https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

A deep-learning-based classification of European air quality monitoring sites for model verification and data assimilation

Mathieu Joly[1], Valentin Petiot[4], Jean-Jacques Hok[2], Oihana Fermond[2], Cécile Cazalet[2], François Besson[2], Augustin Colette[3], Frédérik Meleux[3], and Vincent Guidard[1]

[1]Météo-France, CNRS, Univ. Toulouse, CNRM, Toulouse, France
[2]Météo-France, Direction des Services Météorologiques, Toulouse, France
[3]INERIS, Institut National de l'Environnement Industriel et des Risques, Verneuil-en-Halatte, France
[4]Independent Engineer, 1 route de Lucy, 89480 Crain

Correspondence: Mathieu Joly (mathieu.joly@meteo.fr)

Abstract.

Assessing the spatial representativeness of observational data is a key challenge in atmospheric composition modelling, for model validation and data assimilation. In addition to the valuable metadata provided by network operators, the European CAMS (Copernicus Atmosphere Monitoring Service) regional consortium (Colette et al., 2025) also uses the results of an objective classification method originally based on the work of Joly and Peuch (2012). The method has been significantly improved in 2017, and Section 2 provides the first full published description of the LDA method that has produced site lists for the CAMS Regional Air Quality service each year since 2017. However, the method still had limitations, in particular a high sensitivity to the quality and quantity of data, which led to the rejection of certain time series. This paper presents a new method that aims to be more effective using recent machine learning and deep learning techniques. The new procedure now comprises three stages: i) data pre-processing including gap filling, ii) unsupervised outlier detection based on the statistical features of each annual time series, and iii) a 10-fold supervised classification using deep learning. For each species, the choice of models and hyperparameters has been fine-tuned separately. When applied to the eight years of data from 2018 to 2025, the new method yields improvements for all species and all performance metrics compared to the previous operational version. Case studies show that time series data provide a useful source of information for scrutinising and supplementing the network operator's metadata. This new method is now being used operationally to provide information to the CAMS-regional project and has contributed to a substantial reduction in the amount of data discarded before data assimilation and model validation.

1 Introduction

Over the past decades, air quality monitoring networks have been established and expanded across various European countries. These data have been compiled at European level by the European Environment Agency (EEA), initially in the form of the AIRBASE dataset, and subsequently as Air Quality (AQ) e-reporting, which includes both data validated by each data provider and unvalidated real-time data. At the European level, the measurement network is not evenly distributed. The station density is strongly correlated with population density and emission sources.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Pollutants can be measured close to emission sources (industrial, traffic, biogenic, etc), at distance (background pollution), or in densely populated areas for impact monitoring. Locally, decisions about where to locate monitoring networks are often based on historical and economic factors and vary depending on the species. This is compounded by highly varied topographical and climatological conditions. As a consequence, the spatial representativeness of individual sites varies enormously: a traffic site located in a street canyon may represent an area of only a few hundred square metres, while a rural background site may be representative over hundreds of square kilometres (Piersanti et al., 2015; Martin et al., 2014; Martín et al., 2015). This variability has direct consequences for the main uses of the data: compliance assessment and population exposure estimation (Duyzer et al., 2015; Hohenberger et al., 2021), and, more pertinent to the present study, the validation of chemical transport models and data assimilation systems (Janssen et al., 2012; Solazzo and Galmarini, 2015).

In Europe, the metadata associated with monitoring sites — available through the European Environment Agency (EEA) API — provides a first characterisation of the station type (background, traffic, industrial) and the environment (rural, suburban, urban). Although this classification is based on formal criteria deriving from EU directives, it remains largely subjective and depends on the judgment of national operators, resulting in potential discrepancies at the European scale. Joly and Peuch (2012) showed that the metadata could be supplemented by an objective classification at European level on the premise that the true fingerprint of a monitoring site is encoded in its observations. The method defined statistical indicators describing each time series for each species separately and then applied a Linear Discriminant Analysis (LDA) to separate the different environments of sites. Similar approaches have since been proposed using hierarchical cluster analysis with non-negative matrix factorisation (Malley et al., 2014), k-means clustering of autocorrelation functions (Barrero et al., 2015), frequency distribution analysis with the Gini index (Tapia et al., 2016), and wavelet decomposition-based clustering (Gouveia et al., 2015). More recently, machine learning classifiers – including ensemble methods based on decision trees – have been applied to classify sites from air quality and meteorological data simultaneously (Tüysüzoğlu et al., 2020). These studies share a common finding: time series contain sufficient information to classify sites objectively, often revealing inconsistencies with the existing metadata.

The present study builds on this line of work and on the classification of sites used by the Copernicus Atmosphere Monitoring Service (CAMS) for European regional air quality modelling (Colette et al., 2025). That method, based on an updated and extended version (see Section 2) of the Joly and Peuch (2012) LDA approach, classifies sites annually for seven species (O_3, NO_2, NO, CO, SO_2, PM_{10}, and $PM_{2.5}$) and was used by CAMS-regional from 2017 to 2025 to select the most appropriate sites for model verification and data assimilation. Site selection is indeed a critical step in data assimilation: Hermanns et al. (2025) showed that improving the representativeness of the observational dataset used in the EURAD-IM assimilation system leads to measurable improvements in analysis quality, particularly for CO and NO_2.

The goal of this paper is to replace the previous LDA-based method with a deep learning approach that operates directly on raw annual time series, thereby removing the need for manual feature engineering and potentially capturing more complex patterns in the data. The new method is evaluated against the previous operational approach in terms of both classification quality and practical impact on the number of sites available for CAMS-regional.

Section 2 provides a summary of the reference method, which is an update of Joly and Peuch (2012); sections 3 to 6 then set out in detail the new method proposed in this study. Section 3 describes the preliminary data processing. Section 4 presents the

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

unsupervised outlier detection method, which is essential for the robustness of the results. Section 5 details the classification method and explains how the algorithms and parameters were selected during a preliminary study. Finally, Section 6 presents the validation results, case studies, and the impact on operational use of the results in CAMS-regional.

2 The reference LDA-based classification method

The Joly and Peuch (2012) objective classification was completely revised in 2017, and extended to more species. This constitutes the reference for the present paper, which is outlined here to explain the study's benchmark. For brevity, it will be referred to hereafter as the "LDA reference method". That classification was used operationally within the CAMS-regional service between 2017 and 2025 to provide stations lists used for data assimilation and model validation, accounting for the most recent EEA data.

2.1 Data selection

The LDA reference method uses hourly observations from the EEA AQ e-reporting system for seven species: O_3, NO_2, NO, CO, SO_2, PM_{10}, and $PM_{2.5}$. The species CO and $PM_{2.5}$, absent in Joly and Peuch (2012), are now included. When validated data is not yet available, unvalidated data is included to enable new sites to be integrated into the classification lists more quickly. Time series are clipped based on maximum value thresholds determined from the Probability Distribution Function (PDF) of the values for each species. "Suburban traffic" sites are no longer considered, whereas in Joly and Peuch (2012) they were treated in the same way as "urban traffic" sites to expand this group prior to the LDA. Furthermore, so as not to distort the statistical analysis, the few sites at altitudes above 1,400 metres were excluded, because meteorological and chemical regimes differ substantially from those of lower-altitude sites, where most of the population and emission sources are concentrated (Okamoto and Tanimoto, 2016).

2.2 Feature extraction

To supplement Joly and Peuch (2012), new indicators have been considered in the LDA reference method. Some are based on the PDF of hourly values, e.g. Gini index (see Tapia et al., 2016, study for ozone), skewness, and kurtosis. Following the work of (Barrero et al., 2015), the autocorrelation of the time series has also been added. New methods have been implemented for calculating the indicators relating to diurnal and annual cycles, and for the "weekend effect". In the end, 25 indicators – versus 8 in Joly and Peuch (2012) – were compared. Fourteen were selected for their ability to distinguish between the different groups of sites, using two statistical tests based on different hypotheses: the Welch T-test and the Mann-Whitney U-test. The fourteen selected features allow a qualitative description of the time series (if sufficient data is available).

2.3 Feature transformation

At the heart of the improvements in the LDA reference method are new methods for transforming the indicators before the LDA. Indicators with highly asymmetric distribution are processed with the automatic technique from Box and Cox (1964).

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

After standardisation (with median and inter-quartile range), outlier values are detected for each group separately, based on a multivariate method: Support Vector Machines (SVM) available in Python library `scikit learn` (Pedregosa et al., 2011). Principal Component Analysis (PCA) is then used to reduce the number of dimensions by identifying the directions that maximise the variance. The first seven principal components were selected as they accounted for 93% to 98% of the total variance, depending on the species.

2.4 Multi-group linear discriminant analysis

In the LDA reference method, the LDA is first performed on a more homogeneous subset of sites, excluding those presenting outlier values in their features, and considering a geographical subdomain encompassing latitudes between the Strait of Gibraltar and the Arctic Circle, and longitudes between the west coast of Ireland and the eastern border of Finland. All the data set aside is then projected a posteriori onto the LDA axes. In Joly and Peuch (2012), the LDA was performed using the two most dissimilar populations: rural background on the one hand, versus urban background plus urban or suburban traffic on the other. With the LDA reference method, it became possible to implement a multi-class linear discriminant analysis across the four groups of sites. The multi-class LDA favours the directions that best separate different groups, maximising inter-group distance and minimising intra-group variance.

2.5 Division into 10 classes

In Joly and Peuch (2012), the projection onto the Fisher axis was divided into classes based on the 9 deciles of the distribution of values. This ensured that the classes have the same number of observations. A significant drawback was difficulty in interpreting and comparing the classes obtained for the different species. Since the LDA reference method, the proportion of different families of sites is taken into account when projecting into the 10 classes. This allows the differences between species to be preserved; for example, there are more rural sites and less traffic-related data for ozone than for CO.

2.6 Motivation for the present work

The aim of this section was to show that, with the LDA reference method, time series analysis has become both more detailed and more effective than in Joly and Peuch (2012), but also more complex. With the 2017 update, the LDA-based method has reached its full potential. The main limitations are a high sensitivity to data quality and quantity, which leads to the exclusion of a number of sites. For certain species, the rejection rate under the LDA reference method exceeds 25% of available background sites, as quantified in the results (see Section 6.4), which directly reduces the spatial coverage of the data assimilation and model verification. Besides, for some species, such as SO_2, the time series are sometimes difficult to characterise using the 14 indicators, particularly with low-resolution sensors. The idea behind the present study is to replace this entire pipeline, using new machine learning and deep learning methods to try to simplify and further improve this classification method. Since the method was last updated in 2017, there has been unprecedented progress in machine learning techniques, which are now

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

available in standard Python libraries. There has also been an increase in GPU computational facilities. All of these factors make it highly appealing to explore these new approaches.

3 Data for the deep learning approach

3.1 Observational dataset

As in the LDA reference method, 8 years of data are considered - 2018-2025 in the present case - for the same seven species: O_3, NO_2, NO, SO_2, CO, PM_{10}, and $PM_{2.5}$. All the available data from the EEA API has been downloaded (E1a and E2a fluxes), using "real time" data when and where "validated" data were unavailable, to ensure the dataset is as comprehensive as possible. As shown in Figure 1, a broad geographical area is considered, between latitudes 27°N and 72°N, and between longitudes 32°W and 46°E. This includes sites in Iceland, the Azores archipelago, and the Canary Islands, but also all mountain sites, which were excluded from the LDA reference method (see Section 2.1).

