

Dear Reviewer,

Thank you for carefully reviewing our manuscript and for your helpful comments. We have revised the manuscript and respond to each comment below. Your comments are shown in blue, our responses in normal font and the corresponding revised text in grey.

Reviewer 2

Wang et al. present a generative ML framework to emulate the STILT footprints. The framework generates an ensemble of footprints from different stochastic ML realizations, applies mass conservation screening to the STILT footprints, and performs an OSSE to evaluate if GenGHG could recover prescribed methane fluxes. It is interesting to see the model evaluated against DeterGHG across different dispersive and terrain regimes, under cases with/without STILT mass violations, and within an OSSE for flux inversion. GenGHG generally shows improved performance relative to DeterGHG across these cases, demonstrating the potential value of the generative framework.

Comment 1: My major comment lies in the lack of evaluation of temporally resolved footprints. The current evaluation appears to focus primarily on time-integrated footprints. Have the authors considered evaluating the ability of GenGHG to reproduce the temporal evolution of footprints, for example at hourly resolution? Temporally resolved footprints can be particularly important for applications involving rapidly varying or bidirectional surface fluxes, such as biogenic CO₂ exchange, where daytime uptake and nighttime respiration have opposite signs. Evaluating whether GenGHG preserves this temporal structure would help establish its applicability to a broader class of inverse problems.

Response. Thank you for this important suggestion. We have revised the Discussion to place our work in the context of related studies and to clarify that the present study focuses on time-integrated footprints for satellite snapshots of anthropogenic emission hotspots. We also acknowledge that this representation may not be sufficient for other applications. The revised text recognizes the importance of time-resolved footprints and identifies their generation as an important direction for further model development. We appreciate this recommendation and view it as a valuable next step, while retaining the focus of the current evaluation on the application for which GenGHG was initially developed.

“(Lines 424–431) Second, GenGHG currently generates footprints integrated over a 24 h transport window, a setup used in several urban-scale studies of anthropogenic emissions, particularly for satellite overpasses or snapshot observations (e.g., Wu et al., 2018; Ohyama et al., 2023). We therefore focus on this application. More complex applications, including continuous ground-based observations and rapidly varying or bidirectional surface fluxes such as biogenic CO₂ exchange, may require time-resolved footprints that retain contributions from individual transport intervals (Gourdji et al., 2010; Sargent et al., 2018). Existing ML work has used exponential-decay allocation as a practical strategy for reconstructing temporal structure from time-integrated predictions (Dadheech et al., 2025). Future work should assess whether such strategies can be adapted to GenGHG while preserving transport at each time step and across the full window.”

Comment 2: A more general suggestion is to provide a clearer discussion of the use cases or ranges of applicability for GenGHG, which would be useful for users. In particular, it would be helpful to discuss the spatial and temporal scales over which GenGHG is expected to remain reliable. The current applications appear relatively near-field and are especially relevant to ur-

ban or source-region studies. Within what transport distances, footprint ages, or domain sizes would the authors recommend using GenGHG? Would the current framework also be suitable for regional or global inversions, or would additional training data and/or model development be required? If aiming for regional and global scales, it would be useful to discuss how the framework would handle background concentrations and trajectory endpoints.

Response. We appreciate this comment, which helped us define the applicability of GenGHG and directions for future development more clearly. We revised the latter part of the *Discussion* to describe both the current scope of GenGHG and potential development pathways.

- We specified the near-field scope of GenGHG and outlined the additional development required for regional or continental modeling.

“ (Lines 414–423) The current GenGHG configuration has clearly defined applicability boundaries. The following discussion outlines these boundaries and the corresponding strategies for future model development.

First, GenGHG is designed for near-field transport from localized emission hotspots, covering approximately 300 km over a 24 h backward window. It represents local enhancements above the large-scale background and should not be applied directly at regional or continental scales. Extending the model to these scales would require additional training data covering larger meteorological domains and longer transport histories. The generative framework could be retained, but the neural network architecture would need adaptation. The transport module could operate on a coarser grid to capture large-scale transport, while the subgrid adapter could be omitted where fine-scale reconstruction is unnecessary. These changes could extend the framework to regional source–sink analyses (Hu et al., 2019) and enable explicit treatment of background concentrations and particle trajectory endpoints (Göckede et al., 2010; Karion et al., 2021). ”

- We clarified the time-integrated nature of the current footprints and discussed the value and challenges of developing temporally resolved footprints, as detailed in our response to Comment 1.
- We further discussed global applications and the additional data and model development needed for regions with complex topography.

