30.10.2026

Dear Editor,

Thank you very much for handling our manuscript. We have revised it in response to the referees' comments. The main changes are listed below.

1. We matched the self-supervised model's hidden size to that of the supervised models (512) and adopted the suggested dynamic masking rate.
2. We moved the bootstrap significance analysis to the Appendix
3. We added a table with the KGE' decompositions to the Appendix,
4. We merged Sects. 4.1 and 4.2 into a single Results section, shortened the results of the $_{\text{self-supervised}}$LSTM, and removed Fig. 3.

Data update

During the revision process, we were informed by the authors of the already published QUADICA v2 dataset (Ebeling et al., 2026) that they had identified and corrected errors in its data processing. We therefore reran all models and analyses using the corrected dataset. The corrections mainly affected electrical conductivity and water temperature in selected federal states; most catchments and water-quality variables included in our study remained unaffected. **The updated results show only minor changes in model performance and regional patterns, and the main findings and conclusions remain unchanged.** We have revised all affected text, figures, and tables accordingly. All resulting changes to the manuscript are documented in the track-changes version provided with this revision.

Thank you again for your time and for the reviewers' constructive comments.

On behalf of all co-authors,

Jean-Paul Brede

Ebeling, P., Hubig, A., Wachholz, A., Scharfenberger, U., Haug, S., Nguyen, T., Sarrazin, F., Batool, M., Musolff, A., and Kumar, R.: QUADICA v2: extending the large-sample data set for water QUAlity, DIscharge and Catchment Attributes in Germany, Earth Syst. Sci. Data, 18, 691–712, https://doi.org/10.5194/essd-18-691-2026, 2026.
Updated DOI: https://doi.org/10.4211/hs.11cd7518f3b74225a5b9e66bfbce362e
Until revised publication available under:
http://www.hydroshare.org/resource/11cd7518f3b74225a5b9e66bfbce362e

30.10.2026

The following pages contain the replies to the referees.

Reply to RC1: 'Comment on egusphere-2026-3604', Anonymous Referee #1, 07 Sep 2026

We thank the referee for the thorough review of our manuscript and the detailed, well-founded comments. We reply to each comment below and keep our answers as concise as the comments are precise.

Major Comments

[1] Comments on hidden size:

"It is stated (Lines 201–202) that the "only difference" between models is the learning objective to assess the SSL formulation independently. However, Table 2 states that the baseline LSTM uses a hidden size of 512, while the self-supervised LSTM is increased to 1024. This modification represents a roughly fourfold increase in model capacity. Given that the highlighted gains for PO_4–P and DO are quite modest (delta KGE' of 0.05 and 0.02 respectively, per Line 294), it is possible yet not discussed that these gains are a result of increased capacity rather than the SSL logic.

The possibility that these gains are a result of the increased capacity rather than the SSL logic should be tested before concluding. Ideally, a standard benchmark experiment should be added where a baseline supervised LSTM is trained with a hidden size of 1024 to serve as a proper control. Alternatively the increase in hidden size must be argued and discussed with regards to performance changes." - RC1

We tested several hyperparameter settings for the supervised and for the self-supervised model versions. In this early testing the supervised model did not benefit from the increased hidden size of 1024 while the self-supervised model did. Prompted by the referee's comment we will train new ensembles for the multi-target baseline model with 1024 hidden size and the SSL model with decreased hidden size in order to be able to systematically rule out the influence of model capacity on the performance.

