Comments

The main reason behind my decision is that the models proposed by the authors are developed and tested on a "well-instrumented site". What is the guarantee that the proposed model will work elsewhere when there is no lysimeter or EC data available for calibration? What parameters to use? To show the actual feasibility of this approach, the authors should have tested on other sites not used for calibration. What errors can we expect by applying parameters calibrated elsewhere? What are influential factors hindering generalization? Also, how are results affected by the individual mobile phone used?

Our goal with the manuscript is to show that for a particular site, and even with limited training, smartphone-based observations can provide ET estimates that are as accurate as estimates of lysimeter ET by EC and vice versa. We think this (preliminary) finding is important. With this manuscript, we want to invite other groups to start exploring the use of smartphone-based ET estimation so that we can find answers to the questions posed by the reviewer. Our goal is NOT to claim that the model we used should somehow be the basis of all future smartphone ET estimation. We do believe in the general principle of combining smartphone observation with machine learning, but future studies will be able to do so with more training data, and more sophisticated ML methods. This will be made clearer in a revised version. We are currently exploring the accuracy of the initial simple model when applied to other sites. We are happy to include a validation on a different site (likely at one of the TERENO lysimeter sites in Germany) in a revised version as suggested by the reviewer. A full quantification of errors induced by generalization will however not be feasible, as there will likely be important vegetation effects which can not all be investigated with the lysimeters currently available. But it should be noted that other ET estimation methods, such as those based on airborne or spaceborne thermal remote sensing, will suffer from similar if not larger uncertainties that are rarely quantified. Based on our experience, we believe the effect of individual mobile phones is small compared to the uncertainties associated with the Weatherflow sensor. The FLIR Lepton sensor is generally accurate, and within-sensor variability is likely small compared to the averaging/sampling error made when estimating the average surface temperature over the lysimeter or EC footprint. This will also be discussed in more detail in a revised version.

Other issues of relevance include the inappropriate use of the terms "machine learning" to define multivariate regression done with very limited data (machine learning is data hungry), as well as the lack of information concerning the rationale behind the formula used. Other detailed comments follow below.

Multiple linear regression is typically seen as the simplest form of ML, and therefore the best given the limited amount of data we currently have. However it should be noted that the results are robust, as is shown in Fig 4 by the relatively small spread in the distributions of the model parameters after repeating the fitting 2000 times with different samples. The model was chosen because of its simplicity, and because many of the existing ET formulas are (near)linear combinations of global/net radiation, temperature, and humidity. Again, it should be noted that we do believe in the general principle of combining smartphone observation with machine learning, and that future studies will be able to do so with more

training data, and more sophisticated ML methods. This will be made more clear in a revised version.

Lines 24 - 32: Can the authors refer to existing literature when highlighting these gaps?

This was also noted by the other reviewer. We will add more references in a revised version.

Line 44: built-in

Thanks for the correction.

Line 48: not sure referring to the figure without a thorough explanation is suitable at this point in the introduction. Better in the methodological section?

Good point. We will reconsider the referencing to the figure.

Lines 49-50: "[...] the question is how these estimates can work in concert under field conditions to produce accurate ET." this part is not sufficiently clear.

This will be rephrased.

Line 54: RQ1 seems to be very generic; is the focus beyond that of ET?

RQ1 was formulated this way because sufficient accuracy of the individual variables is assumed to be a prerequisite for accurate ET estimation. This will be motivated better in a revised version.

Line 55: RQ2 comes a bit out of the blue. It is the first time machine learning is mentioned. There is no story leading to it.

Thanks for the suggestion. In a revision will be add a paragraph on the use and potential of ML in ET estimation to better link to the RQ.

Figure 1: Not sure why Figure 1 is made of two parts given that they look quite different. Would it make more sense to seprate them? Also, I don't find the caption particularly informative and sufficiently correlated with the images.

We will reconsider the figure and caption of Fig 1.

Line 82: Add a space before "An overview..."

Will correct.

Line 88: "we use the following multivariate regression as a simple form of machine learning to estimate the evapotranspiration..." <- why calling this machine learning? It is a simple multivariate regression, and there is nothing wrong with it, per se. The authors should refrain from calling this machine learning and change the manuscript accordingly. On the other hand, what is the rationale behind the formula used? What are the parameters that need to be calibrated? There is no explanation.

We will use the term multivariate regression when discussing the method that we followed. However we do want to point out that future studies, which likely will have more data available for training, should use more sophisticated ML methods.

Line 89: The equation has not been numbered.

Will correct

Line 93: how much data is available in total?

Indeed this information was not mentioned in the manuscript. All measurements (18 variables on smartphone, fluxes, meteo observations) are available for 36 moments over 4 days (Fig 1B shows some variables over the whole observation period). This will be added to the manuscript.

Figure 2: the legend and captions are confusing, please amend.

Will change.

Line 113: "training Eq. 1" you do not train equation and that equation is not machine learning.

We will use “fit” rather than “train”

Line 117-120: I think the authors overstate the results they obtained. They are fitting their model to the two target variables, EC or lysimeter using values of said variables for parameters calibration (in the "training" dataset). On the other hand, lysimeter and EC are obtained independently.

It is not clear what point the referee aims to make here. We fit the model to EC and lysimeter data independently. These models are then evaluated both against the same method as well as the other (independent) one.

Line 130: I don't this is a great practice to add comments in parentheses? E.g., "it should be noted...".

Good suggestion, will change.

Line 130-133: This paragraph is convoluted. Please rephrase and consider splitting it.

Thanks for the suggestion. Will do.

Figure 3: The caption is not clear and the figure as well. Why not showing with different markers calibration vs validation data?

Good suggestion, will adapt.

Why the two vertical bars separating the two parts of the image?

Thanks for noticing. The bars are probably a result from the merger of the two panels from separate files. This will be fixed.

Discussion and Outlook: I find that this section lacks a proper discussion on the limitation of the proposed approach. How many precision lysimeters do we need to calibrate a world-wide network of phone-based algorithms? Is a linear method sufficient to generalize to location with no calibration data? Perhaps a non-linear machine learning model would be

more useful to improve model generalization by processing external data (i.e., rural/urban catchment characteristics, see [1] for instance.

We agree that the discussion would benefit from discussing these points in more depth. We did not intend to claim that the linear model should be seen as an end-point, but rather as the simplest starting point in a more generic ML-based approach. Future studies that have more data available for training should use more sophisticated ML methods, that allow for the use of categorical variables.

[1] Kratzert, Frederik, et al. "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets." Hydrology and Earth System Sciences 23.12 (2019): 5089-5110.