Dear Referee,

We would like to thank you for the thoughtful comments which will contribute towards improving the manuscript. The suggestion of automatically optimizing the hyperparameters of ML models has inspired us to navigate many optimization algorisms for hyperparameter tuning that are rarely applied in hydrological modelling studies.

First, I want to apologize with Authors due to my late review. It was due to unexpected issues. The present study presents a numerical analysis to compare the performance of multiple Machine Learning techniques against conceptual models for the hydrological analysis and forecasting of Green Roofs behavior. The aim of the paper is interesting and of relevance for HESS readers.

We appreciate the positive comment about the study.

However, I find that the paper has multiple weaknesses:

- *There are multiple bold statements against the use of physically-based models for GRs analysis, which are not supported by evidence and not needed in the manuscript, which should simply attain to its aim: assessing the performance of ML techniques for GRs analysis. Instead of reinforcing the paper, these statements draw the attention on other aspects, which are highly debatable. There doesn't exist a perfect numerical tools for everything, or one better than the other. It's up to the modeler to choose the right model for the specific modeling task.*

  We can agree that some statements in the current manuscript are debatable and not needed for the study. We will modify the manuscript accordingly.

- *The emulators training is performed by using the trial-and-error technique, which is an outdated and inefficient methodology. This is especially true for this task since the response surface in the hyperparameters' space can be multimodal, thus making it easy to get trapped in local minima. Furthermore, the uncertainty of the estimated hyperparameters should be properly assessed and eventually propagated in the validation step. The way it is handled in the paper (manually changing hyperparameters) is weak.*

  This is an excellent comment that has really inspired us to search for suitable algorithms for hyperparameter tuning. We acknowledge the limitation of trial-and-error techniques using random or grid sampling which we applied following recent hydrological modelling studies (King et al., 2020; Kratzert et al., 2018, 2019; Sattari et al., 2021; Teweldebrhan et al., 2020; Mo Zhang et al., 2020). Based on the reviewer comments, we decided to apply Bayesian optimization for hyperparameters tuning (Snoek et al., 2012) , which is suitable for functions in which evaluating one set of parameters is expensive and time-consuming. This algorithm was applied by Worland et al. (2018) to optimize hyperparameters for several machine learning algorithms to predict low flows for ungauged basins. In Bayesian optimization, the objective function (i.e. the relation between hyperparameters and the error value in the validation data set) is approximated by a probabilistic model (e.g.

Gaussian processes) that is used to select the most promising hyperparameter to evaluate in the true objective function.

- *Since authors calibrate (manually, but still calibrate) the emulators and compare it with a conceptual model, then the latter should be calibrated as well to conduct a fair comparison. This was not done.*

The conceptual green roof retention model describes the processes that control moisture removal from the green roof substrate during dry weather periods. The model is implemented using a physically-based estimate of ET and a measurable physical property of the substrate, Smax. The parameter Smax represents the maximum retention capacity of the green roof or the difference between the field capacity and the permanent wilting point of the green roof substrate (Stovin et al., 2013). There exist standard laboratory tests to physically measure the substrate field capacity (FLL, 2008) and the permanent wilting point (Fassman & Simcock, 2012). It may be argued, therefore, that our conceptual model is physically based, and should not require calibration. No empirical parameters are introduced into the model that require calibration. However, we acknowledge that the Smax values were estimated in this case, and that it is relevant to confirm the extent to which calibration of this parameter would further improve the predictions. Therefore, we agree to calibrate the conceptual model (by tuning the Smax value) for each roof in the study. We will use the data of the training year for Smax tuning and the testing year for model evaluation and comparison with the ML models. As the conceptual model is cheap to run, we could run the conceptual model for all roofs by varying Smax between 1% to 100% of the roof substrate as shown in Figure 1.

