# A Machine Learning Method for Estimating Atmospheric Trace Gas Concentration Baselines

Kirstin Gerrand[1, 2], Elena Fillola[1, 3], Alistair J. Manning[4, 1], Jgor Arduini[5], Paul B. Krummel[6], Chris R. Lunder[7], Jens Mühle[8], Simon O'Doherty[1], Sunyoung Park[9], Ronald G. Prinn[10], Stefan Reimann[11], Dickon Young[1], and Matthew Rigby[1]

[1]Atmospheric Chemistry Research Group, School of Chemistry, University of Bristol, Bristol, UK
[2]Earth Sciences New Zealand, Private Bag 14901, Kilbirnie, Wellington, New Zealand
[3]School of Engineering Mathematics, University of Bristol, Bristol, UK
[4]Hadley Centre, Met Office, Exeter, UK
[5]Department of Pure and Applied Sciences, University of Urbino, Urbino, 61029, Italy
[6]CSIRO Environment, Aspendale, VIC, Australia
[7]NILU, Kjeller, Norway
[8]Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA
[9]Kyungpook Institute of Oceanography, Kyungpook National University, Daegu, Republic of Korea
[10]Center for Sustainability Science and Strategy, Massachusetts Institute of Technology, Cambridge, MA, USA
[11]Laboratory for Air Pollution/Environmental Technology, Empa, Dübendorf, Switzerland

**Corresponding authors:** Kirstin Gerrand (kigerrand@gmail.com) and Matthew Rigby (matt.rigby@bristol.ac.uk)

**Abstract.** Estimates of trace gas baseline mole fractions in high-frequency atmospheric measurement records are crucial for analysing long-term changes in atmospheric composition. Baseline mole fractions are those that would be observed far from emission sources (and hence are representative of background conditions)~~at specific latitudes in the atmosphere~~. Previous methods for inferring baseline mole fractions have used statistical or meteorological approaches, or, if available, co-measured tracer species thought only to be emitted from non-baseline wind sectors. Combinations of these techniques have also been employed in some applications. Statistical methods typically fit a baseline to the observations themselves, while meteorological methods use atmospheric models of varying complexity to categorise air mass origins. In this paper, we present a novel machine learning method for estimating trace gas baseline mole fractions, which benefits from the physical basis of model-based filtering without the need for running an expensive simulator. Our approach offers the accessibility and computational cost-effectiveness of statistical models, without the associated smoothing or difficulty in identifying rapid baseline variations. By training on historical Lagrangian particle dispersion model outputs, our model learns to predict baseline mole fractions directly from meteorological fields. This advancement opens new avenues for low-latency trace gas time series data analysis, reconstruction of historical baseline trends, and improved utilisation of tracer measurement air mass classification methods.

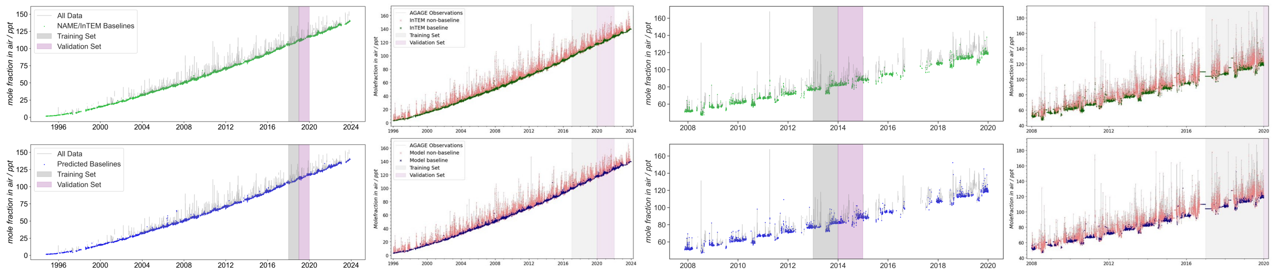
## 1 Introduction

The evaluation of long-term trends in the concentration of atmospheric trace species is important for understanding phenomena such as stratospheric ozone depletion and climate change. However, a challenge associated with the analysis of high-frequency (∼ hourly to daily) trace gas measurements is the separation of “baseline” concentrations from measurements strongly influenced by nearby sources. Here, we define baseline measurements as ~~those representative of~~ concentrations that would be observed ~~far from emission sources at the same latitude as the measurement point~~ at a point in the atmosphere, if it were not influenced by nearby sources or sinks (e.g. Manning et al., 2011). Such baseline time series have become essential for understanding hemispheric or global trends in greenhouse gases (for example, the “Keeling curve” for carbon dioxide (Keeling et al., 2005)), and ozone depleting substances (e.g., Laube et al., 2022; Liang et al., 2022).

Whether the trace gas mole fraction in a particular air mass is characteristic of the regional baseline or non-baseline conditions depends on the interplay of meteorology and fluxes. Advection from regional sources or sinks to a measurement site over timescales of days to weeks will lead to mole fractions that are above or below the baseline. Enhancements above baseline can also be observed under low wind speeds or planetary boundary layer heights (PBLH), when the measurements are particularly sensitive to local fluxes. Even when the influence of regional fluxes on a particular air mass is small, variations in mole fractions are observed due to long-range transport from latitudes or altitudes with very different baseline mole fractions to that of the measurement site (e.g., Arnold et al., 2018; Lunt et al., 2016).

Measurement networks such as the Advanced Global Atmospheric Gases Experiment (AGAGE) collect mole fraction data for numerous greenhouse gases (GHGs) and ozone-depleting substances (ODSs) at several locations around the world at approximately hourly frequency (Prinn et al., 2018). These monitoring stations observe baseline concentrations, overlaid with time-varying enhancements or depletion events for the reasons outlined above. Data filtering is therefore needed to estimate baselines, or categorise the data points that best reflect baseline concentrations.

As an example, Fig. 1 shows AGAGE measurements of the hydrofluorocarbon, HFC-134a ($C_2H_2F_4$), at Mace Head, Ireland, and Gosan, Republic of Korea. Widespread use of this compound as a refrigerant has resulted in increasing global atmospheric abundances (Liang et al., 2022). The measurements at Mace Head, Ireland, are characterised as baseline when air originates from the Atlantic to the ~~West~~west, but these can be overlaid with enhancements as a result, primarily, of the advection of “polluted” air masses from European sources to the ~~East~~ east (raw data are shown as grey lines in Fig. 1, with overlaid coloured crosses indicate baseline/non-baseline). At Gosan, enhancements are seen when air originates from a wider range of wind directions, due to surrounding sources from China to the ~~West~~west, the Korean peninsula to the ~~North~~ north and Japan to the ~~East~~east. Furthermore, during the summer months, intrusions of ~~southern-hemispheric~~ Southern-Hemispheric air are frequently seen, associated with the observation of ~~below-northern-hemispheric~~ below-Northern-Hemispheric mole fractions. The investigator may or may not wish to include baseline mole fractions originating from latitudes that are very different to the measurement station, depending on the application. For example, when estimating the long-term mole fraction trend using observations from Mace Head, Ireland, air masses originating from the tropical Atlantic were removed in Manning et al. (2021), but a summertime baseline more characteristic of the Southern Hemisphere was included in the inverse modelling study using Gosan data in Arnold et al. (2018).




(a) Mace Head, Ireland

(b) Gosan, Republic of Korea

**Figure 1.** Measurement time series for HFC-134a ($CH_2FCF_3$) at a) Mace Head, Ireland, and b) Gosan, Republic of Korea. The observations are shown ~~in grey~~as crosses and coloured according to the baseline/non-baseline label. The top panels show the measurements identified as baseline using the NAME/InTEM footprint-based filtering approach with no statistical filtering applied (green), and the bottom panels show the data points classified by the MLP algorithm ~~’s predictions of baseline points~~ presented here to be baselines (blue). The ~~one-year~~ three-year training period is shown as grey shading and the two validation ~~year is~~ years are indicated by purple shading.

Previous baseline classification or fitting algorithms have broadly used three types of approach: statistical filtering, meteorological filtering or filtering based on co-measurement of some tracer for non-baseline conditions. Measurement-based methods have used species such as $^{222}$Rn or carbon monoxide (CO) to identify air masses substantially influenced by terrestrial sources or anthropogenic activity, respectively (e.g., Chambers et al., 2016, 2013; Yang et al., 2009). Since they require deployment of additional specialist instrumentation, we will not discuss them further here, although we note that our proposed approach could use measurements of these tracers as a training dataset.

