Response to Reviewer Comments:

We thank Dr. Marc Bocquet and an Anonymous Reviewer for their thorough comments, we think they have greatly improved the content of the manuscript.

Reviewer #1 (Dr. Marc Bocquet) Comments:

The authors overlooked the findings that have been reported by Bocquet et al. (2011); Wu et al. (2011). From the results of Bocquet et al. (2011); Wu et al. (2011), I believe that the optimal resolution as seen by the authors is the results of suboptimal choices. What is reported in the present manuscript is nevertheless interesting since those suboptimal choices could be made for the sake of numerical efficiency. It is problematic that the authors are (unintentionally) hiding what actually leads to the appearance of a minimum in the total error curve as a function of resolution.

We have expanded the discussion of Bocquet et al. (2011), Bocquet and Wu (2011), and Wu et al. (2011). (see response to Dr. Bocquet's second comment below).

1.) Frankly, the notations are unfriendly. I understand the authors follow those of Rodgers (2000). Yet, they diverge a lot from standard data assimilation or inverse modelling notations that have been widely adopted in atmospheric chemistry data assimilation. For instance "a" usually refers to the analysis while the authors use it to refer to the prior, when "b" ("f" in a sequential context) is very often chosen. The gain is usually designated as **K**, not **G**; "**H**" is much preferred to "**K**" for the observation/Jacobian/source-receptor operator. That said, the choice of notations belongs to the authors. But, I guess that the present notations would significantly distract potential readers.

We draw on notation from the inverse modelling and trace gas retrieval communities. This notation is standard in the retrieval communities (e.g., Rodgers 2000) where most of the smoothing error discussion has previously taken place (e.g., von Clarmann 2014). This is also the notation that has been used by the Jacob group for the better part of a decade (e.g., Jacob et al. 2002; Jones et al. 2003; Heald et al. 2004; Palmer et al. 2006; Kopacz et al. 2009; Drury et al. 2011; Wecht et al. 2012; Zoogman et al. 2013; Wecht et al. 2014). As such, we prefer to keep the present notation given the history of use in both inverse modelling and trace gas retrieval communities.

2.) One of the results of Bocquet et al. (2011) is that with a proper choice of prolongation operator, one can reduce the smoothing error as much as possible, so that the total error (smoothing+aggregation) is actually a monotonically decreasing function of the resolution. If this is correct, there is no optimal resolution but the finest one (CTM's for instance), except from a numerical efficiency standpoint or if one introduces other sources of scale-dependant errors (such as model errors). The authors presumably obtain such (discrete) optimum because they make an arbitrary choice in the prolongation operator which restricts the transfer of information through scales. Mathematically speaking, this can be seen as an artifact. Had the authors made another implicit choice for the prolongation operator, they

would have found a different result, possibly leading to the finest grid being optimal. If this correct, the authors should clearly acknowledge this and give a fair account of the findings of (Bocquet et al., 2011).

We appreciate Dr. Bocquet's insightful discussion of the error and expanding on the findings of Bocquet et al (2011) and Wu et al. (2011). However, we believe there may have been a slight misunderstanding in the interpretation of the local minimum. Dr. Bocquet assumes the local minimum was due to the use of suboptimal restriction/prolongation operators. As Dr. Bocquet pointed out, this issue was discussed in Bocquet et al. (2011) and Wu et al. (2011). In practice, we are generally unable to specify the true off-diagonal terms in the covariance matrices and typically resort to using a single correlation length scale for the entire domain. In reality, these length scales are not constant. For example, regions dominated by wetland sources (e.g., the Hudson Bay Lowlands) will have large error correlation length scales because the underlying emissions for a large region are driven by a parameterized wetland model while a region like Los Angeles will be largely independent of the surrounding region and should not have a large error correlation length scale. So, while we may not be able to specify the true off-diagonal terms in the covariance matrices we may be able to approximate them. Thus, it was our goal to design our state vector such that it accounts for these off-diagonal terms that are generally missing from native-resolution inversions.

Therefore, the appearance of the local minimum is due to the coarser state vector accounting for off-diagonal terms while the native-resolution inversion does not, thus reducing the error. That said, this is not inconsistent with Dr. Bocquet's reasoning that our local minimum is an "artifact". If the covariance matrices at the native resolution included realistic off-diagonal terms then this approach would indeed be "suboptimal" because our state vector design is not dependent on the Jacobian at the native resolution.

