Answers to comments of Reviewer 1

№	Comment	Answer
1	The authors use just 15 combinations of model parameterization and forcing data to arrive at different conclusions regarding the importance of the two for modeling evaporation. In my opinion, this is severely inadequate for a robust assessment of uncertainty, let alone making any absolute conclusions about the importance of either model parameterization or forcing data, especially for a model which has greater than 20 parameters for modeling evaporation. A systematic uncertainty quantification would involve Monte Carlo simulations with a robust sampling scheme such as the Latin hypercube (by varying model parameters and meteorological inputs). As it stands, the results do not offer any conclusive quantitative evidence and as such is very superficial, and frankly not very useful.	Agreed partly. Currently to our knowledge, there are only three existing BROOK90 setups and all of them are used in the presented study: two automatic frameworks and the manual model itself. 'Automatic' means that the framework is collecting all data necessary to run the model, without expert knowledge. Yes, Monte-Carlo simulations will be advantageous to study the BROOK90 model uncertainty more deeply. This however was not the main intention of the study. We purposely used only real data for forcing and parameterization (best-per-scale) instead of statistical bootstrapping, as no one before did such an analysis for the BROOK90 frameworks. We wanted to emphasize the scale problematic, with the practical outcome. Mainly, in a presence of limited resources and data, do the global regional automatic frameworks deliver plausible E results and where the user should put more attention - accurate parameterization or meteorological input.
2	I have doubts about what the authors term as uncertainty in "model parameterization". From what I can gather, the only difference among the two models (BROOK90 and EXTRUSO) is land cover type and some input datasets. I do not think this is enough to quantify the uncertainty in model parameterization. The difference in the different models would then mainly arise from the difference parameter values of the calibrated and non-calibrated models. I do not understand how this difference can be construed as parameterization uncertainty. Either the authors should choose models which have completely different evaporation models (Penman vs Priestley-Taylor vs Hargreaves etc) or present a more robust quantification of the model parameter uncertainty (Monte Carlo simulations described above).	Agreed partly. We understand the model parameter uncertainty as follows: 'inability to specify exact values of model parameters' (Renard et al 2010 https://doi.org/10.1029/2009WR008328). The model possesses around 100 physically meaningful parameters, however, only around 30 of them are recommended to be changed according to the developer (other parameters refer to as fixed), namely the ones which describe vegetation and soil. Here we wanted to show the impact (uncertainty) of different BROOK90 parameterization schemes on accuracy of E simulations (automatically or manually derived for different scales from different datasets) – in general, not going into deep analysis of single parameters uncertainty. Additionally see comment №6. Incorporation of other models or methods to simulate evaporation will go far beyond the scope of the main topic. Nevertheless, as such an example, in discussion part 5.3 we show a comparison of complex vs simple (BROOK90 vs FAO) model setups.
3	In the same vein, the lack of uncertainty seen due to model forcings is just a function of the 3 datasets (in-situ, RaKliDa, and	Agreed aptly. See comment №1&6.

	ERA5). The present analysis does not provide sufficient evidence that forcing uncertainty is not as important parameterization uncertainty (Vrugt et al. 2008). doi:10.1029/2007WR006720	
4	The attempt to study the differences in the spatial scale of evaporation modeling is commendable. But the authors do not discuss the differences among the different models from the perspective of spatial scales sufficiently. It is quite obvious that a model calibrated with local data would perform better. However, the interesting thing is to understand the differences in the regional and global model. There is no discussion pertaining to this. I would think this is because of the inadequate sample space in which the study operates. I recommend that the authors perform a systematic quantitative assessment of uncertainty.	Agreed, discussion on the difference in model performance between scales (especially for the GBR90 and EXTRUSO frameworks) will be elaborated.
5	Many of the design choices are not explained and seem adhoc, The authors do not explain why a multi-objective optimizer was used here. Why attempt to create a Pareto-optimal solution for calibrating evaporation (growing period vs winter)? Why compare ERA5 hourly and ERA5 daily? Why only 3 input datasets? I can imagine that for Europe there are many observed forcing datasets (such as E-CAD). Why was the BROOK90 and EXTRUSO model chosen for this study? Why were only 20 parameters chosen? Was a sensitivity analysis conducted? Which are the most important parameters which contribute to the uncertainty?	Agreed, argumentation will be elaborated. Pareto-front calibration was used to address two issues. First, as most of E occurs in the vegetation period, it was decided to separate this period from the whole year as winter months should have lesser 'weight' during model fitting. Second we tried to account for possible systematic errors of E-C measurements themselves, which could be different in these two periods (i.e. Hollinger et al 2005 2021 https://doi.org/10.1093/treephys/25.7.873, Widmoser et al 2021https://doi.org/10.5194/hess-25-1151-2021, Twine et al 2000 https://doi.org/10.1016/S0168-1923(00)00123-4). Therefore the pareto front could help to choose an optimal parameter set (i.e. enhance winter month performance with insignificant loss of performance in vegetation period). The three used datasets represent 'state-of-the-art' meteorological datasets for global, regional and local scales for the study sites. RAKLIDA data is far better than E-CAD for regional scale as it was specifically designed and produced for Saxony and it has 1 km resolution (while E-CAD has 0.1-0.25 degree). Hourly ERA-5 data was applied only as an additional dataset, since it is implemented as primary forcing in the original GBR90 framework. However, for the comparability of three dataset's performance, ERA5 was upscaled to daily. Additionally we wanted to test and show the sensibility of the model to hourly vs daily data (see comment 11). Here we presume there is a small misunderstanding with the naming.

