https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


# Benefits of multi-target and self-supervised LSTM models for water quantity and quality

Jean-Paul Brede[1], Pia Ebeling[2], Jens Kiesel[3], Ralf Loritz[1]

[1] Institute for Water and Environment, Karlsruhe Institute of Technology (KIT), Karlsruhe 76131, Germany
[2] Department of Hydrogeology, Helmholtz Centre for Environmental Research – UFZ, Leipzig 04318, Germany
[3] Stone Environmental, 535 Stone Cutters Way, 05602 Montpelier (VT), USA

*Correspondence to*: Jean-Paul Brede (jean-paul.brede@kit.edu) and Ralf Loritz (ralf.loritz@kit.edu)

**Abstract.** Deep learning has become a standard tool for streamflow modeling, but its application to water quality remains challenging due to sparse, irregular, and noisy in-situ observations. Yet, water-quality variables are tightly linked to discharge and to each other through shared hydrological and biogeochemical controls, suggesting that jointly modeling water quantity and quality may help compensate for limited data availability. In this study, we compare single-target, multi-target, and self-supervised LSTM models for the joint simulation of discharge and six water-quality variables ($NO_3$–N, $PO_4$–P, DO, DOC, EC, and WT) across 408 German catchments. Our results highlight that extending the $_{\text{baseline}}$LSTM from single-target discharge prediction to jointly predicting discharge and all six water-quality variables does not substantially degrade discharge performance (across all baseline configurations, median KGE' = 0.84–0.87) and yields median KGE' values between 0.35 ($PO_4$–P) and 0.94 (WT) for the water-quality targets. Interestingly, learning discharge as a co-target consistently outperforms models that use observed discharge as an additional input, indicating that jointly learning water quantity and quality is more effective than using discharge as a predictor. A variable-averaged loss function is key to balance the strongly uneven observation densities of discharge and water-quality variables. Building on this multi-target framework, we further explore whether an alternative training strategy based on self-supervised learning can better exploit the incomplete and heterogeneous nature of environmental observations. Our evaluation reveals that the $_{\text{self-supervised}}$LSTM yields a level of predictive skill comparable to the multi-target supervised $_{\text{baseline}}$ under inference conditions restricted to meteorological drivers, while effectively leveraging cross-variable dependencies to enhance $PO_4$–P and DO reconstructions when contextual water-quality data are provided. Besides showcasing the strong performance of LSTMs in water-quality simulations with sparse and irregular data, our results demonstrate that multi-target learning provides an effective framework for coupled water quantity–quality modeling, while self-supervised learning offers additional flexibility for exploiting incomplete and heterogeneous environmental observations, yielding predictive skill comparable to, and in some cases exceeding, that of calibrated process-based water-quality models and supporting the use of LSTMs as a scalable alternative for regional water-quality simulation in data-sparse environments.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


## 1 Introduction

Deep learning (DL) has become an effective approach for hydrological modeling, offering a data-driven complement to traditional process-based methods. Particularly when trained on large-sample datasets, DL models can provide reliable water quantity estimates that often match or exceed the performance of conceptual benchmarks (e.g., Acuña Espinoza et al., 2024; Kratzert et al., 2019). Reflecting the field's rapid progress, these advancements are already being integrated into regional and global systems for operational flood and drought forecasting (Nearing et al., 2024). In contrast, the uptake of DL in water-quality (WQ) modeling has been slower (Zhi et al., 2024), partially due to the nature of the available data. Unlike water-quantity variables such as discharge or groundwater levels, which are typically observed with a high frequency and automated, water-quality measurements are often sparse and irregular in time. Many observations are based on grab sampling, which can miss short-lived but extreme concentration peaks (Rathjens et al., 2023). In addition, measurement uncertainties are often higher due to laboratory analyses, detection limits, and sample handling. As a result, the limited temporal density, uneven coverage and measurement uncertainties constrain the amount of information available about relevant processes for model training and evaluation, which poses a key challenge for any data-driven model.

Despite these challenges, the complexity of water-quality dynamics may also create opportunities for DL-based approaches (Zhi et al., 2024). Water-quality processes are governed by hydrological, biogeochemical, and anthropogenic interactions that vary across space and time, with solute transport and transformation controlled by factors such as land use (Vaughan et al., 2017), soil or storage-related properties (Musolff et al., 2015), as well as flow paths, temperature, redox conditions, and biological activity (McClain et al., 2003). These interactions create strong dependencies between variables, for example, nitrate, iron, phosphorus, and dissolved organic carbon are linked by redox conditions (Musolff et al., 2017), while temperature and discharge shape seasonal and event-driven dynamics in concentrations (Ebeling et al., 2021; Zhi et al., 2023). For instance, nitrate concentrations often increase with streamflow during seasonal wet periods and storm events due to leaching and runoff (Ebeling et al., 2021; Minaudo et al., 2019), while dissolved oxygen is closely linked to water temperature with seasonal minima during warm summer, low-flow conditions with high biological activity (Zhi et al., 2023). In addition, retention processes in soils and aquifers introduce memory effects and time lags that obscure direct cause–effect relationships (Ehrhardt et al., 2021; Van Meter et al., 2016). Furthermore, water-quality observations typically include a wide range of variables grouped into, e.g., physical, chemical, nutrients, pesticides, metals, and biological, that are tightly linked through shared hydrological flow paths and biogeochemical controls. Although individual time series are often sparse and asynchronous, their interdependence implies that they contain substantial structural information about the coupled system. While such process complexity is difficult to represent and parameterize in traditional process-based models, it aligns well with the strengths of DL, which can learn nonlinear relationships and cross-variable dependencies directly from data. These relationships suggest that jointly modeling multiple variables can increase the effective information content of the data. Realizing this potential, however, requires models that explicitly account for the distinct environmental drivers of each constituent and the interactions between the variables (Xia et al., 2025).

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


Taken together, these characteristics suggest that jointly modeling hydrological states and water-quality variables may provide a promising path forward (Fohrer et al., 2022). By explicitly linking water quantity and quality within a unified modeling framework, it becomes possible to exploit their shared controls and interactions, effectively increasing the information available for learning even under sparse observation conditions. In particular, multi-target approaches that leverage both continuous hydrological signals (e.g., discharge) and irregular water-quality observations can help constrain model behavior and improve the representation of coupled processes (Zhang et al., 2016). Such approaches align naturally with DL architectures that are able to learn shared latent representations across variables and scales. As a result, integrating water quantity and quality in a combined modeling framework may not only mitigate current data limitations but also enable more consistent and statistically robust predictions of hydro-biogeochemical dynamics (Ouyang et al., 2025; Xia et al., 2025).

A promising, yet still underexplored, extension in this context is self-supervised learning (SSL), which has recently been applied to hydrological variables such as soil moisture and streamflow (Ghosh et al., 2022; Wang et al., 2025) but has not yet been adopted for in-situ riverine WQ modeling (potentially due to the scarcity of the observations themselves). Rather than relying exclusively on predefined target variables, SSL leverages the multivariate structure of hydrological data by training models to reconstruct, predict, or contrast partially observed inputs (Liu et al., 2021). In doing so, the model is forced to learn the joint statistical dependencies and temporal dynamics across variables. For instance, reconstructing masked nitrate concentrations from meteorological forcings, discharge, water temperature, and dissolved oxygen requires the model to internalize hydro-biogeochemical relationships rather than fitting a single target from predefined input features. This formulation effectively transforms the learning problem into a structured inference task, where missing information must be inferred from the remaining context. As a result, SSL yields representations that are intrinsically compatible with incomplete and irregularly sampled data, as missing values can be handled directly within the training objective through masking rather than requiring explicit imputation, dedicated missing-value embeddings, or additional preprocessing steps. Beyond predictive performance, this makes SSL particularly attractive as a diagnostic tool, enabling the systematic exploration of process-level relationships and hypotheses within complex hydrological systems.

In this study, we first investigate how the performance of an LSTM-based model changes as additional target variables are incorporated into the learning task. Starting from single-target models for discharge and six individual water-quality variables, we progressively increase the level of integration by jointly simulating pairs of variables and, ultimately, the full set of hydrological and water-quality targets within a multi-target framework. This hierarchical setup allows us to assess whether and how multi-target learning affects predictive skill and the model's ability to exploit cross-variable dependencies and interactions. Building on the best-performing multi-target configuration, we then evaluate the effect of SSL, testing whether SSL-based training strategies can further improve representation learning and downstream predictions, and discussing their advantages and limitations.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


## 2 Data

Our study leverages the Catchment Attributes and MEteorology for Large-Sample Studies in Germany (CAMELS-DE) dataset (Loritz et al., 2024), supplemented by stream water chemistry data from the QUality, DIscharge and Catchment Attributes for large-sample studies in Germany (QUADICA) dataset (Ebeling et al., 2026). We divided the time series into three distinct time periods for training (2000-10-01 to 2010-09-30), validation (1990-10-01 to 1995-09-30) and testing (2010-10-01 to 2019-09-30). Training and testing are the time periods with the highest number of measurements in the dataset. From the 1386 monitoring stations present in QUADICA, 554 are matched to catchments in the CAMELS-DE dataset (Ebeling et al., 2026). Applying additional data availability criteria, a catchment is kept only if the dataset contains sufficient WQ observations in both the training period and the test period (validation period was only used to identify hyperparameters). Specifically, at least three different WQ variables ($NO_3$–N, $PO_4$–P, DO) must each have 10 or more measurements within the training and testing period respectively, ensuring that the basin contains a minimal amount of multivariate water-quality information for both model training and evaluation. The resulting dataset contains 408 catchments distributed over Germany (Fig. 1).

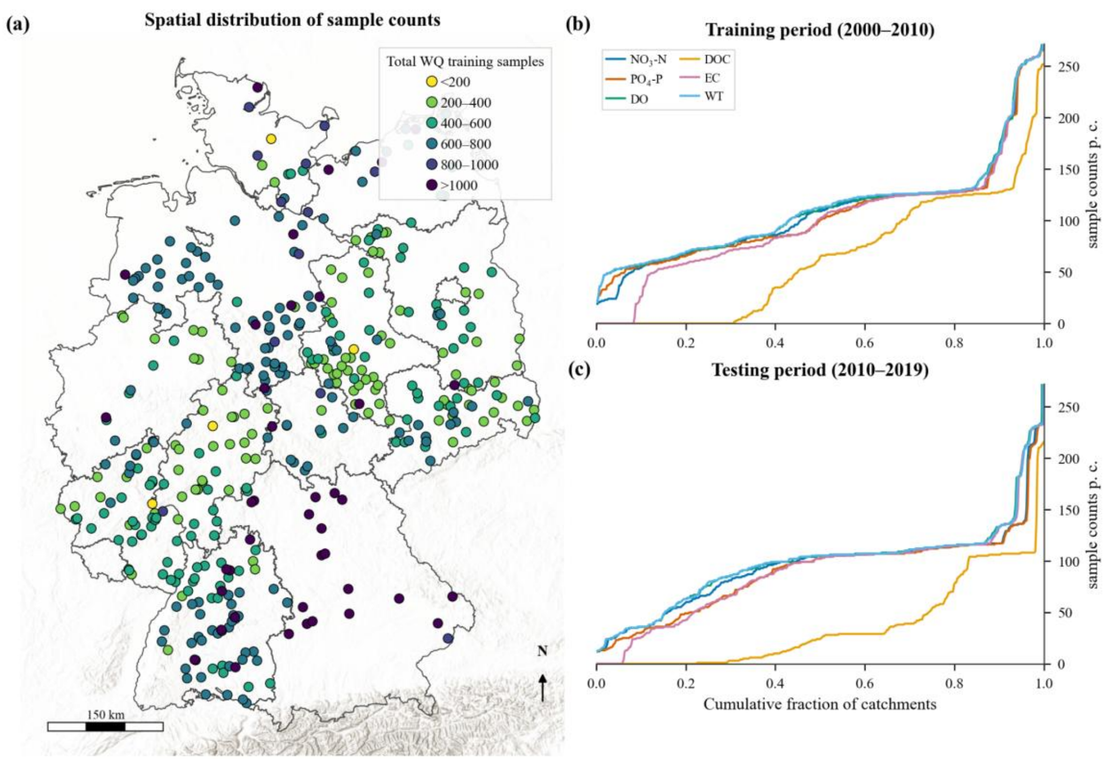



**Figure 1**. WQ data availability across the 408 selected CAMELS-DE catchments. **(a)** Spatial distribution of all WQ grab sample counts per gauging station during the training period (2000–2010). **(b, c)** Cumulative distribution functions for the sample counts per catchments during the training & testing periods. Appendix C contains further metrics for data distributions. Basemap: Esri World Hillshade (sources: Esri, USGS, NOAA). Federal state boundaries: © GeoBasis-DE / BKG.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


### 2.1 Meteorological forcing, catchment attributes and streamflow from CAMELS-DE

To simulate streamflow and water-quality dynamics, the models require information on both the temporal evolution of meteorological conditions and the static characteristics of the catchments. Meteorological forcings provide the time-varying drivers of hydrological and biogeochemical processes, while static catchment attributes describe landscape properties that influence how catchments respond to these forcings.

