Response to Anonymous Referee #2

The manuscript, "Development of the WRF-CO2 4DVar assimilation system", describes exactly that. The authors have created linearized model versions of the WRF-Chem CO2 transport mechanisms, validated their performance against finite difference approximations, and demonstrated their utility in a simplified pseudo-data experiment. The introduction covers much of the relevant literature necessary to get to the same starting point as the authors, and we make a few additional suggestions below. Theadjoint and tangent linear model developments are described thoroughly, and would be helpful for any person working their way through the code at a later time. The adjoint model evaluation falls a little short, and we provide some suggestions for ways it could be improved. The pseudo-data inversion test, while quite simple and unrealistic, demonstrates that the inversion framework is working. It is a first step that undoubtedly took considerable effort, but needs some improvements in the application of the new tool. There is no discussion of the statistical nature of 4D-Var, which is paramount to that method's success with real data and its being labeled a Bayesian inversion technique. We have several specific comments as to how the discussion could be mademore precise and also miscellaneous technical corrections.

We thank the referee for the time they have taken to improve the paper with their insightful and detailed comments. Below is a summary of the major work conducted in response to the referee's comments. The point-to-point response is in the following sections.

- Model code debugging: as the referee point out, the adjoint model was not error free. To address the problem, we systematically debugged the model code. Errors were isolated, identified and corrected. The evaluation through sensitivity calculation confirms that the three model components (NL/TL/AD) match as expected.

- Optimization experiment: in the revised text, synthetic observation data are from 30 vertical levels from bottom up. They were from the bottom level only in the original text.

- In the footprint calculations, receptors are now placed at 1^{st}, 5^{th}, and 10^{th} . They were placed on the 1^{st} level only in the original text.

- TL/AD/FD sensitivity comparison. (1) tangent linear, adjoint, and finite difference sensitivities are calculated for source and receptors cells at different locations in both horizontal and vertical). (2) The receptor cells are placed at the 1^{st}, 5^{th}, and 10^{th} vertical levels at each tower site.

- Cumulus activity indication: extra variables are implemented in the model to track when/where the convective tracer transport is activated during the simulation. This information is plotted and used to ensure there are sources and receptors located within or near the cumulus activity. As the referee pointed out, this is necessary to evaluate the accuracy of the newly developed TL/AD code of the convective chemistry transport scheme (module_ctrans_grell).

Major Comments:

• Section 2.3: Incremental 4D-Var is used to optimize nonlinear systems. But the CO2 tracer simulation is inherently linear. Thus, I don't really see at this point what the benefit would be of an incremental formulation, nor how updating the inner loop with an outer loop integration would provide any additional information. Thus, the importance of including both of these methods and their comparisons for CO2 inversions needs further justification.

We thoroughly examined the model code for the linearity pointed out by the referee. We found that WRF-CO2 is linear except at certain situation when positive definite chemistry transport is used. Our examination shows that the ACMPBL and convective transport (ctrans_grell) are both linear with respect to CO2. With positive definite chemistry advection, nonlinearity can be introduced when the predicted minimum possible CO2 is negative at a grid point. This will trigger a renormalization procedure, which is nonlinear. We confirmed this nonlinearity through examining finite difference sensitivity around grid point where the above mentioned renormalization is artificially triggered. In order to trigger the renormalization, we created large horizontal CO2 gradients. We do not believe such large CO2 gradient is very common in nature, but may be possible for wild fire emission. In the 24-hour simulation used in this paper, the renormalization was not triggered in any grid cell, thus the system is linear.

The outer loop updating used in the original text was necessitated by the error in the adjoint model code, and we mistakenly attributed it to the loss of conjugacy. We greatly appreciate the referee's insightful comment. With the corrected adjoint model, Lanczos-CG based incremental optimization does not need the outer loop update: only one outer loop iteration is applied in all inverse modeling experiments.

Our inverse experiment results (with the corrected adjoint model, and observation at 30 vertical levels) show that Lanczos-CG converge substantially faster than L-BFGS-B. Although we are aware that this performance difference may be specific to our experiment setup, we consider it is necessary to include both optimization schemes in WRF-CO2 4DVar for future applications.

• Fig 8: As the authors surely know, adjoint sensitivities should agree with the tangent linear sensitivities to numerical precision. The differences between these sensitivities vs the finite difference sensitivities, given that the latter match the tangent linear sensitivities, is an indication that the adjoint model is not error free. For the cases tested here, the errors are manageable, yet there is no guarantee that the errors would not grow for longer simulations. The authors should thus continue to debug their adjoint code, possibly by performing this type of test around the tangent linear and adjoint code of individual physics components developed here (such as ACM2 PBL mixing). If they can not resolve the code bugs this way, they should at least perform additional tests using different receptor locations and simulations of increasing (and decreasing) length to examine how the numerical errors may be accumulating.

Yes, we are aware that adjoint model was not error free. We thank the referee for reminding us to correct it. Following the referee's suggestion, we debugged the individual processes in isolation. We also modified the code to test all three models in a single time step mode. Code errors were identified and corrections were made. We evaluated the updated code by comparing the adjoint sensitivity against the tangent linear and finite sensitivities. As suggested by the referee, these sensitivities are calculated with sources and receptors at different grid cells and the receptors cells are placed at multiple vertical levels. (See Fig. 2 next page for the source and receptor placement. We also ensured that there are sources and receptors placed within or near cumulus activities for testing the convective transport code. (See Fig. 3 for cumulus indicator).