“ (Lines 435–439) Another case is complex terrain, as mountainous regions are represented by fewer quality-controlled physical references in the current benchmark. Addressing this limitation will require expanding the benchmark with simulations driven by nested, higher-resolution meteorology beyond the current 0.25° forcing. Region-specific training or fine-tuning, with topography and land–sea masks included as static inputs, could further improve emulation in regions such as Taipei and Los Angeles. ”

Comment 3: Some of the terminology used in the flow-matching description may be difficult to interpret for atmospheric science readers. In particular, terms such as “target velocity” and “prior samples” could be defined more explicitly. Does “velocity” refer to velocity in the generative/data space rather than atmospheric wind velocity?

Response. Yes, the velocity refers to the change in the latent footprint state along generative time, not to atmospheric wind. We revised *Section 2.2* to explain the generation method more clearly by explicitly defining the prior, generative time and latent-space velocity, thereby avoiding ambiguity for readers.

“ (Lines 116–128) **Generative training and ensemble prediction.** During training, GenGHG learns a generative path that connects a prior distribution to the target data distribution of footprints (middle panel of Fig. 1). Each generation path pairs a prior sample h_0 with a target endpoint h_1 . The two states have the same dimensions and are both defined in the latent space, as detailed below under **Neural network**. The

elements of $\mathbf{h}_0$ are sampled independently from an isotropic standard Gaussian distribution, $\mathbf{h}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This prior therefore imposes no prescribed spatial pattern or transport structure on the generated footprint. For a normalized generative time $t \sim \mathcal{U}(0, 1)$, the path evolves from $\mathbf{h}_0$ at $t = 0$ to $\mathbf{h}_1$ at $t = 1$, with the interpolated state $\mathbf{h}_t = (1 - t)\mathbf{h}_0 + t\mathbf{h}_1$. Because this path is linear in t , its velocity $\mathbf{u}_t$, defined as the rate of change of the latent footprint state, is constant along the path:

$$\mathbf{u} = \mathbf{u}_t = \frac{d\mathbf{h}_t}{dt} = \mathbf{h}_1 - \mathbf{h}_0, \quad t \in [0, 1]. \quad (1)$$

Under the FM formulation, GenGHG uses a neural network parameterized by θ to predict the velocity field $\mathbf{u}_\theta(\mathbf{h}_t, t; \mathbf{c})$ from the interpolated state $\mathbf{h}_t$, the generative time t and the meteorological condition $\mathbf{c}$. The network is trained to match the target velocity $\mathbf{u}_t$ using the following loss function:

$$\mathcal{L}_{\text{Gen}} = \mathbb{E}_{\mathbf{h}_1, \mathbf{h}_0, t, \mathbf{c}} \left[\|\mathbf{u}_\theta(\mathbf{h}_t, t; \mathbf{c}) - (\mathbf{h}_1 - \mathbf{h}_0)\|_2^2 \right]. \quad (2)$$

”

Comment 4: I did not quite follow the explanations for the differences between GenGHG and DeterGHG. What do the authors mean by “biased structural collapse after transformation”?

Response. Thank you for identifying this unclear phrase. By “biased structural collapse after transformation,” we meant that converting the single DeterGHG estimate from logarithmic to physical space can contract its value range, whereas averaging GenGHG ensemble members in physical space better preserves the skewed footprint distribution. We removed this phrase and revised *Section 3.1.1* to state this explanation directly.

“(Lines 277–281) These results show that GenGHG better preserves structural consistency with the physical reference by learning a conditional distribution rather than a single estimate. Two factors explain this difference. First, transforming a single DeterGHG estimate from logarithmic space back to physical units can contract the value range. GenGHG samples multiple members from the learned distribution to obtain an ensemble mean in physical space, reducing the distortion caused by transforming strongly skewed data between logarithmic and physical spaces.”

