We thank the Referee #2 for the helpful comments and suggestions. They will help improve the manuscript. We will address comments in the order discussed by the Referee. Our responses are in blue.

Wetlands are the largest natural source of global methane (CH_4) emissions, but with the largest uncertainty. Ying et al. generated a machine learning based regional (>45) wetland CH_4 upscaling dataset. In general, their work is very important, and they provided a new data-driven benchmark dataset constraint by the most eddy covariance observations, with the highest spatial and temporal resolution, compared with previous ML-based wetland CH_4 upscaling products. However, there remain many parts that are not clear or rigorous enough. Detailed comments can be seen as follows:

Major comments:

1. For the feature selection part, why did you choose the first 10 variables? Did you test other numbers of input features? In section 3.1.1, you mentioned that using all the variables and using selected 10 variables showed no significant difference in wetland CH_4. Is this strategy still reasonable? Maybe the strategy in Peltola et al., 2019 could be helpful. They calculated the feature importance of all the variables, but finally chose four variables, because that group achieved the best performance. (*Peltola, O., Vesala, T., Gao, Y., Räty, O., Alekseychik, P., Aurela, M., ... & Aalto, T. (2019). Monthly gridded data product of northern wetland methane emissions based on upscaling eddy covariance observations. Earth System Science Data, 11(3), 1263-1289.)*

   Responses: Yes, all candidate variables were tested together within the random forest model and we ranked the feature importance of all candidate variables as shown in Fig.S3. The model performance converged with the top 10 most important variables that were selected as input variables, as indicated by the out-of-bag R^2 metric and Fig. S3.

   Our modeling framework differed from Peltola et al., 2019 by separating in situ variables and gridded variables and modeling at the site level and grid level respectively. Previous studies trained and validated models with a mix of in situ and gridded variables but performed upscaling with only gridded variables, causing the evaluation metrics not to show the accuracy of upscaling products. In addition, this modeling strategy tended to favor gridded proxies or variables over in situ variables, leading to the real controlling variables of CH_4 fluxes not being selected.

   We improved the strategy by first modeling at the site level with only in situ variables that were available at all wetland EC sites. We then used the gridded version of the selected in situ variables. For some missing controlling factors that were not measured across all sites, we further added remote-sensing-based variables or proxies (e.g., SMAP soil wetness, MODIS reflectance) in a forward selection process and demonstrated the improvement in model predictive performance as shown in

Fig. 3a. As a result, the grid-level modeling and evaluation is consistent with the upscaling and reflecting the accuracy of the upscaling product.

We clarified this on lines 192-200 and 369-375 in the clean version of the revised manuscript.

2. The workflow seems a little bit confusing to me. Please feel free to correct me if I misunderstood. It seems that the feature selection only included the variables you get from the MERA2 dataset. Why are the variables from remote sensing dataset excluded in the feature selection step, but directly added into the final RF model? Is that fair to all the variables?

   Responses: The feature selection was performed at the site level with in situ measurements. At the grid level, we further performed a forward feature selection by adding remote-sensing datasets to the MERRA2 data. By evaluating the impacts of adding constraints from remote-sensing data on the grid-level model performance as shown in Fig.3a), we prove that adding remote-sensing variables can improve the model's ability to explain the average variability in daily CH_4 fluxes across validation sites and reduce prediction errors.

   To better explain the framework, we edited sentences on lines 192-200 and 285-395.

3. The final produced dataset is 0.098*0.098degree, but the spatial resolution of input datasets (e.g., MERRA2) is much lower. Similarly, the wetland extent dataset (WAD2M, GIEMS) also has lower spatial and temporal resolution. Will that lead to uncertainties in your final estimation? At least, some discussion of this issue should be added to the manuscript.

   Responses: Thanks for this helpful suggestion. We interpolated MERRA2 data to 0.098° x 0.098° weighted by MERIT-DEM. We then modeled and upscaled at this spatial resolution. As a result, the model accuracy metrics reflected a portion of the errors due to the scale difference between MERRA2 input and in situ measurements.

   Per your suggestion, we discussed the need for accurate and dynamic wetland maps at high spatial resolution to improve wetland CH_4 estimations in the study area. We suggested incorporating wetland fractions derived from high-resolution thematic maps (e.g. CALU) to improve the use of WAD2M in cold regions.

