We thank the reviewers for their time and their helpful comments. This document outlines the reviewers' comments, our responses, and the changes we made to the manuscript as a result of these comments. The reviewers' comments shown in italic. All line and page number references are to those in the preprint.

In addition to changes initiated by the reviewer comments below, we have made some upgrades to the code since the initial submission, which have also addressed some errors and areas for improvement. These modifications do not change the main conclusions of the paper.

Firstly, we noticed that the test set was incorrectly formulated, meaning the previously quoted metric values were misleading. In the original code, the features were processed prior to splitting into the train, validation and test sets; this meant that the class ratio (achieved through under-sampling of non-baseline points) was the same in all sets. The test set was therefore unrepresentative of the "raw", unfiltered observations. This has since been amended, and the model is now evaluated on unsampled data and so the new metrics provide a more realistic assessment of performance. As would be expected, this has resulted in a reduction in performance against metrics like precision and recall, compared to the first draft. Secondly, we added more functionality to the model tuning through some additional "hyperparameters". These have allowed us to test the models in many different ways, such as through different sample weights and normalisation. These updates are outlined in the new version of the manuscript. Finally, we have added a third model type to the discussion, a gradient boosting classifier, to show the versatility of the method.

Reviewer 1:

1. While the motivation for reducing the computational cost of LPDM-based baseline classification is clear, the manuscript lacks explicit justification of why a ML surrogate is required, as opposed to simpler statistical or reduced-order physical approximations. In particular, quantifying the computational savings, and discussing alternative non-ML approaches would strengthen the motivation for the proposed methodology.

Between lines 49 and 102 of the original manuscript, we contrasted the strengths and weaknesses of these approaches. However, we perhaps did not sufficiently emphasise that, because LPDM-based methods can account for all of the relevant atmospheric transport and emissions processes, they are generally considered the "gold standard" approach for inferring baselines, when direct co-emitted tracer measurements are not available. Their major downside is the computational cost and technical barriers to implementation.

To emphasise this point, we have modified and merged the paragraphs that began on Line 99 (page 4) and Line 103 (page 5):

"Unlike statistical filters, model-based baseline classification does not impose smoothing on the dataset and airmass categories can be justified based on 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, 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. Our approach preserves the benefits of a meteorological filter for a fraction of the computational cost."

On the quantification of the computational cost improvements, we have added the following to line 288:

"The proposed ML surrogate is therefore approximately five orders of magnitude faster than the full-physics approach."

---

2. A measurement is a combination of both background and enhancements from local sources, and this proposed approach performs a binary classification to identify measurements dominated by either background or emissions from local sources. This approach functions as a filtering rather than a decomposition method which can identify their contributions to the given measurement. I think this manuscript will benefit from a clarification on this distinction as well as a reasoning of why authors chose classification approach over decomposition. This will definitely help avoid over-interpretation of the resulting baseline time series.

The reviewer is correct in saying that the method described in the paper purely filters the data, and so the data points assigned as "baseline" are those considered to be representative of the background. The reasons for pursuing this approach are:

a) the primary use cases for baseline mole fractions (global trend analysis or regional inverse modelling) only require a baseline/non-baseline categorisation
b) decomposing an observed mole fraction into multiple components (or, increasing the number of airmass classifications) is a harder problem, which we felt was beyond the scope of this first attempt at applying ML techniques to this area. It is certainly something that could be investigated in future.

To add clarification of the filtering nature of the method, we have added the following to line 118:

"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."

---

3. Line 20: Why the baseline is defined as representative of concentrations far from sources at the same latitude as the measurement site? Shouldn't it depend on the meteorology? For example, if winds are coming from north/south then baseline will not be the representative from the same latitude.

The answer to the question will depend on the application. As we say on lines 43-49:

"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)."

However, we made an error in the original manuscript, which violated our own definition. The InTEM flags that we used for Gosan retained values observed during the summer, where the station regularly intercepts southern hemispheric air (as per the above-mentioned definition from Arnold et al. (2018)). We have decided to keep these flags as they are, to demonstrate the flexibility of the approach, and modified our description of the baseline accordingly.

In place of our original definition on line 20 ("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"), we now write: "Here, we define baseline measurements as concentrations that would be observed at a point in the atmosphere, if it were not influenced by nearby sources or sinks (e.g. Manning et al., 2011).".

We also note on line 131: "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."

---

4. Line 125: Why binary classification and not multiclass classification to classify various categories that authors mentioned for baseline and non-baseline cases?

The binary classification was chosen to simplify the model requirements and to focus on the main use case for this algorithm, for the reasons outlined in response to the reviewer's second comment.

We have elaborated on the non-baseline category simplification by adding the following to line 129:

"For this work, the categories 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."

