

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 $\text{PO}_4\text{-P}$), whereas the gains of the self-supervised model under single-masking are carried by the correlation r (DO and $\text{PO}_4\text{-P}$). The losses under multi-masking are carried by γ .

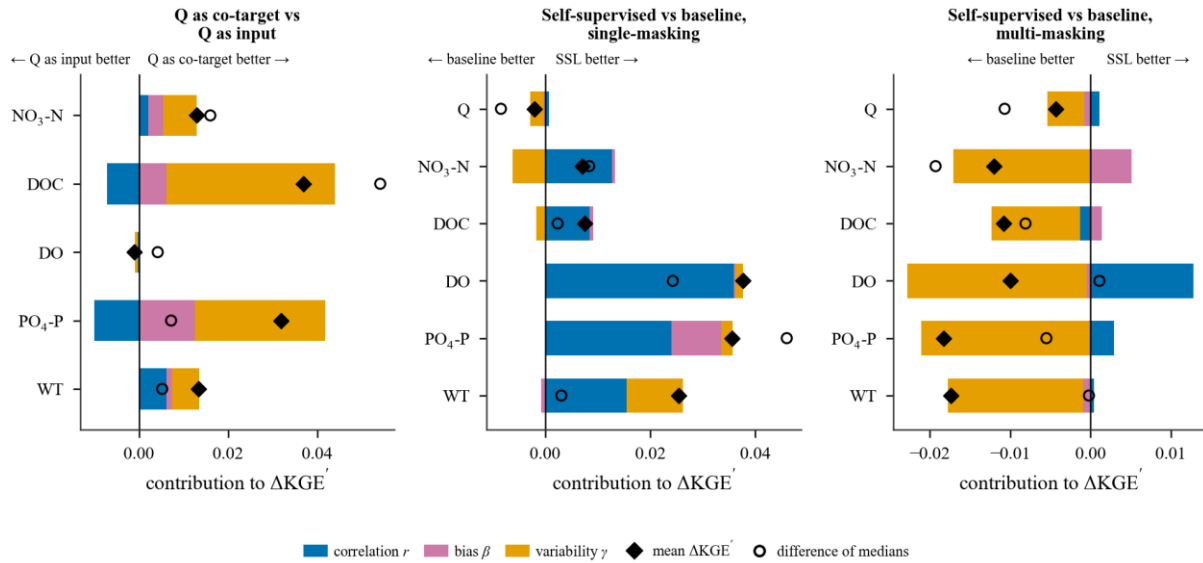


Figure 1: Mean per-catchment contributions of correlation r , bias β and variability γ to the KGE' difference between two models; EC excluded.

[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.