

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 1

The article describes an innovative method of emulating LPDM footprints, leveraging flow-matching based generative models that have been shown successful in meteorological and other applications. I believe this paper is of excellent quality, and almost ready for publication.

The central idea is well motivated, proposing a generative model that builds on documented limitations and strategies of existing deterministic emulators. Accelerating LPDM emulation is a timely task to leverage the large amounts of satellite measurements. The results are evaluated through a nice range of metrics, and compared against the outputs of a deterministic version of the model as a benchmark. The dataset is also a contribution in its own right, given the large amount of locations and dates included. The synthetic out-of-sample inversion experiments illustrate well the downstream application, and the cross-product comparison is a useful demonstration of the potential of a fast emulator where the full physics runs are prohibitive. There are three main areas that I believe need a bit more development:

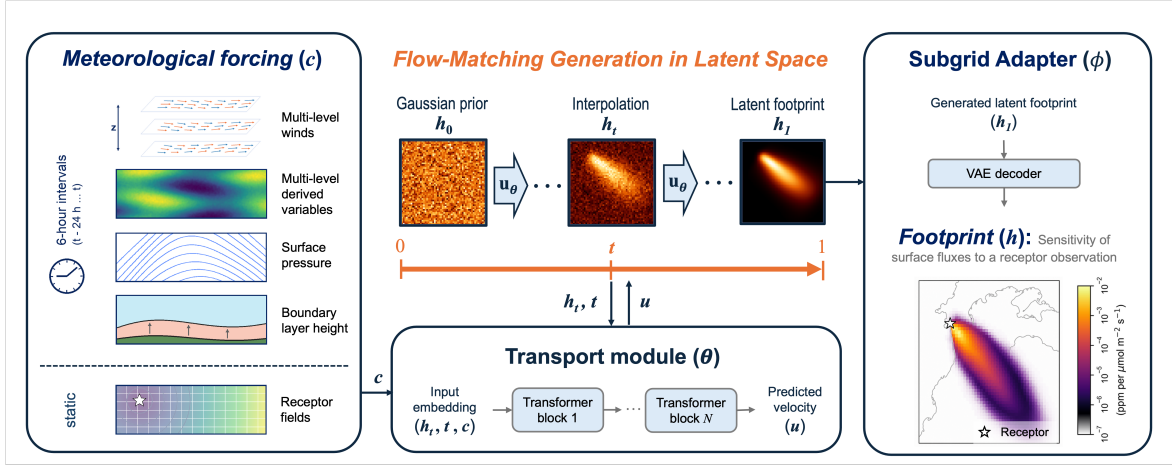
Response. We thank the reviewer for the positive assessment of our manuscript and for recognizing the relevance of this work.

Comment 1: VAE Subgrid-Adapter: The VAE is mentioned in the text, and the architecture described in the Appendix. Figure 1 shows the GenGHG framework but does not show the VAE at all. The figure should include it if it aims to show the complete framework. I am also missing some more details about training (e.g. is the VAE trained with the same dataset as the transformer module?) and more importantly, an evaluation of the VAE outputs (i.e. metrics that evaluate the quality of the reconstructed footprints, and maybe some examples). These metrics will characterise the contribution to the error in the emulated footprints from the VAE reconstruction rather than from the emulation itself. The need for the VAE itself should also be better justified.

Response. We agree that the role and performance of the VAE-based subgrid adapter required clearer documentation. We have clarified its function and training and provided evaluation results for this module. The revisions are listed below.

- We added the VAE-based subgrid adapter to Fig. 1 right part and clarified its purpose and training in the Method.

“(Lines 144–148) The subgrid adapter (right panel of Fig. 1) bridges the resolution difference between the coarser meteorological forcing fields and the finer native footprint grid. Specifically, the adapter reconstructs the latent footprint from the 0.25° grid onto the $1/32^\circ$ grid, which is sufficiently fine to resolve anthropogenic emission hotspots in spatial detail. This design allows the primary transport module to generate footprints in a substantially lower-dimensional space, thereby reducing the computational cost of training and prediction.”

