

Uncertainty assessment in satellite-based greenhouse gas emissions estimates using emulated atmospheric transport

Jeffrey N. Clark¹

jeff.clark@bristol.ac.uk

Elena Fillola^{1,2}

elena.fillolamayoral@bristol.ac.uk

Nawid Keshtmand²

yl18410@bristol.ac.uk

Raul Santos-Rodriguez¹

enrsr@bristol.ac.uk

Matthew Rigby²

matt.rigby@bristol.ac.uk

¹ School of Engineering Mathematics

and Technology

University of Bristol

Bristol, UK

² School of Chemistry

University of Bristol

Bristol, UK

Abstract

Monitoring greenhouse gas emissions and evaluating national inventories require efficient, scalable, and reliable inference methods. Top-down approaches, combined with recent advances in satellite observations, provide new opportunities to evaluate emissions at continental and global scales. However, transport models used in these methods remain a key source of uncertainty: they are computationally expensive to run at scale, and their uncertainty is difficult to characterise. Artificial intelligence offers a dual opportunity to accelerate transport simulations and to quantify their associated uncertainty.

We present an ensemble-based pipeline for estimating atmospheric transport “footprints”, greenhouse gas mole fraction measurements, and their uncertainties using a graph neural network emulator of a Lagrangian Particle Dispersion Model (LPDM). The approach is demonstrated with GOSAT (Greenhouse Gases Observing Satellite) observations for Brazil in 2016. The emulator achieved a $\sim 1,000 \times$ speed-up over the NAME LPDM, while reproducing large-scale footprint structures. Ensembles were calculated to quantify absolute and relative uncertainty, revealing spatial correlations with prediction error. The results show that ensemble spread highlights low-confidence spatial and temporal predictions for both atmospheric transport footprints and methane mole fractions.

While demonstrated here for an LPDM emulator, the approach could be applied more generally to atmospheric transport models, supporting uncertainty-aware greenhouse gas inversion systems and improving the robustness of satellite-based emissions monitoring. With further development, ensemble-based emulators could also help explore systematic LPDM errors, offering a computationally efficient pathway towards a more comprehensive uncertainty budget in greenhouse gas flux estimates.

1 Introduction

Tracking of global climate commitments relies primarily on inventory based (“bottom-up”) national self-reporting of greenhouse gas (GHG) emissions. However, these approaches are increasingly complemented by top-down methods using in situ measurements and/or satellite retrievals, combined with atmospheric transport models. Atmospheric transport models are therefore an increasingly important component of national emissions evaluation systems. However, understanding of their inherent uncertainties remains a major challenge for the accurate quantification of GHG fluxes [1]. Transport uncertainty arises from multiple sources, including errors in meteorological inputs, simplifications in model physics, and interpolation errors in space and time [2], and can affect both regional [3, 4] and global scales [5, 6]. For example, meteorological fields have uncertainties caused by errors and gaps in observations and forecast models, and even small perturbations in wind fields can significantly alter the gas dispersion, and these errors propagate into downstream estimates of emissions [7, 8].

The gold standard approach to evaluating transport models has been controlled tracer-release experiments, where known emissions provide a benchmark for testing model skill [9, 10]. These physical experiments are logically complex, spatially and temporally sparse, and expensive to perform, limiting their use for broad-scale uncertainty characterisation or to the specific locations and observed meteorological conditions, requiring computational methods to understand and quantify uncertainty. Common approaches include ensemble modelling, where perturbations in input meteorology or model parameters generate a spread of outcomes that can be used to approximate transport uncertainty [11]. However, ensembles of physics simulations are particularly computationally costly for Lagrangian Particle Dispersion Models (LPDMs) run at high spatial and temporal resolutions or over extended geographical domains [12]. These challenges are well recognised by the community, with recent policy and science roadmaps emphasising the need to better quantify transport-related uncertainty if we are to realise the full value of emerging GHG observation systems [8, 13, 14].

