

Biogeosciences Discuss., author comment AC2
<https://doi.org/10.5194/bg-2021-256-AC2>, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Reply on RC2

Sparkle L. Malone et al.

Author 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-AC2>, 2021

Authors would like to recognize the thoughtful comments provided by the Reviewer which led to several important changes in our approach. We clarified the goals of this study, we focused on tower infrastructure currently measuring CH₄, and we better explained how we are measuring representativeness.

We used a combination of climate data and dominant land cover types to guide the scientific community on how we can develop a distributed observing network for the US and provide a template for the development of similar networks in other regions. We focus here on EC flux towers because they are essential for a bottom-up framework that bridges the gap between point-based chamber measurements and airborne platforms and are therefore a useful basis for identifying gaps in the current network of CH₄ observations. To understand the landscape representativeness across geographic clusters, we measured dissimilarity based on climate and land cover type, as these two factors together are characteristic of regional resource availability and disturbance regimes. It is important to note that a tower is representative of the ecosystem type and the region where it is stationed (Desai, 2010; Jung et al., 2011; Xiao et al., 2012; Chu et al., 2021); however, the landscape representativeness analysis done here uses a coarser classification of ecosystem types that are more emblematic of regional disturbance regimes, resource availability, and factors that influence how ecosystems function, not the specific ecosystem type where the tower is situated. Chu et al., 2021 examined the land-cover composition and vegetation characteristics of 214 AmeriFlux tower site footprints. They found that most sites do not represent the dominant land-cover type of the ecosystems they exist within and when paired with common model-data integration approaches this mis-match introduces biases on the order of 4%–20% for EVI and 6%–20% for the dominant land cover percentage (Chu et al. 2021), making it essential to consider landscape characteristics in the design and evaluation of network infrastructure. Tower representativeness at the landscape scale is indicative of the capacity to upscale information by climate and the dominant ecosystems of locations within a landscape. We also calculate regional representativeness by the towers' vegetation type to understand the sampling intensity of each vegetation type within a cluster, which is also an essential component of scaling CH₄ fluxes (Knox et al., 2019). In this analysis we used the reported International Geosphere-Biosphere Programme (IGBP) vegetation type that is listed for each tower in the AmeriFlux data base, where we also checked to ensure towers were currently active and providing data to the network.

Limitations of existing land use products: We moved the information on the limitation

of current land cover products from the introduction to the discussion. We did this because having it in the introduction combined with the changes we made in the product gave the impression that the changes we made improved on the current product in a way that would reduce uncertainties. In fact that is not the case, the same uncertainties exist. The product we developed only allowed us to distinguish a few wetland and aquatic classes from each other.

Land cover composition is a response to climate and disturbance regimes. Climatic factors influence resource availability (light, water, temperature) and although it varies geographically, disturbance is tightly linked to land cover types and cover classes have characteristic disturbances that influence the composition of classes within a landscape (Hermosilla et al. 2018). Studies have shown that either changes in climate or disturbance can lead to shifts in landscape composition (Sharp and Angelini 2019; Wang et al. 2020). For example, changes in the plant functional types across Arctic-Boreal ecosystems have been linked to wildfires and climate-driven expansion of woody shrubs (Wang et al. 2020). The interaction of climate and disturbance both influence landscape structure and develop spatial patterns of many ecosystem processes on the landscape (Turner 2010).

Dimensions of MDS: Climate, ecotype, and location (latitude/longitude) were used in a multivariate distance analysis to define the state space of the US (all 50 states & Puerto Rico) at the landscape scale and to divide the US into clusters. The purpose of this analysis is to identify the interrelatedness of all ecological components—biotic, abiotic, terrestrial, and aquatic within a dynamic landscape (Ippoliti et al., 2019). We included location (latitude/longitude) to incorporate the interaction between climate, ecotypes, and most importantly, seasonality. We first developed a dissimilarity matrix by calculating Gower dissimilarity (Gower, 1971; Huang, 1997; Podani, 1999; Ahmad and Dey, 2007; Harikumar and Pv, 2015) using the function `dismix()` from the package *kmed*. We used Gower dissimilarity because it can handle mixed data types. For each variable type in the data set, the dissimilarity metric that works well for that type is used and scaled to fall between 0 and 1. Then, a linear combination using user-specified weights (most simply an average) is calculated to create the final dissimilarity matrix. This approach measures the dissimilarity for each location within the U.S using ecotype, climate, and location information (ecotype, five climate variables, and location) and creating a dissimilarity matrix (20,000 x 20,000) that indicates dissimilarity for a location to every other location in the US.