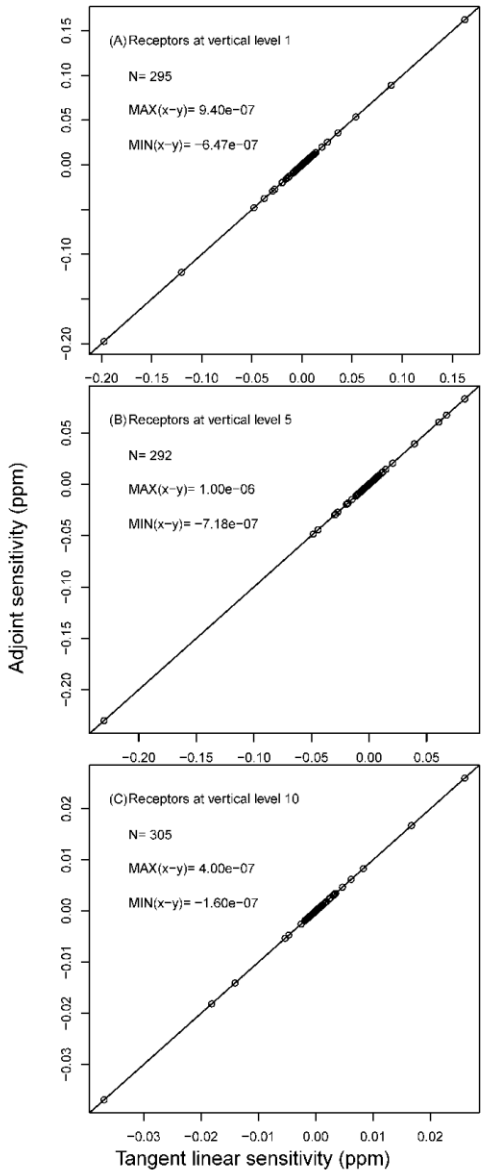


Figure 1. Comparison between tangent linear sensitivity (x-axis) and adjoint sensitivity (y-axis). The sensitivities are organized into three groups for receptors placed at the 1st, 5th, and 10th vertical levels. Because there 35 sources (red starts Fig. 2) and 20 tower sites (red triangles in Fig. 2), there are 700 (35x20) pairs of tangent linear and adjoint sensitivities for each group. Sensitivities with absolute values less than 10^{-10} are not included in comparison, resulting in 295, 292, and 305 pairs of comparisons. The solid lines in each figure are the 1:1 lines. Because the values are very close, we summarize each figure by the minimum and maximum difference between the tangent linear and adjoin sensitivities, instead of the slope and r-squared.

• p. 12, line 7: It is not required that a grid cell be both a receptor and a source to have non-zero sensitivity, as evidenced by Figures 5 and 6. The source and receptor must simply be significant (large source and large perturbation to concentration due to that particular source). Choosing them to be the same grid cell is likely to produce good agreement between the adjoint and finite difference methods even when the advection, PBL, or convective transport adjoint may be incorrect. *Did the authors choose identical grid cells for source and receptor?* If so, then additional tests are required to prove that the adjoint code works as described. Additionally, validating the convective transport adjoint and tangent linear codes requires demonstration during a period when the subgrid cumulus parameterization is active and for sources and receptors near that phenomenon. The authors should present some indicator of cumulus activity in an additional figure.

Placement of the source and receptor cells: In the original manuscript, the source and receptors were the same grid cells. They were the bottom level grid cells where the 20 towers are located.

We agree that code should be tested with sources and receptors placed at different grid cells as long as there are discernible impact during the simulation period. To address this, we conducted the sensitivity calculations with the updated model in a more systematic approach: (1) Receptors are still placed at the 20 tower sites in horizontal, but at each tower site, 3 receptors are placed at different vertical levels: level 1, 5, and 10. (2) Sources are at a different set of 35 grid cells placed around the receptors. The placement of sources and receptors used in the revised manuscript is in Figure 2 below.

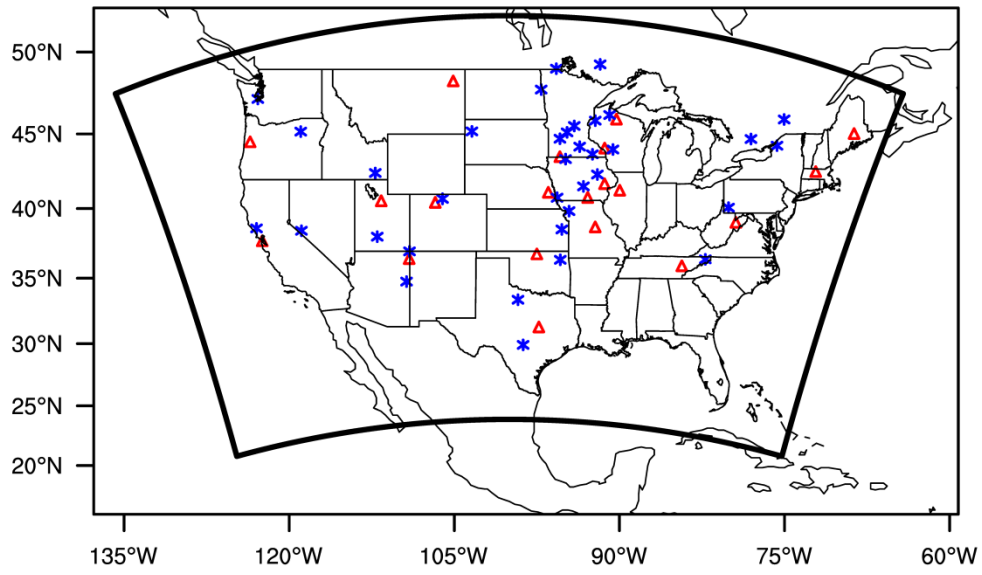
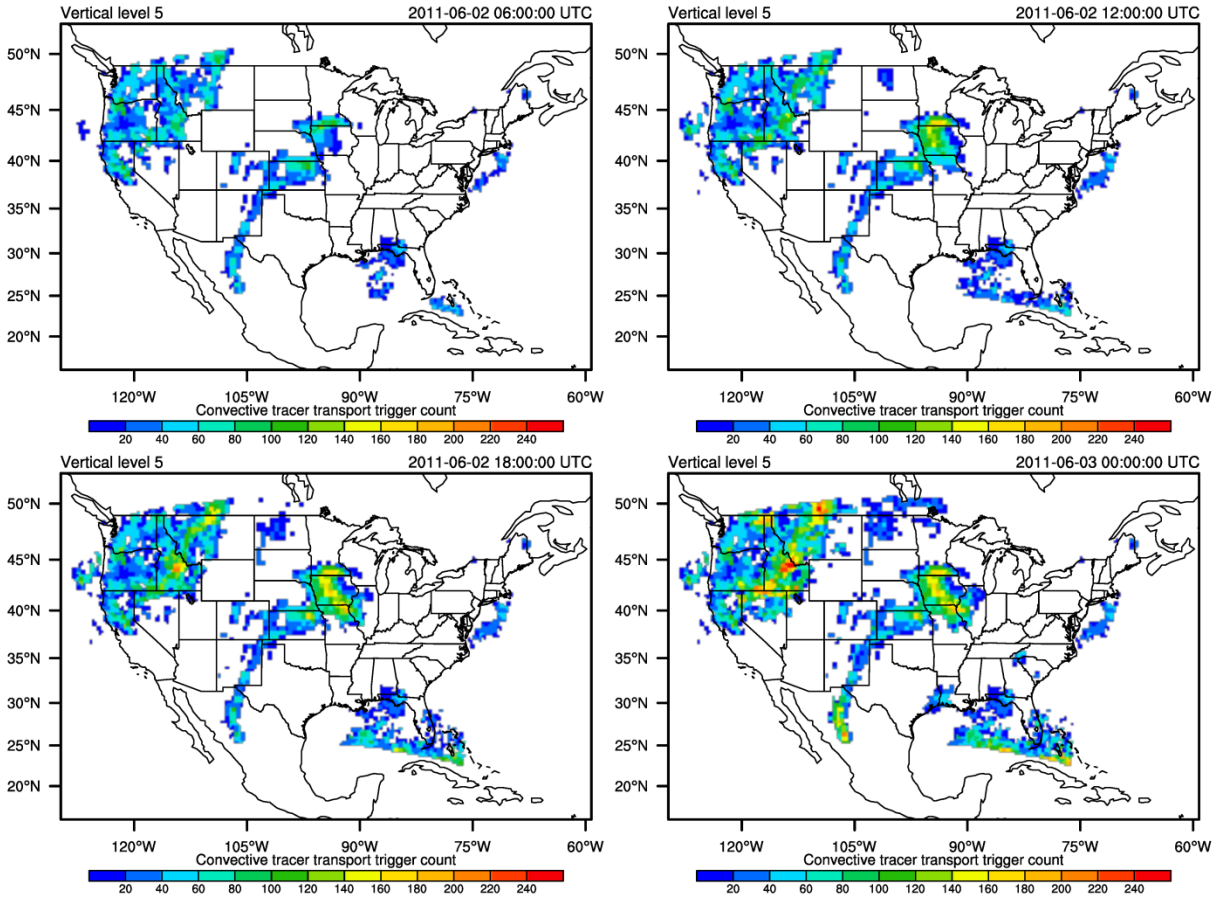


