Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2021-124-RC1, 2021

Comment on hess-2021-124

Anonymous Referee #1

Referee comment on "Evaluating different machine learning methods to simulate runoff from extensive green roofs" by Elhadi Mohsen Hassan Abdalla et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-124-RC1, 2021

Summary
This paper compares the performance of four machine learning algorithms (including a deep learning one) in simulating runoff from green roofs, and provides their benchmarking by also utilizing a conceptual model. The comparison is conducted by using data from sixteen green roofs located in four Norwegian cities, and the comparted algorithms are the Artificial Neural Network (ANN), M5 Model tree, Long Short-Term Memory (LSTM) and k-Nearest Neighbour (kNN) ones. Additional investigations focus on the transferability of the algorithms between different green roofs. The results show that the performance of the investigated algorithms is acceptable; however, the conceptual model should be preferred over the transferred machine and deep learning algorithms.

General comments
Overall, I believe that the paper is meaningful, interesting and mostly well-written with room for improvements.
Although my comments are quite few, I recommend major revisions, as the suggested improvements (mainly those prescribed with specific comment #1) are both important and necessary, to my view, for the model comparison (and the entire paper) to reach the best possible shape.

Specific comments
1) In line 246, it is written that the "methods were evaluated based on the performance on the validation data sets". However, in line 221 it is written that "to avoid overfitting, the performance of changing hyperparameters was observed in the validation periods". As the validation set has been used for hyperparameter selection (i.e., for identifying the best version of its machine learning algorithm), the addition of an extra independent set (i.e., a test set that is not used for model selection) is necessary here. This extra set will serve the independent comparison between machine learning algorithms, as well as the independent comparison between machine learning algorithms and the conceptual model. Therefore, the datasets should be divided into (at least) three independent sets (including different data points), i.e., the training, validation and test sets.
2) Moreover, it would be better (but not strictly necessary, to my view) that the datasets are divided into four independent sets (i.e., the training, validation 1, validation 2 and test

sets), as time lag selection also takes place according to the following lines: “Secondly, the structural parameters were fixed, and different lag values ranging from 1 hour to 200 hours were tested to identify the optimal lag value” (lines 219−220).
3) In lines 217−216, it is written that “BERG1, OSL1, SAN1 and TRD1 roofs were selected to test different hyperparameters to find the optimal parameters for each city”. Would it be better to select different hyperparameters for each roof?
4) In lines 209−211, it written that “data were aggregated into one-hour resolution, and snow accumulation periods were excluded (1 Oct. – 31 Mar.). One year was used for training and one year for validation. The selection of the training year was based on the sum of precipitation as the wettest year between 2015 to 2017 for each roof, and the second wettest year for validation. The rationale for the selection is that the wettest year covers a broader span of precipitation events which improves the generalization performance of the models”. To my view, it would be better if the training and validation periods for all greens roofs were presented in a new table.
5) Also, I think that –at least in the supplement− it would be interesting to show what happens when one uses the entire datasets (i.e., without excluding the snow accumulation periods or other periods), and not selected parts of these datasets.
6) I find that some important literature pieces on data-driven hydrological modelling (e.g., some of the oldest works in the field) are currently missing from the manuscript’s reference list.
7) Lastly, since the manuscript is not typo-free at the moment, a careful reading and typo correction are required. For instance, something is currently wrong with the sections numbering (“2 Data”, “2.1 Machine learning models”, “3 Results and Discussion”). Also, there are typos in the units, symbols and equations, which should be written according to the following conventions:
• Single-letter variables should be written in italics.
• Multi-letter variables should not be written in italics.