Reviewer comments

We thank the reviewer for their helpful comments. Below, for each reviewer we provide a response to each comment and note changes to the manuscript. Points raised by the reviewers are in italics. All line numbers refer to the updated manuscript.

Reviewer 1

General comments:

Fillola et al. developed a machine learning method, GATES, to simulate Lagrangian footprints. They demonstrated the efficiency of such an approach compared to the traditional physics-based LPDM models to fast generate footprints and do an emission inference using satellite GHG data. With the increasing amount of GHG data, these ML-based methods could be increasingly useful in this area. Surprisingly, the input data required to do these ML emulations is much less than those required in traditional LPDM footprint simulations. But they can still generate footprints that have an acceptable quality as the physics-based footprints. For example, in the NAME simulations, it requires 15 met variables at 50 levels at over 200 timesteps for 30 days, whereas the ML emulation only needs 12 hours of met information with 7 levels (only 3 levels for wind information). Does this mean that the traditional LPDM footprints may not need to simulate for 30 days? Instead, they could only run footprints backward in time for 12 hours to generate an acceptable-quality of footprints?

Data-driven models learn a mapping between a set of inputs and the corresponding outputs. The mapping skill is linked to how well the inputs characterise the outputs, with more complete inputs often providing better skill. However, large input datasets might make it more difficult to identify the relevant variables, and will require significantly more computing memory. The strong correlations present between the meteorological variables across time and height mean that a lot of the information present in the dense dataset is redundant for footprint emulation, and that a subset can be used without sacrificing skill. Here, we determine that seven levels and three meteorological timesteps provide GATES with sufficient information to produce footprints. Adding further information might improve accuracy in small amounts, at a large computational cost, although further extensive tuning could highlight specific inputs that present significant improvements. We note that the wind speed and direction is represented in two ways – as x- and y- vectors, provided at seven levels, and as speed and angle, provided at three levels.

Unfortunately, this does not tell us much about how to run an LPDM, which does not learn a mapping, but instead uses the dense meteorological inputs to accurately simulate physics at each timestep (e.g., it will still need to simulate small-scale temperature and wind speed gradients to accurately parameterise turbulence).

It is also unclear to me that, in the traditional LPDM-based inversions, back-trajectories simulated from the LPDM models are often used to estimate background before inversions. But in GATES, they don't emulate back-trajectories. Does this mean that GATES cannot be used alone for inversions, as they don't have estimates of back-trajectories, which are generally required to estimate background? If both GATES and NAMES need to be run for inversion applications, this would not save computation time.

Yes, this is indeed a limitation of the current version (although we note that there are plenty of examples in the literature where boundary conditions are hard-wired in the inversion, or are estimated

directly from the data, so that modelled boundary conditions are not used). We decided to focus the paper on footprint emulation, isolating the impact of the GATES footprints, without introducing an additional or separate emulator. We are working on an extension that will emulate boundary conditions too, and will publish a technical note on this aspect soon.

We have added the following to the text to clarify:

From line 411: Added a brief explanation of the inversion set-up, including the following clarification

"In this work, GATES emulates only the footprints – the boundary sensitivities are used as generated by NAME throughout, to isolate the effect of the emulator."

Line 481: Added the following to the Discussion

"Emulating the boundary sensitivities is a key step to fully integrate GATES into an inversion pipeline. These boundary sensitivities are used to calculate the background concentrations, incoming from outside the domain. In this work, we use the boundary sensitivities generated by the LPDM throughout, isolating the effect of the column-averaged footprint emulator. Future work will address this limitation, developing a boundary condition emulator. "

Specific comments:

Lines 63 - 64: the satellite observations often got coarsened, no matter whether they are used in Lagrangian-based inversions or Eulerian inversion models.

The reviewer is correct in their observation. We have edited the paper to be more specific, saying instead "Two recent studies that use Lagrangian methods for TROPOMI observations over Alaska and Siberia propose methods to coarsen (Thompson et al., 2025) or subsample (Ward et al., 2025) the dense observational dataset, so that it is feasible to run LPDMs."

Lines 79: why 2D? Many LPDM simulations have 3D footprints, including the third dimension in time.

The reference study by Tunnicliffe et al. (2021) uses footprints that aggregate surface interactions over the 30-day simulation period for each observation. They do not produce time-disaggregated footprints, which would take a significant amount of memory to store. We emulate these footprints. The footprint produced for each observation has dimensions latitude-longitude and therefore can be considered 2-dimensional. We have added the following text (here in bold) to clarify:

Line 77: "The model records whenever these particles are near the surface (within 40m in the simulations used here) **throughout the whole time period**, creating an aggregated 2D "influence footprint" **for each observation** that indicates the contribution of a unit surface flux at a particular location to the observed mole fraction."