Machine learning offers attractive alternatives [15], and, among these, ensembles of neural networks have emerged as one of the most effective strategies for uncertainty estimation, with Bayesian Neural Networks (BNNs) [16] and deep ensembles [17] often regarded as the de facto standard. Both approaches work by averaging predictions across multiple models to obtain a predictive distribution, but at the cost of substantial computational overhead. More efficient approximations of full ensembles, such as BatchEnsemble [18] or Monte Carlo Dropout [19], have been proposed to capture similar benefits with reduced overhead, though each comes with trade-offs in accuracy or inference cost. These developments highlight the central role of ensembles in uncertainty quantification and motivate the use of machine learning emulators as a practical surrogate for computationally expensive transport models. Recent work has demonstrated the potential of emulators to reproduce the outputs of complex transport models at significantly reduced computational cost, opening the possibility of using ensembles of such emulators as a proxy for uncertainty quantification [20, 21]. If an emulator struggles to reproduce certain transport behaviours, that might provide a signal that that area or conditions lead to more inherent variability in the LPDM outputs, and therefore more uncertainty. Such approaches could provide a computationally efficient way to characterise error structure, in contrast to conventional physical ensembles.

In this paper, we take the first step towards this goal for GHG flux estimates by quantifying the emulation error of a graph neural network (GNN) LPDM emulator for GHG transport. We examine how well the ensembled emulator captures transport dynamics and use its errors as a diagnostic for uncertainty, over spatial and temporal dimensions. The pipeline and

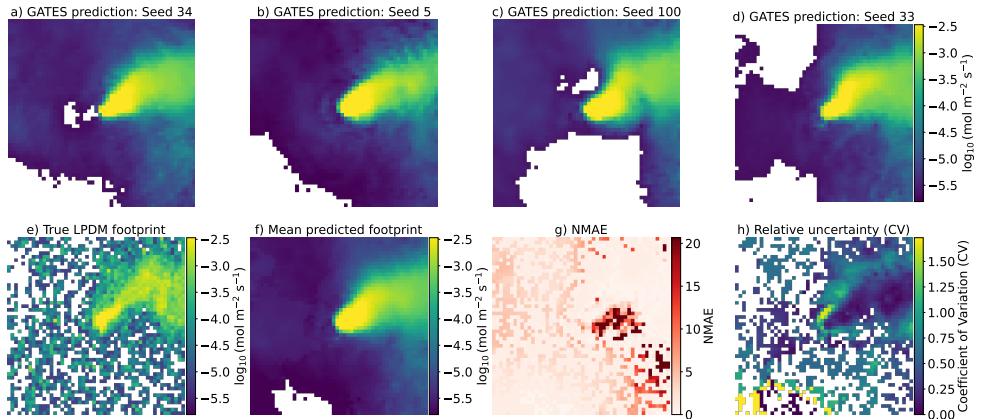


Figure 1: Top row: predicted atmospheric transport footprints generated by the four GATES (LPDM emulator) models for the same randomly selected time point from the test set. Bottom row: A comparison against the true LPDM footprint (e) with the GATES emulator ensemble mean prediction (f), normalised mean absolute error (NMAE) between the two (g), and coefficient of variation (CV) of predictions (h).

analysis that we describe here can be applied more broadly, for example to characterise the uncertainty in LPDM-only ensembles from meteorological or physical perturbations. By integrating machine learning uncertainty quantification with atmospheric transport modelling, our contribution aims to bridge the gap between computational feasibility and robust GHG emissions estimation.

2 Methods

2.1 Dataset

We utilise a GNN model to emulate LPDM-derived transport “footprints” (Figure 1). Each footprint represents the sensitivity of a satellite measurement to surface emissions, computed by releasing thousands of hypothetical air parcels backwards in time for 30 days from the satellite sounding location (measurement point) and altitude using atmospheric state estimates. These parcels record their surface contact, yielding a two-dimensional sensitivity field. In our experiments, this occurs on a regular latitude–longitude grid ($\sim 33 \times 25$ km resolution) spanning 60.98° S to 22.32° N and 91.33° W to 24.8° W over South America, generated with the UK Met Office’s NAME LPDM. Footprint values are log-transformed.

The GNN model utilises 160 input features per grid cell, following recently established methods [10]. The model utilises a range of time-varying meteorological features derived from the Met Office’s Unified Model (UM) global analysis fields, and static features to provide location-specific context. The meteorological features are extracted at seven vertical levels (100 m to 18 km) and at three time steps relative to the observation (0, -6, and -12 hours).

