We thank the reviewer for the kind words and for the good suggestions to present the results in the manuscript more convincingly.

The main concern is by the reviewer named “lack of convergence” in our solution that includes a twenty-year timeseries of global OH variations. We agree with the reviewer that the degree of convergence is an issue of concern, which is why we treat it separately in Section 3.5. Our study has been constrained mostly by pragmatic concerns. To give an indication of the computation time involved: to reach the degree of conversion presented in the manuscript, around thirty wall-clock days were needed. The set-up of the final inversion was inspired by test inversions and, likely, more improvements in this set-up are possible. We expect, however, that new challenges will emerge if we implement these improvements. We decided to stop this cycle and share our findings with the research community. With these inversions we believe that we have reached an acceptable endpoint, which we will argue further here.

In the statistical framework that we have defined, the solution has indeed not fully converged: a smaller cost function would be found, if we would continue to iterate. We have stopped the inversion for two main reasons:

1) **The standard REF solution presented in the manuscript reproduces NOAA surface observations to a satisfying degree.** Although we do not reproduce each individual observation within the error we have prescribed, we suspect that the pre-defined observational + modeling error might have been too small. We discuss this issue in Lines 246-249 of the manuscript. It is difficult to formulate a correct modeling error a priori, because MCF gradients are small and sub-grid variations can be important. That it is difficult to define such an error can be seen in Bousquet et al. (2005), where the modeling error is actually defined in hindsight, to reach a chi-squared of 1. We cannot use the same approach due to computational expenses in our set-up, but we extensively investigate what causes the residual errors and we consider the residuals small enough that we can call our REF solution a satisfactory solution. To put it differently: given the fact that we might have underestimated the model error on observations, iterating further towards a fully converged solution could be considered overfitting.
   In hindsight, our recommendation would be to not assimilate individual observations, but instead assimilate monthly means, because with relatively homogeneous OH and emission fields we cannot reproduce individual observations anyway (lines 235-236). However, since we already reproduce monthly means at most sites, we do not consider implementation of this recommendation a prerequisite for publication. We have included this recommendation in the new version of the manuscript (lines 250-251).

2) **Further iterations result in a lower cost function, but not necessarily in a better solution.** The residual cost function appears to be dominated by intrahemispheric biases and by short-term variations. As discussed above and in lines 243-244, there are simply not enough degrees of freedom in the inverse framework to improve the match with short-term variations.
   The intrahemispheric bias is a different concern. Already, adjustments in the latitudinal OH distribution in the 20-year REF inversion are up to 30%, which corresponds to 3-sigma. In the better-converged 10-year inversion, we find adjustments of 60%, or 6-sigma. As discussed in the manuscript, we consider it unlikely that these adjustments in OH provide the best solution to explain the intrahemispheric biases, but it is the only explanation our inversion can provide us with, in this set-up.
   The reviewer accurately notes that we did not provide convincing evidence that such adjustments are physically unrealistic. We have modified Figure S8, that now compares literature estimates of OH distributions to our prior and posterior OH distributions. It can be observed that, in the ten-year inversions, which came closest to reproducing the intrahemispheric gradients, the ratio between tropical and extra-tropical OH ends up much higher than in estimates from a range of chemistry models. Although we cannot exclude the possibility that tropical OH is too low in the prior distribution, we derive adjustments far outside the (arguable uncertain) prior error settings. One of our noteworthy hypotheses is therefore that a scenario that includes a high-latitude ocean source of MCF is more likely than one that does not.

In summary, we agree with the reviewer that stopping an inversion half-way is suboptimal and statistically inconsistent. However, we emphasize in the manuscript that the error we put on observations might be too small and that our inversion reproduces NOAA surface observations well at most sites. Additionally, where the agreement is poor, we do not think that more iterations will help.

Further general comments:

1. The importance of using ocean fluxes that account for absorption and reemission, compared to a 1st order loss has been well known for almost 20 years, at least for the overall MCF trend, and particularly during the period where emissions were changing rapidly. This article presents a nice demonstration of the influence of different ocean flux parameterizations on the meridional gradient. However, given that it is well established, I'm puzzled as to why the more realistic ocean fluxes weren't used in the main inversions?

