We thank the reviewers for their comments. Our responses are color-coded in blue.

We thank the authors for their detailed rebuttal and appreciate the substantial work undertaken to retrain the model and revise the manuscript. The revisions improve the clarity of the contribution. However, several issues remain that should be addressed before publication to ensure the claims are appropriately supported and not overstated.

Major points

1. Claims of FootNet outperforming STILT remain too strong.

Although the authors acknowledge in the rebuttal that FootNet cannot exceed the physical fidelity of STILT, the abstract still states that the emulator "out-performs STILT." This wording remains misleading because FootNet is trained explicitly to reproduce STILT outputs. Any differences relative to STILT are, by definition, deviations from the training target. If these deviations lead to a better match against independent observations, that is interesting but should not be framed as FootNet outperforming STILT. I strongly suggest softening the language in the abstract and throughout as agreed in the rebuttal.

This is an important point and we disagree with the reviewer here. Our final goal is to conduct GHG flux inversions and our primary metric of success is observations that indicate a successful GHG flux inversion. Our work demonstrates that FootNet matches or exceeds the performance of physics-based models **when used in a flux inversion** and was evaluated against independent GHG observations (see Figures 7 and 8). Again, our flux inversion comparisons are with respect to real independent observations in the atmosphere and, as such, FootNet **can** outperform STILT in such comparisons. We have updated the text in the manuscript to reflect this.

Line 10: Case studies using GHG measurements in the San Francisco Bay Area and Barnett Shale show that FootNet matches or exceeds the performance of physics-based models when used in a flux inversion and evaluated against independent GHG observations.

2. Clarification and consistency on the testing/validation datasets.

It is good to see that the authors have removed 2020 from the training dataset so that tests can be performed on unseen periods during 2020. However, this separation of the training and testing/validation datasets should be enforced throughout the paper. Furthermore, the distinction between validation and test datasets is still unclear. As written, they appear to be the same set?

The rebuttal states that the testing and validation set includes randomly sampled footprints from 2020 and 2021. If 2021 remains included in both the training and validation sets, there will be strong temporal and spatial correlations with those samples, even if footprints are from different receptor points. It would be preferable for the training/validation set to only use data from 2020.

Figures demonstrating performance (e.g., Figure 5) should also be updated to show only 2020, for consistency, and to minimise the influence of data leakage.

We thank the reviewer for this comment. The training, validation, and test datasets are distinct and share no overlapping receptors. Following standard machine learning practices, we randomly sampled training, validation, and test data from a distribution while ensuring

the sets are mutually exclusive, to maintain statistical consistency and prevent data leakage.

Figure 5 shows footprints computed by in-sample FootNet, which is trained on data from both 2020 and 2021. We have now added a supplemental figure (S7) showing footprints computed by out-of-sample FootNet for receptors sampled in 2020, while it was trained on footprints from 2021. We have also updated the Figure 5 caption.

Figure 5: Figure S7 shows a similar comparison for out-of-sample FootNet-predicted footprints for the year 2020.

