Referee:

I had a question on your previous manuscript version on line 225 (the question about the number and distribution of measurements) which I am still not fully understanding. I am not worried about the temporal or spatial comprehensiveness of your measurements, but it would be important to understand when the soil moisture and CH4 flux measurements that were used to train each model (May-July or Aug-Oct models) were taken and how the measurement date could influence your model performance. For example, ideally, you would do your soil moisture measurements across all the sampling locations on the same day, to make sure your measurements are representing the same conditions, and use this information to train your soil moisture model. Now I think your measurements are distributed quite randomly throughout the model period (either May-July or Aug-Oct), making them less comparable with each other. For example, you might have measured one location on a sunny day in early June and another one during a cold and rainy day in late June. Or, you might have measured one sampling location primarily in May, whereas you measured another location only in July. Even though the topographic position, soil texture, and vegetation conditions would be identical, and consequently soil moisture of these two locations should be the same if they would be measured at the same time, the location sampled only in July might suggest that the soil is drier than the other location because it tends to dry out during the summer. What I mean is that some part of the unexplained variability in your soil moisture/CH4 flux models might be related to the measurement time (date) and how warm and rainy it has been before it. And your current predictors do not consider this variability at all since they describe static topographic properties. I would either try to test how the measurement date or air temperature/precipitation conditions prior/during your measurement explain soil moisture or CH4 flux to check how much this could explain the performance issues in your models, or discuss this as a potential reason for the fact that a relatively large amount of variability in soil moisture/CH4 flux remained unexplained somewhere in discussion. You are kind of discussing this on line 620->, but I think you should also mention the comparability issues of your measurements somewhere, if I have understood the sampling scheme correctly.

Authors:

Yes, the referee is right. All the chamber locations were not measured during the same day, but (nearly) all the sample points were always measured during a 5-day-period, and these measurement rounds were repeated approximately every third week in May–October. On average, each sample point was measured every 22 days in May–October. This information was added on lines 148–151. Still, there were differences between the timing of measurements at different locations, which is somewhat inevitable when sampling many locations manually with limited workforce. We tried to overcome the effect of this issue on upscaling by doing two static predictions with the RF model, where the RF models were trained with temporally averaged data. However, despite the averaging, there may have been some leftover variability between the temporal means e.g. due to sampling some locations more during rainy days and others more during hot days. The apparent spatial variability caused by unsynchronised sampling is something that cannot be explained with the topographic properties as the referee points out. This apparent variability would decrease the performance of RF model in explaining the observed spatial variability in soil moisture/CH4 flux. We will add a note on this in the discussion section of the manuscript.

Minor points:

R: 29: “which was enough” sounds a bit strange to me → A: Removed.

R: 30: CH4 -> CH4 → A: corrected

R: 31: “upscaling predicted stronger CH4 uptake”-> do you have an idea why? Because you might have sampled more wetter environments than what their aerial extent truly was? Maybe this could be discussed briefly in the discussion. → A: Yes, our interpretation of these results is that we sampled more wetter environments than what their true areal extent was, and upscaling rectified this sampling bias.

R: 33-35: I still think these points come a bit out of the blue here in the abstract because they haven’t been mentioned before. Maybe you could try to link this sampling strategy comment more specifically (but shortly) to your results listed in the abstract, e.g. to the ones related to the differences in measured vs. predicted fluxes (see my comment above)? The main text mentions the sampling strategy in the introduction and discussion, which is great, but I’m just trying to assure that the reader understands the abstract as well as possible too. → A: We added a sentence about this analysis in the abstract, to the methods part.

R: 35: “..., and the measured fluxes...” seems a bit redundant to be. Doesn’t it always make sense to link fluxes to their environmental drivers? → A: We rewrote the sentence, so that it doesn’t sound like it’s not usually done.

R: 213: distribution was non-normal → A: corrected

R: 237: “The calculations were made with TopoToolbox (Schwanghart and Kuhn, 2010).” -> can be removed, since this is already mentioned on line 234 → A: We removed the reference, but probably good to keep the information that also these calculations were made with TopoToolbox, and not only the DEM pre-processing, which is previously mentioned.

R: 242: What was the spatial resolution of the other gridded data sets mentioned on lines 228? → A: They were 16 x 16 m resolution. This is now added to the text, lines 232–234.

R: 252: You could report here how many measurements and sampling locations you used in these two models. I assume that you had data from 60 sampling locations in all models, but it’s unclear to me how many observations you had in total (i.e. what was the sample size in your models). → A: Yes, we used all the 60 sample points both in May–July and Aug–Oct. We used the sample point averages for the modelling ((on average 7 and 5 measurements for each sample point location during May–July and August–October, respectively)), while the total number of measurements for both seasons were 392 and 320. We clarified this on lines 259–261, and also added the total numbers of measurements for both seasons to line 375.

