Response to Reviewer 2

We thank the reviewers for their constructive feedback on our manuscript (https://www.geosci-model-dev-discuss.net/gmd-2018-20/). The reviewers' comments are shown below in *italics* with our responses directly following.

Anonymous Referee #2

This work documents the workflow of STILT simulations and presents improved physical processes for fine-scale simulations. I appreciate the authors' efforts in addressing overdue problems for the community, in particular those who use STILT extensively. I hope that the authors continue updating their work through GitHub.

I can easily follow the method and think the paper is relatively well written given the conciseness in length.

Thanks for the positive comments. We hope that readers will agree.

I have some questions/concerns in the evaluation of the improved method. In current form, the authors do not characterize the errors, in particular in surface emissions. So it is hard to evaluate the results. The model evaluation is a key result in this study, and the authors need to describe how much they know (or pre- scribed) the errors in surface emissions (and others if prescribed) so that we can be sure that the better results from GWD are due to the improved schemes.

We have added a discussion regarding difficulties in estimating uncertainties in emissions inventories. Please see below for details.

Detailed Comments:

L13 - 21: STITL-R should be applicable to other tracer gases, not only CO2. The authors describe CO2 only, which seem to be strange. This is probably because the authors show an evaluation study using CO2, but this CO2 focus is limited.

STILT's applications certainly exceed only simulating atmospheric CO2. We attempt to describe the use of LPDMs and the STILT model (~p2L9, ~p2L20) using generalized language such as "atmospheric mole fractions", "pollutant concentrations", and model applicability to "observed emissions" and "surface fluxes". We use urban CO_2 as the primary motivation for several reasons: urban CO_2 cycling is the focus of a large and growing body of scientific literature that this model update will play a prominent role in, it allows for the use of novel CO_2 surface flux

inventories purpose-built for the study region (the Hestia model), and it applies well to the case study using the unique data available from the light-rail measurement system.

P2, L21: Need to cite older work about HYSPLIT.

We have added a citation for Draxler, R.R., and G.D. Hess, 1998.

P2, L28 - 29: Need to mention more recent work on city-scale or regional inversion work based on multiple receptors that uses STILT extensively. Literature review here does not represent a full range of the use of the traditional STILT, which I believe is import to for the reader to understand the context, and motivation for the new development.

We have added citations for McKain et al., 2012 and McKain et al., 2015 describing STILT modeling applications in Salt Lake City and Boston as well as Kort et al., 2013 describing STILT's use to assess measurement network design in Los Angeles.

P3, L6: Need to include the reference for R properly. Not doing so is irresponsible because without R this work is not possible.

Thank you for the suggestion. We have added the citation for the R software at ~p3L7.

P3, L20: For large-scale simulations, the users have applied other types of parallelizations in running STILT, e.g., running multiple jobs (each job may represent one receptor for a give period) at the same time taking advantage of high performance computing. The authors need to briefly mention what the difference between the old method and the one introduced here would be although the method described here seems to be similar to what users have been using. Is there a new concept here?

We recognize that we did not adequately describe past efforts to run parallel simulations. While the concept of executing batches of receptors across multiple jobs is not new, users have previously had to write and run separate scripts defining the receptors and relevant data inputs for each job which can require significant manual labor or develop their own methods for batch processing receptors. The manuscript formalizes methods for automatically executing the parallel batches of receptors, with receptor batches distributed between the parallel jobs and managed by the code itself rather than the user. The workflow presented, controlled with *run_stilt.r* and with output saved to simulation ID directories, remains the same for serial and parallel execution with only changing the setting for the number of parallel processes.

To clarify this point, the following text has been added to ~p3L28 :

> However, past methods for parallelizing simulations require users to manually define batches of receptors and relevant meteorological inputs in unique initialization scripts and submitting each script as a separate job to the scheduler. While increasing the

number of parallel threads decreases the size of each simulation batch, the requirements of the user become more complex.

We formalize methods for automatically distributing batches of receptors across many parallel threads managed by the model rather than the user.

