Ozone exceedance forecasting with enhanced extreme instance augmentation
A case study in Germany

Deng, Tuo; Manders, Astrid; Segers, Arjo; Heemink, Arnold Willem; Lin, Hai Xiang

DOI
10.1016/j.envsoft.2024.106162

Publication date
2024

Document Version
Final published version

Published in
Environmental Modelling and Software

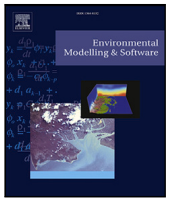
Citation (APA)
Deng, T., Manders, A., Segers, A., Heemink, A. W., & Lin, H. X. (2024). Ozone exceedance forecasting with enhanced extreme instance augmentation: A case study in Germany. *Environmental Modelling and Software*, *181*, Article 106162. https://doi.org/10.1016/j.envsoft.2024.106162

Environmental Modelling and Software 181 (2024) 106162

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft



Position Paper

Ozone exceedance forecasting with enhanced extreme instance augmentation: A case study in Germany

Tuo Deng [a,*], Astrid Manders [b], Arjo Segers [b], Arnold Willem Heemink [a], Hai Xiang Lin [a,c]

[a] Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands
[b] TNO, Department of Climate, Air and Sustainability, Utrecht, The Netherlands
[c] Institute of Environmental Sciences, Leiden University, The Netherlands

ARTICLE INFO

Keywords:
Air quality
Ozone prediction
Random forest
Extreme instance augmentation

ABSTRACT

Accurately forecasting ozone levels that exceed specific thresholds is pivotal for mitigating adverse effects on both the environment and public health. However, predicting such ozone exceedances remains challenging due to the infrequent occurrence of high-concentration ozone data. This research, leveraging data from 57 German monitoring stations from 1999 to 2018, introduces an Enhanced Extreme Instance Augmentation Random Forest (EEIA-RF) approach that significantly improves the prediction of days when the maximum daily 8-hour average ozone concentrations exceed 120 $\mu g/m^3$. A pre-trained machine learning model is used to generate additional high-concentration data, which, combined with selectively reduced low-concentration data, forms a new dataset for training a refined model. This method achieved an improvement of at least 8% in the accuracy of predicting days with ozone exceedances across Germany. Our experiment underscores the approach's value in enhancing atmospheric modeling and supporting public health advisories and environmental policy-making related to ozone pollution.

1. Introduction

Air pollution poses a critical challenge to environmental sustainability, public health, and the vitality of ecosystems. Ground-level ozone stands out as a particularly deleterious component, with well-documented ramifications including respiratory illnesses in humans and damage to agricultural and natural vegetation (Sicard et al., 2019). For the protection of human health, the European Union's (EU) air quality directives and the World Health Organization (WHO) guidelines set thresholds for maximum daily 8-h ozone mean concentrations (MDA8) (EU: 120 $\mu g/m^3$, WHO: 100 $\mu g/m^3$) (Hjellbrekke and Solberg, 2022). According to EU, the number of ozone exceedance days should not be more than 25 days per calendar year. Persistently high ozone concentrations continue to pose ecological threats, negatively impacting plant and animal life as well as human health across North America, Europe, and Asia (Zhang et al., 2019). Even under optimistic emission scenario pathways, ozone pollution will cause additional yield losses for wheat, soya bean and maize of between 0.1 and 11% globally by 2030 (Emberson, 2020). Therefore, it is essential to analyze and simulate long-term ozone changes to prevent air pollution and protect public and ecological health.

Tropospheric ozone is generally attributed to its in situ photochemical production and destruction coupled with regular intrusions of ozone-rich stratospheric air. Worldwide expansions in agriculture, transportation, and industry are producing a growing burden of waste gases, which include nitrogen oxides (especially NO and NO_2) and volatile organic compounds ($VOCs$). These gases enter the atmosphere and exacerbate the photochemical production of ozone. The atmospheric chemistry of tropospheric ozone formation is complex, making it challenging to simulate the process of tropospheric ozone formation (Lu et al., 2019). The formation of ozone can be described by the following reactions. Under the influence of solar radiation, nitrogen dioxide (NO_2) undergoes photodissociation to yield atomic oxygen, a process that initiates the formation of ozone (1). This atomic oxygen rapidly reacts with diatomic oxygen to form the ozone molecule (2).

$$NO_2 + h\upsilon \rightarrow NO + O \quad (1)$$

$$O + O_2 + M \rightarrow O_3 + M \quad (2)$$

In tandem, $VOCs$ play a crucial role in atmospheric chemistry by oxidizing nitric oxide (NO) into nitrogen dioxide, thus fueling the continuous production of ozone. During nightfall, particularly in areas with substantial nitric oxide emissions (e.g., power plants), ozone engages in a chemical reaction with nitrogen dioxide, leading to the formation of

* Corresponding author.
E-mail address: t.deng@tudelft.nl (T. Deng).

https://doi.org/10.1016/j.envsoft.2024.106162
Received 2 March 2024; Received in revised form 16 April 2024; Accepted 18 July 2024
Available online 25 July 2024

the nitrate radical (NO_3) (Finlayson-Pitts and Pitts, 1993).

$$NO_2 + O_3 \rightarrow NO_3 + O_2 \quad (3)$$

Climate change and the emission of ozone precursors directly affect the tropospheric ozone concentration. Recent research underscores the intricate relationship between meteorological conditions and ground-level ozone pollution. Elevated temperatures accelerate ozone formation by enhancing the rate of photochemical reactions, leading to higher ozone concentrations, particularly in urban areas (Li et al., 2023). Relative humidity plays a dual role; while it can facilitate ozone formation by contributing to the formation of precursor compounds, higher humidity generally leads to a decrease in ozone levels due to increased ozone scavenging (Li et al., 2021). Wind patterns significantly influence ozone distribution by dispersing or concentrating ozone and its precursors, affecting both local and regional ozone levels (Liu and Wang, 2020). Solar radiation, especially UV light, is a critical driver of ozone formation, as it initiates the photochemical reactions that produce ozone from its precursors (Fang et al., 2020).

Considering the complex formation of ground-level ozone and the multitude of influencing factors, the role of Chemical Transport Models (CTMs) is essential. These models not only help in predicting long-term ozone trends but also in comprehensively understanding the nuanced effects of various environmental and anthropogenic factors on atmospheric ozone concentrations. Travis and Jacob (2019) investigated the reasons for the difficulty during the ozone diurnal cycle simulation by CTM. The GEOS-Chem model was used to simulate the case study of the NASA SEAC4RS aircraft campaign in the Southeast U.S. It is found that the proper representation of diurnal variations in mixed layer dynamics and ozone deposition velocities is critical in models to describe the diurnal cycle of ozone. Ryu et al. (2019) evaluate and compare the WRF-Chem simulations driven by Rapid Refresh (RAP) and the Global Forecast System (GFS) forecasts over the Contiguous United States (CONUS) for 2016 summer. They found that the ozone concentration mainly responds to changes in the boundary layer height, which directly affects the mixing of ozone and its precursors. The article also shows that the CTM overpredicts the ozone concentration and gives false alarms. The ozone simulation error from the CTM model is not an exception. Manders et al. (2012) found that LOTOS-EUROS underestimates the daily ozone maximum, especially for the highest ozone peaks. According to previous model validation and intercomparison studies (Otero et al., 2018), the ozone simulation error could be caused by the uncertainties (emissions, meteorological parameters, etc.) in the CTM.

The interplay between ozone concentrations and meteorological factors advances, evidenced by research like Otero et al. (2018), the focus expands to include advanced modeling techniques. Traditional statistical learning methods and modern machine learning (ML) models, such as deep learning, are increasingly employed to understand these dynamics. Zhang et al. (2023) developed a 2-D convolutional neural network-surface ozone ensemble forecast (2DCNN-SOEF) system and applied it to 216-h ozone forecast in Shenzhen, China. The model can quantify the uncertainty of surface ozone forecast due to weather forecast uncertainties. Using spatial patterns of weather, the model detects the ozone-meteorology relationship. The XGBoost (Extreme Gradient Boosting) was used to simulate the variability in urban ozone over Doon valley of the Himalaya (Ojha et al., 2021). The model can reproduce the ozone data based on the training with past variations in ozone and meteorological conditions. Given the lack of high-resolution observation, ML simulations can be used to assess the regional impacts of ozone on health and agriculture. Feng et al. (2019) compared three different models – the Weather Research and Forecasting Model coupled with Chemistry (WRF-CAMQ), Random Forest (RF), and Recurrent Neural Networks (RNN) – for forecasting ozone pollution over a 24-h period in Hangzhou. The study highlighted the importance of individual features in the RF model and found that the RNN model outperformed the chemical transport model (CTM) and other ML models in predicting ozone levels.