~~Pure~~ Baseline identification through pure statistical filtering is seen in many studies, with examples including the use of iterative polynomial fitting with exclusion of outliers, and filtering of certain frequencies using Fourier transforms (O’Doherty et al., 2001; Ruckstuhl et al., 2012; Thoning et al., 1989; Novelli et al., 1998). O’Doherty et al. (2001) outlines ~~in detail~~ the AGAGE baseline algorithm, which iteratively fits a polynomial to the measurement time series and excludes points that are more than 3 standard deviations above the median within some time window (121 days). Similarly, Novelli et al. (1998) used a polynomial fit, including harmonic terms, with a low-pass filter to exclude above-baseline measurements. Their approach was subsequently improved to transform the residuals to and from the frequency domain and apply high and low-pass filters in Novelli et al. (2003). Ruckstuhl et al. (2012) describe the “Robust Estimation of Baselines (REBS)” approach, which uses local regression within some defined time window to estimate the baseline and the distribution of its observed values.

Each of these statistical methods has the advantage of requiring no ancillary data or model simulations to apply, and is computationally efficient. However, due to the use of polynomial fitting or moving windows over which data are excluded or regression is performed, they all implicitly or explicitly apply some smoothing to the data, which may not be characteristic of the true baseline variability. Furthermore, they cannot readily identify or remove baseline values that are characteristic of a latitude different from that of the measurement station (as observed, for example, when southerly air masses arrive at Mace Head or Gosan; Fig. 1).

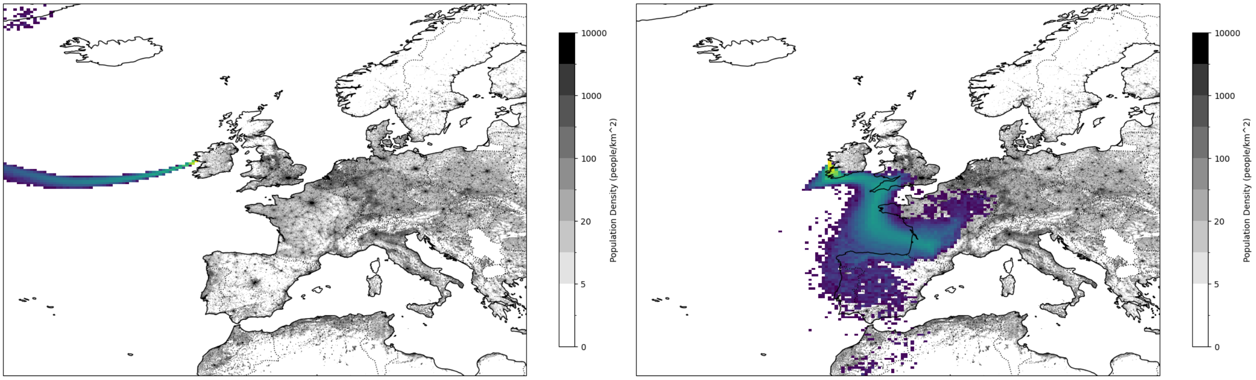
Meteorological filtering techniques have often used air mass back trajectories (estimates of advective transport prior to an observation), or wind sector analysis, to identify air masses that are unlikely to be strongly influenced by regional fluxes. Henne et al. (2008), Lööv et al. (2008), and Salvador et al. (2010) computed back trajectories, and then applied a clustering algorithm to explore patterns in air mass origins. Alternatively, several studies (Derwent et al., 1998b; O'Doherty et al., 2001; Derwent et al., 1998a) have applied a wind sector allocation approach, described in detail by Derwent et al. (1998c) to isolate air masses originating from some wind sector thought not to be strongly influenced by local fluxes.

In an extension of back-trajectory-based methods, recent studies have used trace gas source-receptor relationships ("footprints"), calculated using Lagrangian particle dispersion models (LPDMs), to estimate the full influence of transport and mixing on observed concentrations. A range of particle dispersion models have been applied to baseline estimation, including the UK Met Office Numerical Atmospheric Modelling Environment (NAME) (Ryall et al., 1998; Manning et al., 2011), which we use in this study. LPDMs estimate footprints by considering the transport of an ensemble of hypothetical gas particles backwards in time, driven by archived meteorological fields. For the NAME simulations used in this study, these are obtained from the UK Met Office Unified Model analyses (Cullen, 1993). LPDM footprints quantify the contribution to the observed concentration of a unit emission from the surface at each grid cell surrounding the measurement point (Manning et al., 2011).

Manning et al. (2021) combined LPDM footprints with a population density map to identify air masses that were potentially influenced by anthropogenic emissions (illustrated in Fig. 2). A flux proportional to population density was transported through the NAME model atmosphere to produce a synthetic anthropogenic tracer at each measurement site. When the concentration of this tracer fell below some arbitrary threshold, the corresponding air mass was labelled as being representative of the baseline. This method, part of the Inversion Technique for Emission Modelling (InTEM) (Manning et al., 2011), has been applied in a number of studies (Arnold et al., 2018; Lunt et al., 2021). In addition to baseline classification, the method also categorises air masses further into a set of classes specific to a given site. For Mace Head, Ireland, for example, air masses are either "southerly" (when air masses originate from lower latitudes, specific for European sites), "local" (where local influences may dominate due to low ventilation conditions), "polluted" (where potential anthropogenic influence is high), "mixed" (when there is a combination of source types), "upper troposphere" (when air masses have descended from higher altitudes), or baseline. Figure 2 shows two example footprints for Mace Head, Ireland, superimposed on a population density plot to illustrate the approach. Figure 2(a) shows conditions consistent with baseline mole fractions for compounds whose fluxes are highest over land; the footprint is primarily over the ocean and does not deviate substantially in latitude from the measurement location. Figure 2(b) shows an instance where the air mass originates from more populated and industrialised regions, where enhanced concentrations are typically observed.

Unlike statistical filters, ~~meteorological (or~~ model-based ~~filtering) does~~ baseline classifications do not impose smoothing on the dataset and ~~the baseline classification~~ airmass categories can be justified based on ~~physical considerations. However, the methods based on atmospheric model-derived air histories are computationally expensive and rely on the availability of the appropriate meteorological datasets needed to drive models.~~

a more complete range of physical considerations than simple meteorological filtering (e.g., based on small numbers of trajectories or wind sectors). Therefore, this approach is generally considered the gold standard for baseline classification,




(a) Baseline footprint

(b) Non-baseline footprint

**Figure 2.** Two example NAME footprints showing incoming air masses to Mace Head atmospheric research station that illustrate meteorology consistent with, a) baseline and b) non-baseline observations. The footprint colour refers to the susceptibility of the measurement to emissions from that grid cell on a logarithmic scale; green represents higher values, and purple, lower (colourbar not shown). Population density is also shown ~~,~~ as a proxy for anthropogenic flux magnitude (grey shading).

when co-emitted tracer measurements are not available. However, such methods are computationally costly and technically challenging to implement. Here, we present a baseline classification algorithm using a machine learning (ML) approach that emulates an LPDM-based filter. ~~This~~ Our approach preserves the benefits of a meteorological filter for a fraction of the computational cost.

ML ~~is a branch of artificial intelligence that enables computers to quickly analyse data and identify relationships and patterns within. It~~ has been employed successfully for various applications in atmospheric chemistry, including the prediction of particulate matter concentrations ~~,~~ and nitrogen dioxide modelling ~~and estimating trace gas footprints . Artificial neural networks, a type of supervised learning algorithm whose architecture mirrors the workings of the human nervous system and brain, have been demonstrated to be very useful in atmospheric science, especially when dealing with complex systems with numerous non-linear relationships . The multilayer perceptron (MLP) is one example of a widely used neural network architecture . Decisions are made by sending signals from an input layer to an output layer; data are processed in each step, with activation functions applied to introduce non-linearity to the computation . Another ML algorithm applied in atmospheric chemistry is the random forest; a collection of decision trees each make their own prediction based on a unique subset of data, and the majority vote is the final prediction~~ (Brokamp et al., 2018; Masih, 2019). Also, ML-based LPDM footprint emulation is an increasingly active area of research; recent methodologies such as Fillola et al. (2023), FootNet (He et al., 2025) and GATES (Fillola et al., 2026) demonstrate the potential of ML-based surrogates in this field. Whilst this study focuses on baseline identification rather than full footprint reconstruction, these approaches show the broader capabilities of ML-based frameworks for emulating LPDM-based diagnostics.

The aim of this study is to investigate the application of ML for the classification of trace gas baseline air masses, in order to accurately recreate the air mass categories obtained from the Met Office NAME/InTEM LPDM-based algorithms. The method identifies data points likely representing background conditions, separating them from those with substantial local emission contributions. It does not attempt a quantitative decomposition of individual observations into separate background and local source components. We demonstrate this method at nine different AGAGE sites, in a range of meteorological regimes and for a range of compounds. The baseline classification is based only on meteorological inputs, namely wind speed and direction, boundary layer height and surface pressure.