		BROOK90 is the model, which is the core of all setups. 'Extruso' (like 'Global BROOK90') is a framework (regional) which uses this model. For the calibration we initially took the physical 'location' parameters of the vegetation and flow parameters which are recommended by the developer and other researchers as the most sensible (Vilhar 2016 10.3832/ifor1630-008, Schwärzel et al 2009 https://doi.org/10.1016/j.foreco.2009.03.033, Habel et al 2021 https://forecomon2021.thuenen.de/fileadmin/forecomon/Presentations/132_Puhlmann_2.pdf, Groh et al 2013 http://dx.doi.org/10.5675/HyWa-2013,4-1). Then we conducted manual sensitivity analysis ('try-tests' with the given data) to come up with the chosen 20 parameters.
6	In summary, the study as it stands is very superficial and the authors have to make a strong case for why a qualitative assessment is sufficient to understand the uncertainty in model parameterization and forcings. In my opinion, the evidence provided in the manuscript points to the contrary: uncertainty assessments need far more robust experiment design to weed out spurious conclusions.	The main intention of the study was not to make a detailed assessment of the model's parameters and forcing uncertainty. Rather we want to address the topic mentioned in the main title. Namely, how available parameterization schemes and meteo input and their scales influence BROOK90 performance regarding evaporation simulations and existing model setups. Thus, we suggest to rephrase/omit 'uncertainty' term confusement or use it with caution, pointing out that we did not present a quantitative 'meteo and parameter uncertainty' evaluation and elaborate the last two paragraphs in the intro.
7	The abstract is very vague. What is the main conclusion of the study? What is the main implication of the conclusion?	Agreed, will be elaborated regarding main outcomes.
8	The manuscript needs to be edited to remove some idiosyncratic language use. For example Line 9: "Evaporation occurs on each surface…", Line 26: "...evaporation exposes larger variability…". Line 28: "...deepening knowledge…". Line 41: "The project allocates standardized …". Line 65: "the parameter set or meteorological input" should be "the parameter set and meteorological input".	Agreed, will be checked and corrected.
9	Line 40: I am not sure FLUXNET is an operational measurement network. I would term it as a database which collates measurements from different flux tower sites.	Agreed, will be corrected.
10	Line 215. Do you mean that the goodnes of fit should increase (rather than decrease) from global to local scales?	Agreed, will be corrected.
11	Why did the ERA5 daily outperform ERA5 hourly?	We suppose two main reasons. At first, due to the shortcomings in the interception modules of BROOK90. It runs on subdaily basis and if no subdaily P is passed in, it uses 'daily average rain

		duration in hours' parameter (which varies for each month) to disaggregate daily P into hourly. Furthermore, there are other simplifications (i.e. omitting diurnal cycle of potential evaporation). Federer (model developer) says that the module that uses subdaily P data consistently produces too much interception. Second one could be the poor quality of subdaily precipitation distribution in the ERA5 data for the study region. It was found that on daily, monthly and annual scales, ERA5 did not show a significant difference with the station data, which could account for that amount of differences in daily vs hourly KGE values. Additionally, it could be a case that simulations with hourly P are actually closer to reality, and eddy-covariance measurements themselves systematically underestimate interception. As we do not have enough evidence to check the plausibility of the abovementioned reasoning (five sites in one region, 10-30 years of data), thus we omitted discussion on this topic. However, we could add our suggestions as a discussion statement.
12	Line 355: This is a very absolutist claim. The partitioning of evaporation is a topic of major debate and the 60% estimate from Wei et al. 2017 is just one estimate. There is some uncertainty here varying from 55-85% depending on which study one considers.	Agreed, will be rephrased and elaborated.
13	Figure 7: It does not show which model result is shown in which pie chart.	Here we made an average from all model setups to derive general conclusions on the E partitioning for yearly and seasonal scale. Results for specific model setups are presented in Fig. 8.
14	The results section uses very subjective terms to describe model performance (example, 'fairly good' in Line 404).	Agreed, will be rephrased.
15	Line 449: I do not understand "...underestimation of the real site footprint or by permanent".	Agreed, will be corrected, the last part will be deleted.
16	Line 487: "...parameterization gave us higher spread". Where is this higher spread quantified? I recommend the authors attach some numbers to such claims, just a visual inspection is not enough.	The spread was described quantitatively in section 4.1. (lines 334-344). Sentences will be rephrased to add some numbers (%).