In the manuscript, the authors evaluated the performance of 4 machine learning (ML) approaches in simulating runoff from green roofs. The experiments include 16 green roofs in 4 cities which are of different environmental conditions. Upon modeling retention of green roofs, The authors highlighted the advantage of the ML methods in comparing with a conceptual model. Upon modeling runoff, most models achieved a promising performance (NSE>0.5). The authors also examined the transferability of ML models between green roofs and concluded that those models could be transferred between cities with similar rainfall events characteristics.

As an extra reviewer who accidentally reviewed the out-of-date version, I am impressed by the substantial amount of work that the authors have done to improve the manuscript. The current version of the manuscript is clear-written and resolved most concerns I had in the previous version. I found several technical issues that I will detail following, but most of them are easy to fix in my opinion.

We appreciate the positive feedback of the reviewer. We acknowledge that the thoughtful comments of the reviewers in the first round have significantly contributed to improving the quality of the study. This indeed highlights the vital role of the peer-reviewing process.

L25: I like the highlight of retention and detention, as the first round manuscript did.

Done

L190: The author should write clearly about how the distance is calculated in this work, instead of "such as".

Done

L210: The description does not agree with the precipitation amount in table 2. For example, at Bergen and Oslo the drier year is chose as the training set.

The initial selection of the training periods was based on the amount of precipitation. However, after hyperparameter tuning, we further analyzed the change of ML performance when using the initially selected validations datasets for model training. Some of the validation datasets slightly improved the ML performance and hence were selected as training datasets. The final selection of the training, validation and testing periods is presented in table 2. This has been stated in the MS in lines 209-215.

L234: The dropout citation is not correct here as it is already so widely used before the cited work.

Done. We cited the original paper

L331: Table 5 presented the testing error?

Yes. We clarified that in the revised MS

L335: Typo. Reads BERG2 here but BERG1 in figure 6?

Done

L344: The author only explained why models on Bergen are of better performance comparing to Oslo, but did not reason upon "other roofs in the study" as claimed.

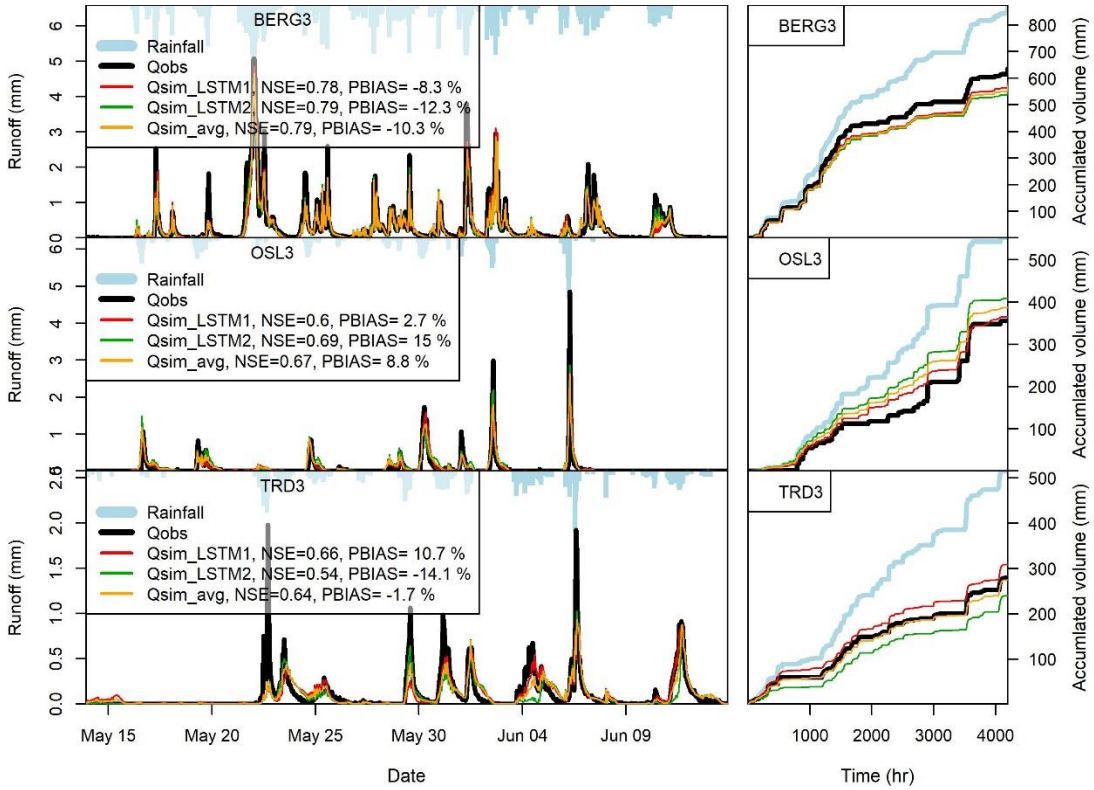
Corrected

L351 and figure 7: The author implied that the TRD site is heavily impacted by snow melting here, which is supported by the calibration results in table 4 (preferred longer lag time). However figure 7 did not present which season/month/date it plotted, and readers who are not familiar with the climate of Norway (like me) may have a hard time understanding the effect of snow storage.

Done.

L362: I am wondering if the conclusion is true for all sites? The three sites presented in figure 8 happen to have positive and negative biases for the two training years, and maybe that is why the sum of two models outplay either one of them. Are there any sites that the two models from two years result in biases of the same direction?

we found few green roofs where the two LSTM models from the two years resulted in biases of the same direction (at BERG3 and OSL3 roofs, as shown below). We clarified that in the revised MS lines 364-366



L383: In the manuscript, the author did not include a process-based or conceptual model upon detention process, and as a result, the work only proved that ML models are better at simulating retention process comparing to conventional methods, other than simulating the runoff. The reason seems to be the detention models are not "convincing" according to the manuscript. However, the authors reviewed detention models using three paragraphs (L39-57) in the introduction. Are none of those reviewed models "convincing"?

The reviewed detention models rely on calibration to estimate their parameter values. Previous studies have highlighted the limitation of transferring calibrated parameters between similar roofs

(Johannessen et al., 2019) and difficulties of identifying clear relationships between model parameters and roof characteristics (Kasmin et al., 2010). The ML methods have shown good potential in modelling detention (high NSE) and transferability between cities with similar climatic condition.

Table 1: There are two BERG2.

Corrected

Figure 4: I would suggest labeling train, validation, and test for each plot.

We prefer to keep the current labelling because the final selection of the training, validation and testing periods is different among the cities (table 2).

figure 6: I think the author presented the testing period other than the validation period? And which are the three months?

Corrected

In general, I find the manuscript is well-written and could be a novel contribution to the community. I have to acknowledge that I am not an expert in green roof modeling, and I may not fully understand the background and contribution of this manuscript, so please consider my review accordingly.

We thank the reviewer for the thoughtful comments and the positive feedback

References

Johannessen, Birgitte Gisvold, Hamouz, Vladimír, Gragne, Ashenafi Seifu, & Muthanna, Tone Merete. (2019). The transferability of SWMM model parameters between green roofs with similar build-up. *Journal of Hydrology*, *569*(October 2018), 816–828. https://doi.org/10.1016/j.jhydrol.2019.01.004

Kasmin, H., Stovin, V. R., & Hathway, E. A. (2010). Towards a generic rainfall-runoff model for green roofs. *Water Science and Technology*, *62*(4), 898–905. https://doi.org/10.2166/wst.2010.352