The authors have considerably extended the manuscript, compared to the initial submission. The methodological description is somewhat clearer, although it can still be improved. The results analysis and interpretation has also been improved. However, they have added a new OSSE experiment, which is terribly designed, and whose results are extremely poorly interpreted. I have some theoretical reservations on the adequacy of their inversion approach, now that I understand it better, but with a well designed experiment, they could help lift these. Unfortunately, their (single) OSSE is designed to “showcase” their setup, rather than help determine its strengths and weaknesses.

I think that, beyond the poor quality of at least some parts of the revision, the problem is that the scope of the paper is not well defined. They resolve a joint meteo-tracer-emission state, with two tracers, using an unpublished 6-hourly sequential data assimilation framework, and make quite a few bold claims about the advantages of their system. It’s a lot of things to prove in a single paper.

The manuscript hasn’t really gotten closer to publication-ready since the first revision, I doubt that the comments below can be addressed in just one or two extra rounds of revisions (even sections 2 and 7, which are now in a better shape, still need significant improvements). For this reason, I recommend a rejection, with an encouragement to resubmit a new manuscript, with a better focus and experimental design.

Main remarks

Inversion technique

As far as I could understand, during an assimilation cycle, the data assimilation system computes an analyzed state that comprise meteorological variables, the atmospheric tracer (CO2 and CH4) concentrations, and emissions which are in fact “surface forcing parameters for the subsequent forecast”. Only the meteo and concentration are observed, so the emission forcings are constrained indirectly, through their (ensemble-based) correlations with the concentrations. The inferred atmospheric state is then used to initialize the next 6-hourly forecast, wich runs with the emission forcings optimized in the previous analysis cycle.

If I understand correctly, this means that there is no causal relationship (in the model or in the real world) between the emissions and the observations that are used to optimize them. This approach is justified in an atmospheric composition forecasting application, where the aim is to obtain the most realistic forecast of the atmospheric composition, and adjusting the emissions during the forecast cycle *p* based on the analysis of cycle *p-1* is an efficient way to do that: if the emissions are overestimated at *p-1*, they will probably also be over-estimated at *p*, though it’s not guaranteed and so it may occasionally degrade the forecast.

I am far less convinced about the appropriateness of this approach for GHGs inverse modeling, where the focus is on the accuracy of the emission estimate, and not on that of the 4D concentration fields: it certainly works, to a degree, but is it really better than making use of the actual, modelable causal relationships between the observations and the emissions anterior to them? It implicitly relies on the fact that emission errors are correlated in time, which is often the case, but not always. These considerations are not stated in the paper.

Another issue with this approach is that the short assimilation window (6 hours) means that a given observation will never be informative of emissions further back in time. This may be justified with short lived tracers (e.g. NO2), but with stable tracers such as CO2 or CH4, emission signals are carried over long distances and time. It may be that the information is still passed, indirectly, through the propagation of the optimized concentrations, but this is not discussed.

OSSE experimental design

I don't think that the remarks above regarding the inversion method should be necessarily blocking. It may be that they don't matter much in practice, or just that I didn't understand the approach (if so, please clarify it). Regardless, since it is a new inversion setup, its performance should be tested, and that's the role of OSSEs.

So, in principle, adding an OSSEs section to the paper was a good idea, the problem is that the single one that was added is really poorly designed. They have created a known truth and a perturbed prior, by removing one entire emission sector, representing >40% of the anthropogenic emissions over Korea, so it's a massive signal, that will have a major impact on the fit to observations. They have then added 50 sites with synthetic observations, distributed downwind of urban areas … closely following the geographical extent of the emission perturbation, and comprehensively covering it (at the part inside South Korea). For reference, 50 sites is on the same order of magniture that the entire GHG monitoring networks in the US or in Europe.

Besides the ensemble perturbations, the only differences between the true and prior simulations is this emission perturbation. The boundary condition is the same, the transport model is the same, the C fluxes in other sectors (including natural) are the same. This can be a useful technical test, but it doesn't provide any meaningful insights on the performance of the DA, so I don't understand the rationale for presenting it.

