



A machine learning emulator for Lagrangian particle dispersion model footprints: a case study using NAME

Elena Fillola¹, Raul Santos-Rodriguez¹, Alistair Manning², Simon O'Doherty³, and Matt Rigby³

¹Department of Engineering Mathematics, University of Bristol, Bristol, UK

²Hadley Centre, Met Office, Exeter, UK

³School of Chemistry, University of Bristol, Bristol, UK

Correspondence: Elena Fillola (elena.fillolamayoral@bristol.ac.uk)

Received: 27 October 2022 – Discussion started: 7 November 2022

Revised: 9 February 2023 – Accepted: 12 March 2023 – Published: 12 April 2023

Abstract. Lagrangian particle dispersion models (LPDMs) have been used extensively to calculate source-receptor relationships (“footprints”) for use in applications such as greenhouse gas (GHG) flux inversions. Because a single model simulation is required for each data point, LPDMs do not scale well to applications with large data sets such as flux inversions using satellite observations. Here, we develop a proof-of-concept machine learning emulator for LPDM footprints over a $\sim 350 \text{ km} \times 230 \text{ km}$ region around an observation point, and test it for a range of in situ measurement sites from around the world. As opposed to previous approaches to footprint approximation, it does not require the interpolation or smoothing of footprints produced by the LPDM. Instead, the footprint is emulated entirely from meteorological inputs. This is achieved by independently emulating the footprint magnitude at each grid cell in the domain using gradient-boosted regression trees with a selection of meteorological variables as inputs. The emulator is trained based on footprints from the UK Met Office’s Numerical Atmospheric-dispersion Modelling Environment (NAME) for 2014 and 2015, and the emulated footprints are evaluated against hourly NAME output from 2016 and 2020. When compared to CH_4 concentration time series generated by NAME, we show that our emulator achieves a mean R -squared score of 0.69 across all sites investigated between 2016 and 2020. The emulator can predict a footprint in around 10 ms, compared to around 10 min for the 3D simulator. This simple and interpretable proof-of-concept emulator demonstrates the potential of machine learning for LPDM emulation.

1 Introduction

To monitor the efficacy of climate agreements and understand climate feedbacks, there is an urgent need to quantify changing greenhouse gas (GHG) fluxes. Flux inference or inverse modelling systems are becoming increasingly popular for GHG flux quantification as they produce estimates of the spatial distribution of methane sources from atmospheric observations using an atmospheric transport model and statistical inversion framework. They have been used, for example, to evaluate methane emissions of the UK and Europe using in situ sensors (Lunt et al., 2021; Bergamaschi et al., 2018), for the investigation of regional CFC-11 emissions from eastern China (Rigby et al., 2019), and for many other applications.

Flux inference inverse methods were traditionally designed for relatively small data sets based on high-precision ground-based measurements (tens of sites globally that together collect \sim thousands of observations per month). However, the growth of surface networks and space-based observations mean that the volume of GHG data has increased by several orders of magnitude in recent years and will continue to grow in the next decade. For example, the TROPOMI instrument onboard the Sentinel-5 precursor, which was launched in 2017, collects around 7 million CH_4 soundings per day (Butz et al., 2012), compared to 10 000 per day from the GOSAT instrument that was launched in 2009 (Taylor et al., 2022). This growth is causing increasingly severe computational bottlenecks for GHG flux inference systems. In particular, systems relying on backward running Lagrangian particle dispersion models (LPDMs) to solve for atmospheric

transport are particularly impacted, as the required number of model evaluations grows with the number of observations.

Flux inference systems primarily use one of two types of systems to simulate atmospheric transport: Lagrangian particle dispersion models (LPDMs) or Eulerian models. The LPDMs simulate trace gas transport by following hypothetical “particles” as they move according to 3D “analysis” meteorology provided by forecasting centres. Their main advantage for GHG flux evaluation is that transport can be run backwards in time. This means that the sensitivity of GHG concentration measurements to upwind emissions, often called the “footprint” of an observation, can be calculated directly. This property makes them relatively simple and flexible to apply to GHG flux evaluation, and, when the number of observations is small, they provide a highly efficient estimate of the sensitivity of those observations to the surrounding high-dimensional flux field. Examples of widely used LPDMs are the Numerical Atmospheric-dispersion Modelling Environment (NAME, Jones et al., 2007), the Stochastic Time-Inverted Lagrangian Transport Model (STILT, Fasoli et al., 2018), and the FLEXible PARTicle dispersion model (FLEXPART, Pisso et al., 2019). Eulerian models, which calculate concentrations throughout a 3D atmospheric grid, do not suffer from the same scaling problem as the number of observations grows. However, because they do not directly calculate source–receptor relationships, they require the development of complex “adjoint” model codes (e.g. Kaminski et al., 1999), low-resolution finite difference schemes (e.g. Zammit-Mangion et al., 2022), or relatively expensive ensemble simulations (e.g. Peters et al., 2005). If LPDMs are to be used in inverse modelling studies using very large data sets, methods must be developed to overcome their poor scaling with the number of observations.

Machine learning has been shown to be useful for efficiently addressing a number of problems in studies using atmospheric dispersion models, including the correction of bias (Ivatt and Evans, 2020) and urban-scale pollution modelling (Mendil et al., 2022). The LPDM emulators have been developed to simulate volcanic ash plumes or releases from nuclear plants. For example, Gunawardena et al. (2021) use linear regression to predict footprints for a range of model configurations, Lucas et al. (2017) use Gradient-boosted regression trees (GBRTs) to predict outputs for a WRF-FLEXPART ensemble, Francom et al. (2019) use empirical orthogonal functions to reduce dimensionality and Bayesian adaptive splines to model the plume coefficients for different release characteristics, and Harvey et al. (2018) use polynomial functions to estimate average ash column loads in nearby locations for different model parameters. These studies all have two main factors in common: they all model forward dispersion rather than backwards, and they focus on a single point source and a single emissions event, looking at the ensemble members produced by different LPDM configurations.

A small number of methods have been developed to efficiently approximate LPDM footprints, mostly using interpolation or smoothing: Fasoli et al. (2018) proposed a method to run the LPDM with a small number of particles and use kernel density estimations to infer the full footprint, Roten et al. (2021) suggested a method to spatially interpolate footprints using nonlinear-weighted averaging of nearby plumes, and Cartwright et al. (2023) developed an emulator that is capable of reconstructing LPDM footprints given a “known” set of nearby footprints, using a convolutional variational auto-encoder for dimensionality reduction and a Gaussian process emulator for prediction. Though more computationally efficient than LPDMs alone, these methods still require running the LPDM a number of times for new predictions. An emulator that is capable of making footprint predictions without needing nearby simulator runs would allow substantial further efficiency gains.

Here, we present a machine learning emulator for backward running LPDM simulations based purely on meteorological inputs. Our emulator outputs hourly footprints for a small (approx. 350 km × 230 km) region around an observation point. Once trained, it does not require any further 3D simulator runs for footprint prediction. The emulator can only be constructed for fixed measurement locations, and therefore it is not applicable for satellite retrievals. However, we present it as a proof-of-concept emulator with a simple and interpretable design that can be built upon to be used for a wider range of measurement platforms. We train and evaluate the emulator by comparing it to NAME for seven sites around the world, training with data from 2014 and 2015 and evaluating predictions for 2016 and 2020. In Sect. 2, we describe NAME, the training and testing data sets, and the observation locations. Section 3 outlines the machine learning model and its characteristics, and Sects. 4 and 5 demonstrate and evaluate the predictive capabilities of the emulators. In Sect. 6, we discuss the applicability of our methodology and potential avenues for further development.

2 Data and observations

2.1 Measurement locations

Our emulator is designed to be applied to the calculation of LPDM footprints for in situ measurement stations. Here, we emulate the NAME model at seven locations. These locations were chosen to emulate a national network so that inverse modelling of national emissions could be performed, and two other locations in different meteorological regimes were chosen to demonstrate versatility. The seven measurement locations are shown in Fig. 1. Five of these sensors, located in the UK and Ireland, belong to the UK DECC (Deriving Emissions linked to Climate Change) Network (Stanley et al., 2018), and the other two sensors belong to the AGAGE (Advanced Global Atmospheric Gases Experiment) network

(Prinn et al., 2018). The stations in the DECC network have previously been used for evaluating the UK methane emissions using inverse modelling (Lunt et al., 2021), and the AGAGE stations, Trinidad Head and Gosan, have been used in various inverse modelling studies (e.g. Ganesan et al., 2014). In Sect. 4.3, we further detail the characteristics of the CH₄ measurements and follow the method used by Western et al. (2021) to infer monthly UK emissions using the predicted footprints and compare the findings to those of the NAME-produced footprints.