Figure 2. Placement of the sources (blue stars) and receptors (red triangles). Receptors are the placed at 1st, 5th, and 10th vertical levels of grid cell of the 20 towers (Table 4 of the manuscript). Being the surface flux source, all sources are place at 1st vertical level.

Indicator of cumulus activities:

To ensure some sources and receptors are placed within or near cumulus activities, we implemented an counting mechanism within the convective transport code (module_ctrans_grell.F in the chemistry directory). In WRF, whether chemistry species at each grid cell is vertically transported at a given time step by the cumulus process is determined by a number of tests and marked by a pair of flags (one for deep convection and another for shallow convection). These two flags are reset at each time step. We added two variables to track the two flags across time steps: each time the convective transport process is triggered at a grid cell, its count increases by one. We refer the new variables "convective tracer transport trigger count" and used it to examine the cumulus activities. Comparison of Figure 3 (below) and Figure 1 confirms there are sources and receptors placed within or near cumulus activities.



Fiure 3. Convective tracer transport trigger count plotted at 6-hour intervals at vertical level 5. These counts are vertical level specific as the convective transport activations are determined for each vertical level. The figures show that deep convections are trigger during a large portion of the 24-hour simulation at the Pacific northwest and Midwest (center round Iowa). A comparison with source/receptor cells placement in Figure 1 confirms that there are

sources/receptors placed around cumulus activities, and thus tangent linear and adjoint code of the convective transport are indeed tested.

• p. 12, lines 16-17: Similar to the comment above, choosing the observations to be in the lowest model layer reduces the importance of having the adjoint and tangent linear treatments of vertical mixing. The observations should be spread more thoroughly, vertically.

We agree with the referee.
To address this issue, new pseudo observation data are generated from the forward model run at the first 30 vertical levels (out of a total of 50 levels). This means that there are 30 observations at each horizontal grid. We emphasize this setting is for the sole purpose of testing the inverse system with ideal error-free synthetic data. The inverse experiment results are shown in Fig. 10-13 of the revised manuscript.

• From a software perspective, I'm a bit confused about the distinction between the present work and that of GH15/16, where the adjoint of the BC tracer is developed. So, essentially the update here is that BC has been changed to CO2, and convective transport has been added? On a broader note, would it be beneficial to the community to view these as two options within a single chemical 4D-Var system, rather than as two different models? I realize this likely results from development of these systems over time, in parallel, but thinking to the future I wonder if a model merge would be in order. To illustrate my point, imagine if rather than a consolidated WRF-Chem model, we had separate WRF-Chem-BC, WRF-Chem-CO2, WRF-Chem-CO, . . . etc. models. That would clearly hinder development of the tool as a whole, which shares many common elements across the different tracer simulations.

Yes, we developed WRFCO2 4dvar system as part of an effort proposing to NASA's carbon science program. Our development is in parallel with G15/16, and no collaboration has been involved yet.
We agree with the referee that coordinated code development will benefit the community. We will contact Dr. Henze's group for collaborating on future code consolidation/merging.

Further Comments:
1. p. 2, line 33: The reference to Streets 2013 is a bit odd, as that paper is specifically a review of remote sensing based constraints on emissions and focuses mainly on reactive trace gases and aerosols. As the present work doesn't seem geared towards remote sensing observations, some other references to literature on regional CO2 inverse modeling would be a better fit here.

This reference has been removed.

2. p. 2, line 9: Probably more correct to say "instead they directly compute the product of the Jacobian with a forcing vector, which is the gradient used for optimizing

the state vector."

This sentence has been corrected following the referee's suggestion.

3. p. 2, line 10: The notion that posterior error can not be calculated from a variational inversion is outdated. Posterior error can be calculated analytically using the Lanczos vectors from a CG minimization in the incremental 4D-Var framework following Fisher and Courtier (1995) - as currently done in operational weather forecast centers such as ECMWF and the UK Met Office - for minimal additional cost. Efficient posterior error estimate for non-incremental 4D-Var frameworks are described in Bousserez et al., QJRMS, 2015, including previous works on Monte Carlo (e.g., Chevallier et al., JGR, 2007) and stochastic (Rabier and Courtier, QJRMS, 1992) methods.
Fisher M, Courtier P. 1995. Estimating the Covariance Matrices of Analysis and Forecast Error in Variational Data Assimilation, Technical Memorandum 45. ECMWF: Reading, UK.

