Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-4-RC2, 2020

Geoscientific Model Development Discussions

Interactive comment on "ML-SWAN-v1: a hybrid machine learning framework for the prediction of daily surface water nutrient concentrations" *by* Benya Wang et al.

Anonymous Referee #2

Received and published: 2 June 2020

The manuscript presents a useful approach to simulate high temporal resolution water quality data required as inputs to receiving water modelling. In my opinion the work is publication worthy and generally well written.

The hybrid approach is novel as far as I'm aware, to predict a one water quality constituent for the current time step, as an input to predict a different constituent. The results demonstrate that this improves model performance compared to predicting the target constituent directly. This suggests that the different water quality parameters are highly correlated, and provide additional information that is not contained within the other inputs (of baseflow, quickflow, rainfall, day of the year and lags or averages of

these). Two additions are recommended related to this observation: 1. Presentation of the relationships between in the inputs and outputs would be informative. As noted in the discussion (line 445) correlation tests can be used to assess the which parameters would be useful for pre-generation. Potentially a scatter plot matrix, with correlation values on the lower diagonal, would be useful to show the relationships concisely. 2. It would be useful for the authors to speculate in the discussion, or even undertake the modelling, if training the model on the observed values for the hybrid model, rather than the generated values, would further improve performance. Possibly this would allow the data driven model to fit the underlying relationships more accurately, even if when implemented less accurate predicted values must be used as the inputs. Or possibly the hybrid model is correcting for some of the errors in the generated inputs if there is a systematic bias, and this would degrade performance.

Line 200: Please provide further information on how the 80:20 data split was implemented. Was the last 20% of the time series used for testing, or was a more complex method used? If this was the approach, does this explain why the DON, and DOC for Murray River, was the most important input, because this data is only present in the testing period? As noted in the manuscript, there appears to be a long term trend in the dataset, and only using this data that is more representative of the testing dataset may improve performance? Possibly data from the 1990s does not represent the responses experienced in the more recent testing period? The authors should further comment on these temporal and data length issues.

Line 316: claims that the generated data enables the hybrid ML to capture long-term trends. Please elaborate on how this is possible, as it is not clear to me. A given set of inputs to the stand alone ML to predict the generated data will give the same result in 1990 as 2018, so how can long term trends be captured? As outlined above, possibly because using inputs only collected in the more recent period is more representative of the testing data?

Section 4.2 moves from considering all nutrients to TN only. Please add a justification

why this was the case. Is this because TN is the more relevant for the application, or for the sake of brevity only one parameter is analysed in more detail.

Analysis of the relative importance of the different inputs is a valuable addition to the manuscript, which is often not considered in machine learning work. Overall, I consider this work of publishable merit subject to these minor additions and clarifications.

Minor points The first sentence of the abstract would benefit from being rewritten. Suggest making this two sentences and reworking, e.g. nutrient data is monitoring, not necessary for monitoring, for example.

I found the term "temporal data" confusing, I was assuming this meant lagged data, Qb,t-1, for example. It wasn't until I found Table 1 that this means Julian day of the year. It is suggested to be more specific, and call this "Julian day" or "seasonal component" or similar.

Line 70: Another relevant study for hybrid modelling may be Hunter et al. (2018) https://doi.org/10.5194/hess-22-2987-2018

Line 164: remove are from GBM model are generally have less. . .

Line 214: . . .. Can be divided into there stages, add s to stages.

Line 268: I was confused by this sentence, in that lower RMSE and increased MEF are both improved performance, but opposite patterns. Suggest "overall, the scaled RMSE improved from LM. . .. And the same pattern was found form MEF. . .

Line 277: suggest add to the end of this paragraph, "as such, the WRTDS results are not directly comparable to the other methods"

Line 389: the temporal data was more useful for the perennial catchment, Murray River. Could this be because seasonal information is captured in other inputs, e.g. Q for the more ephemeral catchment?

Line 399: Please elaborate further on the method used in this section, it is not clear

what has changed. Should "gradually" be "sequentially"?

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-4, 2020.