Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2022-174-RC2, 2022

Comment on gmd-2022-174

Anonymous Referee #2

Referee comment on "Prediction of algal blooms via data-driven machine learning models: an evaluation using data from a well-monitored mesotrophic lake" by Shuqi Lin et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-174-RC2, 2022

The manuscript discusses how combining machine learning (ML) models and elements of a physically based (PB) model can improve the prediction of algal blooms (identified by Chl concentration) in a well-monitored lake (Lake Erken, Sweden). Two ML models are used: Gradient Boosting Regressor (GBR) and Long short-term memory (LSTM), an advanced type of artificial neural network.

Three workflows (either purely ML or hybrid) are tested: 1. trying to obtain the Chl results directly with ML; 2. with an intermediate step (a first ML model predicting nutrients followed by a second model predicting Chl); 3. like the previous one but exploiting a PB model to obtain some physical parameters.

Moreover, several strategies are examined to calibrate (train) the model with variable portions of the available data.

The analysis is quite interesting and timely. The manuscript is well-written.

The goal of discussing how available data and their pre-treatment can affect the results of ML methods is surely important for the scientific community and for lake managers.

Nevertheless, some remarks can be pointed out.

-- Major remarks

1) Possible overfitting

The authors discuss the potential overfitting issues a few times.

L259, “there was overfitting issues in all three workflows, in both GBR and LSTM models, indicated by higher MAE and RMSE in the testing dataset compared to the training dataset especially for GBR”

This statement is not completely accurate: higher error in the testing dataset does not immediately imply overfitting. In particular, the authors discuss the peculiarities of the algal bloom in July-August 2019, which is not properly predicted by any model; as this occurrence is in the testing phase, it means that the errors are expected to be large anyway.

On the other hand, overfitting issues may effectively exist in the model application. In fact, Text S2 reports on the hyperparameters of the LSTM model: by adopting 3 layers and 100 neurons, it approximately implies 300 degrees of freedom (parameters). No information is provided for GBR (please add it).

How does the high number of parameters to be calibrated in the ML model compare with the number of available data? If the number of data is not large enough, the model is intrinsically prone to overparameterization.

Did the authors test different hyperparameters for LSTM, e.g. smaller number of neurons and layers?

2) Intrinsic variability in the model’s results

The authors analyze the variation of the results obtained in the testing period (2019-2020) when shuffling the training years (section 3.4), and of other possible modifications of the dataset, e.g., by artificially reducing the frequency of the data. As I already mentioned, this is very important, and the analysis is well conceived.

Nevertheless, single realizations of ML models may provide non-optimal results. For this reason, it is a common practice to repeat ML runs several times and then average the results (e.g., Piotrowski et al., 2021; Yousefi and Toffolon, 2022). Did the authors account for this?

-- Minor remarks and typos

The Supporting Information contains some data and plots that would fit well in the main text. For instance, Table S1 is useful to understand the procedure used in the analysis.

L270, “Even the LSTM algorithms could not account for previous condition so far back in time”. How long is the expected memory of the model?

L56 “beings”: begins

L136 “modified Kappa” is not a common metric. Please give a short description of what it represents.

L 228 “even though the GBR model usually performs better in Fig. 5c the testing period chosen for use in Fig. 3, showed the opposite result.” Cumbersome sentence, please rephrase it. Moreover, the whole section contains a weird use of commas.

Figure 4. The subplots (b) and (c) are not described in the caption.

Text S3, reference to Wilson et al. (2020): it is not in the bibliography.

Figure S3, specify that boxplots refer to the period 2004-2020. Maybe it would be more interesting to compute the boxplots excluding 2019-2020?

Figure “Penal”: Panel

-- References

Piotrowski, A.P., Osuch, M., Napiorkowski, J.J., 2021. Influence of the choice of stream temperature model on the projections of water temperature in rivers. J. Hydrol. 601, 126629. https://doi.org/10.1016/j.jhydrol.2021.126629.

Yousefi, A., Toffolon, M., 2022. Critical factors for the use of machine learning to predict lake surface water temperature. J. Hydrol. 606, 127418. https://doi.org/10.1016/j.jhydrol.2021.127418