We do not agree that the ocean flux described in Wennberg et al. (2004) would a priori have been a better choice than the simple first-order loss we have used. In Supplement S4, we outline the observational evidence for the low hydrolysis rates at cold temperatures. These low hydrolysis rates are key to the hypothesis of an ocean source, but the evidence is thin: it is based on extrapolation of hydrolysis rates measured above 25°C, and the only study performed at 10°C found higher-than-expected hydrolysis rates. We are not aware of any experimental follow-up studies. We have included a recommendation in the supplement (lines 122-124).
In Rigby et al. (2017) it is argued that inversions of MCF that do not adopt an ocean source derive spurious OH variations particularly around 1998. However, we would argue that during this period uncertainty in emission timing (particularly delayed emissions) could equally well explain the derived variations. Large changes in MCF emissions coincide with the hypothesized onset of an ocean source, which is why we have found it impossible to find evidence for a switch in sign of the ocean flux in the surface network observations during this time. On that note, in Naus et al. (2019), we have performed a two-box inversion of MCF, covering 1994-2014, that included a first-order ocean sink and we were able to reproduce hemispheric averages of MCF without large OH variations.
In the absence of strong evidence for (or against) an ocean source, we choose to adopt the simplest assumption: first-order ocean loss. Having performed our 3D model inversions, we think that we have found convincing evidence in the latitudinal gradients of the surface networks for an ocean source of MCF. Given that we do not consider an ocean source well-established, we present this as one of the main findings of our study.

2. A main conclusion of the paper is that the variation in oxidation magnitude is small (< 3% per year). This does indeed seem to be the case from the point of view of the standard deviation in the solution. However, some year-to-year changes in fact seem to be very large. For example, sometime around 2010 – 2012, the REF inversion shows a change from _-5% to _+5% compared to the prior (Figure 1). Wouldn't a change in tropospheric oxidation of 10% over 2 years actually be considered quite substantial, and have major impacts on, for example, the global methane budget?

For comparison with e.g. Montzka et al. (2011), it would be more appropriate to consider the annual mean k.OH anomalies presented in Figure 8. The change in k.OH in 2011, 2012 and 2013 are +5%, +3% and -4% respectively. The reviewer correctly notes that these variations are large with respect to the standard deviation of the interannual variations (2.4%) and it is striking that we find large variations in subsequent years. However, whether this is statistically unexpected is difficult to say: one 2-sigma deviation is expected in a twenty-year timeseries and the distribution of interannual variations over the 20-year time period is not very different from a normal distribution. Also based on Reviewer 1's comments, we have placed more emphasis on the large 2012 anomaly, but we still consider that overall interannual variability is small.

Furthermore, we do think that such OH variations will have a significant impact when applied to the methane budget. This is an aspect we did not investigate thoroughly, because methane was not included in our simulations. As was shown in Rigby et al. (2017) and Turner et al. (2017), interannual variations in OH of a few percent can significantly affect the most likely interpretation of the methane budget, especially when the interannual variations stack up to larger multi-annual variations. This why we recommend inclusion of our derived OH variations in a methane inversion, even if our timeseries of OH is still highly uncertain. We now provide the methane lifetime in Supplement S1, but the effect of lifetime variations on the CH_4 budget requires further research.

An additional point: converged solutions in Figure S6 seem to show, in general, more variation than the unconverged "main" results. So, again, it would be important to investigate more fully how sensitive this main conclusion is to the lack of convergence in the main results

The difference between the blue solid line (REF inversion, 20y) and blue dashed line (REF inversion, 10y) in Fig. S6 we consider to be within the error margin of our posterior OH estimate. We draw confidence in the 20-year inversion from the high degree of consistency between the timing of OH anomalies from these two inverse results that have reached different degrees of convergence (excepting a spin-down period).

IAV in posterior k.OH anomalies over 1998-2008 is 2.0% and 2.9% for the 20 and 10 year inversions, respectively. Which of these is better depends on how well we want to fit observations, which brings us back to the difficulty of defining a model error. While the converged state corresponding to our inverse framework will likely have an IAV in OH similar to (or even larger than) that of the ten-year inversion, we can already reproduce atmospheric observations within reasonable bounds with the IAV of the twenty-year REF inversion (e.g. Fig. 6). We have added a paragraph to the discussion on convergence that addresses this issue (Lines 378-386). We still consider the estimate of IAV in OH of <3%, i.e. the number in our abstract, to be consistent with our results, even when we account for the increased amplitude in the ten-year inversions.