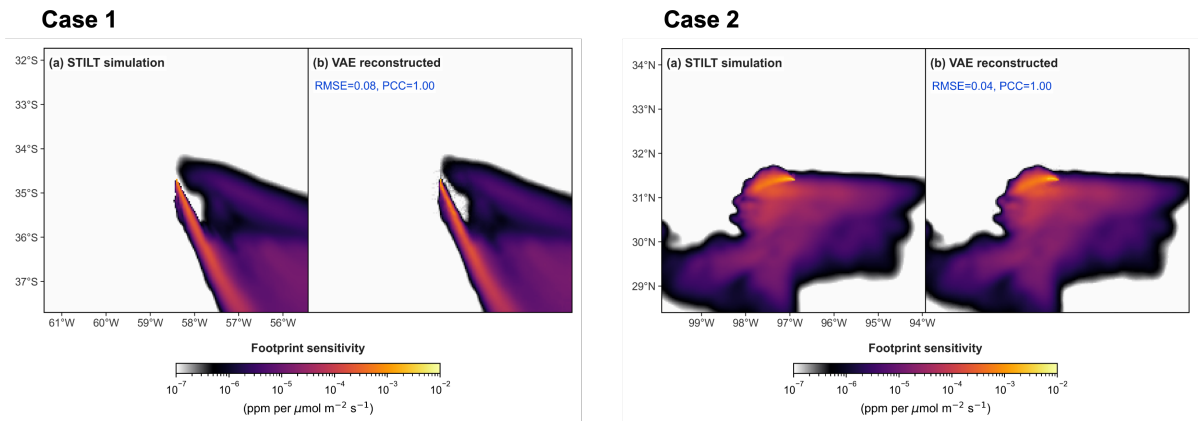


“Figure 1. GenGHG framework for generative footprint emulation. Conditioned on meteorological forcing fields, the transport module moves a Gaussian prior toward a target footprint through flow-matching generation.”

- We made evaluation of the subgrid adapter a prerequisite for the subsequent detailed analysis of the transport emulator. We used the same evaluation metrics and benchmark test partition as for the transport module. The independent reconstruction evaluation shows that the adapter contributes little to the overall emulation error, and two visual examples are provided in the Results and Appendix Fig. 10.

“ (Lines 257–260) We evaluated the subgrid adapter independently to isolate the reconstruction error introduced when mapping latent outputs to the footprint grid. On the benchmark test set, the VAE decoder achieved an MSE of 0.003, an RMSE of 0.053, an MAE of 0.170 and a PCC of 0.999. These results indicate that the reconstruction step contributes little to the overall transport emulation error. We therefore focus the subsequent analyses on the performance of the primary transport module. ”

“ (Lines 308–310) For both cases, the subgrid adapter shared by GenGHG and DeterGHG reconstructed the grid-level footprints with little degradation (Fig. A1). The comparison in Fig. 4 therefore isolates transport emulation. ”



“Figure A1. Latent-to-grid footprint reconstruction by the VAE. For each of the two cases shown in Fig. 4, panels compare (a) the STILT- simulated footprint with (b) its VAE reconstruction after encoding into the latent representation and decoding back to the grid.”

Comment 2: Ensemble evaluation: GenGHG addresses the drawbacks of deterministic models, which tend to produce smoother fields, by introducing a probabilistic element and generating 30-member ensembles for each sample. The evaluation focuses only on the mean of these

members (which is itself smoother). The authors mention that future work will investigate how the spread among ensemble members relates to uncertainty, but I believe that some metrics that characterise the ensemble would improve the quality of this paper (e.g. spread–skill ratio, CRPS). Further work might also evaluate performance under different numbers of members.

Response. We appreciate the suggestion to include probabilistic metrics in the manuscript. We added a separate subsection evaluating the probabilistic skill of GenGHG’s flow-matching design, extending the evaluation from the ensemble mean to the full predictive distribution. GenGHG achieved a CRPS of 0.199 and an SSR of 0.829. We also compared flow matching with two perturbation-based methods to provide context for its relative performance. The revisions are listed below.

- We introduced CRPS and SSR in *Section 2.3* and provided their detailed formulation definitions in the *Appendix B*.

“ (Lines 157–160) Probabilistic performance is evaluated using the continuous ranked probability score (CRPS) and the spread–skill ratio (SSR), as detailed in Appendix B. CRPS assesses the predictive distribution as a whole, with lower values indicating better probabilistic skill. SSR compares the ensemble spread with the realized prediction error, with values close to 1 indicating well-calibrated uncertainty. ”

- We introduced two probabilistic baselines in *Section 2.3* and described their perturbation principles and methodological sources in *Appendix C*.