4. L362-364: I think it is not surprising to see that groups (2) and (3) have lower accuracy, because they only contain features from soil wetness or NBAR, but missed the most important information from the features provided by MERRA2 (which you revealed in the feature selection part). Thus, if my understanding is correct, would it be more reasonable to set the input feature as MERRA2, MERRA2 + NBAR, MERRA2+SMAP, and compare them to MERRA2+ all RS data?

   Responses: Thanks for the suggestion. We redesigned the model input variable settings to accommodate this suggestion and your comment #6. We updated Fig. 3. Please see our responses to comment #6 for a complete description.

5. Did you test uncertainties from MERRA2? Will the estimation and key findings be the same if using different reanalysis datasets?

   Responses: We improved the discussion on the uncertainties in MERRA2 and the potential impacts on emission estimates on lines 791-797:

   "However, lower correlations and overestimated monthly variability were found in the cold season in Pan-Arctic (Herrington et al., 2022). This suggests the impact of the uncertainty in MERRA2 soil temperatures were concentrated in the cold season, when CH_4 fluxes were low. The agreement between ensemble means of soil temperatures from eight reanalysis and land data assimilation system products and station measurements improved in the pan-Arctic region (Herrington et al., 2022), suggesting the potential to reduce upscaling uncertainty forced by the ensemble mean of reanalysis datasets."

6. Figure 6: I am curious why DEM is the most important feature. You mentioned it highly correlated with air pressure, but the importance of air pressure is very low. Please share more explanation of the mechanisms of how DEM affects wetland CH_4.

   Responses: Thank you for this helpful comment. We found that the variable importance ranked by the impurity decreases in RF models affected the interpretation of real controlling variables when covariates existed. The collinearity among input variables (such as temperatures at different depths, DEM and air pressure, air temperature) allows some of the removed variable's information to be retained, potentially distorting its true importance. This highlights the need for careful interpretation of correlated features' importance and is the reason why DEM appeared so important in the previous variable importance analysis. To address this,

   1, We updated input variables by using interpolated MERRA2 variables (at ~10 km spatial resolution) weighted by DEM for modeling and removing DEM from the input predictors. We clarified this on lines 291-296. Accordingly, we updated Table 1 to reflect this modeling spatial resolution change in MERRA2 data.

   2, We improved our design of input feature settings at the grid-level modeling. We first built a baseline grid-level model with independent variables after a pairwise Pearson correlation test (Fig. S14) to exclude covariates. The resulting baseline features included air pressure (pa), latent heat flux (le), sensible heat flux (h), soil temperature (ts2), rootzone soil wetness (sm_r_wetness), slope, spi, and cti. Then we designed four additional different model settings by changing predictor variables, including (1) baseline variables plus covariates, (2) only variables from MODIS NBAR, (3) baseline variables plus NBAR bands, and (4) all predictor variables. In this forward feature selection process, we evaluated the impacts of adding constraint variables from remote sensing products on model performance. RF models can enhance robustness when handling correlated input variables, so these collinear variables shouldn't negatively

affect model performance, only the variable importance assessment. This is due to the RF algorithm randomly selecting subsets of input variables and choosing the best one for splitting nodes during tree constructions. We modified the description on lines 385-295. We demonstrated error reduction and model improvement as new variables, including physically independent MODIS NBAR observations, were added as shown in Fig. 3a.

3, We updated feature importances from the baseline model with non-covariates. We merged Fig. 3 and Fig. 6, and showed the importance of baseline features in Fig. 3b. The new result demonstrated the importance of soil temperature and moisture, as described in the revised manuscript on lines 539-546.

4, We updated the predictive performance metrics of the upscaling model (lines 526-537, 561-592) and ultimately the upscaling results from non-DEM ensemble models (the results section 3.2 Upscaled wetland CH_4 emissions). The new model improved performance at wet tundra sites but enlarged errors at a few fen and bog sites. Overall, it slightly overestimated CH_4 fluxes at the validation sites as shown by positive bias (mean ME). The upscaling results from the new model manifest slightly higher flux intensities in wet tundra and in the summer season (JJA), resulting in increases in the estimates of mean annual emissions by ~2 Tg CH_4 yr^{-1} with WAD2M to ~5 Tg CH_4 yr^{-1} with GLWD v1 and v2. No significant change in the absolute and relative interannual variability in subregions.

