We thank the Referee Frans-Jan Parmentier for the helpful comments and suggestions. We addressed comments in the order discussed by the Referee and improved our manuscript accordingly. Our responses are in blue as shown in the supplement document.

Referee #1
I'm glad to see the revision by the authors, and that they incorporated all of my comments adequately. I'm also sorry that it took this long to deliver my review of this revision. I think that the paper has become a lot better, and I only have a few minor remaining comments.

Line 233: The exclusion of CA-BOU and RU-COK is based on the QC, but it's never really explained in detail on what exact basis they were excluded. Can you elaborate?

Responses:

We communicated with the data manager of the CA-BOU site to help with quality control. We also compared data anomalies in 2018 at the RU-COK site with the data before our study period. We modified line 233-234 to explain that the CA-BOU and RU-COK sites were excluded due to invalid seasonal measurements potentially due to device malfunction:

"Another 2 sites (CA-BOU, RU-COK) were excluded after quality control revealed an instrument anomaly that affected the measurements."

Line 679-680: I'm happy to see this additional analysis, and I know I asked for it, but it would be good if you could add some text for your readers on the reasoning for using maximum annual extent.

Responses:

Thanks for the helpful suggestion. We added an explanation on using the maximum annual extent of WAD2M in line 712 - 714:

"The addition of maximum annual wetland extent further constrains the limitations of seasonal WAD2M extents in underestimating methane emitting surface for northern high latitude wetlands, especially in cold seasons."

Line 735-737: This belongs in the discussion, not the results.

Responses:

Thanks for your suggestion. This point was elaborated in the Introduction paragraph 3 lines 95-97. We deleted this sentence because it does not belong to the result section.

Line 859: The site RU-VRK is not located in the West-Siberian lowlands, it's on the other side of the Ural Mountains.

Responses:

We improved the precision of the sentences in line xxx - xxx:

"For instance, Western Siberian Lowlands, the large wetland complex and the major contributor of interannual variations of CH4 in the region has little data. The nearest site (RU-VRK, not included in this study due to the observations before our study period) is situated on the western side of the Ural mountains, within the Usa River Depression."

Otherwise, I would just like to say that I'm happy to see the vast improvement to this paper, and I think that it'll be a nice addition to the literature.

Frans-Jan Parmentier

We thank the Referee #2 for the helpful comments and suggestions. We addressed comments in the order discussed by the Referee and improved our manuscript accordingly. Our responses are in blue as shown in the supplement document.

The authors put much effort into improving the manuscript and the modeling strategy. Thanks for clarifying the site-level and gridded modeling strategy both in the response letter and the manuscript, making it much easier to understand compared with the first version. However, many parts are still not clear or rigorous enough, making the current manuscript fail to achieve the standard of publishment. Detailed comments are as follows.
Major comments:
1. Comment 1: Please clarify in the manuscript that you chose the top 10 most important variables as input variables because the performance converged by R2. Please add model performance results and the converged R2 plot in the manuscript or supplement.
Responses:

We chose the top 10 most important variables based on the importance of impurity decrease in random forest modeling as shown in Fig. S3. We also added Fig. S4 and a sentence in the manuscript (line 529-530) as suggested to demonstrate the model performance converged as the increment of input variables by the importance rank.