The dataset is split by observation period as follows, training set: 2014–2015 ($\sim 11,165$ footprints); validation set: Jan–Mar 2016 ($\sim 4,314$ footprints); and a test set: Apr–Dec 2016

(~16,945 footprints). This separation ensures that validation and testing contain unseen meteorological conditions, preventing temporal leakage from the training set.

2.2 Model architecture and training

We emulate LPDM-generated footprints using the GATES framework [20], a GNN designed to replicate atmospheric transport footprints at a fraction of the computational cost. In the training phase, the model operates on a 50×50 grid centred on each observation; for the purposes of this uncertainty study inference is performed over the same 50×50 grid, with out-of-domain areas filled with zeros.

The GATES model follows an encoder–processor–decoder structure:

- **Encoder:** Maps grid inputs into an abstract triangular mesh. Local features are aggregated using multi-layer perceptrons (MLPs).
- **Processor:** Performs multiple rounds of message passing [21] across mesh nodes, each connected to six neighbours nodes.
- **Decoder:** Maps mesh features back to the original grid, predicting footprint values per cell.

Four GATES models were independently trained with different random seeds with which to initialise model weights. An ensemble of four models was chosen as the minimum viable quantity to demonstrate the technique. Training of each model takes ~ 10 hours on a 32 GB NVIDIA V100 GPU. Post-processing steps include thresholding near-zero values to reduce noise and applying bias correction via quantile mapping using the validation set. Predictive performance was calculated as the error by subtracting the LPDM footprint values from the mean predicted values.

2.3 Mole fraction calculations

Each footprint (GATES-emulated or LPDM-generated) is convolved with a flux field to obtain above-baseline column methane mole fractions. Results are presented using two different methane flux fields, both re-gridded to the footprint resolution: a bottom-up map, and a uniform map. The bottom-up flux field, the same used in [26] and [20], aggregates anthropogenic emissions (EDGAR v4.3.2 database [20]), biomass burning (GFED v4.1 [27]) and wetlands (SWAMPS [28]). We use the map for June 2016 throughout, to remove seasonal differences in the comparison. The uniform flux emissions field is scaled to the median magnitude of the bottom-up emissions flux field to result in interpretable results. The two flux fields therefore serve complementary roles: the bottom-up field reflects realistic spatial emissions variability, while the uniform field enables clearer attribution of uncertainty to transport processes alone.

2.4 Uncertainty analysis

Uncertainty was quantified for the ensemble of four GATES models two ways: firstly absolute uncertainty was considered by calculating the standard deviation of predictions, secondly relative uncertainty was measured by calculating the coefficient of variation (CV_{pred}) as presented in Equation (1) – an established metric for atmospheric transport uncertainty [29].

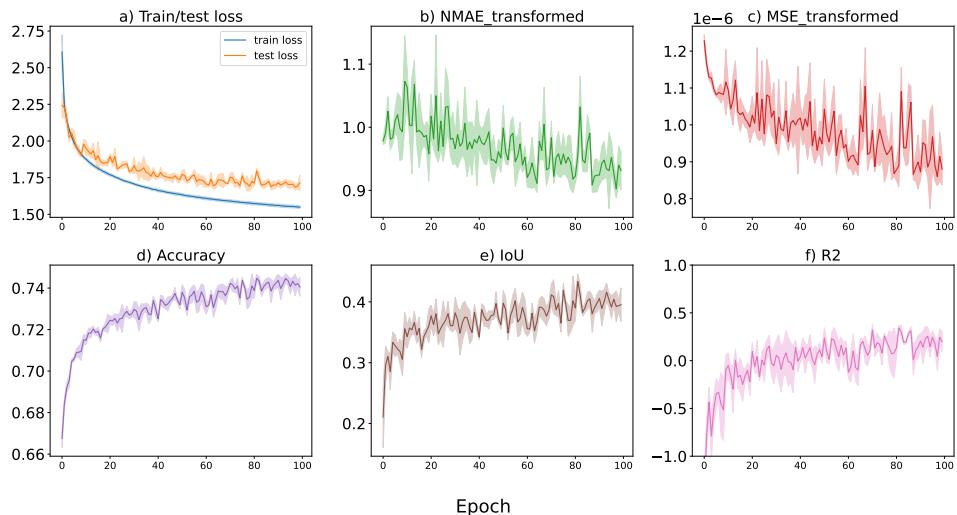


Figure 2: Mean performance during GATES LPDM emulator model development across standard machine learning metrics. Panels b-f present performance on the test set. NMAE = Normalised mean absolute error. MSE = Mean squared error. IoU = Intersection over union. R^2 = coefficient of determination. Error bars represent the standard deviation across the four trained models.

$$CV_{\text{pred}} = \frac{\sigma_{\text{pred}}}{\mu_{\text{pred}} + \varepsilon} \quad (1)$$

where σ_{pred} is the standard deviation across model predictions, μ_{pred} is the predicted mean, and ε is a small constant for numerical stability.

