC computing cluster owned by the federal state of Baden-Württemberg. Model implementation in code was facilitated by guidance from Eduardo Acuña Espinoza. Special thanks go to the reviewers, especially Jonathan Frame, for providing constructive feedback that helped to significantly elevate the paper. Open Access funding enabled and organized by Projekt DEAL.

References

Acuña Espinoza, E., Loritz, R., Álvarez Chaves, M., Bäuerle, N., & Ehret, U. (2024). To bucket or not to bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization. *Hydrology and Earth System Sciences*, *28*(12), 2705–2719. https://doi.org/10.5194/hess-28-2705-2024

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., & Mai, J. (2023). Continuous streamflow prediction in ungauged basins: Long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences*, *27*(1), 139–157. https://doi.org/10.5194/hess-27-139-2023

Barthel, R. (2014). HESS Opinions "Integration of groundwater and surface water research: An interdisciplinary problem?". *Hydrology and Earth System Sciences*, *18*(7), 2615–2628. https://doi.org/10.5194/hess-18-2615-2014

Barthel, R., Haaf, E., Giese, M., Nygren, M., Heudorfer, B., & Stahl, K. (2021). Similarity-based approaches in hydrogeology: Proposal of a new concept for data-scarce groundwater resource characterization and prediction. *Hydrogeology Journal*, *29*(5), 1693–1709. https://doi.org/10.1007/s10040-021-02358-4

Baste, S., Klotz, D., Espinoza, E. A., Bardossy, A., & Loritz, R. (2025). Unveiling the limits of deep learning models in hydrological extrapolation tasks. *EGUsphere*, *2025*, 1–24. https://doi.org/10.5194/egusphere-2025-425

Breuer, L., Eckhardt, K., & Frede, H.-G. (2003). Plant parameter values for models in temperate climates. *Ecological Modelling*, *169*(2–3), 237–293. https://doi.org/10.1016/S0304-3800(03)00274-6

Efron, B., & Hastie, T. (2021). *Computer age statistical inference, student edition: Algorithms, evidence, and data science (Bd. 6)*. Cambridge University Press.

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, *56*(9), e2019WR026793. https://doi.org/10.1029/2019wr026793

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., et al. (2022). Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, *26*(13), 3377–3392. https://doi.org/10.5194/hess-26-3377-2022

Fuente, L. A. D. L., Bennett, A., Gupta, H. V., & Condon, L. E. (2024). A HydroLSTM-based machine-learning approach to discovering regionalized representations of catchment dynamics. https://doi.org/10.22541/essoar.172801404.45473140/v1

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021

Ghosh, R., Yang, H., Khandelwal, A., He, E., Renganathan, A., Sharma, S., et al. (2023). Entity aware modelling: A survey. *(No. arXiv: 2302.08406). arXiv*. http://arxiv.org/abs/2302.08406

Gupta, H. (2024). On machine learning "interpretable" representations of dynamical geoscientific systems. Retrieved from https://www.youtube.com/watch?v=-_z1M6ekS6s

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Heudorfer, B. (2025). Dataset to the paper Are deep learning models in hydrology entity aware? [Dataset]. https://doi.org/10.5281/zenodo.14871156

Heudorfer, B., & Acuña Espinoza, E. (2025). Code to the Paper Are deep learning models in hydrology entity aware? https://github.com/bheudorfer/2025_entity_awareness_hydrology_camels_us

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. https://doi.org/10.1162/neco.1997.9.8.1735

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (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

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., & Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, *24*(3), 1081–1100. https://doi.org/10.5194/hess-24-1081-2020

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, *47*(13), e2020GL088229. https://doi.org/10.1029/2020GL088229

Jr Hodges, J. (1958). The significance probability of the Smirnov two-sample test. *Arkiv för matematik*, *3*(5), 469–486. https://doi.org/10.1007/bf02589501

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., et al. (2022). Uncertainty estimation with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, *26*(6), 1673–1693. https://doi.org/10.5194/hess-26-1673-2022

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. (2019). 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. (2019). 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

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in great britain: A comparison of Long Short-Term Memory (LSTM)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, *25*(10), 5517–5534. https://doi.org/10.5194/hess-25-5517-2021

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., et al. (2022). Hydrological concept formation inside Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *26*(12), 3079–3101. https://doi.org/10.5194/hess-26-3079-2022

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., et al. (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

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

Loritz, R., Hassler, S. K., Jackisch, C., Allroggen, N., Van Schaik, L., Wienhöfer, J., & Zehe, E. (2017). Picturing and modeling catchments by representative hillslopes. *Hydrology and Earth System Sciences*, *21*(2), 1225–1249. https://doi.org/10.5194/hess-21-1225-2017

Loritz, R., Wu, C. H., Klotz, D., Gauch, M., Kratzert, F., & Bassiouni, M. (2024). Generalizing tree–level sap flow across the European continent. *Geophysical Research Letters*, *51*(6), e2023GL107350. https://doi.org/10.1029/2023GL107350

Ma, K., Feng, D., Lawson, K., Tsai, W., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents – Leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, *57*(5), e2020WR028600. https://doi.org/10.1029/2020WR028600

Maier, H. R., Zheng, F., Gupta, H., Chen, J., Mai, J., Savic, D., et al. (2023). On how data are partitioned in model development and evaluation: Confronting the elephant in the room to enhance model generalization. *Environmental Modelling and Software*, *167*, 105779. https://doi.org/10.1016/j.envsoft.2023.105779

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<3237:ALTHBD>2.0.CO;2

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. https://doi.org/10.1016/0022-1694(70)90255-6

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., et al. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, *627*(8004), 559–563. https://doi.org/10.1038/s41586-024-07145-1

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. 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. https://doi.org/10.1175/jhm-d-16-0284.1

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al., (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Reichert, P., Ma, K., Höge, M., Fenicia, F., Baity-Jesi, M., Feng, D., & Shen, C. (2024). Metamorphic testing of machine learning and conceptual hydrologic models. *Hydrology and Earth System Sciences*, *28*(11), 2505–2529. https://doi.org/10.5194/hess-28-2505-2024

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., et al. (2003). 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

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

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (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*, *117*(D3), 2011JD016048. https://doi.org/10.1029/2011JD016048

Yang, Y., & Chui, T. F. M. (2021). Reliability assessment of machine learning models in hydrological predictions through metamorphic testing. *Water Resources Research*, *57*(9), e2020WR029471. https://doi.org/10.1029/2020WR029471

Zhang, M., Liu, N., Harper, R., Li, Q., Liu, K., Wei, X., et al. (2017). A global review on hydrological responses to forest change across multiple spatial scales: Importance of scale, climate, forest type and hydrological regime. *Journal of Hydrology*, *546*, 44–59. https://doi.org/10.1016/j.jhydrol.2016.12.040