2.2 NAME model

The Met Office's NAME model is used to produce the footprints to train and test emulators. Each footprint is produced by releasing 20 000 model particles from the inlet height, following them backwards in time for 30 d and tracking the time particles spent near the surface (defined as being below 40 m a.g.l. as used in Lunt et al., 2016, 2021). Output footprints have a resolution of $0.352^\circ \times 0.234^\circ$ (approximately 35×23 km resolution in mid-latitudes) and cover a domain of 10×10 cells around the measurement site.

NAME was run using UKV meteorology (a UK-specific mesoscale meteorological analysis) over the UK, and global meteorology fields from the UK Met Office's Unified Model (UM, Cullen, 1993) everywhere else. The UKV has a resolution of 1.5 km and 1 h with 57 vertical levels up to 12 km, and UM meteorology has a resolution of 3 h and 25 km up to July 2014, 17 km from then until July 2017 and 12 km thereafter, with 59 vertical levels up to 29 km. The surface meteorology used as inputs is extracted from the UKV meteorology (Met Office, 2013a, 2016a) for UK sites and the global UM meteorology (Met Office, 2013b, 2016b) for the other sites, and the vertical gradients used across all sites are also extracted from the UM. Data from both models are interpolated linearly in time to increase resolution hourly and in space to the same resolution as the footprint, $0.352^\circ \times 0.234^\circ$.

We use the computational domains to produce footprints with NAME used in previous studies (e.g. Lunt et al., 2021; Rigby et al., 2019). The computational domain for Europe (used for all UK and Ireland sites) covers $10.7\text{--}79.1^\circ$ N and 97.9° W– 39.4° E, the USA domain (used for the Trinidad Head site) covers $8\text{--}59^\circ$ N and $140\text{--}39.7^\circ$ W, and the East Asia domain (used for the Gosan site) covers 5.2° S– 74.1° N and 54.5° E– 168.2° W.

The data set consists of NAME footprints calculated every hour throughout 2014, 2015, 2016, and 2020 for each site. We divide this data set into a training set, comprising 2014 and 2015 for all sites, and a testing set used to evaluate the emulators, comprising of data immediately consecutive to the training data (2016) and 5 years after (2020). Each hourly footprint takes about 10 min to be produced.

3 Emulator design

3.1 Formalization

An LPDM f produces a footprint \mathbf{y}_t^ϕ for receptor ϕ and time t , given the location and height of the receptor, the topography around it (both receptor location and topography summarized as ϕ), and a time series of 3D meteorological features (\mathbf{x}_t), so that $\mathbf{y}_t^\phi = f(\phi, \mathbf{x}_t)$.

An LPDM emulator \hat{f} is a statistical approximation of f , built using simulator runs $f(\phi_m, \mathbf{x}_n)$. As this analysis comprises seven independent sites, we instead build site-specific emulators $\hat{f}^\phi(\mathbf{x}_t)$, each trained with data for a single location.

There are many potential approaches to inferring \hat{f} using machine learning techniques: designing a model that can directly output 2D images, like neural networks; using a dimensionality reduction method to decompose \mathbf{y} into a set of features and coefficients and training a model to output coefficients given new inputs (e.g. Francom et al., 2019); or training a number of simple regressors where each one outputs the value at a single location in \mathbf{y} (e.g. Gunawardena et al., 2021). Each of these approaches has certain advantages and disadvantages. Models that are able to output 2D images directly involve deep learning, which can be difficult to design, train, and interpret and are computationally expensive. Decomposing the data to reduce the problem's dimensionality is a common method in the Earth sciences, particularly using empirical orthogonal functions (EOFs). However, Cartwright et al. (2023) demonstrate that EOFs are not able to retain the structural information of footprints as well as a deep learning alternative, which in turn requires additional complexity, including longer training and predicting times and rotating the footprints to reduce spatial variability. A grid-cell-by-grid-cell approach is simpler to design, train and interpret, but it does not implicitly capture the spatial and temporal structure of the output.

3.2 Model design

As the work presented here is of a proof-of-concept emulator demonstrating that a few selected meteorology inputs can be used to produce footprints with reduced computational expense, we demonstrate the use of a grid-cell-by-grid-cell model. As each footprint \mathbf{y}_t^ϕ is a 2D grid, the value of the emulated footprint $\hat{\mathbf{y}}_t^\phi$ at each cell (i, j) is predicted by an independent regression model $r_{i,j}^\phi$ using a subset of the meteorological inputs, such that $\hat{\mathbf{y}}_{t,i,j}^\phi = r_{i,j}^\phi(\mathbf{z}_{t,i,j})$, where $\mathbf{z}_{t,i,j} \subseteq \mathbf{x}_t$.

To reduce computational expense, we calculate the footprint only in a sub-domain of 10×10 cells centred around the receptor, so that $i, j = 1, 2, 3 \dots 10$ with the receptor located at $(5, 5)$. Therefore, each emulator is formed by 100 regressors.



Figure 1. Measurement sites in (a) the UK and Ireland and (b) the rest of the world. Sensors are located at Mace Head (MHD, Ireland, 53.326° N, 9.904° W, inlet is 10 m a.g.l. (metres above ground level)), Ridge Hill (RGL, UK, 51.997° N, 2.540° W, 90 m a.g.l.), Bilsdale (BSD, UK, 54.359° N, 1.150° W, 250 m a.g.l.), Heathfield (HFD, UK, 50.977° N, 0.230° E, 100 m a.g.l.), Tacolneston (TAC, UK, 52.518° N, 1.139° E, 185 m a.g.l.), Gosan (GSN, Korea, 33.292° N, 126.161° E, 10 m a.g.l.), and Trinidad Head (THD, USA, 41.054° N, 124.151° W, 10 m a.g.l.).

We use gradient-boosted regression trees (GBRTs) as regressors since they are easy to build, can handle multicollinearity in the inputs, are highly interpretable, and have been used repeatedly in atmospheric science (e.g. Ivatt and Evans, 2020; Sayegh et al., 2016; Lucas et al., 2017). We use the GBRT implementation from the scikit-learn library (Pedregosa et al., 2011). The GBRTs are built of regression trees, a nonlinear, nonparametric predictive model also known as decision trees. Regression trees partition the input space recursively, once per node, making binary splits on the input data (i.e. for sample z , is the value of feature x_z bigger than value k ?). The input space is therefore divided into regions, where each region corresponds to a terminal node or leaf. For any new data point, the value predicted will be a combination of the all the training samples in that leaf – for example, the mean if using mean squared error as a loss function and the mode if using mean absolute error. Though useful, regression trees alone can be inaccurate and unstable. The GBRTs use boosting to create a more robust regressor: they are a sequence of regression trees, where each tree attempts to predict the errors of the sequence before it (Friedman, 2001).

3.3 Model inputs

Each individual regressor takes meteorological variables ($\mathbf{z}_{t,i,j}$) as inputs from grid cells at two sets of locations: (i) at the cell it is predicting (i, j) and the eight adjacent cells and (ii) at the measurement site ($5, 5$) and the eight adjacent cells. Therefore, each regressor $r_{i,j}^\phi$ will have inputs from 18 locations (which might overlap). This selection of meteorological inputs was chosen because these two regions will dominate the footprint value at a given cell, with the meteorology at the measurement site dictating the overall footprint direction and the local dynamics around a cell affecting the specific behaviour. Testing indicates that this selection produces better predictions than providing the meteorological inputs at all

locations or at a fixed reduced set of locations for all cells. The meteorological inputs used are the x (west–east) and y (south–north) wind vectors at 10 m a.g.l.; planetary boundary layer height (PBLH), all taken both at the time of the footprint and 6 h before; as well as vertical gradients in temperature and x and y wind speed (between 150 and 20 m), taken only at the time of the footprint. Other potential inputs, like sea level pressure and absolute temperature were not used as they did not substantially increase the predictive power of the model.