These metrics are applied to both footprints and methane mole fractions, and across spatial and temporal dimensions.

3 Results

3.1 Model training

Uncertainty during model development is demonstrated across 100 training epochs for the four trained models (Figure 2). Train and test loss (a) decreased steadily with limited separation between models, while both NMAE (b) and MSE (c) exhibited wider error margins, suggesting greater sensitivity to random model initialisation. Accuracy (d) improved gradually with low uncertainty bands, whereas IoU (e) and R^2 (f) converged with uncertainty bands approximately comparable to the variability per epoch within runs. Inference time was ~ 0.75 s per footprint, yielding a $\sim 1,000\times$ speed-up compared with a single LPDM simulation (~ 20 min), enabling large-scale ensemble application.

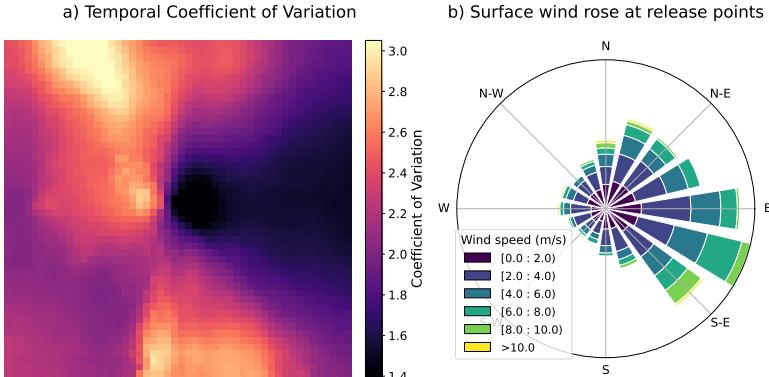


Figure 3: Left: temporal coefficient of variation of the mean prediction over the entire test set. Right: wind rose for surface-level winds, centred around the release point per footprint of the test set.

3.2 Footprints

Predicted footprints reproduced the large-scale structure of LPDM sensitivities (Figure 1a-d, f), capturing the plume extent and orientation of the LPDM footprint (Figure 1e).

Predicted footprints from the four models broadly matched the true footprint's overall structure and each produced qualitatively similar footprints (Figure 1a-d), indicating a broadly consistent prediction in shape and direction of spatial sensitivity. Relative errors and inter-model uncertainty were most apparent at footprint edges and in regions of low sensitivity to surface fluxes (low footprint values) outside of the main footprint, highlighting unstable predictions (Figure 1g,h).

Aggregated temporal and spatial uncertainties revealed structured patterns (Figures 3, 4). On average, per footprint, relative uncertainties were lowest in the east (Figure 3a), aligning with persistent easterly winds (Figure 3b). At the continent scale, lowest average uncertainties were found in north eastern South America (Figure 4c,d). Higher uncertainties occurred in the western regions, such as in the Andes which have more heterogeneous topography, suggesting that dynamically complex meteorological regimes reduce emulator robustness. The coefficient of variation (Figure 4d) further highlighted regions of low sensitivity (Figure 4a,b) but disproportionately high uncertainty, such as the eastern coast of the continent.

3.3 Methane mole fraction predictions

The presented pipeline enables quantification of the mole fraction uncertainty resulting from the footprint emulation error. Timeseries analysis (Figure 5) showed temporal periods of heightened uncertainty between GNN models (blue line). Similarly large temporal fluctuations in mole fraction derived using LPDM footprints occur (red line).