Once we created a dissimilarity matrix, we used multidimensional scaling (MDS) to generate a two-dimensional ordination showing landscape dissimilarity with the *MASS* package in R (Venables and Ripley, 2002). The MDS makes it possible to evaluate dissimilarity in two dimensions, which is essential to our goal to evaluate representativeness. We used the Kruskal method of non-metric scaling with the *IsoMDS* function in the *MASS* package (Venables and Ripley, 2002). *IsoMDS* works best when applied to metric variables (Torgerson, 1958). Torgerson (1958) initially developed this method, which assumes that the data obey distance axioms. It uses eigendecomposition of the dissimilarity to identify major components and axes, and represents any point as a linear combination of dimensions. This is very similar to PCA or factor analysis, but it uses the dissimilarity matrix rather than a correlation matrix as input. Furthermore, the included dimensions are the most important dimensions produced, like PCA which is able to identify all of the dimensions that exist in the original data up to $N-1$, but will retain only the most important ones.

Knowing that regional patterns in climate and land cover will be important for scaling CH_4 to the regional and national scale, we divided the US into clusters to evaluate representativeness. This cluster analysis also allows us to summarize our results within a geographical context, an approach that has been used to delineate spatial sampling domains, to assess the spatial representativeness of networks, and to suggest

arrangements of study sites (Sulkava et al. 2011; Kumar et al. 2016). We used a cluster analysis to determine the optimal number of clusters using the library *cluster* and the function *pam* in R (Reynolds et al. 2006; Schubert and Rousseeuw 2019; Schubert and Rousseeuw 2021). This approach uses the k-medoids algorithm, which partitions data set into k groups or clusters and is a robust alternative to k-means clustering (Kaufman and Rousseeuw 2009). Each cluster is represented by one of the data points in the cluster named the cluster medoid. The medoid has the lowest average dissimilarity between it and all other objects in the cluster. The medoid can be considered a representative example of the members of that cluster. The k-medoid algorithm is less sensitive to noise and outliers, compared to k-means, because it uses medoids as cluster centers. The k-medoids algorithm requires the user to specify k, the number of clusters to be generated. A useful approach to determine the optimal number of clusters is the **silhouette** method. We fit an increasing number of clusters from 2 to 20 to construct a silhouette plot and choose the number of clusters that maximized the average silhouette width. Once we determined the number of clusters and the medoid of the cluster, we calculated the dissimilarity between every location within the cluster to the medoid to create a measure of how different each location was from the medoid condition of each cluster. We utilized the *pointDistance* function in the *raster* package, which provided a unit-less relative measure of dissimilarity that was determined by measuring the difference between the first and second dimensions produced by the isoMDS of each point in a cluster to the dimensions of the medoid. This analysis was repeated 10 times to ensure that the 20,000 pixel subsample would produce similar results in the dimensions and clustering. For simplicity, we show the results of the first analysis.

MDS was conducted only on a subset of 20,000 1km pixels: To measure dissimilarity, we first randomly sampled ($n = 20,000$ 1-km pixels) the US, maintaining the distribution of ecotypes and climate to define dissimilarity between observations. Although there were more than 8 million sampling points available for the U.S, there are limits to the number of samples that can be analyzed by the functions used for the MDS analysis. This analysis was repeated 10 times to ensure that the 20,000 pixel subsample would produce similar results in the dimensions and clustering. For simplicity, we show the results of the first analysis.

To extrapolate the cluster and dissimilarity 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 dimensions using the ecotype and climate layers as predictors. We then created a Random Forest model of the cluster layer using the first and second dimension as the explanatory variables. All models were then projected spatially to produce a spatially explicit cluster layer and a dissimilarity layer beyond the 20,000 sample points that were used in the MDS analysis. This allowed us to measure the importance of the original data on the first and second dimensions defined by the MDS and how the MDS leads to cluster and dissimilarity patterns. This step was essential to producing a spatially explicit cluster and dissimilarity layers for the entire US, since the MDS analysis limits the number of observations that can be analyzed. We added more detail to the text and the results of the random forest model and the importance of the original data in explaining clustering and the first and second MDS dimensions.

What was the purpose of these correlations? This was originally included to explain the main determinants of the MDS and clustering. We removed it from the analysis. The results are redundant with the inclusion of the random forest variable importance plots, which will be included in the supplement.

More discussion of methods and their implication on results are needed:

