**Response to comments of Anonymous Referee #2 (R2)**

**R2:** *The authors present a catchment-scale hybrid model which is leveraged by sap flow data for more accurate hydrological simulations. The results showed that the hybrid model could lead to more realistic soil moisture estimates than the conventional Jarvis-Stewart equation, especially during drought conditions. The hybrid model predictions could match soil moisture and transpiration equally well as model runs using observed sap flow data and more importantly, hybrid model has good potential extrapolation beyond the study site. Such kind of hybrid model approaches which integrate machine learning methods and physical laws could open promising perspectives for more parsimonious process parametrizations.*

*These very interesting results have great potential to benefit the scientific community. With some minor clarification, this manuscript will be considered for publication.*

**Ralf Loritz (RL):** We thank Reviewer #2 (R2) very much for their positive assessment of our work and will address all their comments in a revised manuscript (MS).

Line by line comments:

**R2:** *I just have several specific questions. First of all, they didn't provide the cross-validation results. Secondly, did you also try the normal neural network, not the GRUs?*

**RL:** Good point. In a revised MS we will we present the cross validation results (see also answers to Reviewer #1).

We tried different combinations of artificial neural networks (ANN), gated recurrent networks (GRU) and long short-term memory networks (LSTM). Overall GRUs and LSTMs performed the best. As GRUs need less computational time and have slightly less weights, biases and no cell state, we used GRUs and not LSTMs. We will shortly explain this in a revised MS.

**R2** *(Line 129, 130): So, the 32 trees are evenly distributed in the catchment area? Could you show them on a map?*

**RL:** The sensors are not evenly distributed in the area. The field campaign was designed to capture **"***the typical hydro-pedological characteristics of the Colpach and the Weierbach.".* We will add a map to a revised MS that shows the locations of the sap flow sensors, the soil moisture sensors and the catchment boundaries of the Colpach and the Weierbach.

**R2** *(Line 196, 197): how many predictions time steps? Use 96 hours to predict next hour or next 2 hours? Why not 24h, 48h or 72h?*

***R2*** *(Line 197~198): How did you prove the network which consists of four layers (input, two hidden, output) with 128 cell/hidden state is the most appropriate structure? Will the different dropout rate affect the results significantly, e.g., 5%, 15%, 20%?*

**RL:** Both are hyperparameters and where identified by trial and error. We trained the model on the growing season 2014 and tested different model realization (hidden size, learning rate, sequence length, etc.) as well as different tpyes of ANNs and RNNs in the growing season 2016 (test data). Finally, we validated the model in the growing season 2015 (validation data set) without changing any hyperparameter. We have not tried all options systematically and most likely, you could identify a

model setup that outperforms our model given the current split sampling. This is not well explained in the current MS and we will update this section accordingly.

**R2** *(Line 260): So, you're using data from 2014 and 2016 to train the deep learning model, while use the data of 2015 as the test dataset? Did you try cross validation and set 2014 or 2016 as the test dataset to see the results? Are there significant differences between different catchments and years? Could you show the data distribution, e.g., boxplot, of different years and catchments?*

**RL:** Import point. We state in our current MS: "*The latter was the reason to choose 2015 as test period and not 2016, which would have kept the chronological order and led to overall lower errors without bias.*" We will add the root mean square error for scenarios in which 2014 or 2016 would have been the validation data set.

**R2** *(Section 2.2.4): It seems that the machine learning model is set to point to sap flow directly? Why not just let machine learning model predict the conductance directly? You could also introduce constrains into the loss function by using equation 1 and 2 to constrain the training process.*

**RL:** Comment to Reviewer #1: "*The performance differences between estimating canopy conductances ($g_c$sap) directly with the machine learning model or sap flow and afterwards calculating $g_c$sap are minor. Adding this intermediate step shows, however, that sap flow (an independent observation) can be predicted by a recurrent neural network (RNN) and opens the option to calculate transpiration directly in case catchment averaged plant specific parameters are available (Line 325 - 330). Furthermore, it opens the possibility to validate the model on independent sap flow measurements in case there are new or additional measurements available. We believe that this intermediate step gives the approach additional value and will explain this better in a revised MS as well as record the root mean square error when estimating $g_c$sap directly with an RNN."*

**R2** *(Section 2.2.4): I also suggest you should have a flow chart or schematic map for clearly clarifying the hybrid model. This could be more friendly to the readers.*

**RL:** We will consider adding a flow chart to a revised MS.

**R2** *(Line 293, 294): Could you further explain why the gcDL under- or overestimates on peaks? It seems that the model can't not capture the peak value very well? I think if you let the machine learning model predict the conductance directly with constrains from equation 1 and 2 into the loss function, this problem could be mitigated.*

**RL:** Predicting the canopy conductance directly only slightly improves the simulation results. But we agree that there is a potential to further improve our model results by a more systematic model choice or adding constrains to the model training. We will discuss this in a revised MS.