Despite advancements in atmospheric research, accurately predicting high ozone concentrations, especially exceedances, poses a significant challenge for machine learning models (Eslami et al., 2020). Crucial for addressing environmental and public health concerns due to the serious health risks of high ozone levels (Zhang et al., 2019), this area of research confronts a notable difficulty: the infrequency of high ozone exceedances leads to imbalanced datasets with sparse high-concentration observations. Such scarcity complicates the task of machine learning models in forecasting these rare but vital events (Gong and Ordieres-Meré, 2016; Fan et al., 2022). Therefore, the development of innovative machine learning strategies that effectively handle imbalanced datasets to enhance the precision of high ozone predictions is urgently needed (Chao and Zhang, 2023).

To tackle the challenge of imbalanced datasets in atmospheric science, researchers have explored innovative methods for more accurate predictions. Gong and Ordieres-Meré (2016) investigated various resampling techniques like the Synthetic Minority Over-sampling Technique (SMOTE) alongside machine learning algorithms to balance datasets. Similarly, Tsai et al. (2009) utilized cost-sensitive neural networks, adding weighted losses to focus on rare high ozone days. Recently, Vicente et al. (2024) introduced two distinct approaches to address data imbalance in atmospheric modeling. The first is a threshold-moving method, which adjusts classification thresholds to enhance the model's sensitivity to less frequent high ozone events. The second approach, an error-tolerance increment method, proposes increasing the tolerance for classifying a prediction as an exceedance by lowering the exceedance threshold. These methods enhance model performance by focusing on rare but critical categories, such as extreme ozone pollution, thereby improving the model's prediction accuracy for high ozone concentrations.

To address the challenge of poor prediction accuracy on days with ozone exceedances (MDA8 over 120 $\mu g/m^3$), due to the rarity of such events, we introduce the Enhanced Extreme Instance Augmentation for Random Forest Modelling (EEIA-RFM) method. This approach is designed to overcome the issues of imbalanced datasets and the scarcity of extreme ozone events in the measurement datasets. Through the generation of synthetic data mimicking these rare occurrences, EEIA-RFM substantially boosts model precision. Specifically, our study seeks to contribute to this evolving field by investigating the relationship between Maximum Daily 8-h Average (MDA8) ozone concentrations and meteorological factors in Germany. Utilizing data from 57 monitoring stations spanning from 1999 to 2018, we address the uneven distribution and rarity of high-concentration ozone events. We employ K-means clustering, guided by the Elbow Method, to categorize observation sites based on their geographical coordinates. This step ensures our analysis is sensitive to the distinct environmental conditions of each region. In each cluster, the EEIA-RFM method augments the dataset with synthetic instances of high ozone days. By incorporating more generated exceedance scenarios, the model gains additional insights into the characteristics of ozone exceedances, allowing for more nuanced model training. Collectively, these methods enable more accurate and region-specific forecasting of ozone levels, a crucial factor in understanding atmospheric dynamics and their public health implications.

In Section 2, we delve into the specifics of our data processing methodologies and outline the evaluation criteria that underpin our research. Section 3 details the machine learning methods we have applied, including K-means clustering, Random Forest, and our innovative EEIA-RFM approach. Section 4 presents the outcomes of our clustering analysis and conducts a thorough comparison of different models in forecasting ozone levels. Finally, Section 5 brings together our key findings, offering a comprehensive summary and drawing conclusions from our extensive study.

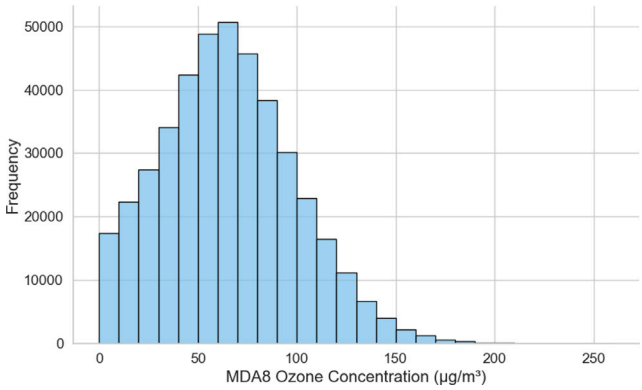


Fig. 1. Distribution of MDA8 ozone concentrations in German (1999–2018).

2. Dataset characterization and model evaluation framework

2.1. Dataset overview

The dataset employed in this study is comprised of meteorological and air pollution data. For meteorological insights, we utilized data from the E-OBS dataset (Cornes et al., 2018), offering daily observations across Europe at a $0.25° \times 0.25°$ spatial resolution. Our analysis included seven critical meteorological features: daily mean temperature (TG), daily maximum temperature (TX), daily precipitation sum (RR), daily averaged sea level pressure (PP), daily averaged relative humidity (HU), daily mean wind speed (FG), and daily mean global radiation (QQ). The comprehensive processing by E-OBS ensured the dataset was free from missing values, facilitating our analysis without the need for additional data imputation methods.

The air pollution component of our study was derived from data provided by the German Environment Agency (Umweltbundesamt, UBA) (Bollmeyer et al., 2015), focusing specifically on hourly ozone measurements from the years 1999 to 2018. For quantifying ozone pollution severity, we calculated the maximum daily 8-h mean (MDA8) concentration of ozone. All meteorological and air pollution data utilized in this study underwent a thorough normalization process to enhance model training speed and stability.

A comprehensive screening process resulted in the selection of 57 monitoring stations across Germany for our study. These stations were chosen for their consistent data quality, each presenting less than 5% missing ozone values annually. For handling missing ozone values, we employed the temporal nearest-neighbor interpolation method. This approach interpolates missing entries by substituting them with the closest available value within the same temporal series from the same monitoring station. The interpolation process was efficiently conducted using the 'Scipy' Python package (Virtanen et al., 2020). Fig. 3 illustrates the distribution of the selected observatories, showcasing a diverse range of environmental settings for comprehensive analysis. This includes three stations near Bremen in the northwest, eleven stations around Berlin in the northeast, and the remainder strategically located across southern Germany. This geographic diversity ensures a robust representation of varied atmospheric conditions in our study. Notably, all these stations are situated at altitudes below 1000 m. This uniformity in altitude helps minimize the influence of elevation on the ozone-meteorological relationship, ensuring a more focused analysis of other environmental factors.

Fig. 1 depicts the MDA8 ozone concentration distribution from 57 monitoring stations across Germany, spanning from 1999 to 2018. The data, organized in 10 $\mu g/m^3$ intervals, primarily peaks within the 50-80 $\mu g/m^3$ range, and shows a significant decrease in frequency for concentrations over 70 $\mu g/m^3$. Instances where ozone levels exceed 100 or 120 $\mu g/m^3$ are particularly rare. This infrequency of high ozone events presents a significant challenge for machine learning algorithms, as it limits the data available for accurately predicting these sporadic exceedance events.

2.2. Evaluation and validation

This section delineates the metrics and validation techniques employed to assess the efficacy of our machine learning models in accurately forecasting ozone concentrations, a crucial aspect with direct implications for public health and environmental monitoring.

Evaluation Metrics: To comprehensively gauge the model's accuracy in ozone prediction, we employ:

- Root Mean Squared Error (RMSE) (Hodson, 2022): This metric provides an average of the model's prediction errors, giving us a measure of the precision of ozone concentration forecasts.
- Coefficient of Determination (R^2) (Chicco et al., 2021): R^2 assesses the proportion of variance in observed ozone concentrations that is predictable from our model, indicating the model's overall fit.
- Prediction Accuracy (PA): Focusing on the public health impact of ozone, PA measures the model's ability to accurately predict days exceeding the critical ozone threshold of 120 $\mu g/m^3$, reflecting its capacity to flag potential health risk days.

