Dear Referee,

We would like to thank you for the thoughtful comments which will contribute towards improving the manuscript.

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.

We appreciate the positive comment about the study

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.

We thank the reviewer for this thoughtful comment which will be included in the revised manuscript. We would like first to clarify the old setup that we applied in the study. We only optimized ML hyperparameters for 4 roofs (one roof in each city). The validation data sets of the four selected roofs were used for model selection (i.e. hyperparameter tuning). For the other 12 roofs in the study, the validation data sets were not used for model selection. However, we agree with point #3 that a hyperparameter tuning should be done for each roof. Therefore, we will modify the results by dividing

our data into three sets for model training, hyperparameter tuning and comparisons as suggested by the reviewer (Figure 1)

old setup			
GR	Data set		
	Training	Validation	Testing
BERG1	2017	2016	2015
BERG2	2017	2016	2015
BERG3	2017	2016	2015
BERG4	2017	2016	2015
BERG5	2017	2016	2015
OSL1	2015	2017	2016
OSL2	2015	2017	2016
OSL3	2015	2017	2016
SAN1	2017	2016	2015
SAN2	2017	2016	2015
SAN3	2017	2016	2015
SAN4	2017	2016	2015
TRD1	2017	2016	2015
TRD2	2017	2016	2015
TRD3	2017	2016	2015
TRD4	2017	2016	2015

new setup			
GR	Data set		
	Training	Validation	Testing
BERG1	2017	2016	2015
BERG2	2017	2016	2015
BERG3	2017	2016	2015
BERG4	2017	2016	2015
BERG5	2017	2016	2015
OSL1	2015	2017	2016
OSL2	2015	2017	2016
OSL3	2015	2017	2016
SAN1	2017	2015	2016
SAN2	2017	2015	2016
SAN3	2017	2015	2016
SAN4	2017	2015	2016
TRD1	2017	2016	2015
TRD2	2017	2016	2015
TRD3	2017	2016	2015
TRD4	2017	2016	2015

Model training
Hyperparameter tuning
independent model evaluation
not used

Figure 1: Old setup vs new setup

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).

We agree with the reviewer that, having a fourth data set could improve the selection of the time lag values. However, we have decided, following the major comment of Reviewer#2, to redo the hyperparameter tuning using Bayesian optimization (Snoek et al., 2012) which is expected to improve the estimation of the hyperparameter values. The time lag will be considered as a tunable hyperparameter, following the study of Kratzert et al. (2019).

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?

We agree with the reviewer that it is better to tune hyperparameters for each roof individually. Accordingly, we will optimize the hyperparameters for each roof using Bayesian optimization.

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.

A table will be provided to present training, validation and testing periods for all the roofs

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.

We initially used the entire data set for ML model training and validation. However, we have decided to remove snow periods to allow for comparison with the benchmark model (the conceptual retention model) which does not account for snow modelling.

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.

We are not aware about the literature pieces that is referred to by the reviewer. We attempted to present a literature review that balances between green roof modelling and the application of ML in hydrological modelling. We have mentioned some of the early work in ML modelling in hydrology e.g. (Daniell, 1991; Hsu et al., 1995; Karlsson & Yakowitz, 1987).

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.

We apologize for any typos in the manuscript and we will modify the errors identified by the reviewer.

References

Daniell. (1991). Neural networks. Applications in hydrology and water resources engineering. *National Conference Publication - Institution of Engineers, Australia*.

Hsu, Gupta, & Sorooshian. (1995). Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water Resources Research*. https://doi.org/10.1029/95WR01955

Karlsson, & Yakowitz. (1987). Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. *Water Resources Research*, *23*(7), 1300–1308. https://doi.org/10.1029/WR023i007p01300

Kratzert, Klotz, Shalev, Klambauer, Hochreiter, & Nearing. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-23-5089-2019

Snoek, Larochelle, & Adams. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*.