An efficient emulator should train with as few samples as possible, while observing sufficient examples of the potential meteorological configurations. We study the data needs of the model by training and evaluating emulators using all the training data set, a half, a quarter, and a sixth of the data set (hourly, 2-hourly, 4-hourly, and 6-hourly footprints, respectively), where the hourly data set has 17 520 samples. We find that there is no difference in emulation quality between training with hourly and 2-hourly data, that the 4-hourly data produce noisier footprints, and that the 6-hourly data have little prediction power. We therefore choose to train the model with 2-hourly data, needing around 8700 footprints to train at each site. As it takes around 10 min to produce one footprint with NAME, the training data set for each site takes around 60 d of CPU time. The predictor for each cell takes under 1 min to train in a 24-core CPU, meaning that the emulator described here for a 10×10 cell domain takes around 90 min and once trained, it can produce a footprint in 10 ms (1.5 min for 1 year of hourly footprints). Therefore, if more than approximately 8700 footprints are needed for a particular site (around 1 year of hourly averages), it becomes more efficient to train the emulator than to perform further 3D model simulations (notwithstanding uncertainties as discussed in Sect. 4).

3.4 Evaluation metrics

This section outlines the evaluation metrics used to assess the quality of the footprints predicted by the emulators and where each of them is applied.

- *R-squared score* (R^2): Also known as the coefficient of determination, R^2 represents the proportion of the variance in the dependent variable that is explained by the independent variables (Chicco et al., 2021). It is defined as $R^2(a, \hat{a}) = 1 - \frac{\sum_{i=1}^m (\hat{a}_i - a_i)^2}{\sum_{i=1}^m (\bar{a} - a_i)^2}$, where \hat{a} denotes the predicted values and a the real values. It can range between $-\infty$ and 1, where 1 represents perfect predictions, 0 means none of the variance is explained (the model predicts the mean of the data at all points), and below 0 means an arbitrarily worse model.
- *Mean bias error* (MBE): MBE measures any systematic errors in the predictions and is defined as $\text{MBE}(a, \hat{a}) = \frac{1}{m} \sum_{i=1}^m (\hat{a}_i - a_i)$. A positive MBE indicates that the model tends to overpredict the output, and a negative MBE means the opposite.
- *Mean absolute error* (MAE) and *normalized MAE* (NMAE): NMAE is the MAE normalized by the mean of the true data, so the metric can be comparable across data sets of different scales. It is defined as $\text{NMAE}(a, \hat{a}) = \frac{1}{\bar{a}} \sum_{i=1}^m |a_i - \hat{a}_i|$. Lower values represent better predictions.
- *Accuracy* (AC): Accuracy is used to calculate the spatial agreement of the footprints, measuring which percentage of cells is correctly emulated to be above or below an absolute threshold b . A binary mask is created where the values of the footprint surpass b , such that

$$\mathbf{Y}_{t,i,j}^{\phi} = \begin{cases} 1 & \text{if } y_{t,i,j}^{\phi} > b \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and similarly $\hat{\mathbf{Y}}_{t,i,j}^{\phi}$ using $\hat{y}_{t,i,j}^{\phi}$. The accuracy of an emulated footprint $\hat{\mathbf{y}}_t^{\phi}$ is therefore calculated with $\text{AC} = 100\% \times \frac{|\mathbf{Y}_t^{\phi} == \hat{\mathbf{Y}}_t^{\phi}|}{|\mathbf{Y}_t^{\phi}|}$ accuracy indicates better spatial agreement above threshold b of the better spatial agreement above threshold b of the emulated and real footprint.

The emulated footprints are evaluated in three different ways: (1) footprint-to-footprint comparison, (2) convolving the footprints with a surface emissions inventory to obtain the above-baseline mole fraction, and (3) conducting a flux inversion to estimate UK methane emissions. We do footprint-to-footprint comparison using accuracy to measure the spatial agreement of the footprints and NMAE. We use R -squared score, NMAE, and MBE to evaluate the predicted mole fractions. The monthly UK methane emissions from the

emulated footprints and the real footprints are compared using MAE.

3.5 Training the model

We tune the hyperparameters for each of the emulators optimizing the R -squared score between the true footprint value at a particular cell and its prediction. This metric is chosen as opposed to other common metrics like mean square error (MSE) or mean absolute error (MAE) because the range of values in each cell varies with its position with respect to the release point. As the R -squared score does not depend on the distribution of the ground truth, it is easily interpretable across regressors.

We use 3-fold validation for 10 random regressors in each site, finding that the chosen parameters barely change across cells and sites, and therefore we select the hyperparameters to be equal throughout the emulators. As expected, we find that deeper trees perform better as they are able to capture better high-order interactions than shallow trees (Friedman, 2001). In this case, we require a maximum depth of at least 50 nodes. We find that at least 150 trees in each GBRT with a learning rate of 0.1 is preferred, with the first trees having most of the predictive power and the bulk of the trees providing small improvements to the R -squared score. We also find that the absolute error is a better loss function than the mean-squared error. This is likely because the data are approximately exponential in distribution, with most of the values for each regressor being 0 or near 0, except a few spikes or outliers. The MAE is a more appropriate metric for the Laplacian-like errors often produced by exponentially distributed variables (Hodson, 2022). Moreover, as shown often in literature, adding randomness to the GBRT also increases the score (Friedman, 2002). We find that training each tree with randomly selected \sqrt{n} features, where n is the total number of features, increases the training score significantly as well as reducing the computational expense. However, there is no benefit to using data subsampling.

4 Results and discussion

The emulator is trained for the seven locations shown in Fig. 1 using footprints every 2 h from 2014 and 2015. In this section, we evaluate the hourly footprints that these emulators produce for 2016 and 2020 (i.e. there is no overlap between the meteorology used to train the emulators and that used to test them) in three different ways: footprint-to-footprint comparison, predicted mole fraction evaluation, and UK inversion results. Figure 2 shows an example of five emulated footprints for the DECC network sites at a particular date. We also train a linear baseline model with the same data and structure, to demonstrate the benefit of using GBRTs compared to a simpler model. More details and the full results are shown in Supplement Sect. S1.

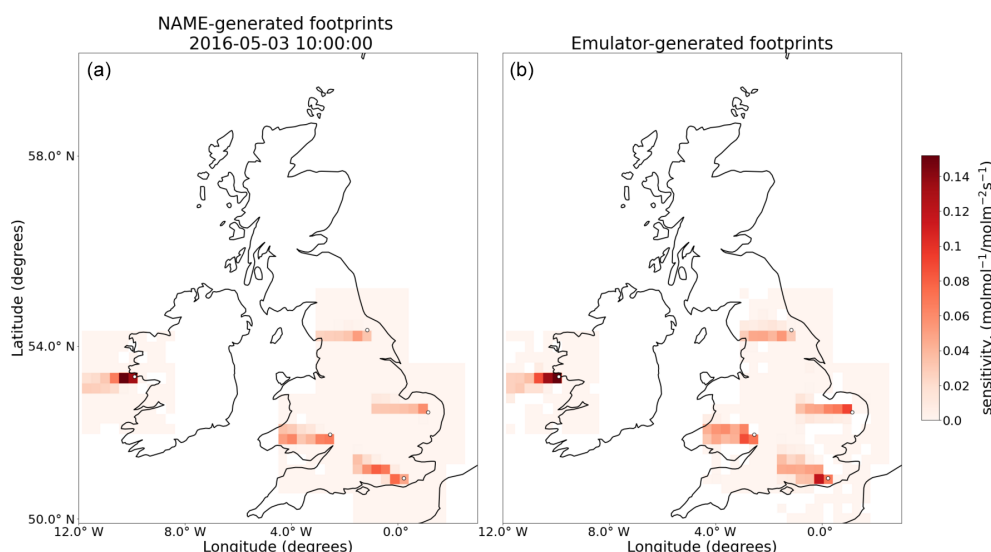


Figure 2. NAME-generated footprints (a) and emulator-generated footprints (b) for the same date (3 May 2016, 10:00:00 GMT) and the five sites in the UK and Ireland (Mace Head, Bilsdale, Ridge Hill, Heathfield, and Tacolneston, marked with a white dot). In the cells where domains for different emulators overlap, the sensitivity represents the maximum value across emulators for that cell.