Confusion Matrix Analysis (Caelen, 2017)**:** To further dissect the model's prediction capabilities, particularly in identifying exceedance days, we analyze:

- True Positives (TP) and True Negatives (TN): Reflecting correctly identified exceedance and non-exceedance days, respectively.
- False Positives (FP) and False Negatives (FN): Indicating instances of misclassification, with FP representing overestimation and FN underestimation of high ozone days.

Cross-Validation Strategies: We adopt three distinct cross-validation methods, each with a default fold number of 5, to ensure a thorough evaluation of our models:

- Traditional Cross-Validation (Browne, 2000): This approach involves randomly dividing our entire dataset into five subsets. In each validation cycle, or 'fold', four subsets are used to train the model, while the fifth subset is reserved for testing its performance. This method offers a well-rounded assessment of the model's accuracy and generalizability across different data segments.
- Year-Based Cross-Validation (Xue et al., 2019): Recognizing the importance of temporal dynamics in atmospheric data, we implement year-based cross-validation. Here, data is partitioned based on distinct calendar years, with each fold representing a different year or set of years. This strategy is crucial for evaluating the model's ability to perform consistently across varying atmospheric conditions that can change from year to year due to environmental and climatic shifts.
- Station-Based Cross-Validation (Xiao et al., 2018): Given the potential for varied environmental conditions at different monitoring stations, this method validates the model's performance by treating each station as a unique fold. This ensures the model is robust and adaptable across different geographic locations and atmospheric backgrounds. When the number of stations is less than five, we adapt by using a leave-one-out approach to maximize the use of available data for validation purposes.

These rigorous evaluation criteria and consistent cross-validation methodologies collectively provide a comprehensive assessment of our machine learning model's performance in long-term ozone forecasting, with a particular emphasis on high ozone concentration events crucial for public health assessment.

3. Machine learning methods

3.1. K-means clustering and Elbow method

To refine the predictive accuracy of our machine learning models, we employed the K-means clustering algorithm for segmenting air pollution observation sites across Germany. K-means is an unsupervised learning technique that categorizes data into a specified number of clusters (Ahmed et al., 2020). It does so by assigning each data point (in our case, observation sites) to the nearest cluster center, also known as the centroid, and then recalculating the centroid of each cluster. This process iteratively continues until the cluster assignments no longer change significantly, ensuring each site is grouped based on geographical proximity. In our experiments, to maximize the similarity in background factors (such as topography, climate, and emissions) among ozone monitoring stations within the same cluster, we chose the geographic coordinates (latitude and longitude) as the feature for clustering.

Determining the optimal number of clusters is crucial for the effectiveness of K-means clustering. We used the Elbow Method to identify this number (Liu and Deng, 2020). This method involves calculating the Within-Cluster Sum of Squares (WCSS), which is the sum of squared distances between each data point and its cluster centroid. By plotting the WCSS against different numbers of clusters, we looked for the 'elbow point' — the point where the rate of decrease sharply changes, indicating a diminishing return in explaining data variance with additional clusters. This point suggests a balance between the model's simplicity and the detailed representation of data, marking the optimal cluster count for our analysis.

The application of K-means clustering, guided by the Elbow Method, ensures that our machine learning models are informed by data sets that are both geographically and environmentally coherent. Such an approach enhances the models' relevance and accuracy in predicting atmospheric conditions specific to different regions.

3.2. Random forest methodology for atmospheric data analysis

The Random Forest (RF) algorithm represents a cornerstone of our machine learning approach, offering robustness and versatility in handling complex atmospheric datasets (Wang et al., 2021). As a method rooted in ensemble learning, RF integrates multiple decision trees to form a comprehensive model. This integration enhances the model's predictive accuracy and reduces the risk of overfitting, which is particularly valuable when dealing with intricate patterns in atmospheric data.

Each decision tree in the RF model is constructed using a random subset of the data and a random selection of features at each split. This randomness introduces diversity among the trees, leading to a more reliable aggregate prediction. When predicting ozone levels, the RF model considers various meteorological factors and produces an outcome that reflects the majority vote or the average prediction across all trees.

One of the key strengths of RF in atmospheric studies is its ability to capture non-linear relationships between variables, such as the complex interactions between different meteorological factors and ozone levels. This capability is crucial for accurate long-term forecasting in environmental science.

Furthermore, the RF model provides insights into feature importance, allowing us to identify which meteorological variables most significantly influence ozone concentrations. This aspect is particularly beneficial for understanding and interpreting the underlying drivers of atmospheric phenomena.

In our application of RF, we tune parameters such as the number of trees and the depth of each tree to optimize the model for our specific dataset. Through this careful calibration, we ensure that the RF model is well-suited to capture the nuances of ozone prediction in varied atmospheric conditions across Germany.

3.3. Enhancing ozone exceedance predictions with EEIA-RFM

The Enhanced Extreme Instance Augmentation for Random Forest Modelling (EEIA-RFM) is an approach we developed to tackle the prevalent challenge in atmospheric science: accurately predicting ozone exceedance days in dataset where high-concentration ozone data is sparse. This method is specifically engineered to boost the predictive performance of machine learning models in environmental datasets where extreme events are under-represented. The method unfolds in several strategic steps:

1. **Data Partitioning:** The dataset is split into a training set of size N and a test set for evaluation.
2. **Extreme Data Isolation:** From the training set, extreme data instances, defined as those with ozone levels exceeding 120 $\mu g/m^3$, are identified and isolated. These instances form a subset of size n.
3. **Model Training on Extreme Data:** The Random Forest model is trained on this extreme data subset (n).
4. **Synthetic Data Generation:** The model predicts new extreme instances using randomly selected meteorological inputs from the extreme subset. This step generates $K \times n$ synthetic extreme instances, where K is a factor determined by cross-validation to achieve data balance.
5. **Dataset Pooling and Balancing:** In this step, the $K \times n$ newly generated exceedance instances from step 4 are combined with a randomly selected subset of the non-exceedance data from the original training dataset (N). Specifically, we randomly select $(N - K \times n)$ non-exceedance instances for inclusion. This approach ensures that the enhanced dataset remains balanced and maintains its original size of N. The final dataset for training the model thus comprises the $K \times n$ newly generated exceedance instances and a representative subset of the original non-exceedance data, providing a balanced and comprehensive set for effective model training.
6. **Model Training on Balanced Dataset:** The newly balanced dataset, resulting from the integration of $K \times n$ newly generated exceedance instances with the selected non-exceedance data, is employed to train a Random Forest model. This dataset now features an increased proportion of extreme data by a factor of K, enhancing the model's focus on predicting ozone exceedances. The training on this enriched dataset is expected to improve the model's accuracy in forecasting high-concentration ozone events.
7. **Performance Comparison:** The model's efficacy is assessed by comparing its prediction accuracy with that of a model trained on the original, unbalanced dataset.

The EEIA-RFM method, by specifically focusing on enhancing the model's exposure to high ozone events, aims to bridge a crucial gap in atmospheric data analysis. Its synthetic data generation and balancing steps are key innovations that enable the model to learn from a more representative sample of ozone exceedance scenarios. This approach not only promises to increase the accuracy in forecasting critical exceedance events but also offers insights into the complex dynamics of ozone behavior under extreme conditions. As such, EEIA-RFM has the potential to be a general tool for dealing with forecasts of rare events in environmental modeling, providing a more reliable tool for public health advisories and environmental policy-making in the face of atmospheric pollution challenges.