3. If emissions are being derived in the main inversions, why was it necessary to "preoptimize" the emissions, assuming constant loss? What happens if you don't do this? If this changes the result substantially, I'd be very concerned, as you're essentially using the observations twice, and, in the first step, you're fixing one of the parameters that you are trying to infer in the second pass. If it doesn't change the results substantially, then wouldn't this step be unnecessary?

We agree that it is not completely statistically sound to pre-optimize emissions with global mean MCF mole fractions, but it was a step necessary for the inversion to converge in a reasonable number of iterations.

Firstly, there is no reliable emission inventory available, especially in the second decade of our inversion, so some arbitrary choice for emissions is necessary. The inverse framework includes relative errors on emissions, because in an inversion with absolute uncertainties we cannot exclude negative emissions, which is problematic when a loss process needs to be separated from emissions. As noted in the reply to Reviewer 1, the relative emission error actually increases from 50% in 2005 to 200% in 2015 linearly, to allow for a wider range of posterior emissions when prior emissions are low (this is now mentioned in lines 120-123 of the manuscript). However, due to the relative error, we still need to define some emissions in later years to allow for a posterior scenario with substantial emissions. We performed a test inversion where we floored emissions at 5 Gg/year: a rather substantial amount. We considered this most fair because it does not include any prior information and allows for many posterior scenarios. In the prior simulation, MCF mole fractions after 2010 (predictably) ended up much too high. The inverse framework compensated with large adjustments in both emissions and OH, because the overestimate was present at all surface sites. We would hope that, eventually, when global mean MCF is captured, the inversion will start matching the small site-to-site gradients and determine that the overestimate is largely driven by too-high emissions.
However, after many iterations we had still not reached that stage, because the large overestimate relative to the small observational error resulted in a very large prior cost that gave the inversion problems. Therefore, we choose to be pragmatic and loosely fit prior emissions to the global MCF growth rate, so that the prior simulation does not drift too far from observations in later years. The prior fit is not even very good (e.g. dashed lines in Fig 4): a "problem" with the pre-optimization that we choose not to correct, because it tests the ability of our inverse system to optimize the global growth rate when the prior simulation is not too poor.

Secondly, we use the global mean mole fractions of MCF now both in the pre-optimization and in the 3D model inversion. However, we argue that the 3D model inversion is driven by additional information contained in the observations: intrahemispheric gradients, vertical gradients (e.g. MLO / KUM), realistic transport, etc. We performed the inversion in the 3D model precisely because we are interested in this latter source of information. Global and hemispheric mean mole fractions of MCF have been explored extensively already in previous box model studies. We considered the double use of global mean mole fractions a price worth paying for easy access to the more interesting information contained in subtle MCF gradients.

Minor comments:

L7: ". . . better reproduce. . ." (than what?)

Adjusted: ".. compared to the prior simulations .." (L7)

L18: "the signals are small". This is too vague. Need to specify what quantities are being referred to.
L19: ". . . better match the global MCF observations. . ." (than what?)

Rephrased to: *While the effect of the derived temporal OH variations on MCF mole fractions is small, these variations do result in an improved match with MCF observations relative to an interannually repeating prior for OH. Therefore, we consider the derived variations relevant for studying the budget of e.g. CH_4.* (L18-20)

L33 – 35: citation needed.

Citations have been added.

L48: What does the phrase: "but artifact-free sampling has become more difficult" mean?

We wanted to convey that declining MCF mole fractions have led to some issues with contamination during sampling in the NOAA network observations, even if the repeatability of flask pair measurements is not affected. Most notably, for this reason, there are no MCF observations available for SPO during 2015-2016. We have adjusted the text to clarify:

Through improvements in measurement techniques, measurement quality has mostly kept pace with the atmospheric decline of MCF. However, artifact-free sampling has become more difficult, because small contamination issues that might have been insignificant years ago become substantially more important as the MCF mole fraction has declined. L49-51

L91 and throughout: phrases like "such as OH" need to specify that it is OH concentration that is being referred to. Perhaps define "[OH]".

We have made adjustments throughout to distinguish between OH, OH concentrations and global mean OH concentrations ($[OH]_{GM}$)

L155: Can you be more specific than saying that model error was "proportional to the 3D spatial gradients". I.e. define what you mean by spatial gradients and let us know if there was a constant of proportionality.