As for Dr. Bocquet's discussion of the total error, in many cases the state vector is not designed using any sort of "intelligent" method like tilings, qtrees, ftrees, PCA, GMM, *k*-means, etc. but is instead designed based on predetermined regions like the TRANSCOM regions (as Review #2 touches on) or a simple coarse-graining scheme. In this case, the simple derivations presented here are useful because they do not require any assumptions about the prolongation operator that Dr. Bocquet claims is hidden in $\mathbf{K}_\omega$ (minor comment #11). We construct the $\mathbf{K}_\omega$'s by perturbing the elements of our reduced state vector, thus explicitly constructing $\mathbf{K}_\omega$ for each different case at many resolutions. We make a choice in designing the restriction operator ($\mathbf{\Gamma}_\omega$) that gives us $\boldsymbol{x}_\omega$ and $\mathbf{K}_\omega$. Correct me if I'm wrong, but I fail to see how we could simply "choose another prolongation operator". The only choice we make is the restriction operator. Thus, the derivations presented here are valuable in that they do not require us to make a choice about a prolongation operator.

We have updated the text to explain this:

Lines 383-394: "Previous work by Bocquet (2009), Bocquet et al. (2011), Bocquet and Wu (2011), Wu et al. (2011), and Koohkan et al. (2012) analyzed the scale-dependence of different grids using the degrees of freedom for signal: $\mathrm{DFS} = \mathrm{Tr}(\mathbf{I} - \mathbf{S}_{a,\omega}^{-1}\hat{\mathbf{S}}_\omega)$. These past works found this error metric to be monotonically increasing. This implies that the native resolution grid will have the least total error and there is no optimal resolution, except from a numerical efficiency standpoint. Here we find a local minimum that is, seemingly, at odds with this previous work. However, the reasoning for this local minimum is that we have

allowed the aggregation to account for spatial error correlations that we are unable to specify at the native resolution. As such, we are taking more information into account and obtaining a minimum total error at a state vector size that is smaller than the native resolution. If the native resolution error covariance matrices were correct then, as previous work showed, the only reason to perform aggregation would be to reduce the computational expense and the grid used here would be suboptimal because it does not depend on the native-resolution grid."

In addition to updating the text to clarify this local minimum, we have added a paragraph discussing the findings of Bocquet (2009), Bocquet et al. (2011), Bocquet and Wu (2011), Wu et al. (2011), and Koohkan et al. (2012) and a distinction of our work:

Lines 57-63: "Previous work by Bocquet (2009), Bocquet et al. (2011), Bocquet and Wu (2011), Wu et al. (2011), and Koohkan et al. (2012) developed optimal grids that allow the transfer of information across multiple scales. These computationally efficient methods (Bocquet and Wu, 2011) generally require the use of the native-resolution grid to derive the optimal representation. They also assume that the native-resolution prior error covariance matrices can be accurately constructed. However, in practice we are generally unable to specify realistic prior error correlations and must resort to simple assumptions."

Minor Comments:

1.) Title: We all know there is no such thing as an "inverse model". This is an abuse of language that I would personally avoid in a title. "Inverse modelling" is almost always preferred.

"Inverse modelling" would awkwardly add another gerund in the title.

2.) p. 1002, l. 4-6: "When the observation vector is large, such as with satellite data, selecting a suitable dimension for the state vector is a challenge". Selecting a suitable dimension for the state vector space is always a challenge, even, and perhaps even more so when the observation vector is small. Let me just mention one paper directly related to what you are discussing and where the observations are in situ and far less abundant than in a satellite retrieval context: Koohkan et al. (2012).

Indeed but that is not really the problem we are addressing. We agree that choosing a suitable state vector is always challenging. However, this work uses satellite data for the example problem and the companion paper performs a "real-inversion" using satellite data. The information content from *in situ* data is, generally, far more intuitive. The observations provide a lot of information near the site and upwind. One could conceivably design a decent state vector by placing many grid cells near the site with fewer upwind. With satellite data (and total column observations) the information content is not immediately clear.

3.) p. 1003, l. 6-7: Same remark as above.

See above.