Lines 81 – 83: this is true except for the XSTILT model

We believe that the X-STILT model, like Ganesan et al. (2017), also calculates LPDM releases for different atmospheric levels, and average them using the satellite's averaging kernel. E.g. see the following from Wu et al. (2018): "To represent the air arriving at the atmospheric column of each OCO-2 sounding, we release air parcels from multiple vertical levels, "column receptors" (Fig. 3e), using the same lat/long coordinates as the satellite sounding at the same time and allow those parcels to disperse backward for 72h."

Lines 91 – 93: HYSPLIT is pretty well-known and has lots of users as well.

The HYPSPLIT model is indeed very well known – we have added the following to the text:

"Other examples of well-known LPDMs that have been applied to similar applications are the FLEXible PARTicle Dispersion Model (FLEXPART (Pisso et al., 2019)), the Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT, (Stein et al., 2015)) and the Stochastic Time-Inverted Lagrangian Transport Model (STILT, (Fasoli et al., 2018))."

Lines 148 – 160: the subpanel labels seem wrong in this paragraph, as there is no f or g in Fig. 1. Also, Fig. 1 is not easy to understand for non-ML experts on what exactly the model contains or how the model is formulated.

We thank the reviewer for spotting the mis-labelling. The references to the figure labels have now been corrected throughout the paragraph. The architecture figure has been designed following common graphics in Machine Learning-specific papers and models.

Table 1: how do you determine the 7 and 3 levels? Can you use 6 levels or 10 levels? Also, how do you determine which level to choose?

A small amount of tuning demonstrated that this configuration produced good skill while reducing computational usage. It is very likely that the model would produce good results with 6 or 10 levels too, with skill changed based on the information contained by each level. For example, our tuning demonstrated that the range of level achieved better performance than only surface levels, or only upper levels. We have added the following in line 213: "The vertical levels and the number of timesteps were decided through tuning, where a small range of hand-crafted configurations were tested."

Lines 224 – 229: Validation: why can't you tune the model hyperparameters during training periods? Also, why do you need to apply an additional bias correction after training using validation datasets? Does this mean your training did not do a good job?

The model is validated against a separate dataset, to assess performance in unseen data. A common method for tuning in ML is cross-validation tuning on subsets of the training data, but this is a more expensive approach and computational resource was limited.

We have added figure B2 in the appendix, showing the distributions of the test data values, the emulated values, and the corrected distribution. The raw outputs of GATES present a high number of background values, rather than zeros – this is because ML regression models don't predict exact values. The bias in the training data values themselves, where there is only a small proportion of high sensitivities, means that the model might struggle to represent these. The bias correction is a simple statistical method to adjust the data distribution which significantly reduces underprediction of sensitivities.

Line 236: Unclear how you treat footprints with zero values, as the log transform of 0 is negative infinite.

The text addresses this in lines 425-247, but we have added the text in bold to clarify: "During training, **the non-zero values in** the footprints are log-transformed and shifted by the minimum value in the training dataset, so that all non-zero values in the original footprints $\boldsymbol{y}$ remain strictly positive in the transformed footprints $\boldsymbol{y'}$**.** The zero values are maintained as such."

Section 4.3.2: what does the size mean? The number of nodes in each layer in the GNN? How do you decide the number of hidden layers you will need in the model?

Yes, it refers to the number of nodes in each layer. We have added the following in lines 254-257 to clarify: "Here, size refers to the number of neurons in each layer. The size of the layers and the number of the hidden layers was decided through tuning, testing a number of configurations. Deeper networks or with higher number of nodes led to the model over-fitting to the training data, whereas shallower networks achieved low skill as they were unable to learn complex dispersion patterns."

Line 257: what does the learning rate of batch size mean? Can you provide more context on this?

The learning rate and the batch size are two separate parameters, often present in many machine learning models to regularise training. The wording the reviewer refers to was unclear. We have updated the text to include a brief explanation of the role of each parameter:

Line 270-272: "This training objective was minimized with the ADAMW (Lam et al., 2023; Loshchilov and Hutter, 2019) optimizer and a learning rate of 5×10^{-5}, which controls the size of the parameter updates during optimization. We use a batch size of 5, which determines how many samples get processed together in each update during training, regularising the learning."