We provide more details in the methods and summarize it here. Knowing that regional patterns in climate and land cover will be important for scaling CH_4 to the regional and

national scale, after creating a dissimilarity matrix that was then scaled to two dimensions, we divided the US into clusters to evaluate representativeness using the first and second dimension from the MDS. This cluster analysis allowed us to summarize our results within a geographical context, an approach that has been used to delineate spatial sampling domains, to assess the spatial representativeness of networks, and to suggest arrangements of study sites (Sulkava et al. 2011; Kumar et al. 2016). We used a cluster analysis to determine the optimal number of clusters using the library *cluster* and the function *pam* in R (Reynolds et al. 2006; Schubert and Rousseeuw 2019; Schubert and Rousseeuw 2021). This approach uses the k-medoids algorithm, which partitions data set into k groups or clusters and is a robust alternative to k-means clustering (Kaufman and Rousseeuw 2009). Each cluster is represented by one of the data points in the cluster named the cluster medoid. The medoid has the lowest average dissimilarity between it and all other objects in the cluster. The medoid can be considered a representative example of the members of that cluster. The k-medoid algorithm is less sensitive to noise and outliers, compared to k-means, because it uses medoids as cluster centers. The k-medoids algorithm requires the user to specify k, the number of clusters to be generated. A useful approach to determine the optimal number of clusters is the **silhouette** method. We fit an increasing number of clusters from 2 to 20 to construct a silhouette plot and choose the number of clusters that maximized the average silhouette width. Once we determined the number of clusters and the medoid of the cluster, we calculated the dissimilarity between every location within the cluster to the medoid to create a measure of how different each location was from the medoid condition of each cluster. We utilized the *pointDistance* function in the *raster* package, which provided a unit-less relative measure of dissimilarity that was determined by measuring the difference between the first and second dimensions produced by the isoMDS of each point in a cluster to the dimensions of the medoid.

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? We will add more detail to the results to show that the **silhouette** plots were used to determine the optimal number of clusters.

Extrapolate the cluster and distance: To measure dissimilarity, we first randomly sampled ($n = 20,000$ 1-km pixels) the US, maintaining the distribution of ecotypes and climate to define dissimilarity between observations. Although there were more than 8 million sampling points available for the U.S, there are limits to the number of samples that can be analyzed by the functions used for the MDS analysis. This analysis was repeated 10 times to ensure that the 20,000 pixel subsample would produce similar results in the dimensions and clustering. For simplicity, we show the results of the first analysis.

To extrapolate the cluster and dissimilarity 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 dimensions using the ecotype and climate layers as predictors. We then created a Random Forest model of the cluster layer using the first and second dimension as the explanatory variables. All models were then projected spatially to produce a spatially explicit cluster layer and a dissimilarity layer beyond the 20,000 sample points that were used in the MDS analysis. This allowed us to measure the importance of the original data on the first and second dimensions defined by the MDS and how the MDS leads to cluster and dissimilarity patterns. This step was essential to producing a spatially explicit cluster and dissimilarity layers for the entire US, since the MDS analysis limits the number of observations that can be analyzed. We added more detail to the text and the results of the random forest model and the importance of the original data in explaining clustering and the first and second MDS dimensions.

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: These correlations were the average, not for clustering but for the first and second dimensions: We removed these results and replaced them with the variable importance from the random forest. It is important to note that the analysis done here will not capture the heterogeneity of the conditions that drive CH₄ fluxes at the ecosystem scale. It is designed to evaluate the sampling intensity of research sites at the landscape scale. In the design of a network, this coarse resolution influences the capacity to scale ecosystem level results to the landscape, region, and to the national level, which is required for the development of budgets and emission strategies.

I am not sure such correlations are meaningful at all. What is the purpose of correlations (positive or negative)? This was removed from the analysis. It was originally included to show the poor relationship between cluster size and the frequency of towers. We agree with the review, this figure is unnecessary and is no longer relevant, since we are no longer including all towers.

It would help to see a map of where these 10 locations are and perhaps a discussion of how well they appear to capture the local methane source/sinks on the ground: Figure 2 (see attachment) shows patterns in dissimilarity across the US with cyan locations having the lowest dissimilarity. We will add additional maps to the supplemental to highlight areas with the lowest dissimilarity. We also agree that it would be informative to show how towers in the medoid capture the local methane source/sinks on the ground. Unfortunately, towers are not currently distributed across cluster/medoid or IGBP to facilitate an evaluation of source sink potential. This highlights the limitations of existing measurement infrastructure which requires strategic augmentation to provide the most valuable information toward reducing uncertainties in future large-scale budget estimations. This analysis complements previous studies based on climatic or vegetation characteristics (Hargrove et al. 2003; Yang et al. 2008; Villarreal et al. 2018), and identifies regions within the US where gaps are limiting the development of upscaling techniques. To accurately understand the impact of climate and land cover change on biogenic CH₄ emissions, we need a long-term, calibrated, and strategic continental-scale CH₄ observatory network. Current gaps in existing measurement infrastructure limit our ability to capture the spatial and temporal variations of biogenic CH₄ fluxes and therefore limit our ability to predict future CH₄ emissions. Maps of potential CH₄ emissions require land cover classification targeted at land cover types like wetlands that are important sources of CH₄ to the atmosphere. Aquatic ecosystems like streams and lakes as well as coastal ecosystems are significant and variable sources of CH₄ not well studied on a long-term basis. Through our analysis using climate, land cover, and location variables, we have identified priority areas to enhance research infrastructure to provide a more complete understanding of the CH₄ flux potential of ecosystem types in the US. For EC tower locations, dissimilarity coverage was lacking for clusters Na, W, and Nb, and currently clusters Na, W, Eb, and Nb are substantially under sampled. All aquatic sites are under sampled within each cluster. An enhanced network would allow for us to monitor both the response of CH₄ fluxes to climate and land use change as well as the impact of future policy interventions and mitigation strategies.