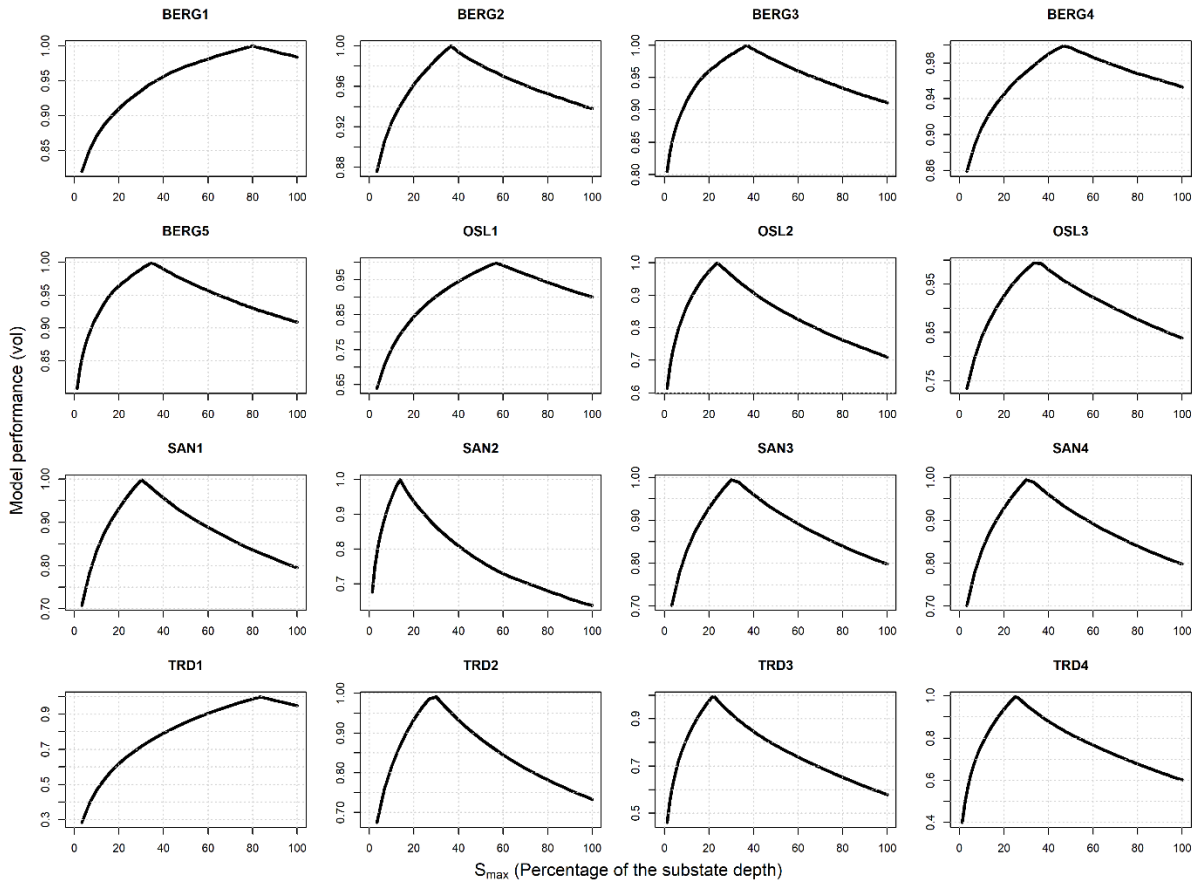

Figure 1: Tuning of Smax. vol is a measure of the volumetric error (equation 15 in the manuscript)

Specific Comments:

L2-5 In my opinion, there is a general misunderstanding in this field, which is reiterated in multiple manuscripts, and it's the idea that conceptual models are always computationally cheaper than physically-based models for the hydrological analysis of GRs. Except particular circumstances, the computational cost is comparable. For instance, the authors can verify by themselves that HYDRUS-1D, a mechanistic hydrological model frequently used in GR analysis, takes less than few seconds for a long-term hydrological simulation. Conversely, for the same task, some conceptual models can be even more computationally expensive if the code is developed in excel or in high-level programming languages. Therefore, I would not build the premise of the work on this.

L2-5 Regarding the complexity, we should first define what is complexity (number of parameters, number of processes, etc). This is again questionable.

> As mentioned earlier, we agree to modify the manuscript by removing debatable statements about physically-based models which is not needed for the study.

Measurements: This is true and implies that conceptual models are not easily generalizable.

L20-25 “Improving quality” is a bold statement. There is an extensive literature about nutrients leaching from GRs.

This will be modified accordingly in the manuscript

L30 Why bold font?

We will remove the bold font

L35-40 I don’t agree with these statements. Mechanistic models actually rely on huge literature body, which can be used to set the model parameters. For instance, parameters of the van Genuchten can be obtained with pedotransfer functions (using particle size distribution and other info from the producer) or set according to several studies which have been already performed. The unsaturated conductivity is needed as the soil water retention curve in the Richards equation, there is no difference. What is the acceptable level of uncertainty depends on the analysis (in dry conditions the magnitude of fluxes is low thus K is not prominent).

L55 Computational cost: As I stated before, I don’t agree with this.

As mentioned earlier, we agree to modify the manuscript by removing debatable statements about physically-based models which is not needed for the study

L75-80 MLs are not uncertainty-free.

We agree. We intended to say that ML models reduce (not eliminate) the structural uncertainty that is caused by inadequate model assumptions (i.e. mathematical equations) in comparison to conceptual or physical models. We acknowledge that the performance of ML models is affected by the selection of suitable hyperparameters, which is a source of uncertainty. We will modify this part.

L115-120 “Green Roof runoff” should be “Green Roof subsurface runoff” to avoid misunderstandings.

We agree. This will be modified accordingly in the manuscript

I would just say “ when observations are not available”

We agree. This will be modified accordingly in the manuscript

L168. “Trial-and-error” This is not true. A correct ANN training should use numerical optimization to identify the right set of hyperparameters since

As mentioned earlier, we will apply the Bayesian optimization for hyperparameter tuning

Section 2.2 I’m not sure that you can basically neglect physical properties of GRs. This might be somehow borderline acceptable for extensive GRs but morphological and hydraulic characteristic will play an important role as the soil substrate depth increases. This is acknowledged also in one of latest

paper from the same authors (Peng et al., 2020), and it is rather intuitive. I would be curious to see how the emulators behave when splitting the sample between thin and thick roofs. This would certainly deliver a more meaningful information to the community.