We thank the referee for correcting us. This statement about posterior error calculation has been corrected accordingly.

4. The introduction states that both bottom-up and top-down approaches are used, but does not say why that is the case. It is recommended to move the first paragraph of Section 4 to the introduction.

The related paragraphs have been rearranged following the suggestion.

5. p. 2., line15: The GEOS-Chem CO2 4D-Var system is also part of JPL's Carbon Monitoring System, e.g. Liu, Bowman, Lee, et al., Tellus B, 2014; Liu, Bowman, and Lee, JGR, 2016.

Thanks for pointing this out. The text has been modified accordingly.

6. p. 2, line 18: Also Chevallier et al., JGR, 2007.

Thanks. This has been fixed.

7. p. 2: 2: For regional CO2 inversions, the list isn't entirely complete, see also Alden, Miller, Gatti, et al., Global Change Biology, 2016; Chan, Chan, Ishizawa, et al., GMDD, 2016. There are others, but I think it suffices to say the literature review could be a bit more comprehensive (or, alternatively, scoped / phrased as to be more narrow).

Agree. The regional CO2 inversion review has been strengthened with the following the two additional literatures.

8. p. 2, line 33: Here and in several other places, the authors use the phrase "influence

function" without every having defined it.

Definition of the influence function has been added.

9. Equations 3,4: define superscript n in this context.
Definition for superscript n has been added.

10. Equations 6,7: I understand incremental 4D-Var, but I think still the authors should rigorously define the superscript n in this context for the sake of completeness, which I believe should differentiate between inner and outer loops. Also, incremental 4D-Var is usually employed with a square-root preconditioning, which I don't see here.

Text has been added to define the superscript n and how it changes differently within the inner and outer loop of the incremental optimization.
The square-root preconditioning is not applied in the pseudo-data based inversion experiments. As explained in Section 3.4 of the manuscript, background error matrix is set to infinity and observation error matrix is set to the identity matrix. This is realized in the code by setting the cost function equal to the observation cost function, and the cost function gradient to the observation cost function gradient. For this setup, we believe the preconditioning does not need to be applied.

11. Fig 1: The way that the observations fit into this diagram doesn't make sense, since currently it implies the arrow coming out of the right side of the Simulated box passes information both to the right and left. Some separate arrows from the Observation box seem to be needed.

Two separate arrows has been added out of the modeled and simulated box to avoid the possible confusion in data flow direction. (See Fig 1 in the updated manuscript.)

12. Fig 2: Despite the caption, this doesn't really show how the CG method is implemented to anyone already not familiar enough with incremental 4D-Var to know that it likes between the "no" and Tangent linear model and involves an updated estimate to the (preconditioned) increment.

We added to two addition boxes (for residual vector and updated gradient respectively) between the tangent linear and adjoint model. This helps explain the data flow between the the TL an AD models in the inner loop. Also the 'exit' box is changed to 'inner loop converged' box to emphasize the condition to exit the inner loop. (See Fig. 2 of the update manuscript.)

13. p. 3, lines 11-13: This sentence is grammatically incorrect, the phrasing is confusing, and the conclusion is drawn weakly. What is the "potential" that online transport based inversion systems have demonstrated? At a bare minimum, add a reference (e.g., Grell and Baklanov, 2011) or remove that statement.

We chose to have the statement removed.

14. p. 3, line 18, p .6, line 3, and p .6, line 23: The original reference for WRFPLUS is Xiao et al. (2008) [DOI: http://dx.doi.org/10.1175/2008MWR2235.1]. The version you use (v3.6) includes the work by Zhang et al. (2013) [mentioned elsewhere in your text], and should be included in these references. Huang et al. (2009) specifically used WRFPLUS for 4D-Var in WRFDA, but did not develop WRFPLUS. Barker et al. (AMS, 2005) originally developed WRFDA (for 3D-Var). These two latter references should be used as references for WRFDA. Barker et al. (2012) is an update on software development for WRFDA, and doesn't even mention WRFPLUS. The appropriate references need only be given at the first mention of these particular models, and do not need to be repeated throughout as references to the entire model. The exception is when discussing a particular aspect of those works.

We really appreciate the referee to clear this up for us. The references are fixed.

15. Throughout, the author should use bold characters for vector notation (i.e., x, y, q, kco2). These would be particularly illustrative on p. 10, lines 21, 24, and 31 to indicate whether the denominator or numerator of @qco2/@kco2 is a vector**.**
All vectors in the equations and inline text have been changed to bold face characters. At place when a symbol can be either vector or scalar, it is kept as non-bold character.

16. p. 5, line 7 and p. 14, line 23. You mention that L-BFGS-B can be used to calculate the posterior covariance, which is true although robustness of this approach with regards to the initial inverse Hessian estimate is an issue when using this algorithm, see Bousserez et al., QJRMS, 2015. As mentioned elsewhere in our review and this manuscript, Lanczos CG can also be used to estimate posterior error due to the eigen decomposition (well documented). Thus, the ability to calculate posterior error estimations is not a valid distinction between these two. Further, calculation of posterior error is not included for either method in this work. So while this could be mentioned in the introduction or discussion of future work, the methods section should only refer to methods that are actually used in this work or ones that provide reasoning for why you used a particular approach.

We thank the referee for correct us on this issue. The text related to posterior covariance calculation between the two optimization schemes have been removed from the manuscript.

17. p. 5, the term "cost function" is used 13 times on this page alone. It is suggested to reduce "cost function gradient" to "gradient".
"Cost function gradient" has been change to "gradient" through the text except where full term is needed to avoid ambiguity.

18. A comparison of Lanczos CG and LBFGS-B based solely on cost function reduction and RMSE is not sufficient. The authors should be more instructive and explicit as to the tradeoffs between them. In regards to p. 6, line 20, and p. 14,

line 21: How much less memory does Lanczos CG require for your particular application, as a percentage? Is that a good reason for choosing it over L-BFGS-B in this case? Are there ways to reduce the memory requirements of each? The most accurate Lanczos CG algorithm requires storing all basis vectors and performing full reorthogonalization after each iteration. Do you include that step? If not, why? *This is a salient topic*, since you discuss the loss of conjugacy later in the manuscript. Also, how can Lanczos CG be adapted for parallel computation in a way that differs from L-BFGS-B? The name of an algorithm or a reference should be included. What are the respective wall-clock times of the two methods? Lastly, It's also not clear why one is more amenable to parallel programming than the other (p. 14, line 21), as both are sequential techniques, unless that is strictly a consequence of the aforementioned memory requirements.