While the raw time series is dense, spatial maps provided clearer insight (Figure 6). Panel 1 of (Figure 6) demonstrates that mole fraction uncertainties derived with bottom-up emissions flux broadly mirrored the footprint uncertainty structure (Figure 4), with higher spread in the north-west. Absolute uncertainty was largest where transport sensitivities and mole fractions were highest (Figure 6 upper panel, plots a,b,d), while the coefficient of variation revealed relative instability in regions of low baseline sensitivity (panel e). Importantly, the

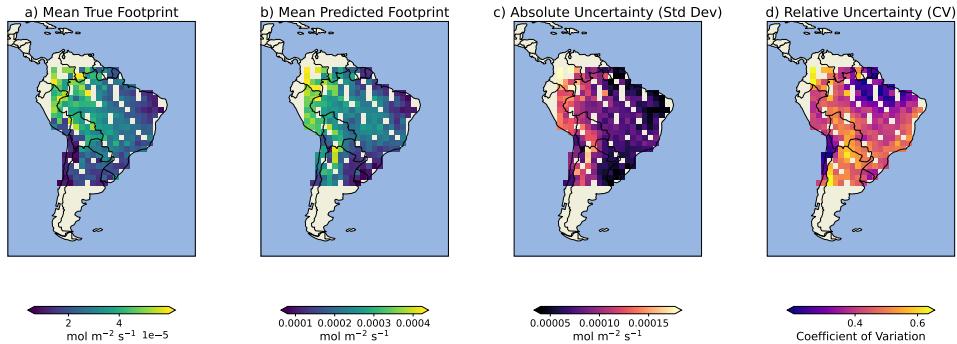


Figure 4: Spatial maps of mean footprint sensitivities across South America. a) LPDM-generated, b) GATES-predicted, c) standard deviation of predictions, and d) coefficient of variation across the four models.

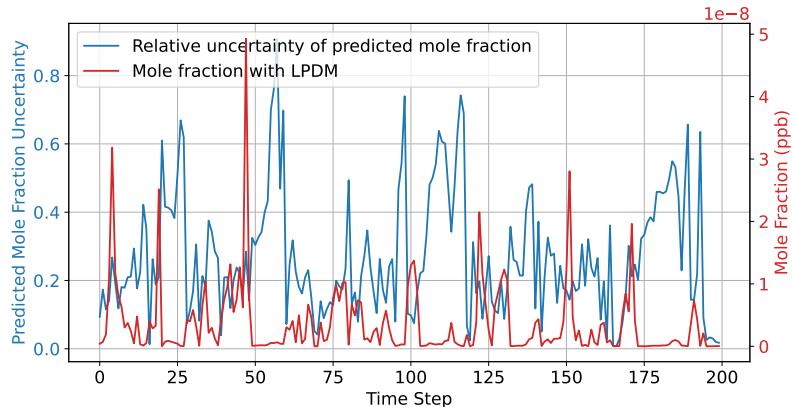


Figure 5: Relative uncertainty for predicted mole fractions across the first 200 time points of the test set, compared against mole fractions using LPDM footprints. Both are calculated using the bottom-up flux field.

spatial pattern of absolute ensemble uncertainty (panel d) qualitatively matches the mean error (panel c), indicating that uncertainty estimates are informative of prediction reliability: regions with higher spread generally coincide with larger deviations from the LPDM-derived mole fractions.

The uniform flux case (lower panel of Figure 6) provides a baseline in which spatial variability arises solely from transport rather than emission heterogeneity. Crucially, ensemble spread continues to align with mean error, confirming that the emulator's uncertainty estimates capture transport-driven variability rather than artefacts of the flux field.