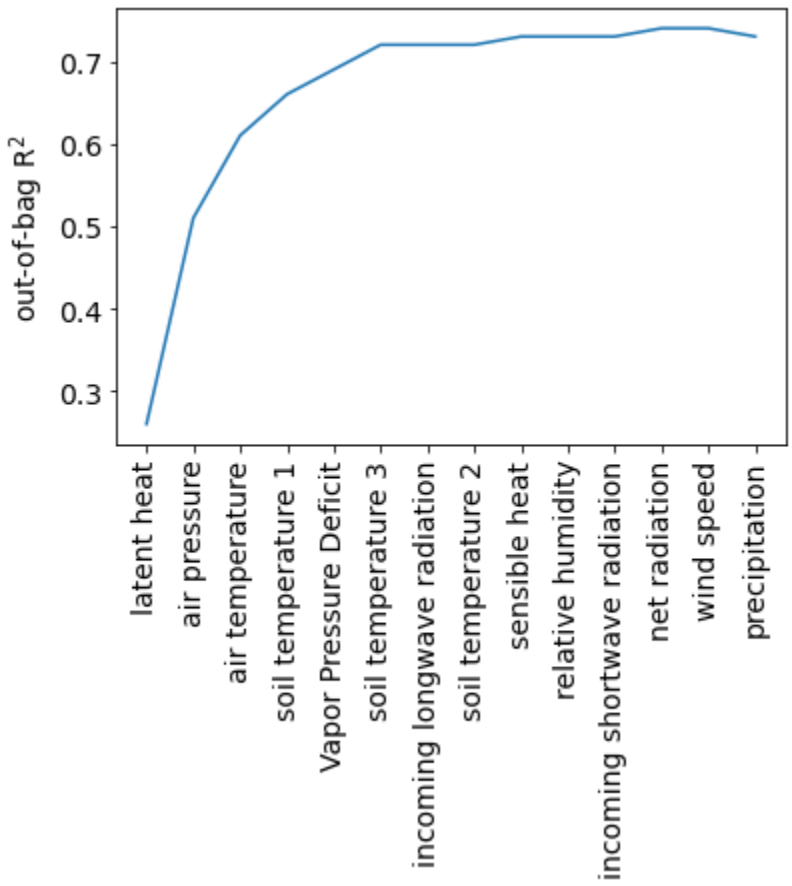

Fig. S4 Site-level model performance (out-of-bag R2) converged as the increment of predictor variables ordered by the importance rank as shown in Fig. S3. Tick labels on the

x-axis represent the addition of a variable to the precedent variables on the left in a recursive modeling.

2. For modeling strategy, you build site-level models and only use them for variable selection. All the upscaling work is based on the grid-level models. Is my understanding correct? (If I misunderstood, please correct it and clarify it in the manuscript.) Why not directly use the gridded model? You can directly establish gridded models, evaluate feature importance (with all candidate variables), and finally use the several most important variables that achieve the best performance to do the upscaling. I think it would be more reasonable and straightforward, and can help establish the best gridded model for upscaling.

Responses:

Yes, all the upscaling work is based on the grid-level models. We modified the objectives in the Introduction section (line 179 - 182) to elaborate this strategy. We agree that directly establishing the best-gridded model is straightforward from the upscaling perspective. However, the antecedent site-level modeling for physical predictor selection helped confirm the site-level controlling factors we learned from the literature and narrow down the gridded candidates. We also compared how the differences in scales and measuring methods between in situ predictors and gridded proxies affect model-learned temporal variability in CH_4 fluxes (line 524-528), which could evaluate the impacts of input spatial resolutions on the model performance.

3. I am still curious whether the site-level model can guide the gridded model in feature selection. Actually, site-level and gridded upscaling models are totally different, with different input features (different data sources and different number of features). Comparing Fig S3 and Fig 3b, feature importance differs significantly in site-level and gridded models. For example, TS_2 showed 1st place in the gridded model, but 8th in the site model. In this situation, can the feature importance by site model still effectively and reasonably guide feature selection for gridded models?

Responses:

The site-level model, site synthesis literature (Knox et al., 2021; Delwiche et al., 2021), and expert knowledge guided our predictor variable selection for gridded models. We used the gridded version of those in situ measured physical variables selected at the site-level modeling. Because the data sources were different, we separated the site-level modeling and grid-level modeling to ensure the data sources within a model were comparable. In this way, the feature importance of gridded models can reflect the relative importance of each gridded variable in the gridded-level and upscaling models. This also applied to the site-level modeling. However, the feature importance also pertains to the input data distribution and random forest model structure. Additional gridded variables from remote sensing products were added to complement the missing controllers from the site-level modeling, as you mentioned “different number of features”. Therefore, the feature importance by site model can help us identify controlling physical variables but would not necessarily translate to the same rank in the feature importance of grid models. We added sentences in line 389-396 to address this comment.