The stated aim is to provide a "technical illustration of the framework behaviour under substantially expanded observational coverage". But it doesn't even do that: first, a reference experiment, without expanded network, would be necessary. Such an experiment, with much fewer sites, would likely be configured with longer localization lengths and eventually … lead to almost exactly the same result, since the perturbation is spatially homogeneous (at least in proportion of the prior). It is heterogeneous at the pixel scale (point sources), but that's not something that even their 50-sites network can resolve anyway.

Regardless, the objective of an OSSE is not to be a "show-case", but to test a refutable hypotheses, and, ideally, in this case, determine what are the limits of the system in terms of its capacity to resolve emission errors, at different spatial and temporal scales. Doing a sensitivity test with a few extra sites to test the impact of the network density can be a good idea … covering the domain of observations not so much.

OSSE results analysis

Even omitting the design flaws of the OSSE experiment presented, its results interpretation needs to be improved. The inversion recovers about half of the prior error, at the national level, and "*the posterior emission increment exhibits a distribution that is broadly consistent with the true difference* …": this goes in the right direction, but the aim is not just to "detect a signal", but use it to accurately quantify the missing emissions. Therefore the success criteria cannot just be that “broad consistency”, it needs to be quantified. There is no presentation or dicussion of the fit residuals to the observations. There is no presentation of the posterior emission error (since the truth is known, that would be a much more useful metric to present than the posterior uncertainty!).

If only about 50% of the missing emissions are recovered, does that mean that the observations are, somehow, not sensitive to the remaining 50%? (with such an extensive coverage)? Is it that adjusting the tracer fields directly offers degrees of freedom that allow fitting the observations without changing the emissions (that wouldn’t be good, in a twin experiment where only the emissions are perturbed!)? Or is it that the observations are improperly fitted? (again, not great ...)

The authors make a lot of claim about the supposed advantages of their system compared to “Lagrangian footprint-based inversions”, both in the introduction and in the discussion. But any not-too-poorly designed Lagrangian inversion should be able to recover the missing emissions here, with such an observational coverage. It may be that there are valid reasons why this is not the case here, but they should be discussed more than in vague terms.

Specific comments

The line numbers refer to the track-changed document.

L73: “Atmospheric inverse modeling frameworks for GHGs typicall employ either Lagrangian or Eulerian methodologies. Lagrangian-based … depend on external meteorological fields … Conversely, Eulerian frameworks solve advection-diffusion equations directly on fixed spatial grids … Notable examples of operational Eulerian systems include the NOAA CarbonTracker ...”: So as an example of a Eulerian model that “solve advection-diffusion equations directly on fixed spatial grids”, you chose a model that precisely doesn’t do that (TM5 is an offline transport model, it relies on ERA5 meteorological forcings … exactly the same way than would do a Lagrangian model).

L195: “The analysis directly updates the prognostic meteorological variables and 3D tracer concentration fields each cycle, while emission adjustments are estimated as surface forcing parameters for the subsequent forecast”: so, emissions at cycle *c+1* are influenced by the model-data mismatches at time *c*. Then, at time *c+1*, a new model-data comparison will be performed (with the tracer and meteo data at time *c+1*), and the concentrations will be adjusted: in a perfect world, where the only error comes from the emissions (which is the case in your OSSE), adjusting the emissions should be sufficient to bring the model in agreement with the observations … which is impossible in this inversion approach… Or am I missing something?

L199-207: I guess that what you mean by “correcting the pronostic tracer concentration state at each analysis time step” means just to initialize the next forecast with the last timestep of the previous analysis? (lots of jargon for saying something simple …). Note that the studies cited here concern NO2, which is a really different tracer (short-lived, highly reactive, very uncertain emissions), whereas CO2 and CH4 are passive and long-lived tracers, I’m not so sure if the justification for NO2 is equally valid for CO2 and CH4 …

Equation 1: To be clear, do I underestand correctly that x_met and x_chem are at analysis cycle *i*, whereas x_flux is at *i+1*?

L247: “H has no sensitivity to flux measurements” you mean it has no sensitivity to fluxes?

L250: “these flux adjustments are estimated as time-indexed forcing parameters rather than being propagated as a dynamic prognostic state”: Sentence is unclear. What do you mean by “time-indexed forcing parameters”?

