Dear authors,

thank you very much for the revised version of your manuscript. Your revisions addressed almost all of the reviewer comments and your manuscript is close to publication now. Reviewer #1 mentions a few minor issues which I would like you to address in one further revision of your manuscript (see corresponding listing of reviewer #1).

If you have any further questions regarding the revision, do not hesitate to contact me directly.

Thank you and best regards,
Wolfgang Kurtz

Response: We thank the editor for these encouraging comments. We have revised the manuscript, supplementary information and figures as per the comments of reviewer 1. Our point-by-point responses to the reviewers' comments are provided below.

Reviewer 1

This revised version presents a significant improvement, including a thorough consideration of the review of the original version. In particular, the framework and code examples are made available to any ambitious reader, such that reproducibility should no longer be an issue, and also the suitability of AI4Water to embrace typical user data is made explicit now.

Response: We thank the reviewer for these comments. Please find our point by point response below.

There are only a few minor and mostly trivial points remaining:

- a small number of typos and grammar errors, e.g. l. 115: "categroical" -> "categorical" or line

337: "indicate" -> "indicates" and a few others, in particular in newly added text. Please re-read carefully and correct accordingly.

Response: We have corrected these typos in the manuscript. We have also reviewed the manuscript for any further typos and have corrected them.

Lines 115: For example, the *hyperopt* sub-module presents the *Real*, *Categorical*, *Integer*, and *HyperOpt* classes.

Lines 337-338: In Fig. S10, a large horizontal bar for a given feature indicates that this feature strongly affected the model's prediction.

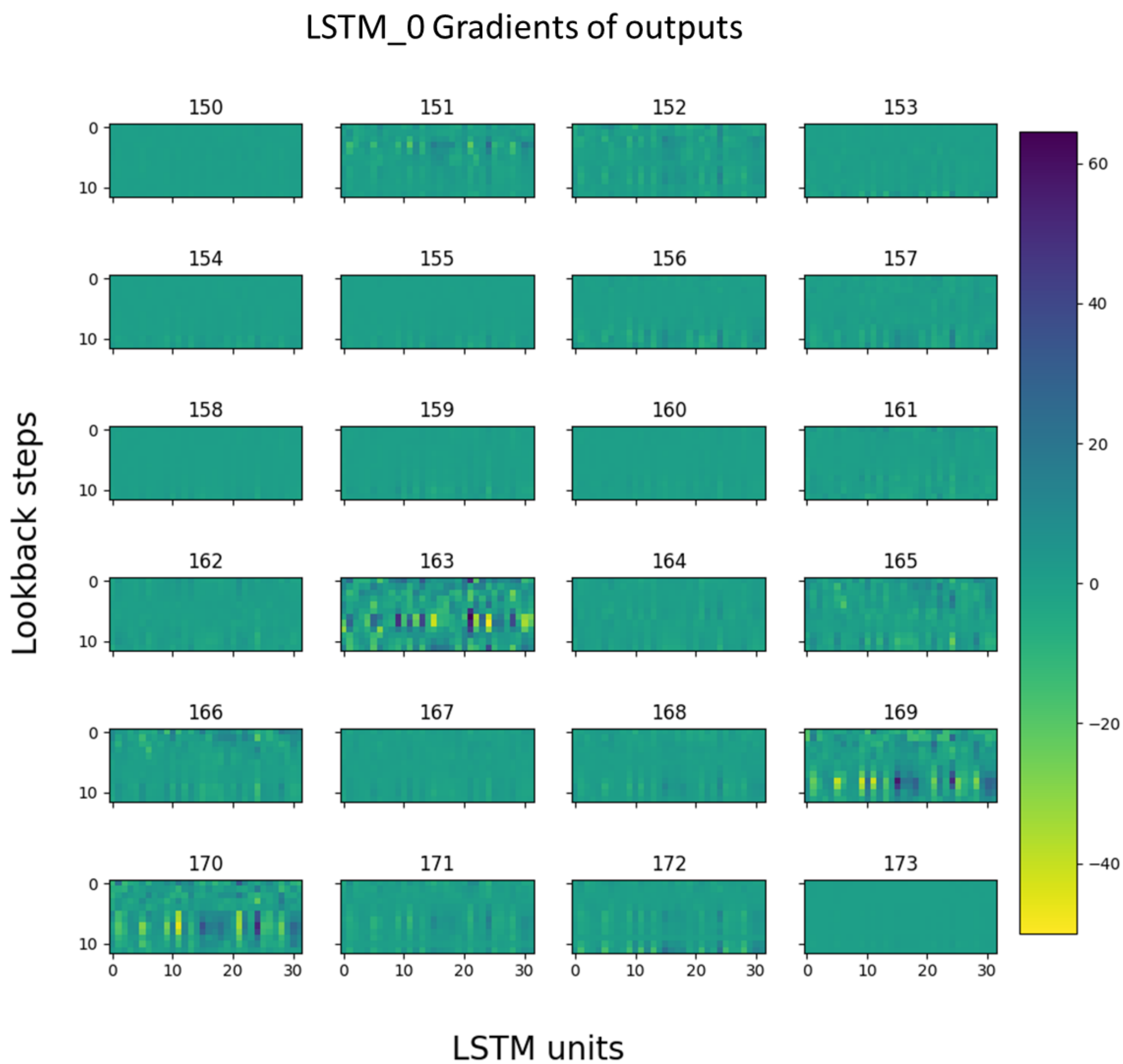
- Supplement, Fig. S1: could you please explain what a "Dense layer" is? It is nowhere mentioned in the main text and not here as well.