4.) p. 1003 l. 18-19: "and may not be able to depart from that knowledge". It all depends on the balance between the observation and background statistics. If the background is not informative enough, the solution may be highly oscillating. In a flux inversion context, the retrieved fluxes would increase around the observations sites, which is all but smoothing. In my humble opinion, the appellation is partially misleading. But I might not have understood its interpretation very clearly (from your manuscript or even Rogders' book), in spite of some experience with inverse modelling.

We have rephrased this and added additional citations:

Lines 34-40: "The inverse solution must then rely on some prior estimate for the state vector and may not be able to depart sufficiently from that knowledge. The associated error is known as the smoothing error (Rodgers, 2000; von Clarmann, 2014) and increases with size of the state vector (Bousquet et al., 2000; Kaminski and Heimann, 2001; Kaminski et al., 2001; von Clarmann, 2014)."

5.) p. 1003 l. 18-19: "smoothing error" lacks a proper definition (although it is given later) and interpretation.

See above.

6.) p. 28-29: "Numerical solutions using variational methods circumvent this problem but not inherently provide error characterisation as part of the solution": we know that this is not true. If that kind of statement was fine a few years ago, I believe it should be nowadays mitigated. Several researchers are using conjugate-gradient and quasi-Newton methods such as BFGS that inherently provide estimation of the posterior errors (for instance Bousserez et al., 2015).

We have updated the text to mention these approximate methods in Sections 1 and 2:

Lines 47-48: "Approximate error statistics can be obtained (e.g., Bousserez et al., 2015) but at the cost of additional computation."

Lines 115-118: "Several approaches have been presented to obtain approximate error characterization (e.g., Courtier et al., 1994; Desroziers et al., 2005; Chevallier et al., 2007; Bousserez et al., 2015) but they can be computationally expensive."

Furthermore, we compared the exact posterior covariance matrix to some of these approximate posterior covariance matrices and found the discrepancies were large to be useful for our analysis, thus we decided not to pursue that approach further.

7.) p. 1004: the literature is incomplete. I believe you have to mention Wu et al. (2011), given it is very close to your objective and analysis and also related to greenhouse gas flux inversions.

Done. See, for example, our response to major comment #2.

8.) p. 1006, l. 5-12: This is incomplete or partially incorrect. The Jacobian can also be computed using the model adjoint, requiring m runs. By the Sherman-Morisson-Woobury lemma, the matrix algebra will scale like m^3. Also, sequential updating by serial processing of observations usually (unless the scheme is sub-optimal) leads to the same numerical cost.

Both points are true. However, typically $m >> n$ so using the adjoint to construct the Jacobian is usually not the most efficient method. As for the latter point, the benefit comes in the reduced memory cost for the matrix operations (i.e. break a large matrix up into smaller matrices so you can perform the necessary matrix operations).

9.) p. 1007, l. 11: "Probabilistic" is one word too many. Bocquet et al. (2011) additionally provide a probabilistic interpretation. But it can be seen as an entirely deterministic process just as the Best Linear Unbiased Estimator (BLUE) formalism. Please remove the word "probabilistic" which conveys the wrong idea in the context of this sentence.

Done. See minor comment #10 (below) for updated text.

10.) p. 1007, l. 12: "However, construction of this prolongation operator is not a well-posed problem because the operator is not unique". Please rephrase the sentence. The construction as defined by Bocquet et al. (2011) is well-defined and well-posed. But in general the choice of the prolongation operator is not unique. Incidentally, you do make a choice for the operator without acknowledging it! That is why I disagree and think that your method might be less robust. But maybe you meant "more practical" rather than "more robust", did you? If you did intend "less robust", please justify your statement with precision.

Our apologies. Yes, "more practical" was the intended meaning. We have corrected this phrasing.

We have updated the text as:

Lines 132-134: "Their analysis relies heavily on the construction of a prolongation operator ($\mathbf{\Gamma}_\omega^\star$) mapping $\boldsymbol{x}_\omega$ back to $\boldsymbol{x}$: $\boldsymbol{x} = \mathbf{\Gamma}_\omega^\star \boldsymbol{x}_\omega$. However, construction of this prolongation operator is not unique. We present here a simpler and more practical method."

11.) p. 1007, Eq.(12): Please define $\mathbf{K}_\omega$ (the source-receptor matrix). That is where you put the definition of the prolongation operator under the carpet... This must be discussed.

