Response to reviewers for egusphere-2024-2591: “Extended range forecasting of stream water temperature with deep learning models”

Referee #1:

The article investigates three main models on their aptitude of predicting the water temperature at specific locations of rivers in Switzerland. It goes on further to compare these models to a set of three simpler, more traditional ML models (RF, ARX and MLP). These models are evaluated in three distinct settings, namely when they were trained on data from all stations and only on a subset of stations while predicting the water temperature on gauged and ungauged stations. As their predictions, the models provide quantile forecasts and therefore directly a measure of uncertainty which is important in real world applications. In addition of investigating the predictive skills of each of the models, the article also provides an analysis of the feature importance for the best DL model (temporal fusion transformer).

All in all this work provides a valuable comparison of multiple model architectures for time series forecasting, probably acting as guidepost for future works.

We appreciate the overall positive evaluation of our manuscript.

Main comments:

1) The model description (starting at L.123) is a bit crammed and for the three deep learning models one architecture illustration each would go a long way to making later aspects more understandable. The fact that NHITS, RNNED and TFT all use encoder and decoders is not clear, and neither is the fact where exactly they use the encoder normalisation. This, however, becomes important on L. 212 (p. 7) where you describe setup B, i.e., the models trained on 20 stations worth of data less and the swap from encoder normalisers to group normalisers. So the suggestion is to include (maybe simpler versions of) diagrams of the models' architectures, aiding the understanding of the later adaptation to the encoder. Another option would be to detail the encoding process for each model where its architecture is shortly described.

We appreciate the suggestion. We expanded the text to clarify important aspects of the models’ architecture, as well as why for setup B we have to change from normalizing the models’ target variable by the long-term average and standard deviation of each station to normalizing by the average and standard deviation of each station during the 64 days of the encoder period. Using the data from the encoder period to normalize the target variable enables the model to generate predictions at stations not included during the training.

2) The description of the date index (DI) L.111 leaves the question of why it includes a shift by one month, s.t., DI=0 is approx. at the end of January instead of at the beginning? A short explanation with a reference would be nice here.

We now clarify this in the text. The index was constructed to be symmetrical: highest at the end of July, and lowest at the end of January. This is done such that June and August, May and September, …, and February and December have the same values. We consider that this is a better representation of climate seasonality in Switzerland than an index without a 1-

month shift, for which e.g. July and May, as well as August and April would have the same values.

3) Lastly, section 2.1 describes the data used for training, with it also mentioning on L.97 that "catchment characteristics" are used. However, a list of which characteristics are considered is only given on L.186, in section 2.3. The suggestion is to also explicitly mention the four static characteristics near L.97.

We now also mention the catchment characteristics near L97.