Response: A "Dense layer" is a fully connected layer which is a linear multiplication of weights with outputs of preceding layer. The purpose of "Dense" layer is to reduce the dimensions of data to match it to the output size. We have elaborated this in manuscript and updated the following lines.

Lines 303 – 304: The four-layered neural network comprises an input layer, two layers of LSTM and a dense layer as output layer (Fig. S1). The Dense layer is a fully connected layer which is used for dimensionality reduction (Chollet, 2018).

- Fig. S3: despite the continous color bar, there seem to be only two colors present - dark red and dark blue - with the exception of day 157. Is there a problem with the range of the colorbar, or some other error here?

Response: We thank the reviewer for pointing out this mistake. The binary representation in Fig. S3 is because of selection of very small range of values (vmin and vmax parameters) during plotting. We have corrected this bug in the code. We have also updated Fig. S3. The colors in updated Fig. S3 are continuous and not binary. The updated Fig. S3 is as below



- Figure S6: the unit of streamflow can't be "m s^-1", very probably it is "m^3 s^-1", this would also corrspond to "cms" used in Fig. S8 which is not explained but very likely means "m^3 s^-1"

Response: We have corrected the units on x and y-axis in Fig. S6 to $m^3 s^{-1}$ which stands for cubic meter per second. We also explicitly mentioned in captions of Fig. S6 and Fig. S8.

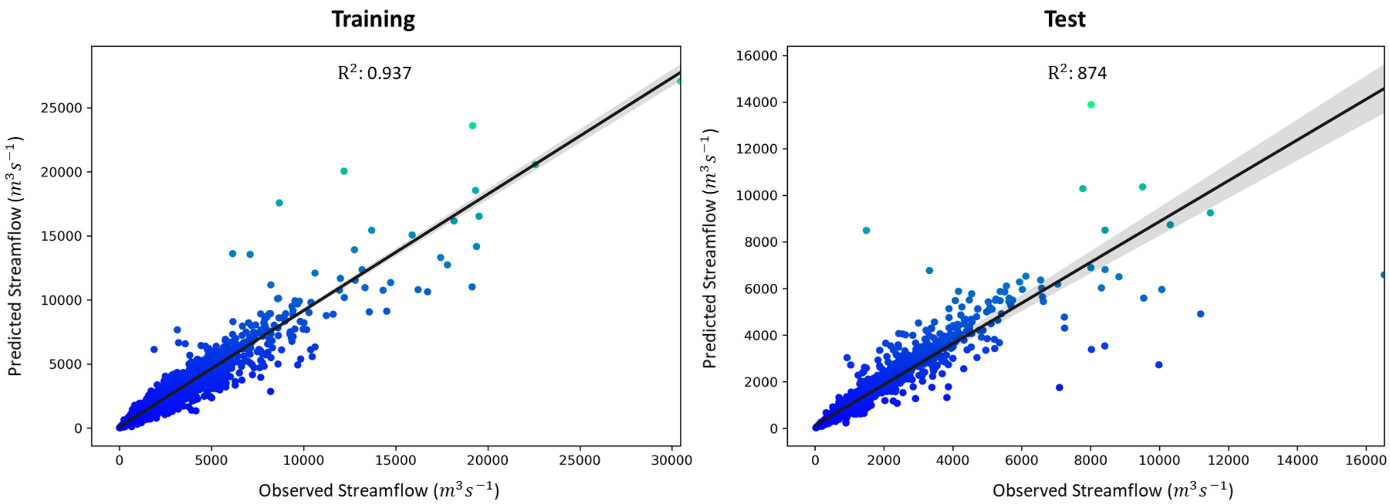


Figure S6: Scatter plots between observed and predicted streamflow for training and test dataset for rainfall-runoff modeling in catchment number '401203' of CAMELS Australia dataset. The units of streamflow are cubic meter per second (m^3s^{-1}).

- Figure S11 (a): there is no scale for f(x), only a horizontal line, presumably indicating zero. Is the model quality related to the distance of f(x) to this zero line? What is f(x)? The explanation "f(x) indicates model's prediction" is very vague.

Response: In Fig. S11 (a), the 52 test examples are sorted by the sum of SHAP values over all features. This results in examples with large positive SHAP values on left side and those with large negative SHAP values on right side. The f(x) in Fig. S11 (a) represents sum of SHAP values of all input features for each of the 52 test examples. The horizontal line represents 0 value. The prediction of machine learning model is equal to sum of these SHAP values and base value. This is because, as per Lundberg and Lee (2017), SHAP values explain model's prediction starting from base value. The base value is mean of total predictions from model on training data (Lundberg et al., 2018). In our example the base value was 4661.082 MPN100 mL-1. Negative f(x) for a given example indicates that the sum of SHAP values of all input features for this example is negative. However, the prediction from model is positive because of addition of base value. We have added labels for y-axis to indicate scale of y-axis. We have also modified the caption of this figure to elaborate these points.