[2] Comments on fixed masking rate and unrepresentative training task:

"The model is trained with a fixed masking rate of exactly 3 variables (Line 193) but is evaluated in a "multi-masking" scenario where 7 variables are masked simultaneously. This creates an unencountered distribution shift at inference that could explain why the SSL model does not outperform (Line 444) its supervised counterpart in that scenario. This technical limitation should be incorporated into the otherwise good discussion better yet the SSL model should be trained with a dynamic masking rate (e.g., randomly masking between 1 and 7 variables per sequence) so it learns to operate under any missingness pattern." - RC1

The referee is correct to point out that the training task and the inference tasks are technically different. However, because of the sparse WQ data, the effective difference is considerably smaller than the nominal masking rates suggest. During the preparation of this study, we tested different settings (e.g. different masking rates, grouped masking, masked forcings), while the majority of options remained unexplored, and could not identify systematic insights, which could (again) relate to the sparse nature of the WQ data. The number 3 for the masked channel hyperparameter during training was chosen based on preliminary tests, but not optimized, which is not explained in the manuscript and we will correct that in the finalized version. The referee is correct in the assessment that the training and the inference task should align. On the referee's suggestion we will also test an ensemble with SSL and a dynamic masking rate of the proposed 1-7 channels and discuss its result in the manuscript.

[3] Comment on “regional” model:

“The study relies on a temporal split (2000–2010 vs. 2010–2019) and regionalized LSTMs to support claims of "regional" modeling (Lines 28) which typically implies the ability to generalize to ungauged catchments. The current setup only demonstrates temporal interpolation within known locations as no spatial holdout is tested. The claims of “regional modeling” should be revised or a spatial holdout tested." - RC1

We use the term "regional" for a single model with one shared parameter set trained jointly on many catchments, in contrast to models calibrated locally for individual catchments. This follows Kratzert et al. (2019a), who refer to a single LSTM trained on 531 CAMELS basins as a regional model and evaluate it on a temporal split of the same basins, while noting that the regional modeling problem is closely related to prediction in ungauged basins (PUB). LSTMs trained in this way have since been shown to generalize to ungauged basins (Kratzert et al., 2019b; Nearing et al., 2024). However, the skill of such models for WQ variables in ungauged basins requires explicit testing as the referee pointed out and will be the subject of future research. We agree that the abstract can be read as a claim of spatial generalization beyond the training set and will revise the sentence accordingly. We will also state explicitly in Sect. 5.2 that our results characterize temporal generalization for gauged catchments.

[4] Comment on bootstrapping:

“The paired bootstrap analysis across catchments (Lines 290, 295) asserts statistical significance for very small metric gains (e.g., delta KGE' of 0.02 for DO and 0.05 for PO_4–P) but is only based on the spatial distribution of the ensemble means. Averaging the 5 members predictions prior to metric calculation flattens temporal variance and discards the epistemic uncertainty of model initialization. The significance test must account for both spatial and initialization variance to prove these minor improvements are not simply artifacts of random seed selection. The methodology should be revised to either reflect only evaluating ensemble mean or ensemble characteristics should be reported and accounted for in statistical tests.“ - RC1

We thank the referee for this comment and agree with their assessment that there is only a spatial bootstrapping in place. We treat the five-member ensemble mean as the model under

evaluation. This follows common practice for LSTM rainfall–runoff models, where averaging members trained with different random initializations mitigates the error arising from the stochasticity of training (Kratzert et al., 2019a, b). The ensemble results in better performance than almost all individual seeds. As suggested by the referee we will add a remark to the methodology to clarify we only evaluate the ensemble means.

To give some additional analysis to the reported significances:

1. Comparing the **median performances of each model seed** for the SSL (single- & multi-mask) with the supervised baseline (multi-target, variable-averaged MSE) results in 25 combinations for each target variable. For PO4-P and DO in the single-mask inference task 100% of the combinations result in better KGE' performances for the SSL model with ranges from +0.014 to +0.081 (PO4-P) and +0.017 to +0.043 (DO). For NO3-N in the multi-mask inference task the SSL set up performs worse in 80% of the seed comparison combinations (ΔKGE' range −0.064 to +0.007) but for EC it only performs worse in 40% (ΔKGE' range −0.080 to +0.024).

2. When conducting the **paired difference bootstrapping for each model seed** combination (n=25) instead of comparing the ensemble means, the reported differences (SSL better in single-mask inference) stay significant for PO4-P in 72% and for DO in 88% of the seed combinations. Individual ensemble members are noisier than the mean, which makes this test stricter.