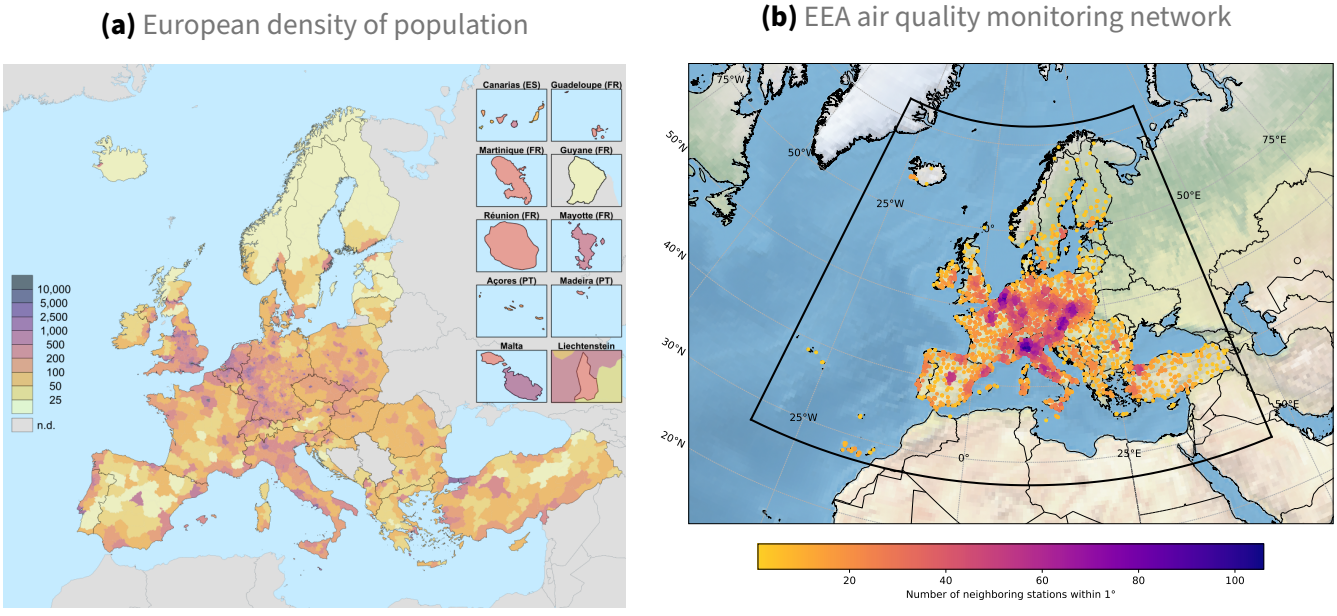


Figure 1. a) Density of population in the European Union in 2017 (inhabitants per km^2), from Wikipedia (2026), b) density of EEA network for all species and for the period considered (number of neighbouring sites within a 1° radius), with the geographical domain considered in black.

Based on the existing metadata (also available through the EEA API), the following groups are derived:

- **T** = traffic urban;
- **U** = background urban;

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

- **S** = background suburban;
- **R** = background rural;
- **X** = others, i.e. either outside the geographical domain, or inside but with metadata industrial, or traffic suburban and rural.

The **X** sites are not considered for classification: for these sites, one must rely on the metadata provided by the network installer. Industrial sites can exhibit a wide variety of behaviours, resulting in time series that are difficult to use. As for traffic sites in rural or suburban areas, it is difficult to categorise them because the influence of the emission source can vary greatly depending on the topographical and meteorological conditions. Consequently, these sites are not, in principle, suitable for the purposes of model validation and data assimilation. They are therefore discarded.

In Joly and Peuch (2012) and in the LDA reference method, a much more selective approach was taken with the data (cf. Section 2), given that LDA can be highly sensitive to outliers in the statistical indicators calculated from the time series. This is no longer the case with the new method, the aim of which is precisely to take into account as much data as possible, as is, and to rely on a high-performance unsupervised outlier detection method (described in Section 4) to automatically identify truly problematic time series.

3.2 Pre-processing

Gaps of up to 24 consecutive hours are considered small gaps and are interpolated. After some tests, the Python `scipy` PCHIP interpolator (Piecewise Cubic Hermite Interpolating Polynomial) was chosen because it preserves monotonicity in the interpolation data and does not overshoot if the data is not smooth. The interpolant uses monotonic cubic splines to find the value of new points (Fritsch and Butland, 1984; Moler, 2004).

Gaps of up to 720 consecutive hours (i.e. 30 days) are considered large gaps. To keep a diurnal cycle, a running mean is computed *for the given hour of the day*, with a window size of 91 values (days), if minimum 20 values are available. The window size of 91 corresponds to a quarter (as with a climatological average), and the threshold of 20 values originates from tests and choices made for the preceding Joly and Peuch (2012) and the LDA reference method.

The authors acknowledge that this gap filling procedure is not perfect: interpolation cannot invent missing data, but the results are much more realistic than zero padding — a common but crude alternative often used before deep learning — and in any case allow to integrate a larger number of "complete" time series into the dataset. In the end, each civil year, for each site, is considered separately, and truncated at 8760 values to process arrays of same shapes (i.e., no leap years). Annual time series with remaining missing values (because of gaps larger than 720 hours) are discarded.

4 Outlier detection

Before classification, it is essential to detect and exclude time series that exhibit truly abnormal behaviour. The aim is both to avoid compromising the training dataset with genuinely erroneous or unrepresentative data and, ultimately, to safeguard

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

the operational systems for data assimilation and forecast model validation. The approach adopted here is fully unsupervised, requiring no a priori knowledge of site type.

4.1 Feature extraction

Until now, in Joly and Peuch (2012) and in the LDA reference method, the implementation of statistical indicators to describe time series relied on knowledge of the behaviour of the time series in question (see Section 2.2). In the present study, the `tsfresh` Python package automatically calculates a large number of time series characteristics, the so-called "features". Using all the features available in the 0.21.0 version of `tsfresh`, one gets 783 values for each time series. Unlike the LDA reference method, which used PCA to reduce the number of dimensions, all features are now passed to the outlier detection without filtering, because the methods under consideration (see below) scale reasonably well to large feature spaces.

4.2 Choice of algorithm

Two unsupervised outlier detection algorithms, both available in the `scikit-learn` Python library (Pedregosa et al., 2011), were evaluated on the feature vectors described above.

The first, Isolation Forest (Liu et al., 2008), exploits a fundamental property of anomalous data points: they are rare and statistically distinct from the bulk of the data, which means they are easier to isolate than normal observations. The algorithm builds an ensemble of random binary trees by recursively partitioning the feature space along randomly chosen dimensions and thresholds. Because anomalous points occupy sparse regions of the feature space, they tend to be separated from the rest of the data after very few partitions, resulting in short average path lengths through the trees. Normal points, embedded in denser regions, require many more splits to be isolated and therefore have longer average path lengths. The anomaly score of each observation is derived from this average path length across the ensemble: a short path indicates a likely outlier. A key advantage of this method is that it operates globally — it identifies points that are anomalous with respect to the entire dataset.

The second algorithm, Local Outlier Factor (LOF; Breunig et al., 2000), takes a fundamentally different approach by comparing each observation to its local neighbourhood rather than to the global distribution. For each point, the local density is estimated from the distances to its k nearest neighbours. A point is considered an outlier if its local density is substantially lower than the densities of its neighbours — in other words, if it occupies a sparser region than the points surrounding it. Unlike global methods, LOF can detect outliers that are anomalous only within their local context and would not be flagged by a distance-based approach applied to the full dataset.

The contamination parameter, which controls the threshold between normal and anomalous observations, was set to its automatic mode in both algorithms, allowing the threshold to be determined adaptively from the score distribution rather than imposing a fixed outlier fraction. This choice is motivated by the fact that the proportion of erroneous time series varies between species and between years as data quality improves, making a fixed contamination rate inappropriate.

In practice, the two algorithms responded differently to the structure of the feature dataset. This is illustrated for NO_2 with Figure 2. For visualisation purposes, the 783-dimensional feature dataset was projected onto the first two principal components. This projection reveals that Isolation Forest tends to form coherent clusters of data points that exhibit similar behaviour in the

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

main features with high covariance. LOF, by contrast, seems more effective at isolating atypical patterns in features with lower variance, more in line with the objective of detecting erroneous or unrepresentative data.

Geographically, Figure 2 (c and d) shows that both methods locate outliers in a very consistent manner. However, the Isolation Forest method tends to form small, denser clusters of outliers (for example, over Italy for NO_2), which is undesirable. The goal is not to discard coherent groups with distinctive pollution patterns, but rather to detect genuinely atypical time series at a more localised level. For this reason, LOF was selected for the outlier detection step.

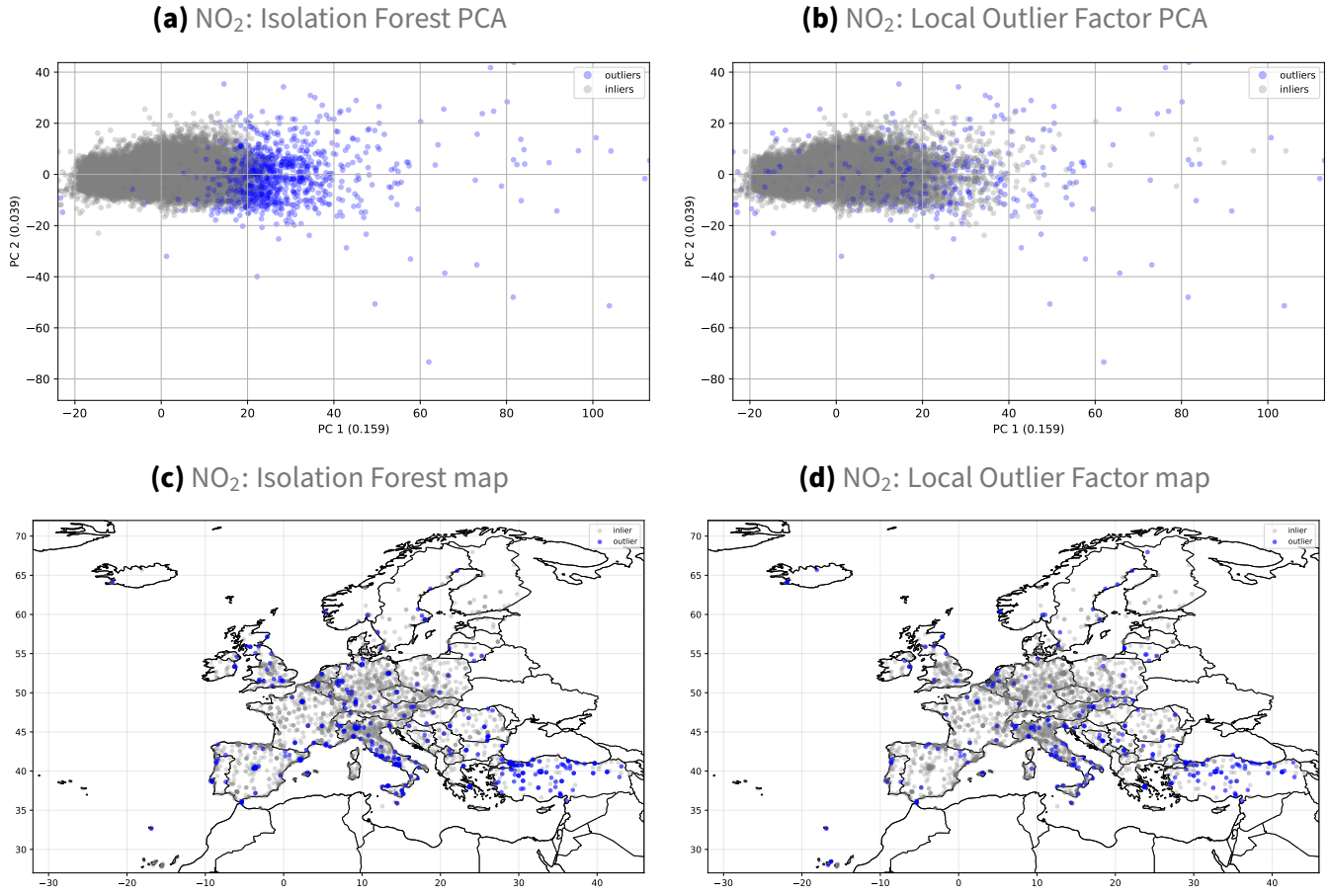


Figure 2. Comparison of outlier detection methods for NO_2: a) and b) projection on the first two axes of a Principal Component Analysis of the time series features for all years ; c) and d) localisation of sites with at least one year of outlier data. Data detected outliers are marked in blue, and the rest are marked in grey.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