Section 4.4.1: if the threshold is set by validation set, does this mean you will need to retrain your data with the test set after setting grid cells with values lower than this threshold 0?

(addressed with the answer below)

Section 4.4.2: same as my question above, do you need to retrain the data with the test set after bias correction?

The raw footprints generated by GATES are then post-processed using statistical methods, without needing further training of the model. The two statistical methods, the thresholding and the bias correction, are calibrated on the validation set, and then applied to the test set. We have added Figure B1 in Appendix B to illustrate the process for a particular footprint. We have added the following to the text in line 281 to address this explicitly: "GATES is not retrained after thresholding and bias correction: these two statistical steps are applied as post-processing of the already emulated footprints. Figure B1 in Appendix B illustrates the steps for a particular footprint.

Line 280: 1000x faster after training, right?

Yes, we are calculating ≈1000x faster by comparing the 20 minutes it takes NAME to generate the footprints, compared to the ≈1s it takes GATES. The training time, of ≈10 hours, is negligible compared to the LPDM running time. We acknowledge that the dataset itself would have taken a significant amount of compute to generate – we leverage an existing dataset, generated for a scientific study over a region of interest (Tunnicliffe et al., 2021), to avoid having to generate new dat.

Line 433: LPDMs, not "LDPMs".

This has been corrected, thank you.

Reviewer 2

Major improvement

To me, the vertical component of the study is quite novel and an innovation over previous work by this team. The skill to create vertical gradients of trace gases transported from the surface with GATES is something I would like to see demonstrated, especially for a complex area like this. I realize that in this application the footprints are kernel-weighted averages over height to make total column sensitivity, but this is likely to mask errors and dampen the impact of an imprecisely learned LPDM. As I started to think about this dampening due to vertical averaging, I also realized that in many places I am unsure whether I was already looking at weighted footprints, and simulated XCH4 enhancements or at surface footprints and CH_4 mole fractions. An example is Figure 2, but also other places in the text. Adding this information would help.

*But also in Section 5.1 and 5.2 I would really like to see back the explicit vertical dimension in your evaluations. Are footprints at 500 hPa learned just as well as those near the surface, or better? What is the skill of GATES at the typical peak sensitivity of GoSAT? If we would construct the vertical profiles of CH_4 *before* collapsing them to an XCH4, would they have differed more between the LPDM and GATES than in XCH4 units (which we are shown in the tables and figures I presume)?*

I suggest to include in Section 5.2 the locations where aircraft data are routinely gathered (Gatti network + Manaus), as violin plots of (LPDM-minus-GATES) CH_4 differences with altitude on the y-axis (based on prior fluxes, or on their respective posterior ones). Even better would be a Hovmoller plot with months on the x-axis, and the differences on a colorbar. Such figures would allow the reader to see GATES performance over a much larger spatiotemporal domain than the provided example time series in Fig 2, including the vertical domain.

If the authors can think of a better way to convince me that their vertical reconstruction of a methane profile is good enough, I also would accept that of course. But I would not be satisfied if GATES is only suited for column-averaging, as one could then never trust the fluxes+GATES to reproduce aircraft profiles or other independent datasets we use for assessment.

We acknowledge the reviewer's point about the value of the vertically resolved footprints. Achieving a version of GATES that can produce footprints for any location along the vertical dimension would provide a more complete representation of atmospheric transport, and be useful for simulating surface sites, aircraft observations, and kernel-averaged footprints.

However, data-driven models are only able to produce data that they have seen during training. The dataset used to train the GATES model described in this work consists of exclusively kernel-weighted footprints, which do not show vertically resolved transport. Therefore, here we evaluate the GATES model against unseen footprints of the same type: kernel-averaged NAME simulations for years outside of GATES' training period. Evaluating the model against aircraft observations would be an unfair comparison, as the model has not (yet) been trained to produce that type of data.

To produce surface footprints, or vertically resolved footprints to use with aircraft data, a new dataset would need to be generated, and a new instance of GATES re-trained for that purpose. Future work can revolve around compiling such datasets, which are not available to us at the moment, and training new versions of GATES, while investigating any necessary changes to the architecture or inputs to suit the new footprint types. While it is very likely that GATES would be able to predict these types of footprints given its skill in kernel-averaged footprints, the development and analysis of these models are beyond the scope of the current paper.

The inversion in the paper demonstrates the performance of GATES in application, deriving estimated emissions and comparing them to the same inversion using the model that is being emulated (NAME). The performance of the inversions themselves is not the focus of this research, but rather the difference between the inversions. As the reviewer points out, inversion outputs are often evaluated by using the posteriors + LPDM footprints to reproduce aircraft profiles or other independent datasets. Although this analysis is also outside the scope of this work, this approach could still be used to test the posteriors against independent datasets, by generating aircraft-specific footprints with NAME, or, in the future, by using a GATES-aircraft model.

