Summary and major concern

The manuscript describes a network of in situ measurements in and around NY state, and uses the network to attempt to constrain methane emissions in the state itself. The inversion methodology is extremely thorough. I find the end result of it all difficult to interpret. The network, in particular areas outside of NYC, is very information constrained. The authors are very thorough about their statistical interpretations, but statistical interpretations are contingent on the uncertainties assigned in the inversion itself. These uncertainties (and the following is a statement about inversions themselves, not a jab at this paper) aren't real uncertainties, but rather mostly arbitrary assignments, and thus I'm never sure how to interpret the confidence of the results using these metrics. My biggest concern in all of this is that a significant portion of these results may just be informed by the prior inventory itself, rather than any data. Because the authors only use one prior throughout the entire study, it's very difficult to tell how much of the end result is not us just looking at a reflection of that prior. In seeing the tower map, combined with the fact that no correlation length scale is used, my gut feeling is that Upstate NY isn't being constrained by the observations at all, and if that's the case, it needs to be made more obvious.

One of my favorite, dumbed-down ways of testing the impact of the prior is to take the existing prior, multiply it by a factor of 2 and 0.5, and rerun the inversion to see how much it returns to my original posterior result. Comparing the three posteriors provides a very straightforward metric on how much we should trust the result of individual grids. Obviously the kernel sensitivity maps do this to some extent, but adjusting the prior does this in a way that a human can actually interpret. I ask the authors to perform this simple test if possible. Make a x2 and a x0.5 prior for NYS (keep the inflow grids the same), run the inversion using your favorite ensemble member, show the two posterior side-by-side with the default inventory posterior, and demonstrate which grids show convergence towards a specific emission rate. That will go a long way to help the reader understand how much the observations are informing the result.

General comments

Figure 1 (Inventory creation in general): Be aware that the EPA Gridded Methane Inventory grossly underestimates emissions from unconventional gas in PA (like a 0.05% vs a 0.40% intensity rate), which could create some issues specifically for the ACT-America PSU tower in northeast PA.

Section 2.1 is in desperate need of an emissions map (similar to Figure 2) that shows a prior for all of the sources combined. It'd be very helpful to see where the emissions are concentrated in NYS and what sources are believed to matter. I see it shows up later in the results section, so maybe just point the reader to that figure somewhere in here if the journal allows it.

Figure 3: I'd love to see a version of this that only uses the afternoon hours prioritized in the inversion.

Figure 4: My first-guess looking at this map is that this network has no hope of solving for anything other than NYC and maybe the Hudson Valley. Upstate NY will barely register a signal, and your few towers in the upstate region are going to have to fill in 100 mile gaps of near nothingness. This is going to make the emissions very sensitive to any sources in the nearfield surrounding the towers, especially without any correlation lengths applied. It just doesn't seem possible that the tower network is going to be able to confidently assign emissions anywhere outside of its local grids.

Line 307: *"Initial attempts in linear space resulted in non-physical negative emissions in select regions."* I would be curious to know what the totals were using linear space and how negative certain grids got.

Table 3: It would be nice to know the mean bias not just in the posterior but in the prior as well to see how much of the bias is being corrected by the inversion itself. I didn't see that info in the supp either.

Line 429: *"Figure 8 shows that the bottom-up inventories we use as prior estimates are consistent with the top-down observations stationary observing system in the whole state for all years, within the uncertainties of the model-observation system."* Your prior inventory appears to have a 95% confidence interval between 600-1700. Saying the posterior is consistent because it falls within that huge range is basically a meaningless statement. It looks like every one of your posteriors consistently end up lower than the prior. Surely that's worth a mention, even if it doesn't reach some arbitrary 95% confidence threshold.

It would also be helpful to break this chart into Upstate vs SE NY, that way it's easier to disentangle the two trends from the full state analysis.

Line 461: "*We consider the statistically significant trends in posterior methane emissions offshore to be essentially meaningless, as their slopes are negligible (on the order of 10−7 Mg CH4 km−2 yr−2), reflecting their orders-of-magnitude lower emission estimates than on-shore emissions."* This sort of demonstrates that the way in which statistical significance is being calculated is flawed, despite being mathematically sound. In general this paper has a fixation on mathematical accuracy, which relies on assumptions that drive the inversion and its metrics which are not actually true or real (we do not actually know the error structures of R and B, as much as we try to pretend). Just a comment. No action required.

Figure 10: Your oil and gas posterior uncertainty is impressively small for a source that's 1/10ths the state's emissions concentrated in a region where you have no observational tower sites. Is it possible your emission estimates aren't actually real here and that they're just constrained by whatever was in the initial prior? If the oil and gas emissions started at 0, would the inversion have detected that and raised those grids accordingly? If the oil and gas emissions were triple, would the inversion have detected that and lowered those grids accordingly? If not, is that really a fair representation of the uncertainty?

Line 561: *“Both landfills are associated with relatively high sensitivity in the averaging kernel diagonal, up to 0.28 for Modern and up to 0.22 for Seneca Meadows, indicating that the stationary observations are sensitive to these major point sources despite their distance from the sources”* They’re sensitive to it because they are large point sources and thus, because of how your errors are structured as a percentage of the total grid value, are sensitive to the inversion itself. If you had put a +-100,000,000 Mg/hr error in that grid, your kernel sensitivity would be off the charts, but it wouldn’t have any more meaning. There isn’t actually any skill here. The tower network is not identifying with confidence the value of a 3 Mg/hr landfill 100+ km away from it.

Section 3.3: You discuss how reductions to the tower network impact the sensitivity kernel. How does it impact the actual emissions estimates? I think to get the sensitivity kernel you had to have run the inversion itself using the same configuration. Did eliminating those towers actually change the posterior emission results significantly? That would be the most obvious way to demonstrate that this loss in tower sites would matter.

Line 620: *“We identified statistically significant ($p < 0.05$) interannual emission trends in central NYS, where emissions are increasing, and in western Long Island outside New York City, where emissions are decreasing.”* It’s possible I missed this above, but is there some sort of physical reason you could point to why the central NYS emissions would be increasing with time? I’m assuming Ag emissions are the main culprit here, so potentially the Ag census shows cow populations increasing in central NYS compared to the rest of the state? Being able to connect observed changes with real world processes helps to make the results seem like more than just noise.