4. Experiments and results

4.1. Cluster result

In our pursuit of developing robust machine learning models for ozone prediction, we initiated our analysis by segmenting air pollution

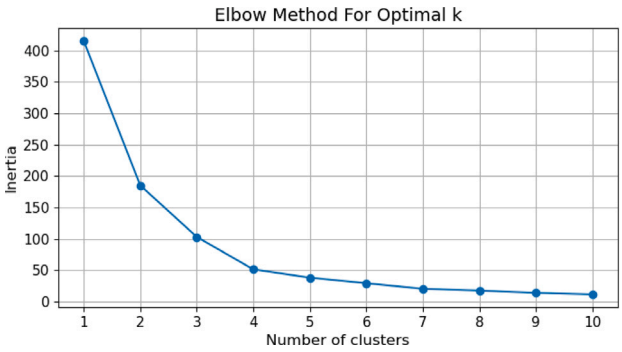


Fig. 2. Optimal cluster determination using the Elbow Method.

observation sites across Germany into distinct clusters. This preliminary step is essential to mitigate the diverse influences of site-specific characteristics on the model's predictive performance. Primarily leveraging geographical location data (latitude and longitude), due to the absence of comprehensive emissions and demographic details, we employed clustering as a foundational strategy to categorize these sites. Here we used K-means algorithm for clustering.

In determining the appropriate number of clusters for the k-means algorithm, the Elbow Method was applied. By plotting the Within-Cluster Sum of Squares (WCSS) against various numbers of clusters, we observed a distinct 'elbow' in the graph. This elbow point, depicted in Fig. 2, was most apparent at the 4-cluster mark. This point is significant as it represents where the addition of more clusters ceases to provide substantial benefits in reducing within-cluster variance, suggesting a balance between model simplicity and explanatory power. Consequently, we chose four clusters to capture the diverse environmental conditions across Germany more effectively, ensuring a nuanced approach in our subsequent modeling.

In Fig. 3, the geographical distribution of air quality monitoring sites across Germany is segmented into four distinct k-means clusters, each representing varied environmental and meteorological conditions. Clusters 2 and 3 are predominantly urban, characterized by high population densities in their respective regions. Cluster 3, situated in northwestern Germany, includes a smaller collection of only three urban stations around Bremen. The coastal proximity of Bremen contributes to a comparatively cleaner environment in this cluster. In contrast, Cluster 2 in northeastern Germany comprises 11 stations, with seven situated around Berlin, representing a mix of urban, suburban, and rural backgrounds. As indicated in Section 4.2 of our study, the diversity in station types does not significantly impact the performance of machine learning models in simulating ozone concentrations. Meanwhile, Clusters 0 and 1, located in the southeast and southwest of Germany respectively, exhibit more dispersed station distributions. Cluster 0 contains 13 stations, some of which are situated in mountainous areas, typically associated with lower emission sources. This contrasts with its urban stations, including three in Munchen. Cluster 1, the largest cluster with 30 stations, presents a complex environmental background. This cluster includes stations in industrial cities with significant emissions from manufacturing and automotive sectors, as well as areas within the Black Forest, indicative of a wide range of emission sources and environmental conditions.

Table 1 presents ozone statistics collected in Germany from 1999 to 2018, further highlighting the regional air quality differences. Cluster 3 (Bremen area) exhibits the lowest average ozone concentration at 58.26 $\mu g/m^3$, along with the fewest exceedance days at 256.67 days, attributed to its cleaner environment. In contrast, Cluster 1 in Southwest Germany records the highest mean ozone level at 66.90 $\mu g/m^3$, accompanied by a substantial 562.13 exceedance days, reflecting its industrial backdrop. This disparity amplifies the need for region-specific machine learning models in ozone prediction, considering the unique

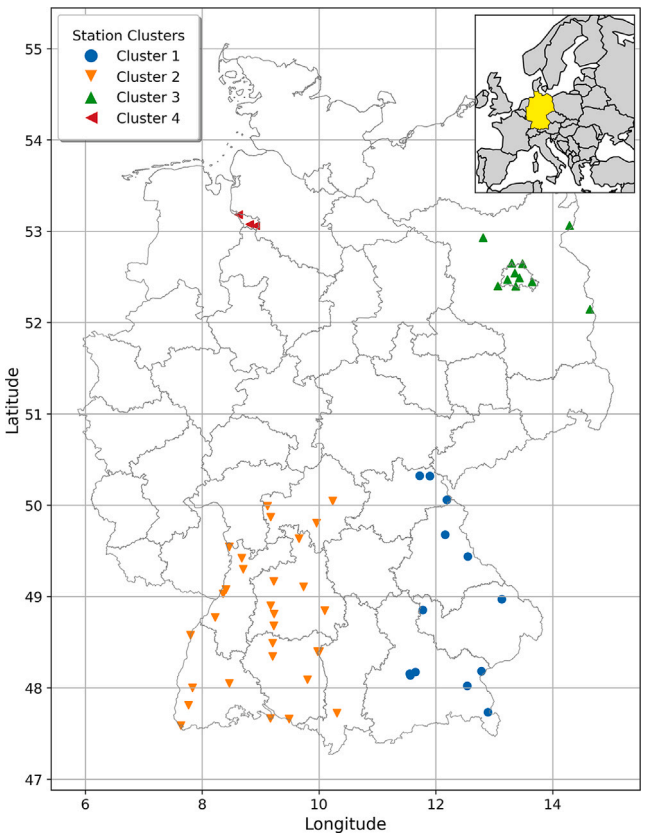


Fig. 3. Geographical distribution of air quality monitoring stations in Germany: K-means clustering results — inset map of Europe in the top right corner highlighting Germany in yellow.

Table 1
German ozone statistics (1999–2018) by clusters.

Cluster	Mean	Std	Max	Number of exceedance days (>120 $\mu g/m^3$) per Station	Proportion of exceedance days over 20 years
Cluster 0	64.24	31.04	183.39	343.69	4.71%
Cluster 1	66.90	34.93	208.67	562.13	7.70%
Cluster 2	65.32	31.38	190.79	378.64	5.19%
Cluster 3	58.26	30.15	206.20	256.67	3.52%

environmental and industrial characteristics of each cluster. Moreover, the data reveals that over the two-decade span, the incidence of ozone exceedances in all clusters remains below 10%, with Cluster 3 experiencing a mere 3.52%. This low frequency highlights the scarcity of ozone exceedance days and underscores the importance of methods like extreme data enhancement in ozone forecasting.

4.2. Comparative analysis of machine learning models in ozone prediction

In this section, we embark on a comparative analysis of machine learning models extensively utilized in ozone research. Our approach involves training each model on data from the four distinct clusters identified previously, employing a 5-fold cross-validation method for evaluation. Given the nature of our dataset, we specifically exclude Recurrent Neural Networks (RNN) (Chang et al., 2020) due to the absence of a time series prediction component in our experiment. Likewise, Convolutional Neural Networks (CNN) (Eslami et al., 2020) are omitted, considering the relatively low dimensionality of our input features. The core objective is to assess the efficacy of selected machine learning models in simulating MDA8 ozone concentrations based on meteorological factors. For this purpose, we focus on three models

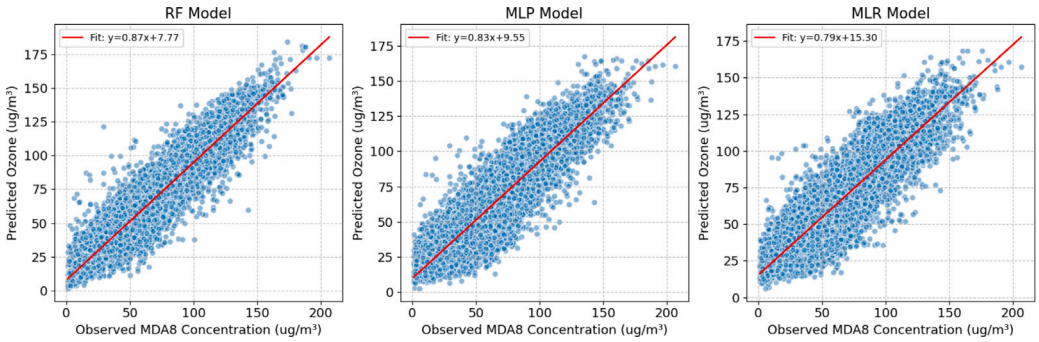


Fig. 4. Comparative scatter plot: Machine learning model predictions in Cluster 2, with best fit line shown in red.

Table 2
Comparison of machine learning model performance by clusters.