The meteorological forcing used in this study includes daily mean, minimum, and maximum precipitation (P, mm) and temperature (T, °C), as well as global radiation (R, W $m^{-2}$) and relative humidity (H, %). The data was taken from the CAMELS-DE large sample dataset where HYDrometeorological RASterdata (HYRAS), provided by the Deutscher Wetterdienst (DWD, German Weather Service), was extracted and spatially aggregated to the catchments. The timeseries are gap-free for both the training and testing periods. For streamflow, we used the specific observed discharge (Q, mm $d^{-1}$) from CAMELS-DE, which also contains no gaps for the selected period and catchments. In addition to these dynamic variables, 17 static attributes are provided to all models as time-invariant inputs. These attributes represent differences in catchment topography, soil properties, land cover, and climatic conditions. This set follows the static attributes selected for the baseline model in CAMELS-DE (Loritz et al., 2024).

### 2.2 Water quality data from QUADICA

We used the grab samples from the QUADICA dataset, containing non-aggregated measurements. From this, we select six relevant modelled WQ variables including the nutrients nitrate ($NO_3$–N, mg $L^{-1}$) and phosphate ($PO_4$–P, mg $L^{-1}$), dissolved oxygen (DO, mg $L^{-1}$), dissolved organic carbon (DOC, mg $L^{-1}$), electrical conductivity (EC, µS $cm^{-1}$), and water temperature (WT, °C). In contrast to the daily streamflow observations (Section 2.1), WQ variables are sparsely sampled. WT and $NO_3$-N are the most frequently measured variables, with an average temporal coverage of approximately 3% relative to the continuous streamflow observations (Tab. 1). DOC has the lowest sample count, with 25,968 measurements in the training period. The median training sample count for the WQ variables per catchment is between 101 and 113, except for DOC which has a median sample count of 88. The maximum sample counts per variable within a single catchment generally do not exceed 273. The minimum number of training samples per catchment for the variables used in the catchment selection procedure ($NO_3$–N, $PO_4$–P, and DO) were 19, 21, and 21, respectively, exceeding the threshold of 10 observations imposed during dataset filtering. For most variables, 77–81% of catchments have a mean sampling interval of 20–50 days, with approximately 10% sampled more frequently (10–20 days) and 10% more sparsely (>50 days). DOC deviates from this pattern, with fewer catchments in the frequent category (6%) and more in the sparse category (17%).

**Table 1.** Number of available WQ samples per variable and sample split period across all catchments. Coverage (%) is calculated relative to the theoretical maximum (408 catchments × number of days in period). For EC and DOC only 373 and 268 catchments respectively contain sufficient samples during the training period.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


| Period | $NO_3$-N | $PO_4$-P | DO | DOC | EC | WT |
|---|---|---|---|---|---|---|
| **Training** | 44,469 (3.0%) | 43,700 (2.9%) | 49,276 (3.3%) | 25,968 (1.7%) | 44,295 (3.0%) | 46,420 (3.1%) |
| **Testing** | 39,476 (2.9%) | 36,637 (2.7%) | 43,701 (3.3%) | 14,888 (1.1%) | 39,572 (3.0%) | 40,915 (3.1%) |

## 3 Models

In this study, we train ensembles of supervised ($_{\text{baseline}}$LSTM) and self-supervised ($_{\text{self-supervised}}$LSTM) LSTMs on the same training data and compare them on the same testing data. For all model configurations, parameters are optimized during training using the mean squared error (MSE) loss. Hyperparameters are selected based on model performance on an independent validation period using the modified Kling–Gupta Efficiency (KGE′; Kling et al., 2012). Final model performance is then evaluated exclusively on the testing period using KGE′. To account for uncertainty arising from random parameter initialization, each configuration is trained as an ensemble and ensemble performance is computed as the mean prediction across ensemble members.

### 3.1 $_{\text{baseline}}$LSTM configuration, targets, and loss functions

We train an ensemble (N=5) of regional $_{\text{baseline}}$LSTMs (per configuration) that are commonly applied in rainfall-runoff modelling (e.g. Arsenault et al., 2023; Kratzert et al., 2019) and also increasingly used in WQ modelling (e.g. Rahmani et al., 2021; Saha et al., 2023; Zhi et al., 2024, 2021). We explored ranges of hyperparameter settings based on previous hydrological studies (Loritz et al., 2024) and found that $_{\text{baseline}}$LSTM performance was relatively stable across a wide range of hyperparameters. We chose hyperparameter settings from Loritz et al. (2024), with modifications made to (a) an increased hidden size (512), (b) an increased number of epochs (50), and (c) the learning rate (Tab. 2). We change the learning rate and number of epochs to make the comparison of training schemes as similar as possible to the $_{\text{self-supervised}}$LSTM (see section 3.2). The remaining hyperparameters are displayed in Table 2.

We consider eight dynamic meteorological features at daily resolution, together with 17 static attributes (see Section 2.1), as input features to the $_{\text{baseline}}$LSTM. As targets, we train different ensemble configurations to assess how model performance is affected by the number of prediction targets and the coupled simulation of WQ and quantity. Specifically, we train ensembles of $_{\text{baseline}}$LSTMs: (i) one ensemble predicting only discharge (Q), (ii) six different ensembles, each jointly predicting discharge and a single WQ variable (Q + $NO_3$-N, Q + DO, etc.), (iii) six different ensembles taking Q as additional input and each model instance trained on a single WQ variable, (iv) same as (iii) but Q excluded from the input, and (v) two ensembles predicting discharge together with all WQ variables (Q + all WQ). This follows the principle of multi-task learning (Caruana, 1997), where training on related targets can improve generalization through shared representations.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


Given this multi-target setup and the strong imbalance in sample coverage between discharge (100%) and the WQ variables (1.7–3.3%), the choice of the loss function becomes critical. We therefore compare two loss formulations: a standard sample-averaged Mean Squared Error ($_{\text{sample-averaged}}$MSE) and a variable-averaged MSE ($_{\text{variable-averaged}}$MSE). The $_{\text{sample-averaged}}$MSE is computed over all available samples in the batch, irrespective of the variable, such that variables with more observations contribute more strongly to the loss. In contrast, the $_{\text{variable-averaged}}$MSE is computed by first calculating the MSE separately for each variable, and then averaging these per-variable errors across variables. This ensures that each variable contributes equally to the batch loss, preventing frequently observed variables (e.g., discharge) from dominating the optimization. A similar issue was addressed in Sadler et al. (2022) via a manually tuned scaling factor on the auxiliary loss. The $_{\text{variable-averaged}}$MSE avoids this additional hyperparameter by averaging per-variable errors directly.

### 3.2 $_{\text{self-supervised}}$LSTM configuration, targets, and loss functions

Similar to the $_{\text{baseline}}$LSTMs, we train an ensemble (N = 5) of regional $_{\text{self-supervised}}$LSTM models. The overall architecture remains largely unchanged (a single LSTM layer that performs temporal feature extraction, followed by a linear output head that maps hidden states to predictions). However, a key modification is that the six WQ variables, together with Q, are not only treated as prediction targets but, in addition to the eight dynamic meteorological features, are also incorporated as dynamic input features, resulting in a total of 15 dynamic input variables and seven targets. The self-supervised training is based on a masked reconstruction objective inspired by masked autoencoder approaches (e.g., He et al., 2021; Li et al., 2023), adapted here to sparse and irregularly sampled multivariate time series. During training, the model operates on a masking-based reconstruction task. For each training sequence, the seven hydrological and water-quality variables (Q, $NO_3$–N, $PO_4$–P, DO, DOC, EC, and WT) are randomly partitioned into two groups. One group is retained as observed input, while the remaining variables are masked by replacing their values with a learned scalar mask token. This token is initialized at zero and perturbed with small additive Gaussian noise (scaled by 0.01). In contrast, meteorological forcing variables are always provided as fully observed inputs and are never masked nor used as reconstruction targets. The number of masked variables per sequence is controlled by a hyperparameter, the masking rate, which is set to 3. This value was determined through hyperparameter tuning and means that, for each individual sequence in a batch, exactly three variables are fully masked (see also Discussion). As a result, variables alternate between serving as contextual inputs and reconstruction targets throughout training. This masking strategy enables the model to explicitly leverage cross-variable dependencies. When a subset of variables is observed, it provides contextual information that supports the reconstruction of the masked variables, encouraging the model to learn coherent relationships between water quantity and water-quality dynamics. At the same time, this formulation accommodates missing data, as unobserved values can be treated as masked inputs, allowing the model to be trained and applied without requiring explicit imputation or gap-filling. Model training, hyperparameter selection, and evaluation follow the same procedure as for the $_{\text{baseline}}$LSTM models. The only difference is the masking-based learning objective, allowing the impact of the self-supervised formulation to be assessed independently of changes in architecture or training data.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


During training, we used a fixed masking rate of three variables per sequence. During evaluation, however, we apply different masking configurations to investigate how well the learned representations can reconstruct variables under varying levels of information availability. We evaluate the $_{\text{self-supervised}}$LSTM under two inference scenarios. In the single-variable masking scenario, all variables except one (Q or a WQ variable) are provided as inputs, and model performance is evaluated on the withheld variable. In the multi-variable masking scenario, only the meteorological forcings and static catchment attributes are provided, while all hydrological and water-quality variables are masked. Model performance is then evaluated separately for each reconstructed target variable.

The single-variable masking scenario assesses the model's ability to exploit cross-variable dependencies when extensive contextual information is available. In contrast, the multi-variable masking scenario represents the most challenging reconstruction setting and mirrors the multi-target prediction setting used for the supervised models, where predictions are generated solely from meteorological forcings and static catchment attributes. Comparing both scenarios therefore provides insight into the extent to which predictive performance originates from information contained in the meteorological forcings versus information shared among hydrological and water-quality variables.

The hidden size is increased to 1024, while all remaining hyperparameter settings of the $_{\text{self-supervised}}$LSTM are identical to those of the $_{\text{baseline}}$LSTM configuration (Tab. 2). During training, model parameters are optimized using the $_{\text{variable-averaged}}$MSE calculated only over the masked variables.