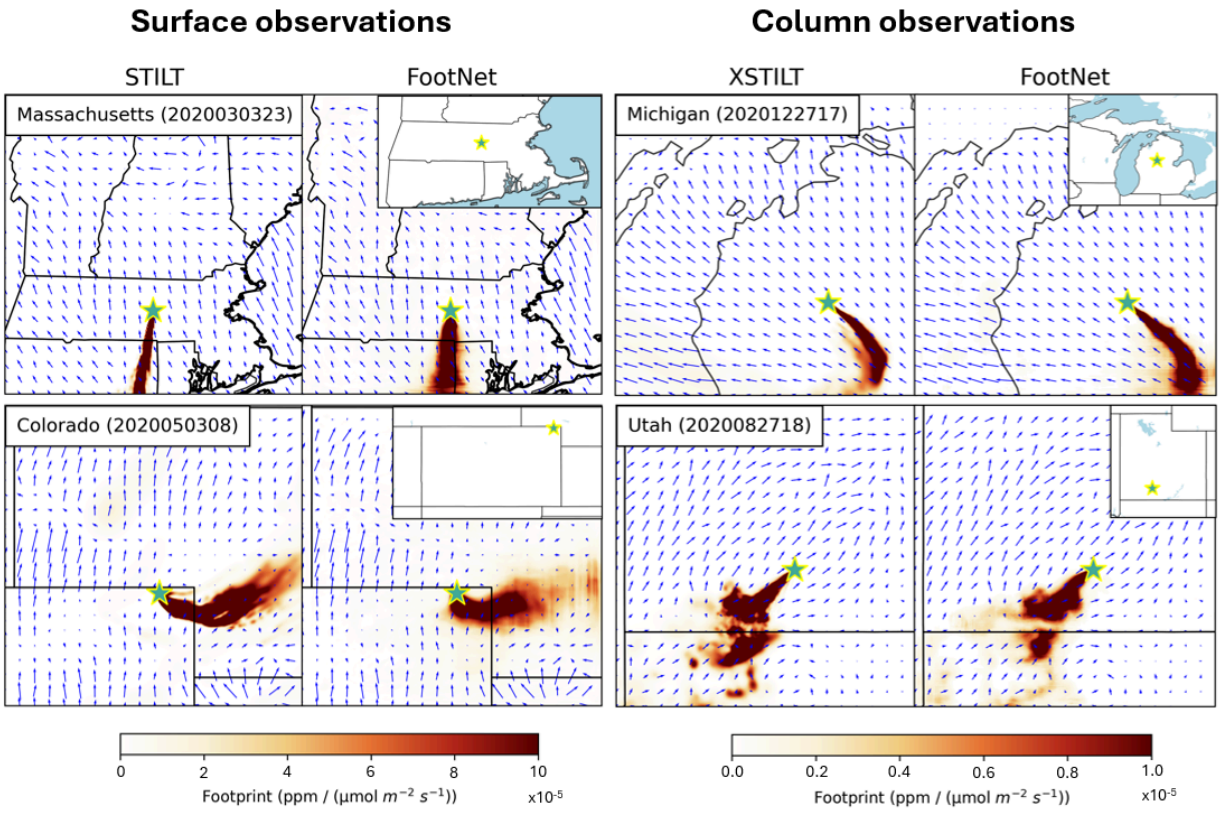

Figure S7: Footprints computed by out-of-sample FootNet for receptors sampled in year 2020. Out-of-sample FootNet was trained on footprints from 2021.

3. Clarification of the advance claimed by the meteorological dataset comparison (GFS vs HRRR).
The revision has added some minor statements to say that GFS and HRRR differ, and I think that it's useful that the authors have demonstrated that their system is relatively insensitive to the meteorological analysis dataset that was used. However, to my mind, the manuscript still over-sells what this comparison demonstrates. It seems to me that this comparison primarily shows that these analysed meteorological products are very similar to one-another. If they weren't, the emulation wouldn't work, since meteorology is the only input.

To address this comment, all that is needed is needed is for statements such as “despite being trained on HRRR, FootNet accurately predicts footprints using GFS” should be softened or not claimed as a particular advance of Footnet itself (e.g., personally, I wouldn’t claim this as a major finding in the abstract).

We thank the reviewer for this comment. As mentioned in our previous response, demonstrating that FootNet can perform well using GFS winds is an important extension to using this model outside CONUS, as HRRR is only available over CONUS. A common question we receive when presenting this work is: *“Can you use other meteorology?”* This comparison clearly demonstrates that we can. As such, it is important to keep this comparison in the text for practitioners who would use FootNet. We have now updated the text in the manuscript.

Line 8: We show that it produces consistent source-receptor relationships when driven by GFS meteorology, even though it was trained with HRRR inputs.

4. Further clarification that Footnet has not “learned the physics”
Related to Point 3, and despite the authors’ agreement with similar comments from the first round of reviews, several statements still imply that FootNet has learned "the underlying physics," when in reality it has learned a statistical relationship between meteorology and footprints. Phrases such as “learns the underlying physical relationship” or “learns the fundamental relationship” should be revised to avoid suggesting that physics-informed ML techniques were used. Removing the word “physical”, or “fundamental” in these contexts would resolve this concern.

While we agree that stating the model has “learned the physics” is confusing, we have clarified the text to state that the model has learned the underlying physical relationship linking the footprints to the meteorology. Further, as mentioned in our previous response, we enforce a penalty to conserve mass. This is exactly what a physics-informed neural network does. We prefer to leave the language as is because this is supported by the methodology, feature importance, and results.

We have gone back through the manuscript to ensure the language is consistent with the phrasing *“learned the underlying physical relationship”* rather than the former statement: *“learned the physics”*. Again, we feel that this phrasing is supported by the methodology, feature importance, and results.

Minor comments:
Figure 1: The legend now explains the coloring, but the notation for the convolution operations remains unclear. A concise description (e.g., “applied convolution kernel size”) would help.

We have now added the convolution kernel size to the Figure 1 caption.

Figure 1: We applied a 3x3 convolution kernel on the convolution layers.

Figure 2: Please clarify what the blue boxes inside Domain B represent.

We thank the reviewer for this suggestion. We have now added the description of the blue boundaries inside Domain B.

Figure 2: The blue boundaries in Domain B show counties in the Barnett Shale basin.

Figure 4 mixes the terms percentage footprint mass difference, percentage footprint sum difference, and normalized footprint mass error. Please use one consistent term and explain precisely what quantity is being plotted.

We thank the reviewer for pointing this out. We now use “percentage footprint sum difference” throughout the manuscript.

Line 94 still contains the term pseudo-observation. This should be replaced with “receptor” or “footprint,” consistent with the manuscript revisions.

We thank the reviewer for this comment. We have now updated the manuscript.

Line 92: Each receptor was simulated using STILT(Lin et al., 2003; Fasoli et al., 2018) for surface footprints and X-STILT (Wu et al., 2018) for column-averaged footprints, using NOAA High Resolution Rapid Refresh (HRRR) meteorology at 3 km resolution regridded to 1 km.

Line 129: The text states that the full training dataset is essential. Can you provide justification that a reduced dataset (e.g., 100,000 samples) is insufficient?

We thank the reviewer for this comment. While validation loss begins to plateau beyond approximately 100,000 samples, we found that models trained on smaller subsets exhibited degraded performance in complex terrain and coastal regions. Using 500,000 samples ensures uniform skill across the diverse meteorological and geographic conditions represented within CONUS.

As an example of this, the model trained using only data from 2021 has 170,000 samples. This model exceeds the 100,000 sample criteria we identified and performs well. This is a confirmation of the analysis mentioned. More training data is always preferable, but 100,000 samples *should* be sufficient for training a FootNet model. We have now updated the text.

Line 124: For generalizable inference with uniform skill across the diverse meteorological and geographic conditions represented within CONUS, the full training set remains essential.