Review of GMD-2022-230

This study applies a functional ANOVA (analysis of variance) method to extract spatially-varying effects related to different experimental factors from carbon dioxide inversion results. The ideas are not completely new, but some computationally efficient tricks are introduced, and the methodology is applied to a few different questions in the context of interpreting fluxes from atmospheric inversions. The goal is to quantify the impact of different sources of data, different methods of data aggregation, and the use of different inversion models. The code has been made public, and appears to be sufficiently documented. (I did not replicate the results, but a cursory inspection of the code repository suggests that this would be possible.) The study is generally well written, and the figures are of high quality. The subject matter is appropriate for publication in GMD once some issues are addressed.

A significant concern about the methodology is related to the use of adjacent months (June, July, and August) as independent replicates to represent the accumulated JJA signal. This seems questionable in two regards:

1. These are not truly independent: There is certainly some correlation between the fluxes of adjacent months, in terms of the data constraint, the geophysical processes involved, and likely even the optimization of the fluxes via a 4DVar approach.
2. These are not realizations of the same quantity: The fluxes for June are not a replicate of the fluxes for August, due to both seasonal cycles (as can be clearly seen in the columns of Figure 1) and differences in short-term anomalies (e.g. June could be anomalously hot and dry while August of the same year could be anomalously cool and wet).

Both of these points call into question whether this choice is valid. This is rather different than the previous studies mentioned in the conclusion, which used different years (not adjacent months!) to infer spatially-varying climate signals and model effects in the presence of interannual variability. (The studies cited that analyzed climate models also used an order of magnitude more replicates.)

The findings seem to suggest that, indeed, these are not appropriate to use as replicates. Lines 242 and 243 point out that the estimated GP standard deviation for the is largest for the error fields (10 times larger, in fact), which indicates that the month-to-month variability within each treatment combination is rather large. Indeed, the standard deviation of the error term is largest for all the experiments, as seen in Tables 3 and 4. This is discussed in the conclusion, with the remark that such a “low-replication, high-variance scenario often translates to higher uncertainty in the other ANOVA effects and a tendency for them to shrink to the assumed population mean”, namely zero. Is there a way that the study could have been conceived in order to have a larger number of replicates?

Regardless, the authors argue that they have been able to extract “significant anomalies”. It is unclear to me how to judge the significance of these results. Is it just because there are coherent spatial patterns? The 95% credible intervals are generally smaller for the error than for the mean, α, β, or interaction terms – does this say something about the relative significance of the results? The authors argue that the large standard deviation of the error fields is offset by the relatively small estimated range parameter of the error term. How is this distance to be interpreted? In some cases, the range parameter is significantly smaller than the size of the models’ pixels (e.g. a median value of 70.5 km for the range parameter of the error for the African region in Table 4, compared to model resolutions an order of magnitude larger, from 4°x5° to 6.7°x6.7°, based on Table 1 of Peiro et al., 2022). It would be helpful to provide some more information about how to interpret these numbers. As another example, the unitless smoothing parameter is reported in the tables, but these results are never discussed in the text. What do these mean?

I think it would be useful to include some discussion about the similarities and the differences between your approach (and results) and those of Cressie et al. (2022). The latter study applied ANOVA to the same two years of very similar OCO-2 MIP results (a previous iteration), also using individual months of a three-month season as replicates (which seems questionable in their study as well). The application was different, however, as they were using it to develop an optimally-weighted multi-model mean for regional fluxes, and considered all seasons and regions covering the whole globe.

This raises the question: why did you decide to focus on JJA, and why only on this one season? How would you expect the results to change if different seasons were analyzed? Why did you only use two years of data, when the OCO-2 MIP results you are using (from Peiro et al., 2022) provide four full years of valid results? What led you to look only at these specific regions? There is some discussion of why North America and Africa were chosen to illustrate the technique, due to contrasting measurement coverage, but there is no clear argument as to why Eurasia was chosen for the first part of the study. Providing a justification for why a given region was chosen to explore the specific question would make the scientific results more compelling.

In general, more scientific interpretation of the spatial patterns that are presented as results would improve the paper. In particular: in the plots of the "Data Source Effect" in Figures 7 and 10, how do these patterns correspond to the location of in situ measurement sites? (Yes, the stations are shown in Figure 1 of Peira et al. (2022), but putting the site locations onto your figures would make this easier to interpret.) Where is there persistent cloud cover and/or good retrievals of OCO-2 data in North America and Africa during the period in question? Interpreting the structure of these "Data Source Effect" figures with reference to the measurements would be helpful.

Another concern regarding the methodology is related to the fact that all the ANOVA analysis in the MIP portion of the study is done on the flux increment, i.e. the posterior flux minus the prior flux for a given space and time, as shown in Equation 1. I believe that this is problematic, because the four models whose results are compared do not have identical priors. Extracted from Table 1 of Peiro et al. (2022), the following prior fluxes were used:

Simulation name	**Prior land**	**Prior ocean**	**Prior fire**
Ames	CASA-GFED4.1s	CT2019OI	GFED4.1s
CMS-Flux	CARDAMOM	ECCOS-Darwin	GFED4.1s
OU	CASA-GFED3	Takahashi	GFEDv3
Baker	CASA-GFED3	Takahashi	GFEDv3

All the inversions are constrained by the same data, and we would expect that if transport is perfect, the data constraint is sufficient, and errors are well-characterized, the four models should converge to the same posterior fluxes. That does not mean that their flux *increments* will be the same. Consider a case where the prior of model A is biased systematically high over a given region, while the prior of model B is biased systematically low. Both models converge to the same response, but when analyzing the increment ($y_{ijk}(\mathbf{s})$- $y_{ijk\,(0)}(\mathbf{s})$), model A's would be negative and model B's would be positive. If the biases were equally biased, the "consensus flux anomaly" of this ensemble of two would be zero. This is written in the text, e.g. in lines 283-284: "Noting again that the analysis is carried out the inversions' deviations from their respective priors, the map depicts a consensus flux anomaly from the collection of inversions analyzed." How can we interpret a "consensus flux anomaly" when the models are starting from different priors? (This does not apply to the first part of the study, only the MIP results.)

