'Dealing with Spatial Heterogeneity in Pointwise to Gridded Data Comparisons' demonstrates a new methodology for comparing point measurements to those that are more spatially representative of an area (e.g., satellite or models) for the purpose of data product validation. This topic takes an important challenge with these types of data comparisons for the purposes of validation and attempts to lay out a solution going forward. The manuscript is very well written and organized and it fits well within the scope of AMT and should be considered for publishing after some minor changes.

We thanks reviewer for their thoughtful and constructive comments. Our response follows:

General comments:

- The paper switches back and forth between the nomenclature of using the method in relation to validation vs. data comparisons. While the methodology and conclusions made related to using point measurements for validating grid-like data are completely valid and important, the language in this paper is dismissive of other potential reasons to compare point-like measurements with something like a model or satellite. Therefore, the suggestion is to clarify that this methodology is for validation and not generalize conclusions for data comparisons.

  Thanks for your comment. To be able to validate volumetric data (such as satellites and models) using pointwise measurements, we have to compare them. This comparison is apples to oranges. Since all apples-to-oranges comparisons are wrong, we, as scientists must be alert to what is importantly wrong with respect to such a comparison. Validation comes with data comparisons, and data comparisons are meant for validation of a product or a hypothesis. They both are inextricably linked together. So we cannot say that it is okay to compare two different things as long as we do not want to use the quantitative statistics made from the comparison.

  - Examples of this type of language include, but are not necessarily limited to:
    - Line 80-81 with the question of whether a 'comparison ever logical' – in the right case the comparison could be logical when trying to learn how satellite data can be interpreted in relation to a ground-based measurements

    To account for this, we added: If one compares a grid box to a point sample (i.e., apples to oranges)

    And:

    can the average of the spatial distribution of the underlying compound be represented by a single value measured at a subgrid location?

    - Line 551: change 'comparisons' to something along the lines of validation

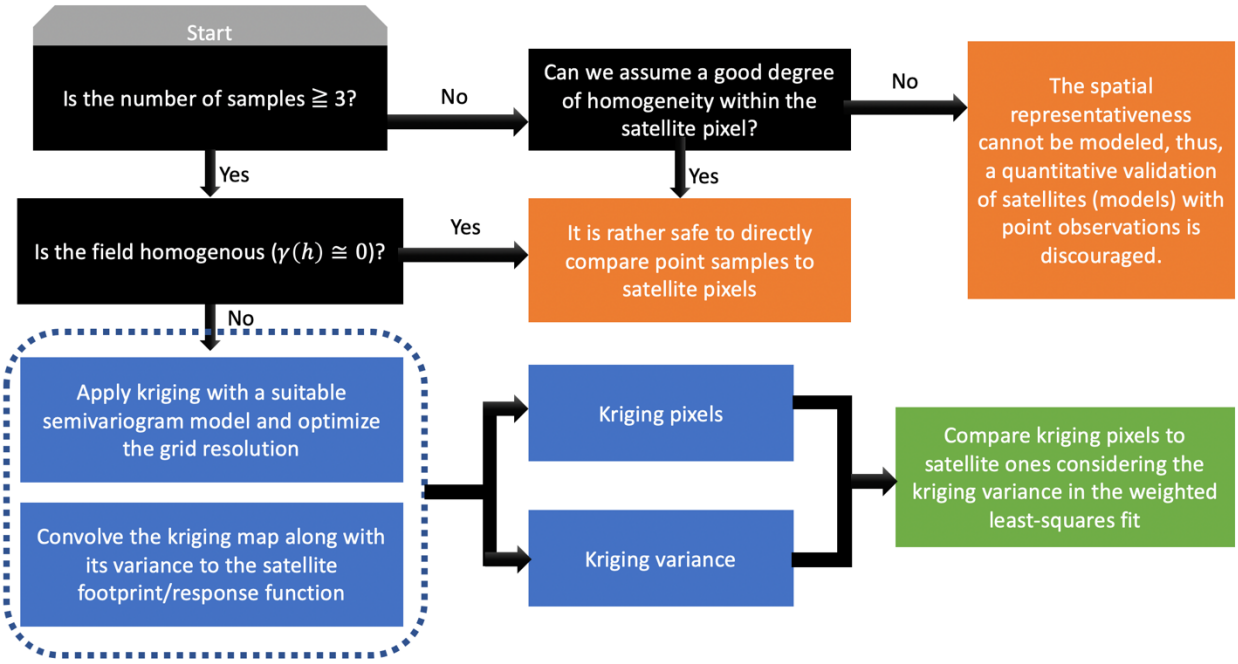
    "The validation against point measurements can be carefully conducted in the following steps:"

    - Line 572: 'point-pixel comparisons' should say pixel validation with point measurements or something along that line.

    We changed it to: "validating satellites/models using pointwise measurements"

- Figure 9: in the right orange box change ‘comparison between satellites and observations’ to ‘validation of satellites (models) with point observations’

We changed it:



- Could the authors comment on the reality of ground-based networks that could actually contribute to satellite/model validation with this methodology? Does the required observational density exist anywhere? What are some paths forward/recommendations?

