Corrigendum to Hydrol. Earth Syst. Sci., 28, 525–543, 2024
https://doi.org/10.5194/hess-28-525-2024-corrigendum

Hydrology and Earth System Sciences
Open Access
EGU

Corrigendum to "On the challenges of global entity-aware deep learning models for groundwater level prediction" published in Hydrol. Earth Syst. Sci., 28, 525–543, 2024

Benedikt Heudorfer[1], **Tanja Liesch**[1], **and Stefan Broda**[2]

[1]Karlsruhe Institute of Technology (KIT), Institute of Applied Geosciences, Kaiserstr. 12, 76131 Karlsruhe, Germany
[2]Federal Institute for Geosciences and Natural Resources (BGR), Wilhelmstr. 25–30, 13593 Berlin, Germany

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

Published: 16 April 2025

A minor error in the data processing code was identified after publication, affecting the cross-validation (called out-of-sample, OOS, in the paper) results reported in the original publication. Specifically, static features were incorrectly assigned to dynamic data during training sample creation. This corrigendum presents corrected figures and updated performance metrics, mostly concerning the TSfeat model variant which experienced larger changes in performance, while all other model variants perform effectively the same. Since the TSfeat model variant is not the primary concern of the original publication, the overall conclusions of the study remain unchanged. This corrigendum maintains the same section names and numbering as the original publication but details the specific adjustments made to the original text and figures.

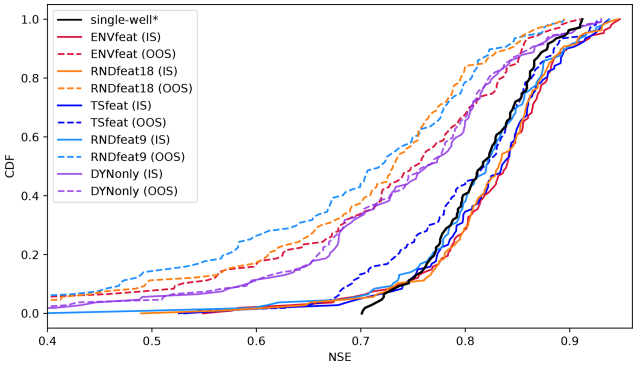

Figure 3. Cumulative distributions of NSE for the model variants ENVfeat, TSfeat, RNDfeat9, RNDfeat18, and DYNonly in in-sample (IS) mode and out-of-sample (OOS) mode against the performance of the single-well models by Wunsch et al. (2022a) (*). Lines represent sorted median NSE scores of 10 ensemble members. A version that includes the ensemble ranges as envelopes is shown in Fig. A5.

3 Methods

3.1 Corrections made to the code

A coding error was discovered in the data pre-processing stage of the original study. In one line of the code responsible for generating cross-validation runs (model_CV.py), a misattribution of station identifiers occurred during the fusion of static to dynamic features to form the overall training samples. Specifically, a wrong iterator–ID vector pair was used when subsetting the station ID vector while matching sets of dynamic and static features. This error resulted in the incorrect pairing of static features with dynamic features for stations, effectively dissolving the intended physiographic correspondence for the stations. The erroneous line of code was corrected, and the affected figures and performance metric table are presented in Sect. 4 of this corrigendum, next to an assessment of these changes. Since the coding error only occurred in the OOS code, all results concerning the in-sample (IS) models remain unchanged.

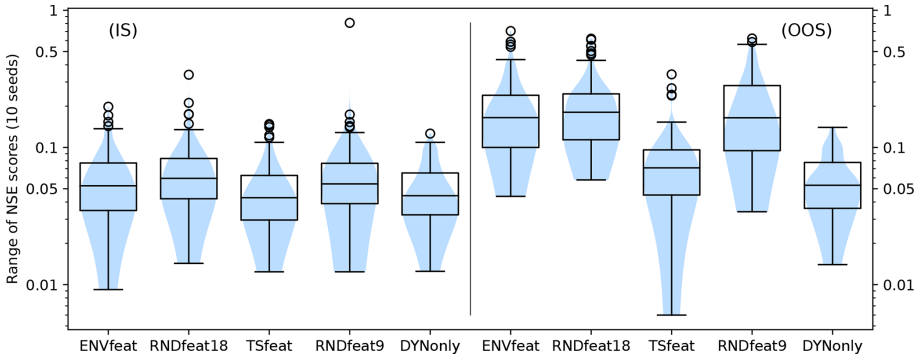

Figure 4. Range of NSE scores of the 10 ensemble members of all model variants in IS mode and OOS mode.

Table 3. Mean (bold font), lower (10 %), and upper (90 %) percentile NSE scores of the 10 ensemble members for all model variants as well as the mean NSE for single-well models (also in bold font) as published in Wunsch et al. (2022a). R^2 and RMSE show similar patterns to NSE and are reported for comparison but not discussed in the text.