(1) Regarding the comparison between L-BFGS-B and Lanczos CG, we added discussion about the memory requirement, and parallel implementation related issue in Section 4.
(2) Regarding the reorthogonalization. Yes the reorthogonaliation is implemented in the code used in the original optimization experiment. The referee is correct that the degradation of the Lanczos CG with increased inner loop iteration was not caused by the loss of conjugacy, but the error in the adjoint model (and thus the calculated gradient vector). After we corrected the adjoint model, the need for the second outer loop does not exist anymore.
(3) Walltime used in our experiment with the two schemes are documented and added in the text.
(4) References for the L-BFGS-B Fortran code (Algorithm 788) compiled in WRF-CO2 4DVar is added.

19. p. 6, line 20: Lanczos CG provides approximations of both the leading eigenvalues and eigenvectors (eigenmodes), not only the former.

Thanks. This statement has been corrected.

20. p. 7, lines 8-9: Does VPRM calculate fluxes at the grid-scale in every time step? You can scale fluxes whether they are provided online or offline.

We agree with the referee that the fluxes can be scaled whether they are from offline data files or calculated by online model (VPRM). Because running the VPRM model requires additional datasets (satellite derived vegetation indexes and land cover classification maps) and some parameter tuning, we choose to use offline CarbonTracker CO2 fluxes instead. Both methods are valid, but using the offline files allows us to focus on the core code development by avoiding some extra input data preparation.

21. p. 7, line 15: So a tagging scheme for source specific CO2 has been implemented as well? This might present an interesting feature for testing the adjoint sensitivities and 4D-Var system, or performing low-dimensional analytic inversions.

The present model code does not include tagging scheme for specific CO2 sources. But as the chemistry 4d variable in all three models (NL/TL/AD) includes separate variables for each individual fluxes, tagging scheme can be implemented with some minor code modification in the future development.

22. p. 7, line 32: Well, to be more precise about the wording, the adjoint is just the backward sweep, although for nonlinear systems it would need information from the forward sweep. I'm not sure what that information would be though, for a linear CO2 simulation.

The statement about the adjoint and forward/backward sweep has been revised.
Regarding the referee's question about what information from the forward sweep would be needed by the adjoint (the backward sweep): (1) Meteorology state variables: In the dynamical core, advection and diffusion of chemistry species are carried in each of the three sub-steps of the Runge-Kuta loop. At a given time step, at the start of the backward sweep, only the meteorology state variables at the last sub-step is available while all three all needed. This requires the forward sweep to save (push to local stack) the meteorology at each sub-step to be used in the backward sweep. (2) CO2 mixing ratio: as explained in our response to the referee's major comment #1, nonlinearity can be present in WRF-CO2 when the positive definite advection predicts the minimum CO2 mixing at a given grid cell to be negative and trigger the renormalization. When such nonlinearity occurs, cost function gradient depends on not only perturbation but also background value of CO2.

23. p. 10, line 29: The emission-normalized sensitivities can also be found by dividing the full sensitivities by the emissions. Two separate simulations are not required. You might say that you calculated them this way, but to suggest "this is done" in general by a specific approach is misleading.

We agree and the text has been revised to avoid misleading the readers.

24. p. 11, line 5: I realize this is more of a numerical demonstration than scientific result, but it is strange to define the adjoint forcing for tower observations to be at the surface rather than tower height, as in practice these types of measurements would have a greater emissions footprint (hence the rational for using a tower . . .).
We agree. We conducted new experiment with the improved code, and set the adjoint forcing at the 1st and 10th vertical level of the WRF model grid. Footprint figures (Figs 8 of the revised manuscript) are redrawn with the new simulation results. We also noted the difference of footprints between the adjoint forcing at the 1st and 10th levels, as the referee pointed out in his comment.

