The authors suggested a classical geostatistical approach (using semivariogram and kriging) to deal with spatial heterogeneity in point and grid data comparisons. The value of the approach was also demonstrated using both theoretical and real-world experiments. Overall, this paper is well written and falls into the scope of AMT. The following comments need to be addressed before publishing.

We thank this reviewer for their positive/generous comments. Our response follows:

The approach suggested in this manuscript is very useful. However, given the complexity and computational cost, it may not always be practical for other groups to apply. I'm wondering if the authors have any suggestions or comments on this?

The semivariogram modeling and calculations along with different types of kriging (ordinary, simple, universal, …) have been implemented in many different programming languages. It is extremely fast to get the results such that all of experiments presented in the paper were done on a personal laptop. A couple of useful links on publicly available geostatistical toolboxes:

MATLAB: https://www.omicron.dk/dace.html
https://www.mathworks.com/matlabcentral/fileexchange/25948-variogramfit
https://www.mathworks.com/matlabcentral/fileexchange/29025-ordinary-kriging
Python: https://geostat-framework.readthedocs.io/projects/pykrige/en/stable/
https://gmd.copernicus.org/preprints/gmd-2021-301/
R: http://www.gstat.org/

The authors suggest that when the sampling is sparse (<3 samples within the field) and a good degree of homogeneity can't be assumed, a quantitative comparison between satellite and observations is discouraged. However, it is common for some ground networks to have sparse data. For example, the TCCON sites are generally far from each other and are often used for satellite evaluation. I'm wondering if the authors have any comments on this.

Mathematically, it is impossible to gain spatial variance from discrete data if we have fewer than 3 point samples. This is really pure math, and is not a suggestion made by the present study. That means those sparse networks such as TCCON will never be able to provide the information on the spatial distribution of greenhouse gases at the scale of satellite footprints. There are two solutions to this problem: i) we should increase the number of point observations for the satellite validation purpose, and ii) we may use higher spatial resolution airborne data such as MethaneAIR (https://amt.copernicus.org/articles/14/3737/2021/amt-14-3737-2021.html) flying over TCCON stations. These data have sufficiently fine spatial resolution such that you can directly compare them to the point measurements without being too stressed about the problem of scale. The bias-corrected airborne observations then can be upscaled to a satellite footprint and further be compared. On a brighter note, there are many air quality related campaigns possessing a dense network of spectrometers. Furthermore, EPA provides a large suite of surface observations with relatively high spatial density.

In Section 2.1, a_0 is used multiple times (e.g., equations 4 and 7). Are they the same? To avoid any potential confusion, if they are the same, please explicitly indicate it. Otherwise please use different symbols.

Thanks for your comment, we have changed it to make sure it won't be confused with a0.

Figure 1: This change is not necessary, but it would be great if you could make the semivariogram plot for C1 to have the same x-axis labels as the C2-C5.

Line 236 Please briefly explain how did you "classify the domain into four zones using the k-mean algorithm" if possible.

We simply classify the magnitude of Z(x) using k-means. The inputs are only the magnitudes.

We classify the domain into four zones by running the k-mean algorithm on the magnitudes of $Z(x)$

Line 276: Please change "distance" to "difference" or "bias" to avoid any potential confusion.

Sure, we changed it to difference.