Minor comments

[1] Comment on PO4-P attribution:

“The performance for PO_4–P (median KGE' 0.35) is attributed to "intrinsic complexity," yet the source context (Line 383) correctly identifies missing "point sources" and legacy effects. In addition, the use of a 365-day sliding window could be hydrologically problematic for nutrients. The authors acknowledge "deep groundwater memory" and "legacy effects" (Line 383), however a stateless model resetting every year as used in this study cannot physically capture these multi-year processes. This should be added as a paragraph acknowledging this structural architectural limitation.” - RC1

Indeed, the referee raises a good point here, the point source controls and legacy effects cannot be covered with the current approach, adding to the intrinsic complexity of PO4-P. We will more clearly discuss this limitation in the manuscript. In fact, this also applies for NO3-N, where long-term legacy from diffuse sources is a well-known phenomenon (e.g. Ehrhardt et al. 2021).

Regarding source controls: One approach is to provide source information (diffuse and point sources) as a model input, this is, however, beyond the scope of the current study but useful to be further explored in the future. We would like to highlight though, that although large-scale datasets exist for Germany (e.g., Sarrazin et al. 2024), they are subject to uncertainties and provide limited temporal resolution providing annual statistics only, hence requiring additional strategies and assumptions to downscale to the daily time steps predicted here.
Moreover, we would like to point out that changes in point source inputs during the training (2000-2010) and testing periods (2010-2019) can be considered lower and relatively constant compared to highest nutrient emissions to streams in the 1970s and 1980s (Sarrazin et al. 2024). In this study, the validation period in the 1990s could be affected more strongly with most distinct point source inputs, however, it was only used for hyperparameter tuning purposes.

Regarding the long-term legacy effects: Indeed, longer model memories could be necessary to reflect long-term improvements in nutrient pollution related to long-term changes in the management. In particular, for NO3-N, legacies of up to decades are well studied (e.g. Ehrhardt et al. 2021), for PO4-P however, more immediate reactions to point source reductions are reported, although legacy storage in sediments and remobilization are possible (e.g. Westphal et al. 2019, 2020). Longer model memories (sequence length) would mainly be useful if temporal data such as cropping patterns and fertilizer inputs are considered in the predictors. A technical solution to reflect longer time periods could be the Multi Frequency LSTM (Acuña Espinoza et al. 2024, 2026) developed within our workgroup which might be of further interest to the referee.

[2] Comment on SSL missing data handling:

“The claim that SSL handles missing data naturally without imputation (Line 199) is technically underspecified given the 97% missingness in WQ variables. When a WQ variable acts as a dynamic input, it remains unclear how the daily-scale dynamic input vector represents those gaps. It should be explicitly stated how the SSL handles these e.g.: are these zero-padded, or are they accompanied by binary companion masks to alert the LSTM to the validity of the input signal?” - RC1

The referee is correctly pointing out a technical underspecification of the handling of missing input data. There is neither any zero-padding nor a binary flag feature. Every NaN value in the input is replaced by the same trainable token that is used for the masking process (similar to zero-padding, all missing values receive the same value, but this value is learned during training and is therefore not confounded with a physical value of zero). This handling makes the signal of missing data in the input identical to the signal when masking is enforced. We will add this information in section 3.2.

[3] Comments on formatting and references:

“Lines 12, 188, 264, 333, 373, and 468 (Inconsistent Chemical Formula Formatting): The chemical formulas for nitrate-nitrogen and orthophosphate-phosphorus are formatted in three conflicting styles across different sections of the article:

- *Subscript with en-dash (NO_3–N and PO_4–P) in Line 12 (Abstract) and Line 188 (Section 3.2).*
- *Subscript with standard hyphen (NO_3-N and PO_4-P) in Line 264 (Figure 2 caption) and Line 333 (Section 5.1).*
- *Plain text with standard hyphen (NO3-N) in Line 373 and Line 468 (Section 6).*

Line 240: In the sentence: “...leads to a slight decrease in runoff performance (median KGE from 0.87 to 0.84)...”, the prime symbol is omitted.