4.1 Footprint-to-footprint comparison

We compute the normalized mean absolute error (NMAE) for every predicted footprint, averaging the error throughout the cells. We find that across all footprints and sites, the NMAE is 0.689 for 2016 and 0.701 for 2020. We also compute the accuracy of the footprints; this estimates which percentage of the cells is correctly emulated to be above or below a footprint value threshold b (see Sect. 3.4). We find that across sites, the emulated footprints have an accuracy of 67.3 % and 64 % for 2016 and 2020 respectively with $b = 0$, and of 88.1 % and 87.8 %, respectively for $b = 0.01$. Figure 3 shows the NMAE and accuracy for each of the sites.

4.2 Predicted mole fractions

The LPDM footprint can be convolved with a map of gridded emissions to provide the expected above-baseline mole fraction for that measurement location and time. This is calculated doing element-wise multiplication (Hadamard product) of the two grids and summing over the area. Here, we generate pseudo-time series of atmospheric methane, but our evaluation could readily have used any other species. When applied to the emulated footprints, this produces an emulated time series of expected CH_4 concentration in the area that can be compared to the NAME-generated CH_4 . Figure 4 shows two month-long examples, March 2016 and October 2020, of the time series obtained from the emulated and NAME-generated footprints.

We use EDGARv6.0 (Crippa et al., 2021) for 2016 as the gridded emissions for both 2016 and 2020 because the 2020 data set has not been released yet. EDGARv6.0 represents the mean yearly emissions on a $0.1^\circ \times 0.1^\circ$ resolution, which

is regridded using an area-weighted scheme to the same resolution as the footprints.

We calculate the NMAE, R -squared score and MBE for every time series. We find that across all sites, the NMAE is 0.308 for 2016 and 0.308 for 2020, the R -squared score is 0.694 and 0.697, respectively, and the MBE is -0.0125 and $-0.0043 \mu\text{mol mol}^{-1}$, respectively. Figure 5 shows these metrics for each of the sites.

Although the MBE for the emulator is small, Fig. 4 shows that the highest mole fractions are often not well predicted. We show how the bias changes across values by dividing the yearly real data into 10 quantiles (10 equal-sized, ordered subsets) and taking the MBE of each. Figure 6 shows that the bias is small for lower molar fraction values, but that the model tends to highly underpredict the higher range, with a similar behaviour across sites.

4.3 UK emissions inversion

To evaluate the performance of the emulator in a common application, we carry out a UK methane flux inversion that has recently been performed in Lunt et al. (2021, 2016) and Western et al. (2020). We follow a hierarchical Bayesian Markov chain Monte Carlo (MCMC) method and use input parameters described by Western et al. (2021) to estimate monthly UK methane emissions for 2016 using the predicted footprints. We use the DECC network sensors, which have measured CH_4 continuously for the period analysed here: Mace Head (Prinn et al., 2022), Ridge Hill, Bilsdale, Heathfield, and Tacolneston (O'Doherty et al., 2020) (Fig. 1). Details on the prior and instruments used for measurements can

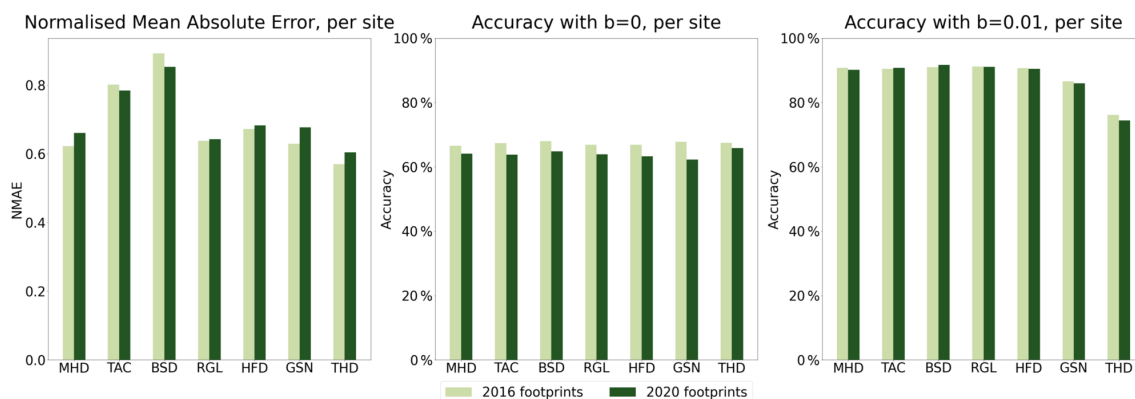


Figure 3. Evaluation of emulators with footprint-to-footprint comparison, using metrics NMAE and accuracy with $b = 0$ (all footprint values) and $b = 0.01$ (high values), shown per site and year.

be found in Lunt et al. (2021), but note that a slightly different inversion method is used in that paper.

Since our emulator is predicting footprints in a small domain around the sensor and the inversion requires a bigger domain, we produce an estimate of the total footprint by using the NAME-calculated footprint in the rest of the domain not within our emulated region. Therefore, while our emulator calculates the most important part of the footprint (i.e. the part with the highest values), it should be noted that there will be some influence of the “true” footprints on the final results for this comparison. We conduct a sensitivity test in Sect. 2 which demonstrates that the inversion is highly sensitive to the emulated area but less outside of this region.

Using the NAME-generated footprints, we find a UK mean for 2016 of 2.03 (1.90 – 2.16) Tg yr^{-1} (uncertainty represents 95 % high density interval), consistent with top-down estimates in Manning et al. (2020) and Lunt et al. (2021) and with the 2016 inventory (Brown et al., 2020). Using the full-domain emulated footprints, the UK mean for 2016 is 2.15 (2.02 – 2.30) Tg yr^{-1} , 5.9 % higher than inferred with the real footprints. The monthly emission rates can be seen in Fig. 7, with a mean monthly difference between the real and predicted inversion of 0.130 Tg or 6.32 %. This increase in inferred emissions, compared to the inversion using the real footprints, is consistent with the emulator generally underestimating the highest mole fractions.

5 Feature evaluation

The design of the emulator, with one regressor per cell, means we can evaluate which input variables are more relevant at each cell and therefore understand the spatial distribution of the feature importance scores, and check if they are physically coherent. Tree-based models like GBRTs are highly interpretable as they can rank the inputs in terms of how much they contribute to building the trees. However, when working with multi-collinearity in the inputs, feature

importance scores are not reliable, as similar information is present across correlated features.

Another common way to rank the features in a model is by calculating the feature permutation importance. In this approach, the values of one or more inputs in a data set are shuffled, effectively adding noise to that feature, and the prediction error of the new data set is compared to the prediction error of the original data set, called baseline error (Molnar, 2022). This approach has similar issues to calculating the importance with the GBRT if it is used on single features that have high correlation to other inputs. However, it can be run on multiple features at a time, meaning that we can calculate the importance of blocks of correlated features.

To calculate the feature importance scores across the domain, we divide the input data for each regressor into nine blocks, one per meteorological input across all locations: x wind (west–east), y wind (south–north), and PBLH at the time of the footprint, the same three inputs 6 h before the sampling time, the x and y wind vertical gradients and the temperature gradient. We calculate the baseline NMAE for each regressor using data for 2016 and then calculate the NMAE when shuffling each of the blocks. The difference between the two, the added error, is a proxy for feature importance scores – the more a regressor relies on a feature, the higher errors it will produce when that feature is noisy. As an example, Fig. 8 shows the most important feature at each cell for Mace Head. We propose that the distribution of importance can be interpreted in physical terms as follows: cells on the S–N axis are more affected by the W–E wind, as low W–E winds would mean higher concentrations in said axis and vice versa, and similarly for the W–E axis and S–N winds. Moreover, winds 6 h before the footprint become increasingly relevant towards the edges of the domain, consistent with the dispersion running backwards in time.