## 2 Methodology

### 2.1 Data

#### 2.1.1 Baseline Flags

For the purpose of training our ML algorithm, a dataset of flags that categorise air masses as “baseline” or “non-baseline” using some independent method is required. Here, we use air mass labels, based on outputs from NAME/InTEM (Manning et al., 2021; Jones et al., 2007). The method categorises air masses bi-hourly, with categories tailored for each site (e.g., “baseline”, “polluted”, “local”, “mixed”, “southerly”, “upper troposphere”, for Mace Head~~, as described above~~). For this work, the categories ~~are~~ were simplified to a binary “baseline” versus “non-baseline” label (grouped non-baseline categories). This restriction was applied to focus on the robust identification of baseline data points for long-term atmospheric composition trend analysis or regional inverse modelling. In the InTEM framework, an additional statistical filter is applied to the derived baseline on a species-by-species basis, to remove any remaining outliers, but the flags used in this study are based only on the footprint-based filter. The algorithm presented here could be retrained with or without the air mass origin conditions described above (e.g., the exclusion of air masses from latitudes different from the measurement site and the Gosan exception), by modifying the training dataset.

#### 2.1.2 Meteorological Data

Meteorological data were obtained from the European Centre for Medium-Range Weather Forecasting (ECMWF) ERA5 re-analysis dataset (Hersbach et al., 2020). These data were downloaded at hourly intervals on a longitude/latitude horizontal grid and hybrid coordinates in the vertical and used as an input to the ML algorithm. The ECMWF meteorological fields were used, rather than the Met Office UM fields that were used to drive NAME, as they were more readily available for longer time periods. An intercomparison of the two products showed close agreement at the AGAGE measurement stations, as would be expected since both products assimilate similar meteorological observations. For example, calculated mean absolute percentage errors gave a less than 10 % error in wind direction across a six-month sample period (January to June 2015) at Mace Head, Ireland~~, in the wind direction~~. However, wind speed saw a slightly higher error (15.3 %) with UM speeds generally surpassing that of the ERA-5 met products (see Supplementary Material).

**Table 1.** Details of selected AGAGE sites, with site designation and coordinates given. The percentage of native baseline flags across the full investigation period (training, validation and testing) is also given.

| *Site Name* | *Station Designation* | *Coordinates* | *Native Baseline (%)* |
|---|---|---|---|
| Kennaook/Cape Grim, Australia | CGO | 40.6833° S, 144.6894°~~E~~ E | 29.633 |
| Monte Cimone, Italy | CMN | 44.1932° N, 10.7014°~~E~~ E | 8.303 |
| Gosan, Republic of Korea | GSN | 33.2924° N, 126.1616°~~E~~ E | 19.424 |
| Jungfraujoch, Switzerland | JFJ | 46.5478° N, 7.9859°~~E~~ E | 20.316 |
| Mace Head, Ireland | MHD | 53.3267° N, 9.9046°~~W~~ W | 24.697 |
| Ragged Point, Barbados | RPB | 13.1651° N, 59.4321°~~W~~ W | 24.509 |
| Cape Matatula, American Samoa | SMO | 14.2474° S, 170.5644°~~W~~ W | 46.449 |
| Trinidad Head, United States | THD | 41.0541° N, 124.151°~~W~~ W | 11.713 |
| Zeppelin, Svalbard | ZEP | 78.9072° N, 11.8867°~~E~~ E | 37.240 |

### 2.1.3 Mole Fraction Observations

Trace species mole fraction data were obtained from AGAGE (Prinn et al., 2025) at approximately two hour intervals. Measurements are provided as dry air mole fractions (in ppt, equivalent to pmol mol$^{-1}$ or ppb, nmol mol$^{-1}$). We demonstrate our algorithm at nine AGAGE sites, covering four continents and spanning a range of remote and relatively "polluted" environments. Details of the chosen sites are outlined in Table 1. All sites have at least 10 years of data, with most having more than 20. This is equivalent to over 200,000 measurements across the nine sites.

## 2.2 Preprocessing

The mole fraction and meteorological datasets were combined by aligning them to the InTEM baseline flag time period. ~~A tolerance of one hour was set to avoid significant extrapolation of the mole fraction data.~~

Initial exploration of the datasets showed a significant imbalance between the baseline and non-baseline classes, with seven of the nine sites seeing baseline points being less than one-third of the dataset. ~~This was mitigated by undersampling the majority class (non-baseline), which we found improved the predictive ability of the models. A random sample of non-baseline points was removed, and we used a baseline:~~The effect of class balance on model performance was explored by varying the percentage of baseline points in the training set (by randomly under-sampling the non-baseline ~~ratio of 4:1 in our training set~~class), and by implementing a sample weight system that assigns more importance to the baseline observations, accounting for the natural class imbalance without needing to remove any training data points. The optimal baseline proportion and sample weighting was found to be both model- and site- specific, although the proportion of baselines was always within 20–40% and the sample weight did not exceed 3. The exact values can be found in the Supplementary Material.

### 2.2.1 Model Inputs

The meteorological parameters included as inputs or features to the ML model were the eastward and northward wind components ($m\ s^{-1}$) at 10 meters above ground level and at two pressure levels, 850 and 500 hPa, surface pressure (Pa), and boundary layer height (m). ~~The wind~~ These data were interpolated to a 17-point grid system around each site, based on two 3x3 grids covering $\pm$ 5 and $\pm$ 10 ° latitude and longitude. To provide the algorithm with further information on changes in ~~wind speed and direction, u- and v-wind components~~ atmospheric conditions, all meteorological variables were also provided ~~for six~~ at time intervals prior to the measurement point. The number and spacing of these intervals were treated as hyperparameters and optimised individually for each model type and site, with intervals up to 72 hours prior to the measurement ~~point~~time tested, to allow the model to capture the recent meteorological history for each location. It was found that, in most cases, adding features at 6, 12, 18 and 24 hours prior to the measurement time yielded the best performance. Additionally, two temporal variables were added into the dataset to represent the time of day and the day of year. These two variables were integers ranging from 0 to 23 and 1 to 366, respectively. ~~This gave a total of 209 input variables.~~ A list of all input variables can be found in the Supplementary Material.

### 2.2.2 ~~Dimensionality Reduction~~

~~To reduce the size of the input dataset and reduce the computational cost of model training and evaluation, a principle component decomposition of the input dataset was used for one of the ML models we tested. As the inputs included a~~ Normalisation of the meteorological inputs was explored to account for the range of physical quantities with different units and dynamic ranges~~, the data was first standardised by normalising by the mean and standard deviation per variable per atmospheric level.~~. Features were scaled separately for each variable category across the 17-point grid (e.g., all eastward wind components at a given height) ~~. The first 20 principal components were used as ML model inputs from this standardised dataset. These components were found to explain over 80 % of the input data variance, whilst reducing the dimensionality of the input dataset by an order of magnitude~~by their mean and standard deviation. This normalisation was also treated as a hyperparameter, allowing testing with all model types.

## 2.3 Machine Learning Models

A multilayer perceptron (MLP) was trained to predict baseline classifications for each of the nine AGAGE sites. The method was also tested using ~~a~~ two tree-based ~~method (~~methods, a random forest ~~algorithm)~~and a gradient boosting algorithm, the results of which can be found in the Supplementary Material. ~~These methods outperformed alternatives, such as gradient boosted trees,~~ The three discussed methods were chosen as they outperformed alternatives in early testing, but it is likely that several other suitable architectures also exist.

~~The random forest method used the reduced dimension dataset. However, the MLPs performed best when provided with the full dataset. These choices were determined by balancing model evaluation metrics and training time.~~

### 2.3.1 Model Training

~~Models were trained individually~~ Models were trained individually for each site. Whilst using one universal model applicable to all sites would be desirable for consistency and reduced training time, early tests showed that a more tailored approach is required here. The classification of baselines is driven by population density maps and site-specific rules, as Manning et al. (2011) does with Mace Head, Ireland, and the transport patterns influencing each site will be strongly driven by the geographical features surrounding that site. Generalisation across sites would require training a more complex model with a substantially larger feature set. The approach taken here, of training a model for each site~~. One year of data was used (approximately 1000~~, means that the geographical characteristics of each location are learned implicitly.

Three years of data were used for model training (approximately 3000 data points)~~for model training; 2018 was arbitrarily chosenfor all sites except Gosan, which experienced substantial periods of down-time that year, so 2013 was used instead. The year~~; 2017-2019 were arbitrarily chosen. The two years following the training ~~year~~ years was used for the validation set (~~2019 for all sites except 2014 for Gosan~~2020 and 2021), and the rest of the dataset was used for testing. Given the substantial auto-correlation on the training dataset, characteristic of synoptic variability, it was important that the training and testing dataset be separated in time by greater than synoptic timescales (∼5-10 days). Three years were chosen for training following an investigation into the balance between length of the training period, model performance, and training time. Improvements seen in model performance when further increasing the volume of training data were too minor to justify the consequent increase in computational demands, as shown in the Supplementary Material. Two years were chosen for validation as one year was found to not be representative enough to generate reliable metrics at sites with a low native baseline ratio (see Table 1); this validation time period was then applied across all sites for consistency. Using the remaining data for testing allowed the models to be evaluated over long periods, testing their ability to account for a wide range of meteorological conditions.