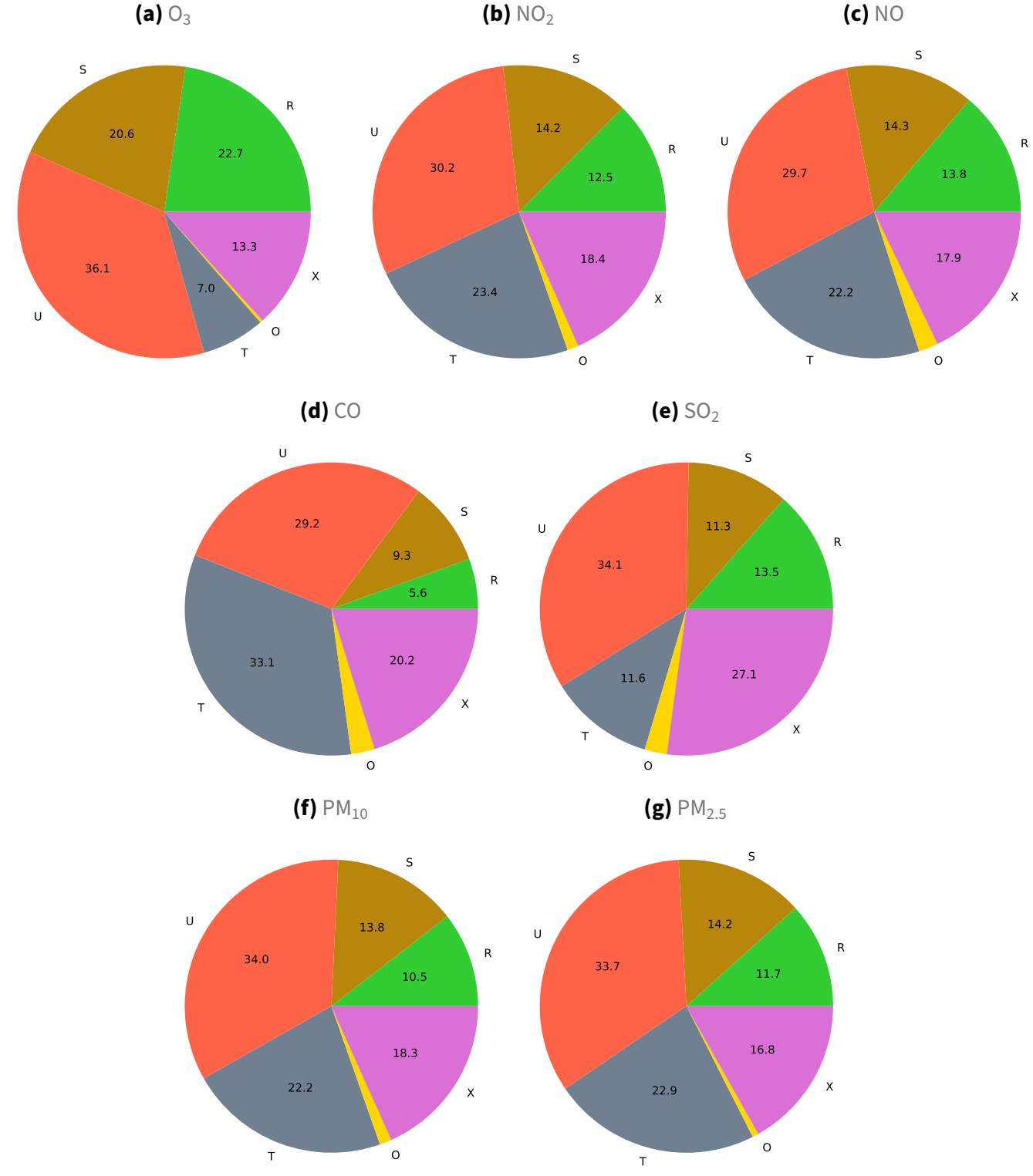


Figure 3. Piechart of the groups obtained with Local Oulier Factor method. Percentage of sites for each group (if superior to 5%).

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

4.3 Typology of the input data

Figure 3 illustrates the data typology after outlier detection. The groups are defined as in Section 3.1, with the addition of "O" representing outlier sites for which all available annual time series were flagged as outliers by the LOF algorithm. The main result is that the fraction of discarded sites remains below 5% for all species.

According to Figure 3, NO_2, NO, PM_{10}, and $PM_{2.5}$ display a similar distribution: approximately 30-34% in the "U" group, about 23% in the "T" group, and about 14% in the "R" and in the "S" groups. For that species, discarded data ("X" and "O") account for roughly 20%. On the contrary, O_3 has a very small "T" group, but larger "R" and "S" groups than the others. CO shows a very large "T" group, and very small "R" and "S" groups. Finally, SO_2 includes a large "X" group, reflecting the prevalence of industrial measurement sites.

To give a quantitative picture of the EEA dataset treated in this section, Table 1 gives the number of selected sites and time series. This shows that the ratio of outlier time series reaches 5% for CO and SO_2, but is close to zero for ozone.

In conclusion, the outlier detection process removes only a relatively small proportion of the available data. Nonetheless — as explained in the introduction — the amounts and typology of the input data varies considerably depending on the species, especially O_3, CO, and SO_2.

Species	R+S+U+T sites	X sites	O sites	available t.s.	outlier t.s.
O_3	2281	141	3	14968	75
NO_2	3188	205	21	20278	416
NO	2227	149	50	13457	510
CO	1007	134	15	5569	303
SO_2	1447	151	38	8407	442
PM_{10}	2594	327	20	14478	385
$PM_{2.5}$	1712	255	15	8538	134

Table 1. Dataset summary: number of sites taken into account (R+S+U+T), and excluded (either X or O). Number of available time series (sufficient data after gap filling), and number of outlier time series detected with LOF method.

5 Classification

The LDA reference method was based on the description of time series using statistical indicators suited to hourly measurements of gas and aerosol concentrations (see Section 2). The rise of deep learning for all kinds of datasets has extended to the field of time series, especially for classification, regression, and forecasting purposes. Rather than basing the analysis on the 783 features calculated for outlier detection, it was decided to apply a supervised classification method directly to the pre-processed annual time series. Indeed, unlike LDA, one of the key strengths of deep learning architectures is the ability to learn directly from the time series.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

The new classification method assigns each background site to one of four environment types (R, S, U, T) on the basis of its annual time series. The algorithm consists of a supervised deep learning classifier applied in a 10-fold cross-validation framework. Before presenting the results, Section 5.1 describes the methodology used to select the optimal deep learning algorithms and hyperparameters for each species. Section 5.2 then describes the full classification procedure applied to the complete dataset.

5.1 Model selection and hyperparameter tuning

5.1.1 Experimental design

For the preliminary study, the data has been divided into two parts:

- "High-potential (HP) data": the most trustworthy measurement sites *a priori*. To do this, the metadata is compared to the previous objective classifications obtained with the LDA reference method for the last three years. Sites which get a good coherence with metadata are selected:
  - metadata="R" and $1 \leq \text{class} \leq 4$ for all three years;
  - metadata="S" and $3 \leq \text{class} \leq 6$ for all three years;
  - metadata="U" and $5 \leq \text{class} \leq 8$ for all three years;
  - metadata="T" and $7 \leq \text{class} \leq 10$ for all three years.
- "Low-potential (LP) data": the remaining measurement sites, for which there is no consistency or no stability in the relationship between metadata and results of the LDA reference method, and also sites that could not be classified for the last three consecutive years.

The deep learning classification has been tuned by performing training on HP data and validation on LP data. The rationale behind this separation of data is that training should be carried out on data for which i) the statistical properties of the time series are informative enough to discriminate between site types, and ii) the metadata label is consistent with the time series, which is preferable for a supervised classification method. The method is then tested on the LP dataset. Performance on LP data is not expected to match that on HP data, but the new method should at minimum match or improve upon the LDA operational method for these ambiguous cases.

Table 2 supplements Table 1 by adding to the tally the number of sites categorised HP or LP, from amongst those selected for the analysis (R, S, U, or T) after the outlier detection. LP data accounts for between 31% (for NO_2) to 47% (for CO) of the total available. This shows that we are applying a significant constraint to the data (particularly for CO and PM_{10}), but this enables us to obtain sufficient data both for training and validation of the deep learning method.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Species	HP sites	LP sites
O_3	1394	887
NO_2	2191	997
NO	1364	863
CO	534	473
SO_2	932	515
PM_{10}	1415	1179
$PM_{2.5}$	1091	631

Table 2. Number of R+S+U+T sites in the HP and LP categories.

5.1.2 Statistical scores

The following metrics have been used to monitor the performance of the classification. In the formulas, tp is the number of true positives, fp the number of false positives, tn the number of true negatives, and fn the number of false negatives.

- Accuracy: ratio (tp + tn) / (tn + tp + fp + fn). Fraction of correct predictions;
- Precision: ratio tp / (tp + fp). The precision is intuitively the ability of the classifier not to label as positive a sample that is negative;
- Recall: ratio tp / (tp + fn). The recall is intuitively the ability of the classifier to find all the positive samples;
- F1 score: harmonic mean of the precision and recall;
- ROC AUC: Compute Area Under the Receiver Operating Characteristic Curve from prediction scores, computed following Hand and Till (2001);
- Matthews correlation coefficient (MCC): computed following Gorodkin (2004), the MCC is used in machine learning as a measure of the quality of binary and multi-class classifications. The MCC provides a balanced measure of classification quality that accounts for all four entries of the confusion matrix. It ranges from -1 (inverse prediction) through 0 (random prediction) to +1 (perfect prediction), and remains informative even when class sizes are unequal — an important property given the imbalanced distribution of site types in the EEA network. This statistic is also known as the phi coefficient;
- Tolerant Accuracy: weighted version of the accuracy defined for our purpose. The confusion matrix is first normalised over the rows, and then multiplied with the matrix W defined in Equation (1). The result is finally divided by $Tr(W)$. This normalises the result to the interval [0, 1], where 1 is the maximum achievable value corresponding to perfect classification across all four groups. Weights are maximum on the diagonal, but they are not zero on either side, which allows for greater tolerance for small differences between predicted and true values (i.e. stations which are predicted one

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

class "away" from their labelled class still contribute to increasing the tolerant accuracy score). Note that the ordering of classes in the confusion matrix rows/columns is important in the definition of this score: classes deemed close to each other must be adjacent.