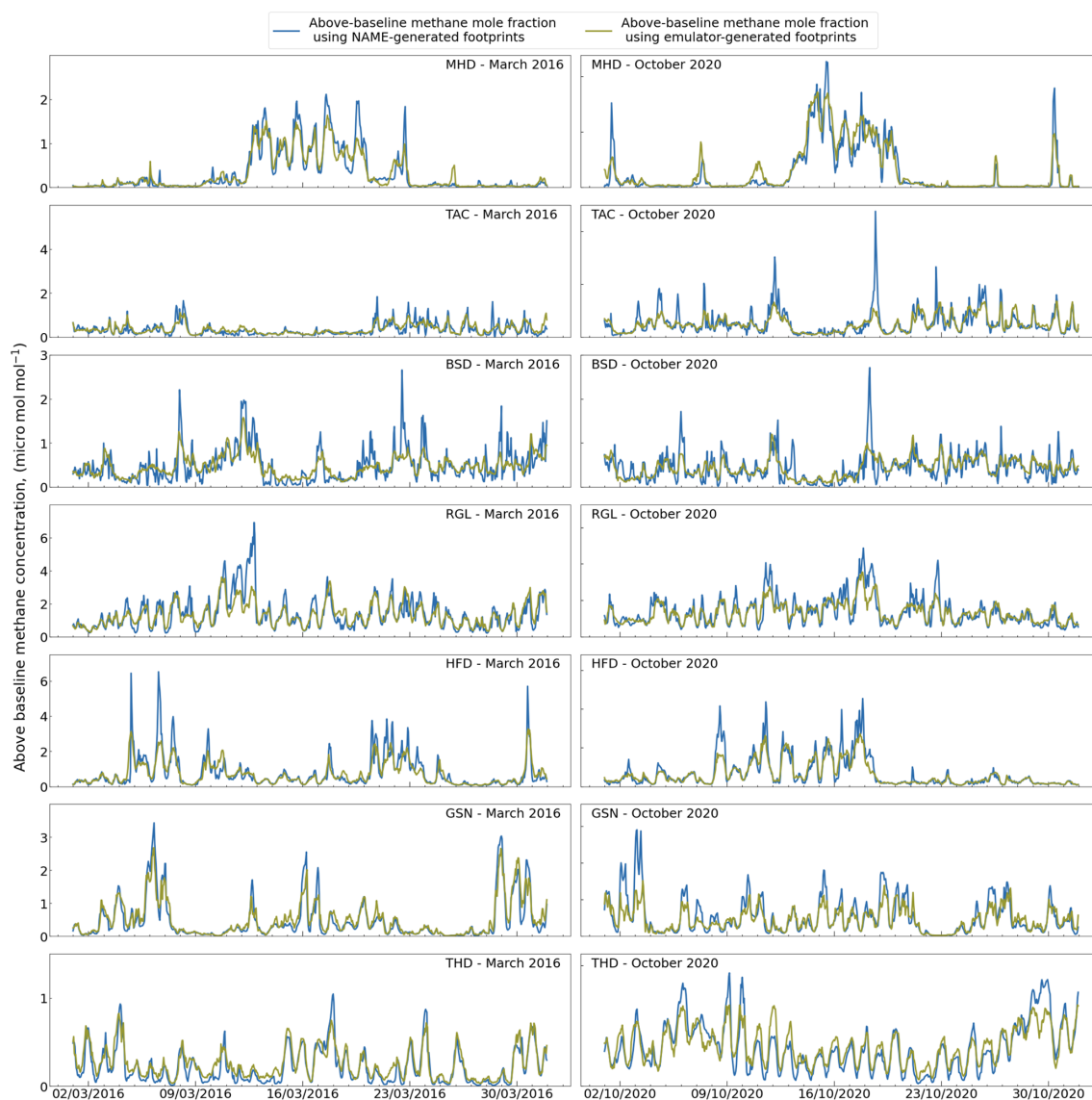


Figure 4. Mole fractions from NAME-generated footprints and predicted footprints for March 2016 (left column) and October 2020 (right column) around each measurement site.

6 Conclusions

We have presented a proof-of-concept emulator for LPDMs, which can efficiently produce footprints using only meteorological inputs, and we demonstrated its performance on seven measurement stations around the world. The emulator offers a considerable speed-up with respect to both normal LPDM runs and interpolation-based methods, because once trained, it does not require further LPDM runs. The emulator can produce footprints that resemble those generated by NAME, with high correlations and low bias for the predicted above-baseline mole fraction at the seven sites investigated here. An inversion of UK methane fluxes performed using our emulated footprints was not statistically different to an inversion using the LPDM footprints. Since there is no decrease in

performance between 2016 and 2020, the emulator appears to have inference capabilities for at least 5 years after the training data, making it a long-term tool that does not require retraining often. Moreover, we use meteorology at different resolutions in different sites (high-resolution, national UKV and coarse-resolution, global UM), but we see no differences in scores across several metrics, meaning the model can be trained and used with different input resolutions. Although not validated in this work, it is likely that performance will be similar when training with different LPDMs, like STILT or FLEXPART.

There are limitations in our emulator that will need to be overcome before it could be used to replace LPDM model evaluations in applications such as inverse modelling. Our



Figure 5. Evaluation of emulators with mole fraction comparison, using metrics NMAE, *R*-squared score, and MBE, shown per site and year.

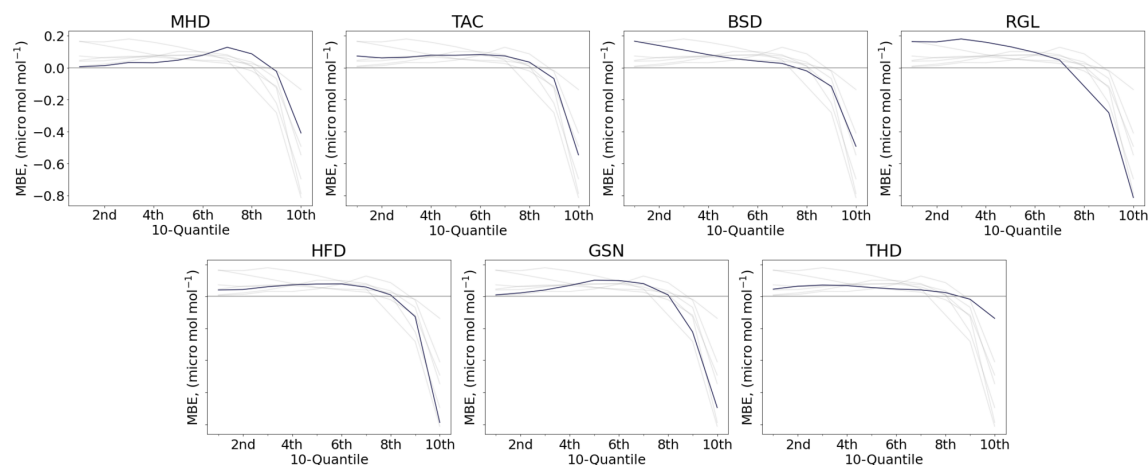


Figure 6. Mean bias error in the above-baseline mole fraction predictions for each of the 10 quantiles of the true data, across sites for 2016. An MBE above 0 indicates overprediction while an MBE below 0 indicates underprediction.

emulator predicts only a small domain around the receptor, which is not big enough for most national-scale inversions. The domain size could be increased by using extra regressors. However, as the training time increases linearly with the number of regressors, strategies to keep training times feasible should be considered. This could include coarsening the grid towards the edges of the domain, as demonstrated in Supplement Sect. S2 or parallel training. Further work would be required to select the most appropriate combination of input data for the added regressors, including, for example, meteorology further back in time. The design of the emulator, chosen due to its simplicity to set up and train, could also be improved by making the regressors dependent, either across time (a regressor's inputs include data from the previous footprint) and/or across space (a regressor's inputs include predictions from nearby regressors).

As well as the known design limitations, the performance of the emulator highlights opportunities for improvement. The high-value bias, present in all sites, could be reduced

with approaches such as bias-reduction methods. Identifying the meteorological conditions in which the model performs more poorly would also be useful, particularly to relate them to conditions usually filtered out in inversions or in which the LPDM is also considered less reliable. For example, low wind conditions, which usually cause high local influence, could coincide with the badly predicted events. Creating a training data set with a more balanced distribution of meteorological conditions may also help to reduce the differences in performance across different situations.