L440: What is the interest of doing a single OSSE with such an unrealistic amount of sites? What can you possibly conclude on the real-world capacity of your system to estimate emissions (with the real existing network)? Especially since you don’t provide a reference experiment to compare it to …

L452: “the corresponding inversion (prior) configuration uses EDGAR … with the power-industry sector removed”: so, in other words, you impose a uniform(ish) bias to the anthropogenic emissions (in proportion of the prior).

L457: “By removing such a major component … detection problem”: That can be a useful test, but what does that tell you, in the end, of the capacity of your system to resolve real-world emission estimates? The real errors are both of much smaller magnitude (for the anthropogenic component), with more complex patterns, and entangled with errors in other emission categories.

L482: By applying such a short localization, you effectively prevent observations to constrain emissions a few hundreds of kms upwinds. With such a dense observation network, it might not be needed, but is this really what you want to do? CO2 and CH4 being stable tracers, long range transport should be modelled reliably enough. Or you don’t trust enough the long-range flux-concentration correlations in your ensemble, but then this is going to be a problem in real inversions, when you have only a handful of sites….

L483: “This configuration mimics a realistic operational capabilities assessment” No it doesn’t! Nothing is close to realistic in that experiment.

L485: “This ensures that the evaluation isolates the impact of expanded network density, rather than the effects of perfect parameter optimization”: how can you claim to isolate anything, without having a reference experiment to compare to?

L488-499: Your inversion infers flux adjustments where there are observations, and where there are prior errors (there are locations with prior errors and no observations, e.g. in China, but no locations with observations and no prior error). That is certainly a good start, but that’s not a proper evaluation of your results: Figure 4 shows the location of prior errors (4a) and analysis increments (4b), but doesn’t show the posterior errors, although based on 4a and 4b, I imagine that you get a degradation of the fit to the true emissions almost everywhere, with even some emission increments over the sea (East of Seoul), where there should not be any emission or emission uncertainty. How do you define, and test, the expectations from such an inversion? With such a dense network, it should be quite high, and needs to be tested through a much more complex prior error structure.

L502-504: So the recovery is only about 50% of the error: What recovery rate should be expected to consider the experiment successful? Are the posterior emissions consistent with the observations (once propagated with the transport model, i.e. in a copy of your “nature” run, but with the posterior emissions)?

L505: “These recovery rates should be interpreted as a conservative lower bound”: How so? You put the model in an un-realistically easy situation, you have a complete observational coverage of the prior error, you have no model error, no boundary condition error, a huge, easily detectable prior error. This is the polar opposite of “conservative”!!!

L506-508: “The short duration of the total assimilation period (14 days) relative to typical flux convergence timescales”: what is this “flux convergence timescale”? You have a comprehensive hourly observational coverage of your emission error, that’s more than sufficient to constrain the emissions over a two weeks periods. If you system cannot do this, it’s a design flaw.

L507-508: “and the use of an operationally representative, conservative filter configuration …”: but in a operational context, you would use these error configurations, and you would have far far fewer observations, so you would have a much much worse error recovery rate.

L508-510: “by avoiding idealized tuning …”: You already have an idealized perturbation, with an idealized network and an idealized model. So maybe you should in fact start with an idealized tuning, as a consistency check to verify that, at least in an idealized case, you recover most of the error. And from there on move towards more complex cases to try identifying the limits of your setup. Your current setup is all but cautious!

L511-512: I guess we are just supposed to believe your word regarding these expectations?

L513-515: “The imposed power-industry signal is highly heterogeneous”: is it, really? It’s a unidirectional bias, likely quite uniform in proportion to the overall anthropogenic emissions signal, and to where your observation sites are (conveniently) located. We indeed cannot expect the inversions to recover the pixel-by-pixel emissions, but the expectation should be much higher in such an idealized case than in a more realistic scenario, such as those you discuss in Section 7, and you never even define what these expectations should be.

L527-535: These whole two paragraphs are complete nonsense, unsubstantiated blabber.

Figure 4: The whole interest of doing an OSSE is that you can use the truth to validate your results. However, a comparison of the analysis to the truth is precisely what you are not showing here. You are also not showing the prior uncertainty structure, so it’s impossible to tell how consistent are these analysis increments with the prior uncertainty. I find it quite worrying to see significant emission increase over the sea (where, if I trust Figure 6 of the original manuscript, there shouldn’t be any emissions).