For multi-class classification, accuracy, precision, recall, and F1 score are computed for each label in a one-vs-rest fashion, and the unweighted mean is taken (`macro` option in `scikit-learn`). Furthermore, ROC AUC is computed using the one-vs-one strategy, that computes the average AUC of all possible pairwise combinations of classes, and is insensitive to class imbalance (`ovo` option).

$$W = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix} \tag{1}$$

5.1.3 Preliminary algorithm selection

We used the `tsai` library (Oguiza, 2023), an open-source Python package built on `PyTorch` and `fastai` that implements a range of deep learning architectures for time series classification. In the preliminary study, using the HP training set and LP validation set (see Section 5.1.1), the following algorithms have been compared with a fixed learning rate (0.001) and batch size (64), and 130 epochs:

- InceptionTime (Ismail Fawaz et al., 2020);
- XceptionTime (Rahimian et al., 2019);
- MiniRocket (Tan et al., 2022);
- ResCNN - 1D-ResCNN (Zou et al., 2019);
- ResNet - Residual Network (Wang et al., 2017);
- FCN - Fully Convolutional Network (Wang et al., 2017);
- mWDN - Multilevel wavelet decomposition network (Wang et al., 2018): 4 levels;
- LSTM-FCN (Karim et al., 2017), 2 flavours: default, shuffle: False;
- LSTM (Hochreiter and Schmidhuber, 1997), 6 flavours: varying layers number 1 to 3, and bidirectional True/False.

The algorithms that gave the best performances according to the statistical indicators defined above were: InceptionTime, XceptionTime, MiniRocket, ResCNN, ResNet, mWDN. It was necessary to set aside the other, less effective algorithms in order to reduce the computational cost and simplify the interpretation of the remainder of the study.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

5.1.4 Best configurations per species

Building on our preliminary tests, the 6 selected algorithms have then been tested with different combinations of learning rate (lr) and batch size (bs), and 140 epochs:

- bs = 64, lr = 1e-4;
- bs = 32, lr = 1e-4;
- bs = 64, lr = 1e-3;
- bs = 32, lr = 1e-3;
- bs = 64, lr = 1e-2;
- bs = 32, lr = 1e-2.

On Figure 4, heatmaps of the statistical scores, defined in Section 5.1.2, allow us to choose the best combination of algorithm and parameters, for each species. Interestingly, no single algorithm outperforms all others across all species, but convolutional architectures (InceptionTime, XceptionTime, ResNet) dominate across all seven species, with none of the wavelet-based (mWDN) or rocket-based (MiniRocket) architectures selected as the best configuration for any species. This suggests that local temporal patterns — such as the diurnal cycle and weekend effect — are more discriminant than long-range dependencies for this classification deep learning task. Furthermore, the performance of a single algorithm can vary significantly depending on the hyperparameters chosen. Table 3 gives the selected combinations. The loss curves (not shown) confirmed that the number of epochs (140) is suitable for all the runs.

Species	Algorithm	Batch Size	Learning Rate
O_3	ResNet	32	1e-4
NO	XceptionTime	64	1e-3
NO_2	InceptionTime	32	1e-3
CO	InceptionTime	32	1e-3
SO_2	InceptionTime	32	1e-3
PM_{10}	InceptionTime	32	1e-2
$PM_{2.5}$	ResNet	32	1e-3

Table 3. Best combination of algorithms and parameters

5.2 Classification procedure

The aim of the preliminary study was to select the optimal settings for each species, and it was based on an arbitrary division of the dataset into HP data for training and LP data for validation. The next step is to define a method based on the deep learning algorithms selected previously, in order to classify the entire dataset.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

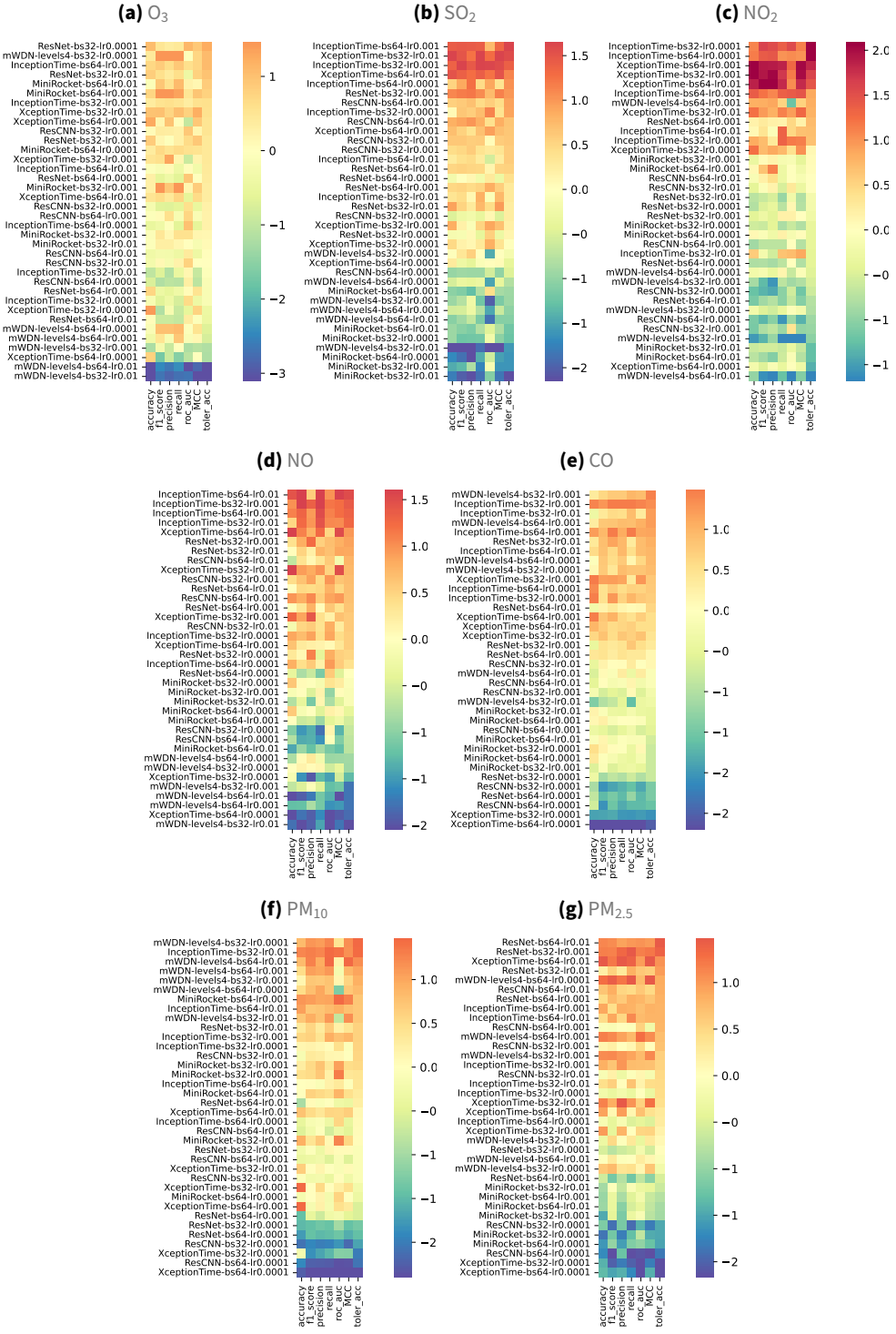


Figure 4. Comparison of algorithms, using a combination of different learning rates and batch sizes. Normalised values of the performance indicators defined in Section 5.1.2.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

The annual time series are now distributed into 10 subsets with similar properties. To do this, for each family (R, S, U, and T), sites are sorted by country (in alphabetical order of station codes) and distributed successively within each subset (in a round-robin fashion). This ensures that each family and each country/region is represented equally in each of the 10 subsets. All annual time series from a given site are assigned to the same subset, ensuring that no site contributes to both training and prediction in any given fold. Ten subsets is an arbitrary choice that strikes a balance between having a large enough training dataset to be representative of the whole, and the computational cost associated with multiple training runs.

Then, the classification method is applied to predict the classes of one subset, based on the data from the remaining 9 subsets. It is therefore assumed that sites whose metadata label is inconsistent with their actual environment are sufficiently rare and heterogeneous that their characteristics will not be systematically learnt by the classifier and propagated to inference.

After classification, each annual time series is given a predicted class. Since 8 years of data are considered, for a given species, each measurement station is assigned 1 to 8 different predicted labels. A weighted average of the different labels is then calculated. First, the values `[1, 2, 3, 4]` are mapped to the labels `[R, S, U, T]`, to order site types along a gradient of anthropogenic influence, from rural background (R) to urban traffic (T), consistent with the main axis used in the LDA reference method. Then, 8 weights `[0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1]` are used in order to give more importance to the most recent time series. Those values were tuned to strike a balance between the higher robustness of an unweighted average and the need to take account of potential evolution in spatial representativeness due to gradual changes in the environment of some sites. The range of the weights was tuned to obtain the best compromise. Finally, the weighted average produces a continuous value between 1 and 4, that is rounded to the nearest integer and mapped back to the corresponding label (R, S, U, or T). The 10-class output of the LDA reference method, which was defined to reflect a continuous gradient of spatial representativeness, is replaced here by a 4-class output aligned with the groups used as input. This simplification is motivated by the supervised nature of the deep learning approach, which learns directly from the metadata labels, and also by the operational needs of CAMS-regional — a prime use case of the classification method developed in this paper —, where the primary selection criterion is whether a site belongs to the background (R, S, U) or traffic (T) category (see Section 6.4).

6 Results

6.1 Performance comparison: LDA vs. deep learning

Performance is evaluated by direct comparison with the baseline, the LDA reference method. As explained in Section 2.5, the classes obtained in the LDA reference method are assigned in such a way that, if the method and the metadata matched perfectly, classes 1–3 would correspond to R sites, 4–5 to S sites, 6–7 to U sites, and 8–10 to T sites. This therefore allows a direct comparison between the two classifications, as the new method predicts directly the labels R, S, U and T.

Figures 5 and 6 compare the confusion matrices for each pollutant. This qualitative validation reveals a marked improvement with the new method, especially for S and U labels. The confusion matrix is much more diagonal. Best results are obtained for

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

NO_2, NO, and O_3, which was already the case with the LDA reference method, suggesting the information content of the time series varies systematically across species, regardless of the algorithm used.

The aim is not to obtain a perfectly diagonal matrix, as this would imply that the metadata and time series are in perfect alignment, or that deep learning has 'overfitted' the input data. Nevertheless, the fact that the matrices are becoming slightly more diagonal shows that deep learning enables information to be extracted from the time series more effectively, leading to a more precise prediction of the different categories. There is inevitably some irreducible error attributable to inconsistencies in the EEA metadata, which is not expected to align perfectly with the objective statistical properties of the time series. This margin of uncertainty varies depending on the species.

A more quantitative way of evaluating the results involves using the statistical scores defined in Section 5.1.2. Figure 7 shows that performance has improved for all scores and all pollutants. As in the LDA reference method, best scores are achieved for NO and NO_2. Performance is lower for SO_2, but significantly improved compared to the LDA reference method. Table 4 quantifies the change in the scores using the values plotted on Figure 7. Interestingly, it is for CO, SO_2, PM_{10}, and $PM_{2.5}$ that the new vs. old ratio is highest. The MCC ratio exceeds 2.0 for SO_2, PM_{10}, and $PM_{2.5}$, indicating that the deep learning method more than doubles the correlation for these species compared to the LDA reference method. This demonstrates the added value of the deep learning for the species that are most difficult to categorise based on the time series.

species	accuracy	precision	recall	f1_score	roc_auc	MCC	toler_acc
O_3	1.29	1.36	1.23	1.27	1.15	1.64	1.10
NO_2	1.21	1.21	1.17	1.19	1.12	1.39	1.08
NO	1.17	1.18	1.14	1.15	1.09	1.30	1.06
SO_2	1.34	1.57	1.21	1.27	1.15	2.46	1.11
CO	1.25	1.57	1.42	1.48	1.13	1.75	1.15
PM_{10}	1.39	1.47	1.29	1.35	1.18	2.13	1.16
$PM_{2.5}$	1.43	1.55	1.36	1.42	1.19	2.50	1.21

Table 4. Ratio of the new DL-based vs. old LDA-based classification scores.