The difference in prediction quality across sites could also provide insights into potential areas of improvement for the model. We find that there are no noticeable differences in accuracy with $b = 0$ across sites, meaning that the spatial distribution of footprints is captured similarly. The improvement in accuracy between thresholds $b = 0$ and $b = 0.01$ likely indicates that the emulators are better at capturing the main shape of the footprint, composed of higher values (i.e. $b = 0.01$), but that the background (captured with $b = 0$) tend

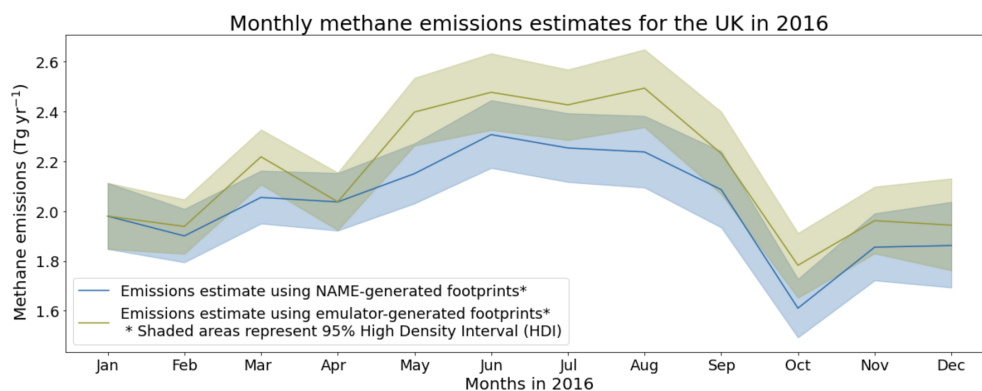


Figure 7. Monthly UK methane emission estimates using the NAME-generated footprints for the five sites in the DECC network (UK and Ireland sites) and the emulator-generated footprints (with NAME-generated values outside of the emulated region) for the same sites.

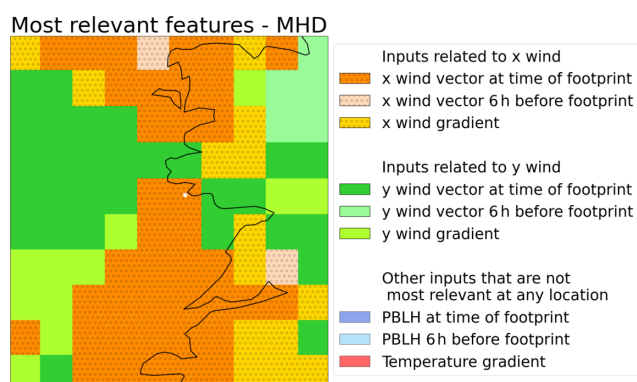


Figure 8. Most important block of meteorological features at each cell for Mace Head, as calculated with the block permutation importance.

to be less well predicted – for example, in Fig. 2, it can be seen that the model confuses very small values with 0. However, we find a difference in performance across sites when evaluating the values predicted with NMAE. We find that the receptors close to ground level (MHD, GSN, and THD are at 10 m a.g.l.) are significantly better predicted than those with higher inlets (TAC is 185 m a.g.l. and BSD is 250 m a.g.l.), both when doing footprint-to-footprint comparison and when evaluating the predicted mole fractions. As most of the inputs are provided at 10 m a.g.l. when the PBLH is low, this meteorology may not be representative of the state of the atmosphere around the taller sensors and therefore lead to higher errors.

To more fully exploit the possibilities of machine learning, it would be desirable to generalize the emulator to any location. For the emulators built here, the effect of the surface surrounding each site is implicitly captured by using site-specific training data. For the emulator to be applied at an arbitrary location, it should have knowledge of the effect of topography and other surface characteristics on dispersion. A well-designed emulator that is trained with data for some

locations should therefore be able to produce footprints for similar, unseen locations. Use of additional variables (e.g. vertical wind speed) and designing the model to read and exploit 2D, 3D, or 4D meteorological fields may improve prediction accuracy. Ideally, more advanced models should also estimate an uncertainty in the predictions, either directly through the model or by choosing a probabilistic method that can be used to build ensembles.

Code and data availability. Code used to train and evaluate the models is available as a free access repository at DOI <https://doi.org/10.5281/zenodo.7254667> (Fillola, 2022b). Sample data to accompany the code, including the trained emulator for Mace Head and inputs/outputs to test it, can be found at DOI <https://doi.org/10.5281/zenodo.7254330> (Fillola, 2022a). The NAME III v7.2 transport model is available from the UK Met Office under licence by contacting enquiries@metoffice.gov.uk. The meteorological data used in this work from the UK Met Office's operational NWP (Numerical Weather Prediction) Unified Model (UM) are available from the UK Centre for Environmental Data Analysis at both resolutions: UKV (<http://catalogue.ceda.ac.uk/uuid/292da1ccfebd650f6d123e53270016a8>, Met Office, 2013a; <https://catalogue.ceda.ac.uk/uuid/f47bc62786394626b665e23b658d385f>, Met Office, 2016a) and global (<http://catalogue.ceda.ac.uk/uuid/220f1c04ffe39af29233b78c2cf2699a>, Met Office, 2013b; <https://catalogue.ceda.ac.uk/uuid/86df725b793b4b4cb0ca0646686bd783>, Met Office, 2016b). The software used for the inversion can be found at DOI <https://doi.org/10.5281/zenodo.6834888> (Rigby et al., 2022). Measurements of methane for the Mace Head station are available at <http://agage.mit.edu/data> (Prinn et al., 2022) and measurements of methane from the UK DECC network sites Tacolneston, Ridge Hill, Heathfield, and Bilsdale are available at <https://catalogue.ceda.ac.uk/uuid/f5b38d1654d84b03ba79060746541e4f> (O'Doherty et al., 2020).

Supplement. The supplement related to this article is available online at: <https://doi.org/10.5194/gmd-16-1997-2023-supplement>.

Author contributions. EF carried out the research and wrote the paper with input from all co-authors. The research was designed by EF, MR, and RSR. AM produced the footprint data and SO provided the atmospheric measurements.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. Elena Fillola was funded through a Google PhD Fellowship 2021. Matt Rigby was funded through the Natural Environment Research Council's Constructing a Digital Environment OpenGHG project (NE/V002996/1). Raul Santos-Rodriguez was funded by the UKRI EPSRC Turing AI Fellowship EP/V024817/1. Measurements from Mace Head were funded by the Advanced Global Atmospheric Gases Experiment (NASA grant NNX16AC98G) and measurements from the UK DECC network by the UK Department of Business, Energy & Industrial Strategy through contract 1537/06/2018 to the University of Bristol. Since 2017, measurements at Heathfield have been maintained by the National Physical Laboratory mainly under funding from the National Measurement System. This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol – <http://www.bristol.ac.uk/acrc/> (last access: April 2023).

Financial support. This research has been supported by the Natural Environment Research Council (grant no. NE/V002996/1), Google (PhD Fellowship 2021), and the UK Research and Innovation, Engineering and Physical Sciences Research Council (Turing AI Fellowship, grant no. EP/V024817/1).

Review statement. This paper was edited by Po-Lun Ma and reviewed by two anonymous referees.