We explicitly constructed the $\mathbf{K}_\omega$'s through perturbations to the reduced state vectors for all the different cases. Thus, $\mathbf{K}_\omega$ was constructed in the same manner as $\mathbf{K}$. So all we needed was a reduced state vector ($\boldsymbol{x}_\omega$) which we defined in Eq. 10. There was no use of a prolongation operator in the construction of $\mathbf{K}_\omega$.

We have updated the text as:

Lines 145-147: “Here $\boldsymbol{y}$ is the observation vector (common in both cases), $\boldsymbol{x}$ and $\boldsymbol{x}_\omega$ are the true values of the native-resolution and aggregated state vectors, and $\mathbf{K}$ and $\mathbf{K}_\omega$ are the native resolution and the reduced-dimension Jacobians.”

12.) p. 1008, l. 7-8: The introduction of the concept of ensemble is cumbersome (just as it is in Rodgers (2000) to be fair). It requires more justification. It appears as a *deus ex machina*.

This concept is not critical to our derivation. Furthermore, this concept has been extensively discussed in the retrieval community. We have included additional references directing the reader to more complete discussions of this concept.

Lines 152-153: “Obtaining the error statistics for $\boldsymbol{\epsilon}_\mathrm{A}$ requires knowledge of the pdf of $\boldsymbol{x}$ for the ensemble of possible true states (cf. Rodgers, 2000; von Clarmann, 2014).”

13.) p. 1008; l. 5: "A" for aggregation, and "a" for background. Really? Why not use "b" for background instead of “a”?

“a” for the a priori is consistent with the notation of Rodgers (2000).

14.) p. 1010, l. 10: Please define the gain $\mathbf{G}_\omega$ explicitly. As I explained, several choices can be made, one being more consistent. Without an explicit definition, you hide what is at the origin of the appearance of the fittest resolution.

Done.

Line 197: “ $\mathbf{G}_\omega = \left(\mathbf{K}_\omega^T \mathbf{S}_\mathrm{O}^{-1} \mathbf{K}_\omega + \mathbf{S}_{\mathrm{a},\omega}^{-1}\right)^{-1} \mathbf{K}_\omega^T \mathbf{S}_\mathrm{O}^{-1}$ ”

15.) p. 1011, l. 6-8: The sum of an increasing and decreasing function does not always possess a minimum.

We have rephrased this line:

Lines 210-213: “Because the smoothing error increases with state vector dimension while the aggregation error decreases, analysis of the error budget can point to the optimal dimension where the total error is minimum.”

16.) Koohkan et al. (2012) discuss how to choose the optimal resolution and how it is impacted by the error balance (observation versus background, section 2.2). Since, ultimately, you end up making the same choice as all the papers I am referring to, that is to say choosing the resolution on a numerical cost basis, Koohkan et al. (2012)’ discussion is relevant and perhaps a bit more precise than only adjustment with respect to the

observation error only.

Thank you for pointing us to Section 2.2 of Koohkan et al. (2012) for the discussion of the error balance, however Section 3 of our manuscript is mostly concerned with simply deriving the different error components. We have added a brief discussion of Koohkan et al. (2012) (See response to minor comments #2 and #19).

17.) p. 1011, l. 12-18: Again, this discussion appears like a *deus ex machina*.

See response to minor comment #4 from Reviewer #2.

18.) p. 1012, l. 5: Actually the adaptive grid method based on tiling was introduced in Bocquet (2009). Moreover, it's worth mentioning that these grid are built to be optimal for the purpose of the inversion.

We have added Bocquet (2009) to the discussion and amended the description:

Lines 225-227: "Analyzing the off-diagonal structure of a precisely constructed prior error correlation matrix would provide the best objective way to carry out the aggregation, as described by Bocquet (2009), Bocquet et al. (2011), and Wu et al. (2011)."

19.) p. 1012, l. 8: Bocquet and Wu (2011) also use PCA coupled to the hierarchical grid to compute an optimal grid in a numerically efficient way yet capturing the variability of the prior. This should be acknowledged.