Model hyperparameters were optimised using grid searches. These grid searches were performed independently for each site, but similar parameter choices were seen in the trained models throughout. For example, all MLP models used the ReLU (rectified linear unit) activation function, which outputs zero for all negative input values but leaves positive ones unchanged to introduce non-linearity to the dataset (Eckle and Schmidt-Hieber, 2019). All sites saw models with ~~either four or five~~ three layers in their optimised MLP, with the number of hidden layers and the number of neurons they contain varying. The Mace Head model, for example, consists of an input layer with ~~209 neurons~~ 682 neurons (all input features plus additional meteorological inputs at 6, 12, ~~two hidden layers with 200~~ 18 and ~~150 neuronsrespectively~~24 hours prior to the measurement time), one hidden layer with 64 neurons, and a single-neuron output layer. Full hyperparameter sets can be found in the Supplementary Material.

A confidence threshold was introduced when making predictions, meaning that the model could only assign a baseline label when the associated confidence exceeded ~~80 %~~this value (plots showing model-derived baseline confidence are provided in the Supplementary Material). This approach was introduced to improve the reliability of the model and reduce the occurrence of false positives. The threshold was treated as a hyperparameter, but optimised as a final step in model training. Again, it was found to be model- and site- specific, and values ranged from 50 to 70 %.

### 2.3.2 Model Evaluation

The model was trained to maximise F1 score (Eq. (1)), a measure of performance in binary classification problems that averages the precision and recall of a model (Eqs. (2) and (3)). Respectively, *precision* indicates the fraction of ~~baselines identified correctly~~ all baselines correctly identified as baseline, and *recall* identifies the fraction of ~~predicted baselines that were "true" baselines . The model is evaluated through the use of these three metrics, as well as with a misclassification rate, which represents the proportion of instances in a dataset that the model incorrectly classifies~~baselines that are correctly identified by the model. The final values that are quoted for these metrics were calculated by considering all ~~the~~ data, except for those points used for training and validation. In most cases, the test set exceeded 20 years of data. Note that these metrics define the model's ability to separate baseline and non-baseline samples based only on the air mass meteorology, and so do not consider associated mole fraction values.

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{1}$$

$$Precision = \frac{N(true\,positives)}{N(true\,positives) + N(false\,positives)} \tag{2}$$

$$Recall = \frac{N(true\,positives)}{N(true\,positives) + N(false\,negatives)} \tag{3}$$

To understand the performance of the ML model in application, the predicted labels can be applied to the atmospheric measurements of a particular compound, and compared to the the NAME/InTEM baseline-only concentration time series. Measurement time series are often aggregated into monthly means in many applications, to remove the influence of short-term fluctuations (Laube et al., 2022; Liang et al., 2022). Therefore, the model was also evaluated by comparing the monthly means derived from the ML emulation to those from the InTEM baseline using three metrics; Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The monthly mean baseline was calculated by collecting and taking the average of all mole fractions labelled as baseline in a calendar month.

To explore the utility of the method across many sites and species, a coefficient of variation (CV) is calculated. This quantifies the noise in the InTEM-derived baseline labels by finding the ratio between the standard deviation and mean across the timeseries, after removing any seasonal variability and the long-term trend. STL decomposition (Seasonal-Trend decomposition based on LOESS) (Cleveland et al., 1990) is applied to the monthly means of the true baseline mole fractions, decomposing each monthly mean $m_t$ into three components:

$$m_t = T_t + S_t + \varepsilon_t \tag{4}$$

where $T_t$ is the long-term trend, $S_t$ is the seasonal component, and $\varepsilon_t$ is the residual. The components $T_t$ and $S_t$ are calculated from the training data (Seabold and Perktold, 2010). As STL decomposition requires a complete time series, any missing months are linearly interpolated prior to decomposition.

For a particular dataset of monthly baseline mole fractions $\mathbf{m} = \{m_t\}_{t=1}^{T}$, and residuals $\boldsymbol{\varepsilon} = \{\varepsilon_t\}_{t=1}^{T}$, the CV is then defined as:

$$CV(\mathbf{m}) = \frac{\sigma(\boldsymbol{\varepsilon})}{\bar{m}} \tag{5}$$

where $\sigma(\boldsymbol{\varepsilon})$ is the standard deviation of the residuals and $\bar{m}$ is the mean of $\mathbf{m}$.

#### 2.3.3 Feature importance

We determine the variables that are most important for ~~determining model performance by carrying out~~ model performance using a permutation importance analysis (Breiman, 2001; Altmann et al., 2010). ~~This approach works~~ Permutation importance is calculated by randomly shuffling the values of a single feature group and observing the resulting change in the model~~'s performance . This~~ 's performance (Fillola et al., 2023). The process is repeated multiple times to ensure robustness~~. The~~ , and the importance of a feature is determined by the extent to which ~~the model 's~~ model performance degrades when the ~~feature's values~~ values of the given feature group are permuted. To account for correlations between individual input features, a known limitation of the permutation importance approach, variables were split into feature blocks prior to analysis (Fillola et al., 2023). Meteorological variables were combined across all times and locations, resulting in five feature groups: the wind u- and v-components, boundary layer height, surface pressure, and temporal features (time of day and day of year). The three most important ~~features~~ feature groups for each site are shown in the Supplementary ~~material~~Material.

## 3 Results and Discussion

~~In~~ The three model types performed similarly across all nine sites, showing that the limitations of the method lie within the input features rather than the model architecture. For simplicity, in this section we evaluate the MLP model only by considering its ability to correctly identify baseline air masses based on meteorological inputs. We then compare time series for a selection of AGAGE gases. Summary results for all sites are presented, and examples are shown for Mace Head, Ireland, and Gosan, Republic of Korea, chosen because of their very different local emissions~~and~~ , meteorological regimes (e.g., Fig. 1) and native baseline ratios. Results for the random forest ~~algorithm~~ and gradient boosting algorithms are presented in the Supplement.

### 3.1 Baseline Identification

Overall model performance for correctly classifying baseline versus non-baseline air masses was analysed by considering confusion matrices for each site. These matrices compare the predicted and "true" classifications of the testing set. Figure 3 shows the MLP confusion matrices for the nine AGAGE sites considered in this work. The outcomes of the confusion matrices

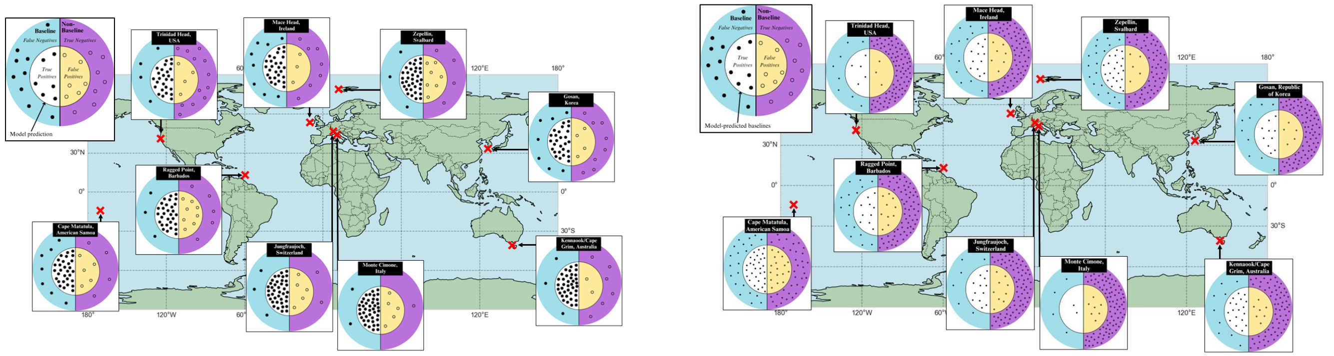



**Figure 3.** A map showing the locations of the nine AGAGE sites, with confusion matrix-derived plots at each location. Each confusion matrix was normalised, to reduce the visual impact of differences in testing set sizes; each point represents approximately ~~2~~ 1 % of the total test set (rounding means that the total number of points on each plot range from ~~49~~ 99 to ~~51~~101). The left half of each circle represents true baseline points, and the right half true non-baseline points. The inner circle shows the MLP model prediction of baseline points. The key in the top left indicates where true or false positives and negatives lie, as explained in the main text.

have been normalised by testing set size and plotted as points on a precision/recall "bullseye" diagram. The inner circles are the MLP model predictions of baseline values. The left half of each diagram represents true baseline points, and the right-hand side are true non-baseline points. Therefore, the correctly identified baselines (true positives) lie in the left-hand segment of the inner circles (white). False positives, or points incorrectly identified as baseline, are in the yellow segment to the right of the inner circle. Points in the blue segment (left-hand side of outer circle) represent missed baselines, or false negatives, whereas the purple area (right-hand side of the outer circle) are true negatives; points identified as non-baseline that are indeed non-baseline. A perfect model would only have points in the white area and the purple area. The fraction of points in the left half of the inner circle (white) ~~,~~ compared to the total number in the inner circle (white and yellow) ~~,~~ is the precision. The recall is the ratio of points in the inner left circle (white) to the total number of points on the left (white and blue).