Model	Cluster 0			Cluster 1			Cluster 2			Cluster 3		
	RMSE	R2	PA[a]	RMSE	R2	PA	RMSE	R2	PA	RMSE	R2	PA
RF	15.17	0.77	53%	15.06	0.79	54%	11.5	0.87	63%	11.26	0.86	65%
MLP	15.29	0.79	53%	17.16	0.75	39%	14.39	0.77	45%	13.87	0.75	49%
MLR	17.98	0.73	32%	18.91	0.69	27%	16.93	0.71	23%	16.74	0.68	18%

[a] Prediction accuracy of ozone exceedances with 120 μg/m^3 threshold.

that have demonstrated utility in atmospheric studies: Multiple Linear Regression (Han et al., 2020), Random Forest (Weng et al., 2022), and Multilayer Perceptron (Chattopadhyay et al., 2019).

In the modeling process, we utilized seven meteorological features (TG, TX, RR, PP, HU, FG and QQ), as delineated in Section 2.1, to serve as input variables. Preliminary attempts to integrate annual average emissions data for NO_X, $NMVOC$, and CH_4 were made; however, this data was ultimately excluded from the final model inputs due to its negligible impact on daily prediction accuracy. The output of these models is the Maximum Daily 8-h Average (MDA8) ozone concentration, directly addressing our objective to enhance the accuracy of ozone exceedance forecasts. Three distinct machine learning models were independently trained on the dataset from the four identified clusters. The results of these individual training are consolidated in Table 2, which demonstrates that the Random Forest (RF) model distinctly outshines the others in performance metrics across all clusters. It records the highest R-squared values, with a peak of 0.87 in Cluster 2 and 0.86 in Cluster 3, and it also exhibits the lowest RMSE, particularly notable at 11.26 in Cluster 3. The clusters with geographically proximate stations, especially Cluster 3, where all stations are situated around Bremen, show improved model performance. This suggests that the closer station locations within these clusters, which likely experience more similar meteorological conditions, contribute to a more accurate modeling outcome. In contrast, Clusters 0 and 1, which have a wider spread of stations over larger regions, exhibit more variable results, indicating that the geographical dispersion of stations may introduce additional complexities to the model's predictive ability. The Multilayer Perceptron (MLP) and Multiple Linear Regression (MLR) models demonstrate predictive potential but lag behind the Random Forest (RF), with the MLP edging out the MLR on RMSE and R2. However, all models show limited Prediction Accuracy for exceedance days. Fig. 4 displays scatter plots from diverse machine learning models for Cluster 2, centered around Berlin, and serves as an illustrative case study. These plots underscore the Random Forest model's heightened precision, notably its alignment with the observed ozone values.

The different performance between the machine learning models can be attributed to the non-linear nature of the ozone-meteorological relationship. RF's ensemble approach captures this non-linearity, as evidenced by its consistent performance, particularly in clusters with closer station proximity. MLR's linear assumptions fail to account for this complexity, resulting in lower performance metrics. MLP's potential is hampered by the absence of broader environmental data, such as emissions and land use, making its training more challenging compared to RF. Given RF's robust predictions with the available data, it has been selected for further experimentation.

Table 3
Comparison of Random Forest model by different CV methods.

Cluster	Year-CV			Station-CV			Sample-CV		
	RMSE	R2	PA	RMSE	R2	PA	RMSE	R2	PA
Cluster 0	16.52	0.76	50%	15.21	0.81	58%	15.17	0.77	53%
Cluster 1	18.15	0.74	52%	16.15	0.77	51%	15.06	0.79	54%
Cluster 2	15.50	0.73	45%	11.15	0.86	65%	11.50	0.87	63%
Cluster 3	15.21	0.75	51%	11.17	0.87	60%	11.26	0.86	65%

The ensemble nature of RF, which mitigates overfitting and adeptly handles environmental data complexity, makes it well-suited for in-depth analysis using both station-based and year-based cross-validation methods. The results from Table 3 suggest that the Random Forest (RF) model's performance is influenced by the cross-validation (CV) method employed, which varies depending on the cluster. Specifically, Station-CV demonstrates enhanced performance, with Clusters 2 and 3 achieving RMSE scores of 11.15 and 11.17, respectively, alongside high R2 values of 0.86 and 0.87. The enhanced performance observed in Station-CV, especially within Clusters 2 and 3, is likely a consequence of the similarity in meteorological and emission conditions among these clusters. This can be associated with the stations being geographically close to each other, implying that they share similar environmental conditions. Conversely, in the two clusters located in southern Germany, the stations within the same cluster are more widely distributed, leading to variations in background conditions. This greater dispersion contributes to larger simulation errors (RMSE) in these clusters.

On the other hand, Year-CV exhibits weaker performance across all clusters, which could be due to the absence of certain dynamic factors in the model's input features. The lack of emission data, along with other influential factors such as population density and traffic patterns, which can change significantly over time, likely hampers the model's ability to account for annual variations in ozone levels. Since emission factors are particularly susceptible to yearly fluctuations due to policy changes, economic growth, and technological advancements, their omission might be particularly impactful on the Year-CV's effectiveness.

In summary, the superior performance of Station-CV reinforces the importance of spatial homogeneity provided by clustering when training machine learning models for environmental applications. The

Table 4
Comparison of RF and EEIA-RF model performance by clusters.

Cluster	RF			EEIA-RF		
	RMSE	R2	PA	RMSE	R2	PA
Cluster 0	15.17	0.77	53%	14.24	0.81	78%
Cluster 1	15.06	0.79	54%	15.97	0.78	69%
Cluster 2	11.50	0.87	63%	11.27	0.88	76%
Cluster 3	11.26	0.86	65%	11.59	0.86	73%

underperformance of Year-CV underscores the need for incorporating time-varying predictors to better capture the annual variability of ozone concentrations, especially in regions like Germany where emissions and other anthropogenic factors can change significantly from year to year.

4.3. Enhanced extreme instance augmentation for random forest modelling

In practical scenarios, accurately forecasting the number of days with ozone exceedances is crucial for assessing ozone pollution. However, the infrequent occurrence of high-concentration ozone days within large datasets presents a challenge for machine learning models. Referring to Table 1, even the cluster with the most exceedance days over a 20-year period—Cluster 1—only accounts for 7.7% of the data, with 562.13 days. Conversely, Cluster 3, with the fewest exceedance days at 256.67, represents a mere 3.52% of the total observations. This imbalance, characterized by a scarcity of high-concentration data points, impedes the ability of data-driven machine learning models to forecast ozone exceedances with high accuracy.

In our experimental assessment, the Enhanced Extreme Instance Augmentation Random Forest (EEIA-RF) method was rigorously tested across each of the four clusters. For each cluster, we partitioned the data into distinct training and test sets. The model training and EEIA processes were exclusively applied to the training set, ensuring that the test set remained independent for unbiased evaluation. This approach provided a focused and tailored assessment for each cluster, enabling a precise evaluation of the method's efficacy. To determine the factor K, indicative of the increased proportion of exceedance data in the EEIA method, we tested values from 2 to 10 through cross-validation. This range was selected to comprehensively assess the method's performance across various levels of data augmentation. The optimal value of K was chosen based on the best model performance. We employed both traditional Random Forest (RF) and EEIA-RF models to elucidate the impact of our augmentation strategy on prediction performance. The assessments in this section were consistently evaluated using traditional (sample-based) cross-validation methods. The results are summarized in Table 4, highlighting the outcomes of this comparative study.

The EEIA-RF approach yielded a pronounced enhancement in Prediction Accuracy (PA) for ozone exceedance days across all clusters. For instance, in Cluster 0, the application of EEIA-RF not only increased the Prediction Accuracy (PA) from 53% (with standard RF) to 78%, but it also led to a reduction in the Root Mean Squared Error (RMSE) and an increase in the R-squared values, indicating enhanced model performance and fit. This pattern of increased PA with negligible changes in RMSE and R2 was consistent in Clusters 2 and 3, with PA improvements of 13% and 8%, respectively. Although Cluster 1 experienced a marginal decrease in R-squared and a small increase in RMSE, the PA still rose substantially from 54% to 69%. The consistent improvements in Prediction Accuracy across all clusters, especially in identifying critical ozone exceedance days, underscore the effectiveness of the EEIA-RF method in enhancing ozone concentration forecasts, a vital factor for atmospheric pollution assessment and public health protection.