**Table 2.** Overview of model architecture, training configuration, and input/output variables. Dynamic inputs were exclusively used as inputs opposed to the targets which were used as inputs and as targets.

| Category | Parameter | Value |
|---|---|---|
| **Architecture** | Model | LSTM supervised / self-supervised |
| | Hidden size | 512 / 1024 |
| | Dropout | 0.3 |
| | Forget gate bias | 3.0 |
| | Layers | 1 |
| **Training** | Epochs | 50 |
| | Batch size | 256 |
| | Sequence length | 365 d |
| | Learning rate (scheduled) | 1e−3 → 5e−4 (ep. 15) → 1e−4 (ep. 30) → 5e−5 (ep. 40) |
| | Loss function | MSE (sample or variable averaged) |
| **Dynamic inputs** | Precipitation (mean, min, max) | mm $d^{-1}$ |

| | | |
|---|---|---|
| (exclusive) | Temperature (mean, min, max) | °C |
| | Humidity (mean) | % |
| | Global radiation (mean) | W $m^{-2}$ |
| **Static inputs** | Topography | area ($km^2$), elev_mean (m a.s.l.) |
| | Soil texture | clay, sand, silt 0–30 cm (%) |
| | Land cover | artificial, agricultural, forest, wetlands, water (%) |
| | Climate indices | p_mean, p_seasonality, frac_snow, high/low_prec_freq, high/low_prec_dur |
| **Targets** | Specific discharge | mm $d^{-1}$ |
| (inclusive) | $NO_3$-N, $PO_4$-P, DO, DOC | mg $L^{-1}$ |
| | EC | µS $cm^{-1}$ |
| | WT | °C |

## 4 Results

In section 4.1 we show the $_{\text{baseline}}$LSTM and analyze how coupled simulations of WQ and quantity affect overall model performance, specifically assessing whether multi-task learning leads to improvements or degradations, and how these effects depend on sample frequency and the number of available measurements. In section 4.2 we focus on assessing the self-supervised models, technical challenges and potentials for further work.

### 4.1 Coupled water quality and quantity simulations using the $_{\text{baseline}}$LSTM

To assess how supervised LSTMs, commonly among the best-performing DL architectures in rainfall–runoff modelling (Liu et al., 2025), respond when predicting multiple targets instead of only streamflow, we evaluate whether their performance declines under multi-task settings or whether they may benefit from shared representations. In this context, all baseline models (Section 3.1) achieve high runoff prediction skill, with median KGE' values between 0.84 and 0.87 (median NSE $\in$ [0.87, 0.89]), consistent with previously reported streamflow performance in Germany (Loritz et al., 2024). Extending the model from predicting only discharge to jointly predicting Q and WQ variables leads to only a small decrease in discharge performance ($\Delta$KGE' $\leq$ 0.03). Overall, this suggests that the $_{\text{baseline}}$LSTM is sufficiently expressive to support multi-target learning without a significant loss in discharge prediction skill. This is consistent with findings from continuous multi-target settings (Ouyang et al., 2025; Sadler et al., 2022). However, the strong imbalance in temporal coverage between discharge

(≈100%) and WQ variables (≈1.7–3.3%) introduces an additional trade-off, which is primarily governed by the choice of the loss formulation.

The dominant factor influencing WQ predictions in the $_{\text{baseline}}$LSTM is the choice of the loss formulation. Switching from the $_{\text{sample-averaged}}$MSE to the $_{\text{variable-averaged}}$MSE leads to a slight decrease in runoff performance (median KGE from 0.87 to 0.84), but consistently improves the prediction of all WQ variables (ΔKGE' ∈ [0.01, 0.06]), except for DO and $PO_4$ (ΔKGE' < 0.01). Henceforth when referencing the multi-target $_{\text{baseline}}$LSTM (Q + 6 WQ), the version that was trained on the $_{\text{variable-averaged}}$MSE loss is meant. Best performance for all WQ variables, except dissolved oxygen (DO), is achieved when training the model jointly on discharge and the respective WQ variable as targets (Fig. A1). Notably, this setup outperforms models that include discharge as an input feature together with the meteorological forcing. This suggests that learning discharge as a co-target helps shape more informative latent representations, which in turn benefits WQ predictions. While the performance gains are modest, they are consistent across all WQ variables (ΔKGE' ∈ [0.01, 0.05]) except for DO (ΔKGE' < 0.01). In contrast, treating discharge merely as an additional input feature does not yield comparable improvements, with performance remaining largely unchanged relative to models driven only by meteorological forcings, except for minor gains in $PO_4$–P and EC.

The predictive performance of the multi-target $_{\text{baseline}}$LSTM (Q + 6 WQ) across the full suite of variables is summarized in Fig. 2. While Q and WT are reconstructed with high fidelity, achieving median KGE' values of 0.84 and 0.94, respectively, the accuracy for chemical constituents exhibits substantial variability. The model demonstrates moderate to high skill for DO (median KGE' = 0.71), DOC (median KGE' = 0.62), and $NO_3$–N (median KGE' = 0.60), whereas $PO_4$–P remains the most challenging target to reproduce, with a median KGE' of only 0.35. The spatial distribution of these metrics (Fig. 2) reveals distinct regional patterns across the study area. For $NO_3$–N (Fig. 2a) and DO (Fig. 2b), the model performs inconsistently across federal states, with notably higher KGE' values observed in northern and eastern Germany. In contrast, the performance for $PO_4$–P (Fig. 2c) is markedly lower and more heterogeneous, characterized by poor predictive skill (KGE' < 0.2) in numerous catchments, particularly in central and southwestern regions. However, we could not identify a clear set of catchment or climatic characteristics that consistently explained the observed spatial patterns, suggesting that multiple interacting effects may be involved.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026

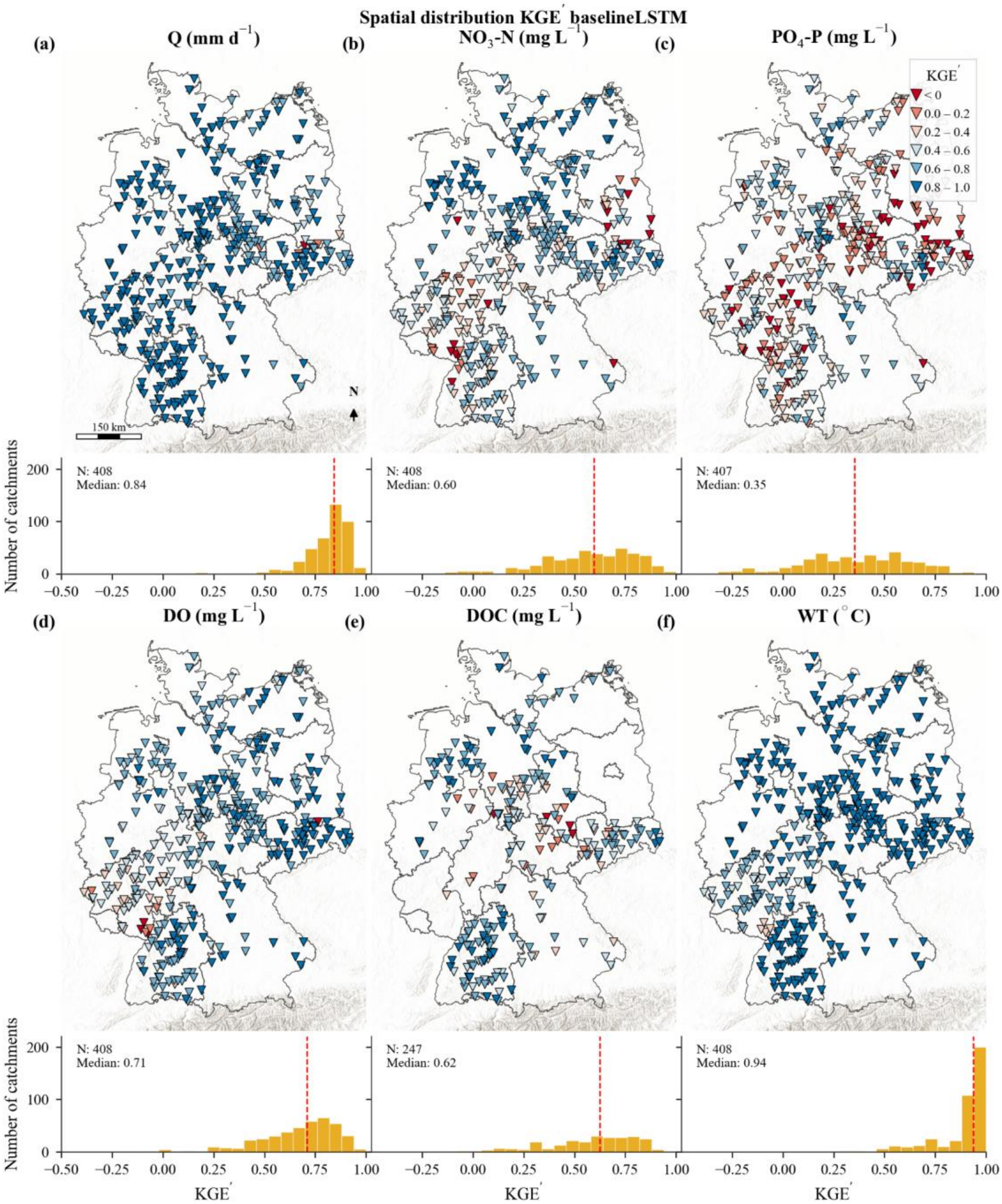




**Figure 2.** Predictive performance of the multi-target $_{\text{baseline}}$LSTM (Q + 6 WQ, variable-averaged loss) on the test period (2010–2019) across 408 catchments. (a–c) Spatial distribution of KGE' for $NO_3$-N, DO, and $PO_4$-P. Symbol colour indicates KGE' magnitude according to the legend; (d–i) Frequency distributions of KGE' across all catchments for Q, $NO_3$-N, $PO_4$-P, DO, DOC, and WT. Dashed red lines indicate the median KGE' across all 408 catchments. EC is not shown due to figure layout constraints and is presented separately in the Appendix. Basemap: Esri World Hillshade (sources: Esri, USGS, NOAA). Federal state boundaries: © GeoBasis-DE / BKG.

Finally, we examine the relationship between data availability and model performance. As expected for data-driven models, predictive skill generally depends on the amount of available training data; however, this relationship is weak and not consistently expressed across variables. Pearson correlations between the number of samples per catchment and variable and their respective KGE' values are low ($\rho < 0.31$; Cohen, 2013). When considering Spearman rank correlations, only $PO_4$–P shows a moderate relationship ($0.36 < \rho < 0.39$; Cohen, 2013). This effect becomes more apparent when comparing groups of basins: the median KGE' for $PO_4$–P increases substantially ($\Delta$KGE' ∈ [0.43; 0.58]) from basins with average sampling intervals of 10–20 days (n = 38) to those with less frequent sampling (≥ 50 days, n = 43). The observed relationships are not attributable to a strong spatial clustering of highly sampled catchments, as sampling densities are distributed across the study area. Furthermore, correlations between test-period sample counts and KGE′ are weak ($\rho < 0.1$ for most variables), suggesting that the reported performance patterns are not primarily an artifact of differing evaluation sample sizes. Overall, while increased data availability is clearly important for DL applications and can improve performance it does not uniformly explain model skill, indicating that other factors, such as process complexity in a given catchment and data noise, also play a significant role (Jeung et al., 2026; Schauer et al., 2025; Xia et al., 2025). This interpretation is supported by a bootstrapping analysis of the test data, in which we resample the observations 1,000 times with replacement and recompute KGE′ for each realization. The resulting performance estimates remain stable across catchments, suggesting that model skill is not overly sensitive to individual observations and extremes.

### 4.2 Coupled water quality and quantity simulations using the $_{\text{self-supervised}}$LSTM

Under the fully masked inference scenario, where both WQ variables and discharge are masked during testing, leaving only meteorological forcings and static catchment attributes as input (multi-masking scenario), the $_{\text{self-supervised}}$LSTM achieves performance broadly comparable to the multi-target $_{\text{baseline}}$LSTM (Fig. 3). Most variables show differences of ≤ 0.01 in median KGE'. The most notable deviations are observed for $NO_3$–N ($\Delta$KGE' = −0.02) and EC ($\Delta$KGE' = −0.02), indicating a modest tendency for the SSL model to underperform the supervised baseline when no observational context is available at inference time. However, according to a paired bootstrap analysis across catchments, only the reduction in $NO_3$–N performance is statistically significant.