The distribution of points in Fig. 3 reflects optimisation of F1 score in the MLP training, in which a high precision and recall is rewarded jointly; in most cases, the majority of predicted baseline points are true positives (precision) ~~, and most~~ and many of the available baseline points have been correctly identified (recall). These statistics are also summarised in Table 2. ~~The tableindicates that precision, recall and F1 score typically lie above 0.8. However~~As seen in the table, there is some variation between sites, with highest overall scores being obtained at Kennaook/Cape Grim, Australia, and poorer performance at Monte Cimone, Italy, which sees particularly low recall (~~0.57~~0.395). These differences are likely a product of both the complexity of the meteorology at the site and the complexity of surrounding fluxes; Kennaook/Cape Grim is a coastal site, whose meteorology is dominated by strongly prevailing westerly flows, whereas Monte Cimone is a mountain site in a region of complex topography. ~~Therefore, our~~ Kennaook/Cape Grim, therefore, sees more baseline meteorological events and so has a higher proportion of baseline data points. Sites with lower baseline proportions (such as Monte Cimone), are more challenging. Our features, which are a subset of meteorological variables on a relatively low-density grid around each site, may not capture

**Table 2.** A tabular summary of the final MLP model outcomes, showing precision, recall and F1 score values for each of the AGAGE sites. Scores can range from 0 to 1, with higher scores indicating higher performance.

| *Site Name* | *Station Designation* | *Precision* | *Recall* | *F1 score* |
|---|---|---|---|---|
| Kennaook/Cape Grim, Australia | CGO | ~~0.942~~ 0.711 | ~~0.930~~ 0.658 | ~~0.936~~ 0.683 |
| Monte Cimone, Italy | CMN | ~~0.897~~ 0.254 | ~~0.962~~ 0.355 | ~~0.928~~ 0.296 |
| Gosan, Republic of Korea | GSN | ~~0.886~~ 0.652 | ~~0.819~~ 0.507 | ~~0.851~~ 0.570 |
| Jungfraujoch, Switzerland | JFJ | ~~0.888~~ 0.575 | ~~0.935~~ 0.661 | ~~0.911~~ 0.615 |
| Mace Head, Ireland | MHD | ~~0.906~~ 0.638 | ~~0.794~~ 0.462 | ~~0.846~~ 0.536 |
| Ragged Point, Barbados | RPB | ~~0.811~~ 0.490 | ~~0.916~~ 0.382 | ~~0.860~~ 0.429 |
| Cape Matatula, American Samoa | SMO | ~~0.866~~ 0.628 | ~~0.822~~ 0.626 | ~~0.844~~ 0.627 |
| Trinidad Head, United States | THD | ~~0.878~~ 0.660 | ~~0.874~~ 0.405 | ~~0.876~~ 0.502 |
| Zeppelin, Svalbard | ZEP | ~~0.844~~ 0.817 | ~~0.928~~ 0.566 | ~~0.884~~ 0.669 |

well the flows in regions of more inhomogeneous topography and meteorology. Furthermore, in some cases, the outer grid (± 10 degrees) may not be enhancing model performance, but may instead add noise, if it is frequently in a different meteorological regime to the measurement site (e.g., in a different meteorological hemisphere, as may sometimes be the case at Ragged Point, Barbados, and Cape Matatula, American Samoa). Restricting the domain to the inner grid (± 5 degrees) and/or replacing the outer grid with a smaller, higher-resolution one, may improve model performance at some sites.

Our feature importance analysis (Section 2.3.3) reveals that wind components are typically the most critical for model performance, with PBLH and surface pressure generally being of lower importance (see Supplementary Material). However, it should be noted that these variables will be strongly correlated with one-another, so it is not possible to fully isolate the influence of each variable individually. Additionally, in most cases, the analysis showed that the models were not learning much from the two temporal features (time of day and day of year), and so these can likely be removed without negatively impacting model performance. This feature importance pattern aligns with that seen in ML-based LPDM footprint emulation studies (Fillola et al., 2023; He et al., 2025).

### 3.2 Baseline mole fractions

Whilst the above analysis demonstrates the model's ability to categorise air masses into baseline or non-baseline points, perhaps the most important test of the algorithm is in its ability to reproduce quantities that are used in real-world applications. Here, we focus on the simulation of baseline mole fractions and evaluate the model's ability to calculate robust baseline monthly means, which are commonly used to track changes in greenhouse gases and ozone depleting substances (e.g. Laube et al., 2022; Liang et al., 2022; Gulev et al., 2021).

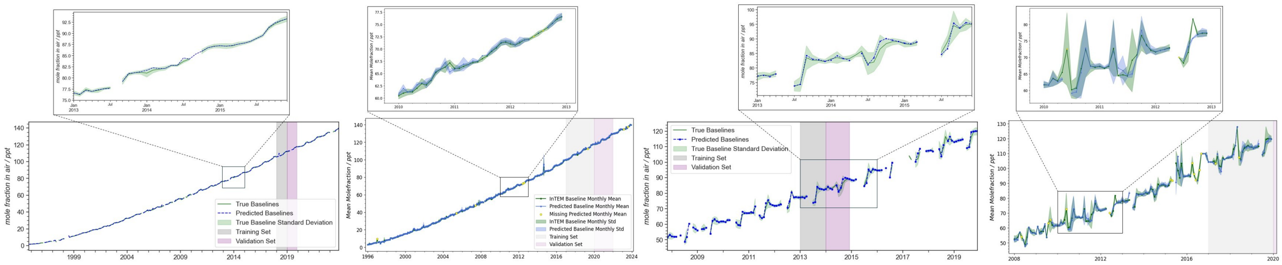
**Table 3.** Atmospheric species used in this study. Atmospheric lifetimes are taken from Burkholder et al. (2023). Primary sources are from Laube et al. (2022); Liang et al. (2022). References describing the AGAGE mole fraction data are provided for some species in addition to Prinn et al. (2018, 2025).

| *Species Name* | *Chemical Formula* | *Atmospheric lifetime (years)* | *Primary sources* | *Data citation* |
|---|---|---|---|---|
| Methane | $CH_4$ | 12 | Wetlands, fossil fuels, agriculture, waste | Cunnold et al. (2002) |
| Nitrous Oxide | $N_2O$ | 109 | Agriculture | |
| CFC-12 | $CCl_2F_2$ | 102 | Refrigerants, aerosol propellants | Cunnold et al. (1983) |
| HCFC-22 | $CHClF_2$ | 12 | Refrigerants | Simmonds et al. (2017) |
| HFC-125 | $CHF_2CF_3$ | 31 | Refrigerants | Simmonds et al. (2017) |
| HFC-134a | $CH_2FCF_3$ | 14 | Refrigerants | Simmonds et al. (2017) |
| Dichloromethane | $CH_2Cl_2$ | 0.5 | Solvents, industrial processes | Simmonds et al. (2006) |
| Methyl bromide | $CH_3Br$ | 0.8 | Fumigants, industrial processes | |
| Sulfur hexafluoride | $SF_6$ | 850–1280 | Electrical insulation, magnesium production | Rigby et al. (2010) |
| Carbon tetrafluoride | $CF_4$ | 50,000 | Aluminium production, semiconductor manufacture | Simmonds et al. (1983) |

The model was applied to a subset of 10 atmospheric trace species measured by the AGAGE network (Prinn et al., 2025). These 10 compounds were chosen to span a range of sources, atmospheric lifetimes, and atmospheric histories (e.g., those that are growing in the atmosphere, such as HFCs, versus those that are declining, such as CFCs).

Examples of the baseline flags applied to HFC-134a measurements at Mace Head and Gosan are shown in Fig. 1. The figure reflects the accuracy of the categorisation shown in Fig. 3; the majority of the InTEM baseline points are identified correctly. The influence of false positives is seen as a number of apparently above-baseline points that are categorised as baseline (noting that a small number of these elevated points ~~appear to be~~ are incorrectly flagged as baseline in InTEM, which is why their full algorithm includes a subsequent statistical filtering step). At Gosan, similar to InTEM, the MLP algorithm identifies baseline values in the summer period, which sees Southern Hemispheric air intrusions. The rapid variation between ~~northern and southern hemispheric~~ Northern and Southern Hemispheric baselines would be very difficult to detect with statistical filters, which rely on baselines being smoothly varying. Similar plots for all other sites and the chosen sample species (those defined in Table 3) are shown in the Supplementary Material.