In our continued analysis with Cluster 2 as a case study, we examine the scatter plots depicting the prediction results for both the RF and EEIA-RF models. As demonstrated in Fig. 5, these scatter plots, particularly the fitted lines within them, reveal a notable similarity in the predictive behavior of the two Random Forest methodologies. These results underscore the EEIA-RF method's efficacy in addressing the data imbalance challenge, particularly in enhancing the model's predictive power for critical exceedance days. Despite some trade-offs, the notable gains in PA confirm that the augmentation method provides a significant advantage in predicting rare but critical high-concentration ozone events, which is of utmost importance for environmental health monitoring and policy-making.

The efficacy of the Enhanced Extreme Instance Augmentation Random Forest (EEIA-RF) method in improving ozone exceedance predictions is rooted in its approach to data balancing and model integration. By augmenting the training dataset with synthetically generated high-concentration ozone instances, the EEIA-RF method corrects for the underrepresentation of such events, leading to a more balanced dataset. This enhancement enables the Random Forest algorithm to better capture the occurrence of extreme ozone levels, leveraging the model's ensemble learning capabilities to discern the complex patterns associated with these events.

In Fig. 6, the box and whisker plots compare the observational data and model predictions within Cluster 2. The general underestimation of ozone concentrations by both the RF and EEIA-RF models is evidenced by the lower quartiles and medians. This trend is pronounced when analyzing the complete test dataset. In contrast, during exceedance days (ozone concentrations > 120 $\mu g/m^3$), the predictive distribution aligns more closely with observations, particularly for the EEIA-RF model which achieves a median prediction (129.69 $\mu g/m^3$) closely approximating the observed median (132.44 $\mu g/m^3$). This comparison highlights the improved predictive capability of the EEIA method for critical exceedance forecasting. It is important to note that both models exhibit limitations in predicting ozone levels above 160 $\mu g/m^3$, resulting in a predictive distribution that appears lower than the observed distribution for exceedance days. Nonetheless, our primary concern aligns with public health objectives—accurately predicting the frequency of days exceeding the 120 $\mu g/m^3$ ozone threshold. This is the level beyond which adverse health effects become a significant concern.

Fig. 7 delineates the outcomes for the Random Forest (RF) and Enhanced Extreme Instance Augmentation Random Forest (EEIA-RF) models, with the latter showing a significant decrease in missed high ozone events from 329 to 199 and an increase in accurately identified exceedance days from 526 to 656. This improvement is crucial for the accurate prediction of ozone pollution, which is vital for public health guidance and environmental policy.

Despite these advances, the EEIA-RF model may incline towards overestimating the number of exceedance days due to its data augmentation technique, which could amplify the frequency of predicted high ozone levels. For instance, an increase in false alerts from 114 to 161 suggests a possible overestimation bias. This conservative stance, while potentially overstating the number of exceedance days, remains a critical area for refinement to ensure the model mirrors true atmospheric behavior.

To avoid such overestimation, it is essential to calibrate the synthetic data generation carefully, ensuring that the augmented instances are representative of the true distribution of extreme events. Moreover, the performance of the EEIA-RF model must be rigorously validated against an independent test set to ensure that any improvements in predictive accuracy are due to genuine learning and not an artifact of artificial data inflation. By continuously evaluating and adjusting the proportion of synthetic data in the training process, the EEIA-RF method can maintain the delicate balance necessary for accurate and reliable exceedance forecasting.

During the training process of the Random Forest model, we were able to ascertain the feature importance of each meteorological variable on ozone concentration predictions. Despite the variations in geography and emissions across clusters, the importance ranking of meteorological features remained consistent. This uniformity in the importance ranking across different geographical areas underscores the

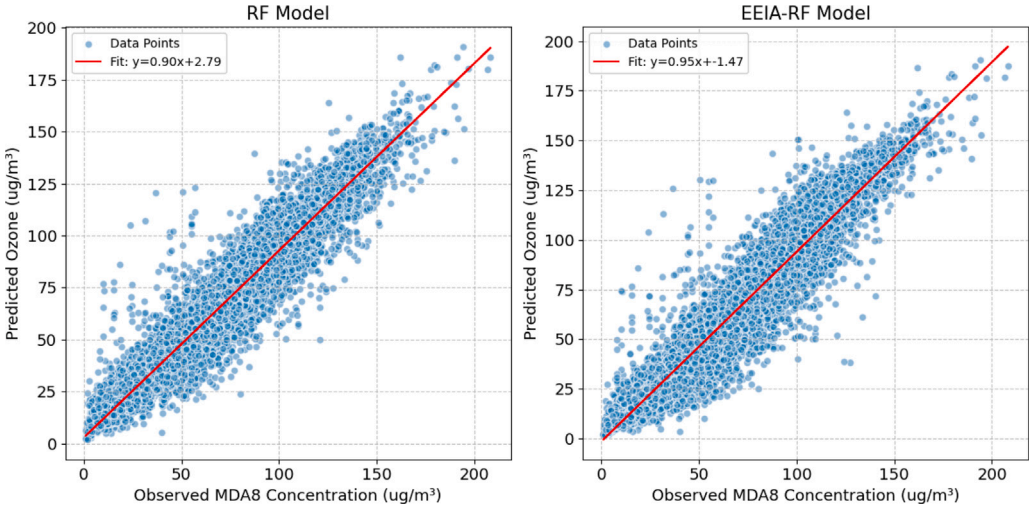


Fig. 5. Scatter plots of RF and EEIA-RF Models for MDA8 ozone forecasting in Cluster 2, with best fit line shown in red.

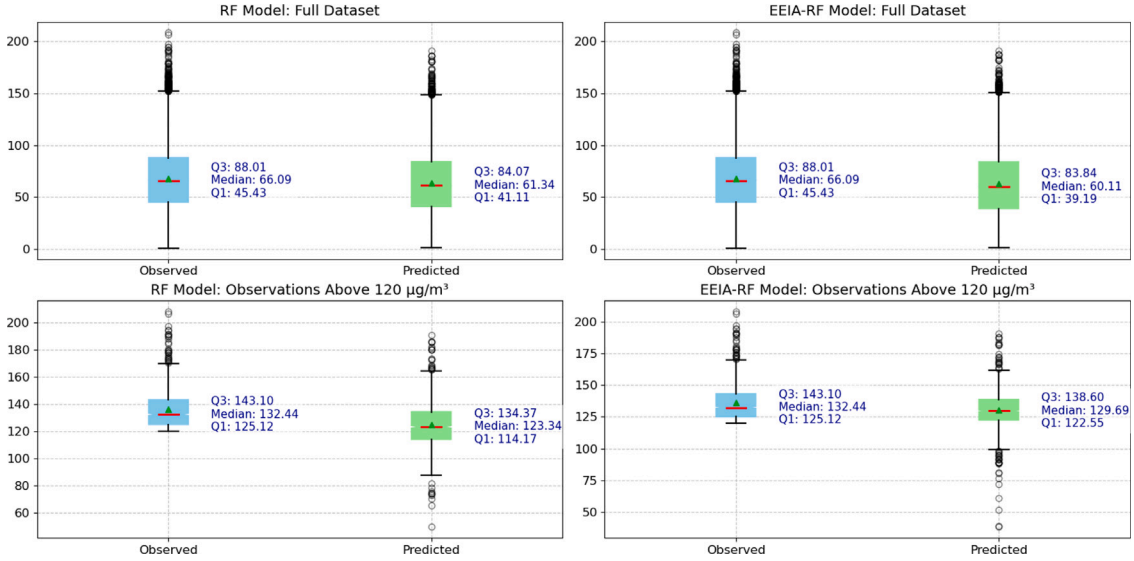


Fig. 6. Box and whisker plot of RF and EEIA-RF model predictions for MDA8 ozone in Cluster 2: The upper subfigures represent the complete test dataset, while the lower subfigures focus on instances with observed values exceeding the 120 μg/m^3 ozone threshold. Red lines indicate mean values, and triangles denote medians, with quartile values detailed within the plots.