5, We added a discussion in the supporting materials Text 6 about the impact of elevation on explaining the intra-site variability within the existing wetland sites of northern high latitudes. We tested the impacts of elevation on model performance in explaining the inter-site variability of CH_4 upon the current locations of wetland EC sites. We recognized that elevation may act as a factor in discerning fen and bog sites with associated wetland attributes that may not be included by other input variables.

Minor comments:

1. L183: The boundary of Arctic-boreal is not exactly the same as '>45 degree'. If your final dataset is >45 area, you cannot say it is Arctic-boreal region. Similar problems appeared several times in the manuscript. Please go through the whole paper and correct them.

   Responses: Thanks for this suggestion. We have rephrased the study area to wetland >45° N.

2. Are the important features the same at different sites? Or are they the same across different wetland types? Did you build separate models for different wetland types? Or use one model for all types?

Responses: We built one model for all types of wetlands as the gridded wetland fraction information by wetland types was not available in the whole study area that was required for wetland fluxes upscaling.

3. Vegetation activity showed significant impacts on wetland CH_4 emissions in many previous studies, especially in the northern wetlands. Why not include proxies of vegetation (such as, LAI, GPP, …) into your feature selection?

   Responses: Previous studies (Peltola et al., 2019; McNicol et al., 2023) evaluated MODIS-derived EVI as a proxy for GPP as a candidate predictor. However, none of them selected EVI (concurrent or lagged) in their upscaling models because the inclusion of EVI did not improve the model performance as much as the temperature-related variables did. Therefore, we did not directly include proxies of vegetation productivity in the feature selection, instead, we included constraints from MODIS reflectance bands that were used to produce GPP, EVI, or LAI as well as surface water indices. We explained this on lines 335-337.

4. Figure 4: Why is monthly prediction much better than that of daily prediction, especially in terms of R_2? Please add more explanation to the manuscript.

   Responses: Model predictive performance on aggregated monthly means of CH_4 fluxes increased by 37% as compared to daily means (R^2 = 0.70, Fig.4, Table S4). Model agreement worsened at daily and weekly timesteps due to higher variability in CH_4 fluxes at finer temporal resolutions (lines 768-769). The amount of noise in the flux data is much higher at a daily resolution while mean monthly fluxes smooth some of this noise away.

5. Figure 7: For carbon-tracker, why did you use natural microbial emissions instead of wetland emissions? It seems that carbon-tracker also has an output layer of wetland CH_4.

   Responses: Natural microbial emissions primarily comprise wetland and aquatic emissions in northern high latitudes. According to Oh et al., 2023, the aquatic CH_4 sources were not discerned from wetland emissions in the current release of CarbonTracker-CH_4. Therefore, we used natural microbial emissions.

   Reference: Youmi Oh, Lori Bruhwiler, Xin Lan, Sourish Basu, Kenneth Schuldt, Kirk Thoning, Sylvia E. Michel, Reid Clark, John B. Miller, Arlyn Andrews, Owen Sherwood, Giuseppe Etiope, Monica Crippa, Licheng Liu, Qianlai Zhuang, James Randerson, Guido van der Werf, Tuula Aalto, Stefano Amendola, Sébastien C. Andra, Marcos Andrade, Nhat A. Nguyen, Shuji Aoki, Francesco Apadula, Ikhsan B. Arifin, Sabrina Arnold, Mikhail Arshinov, Bianca Baier, Peter Bergamaschi, Tobias Biermann, Sebastien C. Biraud, Pierre-Eric Blanc, Gordon Brailsford, Huilin Chen, Aurelie Colomb, Cedric Couret, Paolo Cristofanelli, Emilio Cuevas, Lukasz Chmura, Marc Delmotte, Lukas Emmenegger, Gulzhan Esenzhanova, Ryo Fujita, Luciana Gatti, Elise-Andree Guerette, László Haszpra, Michal Heliasz, Ove Hermansen, Jutta Holst, Tatiana Di Iorio, Armin Jordan, Müller-Williams Jennifer, Anna Karion, Teruo