When ~~downsampled into monthly averages~~ calculating the monthly averages from the baseline-labelled observations, the influence of false positives and false negatives is strongly muted. This is demonstrated in Fig 4, which shows that for almost all months, the error in the MLP model-calculated HFC-134a monthly mean at Mace Head and Gosan is substantially smaller than the baseline variability ($1\sigma$ standard deviation in the baseline) within a month. Where outliers occur, these tend to be




(a) Mace Head, Ireland (b) Gosan, Republic of Korea

**Figure 4.** Baseline monthly means for HFC-134a at a) Mace Head, Ireland; b) Gosan, Republic of Korea. The NAME/InTEM baseline-derived values are ~~in~~ represented by the green line, with associated $1\sigma$ variability within each month shown as green shading. The blue line shows the monthly means of data points that the MLP ~~model-predicted values are in~~ model classified as representative of baseline conditions. The $1\sigma$ variability within the model predicted monthly means is shown as blue shading. A subset of the dataset is shown in more detail in the top panels. Missing months are shown by yellow dots, and arise when a given month has no observations predicted as baseline by the model.

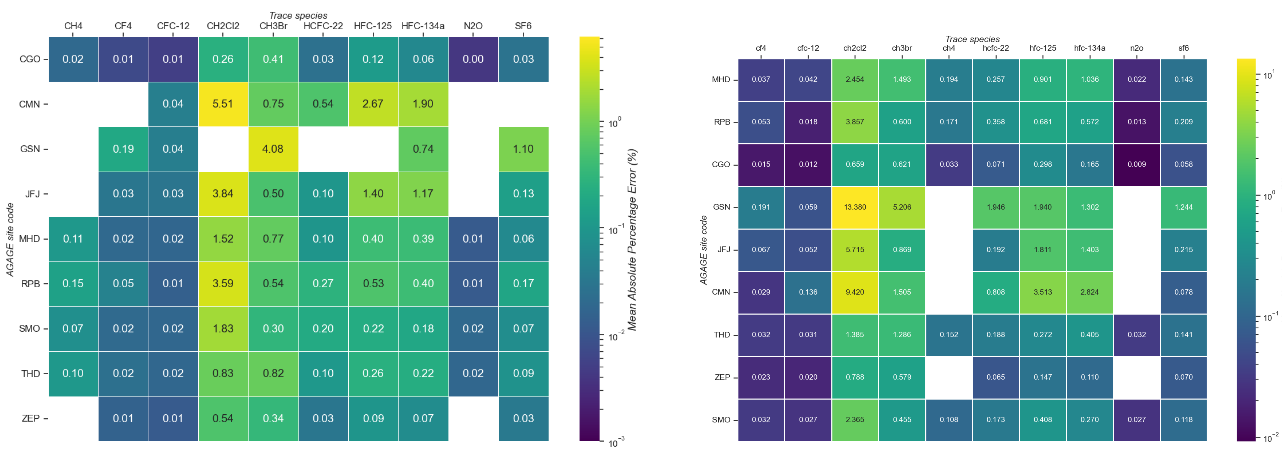



**Figure 5.** Mean Absolute Percentage Errors (MAPE) for the MLP model for selected AGAGE species across the nine monitoring sites. White denotes species that were not available at a particular site.

during months where there were relatively few data points, meaning that any mischaracterisation of the baseline can have a disproportionate influence.

The overall performance of the algorithm applied to each species is summarised in Fig. 5, which shows the ~~absolute~~ percentage error in the MLP-based monthly mean when compared to the InTEM-based equivalent, which is treated as the absolute truth here. ~~The smallest values of MAPE appear to be~~

### 3.3 Coefficient of Variation and Model Utility

The CV metric demonstrates that variability in the true baselines is an indicator of model performance, as quantified by MAPE. As shown in the Supplementary Material, MAPE is generally higher for gases with high variability in their "true" baselines, compared to those with smaller CV. The smallest values of CV are associated with compounds that have relatively small gradients in the background atmosphere (e.g., CFC-12, ~~$CF_4$,~~ $N_2O$)~~, which presumably reduces the impact of false positives or negatives. Likely for the opposite reason, shorter-lived compounds ,~~ . Shorter-lived compounds such as dichloromethane, which tend to exhibit substantial seasonal variability and zonal gradients, tend to show the ~~poorest model performance.~~

~~Mean Absolute Percentage Errors (MAPE) for the MLP model for selected AGAGE species across the nine monitoring sites. White denotes species that were not available at a particular site.~~ highest variability in their true baselines. This pattern is observed across all sites. We therefore recommend caution when applying the method to species with a high CV relative to other species at the same site.

### 3.4 Computation and expected use cases

The computational cost of our baseline classification model is negligible compared to the calculation of LPDM footprints; on a standard desktop computer, the models consistently took less than a minute of wall time to train (~~1 year~~ 3 years of data). Prediction of 1 year of baseline flags for ∼hourly data, for the most part, takes less than 10 seconds. This should be compared to the calculation of LPDM footprints, which takes approximately 10 core minutes for each data point (the calculation of a year of baseline flags by combining the LPDM footprints and a tracer flux field is on the order of seconds to minutes of core time). The proposed ML surrogate is therefore approximately five orders of magnitude faster than the full-physics approach.

For very long time series, such as those presented in Figs. 1 and 4, the main cost associated with our model is the retrieval of subsets of meteorological analyses, which can take a substantial time due to network and storage latency from some archival services (although, of course, substantially larger data volumes are required to run the LPDM).

The envisaged use-cases for the ML algorithm are partly related to its small computational cost, and partly related to its ability to categorise baselines without the need for LPDM simulations. For example, we envisage that this algorithm can be built into data visualisation and quality assurance software, so that mole fraction baseline trends can be examined by data owners with low latency. This is compared to the current situation, in which LPDM runs need to be performed and integrated into the software, often by different research teams to those making the observation data, and often delayed by the availability of appropriate archived meteorology. These lags in the system can delay the analysis by weeks to months. Furthermore, our algorithm allows us to calculate baseline trends further back in time than is currently possible; at present, ~~we only have NAME footprints from 1998~~ footprints have been calculated with NAME from 1989 onwards, and it would require a major resourcing effort (both in terms of staff time and computation) to extend these model runs to the beginning of the AGAGE record (1978). Datasets such as ERA5 cover this period and earlier, and can be used in the model prediction step. Finally, we envisage that this algorithm can be used to extend tracer-based methods beyond the period during which measurements were made. Consider, for example, a campaign in which $^{222}Rn$ was used to derive baseline flags for one year at some location. Our algorithm can use these measurement-based flags in place of the NAME runs, and then be used to provide consistent flagging indefinitely beyond the study period.

These expected benefits must be weighed against the limitations of the model. Whilst we anticipate that improvements in the model architecture or training will be possible, it will always be an approximation of the training data, which is itself, in this case, model-based. Therefore, categorisation errors occur in the form of false positives (points labelled as baseline that are not), or false negatives (baselines that are missed), and so the algorithm should not be used if a precise categorisation is needed of a small number of samples. However, if aggregating over multiple measurements, for example, when calculating a monthly mean of $\sim$ hourly data, we have shown that the model has high skill in the majority of cases.

## 4 Conclusions

We have presented a neural network-based model for baseline air mass classification, based only on meteorological data. The model preserves many of the benefits of LPDM footprint-based classification methods, but at a fraction of the computational cost. Precision ~~and recall~~ values of around ~~0.8~~ 0.6 or higher were achieved for most sites examined here, and baseline monthly means were retrieved with uncertainties that were, in most cases, substantially smaller than baseline variability. We propose that this model can add value in low-latency data analysis and for extending baseline categorisation to time periods for which model simulations or co-measured tracers are not available.

*Code and data availability.* The code for the MLP models, trained models, data processing and evaluation tools are available at https://github.com/openghg/ml-baselines and https://doi.org/10.5281/zenodo.20484075. Training data (InTEM flags and extracted meteorology) are available at https://doi.org/10.5281/zenodo.20483977. Version 20250123 of the AGAGE data was used in this study (Prinn et al., 2025).

*Author contributions.* KG, EF and MR designed the research. KG performed the research under the supervision of EF and MR. AJM provided InTEM baseline flags. All other authors provided observation data, and all authors contributed to writing the manuscript.

*Competing interests.* The authors declare no competing interests.