We have changed the phrasing to be more specific and refer to the study that introduced this error set-up for TM5:

On top of the measurement error, we also included a model representativeness error for each observation. This error is calculated as an absolute average over the mole fraction gradients between the model grid cell that contains an observation and horizontally and vertically adjacent grid cells (Bergamaschi et al., 2005).
L 160-163

From this phrasing we think it is clear that we use no constant of proportionality.

L180: ". . . correlation with the REF inversion" (need to state what the correlation is with respect to)

Adjusted.

L195: I'm not sure what you mean by "would be hard to exclude from a bottom-up perspective". Can you be more explicit?

We want to convey that the absolute magnitude of emissions and emission variations that we derive for these years are very small. While an emission increase from 2012 towards 2013 (derived in the REF inversion) is not expected, the increase is so small that we cannot exclude it based on prior knowledge of emissions during these years. We have adjusted the text:

Firstly, we note that the small emission totals in later years of around 2 Gg/yr, with interannual variations of 0.2 Gg/yr, would be hard to exclude based on prior knowledge of emissions. L282-283

Figure 4: There seems to be a consistent increase in the mismatch in the IH gradient. Any idea what causes this? Seems like a potentially interesting feature. Could this point to a sudden increase in emissions?

An increasing relative contribution of emissions to the MCF budget is a potential explanation for this change. We also derive increasing to near-constant MCF emissions after 2013 in the REF inversion (Fig. 3). However, it is difficult to interpret these small differences in IH gradients for MCF. MCF abundance is lowest in the tropics and so the IH gradient, especially when low-latitude sites are included, might not be the most insightful quantity. For example, we quite well capture the Alert to Cape Grim gradient, which could indicate posterior emissions are realistic (Fig. 5). On the other hand, the observed MCF gradient between Alert and Mauna Loa also increases in recent years, relative to global mean MCF, which our simulations do not capture (Fig. 5). This increase is not seen in the Southern Hemispheric Samoa to Cape Grim gradient, which makes an explanation involving MCF emissions more likely.

However, we consider that correct interpretation of changes in the spatial gradients of MCF is too complicated to do without a model. For example, why would a change in emissions in recent years drive an increase in the ALT to MLO gradient, but not in the ALT to CGO gradient? Therefore, we choose not to hypothesize much on the potential drivers of posterior residuals in the manuscript. We only intended to show with Fig. 4 that the inversion performs well also on those quantities (global and hemispheric averages) that were used in previous box model studies.

Figure 5 and throughout: For the reader not familiar with these site codes, it would be useful to clarify where these stations are (i.e. their latitude is particularly important for the discussion of Figure 5).

For the interested reader Table S2 is available. We now also include latitudes next to the site abbreviation in Figure 5 and corresponding supplemental figures.

L374-375: It would seem important to include the results of this test (scaling global OH) in the supplement.

We would agree if the only difference between the inversion with global scaling of OH and the presented inversions was the degrees of freedom we gave to OH. However, we made other changes too after this test inversion (e.g. to the prior emissions) and so a fair presentation of these results would require an extensive and complex explanation. We are afraid that the storyline of the paper would become even more convoluted.

L379: "These results indicate substantial robustness of the derived OH variations". I'm wondering if this can really be stated so strongly, given that the three main inversion show different OH variations?

We have attenuated the statement somewhat to:

These results indicate that the solution we have derived is robust and consistent with observed gradients in MCF. L396-397

We do think that while the three twenty-year inversions show different OH variations, all other tests indicate robustness, at least for the timing and approximate amplitude of the derived OH variations.

References

1. Wennberg, Paul O., et al. "Recent changes in the air-sea gas exchange of methyl chloroform." *Geophysical research letters* 31.16 (2004).
2. Montzka, Stephen A., et al. "Small interannual variability of global atmospheric hydroxyl." *Science* 331.6013 (2011): 67-69.
3. Turner, Alexander J., et al. "Ambiguity in the causes for decadal trends in atmospheric methane and hydroxyl." *Proceedings of the National Academy of Sciences* 114.21 (2017): 5367-5372.
4. Rigby, Matthew, et al. "Role of atmospheric oxidation in recent methane growth." *Proceedings of the National Academy of Sciences* 114.21 (2017): 5373-5377.
5. Naus, Stijn, et al. "Constraints and biases in a tropospheric two-box model of OH." *Atmospheric Chemistry and Physics* 19.1 (2019): 407-424.