We therefore disagree with the reviewer that not producing vertically disaggregated footprints is a limitation of GATES *per se*, or that it means that the inversions could not be "trusted". If the NAME model is appropriate for this context, and our emulation of those NAME footprints is successful (which is the main focus of the paper), then there is no reason to expect that inversions using the emulated footprints is not "trustworthy". Future work could use the GATES framework and a fit-for-purpose dataset to emulate footprints for each vertical level, as the reviewer suggests. The same model could then be used throughout, to simulate surface data, aircraft data, or to calculate a column mean.

Request for extra material

As a follow-up on the major request above, I found myself often looking for some more reasoning behind the evaluation metrics that are now provided. The choice of 4 locations and several months does not suggest a wide range of geographical and meteorological circumstances, as the text says. What conditions would one expect to encounter in Amazonia that could affect performance? How did you systematically assess these? A more explicit strategy would be nice to see, also in the metrics presented. I especially find the distinction between dry season and wet season conditions of interest, as both fluxes and footprints are likely to differ substantially, the latter especially in their vertical extent.

In this work, the emulated footprints over South America are evaluated in three ways against the LPDM: comparing the footprints themselves, comparing the simulated above-baseline column mole fractions, and within the main anticipated application—by comparing the inversion outputs. We find that the methods offer complementary information about the skill of GATES. In Figure 2, we show four examples of predicted footprints and mole fraction time-series, to illustrate the skill of the model under four different meteorological and topographical conditions. Figure D1 (in appendix D), footprint-wise metrics are shown spatially and seasonally, providing a systematic visualisation of performance under a range of locations and times. It shows, for example, that footprints near the Andes score lower across all metrics, consistent with difficulty emulating footprints in regions with heterogeneous topography. To further analyse GATES' skill in emulating the above-baseline mole fractions, as requested in this comment, we have added figure E1 in Appendix E. It shows the mole fractions modelled with the LPDM and the GATES footprints, as well as the Mean Bias between the two, disaggregated in space and seasonally.

Given the size of the testing dataset, it is difficult to classify the range of emulated footprints into comprehensive set of meteorological conditions. Further analysis could disaggregate the footprint-wise and mole-fraction-wise metrics by wind direction and speed, or some decomposition of the set of meteorological inputs, but we feel that our various metrics and tests are sufficient to demonstrate GATES' overall skill here, and will investigate further testing regimes in the future.

In most of the paper you furthermore show the enhancements over background, but the background itself must also be included for each XCH4 prediction. This comes from the CAMS boundary condition, transported with a footprint that traces back to the boundary of the domain. I am unsure after reading if this BC-sensitivity was also trained and reproduced with GATES, and thus part of the challenge/difference? If so, some results and discussion of the performance would be nice. If not, it must be mentioned that this is not part of the evaluation. Thinking about it more, I would say it would be a bit unfair to ignore the BC-transport in this paper that so nicely introduced GATES capacity in an inverse pipeline, and I would really urge the authors to include it in the effort (if not done already), and in the manuscript.

This is indeed a limitation in this current version of GATES, but we feel that the development of a footprint model alone is enough material for one paper. It is an issue that we are actively working on, and we believe we are close to an emulation of boundary conditions too. We will provide details in a subsequent technical note!

We have added/edited the following line to make this limitation clearer:

Line 411-420: Added a brief explanation of the inversion set-up, including the following clarification

"In this work, GATES emulates only the footprints – the boundary sensitivities are used as generated by NAME throughout, to isolate the effect of the emulator."

Line 481: Added the following to the Discussion

"Emulating the boundary sensitivities is a key step to fully integrate GATES into an inversion pipeline. These boundary sensitivities are used to calculate the background concentrations, incoming from outside the domain. In this work, we use the boundary sensitivities generated by the LPDM throughout, isolating the effect of the column-averaged footprint emulator. Future work will address this limitation, developing a boundary condition emulator. "

If the authors find it of interest, some extra material to show/explain the difference in posterior uncertainty of the flux would be appreciated. It seems that GATES has some 10% larger errors than when using the LPDM. I'd like to read your thoughts about this, either in the Results or Discussion section.

