Dear Referee and Editor,

We are thankful for these helpful comments, that helped to improve the quality of the paper and the future implementation of this system. We hope that the main point about the retrieval biases is now addressed correctly as we do now clearly acknowledge the current limitation of the method. See our response to the comments in italic font below.

Referee #1 comments:

I appreciate the further analyses as well as the additional global view in Figure 9 and also understand that it is hardly possible to carry out a complete and comprehensive analysis of the capabilities and short-comings of the method within the framework of this paper. Therefore, I think that a publication in AMT would be more appropriate to present an introduction of a new interesting technical method but I do not insist on moving the article if the editor has a different opinion.

However, without a more detailed evaluation of the global capabilities some of the statements and conclusions have to be weakened (see below). I would also propose to emphasise more clearly what is actually done already in the title and/or in the abstract: local XCH4 anomalies are detected (by combining satellite measurements and forecasts), which are due either to actual emission anomalies relative to the CAMS emissions or systematic biases in the satellite data or a combination of both.

We have modified the abstract and the title accordingly

The distinction between the two cases has to be made by (subjective) interpretation and is prone to error. I do not think that the decision should be made by solely assessing the persistence of features as suggested in the manuscript. Persistent emissions can also lead to persistent features as there is not always a clear plume structure in satellite GHG data, in particular for non-point sources, when topography causes accumulation, and when using a 30-day time window (see also specific comments). How persistent (in space and time) is a pattern allowed to be to be considered real? You need additional information, e.g. to check if the features are correlated with albedo features (like you do for Figure 13), to classify patterns as retrieval biases. In principle, you also have to check the albedo features in the cases where anomalies are expected (e.g. Permian and Turkmenistan in Figures 10 and 11) to avoid expectation bias. On the other hand, emission patterns could indeed be correlated with albedo (e.g. wetlands or facilities vs. surroundings) complicating the interpretation. Please elaborate more on these issues and discuss them in a more balanced way.

For example, in case of the new example shown in Figure 14, where there is no correlation of the outlier pattern with albedo, I would be cautious to classify this feature as retrieval bias unless another good explanation for a potential retrieval bias has been found.

We agree with the reviewer about the limitations of the method and we have clearly stated them in the abstract. Section 4.3 and the conclusion. We also now have substantially modified and complemented section 4.3 and the associated figures to provide clearer and more accurate discussion about the identification of the retrieval errors.

Specific Comments

Page 2, Lines 58-59: The sentence is a little misleading because the cited papers analyse different regions, but all include the Permian basin. Therefore, I suggest to change it to something like: "... large and extended enhancements in different US oil and gas production regions such as the Permian basin."

The sentence has been adjusted as suggested.

Page 3, Lines 86-87: What is the difference between instrument precision and random error?

We have clarified the statement

Page 4, Lines 117-121: Your data assimilation technique only corrects the concentrations and not the emissions.

But this is exactly the potential problem, isn't it? As a consequence, H(x) (in ppb) potentially depends on patterns observed by IASI and TANSO. Or am I getting something wrong here? Assume there is a (unknown) source, which is observed by TANSO and TROPOMI. Then the concentrations in the forecast are corrected upwards due to TANSO and the difference d to TROPOMI (which also sees an enhancement) is getting smaller because of the assimilation. The other way round, isn't it possible that a potential bias in the IASI or TANSO data, which is assimilated, causes an artificial outlier of your method although emission data bases are actually consistent with the TROPOMI measurements? Along these lines, wouldn't it be better to use a model without assimilation of satellite data as starting point if you want to assess the quality of emission data bases?

The assimilation is having an expected minimal impact for correcting close to sources concentrations when the high-resolution departures are computed for two main reasons:

1. *The assimilation impact close to the surface is weak due to the poor coverage of TANSO and the low sensitivity of IASI close to the surface. See Massart et al., 2014 for details.*
2. *As described in section 2.2 and figure the assimilation is acting at lower resolution (around 25km) at least 4 days before the high-resolution departures are computed.*

This is the way our operational 9km high-resolution ECMWF forecasts are initialized operationally, and we take advantage of this to perform this demonstration of monitoring in a cheap and efficient way. We agree that running another high-resolution suite of runs free of CH4 data assimilation would make things strictly cleaner but more costly for expected very minimal differences close to the surface.