---

5. Line 154: The training dataset seems to preserve the natural imbalance between baseline and non-baseline classes with 4:1 representation ratio. Many ML practitioners use balanced training set (1:1) even in cases where one class has more representation over another (e.g. fraud detection). The manuscript does not provide a justification for this choice or discuss its implications on model performance. It would be helpful for the authors to provide a justification after comparing these two approaches.

During model development we tested different class ratios and found that the combination of under-sampling the non-baseline class and implementing a high confidence threshold to reduce the occurrence of false positives often produced the best model performance. However, we have since run more tests and

found that this is not always the case. Further tests of 1:1 ratios, conserving the natural ratio (of under-sampled baseline points) whilst implementing sample weights, and testing of this confidence threshold showed that the approach is not generalisable and is instead unique to model type and site.

In place of "This was mitigated by under-sampling 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:non-baseline ratio of 4:1 in our training set." on line 152, we have added: "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 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."

---

6. Line 159: Depending on the spatial domain, the air parcels may have a variable time when they entered the domain. It seems only using meteorology up to 6 hours before the measurement may not be enough, especially for large domains when the air parcels may have entered much earlier than 6 hours before measurements (e.g. 2-3 days earlier). This may also be an indication of overfitting if the performance is higher with just meteorology from 6 hours before measurement. The authors should add more meteorology data prior to measurement to strengthen the confidence in the physical representativeness of the approach.

We agree that adding additional meteorology would improve the robustness of the modes. Therefore, we have amended the code to easily add additional meteorology prior to the measurement time, as well as testing different levels of temporal detail. The initial choice to add just six hours allowed us to provide the model with more information whilst keeping the number of input variables low, however, tests have shown that adding additional temporal attributes does not significantly increase training time.

We have added further discussion of the model-specific time interval tuning to line 159:

"To provide the algorithm with further information on changes in atmospheric conditions, all meteorological variables were also provided 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 time tested, to allow the model to capture the appropriate 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."

---

7. Line 179: Why did the authors decide to train the model on only 1 year and test it on 20+ years?

The decision to use a small volume of data for model training was made by considering both the time taken to fit the model and the resulting model performance. We must note, however, that with the newest code iteration, three years of data are now used for training. Three years produce reasonable model performance and a training time of no more than a few minutes (with most cases being less than a minute); tests were done on volume of training data and any increase in performance was minor and did not outweigh the increased computational costs. This test has been added to the Supplement. At most stations, this left 20+ years of data for validation and testing. In each case, one year was allocated to the validation set for model finetuning, and the rest for testing. Using the long testing period allowed the models to be evaluated over a wide range of conditions and with any interannual variability, giving a more robust assessment of their performances. The longevity also demonstrates the capability of the models to be applied far into the future, reducing the need for regular retraining.

We have added an explanation of why the different set periods were chosen to line 184:

"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."

---

8. Line 185: The authors trained separate models for each site instead of a single model. While this is a common approach in the field to train separate models for different sites, it would be interesting to discuss if a generalized model can be developed which can identify baseline across sites.

Creating a universal model applicable to all nine AGAGE sites was briefly explored during the early stages of this work. However, given the unique geographies of each site, we decided to create one model per site giving specialised results. The models receive no inputs relating to site location, surrounding topography, land type or surface roughness, or population densities. The impact of these features is learnt implicitly by each site-specific model. These location-specific features are extremely high dimensional and so would likely require a much more complex model and substantially larger training dataset (likely requiring orders of magnitude more model simulations to be performed, compared to producing a ~3-year training dataset for new measurements sites).

We have added a brief discussion of why a generalised model is not applicable in this case, and noted our non-specialised approach on line 179:

"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, means that the geographical characteristics of each location are learned implicitly."

---

9. Line 221 and Table S5: I think it will also be interesting to analyze the feature importance from temporal perspective. For example, which timestep (measurement time or before) shows stronger feature importance.

When exploring feature importance with permutation importance analysis, the model input variables were put into blocks of similar features. This was not discussed in the original manuscript and has now been added. The permutation importance method does not work well with closely correlated features (such as time-lagged meteorological variables), so we introduce feature blocks to shuffle correlated inputs together, combining meteorological variables by category, such as all u-components of wind at all times and locations. This shuffling does not preserve the temporal structure of the data, therefore, this choice of feature importance analysis does not allow for temporal disaggregation. Despite this, permutation importance analysis was chosen here as it is applicable across all model types, is simple to implement and computationally efficient for the large training datasets.

We have updated our feature importance discussion on line 217:

"We determine the variables that are most important for model performance using a permutation importance analysis (Breiman, 2001; Altmann et al., 2010). Permutation importance is calculated by randomly shuffling the values of a single feature group and observing the resulting change in the model's performance (Fillola et al., 2023). The process is repeated multiple times to ensure robustness, and the importance of a feature is determined by the extent to which model performance degrades when the 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 feature groups for each site are shown in the Supplementary Material."