R: 269: I’m not sure I understand the beginning of this sentence/whether it’s formulated correctly: “In this method, one RF model is developed for each sample point...”. To me it says that you developed a model for each individual sample point (i.e. model sample size n=1). Maybe mention that the validation data came from one sample point/you predicted the model developed with the training data to each individual sample point or something similar. → A: We agree, the wording was a bit misleading. What we mean is that one RF model was developed for each chamber measurement location, not for each observed data point. We will try to clarify this in the text.

R: 278: Perhaps mention here how you calculated uncertainty. Many studies also use a prediction interval of e.g. 2.5^{th}-97.5^{th} percentiles. Why did you choose to use standard deviations? → A: We added the notion that the uncertainty was estimated as standard deviation over the ensemble. If we assume gaussian distribution for the values predicted in the ensemble then the prediction interval (e.g. 2.5^{th}–97.5^{th} percentiles) can be directly calculated from the reported standard deviations. The use of standard deviations was just a practical choice.

R: 279: I think your uncertainty estimate also describes the effect of the distribution of the sampling points. → A: The referee might be right. We added this to the text.

R: 281: Aalto et al. (2018) → A: corrected

R: 283: Do you think it's problematic that your soil moisture predictions are not entirely independent from the topographic indices that you use as predictors of fluxes as well? Ideally, you'd use entirely independent data sets to train your CH4 flux models. Now your soil moisture predictions already have some information about the topographic indices too. → A: Yes, we agree, the variables used in CH4 models are to some degree related to each other. However, note that in the RF model training phase the measured soil moistures were used as predictors for CH4 fluxes, and only in the upscaling phase (i.e. when the RF model is used to predict CH4 fluxes in the study domain) predicted soil moistures (directly related to the topographic indices) were used. In general, the RF algorithm is robust towards redundant predictors when the aim is only to predict the response variable (soil moisture and CH4 flux in our case) as accurately as possible. Furthermore, this approach was suggested by a referee during the previous review round and we opted to follow hers/his suggestion.

R: 438: I'd mention somewhere here that your soil moisture and CH4 flux models seemed to have issues particularly with small and large fluxes, which were over-and underestimated, respectively. This might influence your average predicted fluxes too. E.g., if the area covered by dry conditions which are suitable for net CH4 uptake is high compared to the wet CH4 emitting patches, maybe your predictions are underestimating the net uptake and the difference between measured and predicted fluxes could actually be even greater? → A: Good suggestion, we added a note on this on lines 446–447.

R: 537: I would not repeatedly mention the acronym in the discussion SW-W-3 → A: Yes, SW-W-3 is mentioned several times in this section of the text, however the IDs for the chamber locations are given so that it can be shortly and clearly indicate which point is meant, and hence repetition is to some degree inevitable.

R: 538: Do you have the data about water table depth somewhere? → A: Unfortunately not.

R: 547: Or is it possible that SW-W-3 was measured after a rainy day and other on drier conditions? → A: No, this was not the case. That sample point had a water table level above the surface for most of the time.

R: 566-599: Could this discussion be its own section? Now that it's listed under upscaling, I start to think again about the limitations in using 15-20 sampling points to train a random forest model, even though I understand that this discussion is not related to that. → A: Yes, we followed this suggestion and added a new section '4.3 Representativeness of sample point locations'.

R: 582: You could add a few references here. Also, I guess quite frequently mean flux is calculated for each vegetation type, and then a spatially representative mean flux is calculated based on the spatial extent of the vegetation types? Or is that still rare in boreal forest CH4 flux studies? This method of course has issues related to the vegetation type map used, so I think your method is better! → A: Yes, we think that these type of upscaling methods are quite new, and this mean flux has been the traditional way.

R: 584: "spatially modelled CH4 flux"– are you referring to the modelled estimate for the whole area? → A: Yes, we clarified this accordingly.

R: 612: But how about vegetation indices derived from WorldView, Planet, or Sentinel-2 that have a pixel resolution of 2–10 meters? → A: Yes, these remote-sensing observations have good resolution but to our understanding it is not possible to extract the forest floor signal from these observations. We tried to clarify the sentence a bit more.

R: 613-619: A large part of this paragraph could actually be located in Section 4.1. which focuses on the drivers. → Yes, maybe so, this is partly about the drivers, but here they are discussed as they are linked to the modelling as well.