4. What is the standard for choosing candidate variables in the site model? According to the statements in the main text L268-269, variables that affect CH4 at multi-day to seasonal scales (by previous studies) should be considered. Many studies have revealed considerable impacts of vegetation. For example, in Knox et al. 2021 (you also cited this paper in your manuscript), GPP and RECO significantly controlled CH4 at seasonal scales. Therefore, excluding vegetation proxy in candidate variables for feature selection seems unreasonable. In your response letter, you mentioned excluding vegetation-related variables because neither Peltola nor McNicol used EVI for upscaling, and it seems inconsistent and contradictory with the candidate selection rule you wrote in the main text in L268-269.

Responses:

One standard for choosing candidate variables at site-level modeling is that they are in-situ measured physical variables and are available at most sites. We modify lines 268-269 to stress this standard in the main text. We also acknowledge the significant roles that GPP and RECO play in controlling seasonal CH_4 changes. GPP and RECO are estimated from measured NEE. The partitioning in GPP and RECO is associated with assumptions that yield unknown amounts of uncertainty. Therefore, we did not include GPP and RECO at site-level modeling but included vegetation-related variables in our grid-level modeling and upscaling. We used MODIS NBAR as proxies for vegetation productivity because EVI and GPP can be derived from MODIS NBAR bands, instead of directly using MODIS-based EVI. We modify line 342-344 in Data section 2.2.2 to clarify this point.

5. L386-389: It is unclear how to select variables for the 'baseline' model. 1) What is the standard for choosing the variables? Did you choose the variables with more 'red' grids and less 'white' grids? 2) In Fig S14, you set a threshold of 0.8. Why use that value? 3) Why choosing ts_2 instead of ts_1 or ts_3? What is the difference? 4) Why can NBAR be excluded from the feature selection based on the pairwise Pearson correlation test?

Responses:

We first group significantly correlated variables ($p<0.001$, $r>0.8$, white grids except for those on the diagonal line) in the pairwise Pearson correlation test, forming three groups: SMAP soil moisture variables in group 1 (we also include surface soil moisture that is significantly correlated with the other two soil moisture variables and $r>0.7$), air temperature (tas), downward longwave radiation (rsdl), spfh, soil temperatures (ts1, ts2, and ts3) in group 2, downward shortwave radiation (rsds) and latent heat (le) in group 3. We then select one most important variable in each group according to Fig. S15 for the baseline models. The rest variables out of the groups (air pressure (pa), sensible heat (h), slope, spi, and cti) are included in the baseline models. We added these sentences (line 402-410) to the paragraph about constructing five model settings in section 2.3.1.

6. Figure 3c. Which model did you finally choose to generate the upscaling product? The models with 'all' variables? If so, I think it would be more reasonable to show the feature importance of 'ALL' variables instead of 'baseline' variables. Although collinearity could affect the feature importance, the feature importance of the model with all the variables you finally used for upscaling still should be presented and explained.

Responses:

Yes, we chose the models with 'all' variables as they provided the highest mean R^2 and lowest median errors (Figure 3a) for all validation sites under the LOOCV scheme. We add the feature importance of 'ALL' variables in the supplementary (Figure S15):

"Fig. S15 Variable importance of "ALL" models with all gridded input variables under the LOOCV scheme: mean (last row) and at each validation site (row denoted by site ID). The 'ALL' modeling setting is used to build the upscaling ensemble models."

We explain the feature importance of 'ALL' variables and the agreement with those of the baseline variables (Figure 3b) in line 571-574:

"The average importance of 'all' gridded variables used for upscaling (Fig. S15) was consistent with baseline models, emphasizing the importance of soil temperatures and rootzone wetness. Additionally, air pressure and topography also contributed to explaining the daily variability in CH_4 fluxes."