References

- Bergamaschi, P., Karstens, U., Manning, A. J., Saunio, M., Tsutsumi, A., Berchet, A., Vermeulen, A. T., Arnold, T., Janssens-Maenhout, G., Hammer, S., Levin, I., Schmidt, M., Ramonet, M., Lopez, M., Lavric, J., Aalto, T., Chen, H., Feist, D. G., Gerbig, C., Haszpra, L., Hermansen, O., Manca, G., Moncrieff, J., Meinhardt, F., Necki, J., Galkowski, M., O'Doherty, S., Paramonova, N., Scheeren, H. A., Steinbacher, M., and Dlugokencky, E.: Inverse modelling of European CH₄ emissions during 2006–2012 using different inverse models and reassessed atmospheric observations, *Atmos. Chem. Phys.*, 18, 901–920, <https://doi.org/10.5194/acp-18-901-2018>, 2018.
- Brown, P., Cardenas, L., Choudrie, S., Jones, L., Karagianni, E., MacCarthy, J., Passant, N., Richmond, B., Smith, H., Thistlethwaite, G., Thomson, A., Turtle, L., and Wakeling, D.: UK Greenhouse Gas Inventory, 1990 to 2018: Annual Report for Submission under the Framework Convention on Climate Change, Tech. Rep., Department for Business, Energy & Industrial Strategy, 978-0-9933975-6-1, https://naei.beis.gov.uk/reports/reports?report_id=998 (last access: 28 March 2023), 2020.
- Butz, A., Galli, A., Hasekamp, O., Landgraf, J., Tol, P., and Aben, I.: TROPOMI aboard Sentinel-5 Precursor: Prospective performance of CH₄ retrievals for aerosol and cirrus loaded atmospheres, *Remote Sens. Environ.*, 120, 267–276, <https://doi.org/10.1016/j.rse.2011.05.030>, 2012.
- Cartwright, L., Zammit-Mangion, A., and Deutscher, N. M.: Emulation of greenhouse-gas sensitivities using variational autoencoders, *Environmetrics*, 34, e2754, <https://doi.org/10.1002/env.2754>, 2023.
- Chicco, D., Warrens, M. J., and Jurman, G.: The coefficient of determination *R*-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, *PeerJ Computer Science*, 7, e623, <https://doi.org/10.7717/peerj-cs.623>, 2021.
- Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Lo Vullo, E., Solazzo, E., Monforti-Ferrario, F., Olivier, J., and Vignati, E.: EDGAR v6.0 Greenhouse Gas Emissions, European Commission, Joint Research Centre (JRC) [data set], <https://data.jrc.ec.europa.eu/dataset/97a67d67-c62e-4826-b873-9d972c4f670b> (last access: 1 March 2023), 2021.
- Cullen, M. J. P.: The unified forecast/climate model, *Meteorol. Mag.*, 122, 81–94, 1993.
- Fasoli, B., Lin, J. C., Bowling, D. R., Mitchell, L., and Mendoza, D.: Simulating atmospheric tracer concentrations for spatially distributed receptors: updates to the Stochastic Time-Inverted Lagrangian Transport model's R interface (STILT-R version 2), *Geosci. Model Dev.*, 11, 2813–2824, <https://doi.org/10.5194/gmd-11-2813-2018>, 2018.
- Fillola, E.: Sample dataset for “A machine learning emulator for Lagrangian particle dispersion model footprints”, Zenodo [data set], <https://doi.org/10.5281/zenodo.7254330>, 2022a.
- Fillola, E.: elenafillo/LPDM-emulation-trees: LPDM-emulation-trees v1.0 (v1.0), Zenodo [code], <https://doi.org/10.5281/zenodo.7254667>, 2022b.
- Francom, D., Sansó, B., Bulaevskaya, V., Lucas, D., and Simpson, M.: Inferring atmospheric release characteristics in a large computer experiment using Bayesian adaptive splines, *J. Am. Stat. Assoc.*, 114, 1450–1465, <https://doi.org/10.1080/01621459.2018.1562933>, 2019.
- Friedman, J. H.: Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, 29, 1189–1232, <https://doi.org/10.1214/aos/1013203451>, 2001.
- Friedman, J. H.: Stochastic gradient boosting, *Comput. Stat. Data An.*, 38, 367–378, [https://doi.org/10.1016/s0167-9473\(01\)00065-2](https://doi.org/10.1016/s0167-9473(01)00065-2), 2002.
- Ganesan, A. L., Rigby, M., Zammit-Mangion, A., Manning, A. J., Prinn, R. G., Fraser, P. J., Harth, C. M., Kim, K.-R., Krummel, P. B., Li, S., Mühle, J., O'Doherty, S. J., Park, S., Salameh, P. K., Steele, L. P., and Weiss, R. F.: Characterization of uncertainties in atmospheric trace gas inversions using hierarchi-

- cal Bayesian methods, *Atmos. Chem. Phys.*, 14, 3855–3864, <https://doi.org/10.5194/acp-14-3855-2014>, 2014.
- Gunawardena, N., Pallotta, G., Simpson, M., and Lucas, D. D.: Machine learning emulation of spatial deposition from a multi-physics ensemble of weather and atmospheric transport models, *Atmosphere*, 12, 953, <https://doi.org/10.3390/atmos12080953>, 2021.
- Harvey, N. J., Huntley, N., Dacre, H. F., Goldstein, M., Thomson, D., and Webster, H.: Multi-level emulation of a volcanic ash transport and dispersion model to quantify sensitivity to uncertain parameters, *Nat. Hazards Earth Syst. Sci.*, 18, 41–63, <https://doi.org/10.5194/nhess-18-41-2018>, 2018.
- Hodson, T. O.: Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not, *Geosci. Model Dev.*, 15, 5481–5487, <https://doi.org/10.5194/gmd-15-5481-2022>, 2022.
- Ivatt, P. D. and Evans, M. J.: Improving the prediction of an atmospheric chemistry transport model using gradient-boosted regression trees, *Atmos. Chem. Phys.*, 20, 8063–8082, <https://doi.org/10.5194/acp-20-8063-2020>, 2020.
- Jones, A., Thomson, D., Hort, M., and Devenish, B.: The U.K. Met Office's next-generation atmospheric dispersion model, NAME III, Air Pollution Modeling and Its Application XVII, 580–589, ISBN 978-0-387-68854-1, https://doi.org/10.1007/978-0-387-68854-1_62, 2007.
- Kaminski, T., Heimann, M., and Giering, R.: A coarse grid three-dimensional global inverse model of the atmospheric transport: 1. Adjoint model and Jacobian matrix, *J. Geophys. Res.-Atmos.*, 104, 18535–18553, <https://doi.org/10.1029/1999jd900147>, 1999.
- Lucas, D. D., Simpson, M., Cameron-Smith, P., and Baskett, R. L.: Bayesian inverse modeling of the atmospheric transport and emissions of a controlled tracer release from a nuclear power plant, *Atmos. Chem. Phys.*, 17, 13521–13543, <https://doi.org/10.5194/acp-17-13521-2017>, 2017.
- Lunt, M. F., Rigby, M., Ganesan, A. L., and Manning, A. J.: Estimation of trace gas fluxes with objectively determined basis functions using reversible-jump Markov chain Monte Carlo, *Geosci. Model Dev.*, 9, 3213–3229, <https://doi.org/10.5194/gmd-9-3213-2016>, 2016.
- Lunt, M. F., Manning, A. J., Allen, G., Arnold, T., Bauguutte, S. J.-B., Boesch, H., Ganesan, A. L., Grant, A., Helfter, C., Nemitz, E., O'Doherty, S. J., Palmer, P. I., Pitt, J. R., Rennick, C., Say, D., Stanley, K. M., Stavert, A. R., Young, D., and Rigby, M.: Atmospheric observations consistent with reported decline in the UK's methane emissions (2013–2020), *Atmos. Chem. Phys.*, 21, 16257–16276, <https://doi.org/10.5194/acp-21-16257-2021>, 2021.
- Manning, A., Redington, A., O'Doherty, S., Say, D., Young, D., Arnold, T., Rennick, C., Rigby, M., Wisher, A., and Simmonds, P.: Long-Term Atmospheric Measurement and Interpretation of Radiatively Active Trace Gases – Detailed Report (September 2019 to August 2020), Tech. Rep., Department for Business, Energy & Industrial Strategy, <https://assets.publishing.service.gov.uk/government> (last access 1 June 2022), 2020.
- Mendil, M., Leirens, S., Armand, P., and Duchenne, C.: Hazardous atmospheric dispersion in urban areas: A Deep Learning approach for emergency pollution forecast, *Environ. Modell. Softw.*, 152, 105387, <https://doi.org/10.1016/j.envsoft.2022.105387>, 2022.
- Met Office: Operational Numerical Weather Prediction (NWP) Output from the UK Variable (UKV) Resolution Part of the Met Office Unified Model (UM), NCAS British Atmospheric Data Centre [data set], <http://catalogue.ceda.ac.uk/uuid/292da1ccfebd650f6d123e53270016a8> (last access: 1 March 2022), 2013a.
- Met Office: Operational Numerical Weather Prediction (NWP) Output from the North Atlantic European (NAE) Part of the Met Office Unified Model (UM), NCAS British Atmospheric Data Centre [data set], <http://catalogue.ceda.ac.uk/uuid/220f1c04ffe39af29233b78c2cf2699a> (last access: 1 March 2022), 2013b.
- Met Office: NWP-UKV: Met Office UK Atmospheric High Resolution Model data, Centre for Environmental Data Analysis [data set], <https://catalogue.ceda.ac.uk/uuid/f47bc62786394626b665e23b658d385f> (last access: 1 March 2022), 2016a.
- Met Office: NWP-UKV: Met Office UK Atmospheric High Resolution Model data, Centre for Environmental Data Analysis [data set], <https://catalogue.ceda.ac.uk/uuid/86df725b793b4b4cb0ca0646686bd783> (last access: 1 March 2022), 2016b.
- Molnar, C.: Global Model-Agnostic Methods: Permutation Feature Importance, Christoph Molnar, <https://christophm.github.io/interpretable-ml-book/feature-importance.html>, last access: 1 July 2022.
- O'Doherty, S., Say, D., Stanley, K., Spain, G., Arnold, T., Rennick, C., Young, D., Stavert, A., Grant, A., Ganesan, A., Pitt, J., Wisher, A., Wenger, A., and Garrard, N.: UK DECC (Deriving Emissions linked to Climate Change) Network, Centre for Environmental Data Analysis [data set], <https://catalogue.ceda.ac.uk/uuid/f5b38d1654d84b03ba79060746541e4f> (last access: 1 March 2022), 2020.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 12, 2825–2830, 2011.
- Peters, W., Miller, J. B., Whitaker, J., Denning, A. S., Hirsch, A., Krol, M. C., Zupanski, D., Bruhwiler, L., and Tans, P. P.: An ensemble data assimilation system to estimate CO₂ surface fluxes from atmospheric trace gas observations, *J. Geophys. Res.*, 110, D24304, <https://doi.org/10.1029/2005jd006157>, 2005.
- Pisso, I., Sollum, E., Grythe, H., Kristiansen, N. I., Casiani, M., Eckhardt, S., Arnold, D., Morton, D., Thompson, R. L., Groot Zwaafink, C. D., Evangeliou, N., Sode-mann, H., Haimberger, L., Henne, S., Brunner, D., Burkhardt, J. F., Fouilloux, A., Brioude, J., Philipp, A., Seibert, P., and Stohl, A.: The Lagrangian particle dispersion model FLEX-PART version 10.4, *Geosci. Model Dev.*, 12, 4955–4997, <https://doi.org/10.5194/gmd-12-4955-2019>, 2019.
- Prinn, R. G., Weiss, R. F., Arduini, J., Arnold, T., DeWitt, H. L., Fraser, P. J., Ganesan, A. L., Gasore, J., Harth, C. M., Hermansen, O., Kim, J., Krummel, P. B., Li, S., Loh, Z. M., Lunder, C. R., Maione, M., Manning, A. J., Miller, B. R., Mitrevski, B., Mühle, J., O'Doherty, S., Park, S., Reimann, S., Rigby, M., Saito, T., Salameh, P. K., Schmidt, R., Simmonds, P. G., Steele,