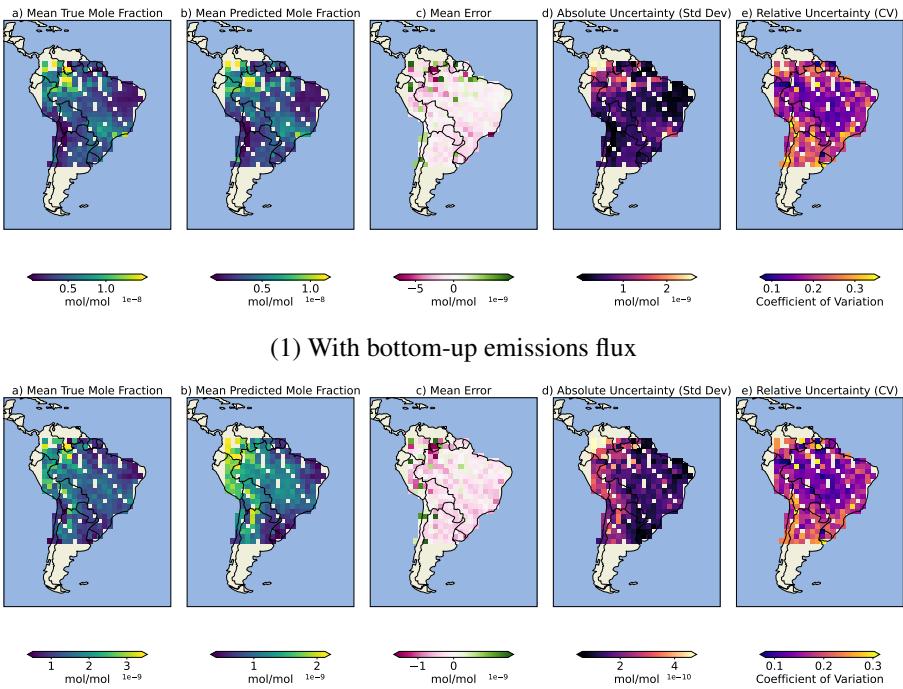


Figure 6: Spatial map of mole fractions and uncertainties with bottom-up emissions flux (1) and uniform flux (2). For each: a) true mole fractions using LPDM-based footprints, b) predicted mole fractions using GATES, c) mean error between the two, d) absolute uncertainty (standard deviation), e) relative uncertainty (coefficient of variation).

4 Discussion

This study demonstrates that ensembles of graph neural network transport emulators can provide both fast and informative estimates of uncertainty in atmospheric footprints and derived mole fractions. The correspondence between ensemble spread and prediction error suggests that model disagreement can serve as a practical indicator of low-confidence predictions. This finding is consistent with other deep learning domains where ensemble spread aligns with true error patterns [31].

Two main insights emerge. First, uncertainty is structured rather than random: it is lowest in regions of persistent easterly flow and highest in complex regions such as the Andes mountains and southern latitudes, following established findings [24]. Second, relative uncertainty (coefficient of variance, CV) complements absolute metrics by identifying unstable predictions in low-sensitivity regions that may otherwise be overlooked.

From an applications perspective, these results have direct relevance for top-down emissions estimation. Current inversion frameworks often assume transport errors are uniform or uncorrelated [24], whereas our approach provides data-driven identification of more uncertain emulated footprints. Incorporating these uncertainty layers could improve inversion robustness and guide prioritisation of satellite retrievals.

Several limitations remain. The ensemble size was small ($n = 4$) and whilst this proved

sufficient to demonstrate that ensemble spread captures regions of low predictive confidence, scaling to larger ensembles (e.g. 10–20 members) would provide more stable estimates. Exploring trade-offs between ensemble size, computational cost, and predictive gains remains an important direction for future work. Our demonstration focused on GOSAT methane retrievals over Brazil in 2016. Although this choice provided a well-constrained test case, further validation is required across different regions, time periods, and gas species. Inversion systems typically combine atmospheric transport, prior fluxes, and satellite retrievals within a Bayesian framework to produce optimised surface fluxes. Our current analysis isolates uncertainty associated with atmospheric transport, as represented by NAME and the GNN emulators. Other possible sources of error within the wider inversion system were not considered. Propagating and combining the multiple sources of uncertainty within a unified inversion framework is a remaining opportunity. Exploration of additional techniques for uncertainty quantification including Bayesian Neural Networks [20] or calibration [24] are warranted. Future work should also explore correlations between ensemble spread and input feature sparsity, systematic model errors, and inversion experiments to quantify the influence of these factors on derived fluxes.

5 Conclusions

We introduce an ensemble-based pipeline for quantifying uncertainty in graph neural network emulators of atmospheric transport footprints. Applied to GOSAT methane retrievals over Brazil, the method revealed structured spatial and temporal uncertainty patterns, with ensemble spread reliably flagging regions and times of low predictive confidence. The approach offers actionable uncertainty estimates using a computationally efficient method. Ensemble spread correlates with emulation error, allowing selective down-weighting of uncertain predictions, using an emulator offering a $\sim 1,000\times$ speed-up compared with NAME LPDM, enabling fast and scalable ensemble-scale analyses. Although demonstrated for an LPDM emulator, the framework could be applied to other atmospheric transport models. By bridging computational feasibility with robust uncertainty quantification, this work supports more reliable satellite-based greenhouse gas monitoring, directly informing climate policy and inventory verification.