We have added Bocquet et al. (2011), Bocquet and Wu (2011), Wu et al. (2011), and Koohkan et al. (2012) to the discussion:

Lines 231-234: "Previous work by Bocquet et al. (2011), Wu et al. (2011), and Koohkan et al. (2012) used tiling and tree-based aggregation methods, while Wecht et al. (2014) used a hierarchal clustering method based on prior error patterns. Bocquet and Wu (2011) also used principal component analysis (PCA) coupled to the hierarchal grid to compute an optimal grid."

20.) p. 1013, l. 15-21: Rodgers (2000) also suggests projection over a specific function basis albeit in a different context.

21.) p. 1016, l. 5: Could you please briefly discuss the numerical cost of the approach?

See response to minor comment #6 from Reviewer #2.

22.) p. 1016: The application of the GMM methods is very interesting. From the methodological standpoint, I believe the fact that the control space is defined with a *probabilistic* mixture is quite novel in this context.

23.) p. 1017: What about the time dimension? Do you apply aggregation in time? I assume you didn't, but you could have.

This same approach should be applicable to the time dimension:

Line 240: "However, the same methods can be used for temporal aggregation."

24.) p. 1017: What if the background error covariance matrices were not diagonal? Could you discuss the issue a little? Apart from the numerical problem, we can see from Eq. (2) that properly transferring information through the scales is more tricky. If one chooses a pragmatical $\mathbf{\Gamma}_{\omega}^{\star}$ as you do (or as I could as well for a very high-dimensional application), it is possible that the resulting "optimal" resolution would be more pronounced.

This is an excellent question and goes back to the main issue discussed in your second major comment. The two main benefits of these methods would be: (1) computational cost and (2) accounting for spatial correlations that are difficult to specify. If one could specify realistic off-diagonal error correlations then the only benefit to a multi-scale approach would be the computational benefits. However, in practice we are unable to specify those realistic off-diagonal error correlations and, thus, have neglected valuable information. See our response to major comment #2 for the added text on this.

25.) p. 1017, l. 18: Please spell out SD (standard deviation?).

Done.

26.) p. 1018, l. 8-13: Your result is not surprising. Because of the baseline results of Bocquet et al. (2011), I was expected that kind of results with a non-pronounced minimum (unless your implicit prolongation operator is badly chosen). Above all, you end up choosing the optimal resolution on a numerical efficiency criterion, just as we did (for not only practical but also theoretical reasons). This should be acknowledged.

See response to major comment #2.

27.) p. 1018-1019: The conclusion should be amended.

Reviewer #2 Comments:

1.) My biggest complaint, however, is the choice of the journal. When I read or review a paper in Atmospheric Chemistry and Physics, my first question is "What have I learned about the physics or chemistry of the atmosphere from this paper?" Unfortunately for this manuscript, the answer to that question is "Nothing!". This is not to say that the work is not

good or not important; it is both, and should be published. However, it is a technical study that will be of relevance only to a class of modelers during their model development, and therefore I think Geoscientific Model Development (from the same publishers) is a much better journal for publishing this work. I would strongly urge the authors to consider submitting this specific work to that journal instead. I do not think this suggestion should come as a surprise to the authors. Previous work on the same problem (which they cite) was published in the Quarterly Journal of the Royal Meteorological Society, and similar technical developments are routinely published in Geoscientific Model Development.

We considered Geoscientific Model Development (GMD) but ultimately chose to Atmospheric Chemistry and Physics (ACP) for three reasons:

1.) This is a companion paper to Turner et al. ACPD (2015) that performs a "real-world" inversion with 2.5 years of GOSAT methane data. It seems fitting to keep the companion papers in the same journal.
2.) The focal point of this paper is about a methodology, not code development. Furthermore, based on the journal scopes, this seems to be a more natural fit for ACP than GMD.
    ACP scope: "The journal scope is focused on studies with general implications for atmospheric science" (see: "http://www.atmospheric-chemistry-and-physics.net/about/aims_and_scope.html").
    GMD scope: "dedicated to the publication and public discussion of the description, development, and evaluation of numerical models of the Earth system and its components." (see: "http://www.geoscientific-model-development.net/about/aims_and_scope.html").
    There have been previous "methods" papers that were published in ACP.
3.) This is a methodology that is widely applicable to both atmospheric chemistry and physics. We also directly apply this methodology to an atmospheric chemistry problem as an example.