Lines 534–535, 566–567, and 614–615: Several references in the bibliography contain compiled duplication errors in their DOI prefixes, which results in broken hyperlinks.” - RC1

We thank the referee for the careful inspection of our manuscript. The referee correctly identified the different uses of en-dashes and standard hyphens in the chemical variable formulations which slipped our attention. We will synchronize them (as well as the plain text instance of NO3-N).
The prime symbol for the KGE in Line 240 will be added.
We will correct the duplicated DOI prefixes.

References

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

Acuña Espinoza, E., Álvarez Chaves, M., Kratzert, F., Klotz, D., Gauch, M., Lang, R., Elfgang, D., Dolich, A., Loritz, R., and Ehret, U.: Everything everywhere all at once: A single-cell LSTM network unifying multi-frequency, missing data, and discharge assimilation for robust operational flood forecasting, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2026-4885, 2026.

Ehrhardt, S., Ebeling, P., Dupas, R., Kumar, R., Fleckenstein, J. H., and Musolff, A.: Nitrate transport and retention in Western European catchments are shaped by hydroclimate and subsurface properties, Water Resour. Res., 57, e2020WR029469, https://doi.org/10.1029/2020WR029469, 2021.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019a.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resour. Res., 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019b.

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., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Sarrazin, F. J., Attinger, S., and Kumar, R.: Gridded dataset of nitrogen and phosphorus point sources from wastewater in Germany (1950–2019), Earth Syst. Sci. Data, 16, 4673–4708, https://doi.org/10.5194/essd-16-4673-2024, 2024.

Westphal, K., Graeber, D., Musolff, A., Fang, Y., Jawitz, J. W., and Borchardt, D.: Multi-decadal trajectories of phosphorus loading, export, and instream retention along a catchment gradient, Sci. Total Environ., 667, 769–779, https://doi.org/10.1016/j.scitotenv.2019.02.428, 2019.

Westphal, K., Musolff, A., Graeber, D., and Borchardt, D.: Controls of point and diffuse sources lowered riverine nutrient concentrations asynchronously, thereby warping molar N:P ratios, Environ. Res. Lett., 15, 104009, https://doi.org/10.1088/1748-9326/ab98b6, 2020.

Reply to RC2: 'Comment on egusphere-2026-3604', Anonymous Referee #2, 08 Sep 2026

We thank the referee for their comments. All points are well founded and we address each specific comment in the following.

Specific Comments

[1] Comment on extreme values:

"Since the strongest impacts of water quality variables typically occur when they reach unusual (i.e., extreme) values, I would have liked to see an assessment of model performance from that perspective, and whether the findings highlighted above also apply during extreme conditions." - RC2

We agree with the referee that the performance at the edges of the distribution is of particular relevance for water quality, since the strongest impacts often occur at unusual values. For discharge, LSTMs have been shown to predict extreme events more reliably than conceptual and process-based models, even when such events were excluded from training (Frame et al., 2022). At the same time, a stand-alone LSTM has a theoretical upper limit on its output, which can lie below the maximum of the training data, and it can show implausible responses under extreme forcing (Acuña Espinoza et al., 2025; Baste et al., 2025). Both studies find that the limitation becomes less relevant when the model is trained on all available data, including rare events, and on larger and more diverse datasets, although this does not guarantee reliable extrapolation (Baste et al., 2025).
We expect the same limitation to apply to water quality, and more strongly so. Water-quality variables are very sparse in our dataset, so extreme concentrations especially during floods are rarely part of the data. With the current sampling, a robust assessment of whether our findings hold under extreme conditions is therefore not possible. We will add this limitation to the discussion, together with the outlook that denser and more complete data, for example from high-frequency sensors, are required to train and evaluate models for extreme water-quality conditions.