Under the single-masking scenario, where all but one hydrological or water-quality variable are provided to the $_{\text{self-supervised}}$LSTM as contextual information, clear improvements emerge for two variables (Fig. 3). $PO_4$–P shows the largest gain ($\Delta$KGE' = +0.05), followed by DO (+0.02), while most other variables remain effectively unchanged, with noteworthy improvements in the lower-performing catchment percentiles for DO and WT. A paired bootstrap analysis confirms that the improvements for $PO_4$–P and DO are statistically significant, whereas differences for the remaining variables are generally not distinguishable from sampling variability. These results suggest that the SSL model is able to exploit cross-variable dependencies when contextual information from related variables is available during inference.

Overall, differences between the SSL and baseline models remain small across both evaluation scenarios, indicating that the self-supervised training strategy does not fundamentally alter predictive performance. Nevertheless, the statistically significant improvements for DO and $PO_4$–P under the single-masking scenario demonstrate that the SSL model can leverage information shared among hydrological and water-quality variables when such contextual information is available. Conversely, the largely unchanged performance under the multi-masking scenario suggests that the learned representations do not consistently translate into improved predictions when meteorological forcings constitute the sole source of information.

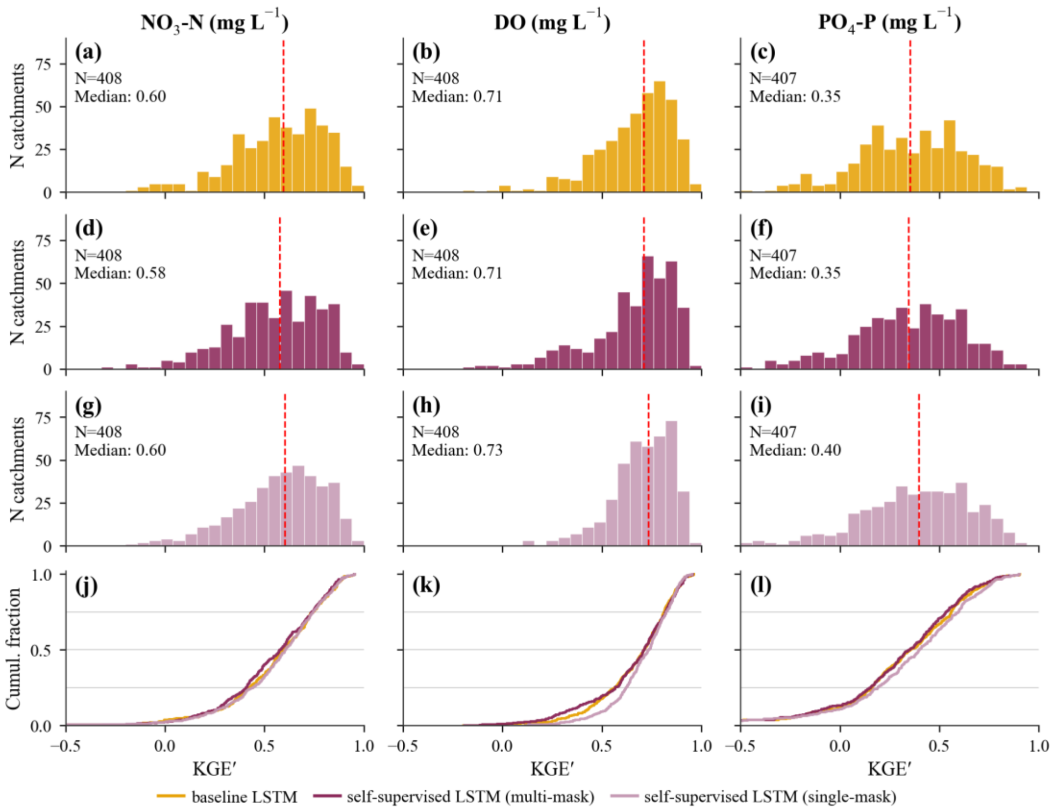



**Figure 3**. (a) - (i): KGE' distributions for $NO_3$-N, DO, and $PO_4$-P across all catchments for the $_{\text{baseline}}$LSTM (target: Q + 6 WQ variables) and two inference scenarios of the $_{\text{self-supervised}}$LSTM. The red dotted line denotes the median of the respective populations. The similarity of the distribution underscores the comparability between the two model types. (j)-(l) Empirical cumulative distribution functions (ECDF) of the same three variables. (k) shows how the $_{\text{self-supervised}}$LSTM outperforms the $_{\text{baseline}}$LSTM in the lower percentiles. Appendix B contains ECDFs for all models and variables.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


## 5 Discussion

The results of this study address two connected questions: 1) whether jointly simulating discharge and multiple water-quality variables in a shared LSTM framework is feasible without compromising predictive skill, and 2) whether SSL can further improve upon this setup. We discuss the first question in section 5.1 and the second question in 5.2.
It must be mentioned that this study uses the modified KGE' (Kling et al., 2012), whereas most comparison studies cited below (Pandit et al., 2025; Xia et al., 2025; Liu et al., 2025) report the original KGE (Gupta et al., 2009). The newer version of the KGE updates the variability ratio term to avoid covariation with the bias ratio. Even though the differences between those two are mostly negligible, the reported values are not directly interchangeable and should be interpreted accordingly, also since sample set, catchment selection and periods differ between studies.

### 5.1 Multi-target learning for coupled water quantity–quality modeling

The model results of the $_{baseline}$LSTM provide several insights into the behavior and potential of multi-target simulation for coupled water quantity–quality modeling. First, we find that extending a LSTM network from single-target discharge prediction to a multi-target setting does not necessarily lead to a deterioration in discharge performance. Across all configurations, discharge skill remains in the range of previously reported benchmarks for Germany (Loritz et al., 2024), indicating that the capacity of this comparatively simple DL model, consisting of a single LSTM layer with a linear output head, is sufficient to accommodate additional targets without sacrificing predictive performance. This is a noteworthy advantage over process-based models, where a streamflow performance trade-off during multi-variable calibration can be more pronounced (Wagner et al., 2025) and in line with other studies that investigated the use of DL for predicting WQ and water quantity (e.g. Sadler et al., 2022).

For nitrate, direct benchmarking remains difficult due to differences in data structure, model setup, and study region. Nevertheless, to provide context for the magnitude of the achieved performance, we compare our results to values reported in the literature. Our multi-target $_{baseline}$LSTM, trained on grab-sample data with approximately 3% temporal coverage, achieves a median KGE' of 0.60 for $NO_3$-N, which falls within a similar performance range with the single-target study by Pandit et al. (2025) despite differences in data density (temporal resolution), model configuration, and geographic setting and outperforms the multi-target supervised LSTM model by Xia et al. (2025), which used a comparable number of catchments and temporal resolution. Their reported median KGE values for nutrient related variables of 0.4 align with our phosphate performance. For dissolved oxygen, Zhi et al. (2023, 2021) present LSTM-based models but do not report KGEs. For the 2021 paper they report a median NSE value of 0.53 for their core evaluation group of 84 basins, which resembles our median NSE value for DO of 0.55. In Zhi et al. (2023) follow-up paper they improved their models performance to a median NSE of 0.76 by training it on more basins (580), a longer period (30 years) and additional input variables like biogeochemical conditions, which alone are attributed a +0.1 NSE increase, and hence not directly comparable with our results. However, Liu et al. (2025), who reproduce

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


these experiments across 236 catchments from Zhi et al. (2021), report a median KGE of approximately 0.70, again consistent with our findings. Water temperature predictions are also in a similar range to previous studies (NSE ≈ 0.96–0.98; Rahmani et al., 2021), although we do not observe their reported improvement when including discharge as an input. For phosphate, our performance (KGE' < 0.4) is comparable to Xia et al. (2025). Nevertheless, in contrast to their findings, nitrate shows higher predictive skill in our setup (KGE' ≈ 0.60–0.62). This overall consistency with the literature suggests that our coupled $_{\text{baseline}}$LSTM, which jointly simulates water quantity and WQ in a multi-target framework, does not compromise predictive performance but it may rather be advantageous when increasing the number of target variables. This contrasts with process-based water-quality models, where adding calibration targets can introduce parameter trade-offs rather than shared gains. For instance, Molina-Navarro et al. (2017) found that a multi-site, multi-variable water quality calibration achieved the best nutrient-load performance when N and P fractions were considered separately, indicating that fully joint calibration can compromise individual-variable performance. For variables such as DOC or electrical conductivity, broader comparisons remain challenging due to the lack of large-sample benchmarks, although individual studies indicate potentially high predictability under well-observed conditions.

An important step in our modelling study was the shift from a sample-averaged ($_{\text{sample-averaged}}$MSE) to a variable-averaged loss ($_{\text{variable-averaged}}$MSE), which introduces a clear but controlled trade-off: a slight decrease in runoff performance is accompanied by consistent improvements across all water-quality variables. This trade-off is consistent with the task-weight sensitivity reported by Sadler et al. (2022), who showed that the optimal balance between primary and auxiliary targets is not universal. The $_{\text{variable-averaged}}$MSE provides a simple solution that ensures equal contribution from all targets, though adaptive weighting strategies may offer further refinement. It must be noted though, that through equally weighting those sparsely observed variables, noise and measurement errors present in them will have a greater impact on the training signal. This highlights that properly balancing the contribution of sparsely observed targets is essential in multi-target settings, especially when target variables exhibit large differences in data availability, noise levels, and the complexity of the underlying processes. In this context, learning discharge as a co-target, rather than using it only as an input, emerges as a particularly effective strategy. For most variables (notably DOC and $NO_3$–N), jointly predicting discharge improves performance compared to models that either exclude discharge or include it only as an input feature. This contrasts with studies that emphasize the importance of discharge primarily as an input (e.g. Rahmani et al., 2021), and suggests that forcing the model to learn discharge internally leads to more informative latent representations. This is consistent with Sadler et al. (2022), who found that multi-task learning outperforms auxiliary-as-input approaches particularly under sparse observation conditions, potentially reflecting a more consistent internal representation of hydrological states (Lees et al., 2022; Ouyang et al., 2025). Interestingly, this effect is not observed for dissolved oxygen, which reflects its stronger dependence on temperature and biological processes rather than direct hydrological forcing (Zhi et al., 2023) compared to DOC and NO3-N concentrations which are largely hydrologically controlled (Ebeling et al., 2021; Saavedra et al., 2022; Zarnetske et al., 2018).

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


The relationship between data availability and model performance is weaker than might be expected and varies across variables, highlighting the difficulty of water-quality prediction. While increased sampling frequency can improve performance, particularly for $PO_4$–P, the overall correlation between sample count and predictive skill remains low, indicating that data quantity alone does not explain model performance. Instead, factors such as measurement noise, sampling bias, and the intrinsic complexity of the underlying processes play an equally important role (Jeung et al., 2026; Xia et al., 2025). This is reflected in the variable-specific and spatial performance patterns across Germany. While discharge and water temperature are reconstructed with high fidelity, performances for chemical constituents are substantially more variable. In particular, phosphorus remains difficult to predict, both in terms of overall performance and spatial consistency. This suggests that phosphorus dynamics are strongly influenced by local controls, including point sources, legacy effects, and biogeochemical interactions in soils and streams, which operate at finer spatio-temporal scales than those represented by the general catchment attributes and large-scale hydroclimatic drivers used in this study (Ebeling et al., 2021; Gu et al., 2017).

By comparison, $NO_3$–N and DOC exhibit a clearer spatial structure, with higher performance in northern Germany. These regions are characterized by agricultural lowland catchments with shallow groundwater systems and high hydrological connectivity, where nutrient concentrations are highly dynamic but strongly linked to discharge patterns. Under wet conditions, shallow solute sources become connected to the stream network, whereas deeper flow paths may transport lower solute concentrations due to longer residence times and retention processes such as denitrification (Ebeling et al., 2021). Conversely, catchments with stronger topographic gradients may be affected by a wider range of event-driven and interacting processes with temporally variable source areas and connectivity patterns (Saavedra et al., 2025; Winter et al., 2024), resulting in more heterogeneous and less predictable water-quality responses.