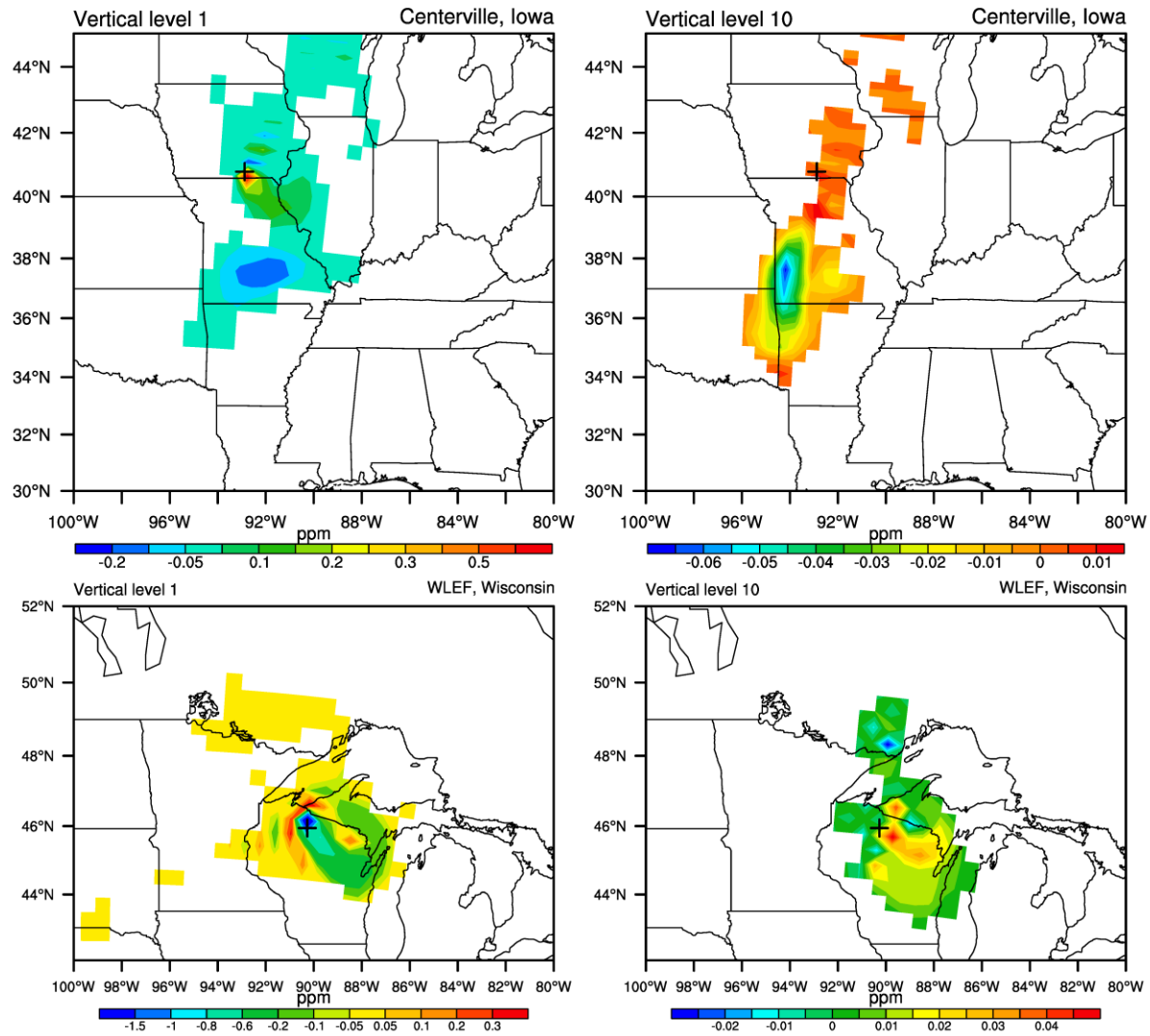


Figure 4. The adjoint sensitivities (footprint) of tower sites at Centerville, Iowa (top panels) and WLEF, Wisconsin (lower panels). At each site, the adjoint sensitivities are calculated twice: one with receptor placed a the 1st vertical level, and another at the 10th level.

25. Section 3.3: It seems like accuracy should be evaluated first, before presenting the sensitivity results in section 3.2.

The order has been switched so that the model accuracy evaluation is presented before the sensitivity spatial pattern.

26. Equation 8: What value used for delta_x? It can sometimes be difficult to find a perturbation value that balances truncation and roundoff error when using this equation to verify adjoint sensitivities.

The delta_x used in final calculation is 0.1. Prior to the final calculation for finite difference sensitivity, we conducted test using delta_x ranging from 0.01 to 1.0 to assess the impact of

magnitude of delta_x. The results indicate there is virtually no difference between 0.01, 0.1, and 1.0 regarding the finite difference value. We attribute this to the fact that CO2 tracer transport is linear. The impact of delta_x on the finite difference sensitivity is documented in Figure 5 below. This figure is included in the supplement document, but not in the manuscript.

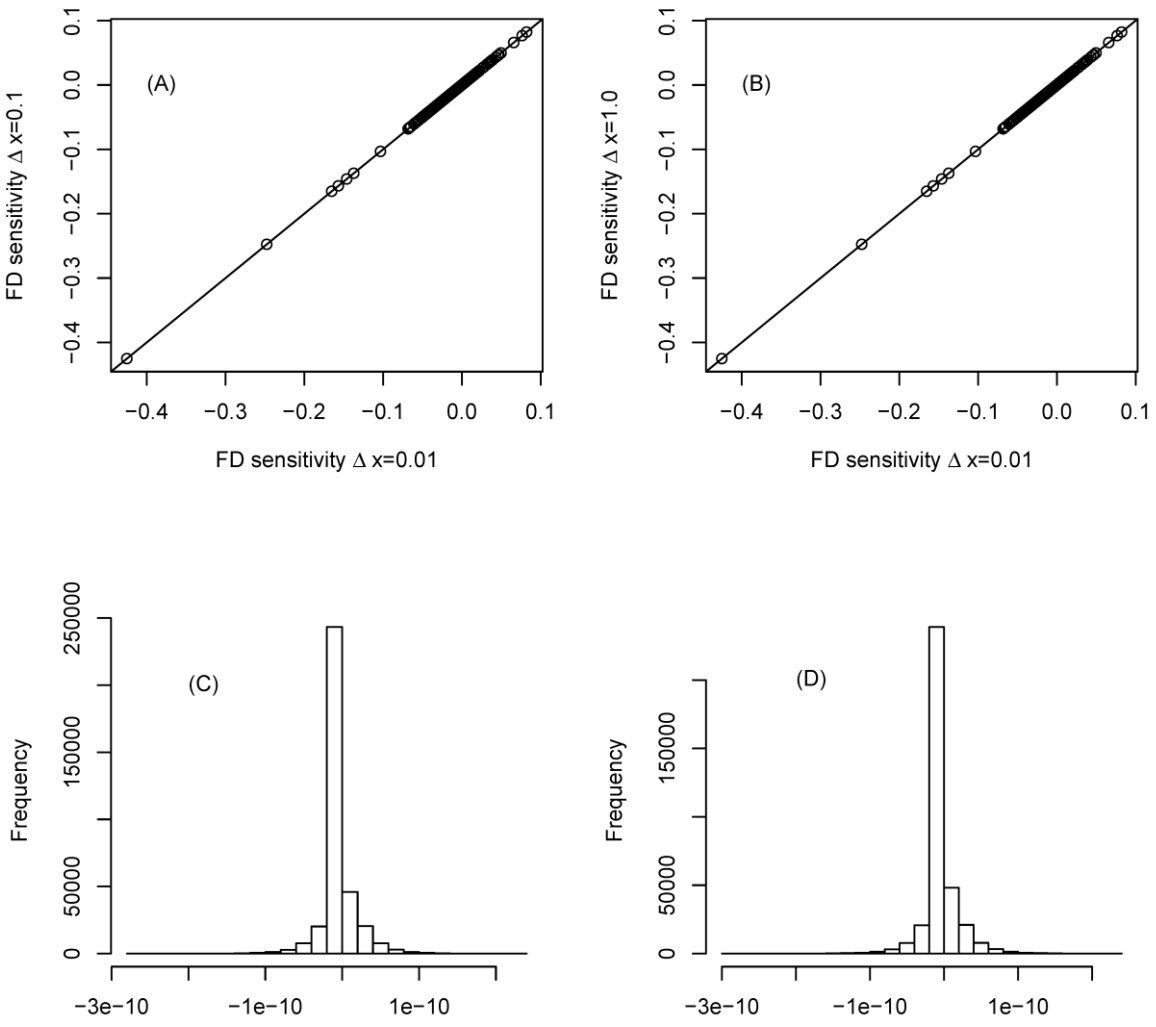


Figure 5. Impacts of the delta_x on finite difference sensitivity calculation. Finite difference of a source is calculated three times, with delta_x set to 1.0, 0.1, and 0.01 respectively. The results are compared grid cell to grid cell. Figure (c) is the histogram of difference in finite difference sensitivity calculated with delta_x=0.01 and delta_x=0.1. Figure (d) is the histogram of difference between detal_x=0.01 and 1.0. Both the scatterplots and histograms show the difference in finite difference caused by delta_x is negligible.