This was done via transferring trained ML models between the green roofs. ML models that are trained using the data of thin roofs, even located in different locations, were used to simulate outflows of thick roofs and vice versa.

L210 The validation should be performed on a drier year to really assess the generalizability of emulators.

We trained the ML models using the wettest years and was validated on the drier year

L210-215 The optimal hyperparameters should be calibrated numerically, since you can easily end up trapped in a local minima (10.1016/j.jhydrol.2005.03.013). This is true for all emulators.

The use of Latin hypercube doesn't make solve the problem. You have a better coverage of parameters' space but, unless you use a global optimization strategy, you can be still trapped in local minima.

As mentioned earlier, we will apply the Bayesian optimization for hyperparameter tuning

L220 What are the structural parameters?

This is meant to refer to the hyperparameters that describe the structure of the network (i.e. number of neuron and hidden layers) to differentiate them from other hyperparameters such the length of input sequence (lag).

L221 What you attempt to do is to investigate how small changes in hyperparameters affect the response of the emulator. Basically, how the uncertainty in the estimated hyperparmeters (you see that ML techniques are not uncertainty free) propagates. This should have been done more correctly by numerically optimizing MLs parameters and estimating (at least) their confidence intervals. Even better would have been using Bayesian inference to estimate posterior uncertainty (e.g., 10.1016/j.jhydrol.2011.09.002).

As mentioned earlier, we will apply the Bayesian optimization for hyperparameter tuning.

L2.3 Why reporting all these equations, which are already mentioned in other studies from the same authors? Cite them and move forward.

This will be done in the revised manuscript

L228 “Without the need of prior calibration…” This sounds puzzling to me. In the Introduction you write “calibration is needed to find their optimal values, unlike physically-based models”, which is true since conceptual models generally needs site-specific calibration. If conceptual model parameters were not previously calibrated in other studies for the same site, then they should be calibrated here to conduct a fair comparison with trial-and-error optimized MLs.

As mentioned earlier, we will calibrate the conceptual model by tuning the Smax value

L3.1 For the reasons that I mentioned above, I consider this way of training emulators not formally correct and scientifically outdated.

As mentioned earlier, we will apply the Bayesian optimization for hyperparameter tuning.

L331 This can be said only when you perform a scientifically sounding calibration and uncertainty assessment of both models. None of the two was carried out, furthermore the conceptual model was not calibrated, thus the comparison is not fair.

This will be achieved after optimizing hyperparameters of ML models using Bayesian optimization and calibrating the conceptual model by tuning the Smax value.

L333-335 Not sure what you refer with “…accommodate complex, multi-layered systems”. These are bold statements not supported by evidence, which actually should be avoided since they don’t contribute to the discussion unless they are proven.

As mentioned earlier, we will remove any bold statements which are not important for the current study

References

Bengtsson, Grahn, & Olsson. (2005). Hydrological function of a thin extensive green roof in southern Sweden. *Nordic Hydrology*, *36*(3), 259–268. https://doi.org/10.2166/nh.2005.0019

Fassman, & Simcock. (2012). Moisture Measurements as Performance Criteria for Extensive Living Roof Substrates. *Journal of Environmental Engineering*. https://doi.org/10.1061/(asce)ee.1943-7870.0000532

FLL. (2008). Guidelines for the Planning , Construction and Maintenance of Green Roofing - Green Roofing Guideline.

(2020). Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-24-4887-2020

Kratzert, Klotz, Brenner, Schulz, & Herrnegger. (2018). *Rainfall – runoff modelling using Long Short-Term Memory (LSTM) networks*. 6005–6022.

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

Sattari, Apaydin, Band, Mosavi, & Prasad. (2021). Comparative analysis of kernel-based versus ANN and

deep learning methods in monthly reference evapotranspiration estimation. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-25-603-2021

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

Soulis, Valiantzas, Ntoulas, Kargas, & Nektarios. (2017). Simulation of green roof runoff under different substrate depths and vegetation covers by coupling a simple conceptual and a physically based hydrological model. *Journal of Environmental Management*. https://doi.org/10.1016/j.jenvman.2017.06.012

Stovin, Poë, & Berretta. (2013). A modelling study of long term green roof retention performance. *Journal of Environmental Management*, *131*, 206–215. https://doi.org/10.1016/j.jenvman.2013.09.026

Teweldebrhan, Schuler, Burkhart, & Hjorth-Jensen. (2020). Coupled machine learning and the limits of acceptability approach applied in parameter identification for a distributed hydrological model. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-24-4641-2020

Worland, Farmer, & Kiang. (2018). Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environmental Modelling and Software*. https://doi.org/10.1016/j.envsoft.2017.12.021

Zhang, Mo, Shi, & Xu. (2020). Systematic comparison of five machine-learning models in classification and interpolation of soil particle size fractions using different transformed data. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-24-2505-2020

Zhang, Zheng, Szota, Fletcher, Williams, & Farrell. (2019). Green roof storage capacity can be more important than evapotranspiration for retention performance. *Journal of Environmental Management*. https://doi.org/10.1016/j.jenvman.2018.11.070