Consistent with this interpretation, we observe higher model performance in catchments with greater internal solute variability, particularly for $NO_3$–N and DOC, and also for WT. For $NO_3$–N, the correlation between KGE' and the coefficient of variation is approximately 0.3. This indicates that variability itself does not necessarily reduce predictability, provided that it follows structured hydrological controls that can be learned by the model. This relationship is not observed for $PO_4$–P, suggesting that phosphorus variability is more strongly governed by processes that are either insufficiently sampled or not well represented by the available predictors. The improved $PO_4$–P performance in more densely sampled catchments further supports this interpretation, indicating that low-frequency sampling may fail to capture the dominant dynamics of phosphorus transport and transformation. This aligns with concepts of spatial and temporal variability in water-quality constituents, which require monitoring strategies adapted to the intrinsic variability of each constituent to capture the relevant processes (Schauer et al., 2025). Relatedly, the “simplicity index” proposed by Xia et al. (2025) shows that model performance strongly correlates with the extent to which a variable’s dynamics can be explained by linear relationships with discharge and seasonal cycles.

### 5.2 Benefits of self-supervised LSTM training

Regarding the SSL setup, we applied a full-variable masking strategy, i.e., entire hydrological or water-quality variables were masked during training rather than individual time steps or short temporal segments. This choice was motivated by the intended application and the sparse nature of the water-quality observations. During training, three of the seven hydrological and water-quality variables are randomly masked, while the remaining variables remain visible. The model is then required to reconstruct the masked variables from the meteorological forcings and the available hydrological and water-quality information. Because the set of masked variables changes from one training sequence to the next, different combinations of variables repeatedly act as contextual inputs and reconstruction targets throughout training. Over time, this encourages the model to learn statistical dependencies among variables, as successful reconstruction requires exploiting information contained in the observed variables. The goal is that the model gradually learns a shared representation of water quantity and water-quality dynamics that can be used to infer missing variables from the available context.

We selected full-variable masking because it aligns naturally with the inference scenarios considered in this study, where complete variables are unavailable and must be reconstructed from the remaining information. We did compare alternative masking strategies, such as masking individual time steps or short temporal segments but found the full-variable masking worked best in our scenario. Although recommended in Li et al. (2023), such approaches may be more appropriate for applications focused on gap-filling or interpolation of partially observed time series. For example, if high-frequency observations of WT and DO are available while $NO_3$–N, DOC, or $PO_4$–P are sampled less frequently, temporally localized masking could be used to learn the reconstruction of missing observations within otherwise observed time series. The number of masked variables per sequence, as well as the selection of which variables are masked jointly, represents an important hyperparameter and defines a continuum between supervised and self-supervised learning. Masking only a small subset of variables moves the training objective closer to a supervised setting, whereas masking a larger number of variables enforces reconstruction from increasingly limited contextual information. The intermediate masking strategy (no masking of the meteorological variables) applied here represents one point in this design space that we found to perform well during hyperparameter tuning. A more systematic exploration of variable-specific masking rates, training duration, and structured masking strategies, where physically related variables are masked jointly, remains an important direction for future work.

From a technical perspective, SSL simplifies both model development and data handling. A single model can be trained once and subsequently applied across a wide range of input–output configurations without retraining, eliminating the need to build separate supervised models for different target combinations. At the same time, missing values do not require explicit treatment, as masking is an inherent part of the training objective. Consequently, incomplete observations can be incorporated directly into the learning process, avoiding the need for imputation, dedicated missing-value encodings, or additional preprocessing steps (e.g. Gauch et al., 2025). This combination of flexibility and robustness substantially reduces

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


computational cost and streamlines the modeling workflow, which is particularly beneficial in hydrological applications characterized by heterogeneous and incomplete data.

From a methodological perspective, SSL provides a framework to systematically investigate relationships between variables. By reconstructing selected variables from different subsets of inputs, the model is forced to learn the underlying dependencies within the hydro-biogeochemical system. This enables targeted analyses of how information is shared across variables, for example by assessing how well a given variable can be inferred from others or how predictive skill changes when specific inputs are removed. In this sense, SSL is one way to extend the role of DL models beyond prediction, allowing them to be used as tools to probe and interpret the structure of complex environmental systems.

In this study, the SSL-trained model performs similar to its supervised counterpart in terms of benchmark performance, but does not outperform. However, there is emerging evidence that SSL can improve generalization under unseen conditions (Wang et al., 2025), suggesting that its benefits may become more apparent in spatially out-of-sample scenarios. We therefore argue that SSL should not primarily be viewed as a means to maximize predictive skill in conventional benchmarks, but rather as a methodological step toward more flexible, scalable, and generalizable hydrological models. More broadly, the SSL training paradigm aligns conceptually with emerging foundation models for Earth system prediction, in which a single model trained on diverse data is subsequently adapted to new tasks or variables with minimal retraining and little computational cost (Bodnar et al., 2025; Lehmann et al., 2025). Taken together, these findings indicate that the potential of SSL may not be fully reflected by conventional benchmark metrics and should also be evaluated in terms of robustness, transferability, and the ability to learn from heterogeneous environmental observations.

## 6 Conclusion

Our results support four main conclusions: i) there is practically no detrimental effect on coupling streamflow with various water quality variables, with Q prediction skill remaining within $\Delta$KGE' $\leq 0.03$ and $\Delta$NSE $\leq 0.02$ across all configurations. This is further supported by performance levels comparable to contemporary studies, irrespective of whether discharge is used as an input or as a target. ii) including streamflow as target rather than input to any other tested WQ variable except DO increases performances. We conjecture this relates to an improved internal hydrological “understanding” of the LSTM, which could be verified by future probing studies. We therefore strongly recommend the usage of streamflow in WQ modelling as an additional target, when using LSTM architectures. This might be extended to other architectures and warrants testing. iii) unequal coverage between targets must be addressed in the loss formulation during model training. In our study, discharge observations are continuous whereas water-quality measurements are sparse and irregular, creating a strongly unbalanced multi-target setting. Across all tested multi-target LSTM configurations, variable-averaging the per-target MSE consistently improved water-quality predictions at only a minor cost to streamflow performance, providing a simple and hyperparameter-free default for applications with strongly differing observation densities among targets. iv) SSL modeling in hydrology shows

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


no systematic performance benefits over the supervised multi-target baseline under comparable inference conditions (multi-masking), with small reductions for $NO_3$-N and EC ($\Delta$KGE' = −0.02 each) and otherwise near-identical medians. When the SSL model is given full cross-variable context (single-masking), small median improvements emerge for $PO_4$-P ($\Delta$KGE' = +0.05) and DO (+0.02), along with robust gains in the lower performance percentiles for DO and WT. Incorporation of missing data points into the training signal is deemed highly promising, especially for modelling outside of large sample dataset coverage where the meteorological forcings are usually incomplete. Testing for that by noise injection or adding adversarial data to the input data can explore whether SSL makes a LSTM model more robust against uncertainty in the input data. With the additional hyperparameters in SSL only touched on here, there is a huge variety of exploratory avenues to optimize for individual use cases. Overall, multi-target training proves viable, discharge prediction is largely unaffected across all configurations, and targeted adjustments to the loss formulation and target configuration yield consistent improvements in water-quality skill. Self-supervised learning, by contrast, offers practical modeling flexibility that is particularly relevant given the sparse and heterogeneous nature of water-quality observations without compromising predictive performance. This training paradigm enables the user-friendly incorporation of a vast amount of related variables for various inference settings with subsets of input data predicting single targets, multiple targets or simply imputing missing data in a single LSTM model. We believe this is a necessary step forward to attain foundation models that are able to generalize not only spatially and over time but also across hydrological subdomains.

**Code and data availability**

The CAMELS-DE dataset is available at https://doi.org/10.5281/zenodo.13837553 (Loritz et al., 2024). The QUADICA dataset is available at https://doi.org/10.4211/hs.c2866cd416b94ca386deb57588834311f (Ebeling et al., 2026), while due to data use licenses raw data cannot be published and are thus deposited in a long-term institutional repository (https://www.ufz.de/record/dmp/archive/16457, last access: 17th of June 2026).

**Author contributions (CRedit)**

Jean-Paul Brede: Conceptualization, methodology, Software, Formal analysis, Investigation, Validation, Visualization, Writing - original draft, Writing - review & editing.
Pia Ebeling: Conceptualization, Data curation, Writing - review & editing.
Jens Kiesel: Writing - review & editing.
Ralf Loritz: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing - original draft, Writing - review & editing

**Financial Support**

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


The results were funded by the Federal Ministry of Research, Technology and Space (BMFTR) as part of the project "AI-based flood prediction in small river basins in Germany" (KI-HopE-De, 16IS24088A). We are grateful to the Volkswagen Foundation for funding the project "Invigorating Hydrological Science and Teaching: Merging Key Legacies with New Concepts and Paradigms" (ViTamins, 0069685).

**Declaration of competing interests**

One of the co-authors is a member of the editorial board of Earth System Science Data. The authors declare that they have no other known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgements**

The authors acknowledge support by the state of Baden-Württemberg through bwHPC. The authors further gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Research, Technology and Space and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program. HoreKa is partly funded by the German Research Foundation (DFG). This work is additionally supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

**AI Usage**

AI tools were used for script generation, code autocomplete, and plausibility checks. Grammar and spelling checks of the manuscript were also supported by AI. Models used for programming-related tasks were Anthropic's Claude (Sonnet and Opus families) for script generation and Microsoft's Copilot for code autocomplete. Natural language-related tasks were model-agnostic.

**Review statement**

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


## References

Acuña Espinoza, E., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Loritz, R., Ehret, U., 2024. Technical note: An approach for handling multiple temporal frequencies with different input dimensions using a single LSTM cell. EGUsphere 1–12. https://doi.org/10.5194/egusphere-2024-3355

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. Hydrol. Earth Syst. Sci. 27, 139–157. https://doi.org/10.5194/hess-27-139-2023

Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J.A., Dong, H., Gupta, J.K., Thambiratnam, K., Archibald, A.T., Wu, C.-C., Heider, E., Welling, M., Turner, R.E., Perdikaris, P., 2025. A foundation model for the Earth system. Nature 641, 1180–1187. https://doi.org/10.1038/s41586-025-09005-y

Caruana, R., 1997. Multitask Learning. Mach. Learn. 28, 41–75. https://doi.org/10.1023/A:1007379606734

Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Routledge, New York. https://doi.org/10.4324/9780203771587

Ebeling, P., Hubig, A., Wachholz, A., Scharfenberger, U., Haug, S., Nguyen, T., Sarrazin, F., Batool, M., Musolff, A., Kumar, R., 2026. QUADICA v2: extending the large-sample data set for water QUAlity, DIscharge and Catchment Attributes in Germany. https://doi.org/https://doi.org/10.5194/essd-18-691-2026

Ebeling, P., Kumar, R., Weber, M., Knoll, L., Fleckenstein, J.H., Musolff, A., 2021. Archetypes and Controls of Riverine Nutrient Export Across German Catchments. Water Resour. Res. 57, e2020WR028134. https://doi.org/10.1029/2020WR028134

Ehrhardt, S., Ebeling, P., Dupas, R., Kumar, R., Fleckenstein, J.H., Musolff, A., 2021. Nitrate Transport and Retention in Western European Catchments Are Shaped by Hydroclimate and Subsurface Properties. https://doi.org/10.1029/2020WR029469

Fohrer, N., Wagner, P.D., Kiesel, J., Haas, M., Guse, B., 2022. A guideline for spatio-temporal consistency in water quality modelling in rural areas. Hydrol. Process. 36, e14711. https://doi.org/10.1002/hyp.14711

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Cohen, D., Gilon, O., 2025. How to deal w___ missing input data. EGUsphere 1–21. https://doi.org/10.5194/egusphere-2025-1224

Ghosh, R., Renganathan, A., Tayal, K., Li, X., Khandelwal, A., Jia, X., Duffy, C., Neiber, J., Kumar, V., 2022. Robust Inverse Framework using Knowledge-guided Self-Supervised Learning: An application to Hydrology. https://doi.org/10.48550/arXiv.2109.06429