27. p. 12, line 23: By assuming B−1 = 0 and R = I, the pseudo-data case ignores how uncertainties will affect the convergence of Lanczos-CG and LBFGS-B. How would the performance of these two approaches differ with imperfect observations? With an unbiased prior? Determining the correct treatment of B and R is an active research area, which the authors do not address. Do the authors plan to explore more realistic covariance definitions in the future? At a minimum, this should be discussed in Section 4.

We agree with the referee that these are very important issues when applying the system to actual observation data. As the referee pointed out, treatment of background and observation error covariance is an active research area. Our objective of the present paper is to develop and test the TL/AD model and the optimization scheme. In-depth discussion on how the two optimization schemes perform with actual data is beyond the present paper's cope. But we did add a statement to remind the readers the nature of the pseudo-data based inverse experiments.

28. p. 12, line 32-24: Did the authors confirm a loss of conjugacy in the Lanczos basis vectors? Also, did the authors make any attempt to force conjugacy through full re-orthonormalization (e.g., Modified Gram Schmidt)? That mechanism is built in to release version 3.6 of WRFDA. While re-orthonormalization uses extra memory, that resource requirement is often very small relative to that of the model integrations. The authors should justify a decision that adds iterations to the optimization. After including full re-orthonormalization, the number of iterations for Lanczos CG to converge in each outer iteration should be proportional to the degrees of freedom (DOF) constrained by the chosen observations (see, e.g., Rodgers, 2000), entirely independent of the conjugacy issue. At that point, the necessity of multiple outer iterations would be caused by a nonlinearity in the forward model, possibly the PBL treatment or convective transport. The authors make no attempt to characterize such a nonlinearity that would necessitate using a nonlinear optimization strategy.

In the original text, we did not examine conjugacy. The loss of conjugacy was a mere guess, and turned out to be a wrong one. We really appreciate the referee pointing it out.
As we explained in debugging the adjoint model, the degradation of the incremental inner loop was caused by the inaccuracy of adjoint model as opposed to the loss of conjugacy.

Yes, re-orthonormalization is implemented WRF-CO2 following the WRFDA code. This means that loss of conjugacy was not possible, but we did not realize this while writing the original manuscript. This has been corrected in the revision.

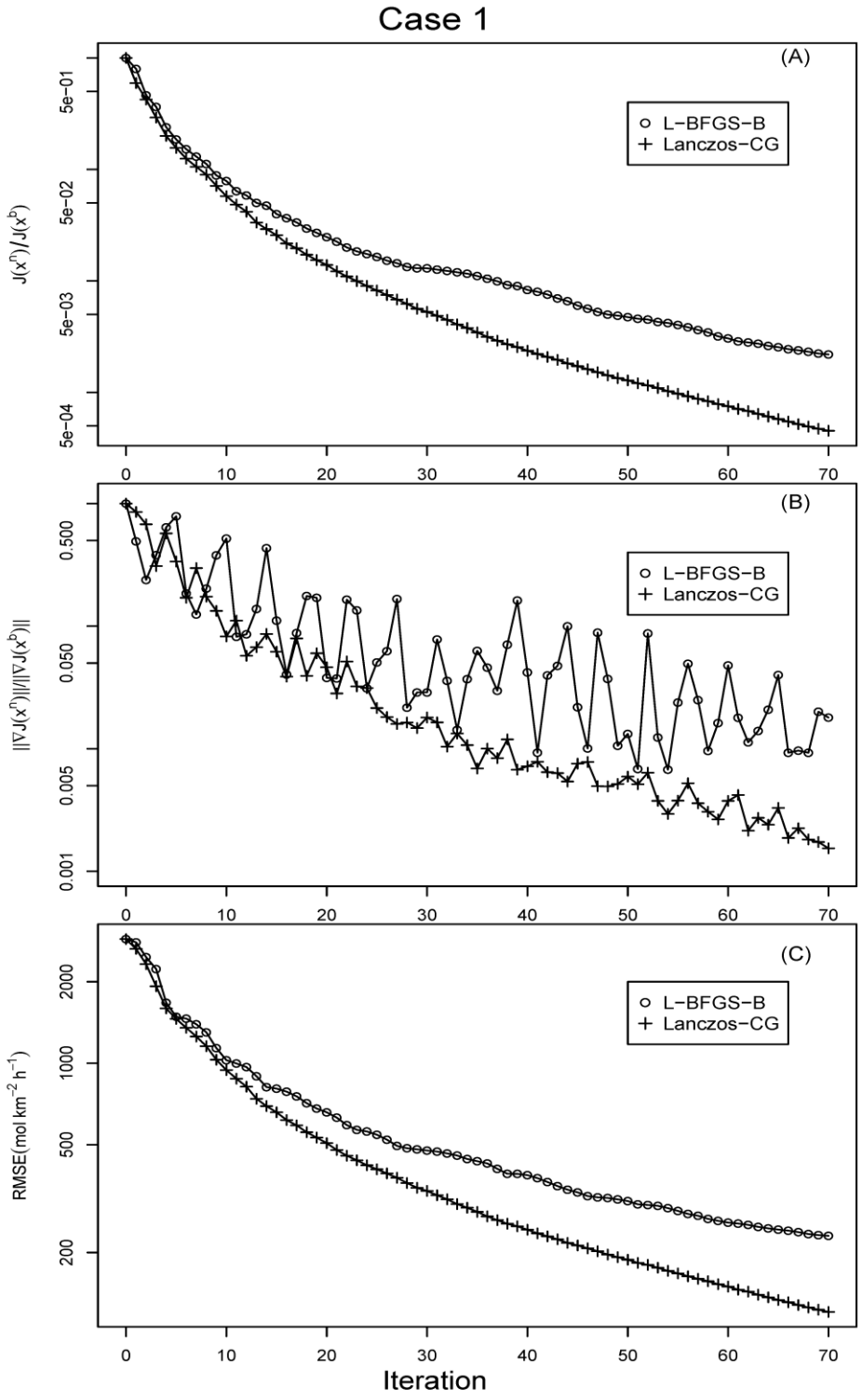


Figure 6. The results of the inverse experiment (Case 1). Reduction of the cost function and gradient norm are expresses as ratio to their respective starting value. Only one outer loop is used in the Lanczos-CG optimization.

