Biogeosciences Discussions

Biogeosciences Discuss., referee comment RC2
https://doi.org/10.5194/bg-2021-256-RC2, 2021

Comment on bg-2021-256

Anonymous Referee #2

Referee comment on "Gaps in network infrastructure limit our understanding of biogenic methane emissions for the United States" by Sparkle L. Malone et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2021-256-RC2, 2021

This study by Malone et. al. presents an representativeness analysis of current Eddy Covariance (EC) CH4 observation sites to understand gaps in data collection and develop guidance for new research infrastructure to reduce the uncertainties in US CH4 budget. It applies a multidimensional cluster analysis based approach to develop and assess CH4 observing network for US.

Study outlines well the limitations and uncertainties associates with current EC CH4 network in US and takes on an important research to better understand its representation. However, there are number of limitations in the data and methodoloy applied in the study, presented results does not support the conclusions and overall it does not provide mcuh improvements over limitations in current literature that manuscript describes in introduction.

Comments on methods:
Method section is brief, lack necessary description and is at time diffcult to understand and follow.

- While cluster analysis method in itself is sound, its unclear if and how the choice of land cover and climate layers chosen to represent primary environmental conditions represent ecosystem scale CH4 fluxes. And if and how they represent the conditions for aquatic sites?
- Authors highlight on Line 75-79 the limitations of existing land use products to identify CH4 source/sink, wetland classifications not suitable to scale and predict CH4 flux rates and processes. BUT they choose to use existing NLCD reclassified to 8 classes and reduced 1km resolution, thus effectively reducing the fidelity of the landcover data. Did do however improve the representation of irrigated vs non-irrigated crop ecotypes, and wetland classes. However, no quantitative analysis has been presented to demonstrate improvements their landcover scheme provide over the existing landcover.

- Line 198-200 states "The land cover and climate layers were chosen to represent the primary environmental conditions that are often indicative of a combination of resource availability and disturbance regimes." What resource availability refers to in this particular context? Also, which layer, if any, capture the disturbance regimes and what type of disturbances?
- Section 2.3 Defining the state space of the US
    - Multidimensional scaling (MDS) was performed to condense ecotype, climate and location variables to two dimension. But unlike climate and location, ecotype is a categorical variable and how was a categorical value handled in the MDS scheme. Why were they condensed down to two dimension, aside from R/MASS implementation's default?
    - However, MDS was conducted only on a subset of 20,000 1km pixels subsampled across the US. What was the purpose and motivation for subsampling? And why 20,000? That sounds like a very small fraction of points across US at 1km resolution, essentially restricting the entire analysis to a very small small and perhaps biased sample of data (even if the 20,000 samples were evenly distributed across ecotypes). This is a severe limitation of the presented study.
    - Lines 218-220 states "We measured the correlation between the ecotypes, climate layers, and locations (latitude/longitude) using the envfit function in the library vegan in R (Oksanen, 2016)." What was the purpose of these correlations? And again how do you calculated correlation against categorical ecotype variable? How meaningful is it to calculated the correlation of climate or ecotype to location lat/lon? These choice of methods need some clarification beyond reference to R packages.
    - Lines 219-223 "This was followed by a cluster analysis to determine the optimal number of clusters using the library cluster in R, which partitions data around medoids (PAM algorithm), using the Gower dissimilarity matrix (Gower, 1971; Huang, 1997; Podani, 1999; Ahmad and Dey, 2007; Harikumar and Pv, 2015)." Its unclear if clustering was done on orignal set of ecotype, climate and location variables or two MDS dimensions? If original variables, I would repeat the need to clarify how the categorical variable was handled? I believe Gower dissimilarity matrix would consider the pairwise ecotype similarity so forest vs irrigated crops will have the same dissimilarity as irrigated vs non-irrigated crops. Is that a correction interpretation, and if so was that intended. Does that approach really help effectively use the ecotypes in this analysis, I believe not. More discussion of methods and their implication on results are needed.
    - Lines 222-224 states "We fit an increasing number of clusters from 2 to 20 to construct a silhouette plot and choose the number of clusters that maximizes the average silhouette width to determine an optimal number of clusters." However, the discussion section jumps to discuss k=10 clusters, but what about 2-9, 11-20? Why 10, why not some other number between 2-20?
    - Lines 236-238 "To extrapolate the cluster and distance layers across the entire US beyond the 20,000-pixel subsample, we fit a Random Forest model with the package randomForest (Liaw and Wiener, 2002) to model the first and second MDS dimension using the ecotype and climate layers as predictors." I am struggling to understand what this statement means. A Random Model model is being fit to model first and second MDS dimension using ecotype and climate layers as predictors. BUT weren't MDS dimensions calculated using ecotype and climate layers in the first place. Is this Random Forest model a replacement for R MASS package. Why, this step makes no sense, unless I am missing something or this statement is somehow mis-written.
    - Lines 239-240 "We then created a Random Forest model of the cluster layer using the first and second dimension as the explanatory variables." Why would you not just cluster the first and second MDs dimension, instead of creating a Random Forest model to identify the cluster layer. This seems like unnecessarily convoluted approach which really makes no sense and is addding methodological complexiy and

model uncertainties.

Comments on results and discussion:
Results and discussion section is more about stating the results and is really lacking in discussion of results, why they were calculated and what they mean for the science question central to the study?

- Lines 260-263 "Latitude (R^2 = 0.95; $p < 0.001$), mean annual temperature (R^2 = 0.84; $p < 0.001$), maximum temperature (R^2 = 0.83; $p < 0.001$), vapor pressure deficit (R^2 = 0.83; $p < 0.001$), minimum temperature (R^2 = 0.82; $p < 0.001$), longitude (R^2 = 0.63; $p < 0.001$) had strong effects on clustering, whereas precipitation (R^2 = 0.10; $p < 0.001$), and ecotype (R^2 = 0.03; $p < 0.001$) showed low correlations."
    - First, I am still not clear what these correlations are?
    - Simply the fact that ecotype and precipitation has very low correlations for clustering is the huge red flag that these clusters are not appropriate for quantifying the representation of CH4 measurement sites. In the introduction section, authors have made strong arguments about importance of landcover, wetlands and agricultural use classifications. By their own measure, clusters that are insensitive to these landcover type are not appropriate for estimating CH4 fluxes. Purpose of study of to identify cluster that represent CH4 and other GHG flues, and not to identify site that reprsent the clusters well. I believe the variable and clustering approach applied are not able to capture the heterogeneities on conditions that drive CO2 and CH4 fluxes, especially in wetlands, croplands and near aquatic sites.
- Lines 270-272 "We found the size of the cluster is not correlated to the number of towers when all towers are included in the analysis but was slightly negatively correlated with the number of EC towers that include CH 4 measurements (Figure 2)." I am not sure such correlations are meaningful at all. What is the purpose of correlations (postive or negative)?
- Mediod locations of 10 clusters would be the theoretically optimal locations for locating an EC site. It would help to see a map of where these 10 locations and are perhaps a discussion of how well they appear to capture the local methane source/sinks on the ground. 10 is a small enough number to present a short and meaninful discussion to show the effectiveness of cluster medioid method.

Few sugestions:

- Please consider including additional variables such as soil moisture, some measure of inundation, soil organic carbon to better capture the CH4 sources/sinks.
- Simplify the methodology and cluster the entire US and not a small 20,000 subsample to make the best use of information and variability captured in the data. Clustering + MDS + RF is unnecessarily complicated and perhaps hurt and not help the analysis.
- It would be of more value to consider the operational vs non-operational status of the EC sites in the analysis, so the results can inform actionable decisions.