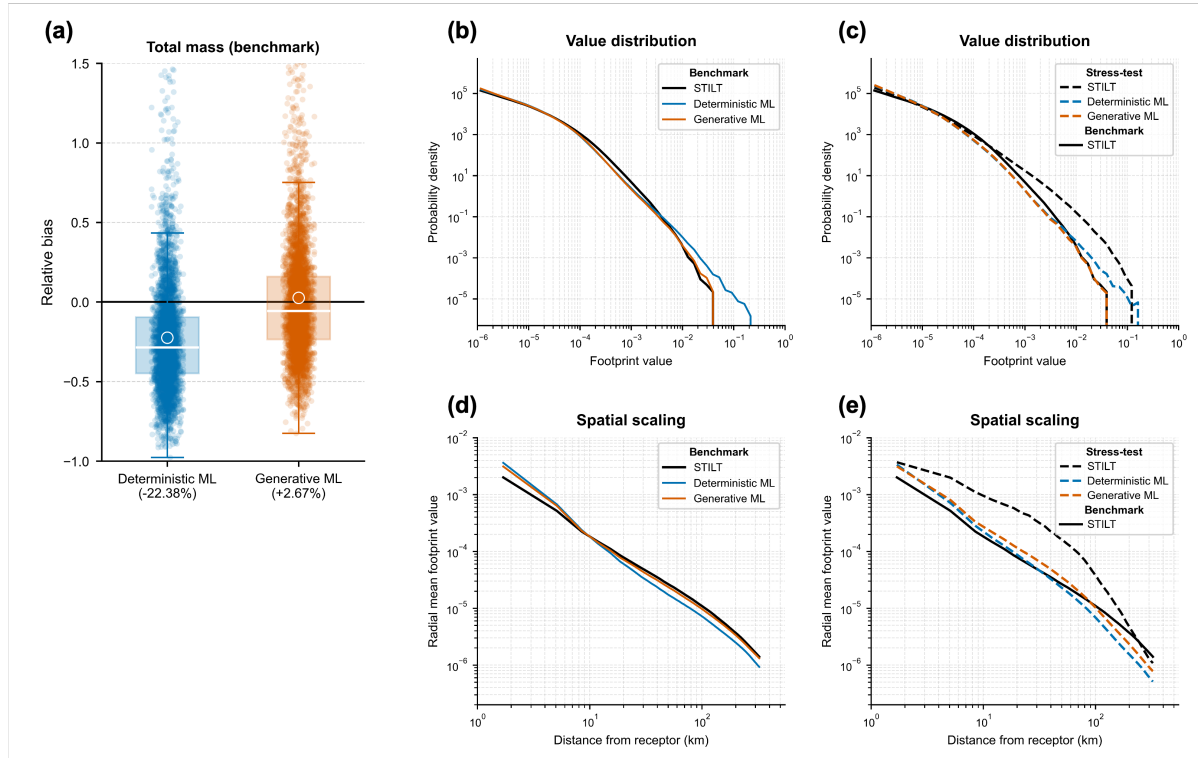
Comment 5: My understanding is that GenGHG was trained using STILT footprints that already passed the dmas screening. If so, it seems unfair to compare the ML results against badly-fail footprints. It would be helpful to add the screened STILT footprint line in Figure 3 next to the dashed lines for the badly-fail cases. The text also appears to suggest that GenGHG learned the mass violation from unfiltered footprints. Please clarify the training data and avoid overstating performance.

Response. Your understanding is correct. Only STILT footprints that passed the dmas screening were used for model training and benchmark evaluation. We agree that our original presentation could have been interpreted as an accuracy comparison against the flagged STILT simulations, which was not our intention. Instead, this experiment was designed as an out-of-distribution stress test to examine whether the ML emulators retain the footprint characteristics learned from the quality-controlled benchmark under conditions associated with STILT mass violations. Because the flagged simulations are not reliable physical references, we now restrict the interpretation to aggregate distributional behavior. We also revised Fig. 3c and e to show the benchmark STILT distributions alongside the stress-test results. The key revisions are summarized below.

- We clarified in *Section 2.4.2* that only screened benchmark footprints, not the mass-violating stress-test cases, were used for training.