2.) My second biggest complaint is the applicability of the technique detailed here. As someone who does atmospheric inversions off and on, my first impulse upon coming across a manuscript of this sort is to wonder "This looks great! Can I apply this technique to my inversions?" From the manuscript, it is not clear that I or any other atmospheric inverse modeler will be able to use the results presented here in real-world inversions. The authors choose the optimal number of state vector elements as the number which minimises the total error in Figure 3. If I understood correctly, generation of Figure 3 required performing the same inversion over and over again with different restriction operators \Gamma, to get the posterior covariance matrices. This was possible for the authors because their native resolution state vector was small, owing to their choice of focussing on the annual average emission over N America. In most real world inversions spanning multiple years with daily/weekly variability in the fluxes, performing the inversion is the most time consuming part, and so performing many inversions just to figure out the optimal size of the state vector seems like a waste of resources. After all, since the authors show that even at the native resolution the smoothing error does not become significant compared to the observational error, what's wrong with just solving at the native (CTM) resolution? I would be happy to be proved wrong on this point, and to be shown that one doesn't need to execute a bunch of inversions to estimate the optimal size. From the current manuscript, however, I do not see how one could use this technique in a real-world inversion, for example any of the CO2 inversions in Peylin et al (Biogeosciences, 2013), or any of the CH4 inversions in Kirschke et

(Nature Geoscience, 2013). This is one more reason why I would prefer to have this manuscript published in a journal dedicated to technical developments (such as GMD) instead of ACP.

We suspect the reviewer may not have noticed that the time period for this manuscript was greatly reduced from our "real-world" inversion presented in the companion paper (Turner et al. ACPD 2015). So the reviewer is correct that we do use Figure 3 to guide the choice of state vector size but we do so using a shorter time period. We have rephrased the text to mention this sooner:

Lines 165-167: "Application of Eq. 17 requires computation of the native-resolution Jacobian **K** but this can be done for a limited test period only. We will give an example below."

> After all, since the authors show that even at the native resolution the smoothing error does not become significant compared to the observational error, what's wrong with just solving at the native (CTM) resolution?

We agree that the native CTM resolution would be ideal if we could specify the off-diagonal error correlations in a realistic manner. However, in practice the off-diagonal error correlations are neglected or simply treated with a single correlation length scale. In this framework we are able to, in essence, prescribe different correlation lengths for different regions (e.g., LA is distinct from the surrounding region whereas the HBL wetlands have a longer correlation length). See our response to Dr. Bocquet's major comment #2.

> From the current manuscript, however, I do not see how one could use this technique in a real-world inversion, for example any of the CO2 inversions in Peylin et al (Biogeosciences, 2013), or any of the CH4 inversions in Kirschke et (Nature Geoscience, 2013)

As for the applicability to "real-world" inversions, I believe this is a perfect example of the applicability. In this manuscript we sampled the full range of possible state vector sizes and determined a reasonable state vector size. We then used that state vector in a "real-world" inversion (Turner et al. ACPD 2015).

Minor Comments:

1.) In the abstract and in section 5 (bottom of p1017), the authors make the point that the GMM method retains resolution of major local features in the state vector. This is true, but only if the prior already has that particular feature. Further, this is not always an advantage, since those major features can sometimes be wrongly located in the prior emission estimate (less of an issue with coal mines and power plants, big issue for wetlands and bovine methane). I would like the authors to mention this.

Thank you for bringing this up. The GMM method can retain major local features if they are

in the prior *or* the adjoint-constraint ("Similarity Vector" number 3; Table 1). Our motivation for including the adjoint-constraint dataset was to allow the state vector to include potential "missing sources" seen by the observations. Further, one could add other datasets (or replace sectors in the prior) if they have reason to believe the other dataset is representative of prior error correlations. We have expanded on this in the manuscript:

Lines 255-256: "the third [similarity vector] represents the scaling factors from the first iteration of an adjoint-based inversion at native resolution"

Lines 259-263: "We choose here to include initial scaling factors from the adjoint-based inversion because we have them available and they can serve to correct any prior patterns that are grossly inconsistent with observations, or to identify local emission hotspots missing from the prior. One iteration of the adjoint-based inversion is computationally inexpensive and is sufficient to pick up major departures from the prior."