L551: “With the controlled proof-of-concept established in the previous section …”: but yet, you use a completely different observation network, different localization settings, etc., so what does the setup describe in Section 6 tell you about the results presented in Section 7? (not much …).

Section 7.1: Here I guess what you call “posterior concentrations” are the concentrations after the analysis step. However, you should also isolate the impact of the emission adjustment. You can do that via a new “nature” run driven by the inversion-optimized emissions, and with everything else kept the same as the prior. In a setup where the DA can adjust the concentration fields directly, I think this is critical.

L571: OH plays absolutely no role at these timescales (the residence time of air over your regional domain being a few days, at most). An incorrect representation of OH in the global model that serves as BC may be one of the drivers for latitudinal bias, but this is just a BC bias among others.

L582-583: yet, it’s not as pronounced in the observations, and in your “posterior” (whatever definition of it is used) concentrations. So your inversion corrects it somehow, and I would like to know how much of it is obtained by adjusting the emissions, and how much is obtained by adjusting the boundary condition (however that is done …). Here the extra diagnostic requested in my comment on Section 7.1 would be necessary.

Figure 6: where is the prior? And again, would be good to isolate the impact of the emission correction (since this is your key target).

Figure 9: Why showing only the posterior flux, which is still largely dominated by emission patterns present in the prior? Why not the difference with the prior?

L666-679: “Figure 9 … Figs 510-113”: all of that is not really an outcome of your inversion, it’s just features of your (forward) modelling setup.

L679-688: “However, the structural differences ...sub-national patterns”: If this is an inverse modeling setup, as in, you are trying to resolve the emissions, representing the tracer concentration at a finer resolution is not necessarily a benefit (it may help, depending on the circumstances, but it shouldn’t be an aim in and of itself. This is a very long paragraph to just explain that using a high resolution transport model leads to finer representation of concentration patterns …

L688-693: “These features … heterogeneity”: yes, coarser resolution model will smooth out fine scale features. But the question is whether it matters or not in your case. For assimilating TROPOMI data, which have a super high resolution, it certainly does (that’s probably what the authors of the studies you cited meant). For your case, with just 3 observation sites … maybe it still matters if you can show that the representation error becomes to large to efficiently assimilate these data (but you don’t step in that direction).

L693-696: “These results illustrate …”: This sentence is completely unsubstantiated. All you have shown so far in this section is that your model is capable of simulating fine scale atmospheric gradients in CO_2 and CH_4. But you haven’t shown whether this leads to higher quality emission increments than what you would obtain with a coarser resolution inversion (you haven’t actually shown emission increments at all, at this stage …).

L778-782: “By coupling meteorology and tracer transport … terrain”: This statement reflects seems to reflect the thought process behind the design choices of this inversion framework, but I don’t think it’s supported by facts. At least it’s not demonstrated in the paper that updating both the meteorology and transport provides any meaningful advantage compared to simply using prescribed meteorological data from a reanalysis (or from a reanalysis ensemble). Perhaps there are practical reasons behind this choice, and it probably doesn’t really hurt, although I can imagine it somewhat limits the number of degrees of freedom of the assimilation, since it has to fit jointly meteo and tracer concentrations. Likewise, the choice of a dual-tracer inversion (CO_2 + CH_4) is not really justified.

L792-803: This entire paragraph is a long list of theoretical qualities that this inversion setup would have and that would constitute advantages over “footprint-based Lagrangian inversions”, but if it’s not supported by any facts and demonstrations … it’s just pedantic. Especially since it denotes a certain lack of understanding of the GHG inverse modelling field in general (which is quite diverse and not limited to using Lagrangian models, which themselves can be employed in a wide range of manners). If you think that solving jointly meteo and tracer concentrations has an advantage over using prescribed meteorology, prove it, you have the right tool for it! If the DART inflation framework provides a quantifiable advantage over some other mechanism that would not be “directly coupled to sequential filter dynamics”, please set up an experiment to demonstrate it, otherwise this sentence has nothing to do in the conclusions.