[2] Comment on KGE' decomposition:

"Given that KGE' is used to evaluate model performance, it would be interesting to know which component of KGE' (correlation, bias or variability) is most responsible for the improvement in the predictions." - RC2

The referee's comment is appreciated. We will add a Table with the KGE' decompositions for all models and variables to the Appendix, such that improvements between model families can be deduced. Below is an example for such a deduction with the comparison of different models. The decomposition shows that the gains of the co-target models over the models with

Q as input are carried mainly by the variability term γ (DOC and PO_4-P), whereas the gains of the self-supervised model under single-masking are carried by the correlation r (DO and PO_4-P). The losses under multi-masking are carried by γ.

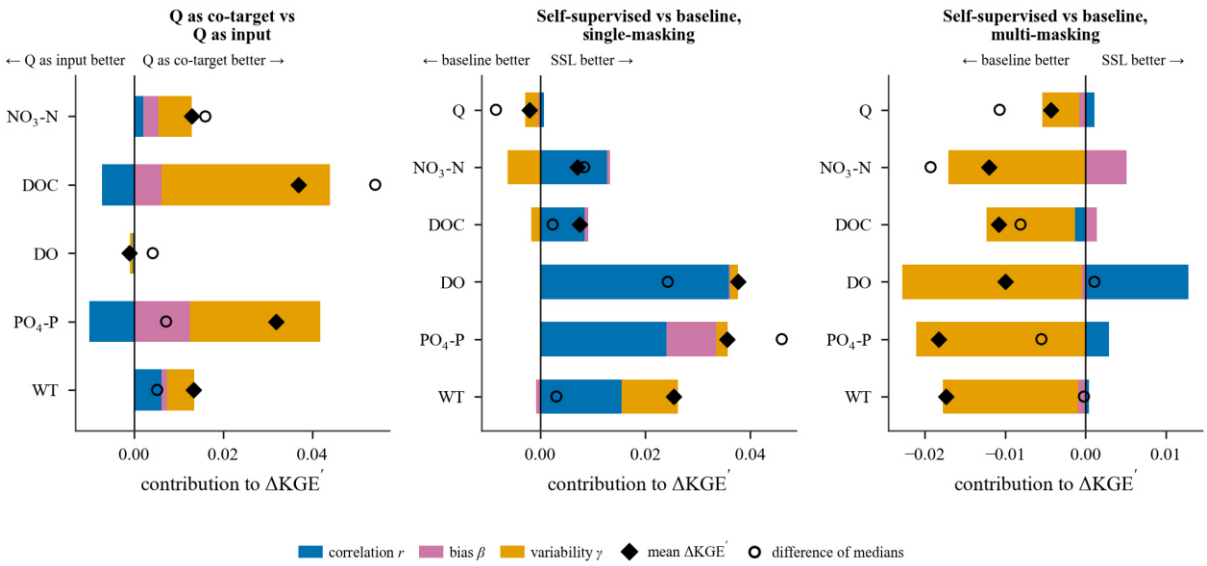

Figure 1: Mean per-catchment contributions of correlation r, bias β and variability γ to the KGE' difference between two models; EC excluded.

[3] Comment on EC:

"The caption of Figure 2 mentions: "Electrical conductivity (EC) is not shown due to figure layout constraints and is presented separately in the Appendix". However, I was unable to find this figure in the Appendix." - RC2

We thank the referee for pointing this out this was our mistake. We will add the KGE' map for EC to the Appendix. Meanwhile the boxplot in Appendix B1 contains the performance spread for all models and their respective target variables.

References

Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., and Ehret, U.: Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events, Hydrol. Earth Syst. Sci., 29, 1277–1294, https://doi.org/10.5194/hess-29-1277-2025, 2025.

Baste, S., Klotz, D., Acuña Espinoza, E., Bardossy, A., and Loritz, R.: Unveiling the limits of deep learning models in hydrological extrapolation tasks, Hydrol. Earth Syst. Sci., 29, 5871–5891, https://doi.org/10.5194/hess-29-5871-2025, 2025.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrol. Earth Syst. Sci., 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.