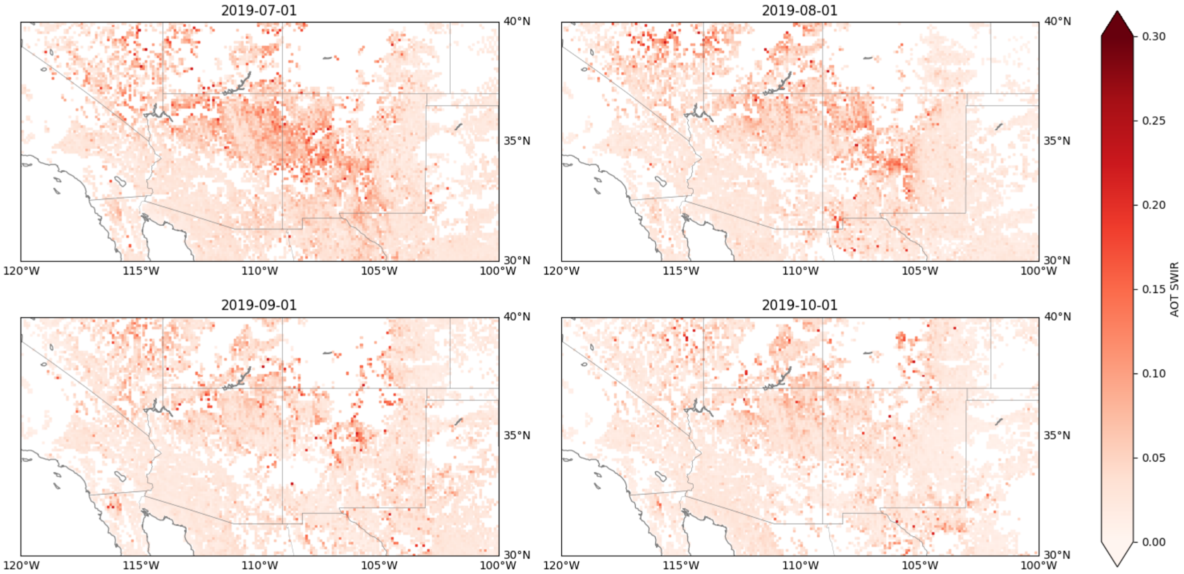
We have now added clarifications at the end of the paragraph.

Page 6, Lines 166-168: Is the averaging kernel function as a function of pressure really discussed in the cited paper? Moreover, the paper analyses a different algorithm than the one used here. Please cite a paper describing the averaging kernels of the operational TROPOMI algorithm if possible.

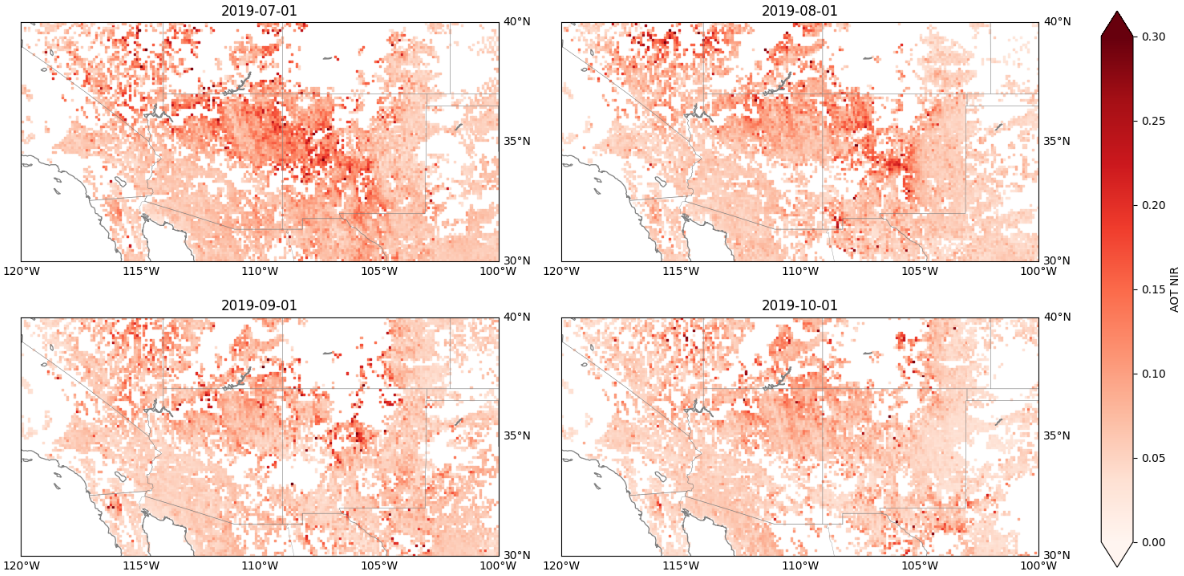
We now put the correct reference (using Hu et al., 2016).

Page 9, Lines 265-267: Please check if there is actually a correlation with surface albedo features (as in the case of Figure 13).