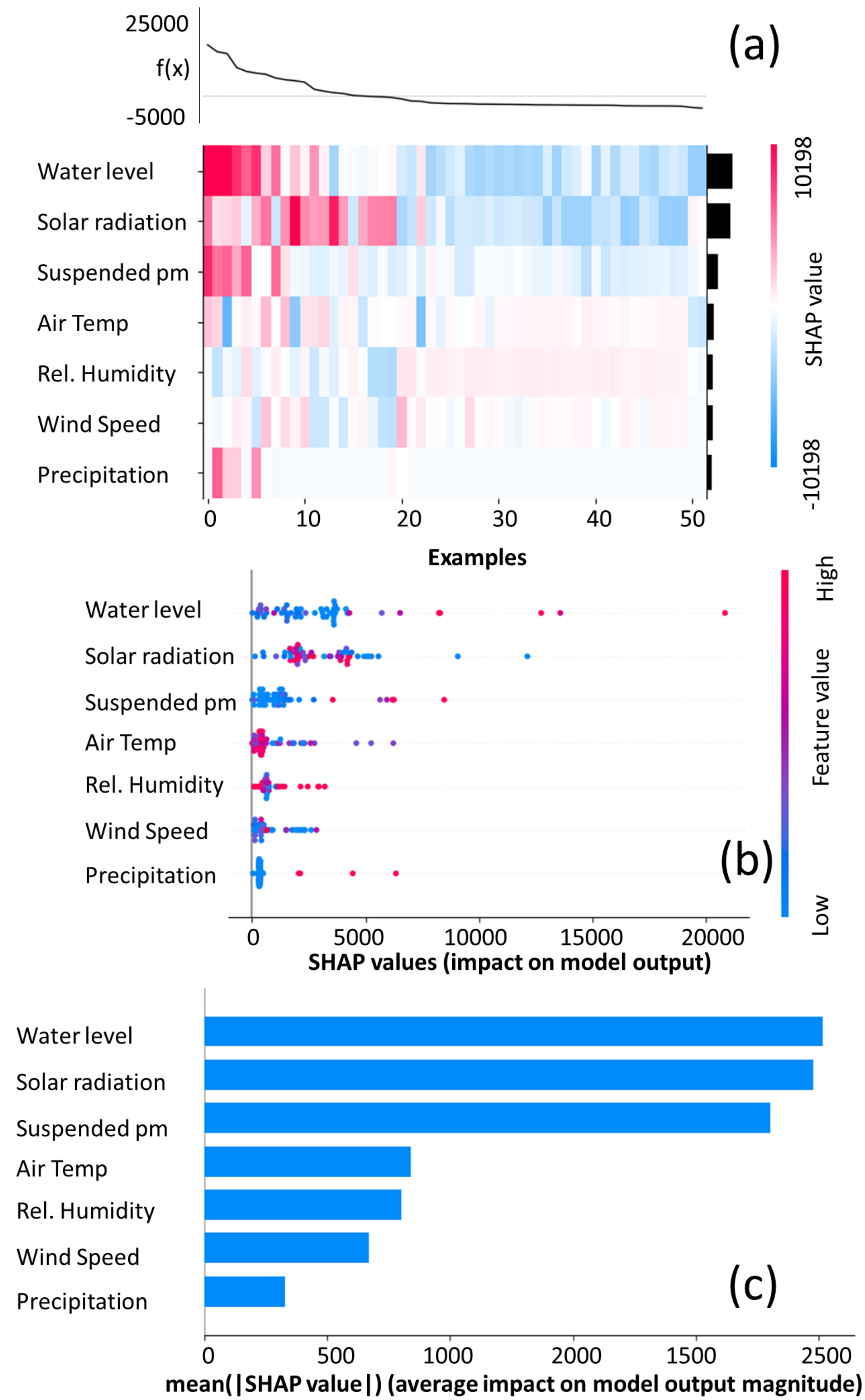


Figure S11. Explanation of XGBoost model for *E. coli* prediction using SHAP method. The explanations show the importance for each input feature by the model. (a) SHAP values as heat map. $f(x)$ indicates sum of SHAP values of all input features. The prediction of machine learning model is equal to sum of $f(x)$ and base value. The base value is the sum of models' prediction on training data which was 4661.082. (b) SHAP values for individual examples in test data, (c) global feature importance based upon SHAP values.

The paper is very close to being publishable. The reviewer expresses hope that AI4Water will be used in the hydrology community soon.

References

Chollet, F. (2018) Deep learning with Python, Manning Publications Co.

Lundberg, S.M., Erion, G.G. and Lee, S.-I. 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Lundberg, S.M. and Lee, S.-I. 2017 A unified approach to interpreting model predictions, pp. 4768-4777.