The reviewer's observation about the difference in uncertainty is appreciated. It should be noted that we use a hierarchical Bayesian inversion, which explores model uncertainty within the inversion (it is not hard-wired as in a traditional Bayesian inversion). Therefore, the higher flux uncertainty is likely reflecting the additional uncertainty that comes from the emulation process. We have noted this in the text in line 417:

"Any errors in the GATES emulation will be propagated to the inversion: the GATES-derived emissions are generally higher than the LPDM-derived ones, consistent with footprints being overall underpredicted in the model. There is also a difference in the magnitude of the uncertainties themselves, with the GATES-driven inversion presenting approximately 15% higher uncertainties in the mean yearly estimates. This additional uncertainty likely reflects the extra error due to the emulation, which is propagated through to the fluxes via our hierarchical Bayesian approach that explores model-data uncertainty as part of the inversion."

Finally, in addition to the Data Statement at the end, the work could benefit from adding a paragraph for a prospective user on how to leverage GATES. What would they do? What would they need? What resources can they expect to help them create their own footprints with GATES?

We thank the reviewer for this suggestion – a sentence describing the usability of GATES and the requirements will be added to the data available statement.

Minor suggestions

Individual minor remarks were left in an annotated PDF.

The updated paper addresses the remarks and comments left in the PDF by the reviewer, including clarification on the metrics used, and improved explanations of the thresholding and bias correction process (Section 4), illustrated by new figures B1 and B2 in Appendix B.

line 88: Since you already wrote this sentence above when explaining the general principle, I would like to have the actual number in your settings, per height. Is the number of heights fixed?

We have added further details about the set-up of NAME to section 21, including the number of vertical levels in line 87:
*"*They [Tunnicliffe et al., Ganesan et al.] perform NAME simulations for 20 vertical levels, releasing 1000 particles per hour for levels 1–8 and at 100 per hour for levels 8–17. NAME cannot emulate dispersion for the upper three column levels (18-20) and therefore the prior column mole fractions are used. "

line 139: not sure what this word signifies here. What is integrated over time?

The LPDM simulates particle transport over 30 days, and produces a footprint that aggregates all interactions with the surface over that period, rather than outputting the particle locations or surface interactions at each timestep. This is common practice in studies of methane and inert gases, like Ganesan et al. (2014), Thompson et al. (2025).

line 345: Less …data to train…or is the meteorology easier to capture? Or is the vertical structure less complex?

We have rephrased the sentence in line 366 to: "In contrast, the July-August-September season, consistently achieves better metric scores, likely due to a higher number of available training samples, or potentially because the meteorology in this season is easier to capture (Fig D1)."

line 349, Table 2: I would appreciate more subsetting of the data to find metrics specific to seasons/heights/meteo situations.

We have added figure E1 in Appendix E, showing the mean above-baseline mole fractions derived with NAME and with GATES, as well as the mean bias between the two. They are shown disaggregated in space and seasonally, showing that there are clear spatial patterns in the bias. The following has been added to the text in Section 5.2: "Figure E1 in Appendix E shows the modelled mole fractions in space and seasonally, revealing spatial patterns in the mean bias. The model over-predicts mole fractions in west Amazonia and in the South of Brazil, and underpredicts mole fractions over the North Andes, in Peru and Bolivia. These regions had already been identified as having higher errors the footprint-wise metrics (figure D1)."

We welcome the reviewer's suggestion - future work will investigate other approaches to subset the data, including evaluating performance under different meteorological conditions.

Line 389: True, but panel (c) does show a pattern of east-west differences that is of similar magnitude as the flux adjustments made in panel (b) (they even use the same color scale). If we assume that this pattern is purely due to transport then the large blue-ish patch in western Amazonia could be significant. What is your take on this? Can you at least mention the difference and discuss it?

The reviewer's observation is very valid, and has been addressed in line 428: "The GATES posterior shows larger emissions coming from Western Amazonia than the LPDM-driven posterior. This region consistently shows higher errors in the footprint-wise metrics (figure D1) and consistent underprediction of the mole fractions (figure E1), consistent with deriving higher emissions." Figure E1, showing the spatial and seasonal distribution of the derived mole fractions and their bias, has been added to expand the analysis and support these findings.

line 457: Perhaps it was mentioned, but what determines the availability of LPDM footprints for training? Was this set created previously based on GoSAT coverage?

Yes, the footprints used for training and testing had been generated for valid GOSAT measurements. We have added the following line in section 4 to clarify "Tunnicliffe et al. produced NAME footprints for GOSAT observations that pass the quality threshold over an area defined by 35.8◦ S to 7.3◦ N and from 76.0 to 32.8◦ W."