“ (Lines 165–168) We also compare GenGHG with two probabilistic baselines that use the same ML backbone but introduce randomness at different stages to generate ensemble predictions. One perturbs the meteorological input conditions, whereas the other perturbs their encoded features in latent space, as detailed in Appendix C. Comparing these perturbation-based ensembles with FM generation in GenGHG isolates how the ensemble construction method affects probabilistic skill. ”

- The metric results and comparisons are detailed in *Section 3.1.3*, and Figure 5 shows performance across all test samples.

“ (Lines 316–327) We compared GenGHG with two probabilistic baselines to determine which stochastic design best supports ensemble transport prediction (Fig. 5). The ensemble means showed similar deterministic skill, although the FM design adopted in GenGHG achieved a slightly lower RMSE and a higher PCC. The differences were clearer in probabilistic performance. FM achieved the lowest CRPS (0.199, compared with 0.226 and 0.233) and the highest SSR (0.829, compared with 0.528 and 0.465). Unlike deterministic error metrics, CRPS evaluates the full predictive distribution by balancing the errors of individual members against their pairwise spread. The lower CRPS therefore shows that the greater diversity among GenGHG members improved probabilistic skill without sacrificing accuracy. Its ensemble spread also more closely matched the realized footprint error, although the ensemble remained slightly underdispersive, with a spread smaller than the realized error. By contrast, perturbing either the meteorological inputs or their embedded features produced more similar ensemble members and substantially underestimated the realized error. These results identify multistep FM generation as the most effective approach for producing stochastic footprint ensembles. The ensemble spread can therefore provide a case-specific indication of potential ML emulation error, allowing predictions with greater spread to be interpreted more cautiously.”