We didn't find clear evidence of albedo features corresponding to the enhancements. We also do not identify facilities that could be the reason of those enhancements either. For the northern Baja California we have identified that the feature could be correlated with scattering parameters see figures below. We have then changed the statement accordingly.



Aerosol optical thickness (AOT) in the short-wave infrared band (SWIR) values provided with the TROPOMI CH4 retrievals. Maps provides averages corresponding to the windows used in figure 10.



Aerosol optical thickness (AOT) in the near infrared band (NIR) values provided with the TROPOMI CH4 retrievals. Maps provides averages corresponding to the windows used in figure 10.

Page 10, Lines 301-303: This statement is too strong. Please write e.g. "potential retrieval error artefacts". Persistence isn't everything because plumes are not always visible in daily GHG data and may disappear when using multi-day time windows if the wind direction changes. As a consequence, it is possible that you only get an anomaly right above the source with your method (see also general comments). Please revise this section accordingly.

We have clarified the statement further. We however still want to emphasize that such a very similar shape seen an extended period of time as in Fig. 13 is extremely unlikely to be an emission signature. Section 4.3 has been re-written, figures amended with additional plots displaying albedo and scattering features retrieved from TROPOMI to complement the discussion.

Page 10, Lines 305-307: I would be cautious to classify this feature as retrieval bias when there is no correlation with albedo features. Are there other potential explanations? (Other features causing biases? Could it be a real signal?)

Section 4.3 has been re-written, figures amended with additional plots displaying albedo and scattering features retrieved from TROPOMI to make the point clearer. The match for the Siberian feature is quite striking.

Page 11, Lines 317-318: Please add a sentence that the distinction between over-/under-reported sources and local retrieval errors is challenging and needs correlation analyses with external data sets such as albedo.

We have added the required sentence.

Figures 9-14: The colours of the four categories are sometimes hard to distinguish in the maps (in particular with the updated colours in the revised version). Please consider to use different colours or to code the classes additionally in a different way (for example by different symbol shapes or hatching).

The colours have been changed already as request by reviewer #2 in the first round of review where gold colours have been made darker. We are not sure to what more differentiable colour between red,green,blue and yellow (gold) we could possibly make the scatter plot with. We are also not sure that different shapes or hatching will make things clearer as the number of data points are high. We then increase the size of the dots and make the colours less faint in the scatter plots to improve the readability.

Editor comments:

1) I'm not a modeller which is probably why the discussion of the monitoring suite and the departure is confusing to me:

a. What do you mean by "trajectory" – for me, that is a Lagrangian calculation of the track of an air mass but here it seems to have another meaning

Trajectory in the variational data assimilation sense is a model forecast within the data assimilation window (here 12 hours) in order to compute the observation departures. Which is different from the model forecasts as this is a model run of few days initialized from the data assimilation analyses. We clarified the text accordingly.

b. What is the difference between what is done in the monitoring suite and what I would have naively done, namely comparing the model forecast for the time of measurement with the measurement after applying the measurement operator to the model profile?

There is no difference. We use part of the data assimilation system to perform this monitoring at the exact model time step.

c. What is the time step of the monitoring suite?

It is 450 seconds; we now specify this in the text.

2) I do not understand your outlier classification:
a. What do you mean by "positive filtered observations"? Do you mean filtered observations, which are positive? If so, then the description of the green and golden classes does not match what I see in Figure 9.

We have clarified the description of the classes.

b. What is the difference between first-guess and forecast? If there is none, please just use "forecast"

We have now changed first-guess to forecast

c. Why is the green category representative of "over-reported or under-reported plumes" as stated in line 256?

We have corrected the sentence.

d. I have trouble seeing the benefit of separating into four instead of just two groups. It would be good to provide for each of these categories an example of an application to illustrate its usefulness.

Separating into only two categories, I assume positive and negative departures only, will not allow to disentangle when anomalies are due to observations or due to the forecasts.

We also now provide example using the low observation category (blue), illustrating its usefulness. For the low forecast category (gold) the number of occurrences is very low and very sparse and probably not significant in our current monitoring dataset we then decide not to illustrate it in our paper. This category could be however important to identify low model/inventory biases if they had to occur.

3) As also pointed out by the reviewer, the use of this method to evaluate emission inventories is undermined by the assimilation of other satellite data. Comparison to a control run would be a cleaner way of achieving this goal.

Please see the response to the reviewer's comment.

4) I also agree with the reviewer that this type of comparison is useful to identify retrieval artefacts but it is not straight forward as it is difficult to disentangle differences coming from deficiencies in the emission inventories and modelling set-up from retrieval problems.

Please see the response to the corresponding reviewer comments. We have now detailed the text accordingly to take into account this important point.