Acknowledgements

We thank the Met Office, for permitting the usage of the NAME model to generate the footprints, and the Unified Model to extract the meteorology. We thank Rob Parker and the University of Leicester team for providing GOSAT satellite XCH4 retrievals. The development and training of models, and all the analysis shown here, were carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol. This work was supported by UK Research and Innovation grant NE/Z504294/1 (JNC, NK, MR), a Google PhD Fellowship, 2021 (EF), and Turing AI Fellowship grant EP/V024817/1 (RSR).

References

- [1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mhammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.
- [2] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vincius Flores Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Çağlar Gülcühre, H. Francis Song, Andrew J. Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey R. Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matthew M. Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *CoRR*, abs/1806.01261, 2018. doi: arXiv: 1806.01261. URL <http://arxiv.org/abs/1806.01261>. arXiv: 1806.01261.
- [3] P Bergamaschi, M Corazza, A Segers, A Vermeulen, A Manning, M Athanassiadou, R Thompson, I Pison, P Bousquet, and U Karstens. Top-down estimates of european ch4 and n2o emissions based on 5 different inverse models, April 2011. Netherlands.
- [4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.
- [5] Frédéric Chevallier, Paul I Palmer, Liang Feng, Hartmut Boesch, Christopher W O’Dell, and Philippe Bousquet. Toward robust and consistent regional co2 flux estimates from in situ and spaceborne measurements of atmospheric co2. *Geophysical Research Letters*, 41(3):1065–1070, 2014.
- [6] Committee on Earth Observation Satellites (CEOS) and Coordination Group for Meteorological Satellites (CGMS). CEOS-CGMS Roadmap for a coordinated implementation of carbon dioxide and methane monitoring from space, Issue 2, v1.0. Technical Report Issue 2, v1.0, CEOS-CGMS Working Group on Climate, October 2024. URL https://ceos.org/document_management/Publications/Publications-and-Key-Documents/Atmosphere/CEOS_CGMS_GHG_Roadmap_Issue_2_V1.0_FINAL.pdf. Endorsed by CEOS Plenary-38 in 2024.
- [7] Andy Delcloo and Pieter De Meutter. Quantification of uncertainty in lagrangian dispersion modelling, using ecmwf’s new era5 ensemble. In *International Technical Meeting on Air Pollution Modelling and its Application*, pages 343–346. Springer, 2018.
- [8] Aijun Deng, Thomas Lauvaux, Kenneth J Davis, Brian J Gaudet, Natasha Miles, Scott J Richardson, Kai Wu, Daniel P Sarmiento, R Michael Hardesty, Timothy A Bonin, et al. Toward reduced transport errors in a high resolution urban co2 inversion system. *Elem Sci Anth*, 5:20, 2017.
- [9] Richard J Engelen, A Scott Denning, and Kevin R Gurney. On error estimation in atmospheric co2 inversions. *Journal of Geophysical Research: Atmospheres*, 107(D22): ACL–10, 2002.

[10] Elena Fillola, Raul Santos-Rodriguez, Rachel Tunnicliffe, Jeffrey Clark, Nawid Keshtmand, Anita Ganesan, and Matthew Rigby. Enabling fast greenhouse gas emissions inference from satellites with gates: a graph-neural-network atmospheric transport emulation system. Under review with EGUsphere, 2025.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[12] Ethan Goan and Clinton Fookes. *Bayesian Neural Networks: An Introduction and Survey*, pages 45–87. Springer International Publishing, Cham, 2020.

[13] Jennifer Hegarty, Roland R Draxler, Ariel F Stein, Jerome Brioude, Marikate Mountain, Janusz Eluszkieiewicz, Thomas Nehrhorn, Fong Ngan, and Arlyn Andrews. Evaluation of lagrangian particle dispersion models with measurements from controlled tracer releases. *Journal of Applied Meteorology and Climatology*, 52(12):2623–2637, 2013.