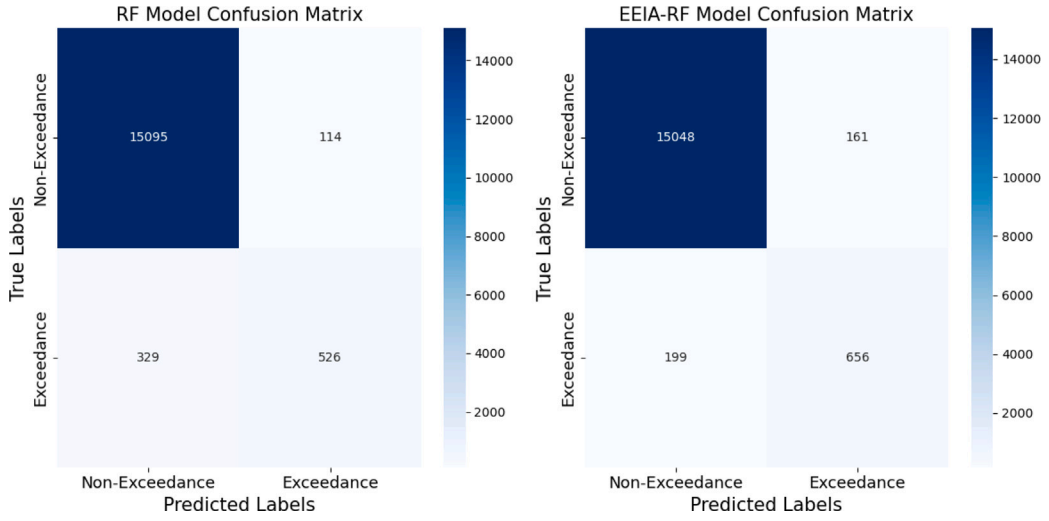


Fig. 7. Comparative confusion matrices for RF and EEIA-RF models in Cluster 2 (2015–2018 validation period).

Table 5
Comparison of confusion matrices for different methods in cluster 2.

Method	Confusion matrix (Unnormalized)				Confusion matrix (Normalized)			
	TN	FP	FN	TP	TN %	FP %	FN %	TP %
Ori-RF	15 001	84	412	574	99.44%	0.56%	41.79%	58.22%
EEIA	14 884	201	196	790	98.67%	1.33%	19.88%	80.12%
SMOTE	14 893	192	233	753	98.73%	1.27%	23.63%	76.37%
Cost-sensitive (Weighted Loss)	14 961	124	348	638	99.18%	0.82%	35.29%	64.71%
Error tolerance increment	14 870	215	239	747	98.57%	1.43%	24.24%	75.76%

significant potential impact of incorporating omitted variables, such as emissions data and geographic background. Notably, solar radiation, temperature, and humidity emerged as the most influential predictors. With the integration of the EEIA method, temperature's influence increased, reflecting its role in enhancing photochemical reactions under high ozone scenarios, a finding that aligns with existing literature (Li et al., 2023). The omission of emissions data in our model precludes a comprehensive simulation of the complex interplay of factors affecting ozone formation, thus a deeper exploration of meteorological elements' impact on ozone concentrations was not within the scope of this study.

To comprehensively evaluate the efficacy of the Enhanced Extreme Instance Augmentation (EEIA) approach, our study involves a comparative analysis with both established and novel methods for addressing data imbalance in atmospheric studies. We focus on Cluster 2 as a case study to juxtapose the EEIA method with the Synthetic Minority Over-sampling Technique (SMOTE) (Gong and Ordieres-Meré, 2016) and Cost-Sensitive (Weighted-Loss) methods (Tsai et al., 2009). Additionally, due to similar results observed between threshold moving and error tolerance increment methods, we present only the latter for clarity and conciseness (Vicente et al., 2024). Each method employed is based on a Random Forest model, with hyperparameters optimized through cross-validation.

Our findings are detailed in Table 5, featuring both unnormalized and normalized confusion matrices. The analysis primarily focuses on two crucial metrics: the True Positive (TP) rate, reflecting the accurate prediction of ozone exceedance days, and the False Positive (FP) rate, indicative of mistakenly predicted non-exceedance days as exceedances. The EEIA-RF model demonstrates a superior TP rate of 80.12%, surpassing SMOTE's 76.37%, Cost-Sensitive's 64.71%, and Error Tolerance Increment's 75.76%. In contrast, the EEIA-RF method exhibits an FP rate of 1.33%, which, while lower than the Error Tolerance Increment method (1.43%), is marginally higher than SMOTE (1.27%) and Cost-Sensitive (0.82%). The EEIA-RF model predicts 991 exceedance days (FP + TP), closely aligning with the observed 986 days (FN + TP), thereby highlighting its effectiveness in providing accurate statistical representations for ozone exceedance analysis. This alignment affirms EEIA-RF's potential as a powerful tool in ozone pollution research.

5. Conclusion

This research focused on utilizing machine learning techniques to predict ozone levels across Germany, leveraging data spanning from 1999 to 2018 obtained from 57 monitoring stations. Central to our study was the application of the Random Forest (RF) model and our novel Enhanced Extreme Instance Augmentation for Random Forest Modelling (EEIA-RFM) approach, tailored specifically for the accurate prediction of high ozone concentration events, which are critical for public health considerations.

The initial phase of our study involved employing K-means clustering and the Elbow Method to segment the extensive dataset, ensuring that the models were attuned to the distinct environmental characteristics of various regions in Germany. This pre-processing step was crucial in enhancing the geographical relevance of our models.

The RF model demonstrated robustness in capturing the intricate, non-linear relationships characteristic of atmospheric data. Our major advancement, however, was the introduction of EEIA-RFM. This innovative approach addressed the under-representation of high-concentration ozone events in environmental datasets. By generating and integrating synthetic extreme ozone event data, the EEIA-RFM method significantly improved the model's accuracy in predicting exceedance days.

In summary, this study presents a significant step forward in the application of machine learning for atmospheric data analysis. By enhancing the predictive accuracy of ozone exceedance events, our research supports the development of more reliable forecasting models. These models can be useful tools for informing public health advisories and shaping environmental policies in Germany, given the critical nature of ozone levels in atmospheric science.

While our model advances prediction accuracy for ozone exceedances, it encounters specific challenges. Notably, there is a tendency to overestimate the number of days when the MDA8 ozone concentration exceeds the 120 $\mu g/m^3$ threshold, evidenced by an increase in false alerts. Moreover, the model's current reliance on a limited array of meteorological inputs, to the exclusion of emissions data and geographic background, constrains its predictive capability. The use of annual average emission data for Germany proves insufficient for our daily predictive needs. To address this issue, we aim to incorporate daily regional emissions data and geographical factors in future model versions, dependent on the availability of such detailed datasets. Furthermore, it could be interesting to split the study period into segments of a decade (e.g.,1999–2008, 2009–2018) to explore the impact of environmental policies on ozone levels in the last decades. This approach can help analyze the temporal trends in ozone levels, offering insights and guidance for more effective future strategies.

Software and data availability

All methods were written in the Python language and in the Visual Studio Code environment (https://code.visualstudio.com/). Machine learning methods were developed using the freely available scikit-learn library (Pedregosa et al., 2011). The source code of Enhanced Extreme Instance Augmentation for Random Forest Modelling is available on Github at https://github.com/td5060/EEIA-RFM. The monitoring air pollution data used in this study was accessed through German Environment Agency (UBA) and is available at https://eeadmz1-cws-wp-air02.azurewebsites.net/index.php/users-corner/download-e1a-from-2013/.

CRediT authorship contribution statement

Tuo Deng: Writing – review & editing, Writing – original draft, Methodology, Investigation. **Astrid Manders:** Supervision, Resources, Conceptualization. **Arjo Segers:** Writing – review & editing, Formal analysis. **Arnold Willem Heemink:** Writing – review & editing, Supervision. **Hai Xiang Lin:** Writing – review & editing, Supervision, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We extend our gratitude to the German Environment Agency (UBA) for providing the air pollution data essential for this study. This data played a crucial role in training the machine learning models utilized in our case study. Special thanks to the China Scholarship Council for their financial support, which greatly assisted the research efforts of the author Tuo Deng. Additionally, we acknowledge the support received from the Timman foundation, which partially funded the work of Hai Xiang Lin. Their contributions were invaluable in the advancement of this research.