Kawasaki, Victor Kazan, Petri Keronen, Seung-Yeon Kim, Tobias Kneuer, Katerina Kominkova, Elena Kozlova, Paul Krummel, Dagmar Kubistin, Casper Labuschagne, Ray Langenfelds, Olivier Laurent, Tuomas Laurila, Haeyoung Lee, Irene Lehner, Markus Leuenberger, Matthias Lindauer, Morgan Lopez, Reza Mahdi, Ivan Mammarella, Giovanni Manca, Michal V. Marek, Martine D. Mazière, Kathryn McKain, Frank Meinhardt, Charles E. Miller, Meelis Mölder, John Moncrieff, Heiko Moosen, Caisa Moreno, Shinji Morimoto, Catherine L. Myhre, Alberth C. Nahas, Jaroslaw Necki, Sylvia Nichol, Simon ODoherty, Nina Paramonova, Salvatore Piacentino, Jean M. Pichon, Christian Plass-Dülmer, Michel Ramonet, Ludwig Ries, Alcide G. di Sarra, Motoki Sasakawa, Daniel Say, Hinrich Schaefer, Bert Scheeren, Martina Schmidt, Marcus Schumacher, Mahesh K. Sha, Paul Shepson, Dan Smale, Paul D. Smith, Martin Steinbacher, Colm Sweeney, Shinya Takatsuji, Gaston Torres, Kjetil Tørseth, Pamela Trisolino, Jocelyn Turnbull, Karin Uhse, Taku Umezawa, Alex Vermeulen, Isaac Vimont, Gabriela Vitkova, Hsiang-Jui (Ray) Wang, Doug Worthy, Irène Xueref-Remy. CarbonTracker CH4 2023, 2023.DOI: 10.25925/40jt-qd67

6. What GCP models did you include in comparison? All the top-down and bottom-up models in Saunois et al., 2019? It would be better to give more information of what model did you used in the supplementary. Or at least, add the citation of GCP models.

   Responses: Thanks for the helpful suggestion. We added information about the GCP models in the supporting materials:

   The bottom-up estimates we used for comparison were from sixteen wetland CH_4 models ($CH4MOD_{wetland}$, CLASSIC, DLEM, ELM-ECA, ISAM, JSBACH, JULES, LPJ-MPI, LPJ-wsl, LPJ-GUESS, LPX-Bern, ORCHIDEE, SDGVM, TEM-MDM, VISIT, TRIPLEX-GHG) in the Global Carbon Project (GCP) Methane Budget (Z. Zhang et al., 2024).

7. Figure 10: Why exclude $WetCH_4$-GIEMS?

   Responses: We updated the figure by adding interannual variability estimated in $WetCH_4$-GIEMS.

8. Figure 8d: Please give more description of land and CALU data, and explain how you generate wetCH4-land and wetCH4-CALU, and why did you use them.

   Responses: We modified the manuscript as below:

   "Given that the wetland area in this region is uncertain (Miller et al., 2016), we computed mean seasonal cycles over the land assuming all land in this area is water saturated in the soil, over freshwater wetlands of CALU, and over WAD2M and Hydrolakes, representing three different scenarios. In the lowland area of the North Slope (74295 km^2 spanning between 69.8°N - 71.4°N, 164.4°W - 152.7°W), the wetland area was estimated at 10611 km^2 from CALU, 4800 km^2 from GLWDv2, and 4049 km^2 from the maximum extent month in July of WAD2Mv2, respectively."

9. Add citations: L78-80, L99-104.

   Responses: Thanks for your suggestion. We added citations:

   "The uncertainties in the estimates of wetland CH_4 emissions are primarily attributed to challenges in mapping vegetated wetlands versus open water leading to double counting (Thornton et al., 2016), seasonal wetland dynamics and uncertainties in estimates on flux rates."

   "Field observations of gas fluxes typically measure CH_4 exchange between the land and atmosphere at sub-meter to ecosystem (100s of m to km) scales (Bansal et al., 2023; Chu et al., 2021)."

10. L921-928: Font style.

    Responses:

    Thanks for pointing this out. We edited the Font in this paragraph.