Minor comments:

1. Fig. S3. I suggest adding the full name of the variables in the caption to make it more readable.
Responses: We updated Fig. S3 to label the full names of variables.

2. Fig. S3 and Fig 1: Are the top 10 variables in Fig S3 finally selected? If so, SW_IN should not be included (it is the 11th). However, in Fig 1, shortwave radiation is included in the models.
Responses: We used MERRA2 surface specific humidity as proxies for vapor pressure deficit (VPD) and relative humidity (RH). Therefore, we selected 10 MERRA2 variables as the gridded version of the top 11 variables from site-level modeling.

3. L196-198 'based on literature and expert knowledge': Do you mean you chose these variables based on previous studies? If so, please add citations here.
Responses: We add citations for the literature references.

4. L202-203: you mentioned the ML with the highest R2 and lowest ME was chosen. But from Fig 3a, it is hard to tell 'Base+CoVar' or 'All' which is better, they look almost the same in the figure. Which one did you finally use for upscaling? Please clarify it in the manuscript.
Responses:

The mean/median R2 of 'All' (0.506/0.55) surpassed those of 'Base+CoVar' (0.505/0.53). The median errors of 'All' (MAE=14.1, RMSE=19.8, ME=4.0, unit: nmol m^{-2} s^{-1}) were lower than the 'Base+CoVar' (MAE=16.0, RMSE=23.6, ME=6.2, unit: nmol m^{-2} s^{-1}), while the mean errors of these two modeling settings were comparable. Therefore, we used the 'All' model setting for upscaling. We modified line 202 - 205 in the manuscript to clarify this and added the R2 and error values in line 558-561.

5. Fig S6 is missing in the supplementary.
Responses:

Thanks for catching this. We added figures in the supplementary and relabeled the figure numbers. We updated the associated figure numbers in the text of our manuscript.

6. Fig 6. Although you clarified that you used the ensemble mean of bottom-up models in GCP in the manuscript, I suggest using the 'GCP Bottom-up ensemble mean' or 'GCP BU ensemble mean' as the title of the subplot. The 'GCP ensemble mean' can be easily misunderstood as the ensemble mean of all GCP methane models, including both bottom-up and top-down models. The same issue for other figures.
Responses:

We updated Fig.6 with your suggested subplot title for 'GCP Bottom-up ensemble mean'. We also updated the captions regarding the bottom-up GCP ensemble mean in other figures.

7. Figure 7d. Why excluding wetCH4_giems and wetCH4_GLWD?
Responses:

We estimated seasonal emissions with $WetCH_4$_giems and $WetCH_4$_GLWD in the North Slope region. GIEMS2 significantly underestimated wetland areas in this region in the summer and estimated zero wetland areas in cold seasons. Fig.7d $WetCH_4$ land (black line) assumed all land in this region as methane emitting surface to represent the potential maximum seasonal CH_4 emissions that $WetCH_4$ could estimate. $WetCH_4$ GLWD emissions were between $WetCH_4$ CALU and $WetCH_4$ land. We selected emission estimations with WAD2M and all land that matched the minimum and maximum range of CARVE estimates.

8. If you used microbial emissions instead of wetland emissions for CarbonTracker, would that still be comparable with WetCh4, GCB ensemble, and WetCHARTs? You also mentioned in L662-663 that non-wetland emissions also have considerable contributions, but the current results by Carbon Tracker cannot distinguish them.
Responses:

The contributions of ruminant animals and winter open water to the microbial emissions estimated by CarbonTracker-CH_4 were negligible in the northern high latitudes (60° - 90° N). When we compared $WetCH_4$, GCP bottom-up ensemble mean, and WetCHARTs in the northern high latitudes, we found an agreement in the phase of the seasonal pattern (Fig. 7b). The differences in the magnitudes in the summer may be attributed to aquatic emissions (Johnson et al., 2022).