References

Ahmed, M., Seraj, R., Islam, S.M.S., 2020. The k-means algorithm: A comprehensive survey and performance evaluation. Electronics 9 (8), 1295.

Bollmeyer, C., Keller, J., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., et al., 2015. Towards a high-resolution regional reanalysis for the European CORDEX domain. Q. J. R. Meteorol. Soc. 141 (686), 1–15.

Browne, M.W., 2000. Cross-validation methods. J. Math. Psychol. 44 (1), 108–132.

Caelen, O., 2017. A Bayesian interpretation of the confusion matrix. Ann. Math. Artif. Intell. 81 (3–4), 429–450.

Chang, Y.-S., Chiao, H.-T., Abimannan, S., Huang, Y.-P., Tsai, Y.-T., Lin, K.-M., 2020. An LSTM-based aggregated model for air pollution forecasting. Atmosp. Pollut. Res. 11 (8), 1451–1463.

Chao, X., Zhang, L., 2023. Few-shot imbalanced classification based on data augmentation. Multimedia Syst. 29 (5), 2843–2851.

Chattopadhyay, G., Midya, S.K., Chattopadhyay, S., 2019. MLP based predictive model for surface ozone concentration over an urban area in the Gangetic West Bengal during pre-monsoon season. J. Atmos. Sol.-Terr. Phys. 184, 57–62.

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, e623.

Cornes, R.C., van der Schrier, G., van den Besselaar, E.J., Jones, P.D., 2018. An ensemble version of the E-OBS temperature and precipitation data sets. J. Geophys. Res.: Atmos. 123 (17), 9391–9409.

Emberson, L., 2020. Effects of ozone on agriculture, forests and grasslands. Phil. Trans. R. Soc. A 378 (2183), 20190327.

Eslami, E., Choi, Y., Lops, Y., Sayeed, A., 2020. A real-time hourly ozone prediction system using deep convolutional neural network. Neural Comput. Appl. 32, 8783–8797.

Fan, K., Dhammapala, R., Harrington, K., Lamastro, R., Lamb, B., Lee, Y., 2022. Development of a machine learning approach for local-scale ozone forecasting: Application to Kennewick, WA. Front. Big Data 5, 781309.

Fang, X., Xiao, H., Sun, H., Liu, C., Zhang, Z., Xie, Y., Liang, Y., Wang, F., 2020. Characteristics of ground-level ozone from 2015 to 2018 in BTH Area, China. Atmosphere 11 (2), 130.

Feng, R., Zheng, H.-j., Gao, H., Zhang, A.-r., Huang, C., Zhang, J.-x., Luo, K., Fan, J.-r., 2019. Recurrent neural network and random forest for analysis and accurate forecast of atmospheric pollutants: a case study in Hangzhou, China. J. Clean. Prod. 231, 1005–1015.

Finlayson-Pitts, B., Pitts, Jr., J., 1993. Atmospheric chemistry of tropospheric ozone formation: scientific and regulatory implications. Air Waste 43 (8), 1091–1100.

Gong, B., Ordieres-Meré, J., 2016. Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: Case study of Hong Kong. Environ. Model. Softw. 84, 290–303.

Han, H., Liu, J., Shu, L., Wang, T., Yuan, H., 2020. Local and synoptic meteorological influences on daily variability in summertime surface ozone in eastern China. Atmos. Chem. Phys. 20 (1), 203–222.

Hjellbrekke, A.-G., Solberg, S., 2022. Ozone Measurements 2020. NILU.

Hodson, T.O., 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geosci. Model Dev. 15 (14), 5481–5487.

Li, H., Yang, Y., Jin, J., Wang, H., Li, K., Wang, P., Liao, H., 2023. Climate-driven deterioration of future ozone pollution in Asia predicted by machine learning with multi-source data. Atmos. Chem. Phys. 23 (2), 1131–1145.

Li, M., Yu, S., Chen, X., Li, Z., Zhang, Y., Wang, L., Liu, W., Li, P., Lichtfouse, E., Rosenfeld, D., et al., 2021. Large scale control of surface ozone by relative humidity observed during warm seasons in China. Environ. Chem. Lett. 19, 3981–3989.

Liu, F., Deng, Y., 2020. Determine the number of unknown targets in open world based on elbow method. IEEE Trans. Fuzzy Syst. 29 (5), 986–995.

Liu, Y., Wang, T., 2020. Worsening urban ozone pollution in China from 2013 to 2017–Part 1: The complex and varying roles of meteorology. Atmos. Chem. Phys. 20 (11), 6305–6321.

Lu, H., Lyu, X., Cheng, H., Ling, Z., Guo, H., 2019. Overview on the spatial–temporal characteristics of the ozone formation regime in China. Environ. Sci.: Process. Impacts 21 (6), 916–929.

Manders, A., Van Meijgaard, E., Mues, A., Kranenburg, R., Van Ulft, L., Schaap, M., 2012. The impact of differences in large-scale circulation output from climate models on the regional modeling of ozone and PM. Atmos. Chem. Phys. 12 (20), 9441–9458.

Ojha, N., Girach, I., Sharma, K., Sharma, A., Singh, N., Gunthe, S.S., 2021. Exploring the potential of machine learning for simulations of urban ozone variability. Sci. Rep. 11 (1), 22513.

Otero, N., Sillmann, J., Mar, K.A., Rust, H.W., Solberg, S., Andersson, C., Engardt, M., Bergström, R., Bessagnet, B., Colette, A., et al., 2018. A multi-model comparison of meteorological drivers of surface ozone over Europe. Atmos. Chem. Phys. 18 (16), 12269–12288.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Ryu, Y.-H., Hodzic, A., Descombes, G., Hu, M., Barré, J., 2019. Toward a better regional ozone forecast over CONUS using rapid data assimilation of clouds and meteorology in WRF-Chem. J. Geophys. Res.: Atmos. 124 (23), 13576–13592.

Sicard, P., Khaniabadi, Y.O., Perez, S., Gualtieri, M., De Marco, A., 2019. Effect of o 3, PM 10 and PM 2.5 on cardiovascular and respiratory diseases in cities of France, Iran and Italy. Environ. Sci. Pollut. Res. 26, 32645–32665.

Travis, K.R., Jacob, D.J., 2019. Systematic bias in evaluating chemical transport models with maximum daily 8 h average (MDA8) surface ozone for air quality applications: a case study with GEOS-Chem v9. 02. Geosci. Model Dev. 12 (8), 3641–3648.

Tsai, C.-h., Chang, L.-c., Chiang, H.-c., 2009. Forecasting of ozone episode days by cost-sensitive neural network methods. Sci. Total Environ. 407 (6), 2124–2135.

Vicente, D., Salazar, F., López-Chacón, S., Soriano, C., Martin-Vide, J., 2024. Evaluation of different machine learning approaches for predicting high concentration episodes of ground-level ozone: A case study in Catalonia, Spain. Atmosp. Pollut. Res. 15 (3), 101999.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat. Methods 17, 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2.

Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q., Zhang, H., 2021. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environ. Res. 202, 111660.

Weng, X., Forster, G., Nowack, P., 2022. A machine learning approach to quantify meteorological drivers of ozone pollution in China from 2015 to 2019. Atmos. Chem. Phys. 22 (12), 8385–8402.

Xiao, Q., Chang, H.H., Geng, G., Liu, Y., 2018. An ensemble machine-learning model to predict historical PM2. 5 concentrations in China from satellite data. Environ. Sci. Technol. 52 (22), 13260–13269.

Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., Zhang, Q., 2019. Spatiotemporal continuous estimates of PM2. 5 concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations. Environ. Int. 123, 345–357.

Zhang, A., Fu, T.-M., Feng, X., Guo, J., Liu, C., Chen, J., Mo, J., Zhang, X., Wang, X., Wu, W., et al., 2023. Deep learning-based ensemble forecasts and predictability assessments for surface ozone pollution. Geophys. Res. Lett. 50 (8), e2022GL102611.

Zhang, J., Wei, Y., Fang, Z., 2019. Ozone pollution: a major health hazard worldwide. Front. Immunol. 10, 2518.