- L. P., Vollmer, M. K., Wang, R. H., Yao, B., Yokouchi, Y., Young, D., and Zhou, L.: History of chemically and radiatively important atmospheric gases from the Advanced Global Atmospheric Gases Experiment (AGAGE), *Earth Syst. Sci. Data*, 10, 985–1018, <https://doi.org/10.5194/essd-10-985-2018>, 2018.
- Prinn, R. G., Weiss, R. F., Arduini, J., Arnold, T., Fraser, P. J., Ganesan, A. L., Gasore, J., Harth, C. M., Hermansen, O., Kim, J., Krummel, P. B., Li, S., Loh, Z. M., Lunder, C. R., Maione, M., Manning, A. J., Miller, B. R., Mitrevski, B., Mühle, J., O'Doherty, S., Park, S., Reimann, S., Rigby, M., Salameh, P. K., Schmidt, R., Simmonds, P., Steele, L. P., Vollmer, M. K., Wang, R. H., and Young, D.: The ALE/GAGE/AGAGE Data Base [data set], <http://agage.mit.edu/data>, last access: 1 March 2022.
- Rigby, M., Park, S., Saito, T., Western, L. M., Redington, A. L., Fang, X., Henne, S., Manning, A. J., Prinn, R. G., Dutton, G. S., Fraser, P. J., Ganesan, A. L., Hall, B. D., Harth, C. M., Kim, J., Kim, K.-R., Krummel, P. B., Lee, T., Li, S., Liang, Q., Lunt, M. F., Montzka, S. A., Mühle, J., O'Doherty, S., Park, M.-K., Reimann, S., Salameh, P. K., Simmonds, P., Tunnicliffe, R. L., Weiss, R. F., Yokouchi, Y., and Young, D.: Increase in CFC-11 emissions from eastern China based on atmospheric observations, *Nature*, 569, 546–550, <https://doi.org/10.1038/s41586-019-1193-4>, 2019.
- Rigby, M., Tunnicliffe, R., Western, L., Chawner, H., Ganesan, A., Ramsden, A., Jones, G., Young, D., Ward, R., Stell, A., Nickless, A., and Pitt, J.: ACRG-Bristol/acrg: ACRG v0.2.0 (v0.2.0), Zenodo [code], <https://doi.org/10.5281/zenodo.6834888>, 2022.
- Roten, D., Wu, D., Fasoli, B., Oda, T., and Lin, J. C.: An interpolation method to reduce the computational time in the stochastic Lagrangian particle dispersion modeling of spatially dense XCO₂ retrievals, *Earth and Space Science*, 8, e2020EA001343, <https://doi.org/10.1029/2020ea001343>, 2021.
- Sayegh, A., Tate, J. E., and Ropkins, K.: Understanding how roadside concentrations of NO_x are influenced by the background levels, traffic density, and meteorological conditions using boosted regression trees, *Atmos. Environ.*, 127, 163–175, <https://doi.org/10.1016/j.atmosenv.2015.12.024>, 2016.
- Stanley, K. M., Grant, A., O'Doherty, S., Young, D., Manning, A. J., Stavert, A. R., Spain, T. G., Salameh, P. K., Harth, C. M., Simmonds, P. G., Sturges, W. T., Oram, D. E., and Derwent, R. G.: Greenhouse gas measurements from a UK network of tall towers: technical description and first results, *Atmos. Meas. Tech.*, 11, 1437–1458, <https://doi.org/10.5194/amt-11-1437-2018>, 2018.
- Taylor, T. E., O'Dell, C. W., Crisp, D., Kuze, A., Lindqvist, H., Wennberg, P. O., Chatterjee, A., Gunson, M., Eldering, A., Fisher, B., Kiel, M., Nelson, R. R., Merrelli, A., Osterman, G., Chevallier, F., Palmer, P. I., Feng, L., Deutscher, N. M., Dubey, M. K., Feist, D. G., García, O. E., Griffith, D. W. T., Hase, F., Iraci, L. T., Kivi, R., Liu, C., De Mazière, M., Morino, I., Notholt, J., Oh, Y.-S., Ohyama, H., Pollard, D. F., Rettinger, M., Schneider, M., Roehl, C. M., Sha, M. K., Shiomi, K., Strong, K., Sussmann, R., Té, Y., Velasco, V. A., Vrekoussis, M., Warneke, T., and Wunch, D.: An 11-year record of XCO₂ estimates derived from GOSAT measurements using the NASA ACOS version 9 retrieval algorithm, *Earth Syst. Sci. Data*, 14, 325–360, <https://doi.org/10.5194/essd-14-325-2022>, 2022.
- Western, L. M., Sha, Z., Rigby, M., Ganesan, A. L., Manning, A. J., Stanley, K. M., O'Doherty, S. J., Young, D., and Rougier, J.: Bayesian spatio-temporal inference of trace gas emissions using an integrated nested Laplacian approximation and Gaussian Markov random fields, *Geosci. Model Dev.*, 13, 2095–2107, <https://doi.org/10.5194/gmd-13-2095-2020>, 2020.
- Western, L. M., Ramsden, A. E., Ganesan, A. L., Boesch, H., Parker, R. J., Scarpelli, T. R., Tunnicliffe, R. L., and Rigby, M.: Estimates of North African methane emissions from 2010 to 2017 using GOSAT observations, *Environ. Sci. Tech. Lett.*, 8, 626–632, <https://doi.org/10.1021/acs.estlett.1c00327>, 2021.
- Zammit-Mangion, A., Bertolacci, M., Fisher, J., Stavert, A., Rigby, M., Cao, Y., and Cressie, N.: WOMBAT v1.0: a fully Bayesian global flux-inversion framework, *Geosci. Model Dev.*, 15, 45–73, <https://doi.org/10.5194/gmd-15-45-2022>, 2022.