6.2 Geographical distribution

A key aim of the classification approach since Joly and Peuch (2012) has been to harmonise the categorisation of monitoring sites at the European scale. This harmonisation may lead to category changes that follow a geographical pattern. This is explored in Figure 8, which focuses on the most significant inconsistencies between the metadata and the predicted label (S stations are therefore not considered in this section). Reassuringly, the discrepancies do not form significant clusters, but are scattered across the entire area. Even specific regions, such as the Canary Islands or the Alpine valleys, are not collectively sidelined into a category that is too distinct from the metadata. Table 5 summarises the anomalies shown in Figure 8. It can be seen that for ozone and NO_2, underestimation predominates (U or T sites classified as R), whilst for the other species – particularly CO, SO_2 and $PM_{2.5}$ – overestimation prevails. Interpreting these results is challenging due to the many factors that

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

may come into play, such as the titration of ozone by NO_x in urban areas, its long-range transport, and, for primary pollutants, topography, climatology and the presence of emission sources not accounted for in the metadata.

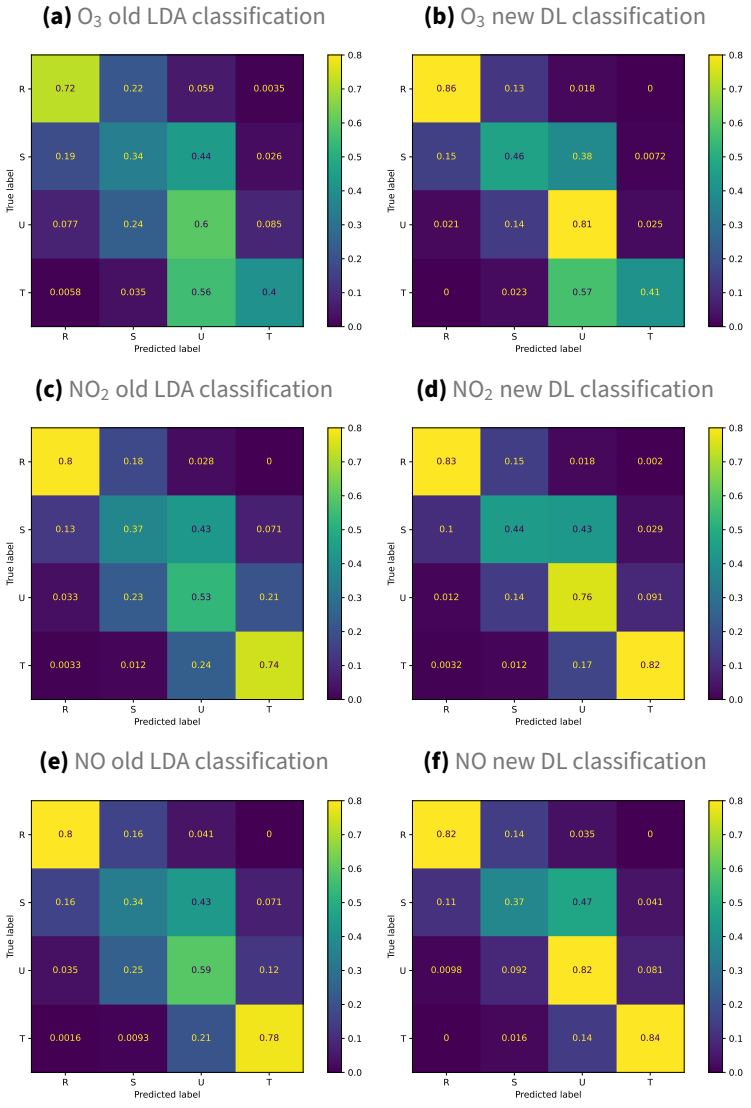


Figure 5. Comparison of the old LDA-based *versus* new DL-based classifications: O_3, NO_2, NO. Confusion matrices of the predicted (x-axis) *versus* "True" (i.e. medatada, y-axis) labels.

6.3 Case studies

Since the best statistical scores are achieved for the classification of NO_2 and NO (cf. Figure 7), it is instructive to examine more closely those discrepancies in Figure 5 where a few U or T sites are classified R, and conversely a few R sites are classified

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

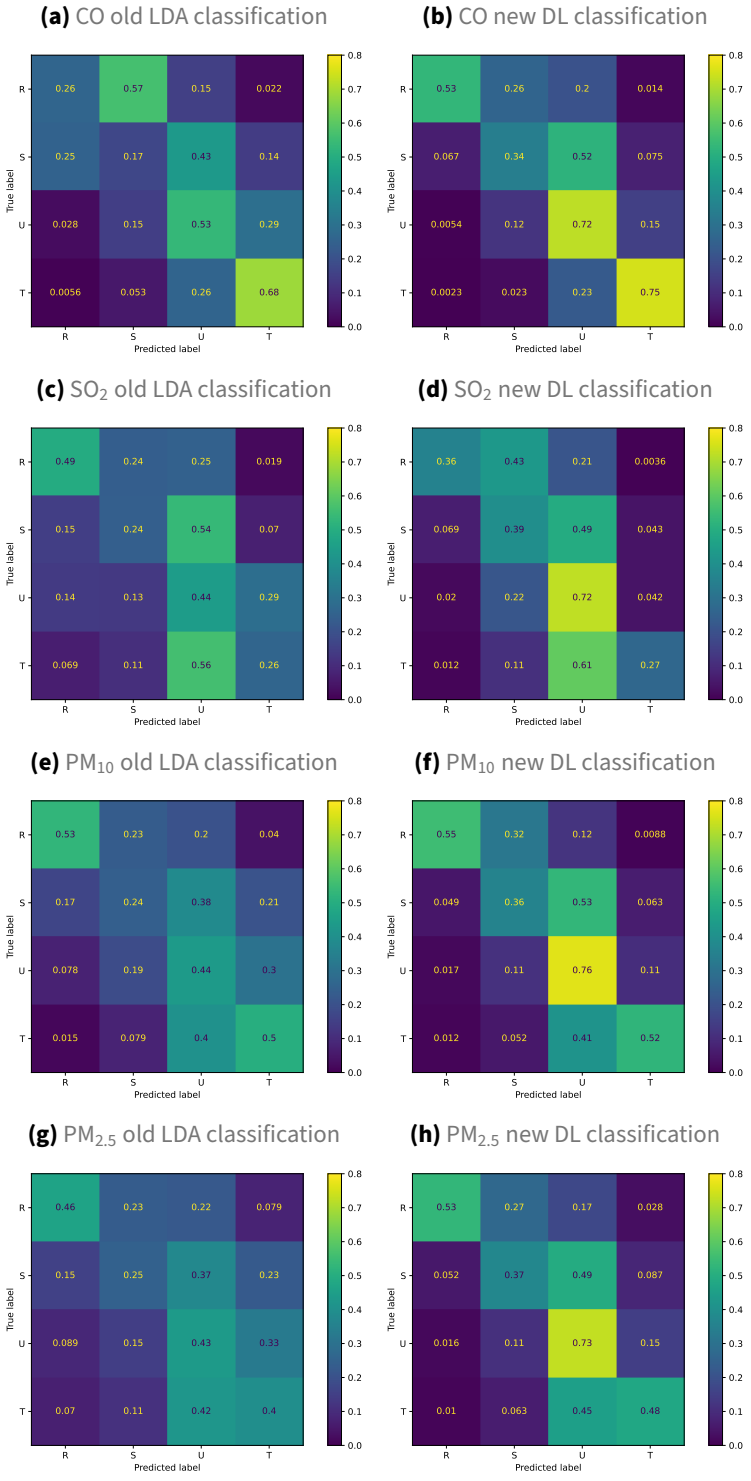


Figure 6. Comparison of the old LDA-based *versus* new DL-based classifications: CO, SO_2, PM_{10}, $PM_{2.5}$. Confusion matrices of the predicted (x-axis) *versus* "True" (i.e. metadata, y-axis) labels.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

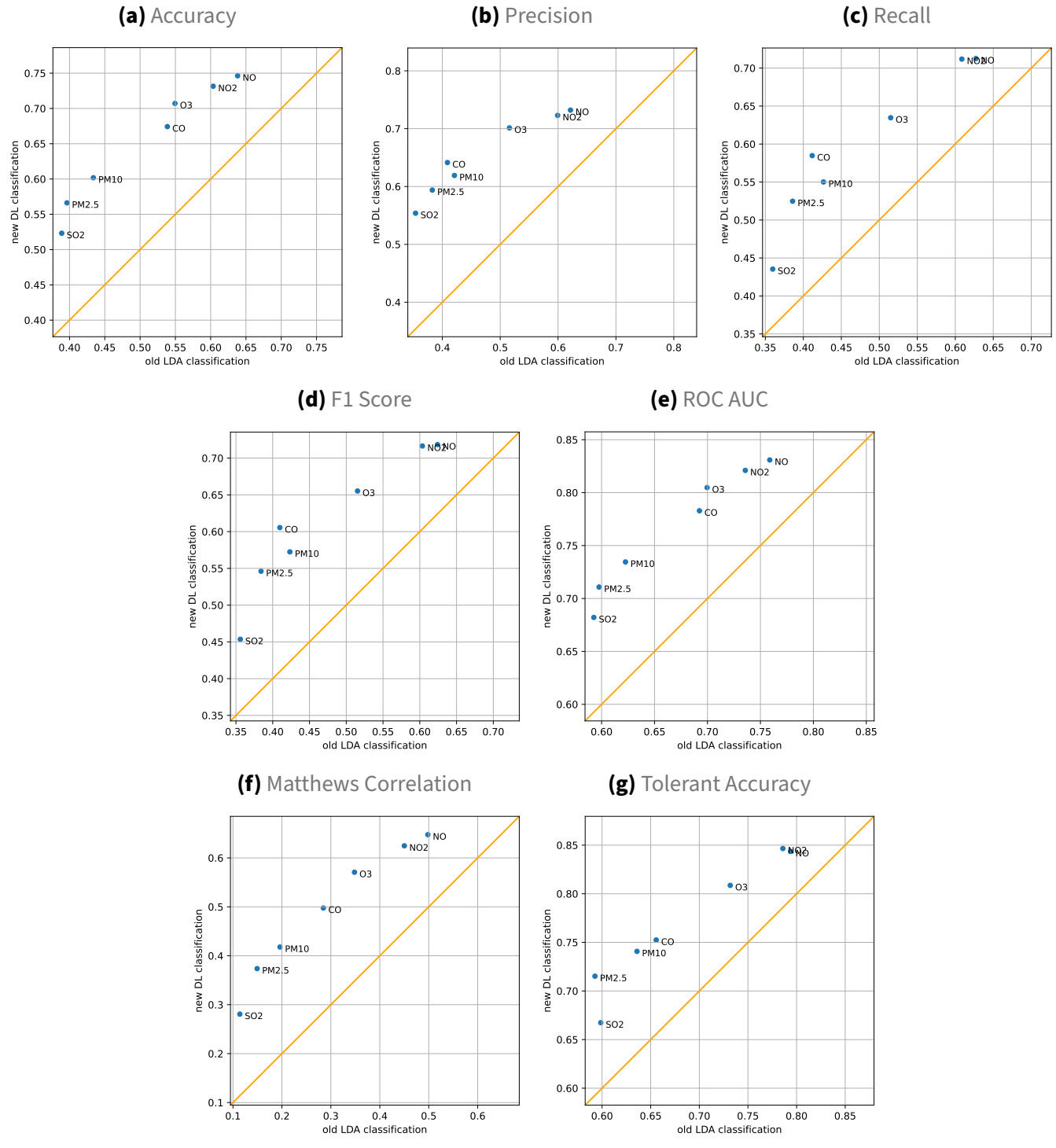


Figure 7. Comparison of the old LDA-based (x-axis) *versus* new DL-based (y-axis) classifications. Scores defined in Section 5.1.2.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