2.) On page 1003, near line 25, the authors say that an additional cost of using a large state vector is the increased computational cost of the inversion. This is not correct. In fact, in most inversions beyond TRANSCOM-style basis region inversions, the costliest part of the inversion is the evaluation of the forward model F (and its adjoint, if needed), be it a CTM in variational/EnKF systems, or an LPDM for "batch" inversions. Irrespective of the aggregation chosen for the state vector, the atmospheric transport still needs to be run at the native resolution, which is the time limiting step.

We were specifically referring to analytical inversions like the one performed here. In an analytical inversion a larger state vector could increase the computational cost of the inversion in two ways: (1) it will increase the size of the matrices that need to be multiplied and inverted and (2) it may increase the number of forward run "batches". For example, simulating atmospheric transport at high resolution generally requires a lot of memory. A modeler can quickly reach the allowable memory limits if they are running thousands of forward runs to construct the Jacobian. This was the main limiting factor in our simulations; ultimately, we could only run a few hundred forward simulations at a time due to memory constraints on the cluster. Thus, we had to run multiple batches of forward runs. The second point does not apply to LPDM inversions. We have rephrased the text to clarify this.

Lines 43-44: "An additional drawback of using a large state vector is that analytical solution to the inverse problem may not be computationally tractable. Analytical solution requires…"

3.) On page 1008, near line 15, the authors mention the assumption that the prior is unbiased. While this is an assumption widely adopted theoretically, in practice it is rarely true. A biased prior leads to a biased posterior, a fact inverse modellers grudgingly live with, as long as they think that the posterior bias is lower than their posterior uncertainty estimate. I would like to know what the consequence of a biased prior is for determining the optimal length of the state vector. Is that estimate expected to change?

Absolutely, a biased prior will bias the posterior. The only reason that an additional error term would impact the optimal state vector length is if it exhibited scale-dependence. Intuitively, a bias term would not exhibit scale-dependence and would, presumably, behave

like the observation error term. As such, a bias in the prior will bias the posterior but not affect the choice of the optimal state vector size. We have added text to reflect this.

Lines 208-209: "A bias term should exhibit similar scale-dependence to the observation error term and could be included by following the derivation from Rodgers (2000)."

Lines 101-103: "We have assumed here that errors are unbiased, as is standard practice in the inverse modeling literature. An observational error bias $\boldsymbol{b}_O$ would propagate as a bias $\mathbf{G}\boldsymbol{b}_O$ in the solution $\hat{x}$ in Eq. 8."

4.) On page 1011, near line 15, the authors have a caveat, which, if I understand correctly, says that one of the assumptions is that the error covariance matrix of the true state is the same as the error covariance matrix for the prior state. Did I understand correctly? If so, then that's a big assumption; knowing the error covariance of the true state before doing an inversion seems like a big ask! If I misunderstood, I will be happy to be corrected.

This is, indeed, correct. In Section 5 we present a simple experiment where we pretend to know the true emissions, so for the purposes of this experiment our assumptions are valid. These expressions are useful for similar experiments where one wants to diagnose the different error components. However, these expressions should not be used to diagnose errors in a "real-world" inversion because that assumption will not hold. Rodgers (2000, p. 49) and von Clarmann (2014) present a detailed discussion of this exact issue. We have rephrased this paragraph:

Lines 215-221: "A caveat in the above expressions for the aggregation and smoothing error covariance matrices is that they are valid only if the prior $\boldsymbol{x}_a$ is the mean value $\bar{x}$ for the pdf of true states and if the error covariance matrix Sa is the covariance matrix for that pdf ($\mathbf{S}_e = \mathbf{S}_a$). Rodgers (2000, p. 49) and von Clarmann (2014) provide a detailed discussion of the errors induced by failing to meet this assumption. As such, these conditions define the assumption for the prior, so the expressions can be taken as valid for the purpose of selecting an appropriate state vector dimension in an inverse problem. However, they should not be used to diagnose errors on the inversion results."

5.) One aggregation technique the authors do not discuss is K-means clustering. If we choose the number of clusters to be equal to the optimal number of state vector elements, and use the same 14 variables as the GMM model to determine the clusters, how would the smoothing and aggregation errors compare to the GMM+RBF case? Did the authors already look into that? If so, I would love to see the results.