Gu, S., Gruau, G., Dupas, R., Rumpel, C., Crème, A., Fovet, O., Gascuel-Odoux, C., Jeanneau, L., Humbert, G., Petitjean, P., 2017. Release of dissolved phosphorus from riparian wetlands: Evidence for complex interactions among hydroclimate variability, topography and soil properties. Sci. Total Environ. 598, 421–431. https://doi.org/10.1016/j.scitotenv.2017.04.028

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. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2021. Masked Autoencoders Are Scalable Vision Learners. https://doi.org/10.48550/arXiv.2111.06377

Jeung, M., Her, Y., Baek, S.-S., Yoon, K., 2026. Sensitivity of hydrological machine learning prediction accuracy to information quantity and quality. Hydrol. Earth Syst. Sci. 30, 1077–1095. https://doi.org/10.5194/hess-30-1077-2026

Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J. Hydrol. 424–425, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

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. Hydrol. Earth Syst. Sci. 23, 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

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

Lehmann, F., Ozdemir, F., Soja, B., Hoefler, T., Mishra, S., Schemm, S., 2025. Finetuning a Weather Foundation Model with Lightweight Decoders for Unseen Physical Processes. https://doi.org/10.48550/arXiv.2506.19088

Li, Z., Rao, Z., Pan, L., Wang, P., Xu, Z., 2023. Ti-MAE: Self-Supervised Masked Time Series Autoencoders. https://doi.org/10.48550/arXiv.2301.08871

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. Hydrol. Earth Syst. Sci. 29, 6811–6828. https://doi.org/10.5194/hess-29-6811-2025

Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., Tang, J., 2021. Self-supervised Learning: Generative or Contrastive. IEEE Trans. Knowl. Data Eng. 1–1. https://doi.org/10.1109/TKDE.2021.3090866

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 Syst. Sci. Data 16, 5625–5642. https://doi.org/10.5194/essd-16-5625-2024

McClain, M.E., Boyer, E.W., Dent, C.L., Gergel, S.E., Grimm, N.B., Groffman, P.M., Hart, S.C., Harvey, J.W., Johnston, C.A., Mayorga, E., McDowell, W.H., Pinay, G., 2003. Biogeochemical Hot Spots and Hot Moments at the Interface of Terrestrial and Aquatic Ecosystems. Ecosystems 6, 301–312. https://doi.org/10.1007/s10021-003-0161-9

Minaudo, C., Dupas, R., Gascuel-Odoux, C., Roubeix, V., Danis, P.-A., Moatar, F., 2019. Seasonal and event-based concentration-discharge relationships to identify catchment controls on nutrient export regimes. Adv. Water Resour. 131, 103379. https://doi.org/10.1016/j.advwatres.2019.103379

Molina-Navarro, E., Andersen, H.E., Nielsen, A., Thodsen, H., Trolle, D., 2017. The impact of the objective function in multi-site and multi-variable calibration of the SWAT model. Environ. Model. Softw. 93, 255–267. https://doi.org/10.1016/j.envsoft.2017.03.018

Musolff, A., Schmidt, C., Selle, B., Fleckenstein, J.H., 2015. Catchment controls on solute export. Adv. Water Resour. 86, 133–146. https://doi.org/10.1016/j.advwatres.2015.09.026

Musolff, A., Selle, B., Büttner, O., Opitz, M., Tittel, J., 2017. Unexpected release of phosphate and organic carbon to streams linked to declining nitrogen depositions. Glob. Change Biol. 23, 1891–1901. https://doi.org/10.1111/gcb.13498

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T.Y., Weitzner, D., Matias, Y., 2024. Global prediction of extreme floods in ungauged watersheds. Nature 627, 559–563. https://doi.org/10.1038/s41586-024-07145-1

Ouyang, W., Gu, X., Ye, L., Liu, X., Zhang, C., 2025. Exploring Hydrological Variable Interconnections and Enhancing Predictions for Data-Limited Basins Through Multi-Task Learning. Water Resour. Res. 61, e2023WR036593. https://doi.org/10.1029/2023WR036593

Pandit, A., Golden, H.E., Christensen, J.R., Lane, C.R., Husic, A., 2025. Deep Learning Prediction and Interpretation of Riverine Nitrate Export Across the Mississippi River Basin. Water Resour. Res. 61, e2024WR039207. https://doi.org/10.1029/2024WR039207

Rahmani, F., Shen, C., Oliver, S.K., Lawson, K., Appling, A.P., 2021. Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. Hydrol. Process. 35, e14400. https://doi.org/10.1002/hyp.14400

Rathjens, H., Miguez, M.B., Winchell, M., Sur, R., 2023. Evaluation of SEAWAVE–QEX in a high agricultural intensity catchment in Belgium. Integr. Environ. Assess. Manag. 19, 513–526. https://doi.org/10.1002/ieam.4688

Saavedra, F., Vergopolan, N., Musolff, A., Merz, R., Wang, Z., Winter, C., Tarasova, L., 2025. From Soil Moisture Spatial Patterns to Catchment Nitrate Dynamics Using Explainable AI. Water Resour. Res. 61, e2025WR040295. https://doi.org/10.1029/2025WR040295

Saavedra, F.A., Musolff, A., von Freyberg, J., Merz, R., Basso, S., Tarasova, L., 2022. Disentangling scatter in long-term concentration–discharge relationships: the role of event types. https://doi.org/https://doi.org/10.5194/hess-26-6227-2022
Sadler, J.M., Appling, A.P., Read, J.S., Oliver, S.K., Jia, X., Zwart, J.A., Kumar, V., 2022. Multi-Task Deep Learning of Daily Streamflow and Water Temperature. https://doi.org/10.1029/2021WR030138
Saha, G.K., Rahmani, F., Shen, C., Li, L., Cibin, R., 2023. A deep learning-based novel approach to generate continuous daily stream nitrate concentration for nitrate data-sparse watersheds. Sci. Total Environ. 878, 162930. https://doi.org/10.1016/j.scitotenv.2023.162930
Schauer, L.S., Jawitz, J.W., Cohen, M.J., Musolff, A., 2025. Spatial and Temporal Variability of River Water Quality. Hydrol. Process. 39, e70154. https://doi.org/10.1002/hyp.70154
Van Meter, K.J., Basu, N.B., Veenstra, J.J., Burras, C.L., 2016. The nitrogen legacy: emerging evidence of nitrogen accumulation in anthropogenic landscapes. Environ. Res. Lett. 11, 035014. https://doi.org/10.1088/1748-9326/11/3/035014
Vaughan, M.C.H., Bowden, W.B., Shanley, J.B., Vermilyea, A., Sleeper, R., Gold, A.J., Pradhanang, S.M., Inamdar, S.P., Levia, D.F., Andres, A.S., Birgand, F., Schroth, A.W., 2017. High-frequency dissolved organic carbon and nitrate measurements reveal differences in storm hysteresis and loading in relation to land cover and seasonality. Water Resour. Res. 53, 5345–5363. https://doi.org/10.1002/2017WR020491
Wagner, P.D., Duethmann, D., Kiesel, J., Pool, S., Hrachowitz, M., Ceola, S., Herzog, A., Houska, T., Loritz, R., Spieler, D., Staudinger, M., Tarasova, L., Thober, S., Fohrer, N., Tetzlaff, D., Wagener, T., Guse, B., 2025. The Unexploited Treasures of Hydrological Observations Beyond Streamflow for Catchment Modeling. WIREs Water 12, e70018. https://doi.org/10.1002/wat2.70018
Wang, L., Shi, L., Reimers, C., Wang, Yanling, He, L., Wang, Yujie, Reichstein, M., Jiang, S., 2025. A self-supervised deep learning model for enhanced generalization in soil moisture prediction. J. Hydrol. 662, 133974. https://doi.org/10.1016/j.jhydrol.2025.133974
Winter, C., Jawitz, J.W., Ebeling, P., Cohen, M.J., Musolff, A., 2024. Divergence Between Long-Term and Event-Scale Nitrate Export Patterns. Geophys. Res. Lett. 51, e2024GL108437. https://doi.org/10.1029/2024GL108437
Xia, X., Liu, X., Liu, J., Fang, K., Lu, L., Oymak, S., Currie, W.S., Liu, T., 2025. Identifying trustworthiness challenges in deep learning models for continental-scale water quality prediction. Nexus 2, 100104. https://doi.org/10.1016/j.ynexs.2025.100104
Zarnetske, J.P., Bouda, M., Abbott, B.W., Saiers, J., Raymond, P.A., 2018. Generality of Hydrologic Transport Limitation of Watershed Organic Carbon Flux Across Ecoregions of the United States. Geophys. Res. Lett. 45, 11,702-11,711. https://doi.org/10.1029/2018GL080005
Zhang, S., Yang, H., Yang, D., Jayawardena, A.W., 2016. Quantifying the effect of vegetation change on the regional water balance within the Budyko framework. Geophys. Res. Lett. 43, 1140–1148. https://doi.org/https://doi.org/10.1002/2015GL066952
Zhi, W., Appling, A.P., Golden, H.E., Podgorski, J., Li, L., 2024. Deep learning for water quality. Nat. Water 2, 228–241. https://doi.org/10.1038/s44221-024-00202-z
Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., Li, L., 2021. From Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale? Environ. Sci. Technol. 55, 2357–2368. https://doi.org/10.1021/acs.est.0c06783
Zhi, W., Ouyang, W., Shen, C., Li, L., 2023. Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers. Nat. Water 1, 249–260. https://doi.org/10.1038/s44221-023-00038-z

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


## Appendix A: Complete model performance overview

Appendix A includes tabular overviews over the different model types and the median KGE' and NSE performances of all models (ensemble mean). The color coding presented in Table A1 is applied to all figures.

**Table A1.** Model configurations shown in Figs. A1–A2. The colored square on the left of each heatmap row indicates the model family. Empty cells (—) indicate that a variable was not a training target for that configuration.

| Color | Model type | Inputs | Target(s) | Variants |
|---|---|---|---|---|
| | baselineLSTM multi-target | meteorological (8) | Q + 6 WQ | 2 (sample-avg. / variable-avg.) |
| | baselineLSTM paired | meteorological (8) | Q + 1 WQ | 6 (one per WQ variable) |
| | baselineLSTM single-target | meteorological (8) | 1 WQ | 7 (Q, $NO_3$-N, DOC, DO, $PO_4$-P, EC, WT) |
| | baselineLSTM single-target (+ Q in) | meteorological + Q (9) | 1 WQ | 6 ($NO_3$-N, DOC, DO, $PO_4$-P, EC, WT) |
| | self-supervisedLSTM | meteorological (8) | Q + 6 WQ | 1 [8 different inference scenarios: single-mask(7) / multi-mask(1)] |