P3, L27: Not all systems use SLURM although it is popular. Is there an option for a different job scheduling tool?

As of writing, SLURM is the only cluster job scheduler that has been implemented. SLURM is open source and utilized heavily by the high performance computing (HPC) systems at the University of Utah. Due to limited availability of HPC clusters, SLURM is the only job scheduler that has been validated. However, modifications to the project scaffolding described in this manuscript that facilitate parallel computation within single-node and SLURM-scheduled environments opens the doors to other queue managers as well. We encourage future collaboration with users who have access to these job schedulers and would be willing assist with testing development code on their systems. To clarify this in the text, we have added the following statement:

While SLURM is the only cluster job scheduler that has been implemented to date, the open source code can be modified to run on systems managed by other job schedulers including TORQUE/OpenPBS, Sun Grid Engine, OpenLava, Load Sharing Facility, or Docker Swarm using methods described by Lang et al. (2017).

P4. L4 - 22: In many cases, PBL heights from meteorological models (e.g., WRF) are directly used to represent z_pbl. The authors need to clarify this and describe more on the use of WRF PBL related to equations (1) and (2). For HNF simulations, WRF needs to be run at a similarly fine sale, which is really expensive? If not, what would be the impact on h = min(h',hˆ)?*

The formulation for the HNF vertical mixing depth adjustment h' is intended to fix systematically low footprints without needing to explicitly resolve z_{pbl} at HNF resolutions. It provides an estimate for the effective mixing depth based on homogeneous turbulence theory without requiring meteorological inputs (e.g. WRF) to be at a scale that explicitly defines the fine variations in PBL height within a city. However, the meteorological data are used outside of the HNF domain to calculate h* using a modified Richardson number method that has been extensively validated for the traditional "near-field" domain.

P5, L1-2: Reading this, my immediate thought was if this would require more simulation time to estimate the weighted influence. It would be nice to mention the cost.

Agreed. Calculating the footprint field using smoothing methods involves a cost tradeoff with a larger particle ensemble. While it is almost always less expensive to apply smoothing methods compared to calculating particle trajectories, quantifying the advantage is difficult. The cost to

calculate particle trajectories varies depending on model configuration, meteorological data source, the size of the meteorological domain, and the size of the ensemble while the cost to apply smoothing depends on the method and the spatial and temporal domain of the output footprint.

To clarify this point, the following text has been added to ~p5L1:

> Computing trajectories of large particle ensembles ($N > 10^4$) is computationally expensive. To lessen the cost of each simulation, footprint fields are often calculated from smaller particle ensembles by applying smoothing methods to compensate for the smaller ensemble size. These smoothing methods are less computationally expensive than calculating trajectories for a larger ensemble but vary in their ability to reproduce the robust footprint field of the large particle ensemble.

P6, L32: Should not include a paper in preparation.

Agreed. We removed the citation since the manuscript is still in preparation.

P7, L5: 24-h backward in time seems to be too short. How was the upstream boundary condition treated? I see a short description from L17. Boundary conditions are complex due to wind directions. Is the wind consistent from one direction? I would like to see a more description on this.

We find that particles exist within the footprint domain for 11 hours on average. The meteorological domain encompasses a larger area than the footprint domain and fluxes from outside of the footprint domain are assumed to be resolved by the background atmospheric signal described at p8L6.

To clarify this point, the following text has been added to ~p7l24:

> Urban development and expansion in the area surrounding SLC is limited by the mountainous topography surrounding the city and the Great Salt Lake which restrict the expansion of the city and suburbs. This confines large anthropogenic and biologic sources into a relatively small area surrounding the SLV and simplifies boundary conditions for SLV-centric modeling efforts. From each receptor, 24 h backward trajectories of 200 particle ensembles were calculated using meteorological fields from the HRRR model, available at an hourly interval with a 3 km grid resolution. On average, particles travel within the model domain for 11 h. Computation of the 33,608 particle trajectories and a single set of footprints completed in 5.5 hours utilizing 80 parallel threads across 5 nodes, each equipped with 64 GB of memory with two 8-core Intel XEON E5-2670 2.6 GHz processors. 6.7% of the simulations were not completed due to short-term outages in the HRRR data product.