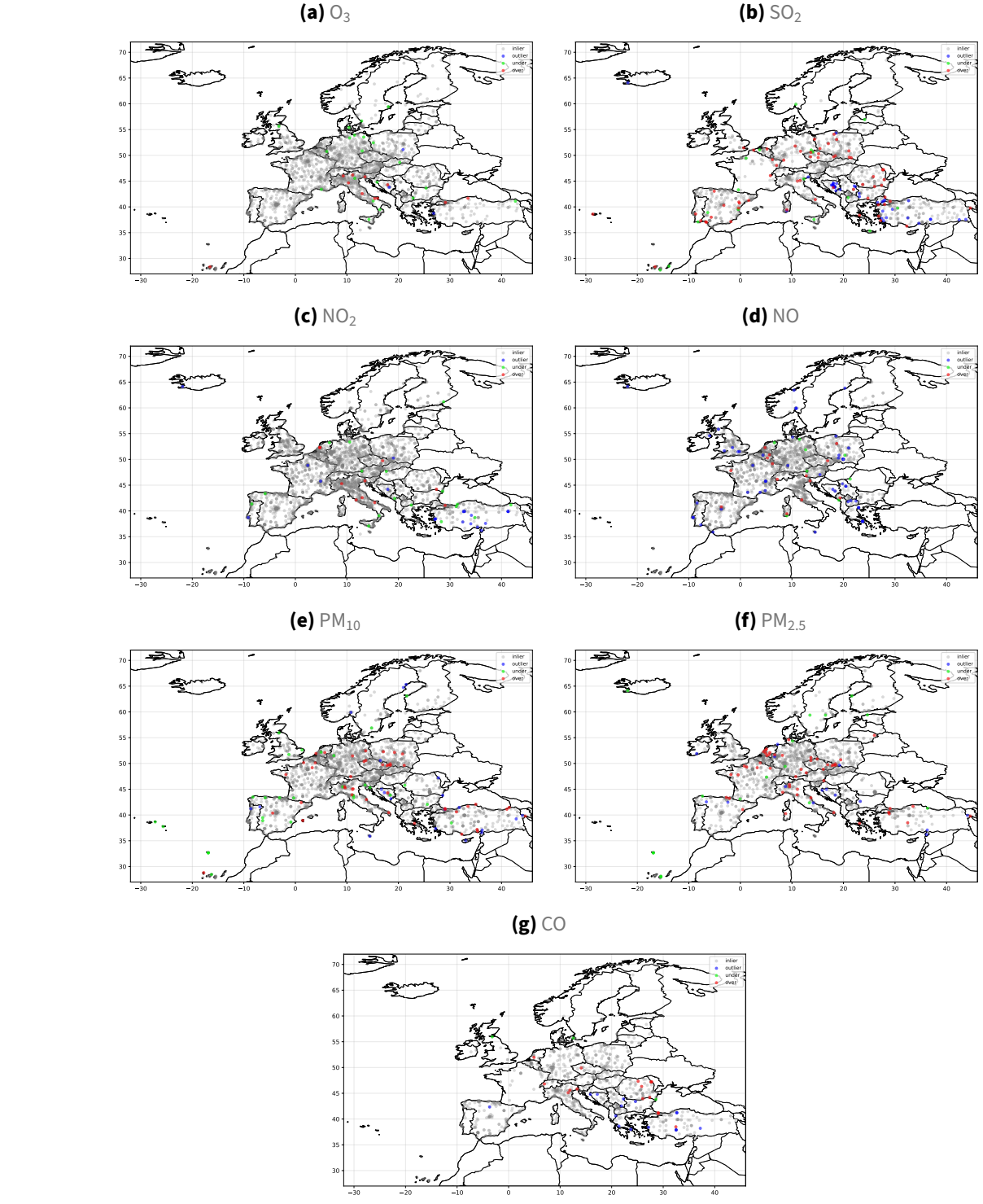


Figure 8. Map of the main discrepancies between the predicted *versus* true labels. *over* (red dots) is for R sites predicted U or T, and *under* (green dots) is for U or T sites predicted R. Sites detected outliers are marked in blue, and the rest are marked in grey.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

Species	Outlier	Over	Under
O_3	3	11	20
NO_2	21	10	17
NO	50	14	8
CO	15	15	3
SO_2	38	60	17
PM_{10}	20	44	28
$PM_{2.5}$	15	50	16

Table 5. Counts of the main discrepancies between the predicted *versus* "True" (i.e. medatada) labels. *over* is for R sites predicted U or T, and *under* is for U or T sites predicted R.

U or T. Google Earth (imagery in date of February 2026) was therefore used to visualise the surroundings of each of the 49 sites identified 'over' or 'under' on Figures 8 c) and d).

This approach does not constitute a quantitative validation, because a simple aerial photograph does not allow for an evaluation of emission sources in the vicinity of a monitoring station. Nevertheless, this provides a qualitative illustration of the types of environment where metadata and time-series-based classification diverge. Furthermore, the 49 images cannot all be included in this article, but Tables 6 and 7 list the selected sites with their coordinates. A brief summary of relevant remarks is provided here.

Concerning R sites predicted U or T (i.e. 'over' sites in Table 6). It was observed that some "rural background" sites (according to the metadata) are sometimes very close to a road and/or a substantial town (e.g., IT1174A and IT0884A). In some cases, the "suburban" type would be more appropriate. Sometimes, the site is close to an airport (e.g., NL00564 and NL00565), or a coastline with boat traffic (e.g., TR340411, BETN046, or FR01060). There are also ambiguous cases (e.g., CZ0MLOS) in which the precise metadata is "rural-nearcity background", and the classification predicts U: neither is unreasonable given the location. Finally, in some cases, the "rural background site" is surrounded with large cultivated fields (e.g., CZ0MLOS, CH0024A, or PL0068A), probably exposed to NO_x emissions by the use of fertilisers (Oikawa et al., 2015; Lieb et al., 2024).

Concerning the U or T sites predicted R (i.e. 'under' sites in Table 7). It was observed that in many cases, those "urban background" sites (according to the metadata) are located in a residential area, and the "suburban" category would be more appropriate. In some cases, those "urban background" sites are at the interface with the countryside (e.g., DEBY079, ES1572A, or TR0020A), in an environment quite similar to some "rural background" sites classified R in the previous case study. Depending on the main winds, and the orography, the influence of emission sources on measured time series can be different.

These case studies illustrate that the difficulty of classification is inherent to the strong complexity of the environments and the various influences of monitoring sites. Both the metadata and the time-series-based approaches carry complementary information, and neither should be considered definitive in isolation.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

code	metadata	classification	species	latitude	longitude
IT2132A	rural background	U	NO_2 and NO	45.83643	13.50112
NL00565	rural background	U	NO_2 and NO	52.27995	4.77067
IT0989A	rural background	U	NO_2	42.57259	12.96206
IT1174A	rural background	U	NO_2	45.28278	8.98944
IT2141A	rural background	U	NO_2	41.63083	15.38667
IT2248A	rural background	U	NO_2	42.13731	11.79317
NL00564	rural background	U	NO_2	52.32746	4.71501
RO0213A	rural background	U	NO_2	44.19681	27.39134
AT72912	rural background	U	NO	46.81911	12.76603
BETN046	rural background	U	NO	50.8823	5.61887
BETN093	rural background	U	NO	50.27448	5.23517
CH0024A	rural background	U	NO	46.139	7.14833
ES1805A	rural background	U	NO	40.78063	-3.70215
IT0884A	rural background	U	NO	42.15786	11.90895
IT1375A	rural background	U	NO	39.44361	9.01528
ME0012A	rural background	U	NO	42.5176	19.32774
PL0068A	rural background	U	NO	53.08065	18.68426
TR340411	rural background	T	NO_2	41.0746	29.05899
FR01060	rural-nearcity background	U	NO	49.16885	6.21267
CZ0MLOS	rural-nearcity background	U	NO_2 and NO	49.74792	16.92578
FR19008	rural-remote background	U	NO	47.83192	-1.83633

Table 6. NO_2 and NO: site information for the main discrepancies between the predicted *versus* "True" (i.e. medatada) labels (cf. maps c and d on Figure 8). *over* case: for R sites predicted U or T.

6.4 Implications for CAMS-regional operations

In the CAMS-regional service (Colette et al., 2025), for operational tasks of data assimilation and model verification, the observational data is first filtered by selecting only the “background” sites according to the metadata. Then, the sites are further selected using the results of the objective classification, taking classes 1-7 (with the LDA reference method) or labels R, S, or U with the new method.

In this subsection, a theoretical analysis is carried out for the sites selected from the complete dataset, bearing in mind that, in operational use, daily unvalidated real-time data is less comprehensive, as some suppliers provide their data with a certain delay. Table 8 shows that in the end, the number of sites that could potentially be rejected is much lower with the new classification (below 15%). This is particularly striking for CO, SO_2, PM_{10} and $PM_{2.5}$, for which 26% to 34% of sites were

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

code	metadata	classification	species	latitude	longitude
FI00424	urban traffic	R	NO_2	61.18892	28.77012
IT2111A	urban traffic	R	NO_2	39.07497	16.23145
ME0011A	urban traffic	R	NO_2	42.42021	19.25525
DEBY079	urban background	R	NO_2 and NO	47.73344	12.89124
NL00938	urban background	R	NO_2 and NO	53.2465	6.60904
DESH058	urban background	R	NO_2	53.38234	10.54052
ES1572A	urban background	R	NO_2	43.3727	-5.8728
HU0025A	urban background	R	NO_2	47.67611	17.64974
IT2207A	urban background	R	NO_2	37.07569	14.22384
MK0051A	urban background	R	NO_2	41.13914	22.49604
PT01052	urban background	R	NO_2	41.35389	-8.46083
RO0226A	urban background	R	NO_2	43.81685	28.56623
TR0020A	urban background	R	NO_2	41.17528	41.81829
TR090641	urban background	R	NO_2	37.90858	28.32192
TR500241	urban background	R	NO_2	38.71616	34.84731
TR670541	urban background	R	NO_2	41.27647	31.43342
TR810341	urban background	R	NO_2	40.86073	31.22865
DEMV028	urban background	R	NO	53.90226	11.47596
DEST066	urban background	R	NO	51.87085	12.66366
IT1947A	urban background	R	NO	39.06667	9.00889
ME0008A	urban background	R	NO	42.10383	19.09431
PL0698A	urban background	R	NO	50.80961	20.26603
RO0096A	urban background	R	NO	46.19046	21.30253

Table 7. NO_2 and NO: site informations for the main discrepancies between the predicted *versus* "True" (i.e. medatada) labels (cf. maps c and d on Figure 8). *under* case: U or T sites predicted R.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

previously excluded, compared with 6% to 14% now. This means more sites are likely to be included for assimilation and verification, improving spatial coverage.