[14] S Houweling, I Aben, F-M Breon, F Chevallier, Nicholas Deutscher, R Engelen, C Gerbig, David Griffith, K Hungershofer, Ronald Macatangay, et al. The importance of transport model uncertainties for the estimation of co 2 sources and sinks using satellite measurements. *Atmospheric chemistry and physics*, 10(20):9981–9992, 2010.

[15] Huellermeier, Eyke and Waegeman, Willem. Aleatoric and epistemic uncertainty in machine learning : an introduction to concepts and methods. *MACHINE LEARNING*, 110(3):457–506, 2021. ISSN 0885-6125. URL <http://doi.org/10.1007/s10994-021-05946-3>.

[16] G. Janssens-Maenhout, M. Crippa, D. Guizzardi, M. Muntean, E. Schaaf, F. Dentener, P. Bergamaschi, V. Pagliari, J. G. J. Olivier, J. A. H. W. Peters, J. A. van Aardenne, S. Monni, U. Doering, A. M. R. Petrescu, E. Solazzo, and G. D. Oreggioni. EDGAR v4.3.2 Global Atlas of the three major greenhouse gas emissions for the period 1970–2012. *Earth System Science Data*, 11(3):959–1002, 2019. doi: 10.5194/essd-11-959-2019. URL <https://essd.copernicus.org/articles/11/959/2019/>.

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[18] John C Lin, Dominik Brunner, and Christoph Gerbig. Studying atmospheric transport through lagrangian models. *Eos, Transactions American Geophysical Union*, 92(21): 177–178, 2011.

[19] Junjie Liu, Inez Fung, Eugenia Kalnay, and Ji-Sun Kang. Co2 transport uncertainties from the uncertainties in meteorological fields. *Geophysical Research Letters*, 38(12), 2011.

[20] SM Miller, MN Hayek, AE Andrews, I Fung, and J Liu. Biases in atmospheric co 2 estimates from correlated meteorology modeling errors. *Atmospheric Chemistry and Physics*, 15(5):2903–2914, 2015.

[21] Saqr Munassar, Guillaume Monteil, Marko Scholze, Ute Karstens, Christian Rödenbeck, Frank-Thomas Koch, Kai U Totsche, and Christoph Gerbig. Why do inverse models disagree? a case study with two european co₂ inversions. *Atmospheric Chemistry and Physics*, 23(4):2813–2828, 2023.

[22] Paul I Palmer, Liang Feng, David Baker, Frédéric Chevallier, Hartmut Bösch, and Peter Somkuti. Net carbon emissions from african biosphere dominate pan-tropical atmospheric co₂ signal. *Nature communications*, 10(1):3344, 2019.

[23] Ronny Schroeder, Kyle C. McDonald, Bruce D. Chapman, Katherine Jensen, Erika Podest, Zachary D. Tessler, Theodore J. Bohn, and Reiner Zimmermann. Development and Evaluation of a Multi-Year Fractional Surface Water Data Set Derived from Active/Passive Microwave Remote Sensing Data. *Remote Sensing*, 7(12):16688–16732, 2015. ISSN 2072-4292. doi: 10.3390/rs71215843. URL <https://www.mdpi.com/2072-4292/7/12/15843>.

[24] Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023.

[25] PG Simmonds, PI Palmer, M Rigby, A McCulloch, S O'Doherty, and AJ Manning. Tracers for evaluating computational models of atmospheric transport and oxidation at regional to global scales. *Atmospheric environment*, 246:118074, 2021.

[26] Rachel L Tunnicliffe, Anita L Ganesan, Robert J Parker, Hartmut Boesch, Nicola Gedney, Benjamin Poulter, Zhen Zhang, Jošt V Lavrič, David Walter, Matthew Rigby, et al. Quantifying sources of brazil's ch 4 emissions between 2010 and 2018 from satellite data. *Atmospheric Chemistry and Physics*, 20(21):13041–13067, 2020.

[27] G. R. van der Werf, J. T. Randerson, L. Giglio, T. T. van Leeuwen, Y. Chen, B. M. Rogers, M. Mu, M. J. E. van Marle, D. C. Morton, G. J. Collatz, R. J. Yokelson, and P. S. Kasibhatla. Global fire emissions estimates during 1997–2016. *Earth System Science Data*, 9(2):697–720, 2017. doi: 10.5194/essd-9-697-2017. URL <https://essd.copernicus.org/articles/9/697/2017/>.

[28] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.