Going back to the theoretical model A and model B from before: these increments would presumably show up as negative and positive flux model differences with a large spatial range, perhaps similar to the all-red (Ames) and all-blue (CMS-Flux) posterior mean maps seen for α for JJJA in North America (Figure 7). However, the authors attribute this to "large-scale regional flux differences among inversion systems with different driving atmospheric transport" (L275-276). This could be the case, but without having common priors, the effects of different priors and different transport cannot be teased apart using this method. Ideally, such an analysis would be carried out on inversion results with identical priors, so that we could truly say something about the influence of the transport model and optimization. Lacking that, this limitation needs to be made completely clear, so that the reader is not misled.

As a more minor/technical note: The treatment of the different spatial resolutions in the MIP part of the study should be explicitly explained. For the first part of the study, only results from the CMS-Flux model are used, and the plots appear to match the 4°x5° resolution of the model. For the MIP, models of different resolutions are compared (but none with a higher resolution than the still relatively coarse 4°x5° resolution of CMS-Flux), but the data appear to have been resampled onto a higher spatial grid in Figures 6, 8 and 9. In Figures 7, 10 and 11, the resolution appears smoother still. What procedure was followed?

Minor/typographical comments:

L1-2: This sentence is a bit odd, making it sound as if measurements made today (present tense) span several years. Perhaps: "has now produced atmospheric greenhouse gas concentration estimates covering a period of several years"?

L8: Not clear what is meant by "mode differences" here.

L10: For inversions over different continents, or for fluxes over different continents?

L17: The language here makes it sound as if the satellites are providing an estimate of global CO2 concentrations frequently, which is not the case. Instead of "frequent", perhaps quantify it? How many good quality measurements per year, on average?

L18: Has flux inversion become demonstrably more tractable (not quite sure what that means here) and precise due to satellite measurements? If so, please provide a reference providing evidence of this!

L24-25: Not clear what is meant by "dependence can be exploited in quantifying uncertainty". Dependence of what on what? Please explain.

L39: collection Earth -> "collection of Earth"

L41: Hyphenate "multiple-component" as a compound adjective, also "spatially-aggregated" on L45.

L47: with associated -> "with an associated"

L55: Is the year missing from Sain et al. reference here? Is this consistent with the Copernicus style guide?

L56: Would replace the comma with a period here, breaking into two sentences. (Or at least a semicolon.)

L57: Would be easier to parse if "used in the current work" were preceded and followed by a comma.

L70: Which version of ACOS?

L73-75: Please clarify about the creation of super-observations through aggregation. How it's written here, it sounds as if the data are aggregated across the whole orbit. Aggregation to super-observations is not the same as what is generally considered a L3 product, which is rather gap-filled concentration fields, perhaps through data assimilation to optimize atmospheric state (instead of fluxes). This is explained further in 2.1, but seems inconsistent with what is written here.

L101: What is SSDF? I believe this has not been defined.

L103: variables of interest?

L120: Perhaps specify that you mean ACOS version 9?

Caption Figure 2: remove "based"

L181: and: should maybe be changed to ", which" or "that"?

L224: parameters -> parameters'

L225: are -> is

L227: is -> are

L231: is -> was

L232: The wording here is a bit dense. Perhaps better: "which makes the MCMC algorithm more computationally efficient."

Caption Table 3: Not clear what "in the records of [the] fused CO2 experiment" means here. Perhaps just "for the fused CO2 experiment"? Also, "of 95% posterior credible intervals" -> "of the 95% credible intervals of the posterior". The same comment applies to the caption of Table 4 as well. In general, I was a bit confused by the use of "the records of fused CO2 experiment", also in the captions of Figures 4 and 5.

Figure 4: Units are missing. Somewhere in the caption it should be made clear that the "mean" is actually the mean difference from the prior for JJA over 2015-2016, rather than the mean flux itself. The label on the lower right panel should be "Interaction" rather than "Interact".

L270: Differences in behaviour, or differences in data coverage?

L283: Missing the word "on" (carried out on the inversions'…)?

L315-316: Really? The probability that the difference across years was larger than the aggregation method effect did not exceed 0.6. This does not convince me that using such a fused product provides internally consistent information for analyzing e.g. interannual variability in fluxes. Perhaps I have misunderstood something.

L339-340: Is the word "variance" (or something else) missing? i.e. "Alternative parameterizations with heterogeneous *variance* across space could be developed."

Supplement:

L32: model -> models

References

Cressie, N., Bertolacci, M., and Zammit-Mangion, A.: From Many to One: Consensus Inference in a MIP, Geophys. Res. Lett., 49, e2022GL098 277, https://doi.org/10.1029/2022GL098277, 2022.

Peiro, H., Crowell, S., Schuh, A., Baker, D. F., O'Dell, C., Jacobson, A. R., Chevallier, F., Liu, J., Eldering, A., Crisp, D., Deng, F., Weir, B., Basu, S., Johnson, M. S., Philip, S., and Baker, I.: Four years of global carbon cycle observed from OCO-2 version 9 and in situ data, and comparison to OCO-2 v7, Atmos. Chem. Phys., 22, 1097–1130, https://doi.org/10.5194/acp-22-1097-2022, 2022.