*Acknowledgements.* The authors are grateful to the AGAGE team for their dedication in providing long-term records of high-precision, high-frequency observations. KG and MR were funded by the Natural Environment Research Council (NERC) InHALE Highlight Topic (Investigating HALocarbon impacts on the global Environment, NE/X00452X/1). MR was also funded by the UK Research and Innovation-funded projects Self-learning Digital Twins for Sustainable Land Management (EP/Y00597X/1) and Greenhouse gas Emissions Measurement and Modelling Advancement (GEMMA, NE/Y001761/1). EF was supported by a Google Research PhD Scholarship. We thank the NASA Upper Atmosphere Research Program for its continuing multi-decadal support of AGAGE, including full support of THD and SMO, and partial support of MHD, RPB and CGO stations, through grants 80NSSC21K1369 to MIT and 80NSSC21K1210 and 80NSSC21K1201 to SIO and earlier grants. The Department for Energy Security and Net Zero (DESNZ) in the United Kingdom supported the University of Bristol for operations at Mace Head, Ireland (contracts 1028/06/2015, 1537/06/2018, 5488/11/2021, and PRJ_1604) and through the NASA award to MIT with the sub-award to University of Bristol for Mace Head and Barbados (80NSSC21K1369). The National Oceanic and Atmospheric Administration (NOAA) in the United States supported the University of Bristol for operations at Ragged Point, Barbados (contracts 1305M319CNRMJ0028, and 1305M324P0411). Observations at Gosan, South Korea, and S.P. are supported by the Korea Meteorological Administration Research and Development Program (Grant No. RS-2025-02313790). Measurements at Jungfraujoch are supported by the Swiss National Programs HALCLIM and CLIMGAS (Swiss Federal Office for the Environment, FOEN), by the International Foundation High Altitude Research Stations Jungfraujoch and Gornergrat (HFSJG), and by the European infrastructure projects ICOS and ACTRIS. Measurements at Zeppelin are supported by the Norwegian Environment Agency.

## References

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T.: Permutation importance: A corrected feature importance measure, Bioinformatics, 26, 1340–1347, https://doi.org/10.1093/bioinformatics/btq134, 2010.

Arnold, T., Manning, A. J., Kim, J., Li, S., Webster, H., Thomson, D., Mühle, J., Weiss, R. F., Park, S., and O'Doherty, S.: Inverse modelling of CF4 and NF3 emissions in East Asia, Atmospheric Chemistry and Physics, 18, 13 305–13 320, https://doi.org/10.5194/acp-18-13305-2018, 2018.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brokamp, C., Jandarov, R., Hossain, M., and Ryan, P.: Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model, Environmental Science and Technology, 52, 4173–4179, https://doi.org/10.1021/acs.est.7b05381, 2018.

Burkholder, J., Hondnebrog, O., McDonald, B., Orkin, V. L., Papadimitriou, V., and Hoomissen, D. V.: Summary of Abundances, Lifetimes, ODPs, REs, GWPs, GTPs, Scientific Assessment of Ozone Depletion 2022 [online], https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936562, (Accessed 20 August 2025), 2023.

Chambers, S. D., Zahorowski, W., Williams, A. G., Crawford, J., and Griffiths, A. D.: Identifying tropospheric baseline air masses at Mauna Loa Observatory between 2004 and 2010 using Radon-222 and back trajectories, Journal of Geophysical Research Atmospheres, 118, 992–1004, https://doi.org/10.1029/2012JD018212, 2013.

Chambers, S. D., Williams, A. G., Conen, F., Griffiths, A. D., Reimann, S., Steinbacher, M., Krummel, P. B., Steele, L. P., van der Schoot, M. V., Galbally, I. E., Molloy, S. B., and Barnes, J. E.: Towards a universal "Baseline" characterisation of air masses for high- and low-altitude observing stations using radon-222, Aerosol and Air Quality Research, 16, 885–899, https://doi.org/10.4209/aaqr.2015.06.0391, 2016.

Cleveland, R. B., S., C. W., E., M. J., and I., T.: STL: A Seasonal-Trend Decomposition Procedure Based on Loess, Journal of Official Statistics, 6, 3–73, 1990.

Cullen, M. J. P.: The unified forecast/climate model, The Meteorological Magazine, 122, 81–94, 1993.

Cunnold, D. M., Prinn, R. G., Rasmussen, R. A., Simmonds, P. G., Alyea, F. N., Cardelino, C. A., and Crawford, A. J.: The Atmospheric Lifetime Experiment 4. Results for CF2Cl2 based on three years data., Journal of Geophysical Research, 88, 8401–8414, https://doi.org/10.1029/JC088iC13p08401, 1983.

Cunnold, D. M., Steele, L. P., Fraser, P. J., Simmonds, P. G., Prinn, R. G., Weiss, R. F., Porter, L. W., O'Doherty, S., Langenfelds, R. L., Krummel, P. B., Wang, H. J., Emmons, L., Tie, X. X., and Dlugokencky, E. J.: In situ measurements of atmospheric methane at GAGE/AGAGE sites during 1985–2000 and resulting source inferences, Journal of Geophysical Research: Atmospheres, 107, https://doi.org/10.1029/2001jd001226, 2002.

Derwent, R. G., Simmonds, P. G., O'Doherty, S., Ciais, P., and Ryall, D. B.: European source strengths and Northern Hemisphere baseline concentrations of radiatively active trace gases at Mace Head, Ireland, Atmospheric Environment, 32, 3703–3715, https://doi.org/10.1016/S1352-2310(98)00093-4, 1998a.

Derwent, R. G., Simmonds, P. G., O'Doherty, S., and Ryall, D. B.: The impact of the Montreal Protocol on halocarbon concentrations in northern hemisphere baseline and European air masses at Mace Head, Ireland over a ten year period from 1987–1996, Atmospheric Environment, 32, 3689–3702, https://doi.org/10.1016/S1352-2310(98)00092-2, 1998b.

Derwent, R. G., Simmonds, P. G., Seuring, S., and Dimmer, C.: Observation and interpretation of the seasonal cycles in the surface concentrations of ozone and carbon monoxide at mace head, Ireland from 1990 to 1994, Atmospheric Environment, 32, 145–157, https://doi.org/10.1016/S1352-2310(97)00338-5, 1998c.

Eckle, K. and Schmidt-Hieber, J.: A comparison of deep networks with ReLU activation function and linear spline-type methods, Neural Networks, 110, 232–242, https://doi.org/10.1016/j.neunet.2018.11.005, 2019.

Fillola, E., Santos-Rodriguez, R., Manning, A., O'Doherty, S., and Rigby, M.: A machine learning emulator for Lagrangian particle dispersion model footprints: a case study using NAME, Geoscientific Model Development, 16, 1997–2009, https://doi.org/10.5194/gmd-16-1997-2023, 2023.

Fillola, E., Santos-Rodriguez, R., Tunnicliffe, R., Clark, J. N., Keshtmand, N., Ganesan, A., and Rigby, M.: Enabling fast greenhouse gas emissions inference from satellites with GATES: a Graph-Neural-Network Atmospheric Transport Emulation System, Geoscientific Model Development, 19, 1893–1915, https://doi.org/10.5194/gmd-19-1893-2026, 2026.

Gulev, S., Thorne, P., Ahn, J., Dentener, F., Domingues, C., Gerland, S., Gong, D., Kaufman, D., Nnamchi, H., Quaas, J., Rivera, J., Sathyendranath, S., Smith, S., Trewin, B., von Schuckmann, K., , and Vose, R.: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, chap. 2 - Changing State of the Climate System, pp. 287–422, Cambridge University Press, 2021.

He, T.-L., Dadheech, N., Thompson, T. M., and Turner, A. J.: FootNet v1.0: development of a machine learning emulator of atmospheric transport, Geoscientific Model Development, 18, 1661–1671, https://doi.org/10.5194/gmd-18-1661-2025, 2025.

Henne, S., Klausen, J., Junkermann, W., Kariuki, J. M., Aseyo, J. O., and Buchmann, B.: Representativeness and climatology of carbon monoxide and ozone at the global GAW station Mt. Kenya in equatorial Africa, Atmospheric Chemistry and Physics, 8, 3119–3139, https://doi.org/10.5194/ACP-8-3119-2008, 2008.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/QJ.3803, 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, pp. 580–589, https://doi.org/10.1007/978-0-387-68854-1_62, 2007.

Keeling, C. D., Piper, S. C., Bacastow, R. B., Wahlen, M., Whorf, T. P., Heimann, M., and Meijer, H. A.: Atmospheric $CO_2$ and $13CO_2$ Exchange with the Terrestrial Biosphere and Oceans from 1978 to 2000: Observations and Carbon Cycle Implications, pp. 83–113, Springer New York, New York, NY, ISBN 978-0-387-27048-7, https://doi.org/10.1007/0-387-27048-5_5, 2005.

Laube, J. C., Tegtmeier, S., Fernandez, R. P., Harrison, J., Hu, L., Krummel, P., Mahieu, E., Park, S., Western, L., Atlas, E., Bernath, P., Cuevas, C. A., Dutton, G., Froidevaux, L., Hossaini, R., Keber, T., Koenig, T. K., Montzka, S. A., Mühle, J., O'Doherty, S., Oram, D. E., Pfeilsticker, K., Prignon, M., Quack, B., Rigby, M., Rotermund, M., Saito, T., Simpson, I. J., Smale, D., Vollmer, M. K., and Young, D.: Update on Ozone-Depleting Substances (ODSs) and Other Gases of Interest to the Montreal Protocol, in: Scientific Assessment of Ozone Depletion: 2022, edited by Engel, A. and Yao, B., pp. 53–113, World Meteorological Organization, Geneva, GAW Report No. 278, 2022.