Few suggestions:

Please consider including additional variables such as soil moisture, some measure of inundation, soil organic carbon to better capture the CH₄ sources/sinks: While measures of soil moisture, inundation, and soil organic carbon are important drivers of ecosystem CH₄ production and consumption, at the scale we are working on climatic conditions are more appropriate, as to not suggest we are able to represent those layers in a meaningful way at a 1 km resolution. The landscape representativeness analysis done here uses a coarser classification of ecosystem types

that are more emblematic of regional disturbance regimes, resource availability, and factors that influence how ecosystems function, not the specific ecosystem type where the tower is situated. Chu et al., 2021 examined the land-cover composition and vegetation characteristics of 214 AmeriFlux tower site footprints. They found that most sites do not represent the dominant land-cover type of the ecosystems they exist within and when paired with common model-data integration approaches this mis-match introduces biases on the order of 4%–20% for EVI and 6%–20% for the dominant land cover percentage (Chu et al. 2021), making it essential to consider landscape characteristics in the design and evaluation of network infrastructure. Tower representativeness at the landscape scale is indicative of the capacity to upscale information by climate and the dominant ecosystems of locations within a landscape. We also calculate regional representativeness by the towers' vegetation type to understand the sampling intensity of each vegetation type within a cluster, which is also an essential component of scaling CH₄ fluxes (Knox et al., 2019).

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: To measure dissimilarity, we first randomly sampled ($n = 20,000$ 1-km pixels) the US, maintaining the distribution of ecotypes and climate to define dissimilarity between observations. Although there were more than 8 million sampling points available for the U.S, there are limits to the number of samples that can be analyzed by the functions used for the MDS analysis. This analysis was repeated 10 times to ensure that the 20,000 pixel subsample would produce similar results in the dimensions and clustering. For simplicity, we show the results of the first analysis.

To extrapolate the cluster and dissimilarity 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 dimensions using the ecotype and climate layers as predictors. We then created a Random Forest model of the cluster layer using the first and second dimension as the explanatory variables. All models were then projected spatially to produce a spatially explicit cluster layer and a dissimilarity layer beyond the 20,000 sample points that were used in the MDS analysis. This allowed us to measure the importance of the original data on the first and second dimensions defined by the MDS and how the MDS leads to cluster and dissimilarity patterns. This step was essential to producing a spatially explicit cluster and dissimilarity layers for the entire US, since the MDS analysis limits the number of observations that can be analyzed. We added more detail to the text and the results of the random forest model and the importance of the original data in explaining clustering and the first and second MDS dimensions.

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: Thank you for this comment. We made changes to the towers used in this analysis. Although we have information on the location of existing EC tower infrastructure that is a part of AmeriFlux ($N=223$), NEON ($N=47$), and known, independent PI-managed sites ($n=141$). We focus this analysis on the towers measuring CH₄ ($n=100$) and we distinguish between towers providing data to AmeriFlux (yes = 49, no = 51) and tower activity (active = 70; inactive = 30). We understand that additional towers exist within the U.S, but because these towers are not reporting or providing data to the flux community, we cannot include them in this analysis.

Table 4. The R_{cluster} for CH₄ towers that are active and providing data to AmeriFlux, the R_{cluster} for all active CH₄ towers and the R_{cluster} for all active towers in addition to NEON towers.

Cluster	CH₄ Towers (Data Providing)	CH₄ Towers (All)	NEON Towers
Na	3.0	34.9	35.5
NW	-	0.1	26.3
NEb	19.8	60.6	65.9
Ea	0.01	63.1	89.4
Eb	88.1	88.1	88.1
SW	2.0	3.3	17.3
W	0.01	0.01	38.8
NEa	79.3	79.3	79.3

Nb	21.3	21.3	21.3
SE	-	23.6	50.8

There were important gains in $TR_{cluster}$ when considering all CH_4 towers regardless of if they were providing data to Ameriflux (Table 4 and Figure 4). The clusters with substantial gains in representativeness ($> 10\%$) include Na, NEb, Ea, and the SE. The $TR_{cluster}$ of the NW, Ea, SW, W, and the SE would be enhanced by more than 10% with the addition of CH_4 instrumentation at NEON sites.

A detailed response to the review comments is attached.

Please also note the supplement to this comment:

<https://bg.copernicus.org/preprints/bg-2021-256/bg-2021-256-AC2-supplement.pdf>