Variant	NSE (Q_{10})	NSE (Q_{50})	NSE (Q_{90})	R^2 (Q_{50})	RMSE (Q_{50})
Single-well	–	0.8134	–	0.8255	0.2961
ENVfeat (IS)	0.8026	**0.8213**	0.8397	0.8418	0.2656
RNDfeat18 (IS)	0.7909	**0.8215**	0.8457	0.8354	0.2673
TSfeat (IS)	0.8028	**0.8229**	0.8395	0.8402	0.2677
RNDfeat9 (IS)	0.7777	**0.8135**	0.8399	0.8274	0.2746
DYNonly (IS)	0.7094	**0.7347**	0.7554	0.7670	0.3580
ENVfeat (OOS)	0.6392	**0.7105**	0.7646	0.7663	0.3726
RNDfeat18 (OOS)	0.6059	**0.6867**	0.7434	0.737	0.3900
TSfeat (OOS)	0.7712	**0.7995**	0.8238	0.8312	0.3250
RNDfeat9 (OOS)	0.5837	**0.6619**	0.7249	0.7187	0.4033
DYNonly (OOS)	0.7095	**0.7319**	0.7508	0.7749	0.3708

4 Results

4.1 Performance comparison of model variants

Originally, the study showed that all static feature models performed similarly poorly in OOS predictions, suggesting that static features did not contribute significantly to model generalization. However, after correcting the error we re-ran the experiments, which yielded a somewhat elevated OOS performance that was however marginal for most model variants (compare Fig. 3 and Table 3 in the original paper and this corrigendum). The models remained on a comparable level, and the overall conclusions from these results do not change. But now a notable exception is revealed, namely the TSfeat model (with time-series-derived static features), which competes almost favourably with IS models now, suggesting that time-series-derived static features provide valuable information for spatial generalization. This corrects the original assertion that all static feature models performed equally poorly in OOS settings. With regard to the ENVfeat model, the updated results affirm that it still performs on a similar level to the RNDfeat models in the OOS setting, and the ENVfeat model is still outperformed by the DYNonly model (without static features), confirming that environmental (and of course random) static features do not enhance OOS performance.

The initial hypothesis that the TSfeat model would outperform the ENVfeat model is now partially supported, at least in the OOS setting. The original text stated that the hypothesis was overall incorrect. Now it is only incorrect for the IS setting. The corrected results indicate that time-series static features provide more robust information for generalization. However, this indication still remains limited in its applicability, since time-series features can only be derived in gauged locations, defeating the purpose of OOS models to provide prediction in ungauged locations.

Appendix A

Figure A1 shows the same hierarchy of model performance between the LSTM (long short-term memory) and CNN (convolutional neural network) models as in the original paper. The LSTM still outperforms the CNN in exactly the same way, still underlining the choice of the LSTM over the CNN as used in previous studies.

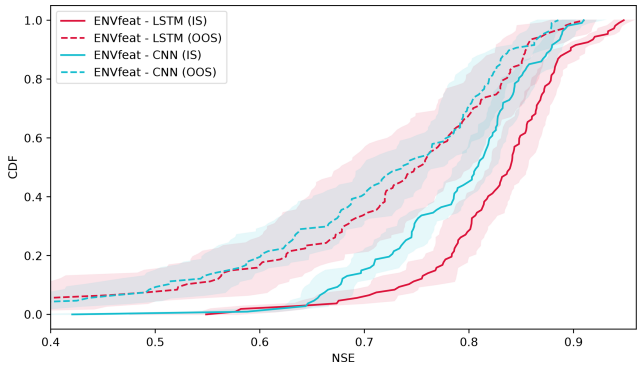

Figure A1. Comparing the performance of LSTM with CNN on the basis of the ENVfeat model variant. For the CNN model, the LSTM layer in the dynamic model thread is replaced with a CNN layer (followed by batchnorm and maxpool1D). The figure shows the performance of the models in an IS and an OOS set-up. While CNNs and LSTMs perform almost the same in the OOS mode, CNNs are clearly inferior to LSTMs in the IS mode. Thus, LSTMs were used in this study.

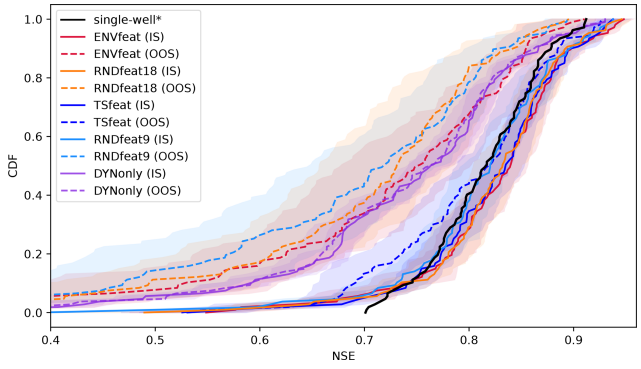

Figure A5. Cumulative distribution function of NSE of the model variants ENVfeat, TSfeat, RNDfeat, and DYNonly in IS mode and OOS mode against the performance of the single-well models by Wunsch et al. (2022a) (*). Lines represent sorted median NSE scores of 10 ensemble members, and envelopes represent ranges of the ensemble forecasts excluding the worst and best members.

Figure A5 shows the updated ranges of performance among the seeds. There is no major change except a reduced uncertainty range for the TSfeat model, as also reflected in Fig. 4.