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026

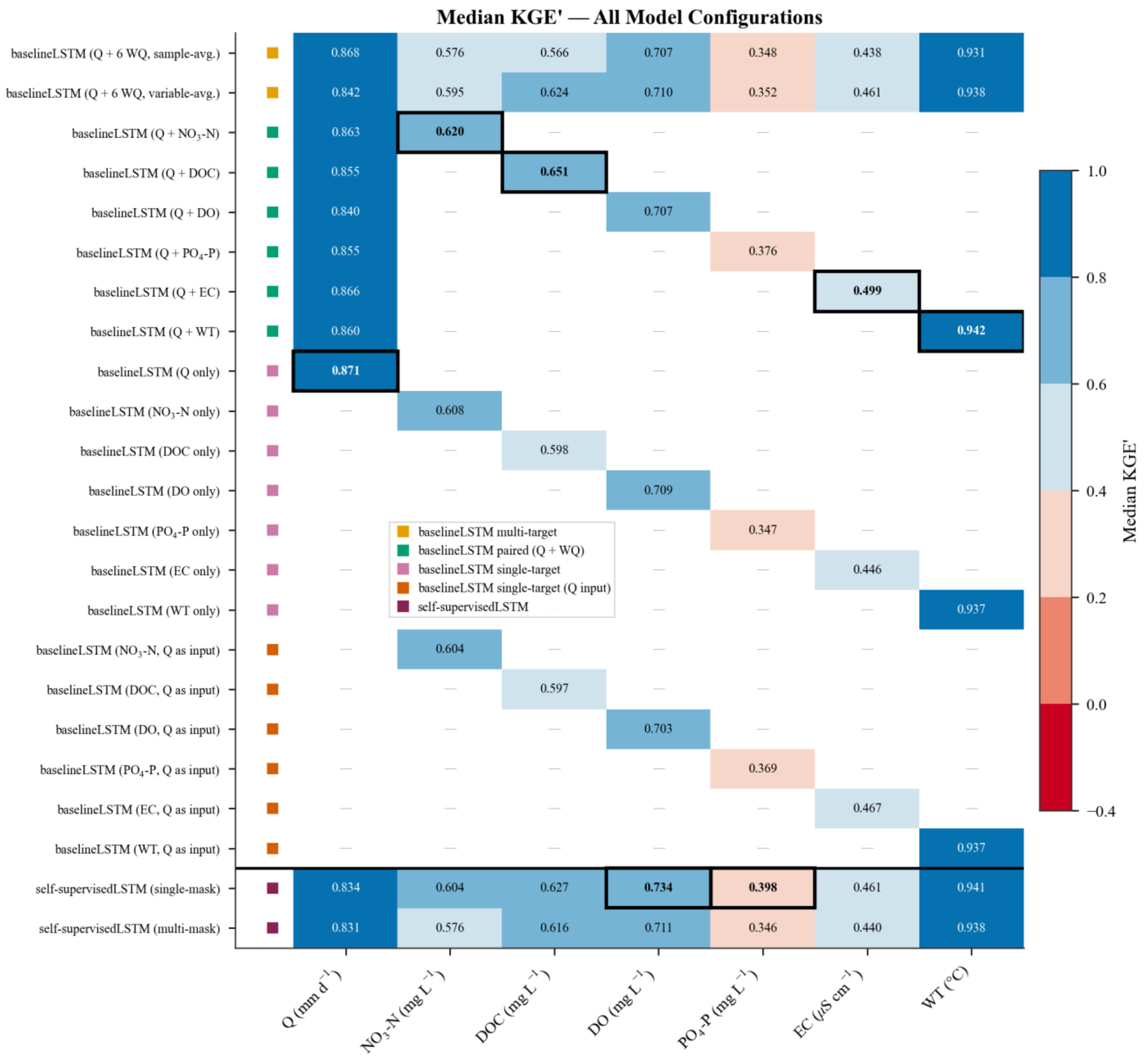




**Figure A1**. Median KGE′ on the test period (2010–2019) for all model configurations and target variables. Each row corresponds to one model configuration; columns represent the seven target variables. Bold values with a black border mark the best-performing configuration per variable. Dashes (—) indicate that the variable was not a training target for that configuration. Colored squares on the left y-axis indicate the model family (see Table A1).

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026

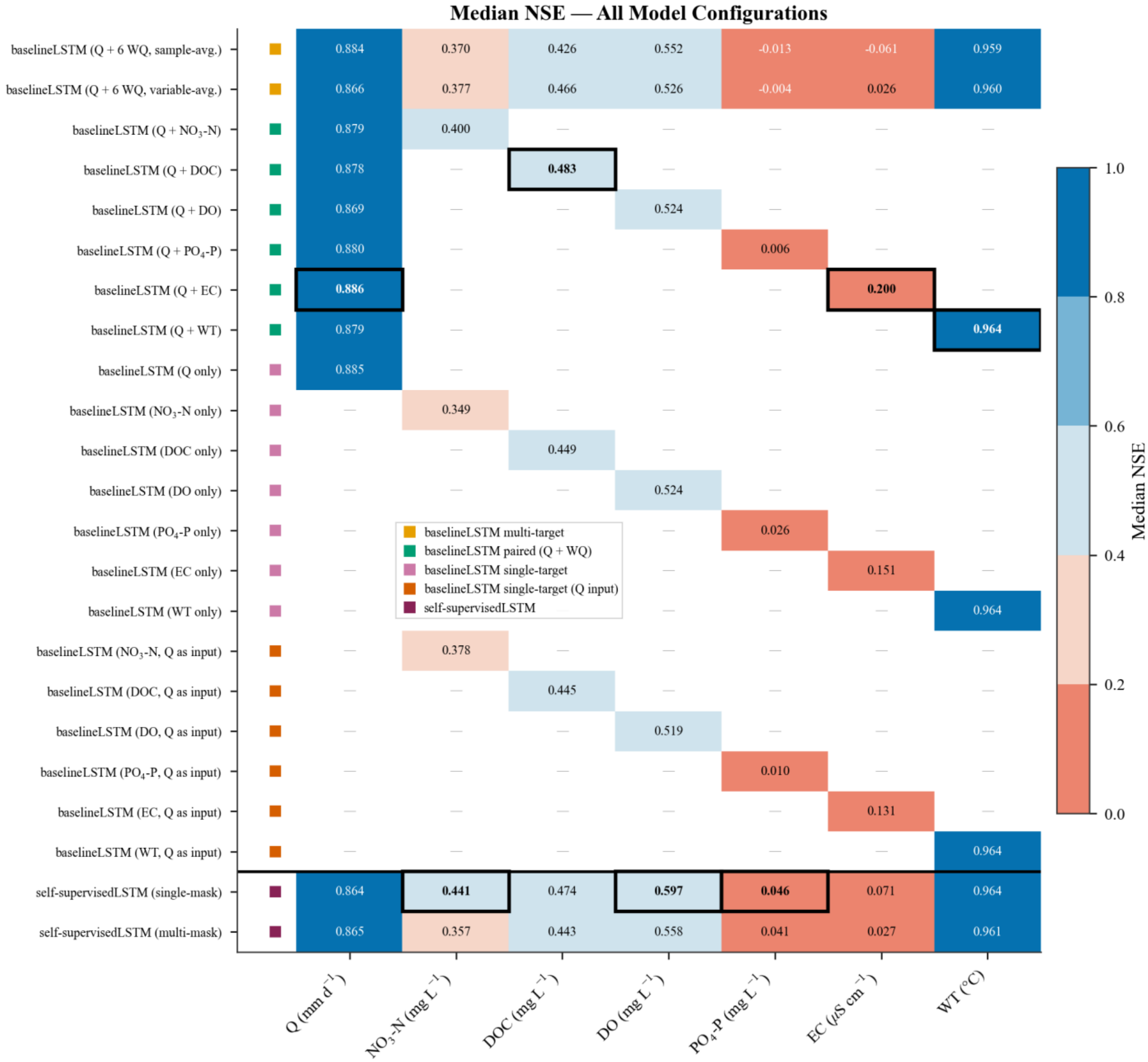




**Figure A2.** As Fig. A1, but showing median NSE. Shows how a different metric can shift the analysis completely.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


## Appendix B: Performance distributions

Appendix B covers KGE' and NSE performance distribution for all models.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026