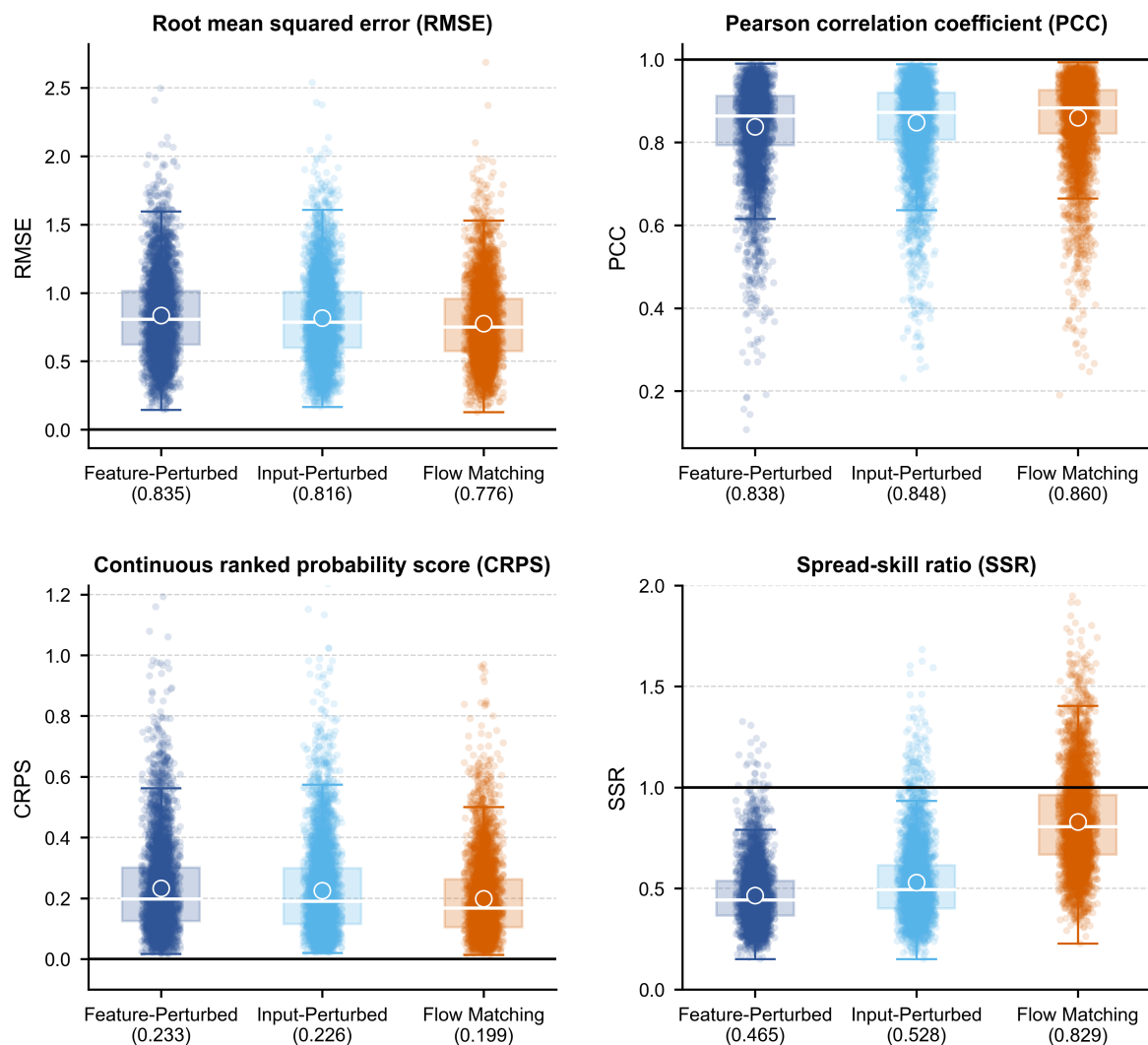


Figure 5. Deterministic and probabilistic skill of the three ensemble methods. Black horizontal lines mark the ideal value for each metric (0 for RMSE and CRPS; 1 for PCC and SSR)."

Comment 3: The mass violation filter: filtering the dataset by the STILT dmas threshold is an interesting choice. The footprints in the “pass” category have a significantly different distribution from the ones in the “badly-fail” category, and it is reasonable to treat the “badly-fail” ones as a stress out-of-distribution test. However, it is difficult to demonstrate that the emulated footprints for those “badly-fail” samples are accurate. The paper also does not show that they could be used within an inversion without distorting the results, or suggest a method for filtering them out from future emulated datasets. In the synthetic inversion example, were any footprints labelled as fail/badly-fail, and were they used within the inversion? The filter also leads to a sampling bias in mountainous regions—are there mountainous regions that are fully or mostly filtered out of the dataset? I do not expect newly trained models or inversions, but the discussion of this choice could be more aware.

Response. Thank you for this careful assessment. We agree that flagged STILT simulations cannot validate the accuracy of individual emulated footprints. We therefore confirm that the synthetic inverse modeling experiment uses only screened STILT footprints, frame the flagged cases only as an out-of-distribution stress test, and expand the discussion of prediction diagnostics and terrain-dependent sampling bias. The detailed revisions are listed below.

- We clarified that only screened STILT footprints form the reference transport operator in the synthetic inverse modeling experiment.

“(Lines 235–237) The reference operator H_{true} was constructed from STILT footprints driven by GFS meteorology that passed the mass-conservation screening, thereby excluding identified mass-conservation issues from the physical reference. ”

- We revised the analysis in Section 3.1.1 to clarify that it evaluates overall distributional behavior rather than providing an accuracy reference.

“(Lines 291–294) 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. ”

- We added ensemble spread as a case-specific diagnostic of potential emulation error in *Section 3.1.3*, supported by the new SSR evaluation.

“(Lines 326–327) The ensemble spread can therefore provide a case-specific indication of potential ML emulation error, allowing predictions with greater spread to be interpreted more cautiously. ”

- We quantified the terrain-dependent effect of screening in *Section 3.1.4* and Fig. 6a. In the *Discussion*, we identified the limited representation of mountainous regions as a benchmark limitation and outlined strategies for addressing it.

“(Lines 331–333) The fraction of STILT simulations passing the mass-conservation screening decreased markedly from coastal lowlands (87.5%) and inland plains (89.1%) to mountainous regions (18.3%) (Fig. 6a). ”

“(Lines 435–438) 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.”

Other questions the authors might consider addressing:

Comment 4: Why were topography or land–sea masks not included as an input? Does the model receive any label for the city, or coordinates?

Response. Thank you for raising this useful question. GenGHG’s static inputs contain only the distance from each surrounding source grid cell to the receptor, providing relative geographic information; the model receives neither a city label nor absolute geographic coordinates. The current version omits explicit location-specific variables, such as topography or land–sea masks, to support transfer to unseen global locations beyond the 60 cities selected for our benchmark. We agree, however, that topography and land–sea masks can benefit region-specific model development, particularly over complex terrain. We therefore added to the Discussion that future region-specific training or fine-tuning could evaluate these variables as static inputs.