P7, L30: Please use rˆ2 and state which method was used in calculating r. Pearson's method? How are these rˆ2 values statistically different? The simulations from GWD is distinguishably

from a different distribution from the other two so that we have more confidence in GWD? Note that in this evaluation, we want to clearly see better results from GWD. Right?

As recommended, we have modified the text to use r^2 instead of r to explain model variance and have clarified that it is based on Pearson's method.

While there is likely no statistical significance in the differences between GND and LEG for this case study, we show that GND agrees better with the physical "ideal" case and may give improvements that depend on the locations of differences between GND and LEG relative to the locations of surface fluxes. With the vertical dilution correction (GWD), the results agree more closely with measurements in both time and space.

P8: L1: I think this is probably the most important single statement in this paper. I would like to know how the authors determined the uncertainty in the surface fluxes. Without precise uncertainty characterization, the results are not reliable. What if the inventory is systematically low and GWD overestimated the mole fraction, which could be shown to be closer to the observations than the other two methods? I believe that the authors have considered this point, but I don't see the details here to the level that I can clearly see the outperformance of GWD. Also we need to note that the r^2 values are all low and similar to each other.

We agree that it is important to investigate uncertainty in inventory estimates. While we can show improvements to footprint smoothing algorithms using physically constrained "ideal" cases, uncertainty estimates within emissions inventories remains an unresolved question within the emission inventory scientific community. Developers of the Hestia inventory have documented that "a devoted effort is needed to generate uncertainty and propagate those uncertainties through the Hestia approach to provide an improved understanding of where results are more or less certain in space and time. This remains a high priority for future research" (Patarasuk et al., 2016) and determination of GHG fluxes and uncertainty bounds is one of the primary goals in the ongoing Indianapolis Flux Experiment (http://sites.psu.edu/influx/). Improvements to LPDMs can help future inverse modelling frameworks that would be better equipped to quantify uncertainties in flux inventories.

To further clarify this, we have added a discussion regarding the difficulties in assessing emission inventory uncertainties. Both of the inventories we discussed in the manuscript (Hestia and ODIAC) agree on the total emissions within the SLV domain which is evidence one inventory is not systematically lower than the other. However, mapping uncertainty to a moving receptor using two emissions inventories that encompass different spatial domains and allocate fluxes using different methods in time and space is a difficult question that requires more tools and analysis than are available in our present manuscript and should be the focus of future work.

The following text has been added to ~p7L20:

> Within the SLV domain where the inventories overlap, Hestia and ODIAC agree on the total anthropogenic emissions to within 1.5% during our study period. However, uncertainties of fluxes applied to our analyses are likely larger since the two inventories allocate fluxes differently in space and time. Further, only Hestia is used to represent the SLV whereas ODIAC is used outside of the SLV to account for regional-scale emissions. Uncertainties in inventory estimates are difficult to quantify in time and space and require a devoted effort within the emission inventory scientific community to propagate uncertainties through underlying assumptions within each inventory (Patarasuk et al., 2016; Lauvaux et al., 2016).

P8, L6: Please be more quantitative. It is not clear what has been reproduced. C3

We have changed the text to generalize that the model sees enhancements downwind from major roadways and introduced a caveat that better details the limitations regarding model resolution.

The following text was added to ~p8L21:

> The model generally produced mole fraction enhancements (ΔCO_2) for grid cells containing or downwind from major roadways (Fig. 7). However, modeling intersection scale enhancements would require finer grid spacing capable of resolving sub-city-block spatial scales that is not yet feasible given current constraints on inventories, meteorological data, and computing resources.

P8, L10 - 15: The simulated mole fractions are a combined result of transport and surface flux emissions. The authors, as mentioned, need to say how much we know about the surface emissions (used here) related to this discrepancy as well as the transport arguably improved from this work.

See comments relating to p8L1.