29. Section 3.4: While interesting and valuable, numerically, there should be some statement with regards to the unphysical nature of the test setup, to emphasize that this is strictly a numerical test and not the expected level of performance (in terms of cost function reduction or RMSE) that would be obtained in a real inversion.

Thanks for pointing this out. We added description about the unrealistic nature of the experiment setup and that real observation data application will need more careful treatment of the errors.

30. In Section 4, the third paragraph needs a topical sentence. Also, the two sentences "We evaluated ... sensitivity." should be combined into one and made more concise.
A topic sentence has been added and the two sentences have been combined.

31. p. 15, lines 1-11: While mentioning the ability to use different aggregation techniques may illuminate a budding area of research to the reader, the details given should be both accurate and concise. The authors' discussion of smoothing and aggregation error (i.e., from Turner et al., 2015) are based in the assumption that no correlation is used in B. Taken out of context, this could be very confusing to the reader. Full non-ambiguous coverage of that topic would require more than a paragraph, but such a description is not appropriate for this section. Indeed, large portions of Section 4 (paragraphs 4, 5, and 6) ought to be rewritten or removed. Many of the references are out of date, and do not represent the state of the science.

The discussion about aggregation and errors has been removed in the revised manuscript, as they are not central to the present paper's objectives. In its place, we added detailed discussion about memory and computation requirement of the two optimization schemes.

32. p. 15, line 18-21: Other areas to improve upon would be more accurate treatment of data and model (i.e. transport and representational) errors in R, and error correlations in B, and posterior error estimation.

Thanks! These suggestions have been added in the revised manuscript.

33. Section 5: I'm not sure this meets the requirements of GMD, and may delay the publication of this work until the code is publicly available.

Accept. The source code has been submitted to zenodo at https://doi.org/10.5281/zenodo.839260
Future development will be made available to the public access in the same fashion too.

Technical Corrections:
1. Add appropriate punctuation to Eqs. 1, 2, 3, 4, 6, and 7.

Fixed.

2. p. 1, line 21: Remove “inversion” at the end of the sentence, as it is implied in the first half of this statement.
Fixed.

3. p. 2, line 16: “LDMZ” should be changed to “LMDZ”
It is corrected.

4. p. 2, line 17: Change “inverse” (noun) to “invert” (verb).
It is fixed.

5. p. 3, line 3: “LPDM” is undefined. Possibly define and change “Lagangian particle backward trajectory model” to “Lagangian particle dispersion model (LPDM)” on p. 2, line 33.
Definition for LPDM is added, and the Lagrangian particle backward trajectory model is changed to LPDM.

6. p. 4, line 12: “Where” to “where”
Fixed.

7. The first term in parentheses in Eqs. 3, 4, and 6 need transpose operators. Additionally, it would be less confusing if brackets and braces are used in addition to parentheses where warranted.
Thanks for pointing out the missing transpose operators. They are added. The two equation presentation has been improved by using brackets and braces.

8. p. 6, line 9: Correct “innoviation” to “innovation”
Fixed.

9. p. 6, line 11: Remove “Eq. 7”, since you are referencing the very next line of the text.
Fixed.

10. p. 6, line 20: Correct “lead” to “leading”
Fixed

11. p. 7, lines 12, 16: “inner” to “inert”
Fixed

12. p. 7, lines 14-16. Combine the two sentences that both state this category does not apply to CO2.
Fixed

13. p. 7, line 20: “(Zhang et al.,” to “Zhang el al., (”

Fixed

14. p. 7, line 30: Correct “simplied” to “simplified”
Fixed

15. p. 8, lines 12 and 16: Correct “inner” to “inert”
Fixed

16. p. 8, line 12: Correct “use” to “uses”
Fixed

17. p. 9, line 1: Change “chemistry vertical mixing” to “vertical mixing of chemical species”
Fixed

18. p. 9, line 3: “dynamical” to “the dynamical”
Fixed

19. p. 9, line 24: “set up” to “setup”
Fixed

20. p. 10, line 5: “simulation spans” to “simulations span”
Fixed

21. p. 10, line 7: “condition” to “conditions”
Fixed

22. p. 11, line 3: The reference should be to Gerbig et al. (2008). Also, use the correct parenthetical format for inline references.
The Reference is fixed.

23. p. 11, line 3: Change “footprint at a receptor” to “footprint of a receptor”
Fixed

24. p. 11, lines 9, 11, 12, 15: The figure references are off by 1.
Fixed

25. p. 11, line 18: Correct “no shown” to “not shown”
Fixed

26. p. 12, line 23: Correct “identify” to “identity”
Fixed

27. p. 12, line 28: Correct “facotr” to “factor”
Fixed

28. p. 12, line 30-31: Lanczos-CG is repeated twice. Also, use either “Lanczos CG” or “Lanczos-CG” throughout the document.
Fixed.

29. p. 13, lines 2 and 21: cost function needs an article, such as “the”
Fixed.

30. p. 13, lines 2-4: Add commas before and after “by the 30th iteration”.
Fixed

31. p. 13, lines 16-17: The opening to this sentence, “Starting at 2336.5 mol km-2 h-1,” is confusing or out of place.
The sentence has been rephrased.

32. p. 13, lines 24: change “the Lanczos” to “Lanczos” for consistency
Fixed.

33. p. 13, line 26: extra “the”
Fixed.

34. p. 14, line 8: Change “system” to “systems”
Fixed

35. p. 14, line 10: Modify, “Such configuration”, which is grammatically incorrect.
Text is modified to correct the grammatical error.

36. p. 14, line 11: Change “incurring” to “requiring”
Fixed