Several DISCOVER-AQ campaigns such as Texas 2013 and Colorado 2014 possess an adequate number of Pandora samples (~10-12) to validate sensors like OMI or GOME. We are obviously in better shape when it comes to model validation. EPA provides a very dense network of NO2, O3, and PM2.5 in the U.S. providing a sufficient number of samples to extract a good degree of spatial variance. It is also desirable to compare surface NO2 concentrations to satellite columns which can be effortlessly done with this method. Indeed, there are networks that do not meet the requirement on the availability of the observations at small scales. Off the top of our head, the FTIR HCHO network is one of them. For those cases, we recommend oversample columns from a high-resolution sensor to gauge the level of spatial heterogeneity in the field at the resolution of a coarser satellite/model. That level of heterogeneity should be accounted for in the comparisons made between pixel vs points; for example, over a city, a larger deviation between point and pixel from the identity line can partly be explained by a larger level of spatial heterogeneity suggested by the oversampled field from the fine sensor. Our method won’t be applicable for this case, but there are ways to articulate that the problem of scale exists and correlates with the discrepancies we observe between the two different datasets.

- Section 4: Why is v3.0 OMI data using instead of the most up to date v4.0?

Lamsal, L. N., Krotkov, N. A., Vasilkov, A., Marchenko, S., Qin, W., Yang, E.-S., Fasnacht, Z., Joiner, J., Choi, S., Haffner, D., Swartz, W. H., Fisher, B., and Bucsela, E.: Ozone Monitoring Instrument (OMI) Aura nitrogen dioxide standard product version 4.0 with improved surface and cloud treatments, Atmos. Meas. Tech., 14, 455–479, https://doi.org/10.5194/amt-14-455-2021, 2021.

Our first attempt to validate the tropospheric OMI NO2 from Goddard was to use the most updated version. Unfortunately this new dataset showed some unrealistic distributions of NO2 at the northeast of the domain. We closely look at SCD and AMF variables and realized that the artifact was not induced by SCD but rather it originated from the shape factors. Unfortunately at that point, we did not have the CTM outputs used in Souri et al. [2016] to rectify the issue. Luckily, the older version of OMI data whose AMFs had already been recalculated using high resolution WRF-CMAQ model presented in Souri et al., [2016] were available at that time. So we resorted to using that version which was free of the artifact. Regardless of the version, OMI shows incredible accuracy over the region which explains why the study of Souri et al. [2016] was successful at constraining emissions.

Specific comments:

- The first sentence in the abstract implies that the two communities have zero realization of the point vs grid problem, which is not true. This is a known problem with the lack of an easy solution. Please consider rephrasing.

We changed the subject to be less direct and used "assume" to be less harsh: "Most studies on validation of satellite trace gas retrievals or atmospheric chemical transport models assume that pointwise measurements…".

- Lines 28-29: This study demonstrates a method but it doesn't actually prove that the only available method 'must taking kriging variance…'etc. State what the paper demonstrates without implying there is no other alternative to this exact method.

  We agree. We added: "This study suggests that satellite validation procedures using the present method must take kriging variance and satellite spatial response functions into account."

- Line 50: consider adding some clarity to the hypothetic scenario by adding after 'atmospheric model' the phrase 'simulating CO2 emissions'…

  Sure, we added: "simulating CO_2 concentrations".

- Line 191: is there a reference for the terminology of 'the sill'?

  Yes, we added "[Chilès and Delfiner, 2009]"

- Paragraph spanning lines 232-239: Use the word stratified somewhere to connect to the second row of the figure.

We added: "As a remedy, it might be advantageous to group the domain into similar zones and randomly sample from each, which is commonly known as stratified random selection"

- Line 385: add that this is also a roadmap for model evaluation as well.

  Added.

- Line 414: add a reference for the length scale of NO2 if connecting it to the range found in those months.

  Sure, "These numbers strongly coincide with the seasonal lifetime of NO_2 [Shah et al., 2020]"

- Lines 433-440: More clearly explain the rational more clearly between 10 vs. 15 vs. 20. In the figure alone it isn't all that clear. Was the choice of 15 quantified somehow as the best option or was it subjective?

  We accounted the cost of having more spectrometers. To clarify:

  "The difference between kriging estimate and the TROPOMI observations using 20 samples does not substantially differ in comparison to the one using 15 samples. Therefore, to keep the cost low, a preferable strategy is to keep the number of spectrometers as low as possible while achieving a reasonable accuracy. Based on the presented results, the optimized tessellation using 15 samples is preferred among others because it achieves roughly the same accuracy as the one with 20 samples."

- Line 449: Should be Herman et al. 2009

  Fixed.

- Line 454-456: what resolution is OMI oversampled to?

  0.2º. We had mentioned that:" Following Sun et al. [2018], we oversample high quality pixels in the month of September 2013 over Houston at 0.2º resolution."

- Lines 456-457 and Figure 12 caption. The wording states that the total column was subtracted from the stratospheric column, when it is the stratospheric column that should be subtracted from the total column. Please fix this wording.

  Indeed, we fixed it.