Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-124-RC2, 2021

Comment on hess-2021-124

Anonymous Referee #2

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-RC2, 2021

General Comments:

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. 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.
- The emulators training is performed by using the trial-and-error technique, which is an outdated and inefficient methodology.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.
- 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.

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.

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.

L30 Why 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.

L75-80 MLs are not uncertainty-free.

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

- I would just say " when observations are not available"

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

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.

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

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.

L220 What are the structural parameters?

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).

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

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.

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the reasons that I mentioned above, I consider this way of training emulators not formally correct and scientifically outdated.

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.

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.