“(Lines 438–439) 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 5: How would this algorithm perform at larger scales (e.g. regional or continental) rather than the demonstrated super-resolution? Would any changes to the architecture be needed?

Response. Thank you for raising this point. The current version of GenGHG is designed for local emission hotspots. We revised the manuscript to clarify this scope and discuss the adaptations required for regional or continental scales. In *Discussion*, we clarified the current spatial scope and discussed potential development at larger scales.

“(Lines 416–423) 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).”

Comment 6: Are the per-grid-cell metrics calculated for all grid cells, or only for non-zero cells?

Response. The metrics are calculated across all footprint grid cells. Because both false negatives (reference footprint > 0 but prediction $= 0$) and false positives (reference footprint $= 0$ but prediction > 0) occur, all grid cells are included. We have made this evaluation domain explicit in the current *Section 2.3* to clarify the relevant evaluation details.

“(Lines 153–154) Each metric is evaluated per-sample across all footprint grid cells, and then averaged over all test samples.”

Comment 7: What is the inversion state vector size?

Response. Its 32x40 grid boxes for each daily emission task, 9 days in total. We have added the state vector information in *Section 2.6*.

“(Lines 234–235) For each daily inverse model run, the state vector s comprised 1280 flux elements on a 32×40 latitude–longitude grid. Each receptor footprint was mapped to this grid to form one row of the corresponding transport operator.”

Comment 8: The tests with multiple meteorology products would be better supported if the resulting footprints were compared against a sample set of STILT runs driven by one or more alternative meteorological products. Currently, the mass difference between products mixes real forcing differences with unknown emulator error.

Response. We agree that cross-product differences combine forcing-dependent transport variation with ML emulation error. In the revised paper, we moderated the interpretation and avoided attributing these differences entirely to product-specific effects, while retaining an indication of their overall magnitude.

“(Lines 377–381) Using GFS-driven STILT footprints as the base reference, the ECMWF-driven predictions exhibit higher footprint total mass, with relative biases of +9.63% for ERA5 and +10.14% for IFS. These differences combine product-dependent transport variation and ML emulation error. Under the assumption of comparable emulation error across products, the cross-product differences exceed the mean GenGHG bias under GFS forcing (+2.67%), suggesting that meteorological forcing may contribute more to transport-pattern variation than the ML approximation itself.”

Comment 9: How long did it take to generate a STILT footprint, and how long did it take to

generate a full ensemble for a single sample?

Response. Thanks for mentioning this. A STILT footprint (moderate particles for a column receptor, using 3000) takes 14 min and a 30-member ensemble for our GenGHG takes 3.3 seconds. We have added this information at the beginning of *Section 3.1*, before the subsequent transport emulation evaluation.

“ (Lines 251–256) Before assessing transport emulation performance, we first examined two practical prerequisites for GenGHG: computational efficiency and the accuracy of the auxiliary VAE-based sub-grid adapter. GenGHG reduced the computational cost by more than 250-fold relative to STILT. On an NVIDIA RTX 3090 graphics processing unit (GPU), generating a complete 30-member ensemble with 50 integration steps took 3.3 s using batched inference. By comparison, a STILT simulation (with a moderate number particle count of 3000 ensemble) took approximately 14 min on an Intel Xeon Silver 4210 Central Processing Unit (CPU) at 2.20 GHz. ”

Comment 10: How are the input embeddings derived? You state that the meteorological inputs are first embedded into 512 feature channels before entering the transformer module, but do not describe the architecture.