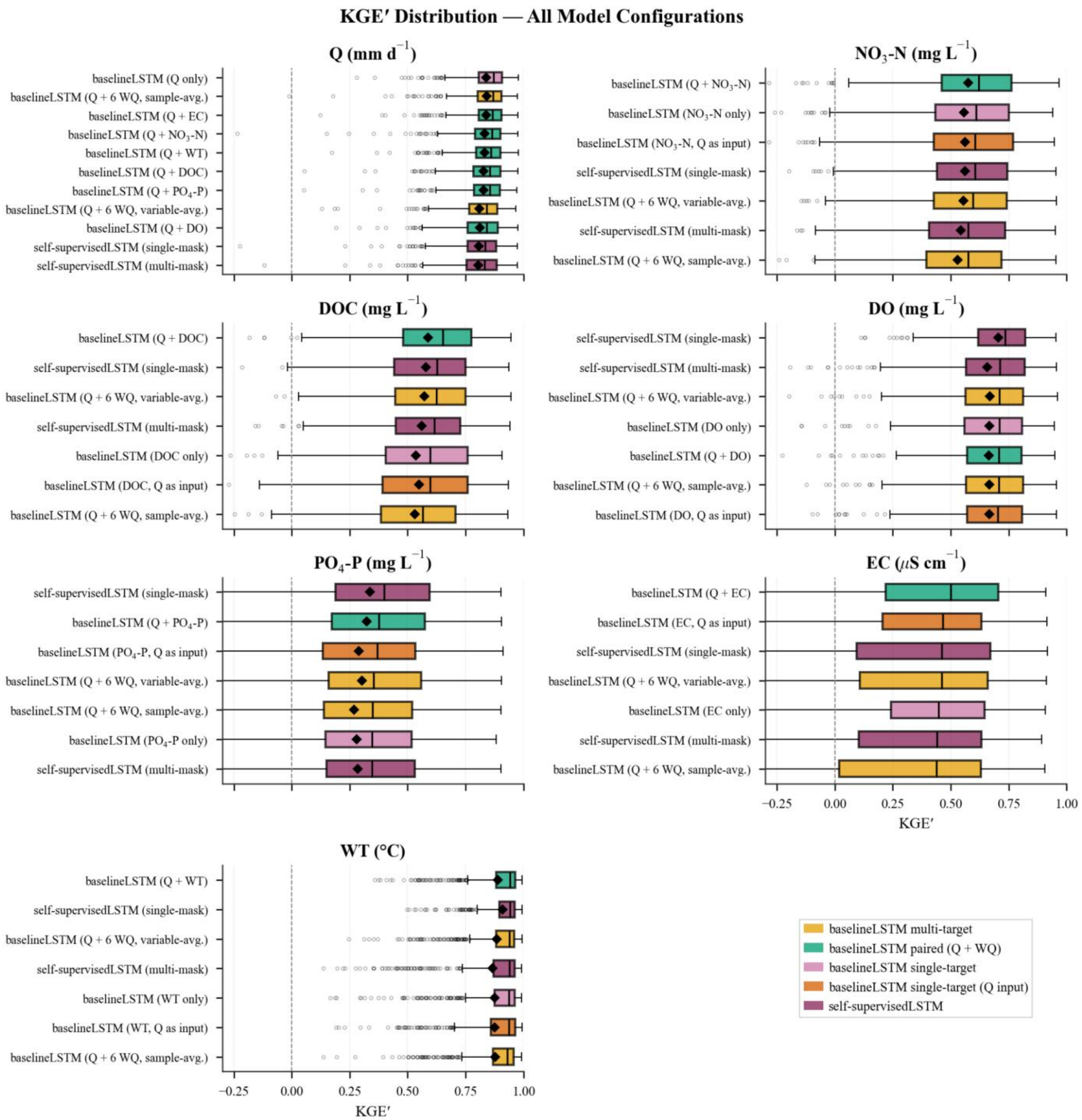



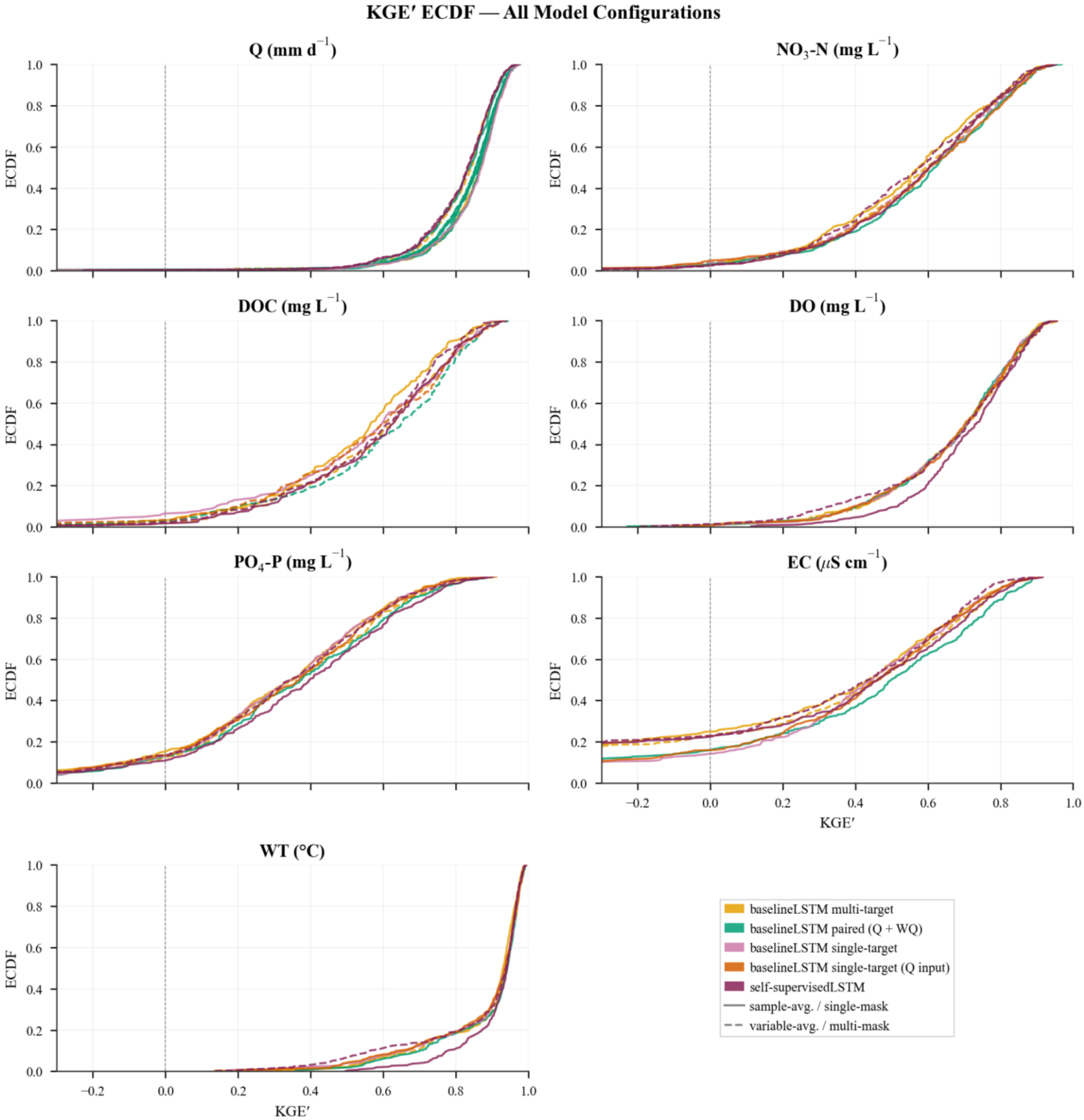

**Figure B1.** Distribution of test-period KGE′ across the catchments for all model configurations and target variables (2010–2019). Each panel shows one target variable; horizontal boxes correspond to individual model configurations, sorted by ascending median KGE′ (best on top). Colour indicates model family (legend; see Table A1). Box centres mark the median, hinges span the interquartile range (IQR), whiskers extend to 1.5× IQR, black diamonds mark the means, and grey points show individual catchments beyond the whiskers. The dashed vertical line at KGE′ = 0 separates positive from negative scores.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026



**Figure B2.** As Fig. B1, but showing distributions of test-period NSE.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


683
684 **Figure B3.** Empirical cumulative distribution functions (ECDFs) of test-period KGE′ across the catchments for all model configurations and target variables
685 (2010–2019). Each panel shows one target variable; one ECDF curve per model configuration. Line colour indicates model family (legend; see Table A1).
686 Within the multi-target baselineLSTM and the self-supervisedLSTM families, solid lines denote the sample-averaged loss and single-mask scenario, dashed
687 lines the variable-averaged loss and multi-mask scenario, respectively. The dashed vertical line at KGE′ = 0 separates positive from negative scores.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026

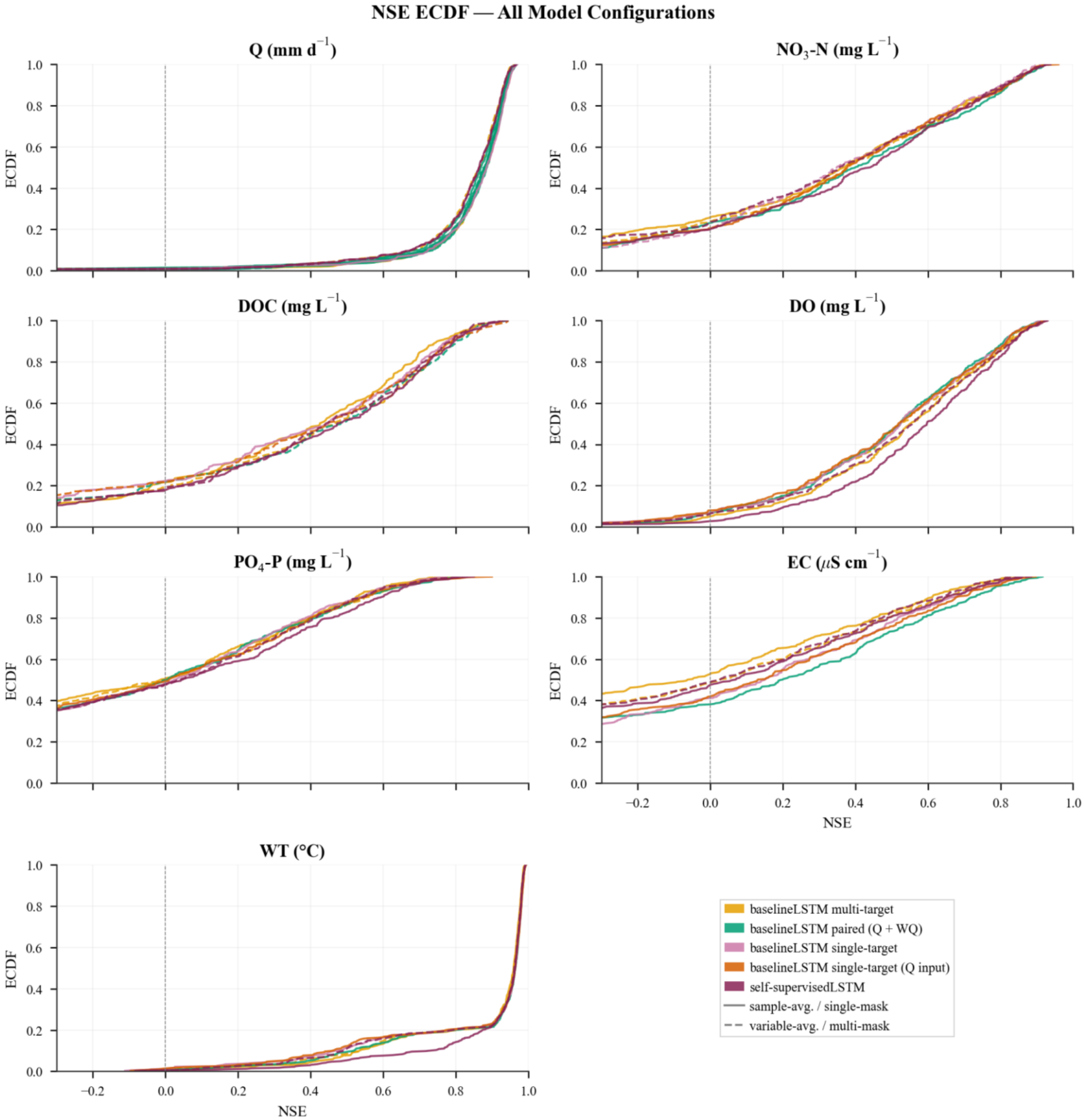




**Figure B4.** As Fig. B3, but showing ECDFs of test-period NSE.

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


## Appendix C: Data distribution

Appendix C gives details to all model inputs.

**Table C1.** Overview of all data used in the modeling process including variables descriptions and median and maximum sample count per variable and sample split period.

| | | | | Per-catchment sample count | | | |
|---|---|---|---|---|---|---|---|
| | | | | Training | | Testing | |
| | Variable | Unit | Description | Median | Max | Median | Max |
| **CAMELS-DE** | | | | | | | |
| *Catchment attributes* | | | | | | | |
| | Area | km² | Catchment area | – | – | – | – |
| | Elevation | m a.s.l. | Mean catchment elevation | – | – | – | – |
| | Clay fraction | % | Mean clay content (0–30 cm depth) | – | – | – | – |
| | Sand fraction | % | Mean sand content (0–30 cm depth) | – | – | – | – |
| | Silt fraction | % | Mean silt content (0–30 cm depth) | – | – | – | – |
| | Artificial surfaces | % | Fraction of artificial/urban land cover | – | – | – | – |
| | Agricultural areas | % | Fraction of agricultural land cover | – | – | – | – |
| | Forests & semi-natural | % | Fraction of forest and semi-natural land cover | – | – | – | – |
| | Wetlands | % | Fraction of wetland land cover | – | – | – | – |
| | Water bodies | % | Fraction of water body land cover | – | – | – | – |
| *Climatic indices* | | | | | | | |

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Mean precipitation | mm d⁻¹ | Long-term mean daily precipitation | – | – | – | – |
| | Precipitation seasonality | – | Seasonality index of precipitation | – | – | – | – |
| | Snow fraction | – | Fraction of precipitation falling as snow | – | – | – | – |
| | High precipitation frequency | d yr⁻¹ | Frequency of high-precipitation days | – | – | – | – |
| | Low precipitation frequency | d yr⁻¹ | Frequency of low-precipitation days | – | – | – | – |
| | High precipitation duration | d | Mean duration of high-precipitation events | – | – | – | – |
| | Low precipitation duration | d | Mean duration of low-precipitation events | – | – | – | – |
| *Meteorological forcing* | | | | | | | |
| | Precipitation (mean) | mm d⁻¹ | Daily mean areal precipitation | daily | daily | daily | daily |
| | Precipitation (min) | mm d⁻¹ | Daily minimum areal precipitation | daily | daily | daily | daily |
| | Precipitation (max) | mm d⁻¹ | Daily maximum areal precipitation | daily | daily | daily | daily |
| | Humidity | % | Daily mean relative humidity | daily | daily | daily | daily |
| | Global radiation | W m⁻² | Daily mean global radiation | daily | daily | daily | daily |
| | Temperature (mean) | °C | Daily mean air temperature | daily | daily | daily | daily |
| | Temperature (min) | °C | Daily minimum air temperature | daily | daily | daily | daily |
| | Temperature (max) | °C | Daily maximum air temperature | daily | daily | daily | daily |
| *Streamflow* | | | | | | | |

https://doi.org/10.5194/egusphere-2026-3604
Preprint. Discussion started: 14 July 2026


| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Specific discharge | mm $d^{-1}$ | Observed specific discharge | daily | daily | daily | daily |
| **QUADICA** | | | | | | | |
| *Nutrients* | | | | | | | |
| | $NO_3$-N | mg $L^{-1}$ | Nitrate-nitrogen concentration | 109 | 273 | 105 | 1,164 |
| | $PO_4$-P | mg $L^{-1}$ | Orthophosphate-phosphorus concentration | 101 | 273 | 103 | 234 |
| *Organic Carbon* | | | | | | | |
| | DOC | mg $L^{-1}$ | Dissolved organic carbon concentration | 65 | 252 | 24 | 216 |
| *Physical properties* | | | | | | | |
| | WT | °C | Water temperature | 113 | 885 | 105 | 1,175 |
| | EC | µS $cm^{-1}$ | Electrical conductivity | 102 | 3,327 | 102 | 2,964 |
| *Dissolved gases* | | | | | | | |
| | DO | mg $L^{-1}$ | Dissolved oxygen concentration | 111 | 3,392 | 105 | 3,041 |