This is illustrated with Figures 9 and 10. The new method improves the spatial coverage of the observational dataset potentially available for the CAMS regional service. Certain regions, such as Turkey, the Balkans and Poland, are now much better represented, and the improvements apply across the whole of Europe for CO and $PM_{2.5}$. This is partly due to the new method's much greater tolerance of missing values. The LDA reference method required very comprehensive hourly data in order to calculate certain indicators, such as the weekend effect (which requires a minimum number of quasi-complete Sunday–Monday pairs).

Species	Background	LDA Rejected	DL Rejected
O_3	2111	198 (9%)	30 (1%)
NO	1618	262 (16%)	97 (6%)
NO_2	2257	377 (17%)	138 (6%)
CO	574	196 (34%)	78 (14%)
SO_2	1240	318 (26%)	76 (6%)
PM_{10}	1873	512 (27%)	160 (9%)
$PM_{2.5}$	1242	359 (29%)	148 (12%)

Table 8. Among Background sites (according to metadata), number of sites potentially rejected by the CAMS-regional service (outlier or predicted T label) with the LDA reference method and with the new deep learning method.

7 Conclusions

This study presents a new method for the objective classification of European air quality monitoring sites, based on eight years of hourly observations of seven air pollutants from the EEA AQ e-reporting network. The method replaces the LDA-based algorithm that was used operationally to provide site lists to the CAMS-regional service, with a three-stage approach: gap filling and pre-processing of the raw time series, unsupervised outlier detection using the Local Outlier Factor algorithm applied to a comprehensive set of automatically extracted statistical features, and supervised classification using deep learning architectures operating directly on the annual time series. For each species, the optimal algorithm and hyperparameters were selected through a preliminary study based on a separation of sites into high-potential and low-potential subsets, ensuring that the training data was as reliable as possible given the inherent subjectivity of the EEA metadata. Section 2 also provides the first full published description of the LDA reference method that has been used for CAMS-regional from 2017 to 2025, serving as the quantitative baseline for the evaluation.

The new method yields consistent improvements across all seven species and all performance metrics compared with the LDA reference method. The gains are most substantial for the species that were previously most difficult to classify — SO_2, PM_{10}, and $PM_{2.5}$ — where the Matthews correlation coefficient improves by more than a factor of two. The improvement is

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

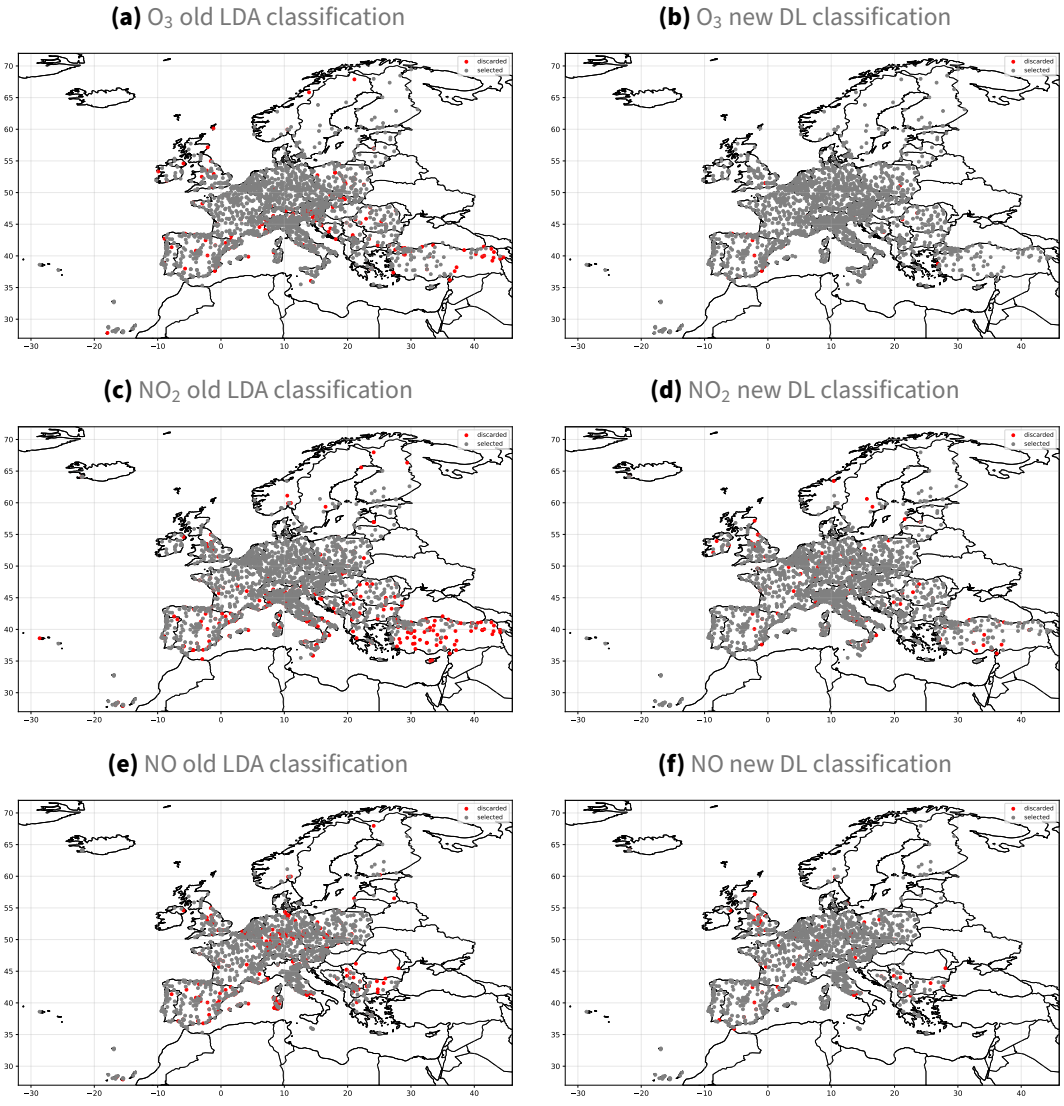


Figure 9. Localisation of Background sites (according to metadata) potentially selected by the CAMS-regional service (R, S, or U predicted label ; in grey), or rejected (outlier or predicted T label ; in red): O_3, NO_2, NO.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

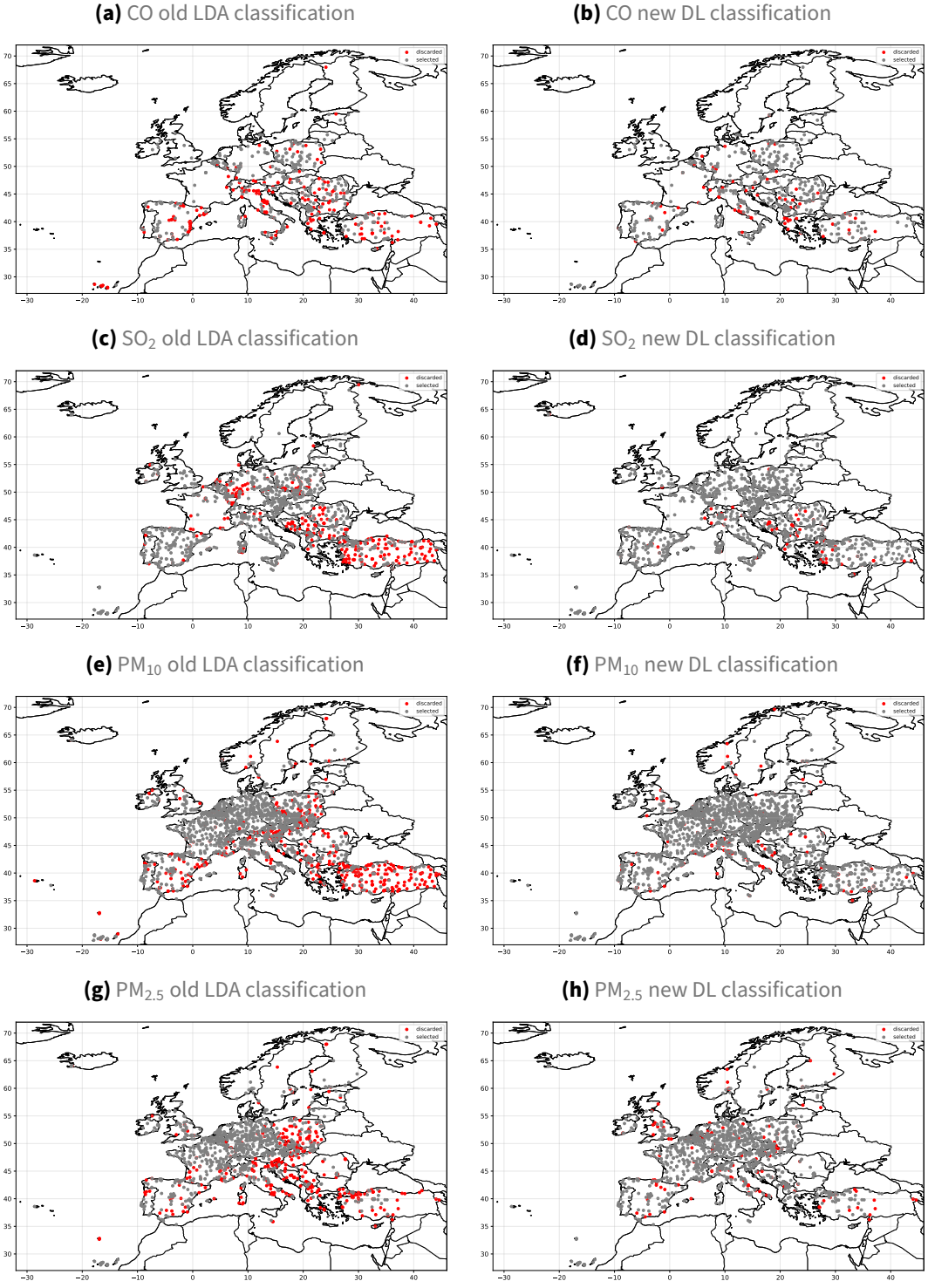


Figure 10. Localisation of Background sites (according to metadata) potentially selected by the CAMS-regional service (R, S, or U predicted label ; in grey), or rejected (outlier or predicted T label ; in red): CO, SO_2, PM_{10}, $PM_{2.5}$.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

attributable to two complementary factors: the ability of deep learning architectures to extract discriminant information directly from the time series without manual feature engineering, and the greater tolerance of the method for missing data, which was a significant source of site rejection in the LDA approach. As a direct consequence, the fraction of background sites rejected in the list of sites provided to CAMS-regional drops from between 9% and 34% under the LDA method to between 1% and 14% under the new approach, resulting in substantially improved spatial coverage across Europe, particularly in regions such as Turkey and parts of eastern Europe where data completeness is lower.

Qualitative case studies based on aerial imagery confirm that the residual discrepancies between the predicted labels and the EEA metadata are largely attributable to genuine ambiguities in the classification of sites: stations located at the urban-rural interface, rural sites exposed to agricultural NO_x emissions, or urban background sites whose time series are more characteristic of a suburban environment. These cases illustrate both the intrinsic difficulty of the classification problem and the value of a time-series-based objective method as a complement to operator-assigned metadata: neither source of information is definitive in isolation, and the two are most informative when used together.