---

10. Line 250: The observed feature importance patterns are consistent with recent work on ML-based emulation of LPDM footprints (e.g. FootNet). Citing and briefly discussing these related methods would help place the present results in the context of ML emulations studies.

The reviewer is correct, adding discussion of LPDM ML emulation would add important context to the study.

We have added the following to the text from line 107 to introduce and compare to LPDM footprint emulation studies:

"ML has been employed successfully for various applications in atmospheric chemistry, including the prediction of particulate matter concentrations and nitrogen dioxide modelling (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."

We have also linked the feature importance pattern with ML footprint emulation studies on line 253:

"This feature importance pattern aligns with that seen in ML-based LPDM footprint emulation methods previously introduced (Fillola et al., 2023; He et al., 2025)."

---

11. Figure 1 & 4: It would help to explicitly clarify in both the figure captions that this ML model only classifies a measurement as a baseline or non-baseline rather than predicting the baseline mole fractions. As currently presented, Figures 1 and 4 could be interpreted as showing ML-predicted baseline mole fractions, whereas the concentrations shown seem to be the original observations filtered using the ML classification. Clarifying this would help avoid potential over-interpretation of the results

The baseline classification rather than mole fraction prediction has been clarified in both figure captions.

Figure 1 caption:

"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 presented here to be baselines (blue)."

Figure 4 caption:

"The blue line shows the monthly means of data points that the MLP model classified as representative of baseline conditions."

Reviewer 2:

The "mole fraction in air" time series predicted baselines often include what appear to be massive deviations from the other baseline points. This isn't addressed in the paper, but seems to be a rather glaring showstopper for many species. E.g., CF4 from Gosan, HFC-134a in Monte Cimone, especially as this contrasts the comment on line 279 that suggests that CF4 should have fewer false positives or negatives. I understand that the prevalence of false positives is quantified, and used as a justification that the monthly mean may be the useful component, but it seems that the expectation for the ML method output is being set rather low, and further, there isn't sufficient discussion on what is useful and what is not. Are there species for which this is not a useful methodology, and/or locations where there are too many local sources for this to be a viable methodology?

These types of mis-categorisation can be separated into two types. The first contains data points that the ML-emulator classifies wrongly with respect to the InTEM label. The second mis-categorisation category includes the non-baseline points that the InTEM method we are trying to emulate also classifies incorrectly; there are some clear examples of these in the Gosan plots. Clearly, our method does not address the latter category (the best we can do is emulate the InTEM flags, even if they are "wrong"). As we note in the paper, InTEM includes a second statistical filtering step that removes many of these "anomalous" points, which could also be employed with our algorithm, if required.

The mis-categorisation from the ML algorithm (the first category discussed above) is reflected in our MAE/RMSE/MAPE labels as we treat the InTEM labels as the "true" baselines. It is possible that sometimes the model does not have enough information to distinguish between similar baseline and non-baseline conditions, and so further, more complex inputs may be required for these points to become "separable". We implemented a confidence threshold so that only points that the model was more than 60-90% confident in would be classified as baseline (threshold optimised as part of training and depends on specific model), hence increasing the prevalence of false negatives with the aim to reduce false positives and obtain more accurate results. We found that this positively impacted model performance.

The reviewer makes a good point that a metric that describes when the algorithm is "useful" would be valuable. Here, we consider that the main likely use case for such an algorithm; the estimation of baseline monthly means. We added a coefficient of variation (CV) metric to quantify the variability in the "true" baselines. This considers the variability in the baselines that the ML model is emulating whilst removing any seasonality and long-term trend. We found that there is a clear correlation between this CV and the MAPE across all sites and species; a plot showing this relationship at each site has been added to the supplement. This could, therefore, be used as an indicator of model performance with species below a certain threshold being considered "not useful", if needed.

Whilst there are clearly areas where the model performs better or more poorly, we strongly disagree with the reviewer's assessment that there are "showstoppers" that prevent the algorithm from being useful. The algorithm provides monthly means that have sufficient consistency with the InTEM estimates for all species and sites to be useful for, e.g., global trend evaluation. Furthermore, we would emphasise that: a) an additional statistical filtering step could readily be performed (as is done with the full InTEM algorithm),

to remove obvious non-baseline events that are not picked up by either algorithm; b) this is the first generation of such an ML-based algorithm, and improvements can, and almost certainly will, be made by different architectures, feature selection, metrics, and training data.

We have added an explanation of the additional true baseline metrics to line 215:

"To explore the utility of the method across 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:"

We have added an additional small section discussing the relationship between CV and MAPE to line 282:

"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, N_2O). Shorter-lived compounds such as dichloromethane, which tend to exhibit substantial seasonal variability and zonal gradients, tend to show the 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."

We have also removed the use of CF_4 as an example in line 279, as we agree it was misleading.