Liang, Q., Rigby, M., Fang, X., Godwin, D., Mühle, J., Saito, T., Stanley, K. M., Velders, G. J. M., Bernath, P., Derek, N., Reimann, S., Simpson, I. J., and Western, L.: Hydrofluorocarbons (HFCs), in: Scientific Assessment of Ozone Depletion: 2022, edited by Montzka, S. A. and Vollmer, M. K., pp. 117–151, World Meteorological Organization, Geneva, GAW Report No. 278, 2022.

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, Geoscientific Model Development, 9, 3213–3229, https://doi.org/10.5194/gmd-9-3213-2016, 2016.

Lunt, M. F., Manning, A. J., Allen, G., Arnold, T., Bauguitte, 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), Atmospheric Chemistry and Physics, 21, 16 257–16 276, https://doi.org/10.5194/acp-21-16257-2021, 2021.

Lööv, J. M. B., Henne, S., Legreid, G., Staehelin, J., Reimann, S., Prévôt, A. S., Steinbacher, M., and Vollmer, M. K.: Estimation of background concentrations of trace gases at the Swiss Alpine site Jungfraujoch (3580 m asl), Journal of Geophysical Research Atmospheres, 113, D22 305, https://doi.org/10.1029/2007JD009751, 2008.

Manning, A. J., O'Doherty, S., Jones, A. R., Simmonds, P. G., and Derwent, R. G.: Estimating UK methane and nitrous oxide emissions from 1990 to 2007 using an inversion modeling approach, Journal of Geophysical Research Atmospheres, 116, D02 305, https://doi.org/10.1029/2010JD014763, 2011.

Manning, A. J., Redington, A. L., Say, D., O'Doherty, S., Young, D., Simmonds, P. G., Vollmer, M. K., Mühle, J., Arduini, J., Spain, G., Wisher, A., Maione, M., Schuck, T. J., Stanley, K., Reimann, S., Engel, A., Krummel, P. B., Fraser, P. J., Harth, C. M., Salameh, P. K., Weiss, R. F., Gluckman, R., Brown, P. N., Watterson, J. D., and Arnold, T.: Evidence of a recent decline in UK emissions of hydrofluorocarbons determined by the InTEM inverse model and atmospheric measurements, Atmospheric Chemistry and Physics, 21, 12 739–12 755, https://doi.org/10.5194/acp-21-12739-2021, 2021.

Masih, A.: Application of Random Forest Algorithm to Predict the Atmospheric Concentration of $NO_2$, in: Proceedings - 2019 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology, pp. 252–255, Institute of Electrical and Electronics Engineers Inc., https://doi.org/10.1109/USBEREIT.2019.8736679, 2019.

Novelli, P. C., Masarie, K. A., and Lang, P. M.: Distributions and recent changes of carbon monoxide in the lower troposphere, Journal of Geophysical Research Atmospheres, 103, 19 015–19 033, https://doi.org/10.1029/98JD01366, 1998.

Novelli, P. C., Masarie, K. A., Lang, P. M., Hall, B. D., Myers, R. C., and Elkins, J. W.: Reanalysis of tropospheric CO trends: Effects of the 1997-1998 wildfires, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002jd003031, 2003.

O'Doherty, S., Simmonds, P. G., Cunnold, D. M., Wang, H. J., Sturrock, G. A., Fraser, P. J., Ryall, D., Derwent, R. G., Weiss, R. F., Salameh, P., Miller, B. R., and Prinn, R. G.: In situ chloroform measurements at Advanced Global Atmospheric Gases Experiment atmospheric research stations from 1994 to 1998, Journal of Geophysical Research Atmospheres, 106, 20 429–20 444, https://doi.org/10.1029/2000JD900792, 2001.

Prinn, R., Weiss, R., Arduini, J., Choi, H., Engel, A., Fraser, P., Ganesan, A., Harth, C., Hermansen, O., Kim, J., Krummel, P., Lo, Z., Lunder, C., Maione, M., Manning, A., Mitrevski, B., Mühle, J., O'Doherty, S., Park, S., Pitt, J., Reimann, S., Rigby, M., Saito, T., Salameh, P., Schmidt, R., Simmonds, P., Stanley, K., Stavert, A., Steel, P., Vollmer, M., Wagenhäuser, T., Wang, H., Wenger, A., Western, L., Yao, B., Young, D., Zhou, L., and Zhu, L.: The dataset of in-situ measurements of chemically and radiatively important atmospheric gases from the Advanced Global Atmospheric Gas Experiment (AGAGE) and affiliated stations (Version 20250123), [Data set]. NASA Langley Research Center (LaRC) Data Host Facility (DHF), https://doi.org/10.60718/0FXA-QF43, 2025.

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 System Science Data, 10, 985–1018, https://doi.org/10.5194/essd-10-985-2018, 2018.

Rigby, M., Mühle, J., Miller, B. R., Prinn, R. G., Krummel, P. B., Steele, L. P., Fraser, P. J., Salameh, P. K., Harth, C. M., Weiss, R. F., Greally, B. R., O'Doherty, S., Simmonds, P. G., Vollmer, M. K., Reimann, S., Kim, J., Kim, K. R., Wang, H. J., Olivier, J. G., Dlugokencky, E. J., Dutton, G. S., Hall, B. D., and Elkins, J. W.: History of atmospheric $SF_6$ from 1973 to 2008, Atmospheric Chemistry and Physics, 10, 10 305–10 320, https://doi.org/10.5194/acp-10-10305-2010, 2010.

Ruckstuhl, A. F., Henne, S., Reimann, S., Steinbacher, M., Vollmer, M. K., O'Doherty, S., Buchmann, B., and Hueglin, C.: Robust extraction of baseline signal of atmospheric trace species using local regression, Atmospheric Measurement Techniques, 5, 2613–2624, https://doi.org/10.5194/amt-5-2613-2012, 2012.

Ryall, D., Maryon, R. H., Derwent, R. G., and Simmonds, P. G.: Modelling long-range transport of CFCs to Mace Head, Ireland, Quarterly Journal of the Royal Meteorological Society, 124, 417–446, https://doi.org/10.1002/qj.49712454604, 1998.

Salvador, P., Artíñano, B., Pio, C., Afonso, J., Legrand, M., Puxbaum, H., and Hammer, S.: Evaluation of aerosol sources at European high altitude background sites with trajectory statistical methods, Atmospheric Environment, 44, 2316–2329, https://doi.org/10.1016/j.atmosenv.2010.03.042, 2010.

Seabold, S. and Perktold, J.: statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference, 2010.

Simmonds, P. G., Simmonds, P. G., Alyea, F. N., Cardelino, C. A., Crawford, A. J., Cunnold, D. M., Lane, B. C., Lovelock, J. E., Prinn, R. G., and Rasmussen, R. A.: The Atmospheric Lifetime Experiment 6. Results for carbon tetrachloride based on 3 years data., Journal of Geophysical Research, 88, 8427–8441, https://doi.org/10.1029/JC088iC13p08427, 1983.

Simmonds, P. G., Manning, A. J., Cunnold, D. M., McCulloch, A., O'Doherty, S., Derwent, R. G., Krummel, P. B., Fraser, P. J., Dunse, B., Porter, L. W., Wang, R. H., Greally, B. R., Miller, B. R., Salameh, P., Weiss, R. F., and Prinn, R. G.: Global trends, seasonal cycles, and European emissions of dichloromethane, trichloroethene, and tetrachloroethene from the AGAGE observations at Mace Head, Ireland and Cape Grim, Tasmania, Journal of Geophysical Research Atmospheres, 111, https://doi.org/10.1029/2006JD007082, 2006.

Simmonds, P. G., Rigby, M., McCulloch, A., O'Doherty, S., Young, D., Mühle, J., Krummel, P. B., Steele, P., Fraser, P. J., Manning, A. J., Weiss, R. F., Salameh, P. K., Harth, C. M., Wang, R. H., and Prinn, R. G.: Changing trends and emissions of hydrochlorofluorocarbons (HCFCs) and their hydrofluorocarbon (HFCs) replacements, Atmospheric Chemistry and Physics, 17, 4641–4655, https://doi.org/10.5194/acp-17-4641-2017, 2017.

Thoning, K. W., Tans, P. P., and Komhyr, W. D.: Atmospheric carbon dioxide at Mauna Loa Observatory. 2. Analysis of the NOAA GMCC data, 1974-1985, Journal of Geophysical Research, 94, 8549–8565, https://doi.org/10.1029/JD094iD06p08549, 1989.

Yang, C. F. O., Lin, Y. C., Lin, N. H., Lee, C. T., Sheu, G. R., Kam, S. H., and Wang, J. L.: Inter-comparison of three instruments for measuring regional background carbon monoxide, Atmospheric Environment, 43, 6449–6453, https://doi.org/10.1016/j.atmosenv.2009.09.026, 2009.