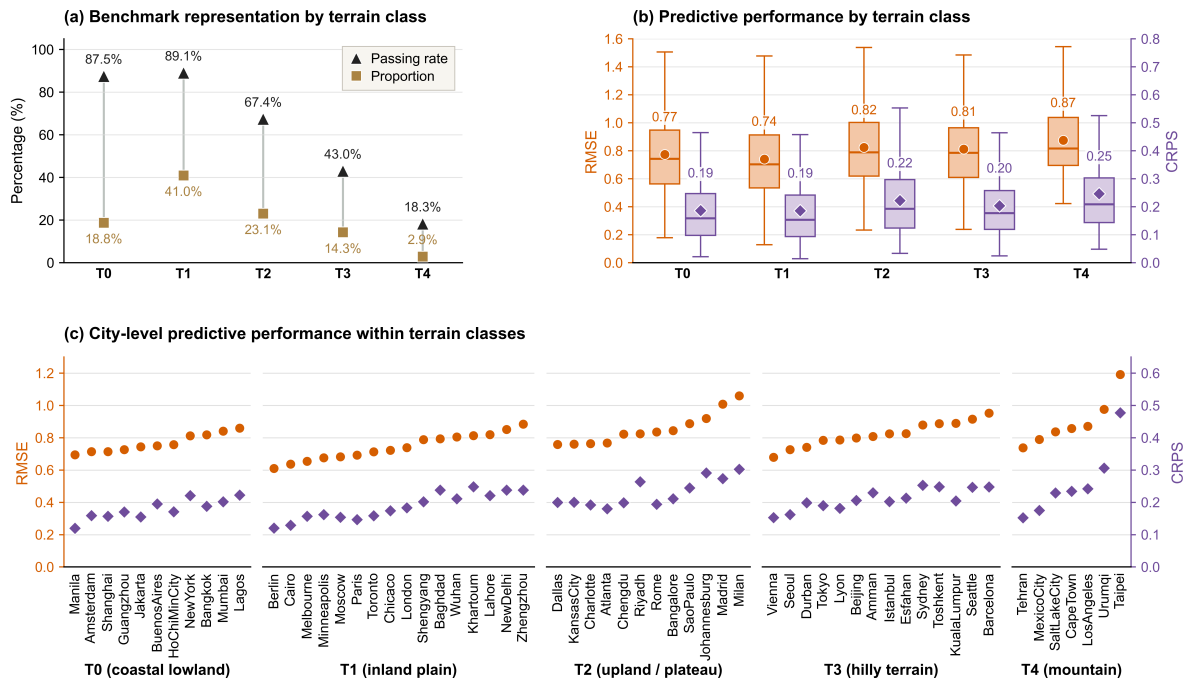
Response. They are derived by convolutional neural network blocks. To present the input-embedding method more clearly, we have expanded *Appendix A* with a clearer description of the input-embedding architecture.

“ (Lines 445–450) **Input embeddings.** Separate convolutional blocks first project the latent footprint state h_t and meteorological forcing conditions c to 512 feature channels. The meteorological features are then reweighted by a squeeze-and-excitation block (Hu et al., 2018) to capture interdependencies among the input variables and combined with an embedding of the normalized generative time t , thereby conditioning the model on the current generation stage. This condition embedding is finally concatenated with the footprint-state embedding and linearly projected before entering the transformer. ”

Comment 11: How does performance vary across hotspot locations?

Response. We appreciate this question, as location-to-location variability is important for assessing GenGHG across the global benchmark. We therefore organized the hotspot-level evaluation into a dedicated *Section 3.1.4*, covering all 60 benchmark cities grouped into five terrain classes. We use RMSE and CRPS to characterize deterministic and probabilistic performance, respectively, and added Fig. 6c to show city-level variation within each terrain class.

“ (Lines 337–343) City-level results revealed additional variation within each terrain class (Fig. 6c). Taipei (RMSE 1.192, CRPS 0.477) and Milan (RMSE 1.059, CRPS 0.302), both low-elevation emission hotspots located near mountainous terrain with pronounced elevation gradients, showed the largest errors. Berlin (RMSE 0.610, CRPS 0.1206) and Cairo (RMSE 0.637, CRPS 0.129), both situated in broad plains, showed the smallest errors. Across all cities, we found no consistent performance gradient between coastal and inland locations, between hemispheres, or with latitude. Together with the performance differences associated with footprint dispersiveness (Table 2), these results show that the challenges to GenGHG emulation arise mainly from local transport complexity and the surrounding terrain. ”



“Figure 6. Benchmark representation and GenGHG performance across terrain classes. (a) Percentage of STILT simulations passing the mass-conservation screening and the resulting proportion of the quality-controlled benchmark in each terrain class. **(b)** Distributions of RMSE (orange) and CRPS (purple) across test samples grouped by terrain class. **(c)** City-level RMSE and CRPS within each class; symbols indicate class means.”

Typos and smaller comments

Comment 12: The paper uses LDPMs instead of LPDMs at several points; please correct this.

Response. We corrected this terminology throughout the manuscript.

Comment 13: Gerbig 2003a and Gerbig 2003b are the same citation. This is probably a mistake.

Response. We corrected the duplicate citation.

Comment 14: “The primary transport module first encodes” should replace “first encode”.

Response. We corrected this grammatical error.

Comment 15: The Code and Data Availability section is missing its heading.

Response. We restored the section heading.