“ (Lines 200–203) Only simulations classified as “pass” were retained in the benchmark for GenGHG training and evaluation. Simulations classified as “badly fail” formed a stress-test set in which STILT exhibits severe mass-conservation violations. We use this set to evaluate whether GenGHG preserves the footprint structures learned from the benchmark under these extreme conditions. ”

- Following your suggestion, we revised Fig. 3c and e by using dashed lines for all stress-test results and solid lines for the benchmark STILT results.



“Figure 3. Footprint total mass, value distribution and spatial scaling. (a) Relative bias in total mass across benchmark test footprints. Points represent individual footprints, boxplots summarize the distributions and circles indicate means. (b,c) Grid-cell footprint-value distributions for the benchmark and stress-test cases, respectively. (d,e) Receptor-centered radial mean footprint values as a function of distance for the benchmark and stress-test cases, respectively.”

- We clarified in *Section 3.1.1* that this comparison assesses overall distributional behavior rather than the accuracy of individual footprints, and shows whether GenGHG retains the characteristics learned from the quality-controlled benchmark.

“ (Line 287–294) By contrast, the ML emulators preserve the footprint-value distribution and radial scaling learned from the benchmark. Although their predictions diverge from the stress-test STILT simulations, they remain closer to the benchmark STILT distribution, with GenGHG retaining an advantage over DeterGHG in reproducing the footprint-value distribution.

This comparison evaluates overall distributional behavior rather than the accuracy of individual emulated footprints, because the flagged STILT simulations are not reliable physical references and the benchmark contains no case-matched alternatives. Nevertheless, the results show that the ML emulators retain the total-footprint characteristics learned from the quality-controlled benchmark when applied to challenging cases, including those in which STILT exhibits mass violations. ”

Comment 6: The degraded performance for highly dispersive footprints is interesting. Could this partly reflect limitations of the assumed Gaussian prior distribution used to initialize the generative process, particularly when the target footprint is very broad or multimodal? Some discussion of whether the prior distribution contributes to the difficulty in these cases would be

useful.

Response. Thank you for this suggestion. The Gaussian prior is isotropic in the latent space and neither represents a Gaussian plume nor prescribes a compact or unimodal footprint in geographic space. Therefore, our current experiments cannot identify the prior choice as the cause of degraded performance in highly dispersive cases. Nevertheless, clarifying this distinction helps avoid misunderstanding, and we made the following revisions.

- We clarified the generative path and Gaussian prior in *Section 2.2*.

“(Lines 119–120) The elements of h_0 are sampled independently from an isotropic standard Gaussian distribution, $h_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This prior therefore imposes no prescribed spatial pattern or transport structure on the generated footprint. ”

- We added a cautious interpretation of highly dispersive cases in the *Discussion*. Based on the spread-to-skill evaluation introduced in *Section 3.1.3*, we recommend using ensemble spread as a case-specific indicator of potential emulation error when interpreting these samples.

“(Lines 432–435) Third, although GenGHG is globally applicable, its predictions under certain conditions should be interpreted cautiously. One such case involves highly dispersive footprints produced by complex transport processes. Although GenGHG uses a common noise prior across all conditions rather than distinct priors based on manually defined transport regimes, we suggest using the ensemble spread as a case-specific indicator of potential emulation error. ”

Comment 7: I am curious about the effects of dmas screening under different meteorological inputs. Does the dmas screening result in more or fewer usable footprints relative to GFS?

Response. Our results suggest that complex terrain may be a more important determinant of dmas screening outcomes than the choice of meteorological product, as shown in *Section 3.1.4* and Fig. 6a. However, the current experiments do not establish how passing rates vary across products because the cross-product analysis in *Section 3.3* is an out-of-benchmark inference test and does not include the product-specific STILT simulations needed for dmas screening.

Comment 8: The manuscript seems to miss several citations/references, for example for FLEXPART and NAME when they are mentioned.

Response. We have added the appropriate references accordingly.

“(Lines 171–173) Several LPDMs can be used for atmospheric dispersion and footprint estimation, including the FLEXible PARTicle dispersion model (FLEXPART) (Stohl et al., 2005) and the Numerical Atmospheric-dispersion Modelling Environment (NAME) (Jones et al., 2007). ”