The main remaining limitation of the method is the requirement for at least one full year of observations before a site can be classified, which means that newly commissioned monitoring sites cannot be incorporated immediately. This constraint is shared with all time-series-based classification approaches and is unlikely to be resolved without introducing external data sources. A promising direction for future work would be to combine the time-series-based approach proposed here with high-resolution geospatial data — population density maps, topographic data, land-use classifications, and emission inventories — to produce a classification that is less dependent on the EEA metadata, which in the present study necessarily acts as both training label and validation reference. Renna et al. (2026) demonstrated the feasibility of a complementary approach on a global scale, using convolutional neural networks trained on satellite land cover imagery to classify particulate matter monitoring sites: their method captures the static environment of a site, while the approach presented here captures the integrated dynamical influence of that environment on observed concentrations. Combining both sources of information within a unified framework represents a natural and scientifically well-motivated direction for further development of the classification method.

Code availability. The method has been coded as a Python package and made publicly available in HAL (Joly, 2026), under the CeCILL-C free software licence agreement.

Data availability. EEA AQ e-reporting data (E1a and E2a fluxes) are publicly available at https://aqportal.discomap.eea.europa.eu/. Data were downloaded in January 2026.

Author contributions. MJ conceived and designed the study (Conceptualization), developed the classification methodology (Methodology), implemented all Python code (Software), performed the statistical analysis (Formal analysis), conducted the investigation (Investigation),

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

produced all figures (Visualization), and wrote the manuscript in its entirety (Writing – original draft preparation). VP developed and maintained the scripts used to download and format the EEA observational dataset prior to this study (Software) and contributed to manuscript review (Writing – review and editing). JJH carefully reviewed the manuscript and provided detailed comments, rephrasing, and critical questions (Writing – review and editing). CC read the manuscript and proposed minor revisions (Writing – review and editing), and contributed to funding acquisition for the CAMS2_40_bis project (Funding acquisition). OF, FB, AC, FM, and VG contributed to project administration and supervision of the broader CAMS2_40_bis project within which this study was conducted (Project administration, Supervision).

Competing interests. The authors declare that they have no conflict of interest.

Disclaimer. This document describes the improvement of a method used to provide data to the Copernicus Atmosphere Monitoring Service implemented by ECMWF on behalf of the European Commission with funding from the European Union. The views and opinions expressed herein are those of the authors only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

Acknowledgements. The authors acknowledge the European Environment Agency (EEA) for providing the AQ e-reporting dataset used throughout this study. The authors thank Météo-France CNRM for providing the computing resources used to train the deep learning models.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.

References

Barrero, M., Orza, J., Cabello, M., and Cantón, L.: Categorisation of air quality monitoring stations by evaluation of PM10 variability, Science of The Total Environment, 524, 225–236, https://doi.org/10.1016/j.scitotenv.2015.03.138, 2015.

Box, G. E. and Cox, D. R.: An analysis of transformations, Journal of the Royal Statistical Society Series B: Statistical Methodology, 26, 211–243, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x, 1964.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J.: LOF: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 93–104, https://doi.org/10.1145/335191.335388, 2000.

Colette, A., Collin, G., Besson, F., Blot, E., Guidard, V., Meleux, F., Royer, A., Petiot, V., Miller, C., Fermond, O., et al.: Copernicus Atmosphere Monitoring Service–Regional Air Quality Production System v1. 0, Geoscientific Model Development, 18, 6835–6883, https://doi.org/10.5194/gmd-18-6835-2025, 2025.

Duyzer, J., van den Hout, D., Zandveld, P., and van Ratingen, S.: Representativeness of air quality monitoring networks, Atmospheric Environment, 104, 88–101, https://doi.org/10.1016/j.atmosenv.2014.12.067, 2015.

Fritsch, F. N. and Butland, J.: A method for constructing local monotone piecewise cubic interpolants, SIAM journal on scientific and statistical computing, 5, 300–304, https://doi.org/10.1137/0905021, 1984.

Gorodkin, J.: Comparing two K-category assignments by a K-category correlation coefficient, Computational biology and chemistry, 28, 367–374, https://doi.org/10.1016/j.compbiolchem.2004.09.006, 2004.

Gouveia, S., Scotto, M. G., Monteiro, A., and Alonso, A. M.: Wavelets-based clustering of air quality monitoring sites, Environmental monitoring and assessment, 187, 694, https://doi.org/10.1007/s10661-015-4901-z, 2015.

Hand, D. J. and Till, R. J.: A simple generalisation of the area under the ROC curve for multiple class classification problems, Machine learning, 45, 171–186, https://doi.org/10.1023/A:1010920819831, 2001.

Hermanns, A., Lange, A. C., Kowalski, J., Fuchs, H., and Franke, P.: Data clustering to optimise the representativity of observational data in air quality data assimilation: a case study with EURAD-IM (version 5.9. 1 DA), Geoscientific Model Development, 18, 9417–9432, https://doi.org/10.5194/gmd-18-9417-2025, 2025.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hohenberger, T. L., Che, W., Fung, J. C., and Lau, A. K.: A proposed population-health based metric for evaluating representativeness of air quality monitoring in cities: Using Hong Kong as a demonstration, PLoS One, 16, e0252 290, https://doi.org/10.1371/journal.pone.0252290, 2021.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., and Petitjean, F.: Inceptiontime: Finding alexnet for time series classification, Data mining and knowledge discovery, 34, 1936–1962, https://doi.org/10.1007/s10618-020-00710-y, 2020.

Janssen, S., Dumont, G., Fierens, F., Deutsch, F., Maiheu, B., Celis, D., Trimpeneers, E., and Mensink, C.: Land use to characterize spatial representativeness of air quality monitoring stations and its relevance for model validation, Atmospheric Environment, 59, 492–500, https://doi.org/10.1016/j.atmosenv.2012.05.028, 2012.

Joly, M.: A Deep Learning-Based Python Package for Classifying Air Quality Monitoring Sites, https://meteofrance.hal.science/meteo-05728373, 2026.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.



Joly, M. and Peuch, V.-H.: Objective classification of air quality monitoring sites over Europe, Atmospheric Environment, 47, 111–123, https://doi.org/10.1016/j.atmosenv.2011.11.025, 2012.

Karim, F., Majumdar, S., Darabi, H., and Chen, S.: LSTM fully convolutional networks for time series classification, IEEE access, 6, 1662–1669, https://doi.org/10.1109/ACCESS.2017.2779939, 2017.

Lieb, H. C., Maldonado, M., Ruiz, E., Torres, C., Olmedo, L., Walters, W. W., and Faloona, I. C.: Nitrogen isotopes reveal high NOx emissions from arid agricultural soils in the Salton Sea Air Basin, Scientific Reports, 14, 28 725, https://doi.org/10.1038/s41598-024-78361-y, 2024.

Liu, F. T., Ting, K. M., and Zhou, Z.-H.: Isolation forest, in: 2008 eighth ieee international conference on data mining, pp. 413–422, IEEE, https://doi.org/10.1109/ICDM.2008.17, 2008.

Malley, C. S., Braban, C. F., and Heal, M. R.: The application of hierarchical cluster analysis and non-negative matrix factorization to European atmospheric monitoring site classification, Atmospheric research, 138, 30–40, https://doi.org/10.1016/j.atmosres.2013.10.019, 2014.

Martin, F., Fileni, L., Palomino, I., Vivanco, M. G., and Garrido, J. L.: Analysis of the spatial representativeness of rural background monitoring stations in Spain, Atmospheric Pollution Research, 5, 779–788, https://doi.org/10.5094/APR.2014.087, 2014.

Martín, F., Santiago, J. L., Kracht, O., García, L., Gerboles, M., et al.: FAIRMODE Spatial representativeness feasibility study, Report EUR, 27385, https://doi.org/10.2788/49487, 2015.

Moler, C. B.: Numerical computing with MATLAB, SIAM, https://doi.org/10.1137/1.9780898717952, 2004.

Oguiza, I.: tsai - A state-of-the-art deep learning library for time series and sequential data, Github, https://github.com/timeseriesAI/tsai, 2023.

Oikawa, P., Ge, C., Wang, J., Eberwein, J., Liang, L., Allsman, L., Grantz, D., and Jenerette, G.: Unusually high soil nitrogen oxide emissions influence air quality in a high-temperature agricultural region, Nature communications, 6, 8753, https://doi.org/10.1038/ncomms9753, 2015.

Okamoto, S. and Tanimoto, H.: A review of atmospheric chemistry observations at mountain sites, Progress in Earth and Planetary Science, 3, 34, https://doi.org/10.1186/s40645-016-0109-2, 2016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, the Journal of machine Learning research, 12, 2825–2830, https://jmlr.org/papers/v12/pedregosa11a.html, 2011.

Piersanti, A., Vitali, L., Righini, G., Cremona, G., and Ciancarella, L.: Spatial representativeness of air quality monitoring stations: A grid model based approach, Atmospheric Pollution Research, 6, 953–960, https://doi.org/10.1016/j.apr.2015.04.005, 2015.

Rahimian, E., Zabihi, S., Atashzar, S. F., Asif, A., and Mohammadi, A.: Xceptiontime: A novel deep architecture based on depthwise separable convolutions for hand gesture classification, arXiv preprint arXiv:1911.03803, https://doi.org/10.48550/arXiv.1911.03803, 2019.

Renna, S., Rodriguez-Pardo, C., and Aleluia Reis, L.: A dataset of harmonized global air quality monitoring metadata, Scientific Data, 13, 466, https://doi.org/10.1038/s41597-026-06797-0, 2026.

Solazzo, E. and Galmarini, S.: Comparing apples with apples: Using spatially distributed time series of monitoring data for model evaluation, Atmospheric Environment, 112, 234–245, https://doi.org/10.1016/j.atmosenv.2015.04.037, 2015.

Tan, C. W., Dempster, A., Bergmeir, C., and Webb, G. I.: MultiRocket: multiple pooling operators and transformations for fast and effective time series classification: CW Tan, Data Mining and Knowledge Discovery, 36, 1623–1646, https://doi.org/10.1007/s10618-022-00844-1, 2022.

https://doi.org/10.5194/egusphere-2026-5616
Preprint. Discussion started: 22 November 2026
© Author(s) 2026. CC BY 4.0 License.



Tapia, O., Escudero, M., Lozano, Á., Anzano, J., and Mantilla, E.: New classification scheme for ozone monitoring stations based on frequency distribution of hourly data, Science of The Total Environment, 544, 1–9, https://doi.org/10.1016/j.scitotenv.2015.11.081, 2016.

Tüysüzoğlu, G., Birant, D., et al.: Classification of air quality network based on meteorological and pollutant factors, in: 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–6, IEEE, https://doi.org/10.1109/HORA49412.2020.9152910, 2020.

Wang, J., Wang, Z., Li, J., and Wu, J.: Multilevel wavelet decomposition network for interpretable time series analysis, in: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2437–2446, https://doi.org/10.1145/3219819.3220060, 2018.

Wang, Z., Yan, W., and Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline, in: 2017 International joint conference on neural networks (IJCNN), pp. 1578–1585, IEEE, https://doi.org/10.1109/IJCNN.2017.7966039, 2017.

Wikipedia: File:Population density by NUTS 3 region (2017).svg — Wikimedia Commons, the free media repository, https://commons.wikimedia.org/w/index.php?title=File:Population_density_by_NUTS_3_region_(2017).svg&oldid=1193633626, [Online; accessed 30-June-2026], 2026.

Zou, X., Wang, Z., Li, Q., and Sheng, W.: Integration of residual network and convolutional neural network along with various activation functions and global pooling for time series classification, Neurocomputing, 367, 39–45, https://doi.org/10.1016/j.neucom.2019.08.023, 2019.