Excellent question. As the reviewer may have surmised, there was a very large computational expense associated with explicitly constructing the Jacobian multiple state vector sizes with multiple methods. As such, we considered including *k*-means clustering and performed some preliminary analysis with *k*-means. The figure below shows example clusters created using the same similarity matrix and criteria as in Figure 1. The *k*-means clustering used 100 replicates (different initializations) with 1000 iterations per replicate. We ultimately decided to include the course-graining method instead of *k*-means. However, the different methods perform comparably so, presumably, *k*-means would also give similar results.

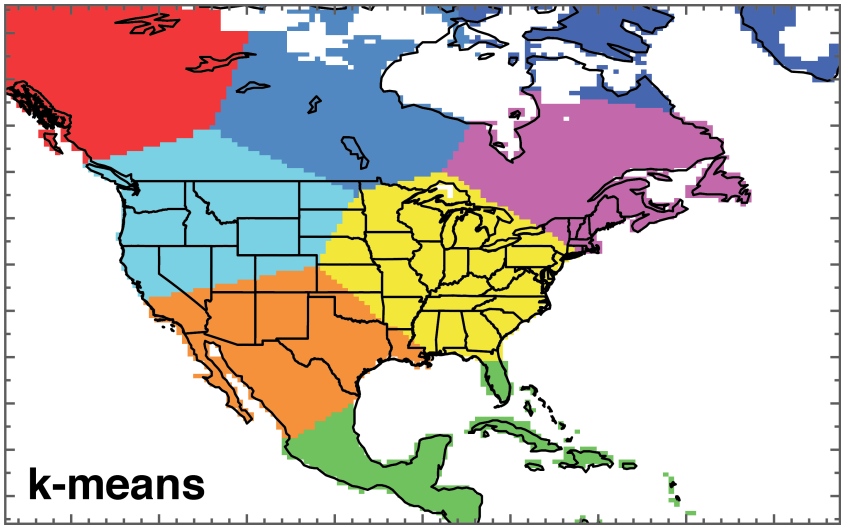

Figure 1: Same as Figure 1 from the manuscript but for *k*-means clustering.

We have added the following text:

Lines 365-368: "The different aggregation methods of Sect. 4 yield very similar smoothing errors, suggesting that any reasonable aggregation scheme (such as *k*-means clustering (cf. Bishop 2007)) would perform comparably."

6.) On page 1016, line 4, the authors say that equations (32)-(35) are iterated until convergence. What counts as convergence, i.e., what is the convergence criterion?

We used an absolute tolerance of $\tau < 10^{-10}$ where:

$$\tau = \sum_i \sum_j \left| \mathcal{M}_{i,j} - \mathcal{M}^{\star}_{i,j} \right| + \sum_i \sum_j \sum_k \left| \mathcal{L}_{i,j,k} - \mathcal{L}^{\star}_{i,j,k} \right| + \sum_i \left| \pi_i - \pi^{\star}_i \right|$$

and the superscript star indicates the value from the previous iteration. We didn't use a relative tolerance because the true value of one of the parameters could, potentially, be zero. In any case, preliminary tests were insensitive to using a tolerance of 10^{-4}. In response, we have added the following text to the manuscript:

Lines 319-326: "The computational complexity for the expectation-maximization algorithm is $\mathcal{O}(nK + pn^2)$ (Chen et al., 2007), however the actual runtime will be largely dictated by the convergence criteria. Here we use an absolute tolerance of $\tau < 10^{-10}$ where

$$\tau = \sum_i \sum_j |\mathcal{M}_{i,j} - \mathcal{M}^{\star}_{i,j}| + \sum_i \sum_j \sum_k |\mathcal{L}_{i,j,k} - \mathcal{L}^{\star}_{i,j,k}| + \sum_i |\pi_i - \pi^{\star}_i|$$

and the superscript star indicates the value from the previous iteration."

7.) On page 1017, line 26, the authors say that RBF weighting performs slightly better than GMM clustering. Is this a general statement about RBF vs clustering, or is it because the 14 variables used to construct the similarity matrix (table 1) are strongly correlated with CH4 fluxes?

This is a general statement about RBF weighting vs. GMM clustering (as well as coarse-graining and PCA clustering). However, it's not necessarily a general statement about RBF weighting vs. other clustering methods. That said, we suspect that RBF weighting would perform favorably against more clustering methods but we have only tested a small subset of clustering methods here.