ACS ES&T Air

pubs.acs.org/estair

Article

An Ensemble Machine Learning Approach for Predicting Sources of Organic Aerosols Measured by Aerosol Mass Spectrometry

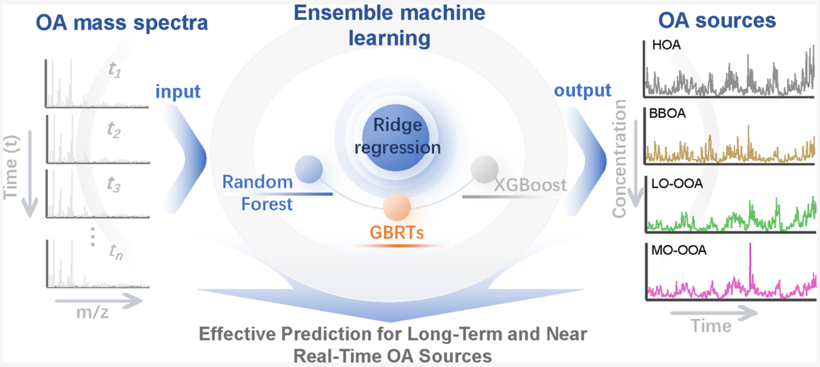
Published as part of ACS ES&T Air special issue "Elevating Atmospheric Chemistry Measurements and Modeling with Artificial Intelligence".

Yunjiang Zhang,* Jie Fang, Qingxiao Meng, Xinlei Ge, Hasna Chebaicheb, Olivier Favez, and Jean-Eudes Petit*

Cite This: ACS EST Air 2025, 2, 378−385

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Long-term source apportionment of atmospheric organic aerosols (OA) is essential for supporting air pollution management strategies. While aerosol mass spectrometry (AMS) combined with traditional source apportionment tools can accurately identify various OA sources, they face efficiency challenges when processing large volumes of long-term data. This study proposes an ensemble machine learning approach to efficiently apportion OA sources from long-term AMS measurements. Using six-year observation of a simplified version of AMS (i.e., ACSM) in the Paris region along with OA factor data derived from positive matrix factorization analysis, we developed an ensemble machine learning source apportionment model. Compared to individual machine learning algorithms, the ensemble model substantially reduced the root-mean-square error (RMSE) and increased the correlation coefficient in predicting OA sources by approximately 30% and 5%, respectively. Sensitivity analysis with five years of baseline data revealed that model performance relatively stabilizes when the training data exceeds two years, with RMSE values for primary and secondary OA factors at 0.31−0.45 μg/m^3 and 0.69−0.84 μg/m^3, respectively. This ensemble model not only enhances the efficiency of long-term OA source apportionment but also holds potential for near-real-time online applications.

KEYWORDS: *Organic aerosols, aerosol mass spectrometry, source apportionment, long-term, real-time, machine learning*

1. INTRODUCTION

Organic aerosols (OA) represent a crucial component of atmospheric fine particulate matter,[1,2] significantly influencing human health,[3] air quality,[4] and climate systems.[5,6] OA is typically categorized into primary organic aerosols (POA) and secondary organic aerosols (SOA). POA is directly emitted from various sources, while SOA results from atmospheric oxidation processes involving precursor compounds.[6,7] The identification of OA sources is critical for improving air quality and has become a central focus in regions with severe air pollution, including the United States,[1,2,8,9] Europe,[3,10] and Asia.[11−13] Accurately quantifying these sources and understanding their temporal variations are essential for addressing global air quality challenges.

To address OA sources, several methodologies have been developed, which can be broadly categorized into offline[4,14] and online[15,16] approaches. Offline techniques involve the collection and analysis of OA samples for specific tracer molecules, followed by statistical estimations of their sources.[17] While effective, these methods do not meet the needs for high time resolution and real-time analysis. In contrast, an aerodyne

Received: September 25, 2024
Revised: January 19, 2025
Accepted: January 21, 2025
Published: January 31, 2025

https://doi.org/10.1021/acsestair.4c00262

aerosol mass spectrometry (AMS) has emerged as a powerful tool for source apportionment,[16] offering stability, suitability for extended operation, and high-time-resolution mass spectral data.[18] AMS has been extensively adopted for long-term online observations worldwide,[10,19,20] contributing valuable data for trend analysis, air quality assessment, and model validation. Despite these advancements, traditional positive matrix factorization (PMF) model[21] used with AMS data still require offline processing,[16,22] which is labor-intensive and insufficient for long-term source apportionment.

In past years, advancements in receptor modeling reflect ongoing efforts to address these limitations. For example, previous studies have explored AMS-derived organic tracer fragment ions for estimating OA sources.[23−25] However, their empirical methods exhibit discrepancies compared to the PMF results. Canonaco et al. (2013)[22] introduced the multilinear engine 2 (ME-2) algorithm for AMS data analysis and developed a commercial and closed-source tool[26−28] to enhance the efficiency of the source apportionment process. Despite these innovations, uncertainties persist, particularly in SOA apportionment when compared to detailed manual PMF analysis.[29] In parallel, the integration of machine learning algorithms into atmospheric chemistry has gained traction. Recent studies have demonstrated the potential of machine learning techniques for OA source apportionment,[30,31] opening new avenues for understanding aerosol sources.

In this study, we address the need for more efficient and compatible methods for long-term or near-real-time analysis by developing an ensemble machine learning approach for OA source apportionment using traditional PMF-derived OA source data from online AMS measurements. This method aims to enhance the effectiveness of long-term and potentially near-real-time online source apportionment of OA.

2. DATA AND METHODS

2.1. OA Sources and Metrological Data. This study utilizes OA mass spectrometry data continuously measured by an aerodyne aerosol chemical speciation monitor (ACSM)[32] at the SIRTA site (2.15°E, 48.71°N, 150 m.a.s.l, https://sirta.ipsl.fr/) in the Paris region from November 2011 to March 2018. The OA source factors include two POA factors—hydrocarbon-like OA (HOA) and biomass burning OA (BBOA)—and two oxygenated organic aerosol (OOA) factors—less oxidized oxygenated OA (LO-OOA) and more oxidized oxygenated OA (MO-OOA), both of them commonly considered as representing the SOA fractions. These 4 factors were derived using the PMF model, driven by the ME-2 algorithm with an Igor-based SoFi toolkit software.[22] The two POA factors were partially constrained using reference source profiles, whereas the two SOA factors were resolved without constraints (Zhang et al., 2019).[20] For the comprehensive 6-year OA source apportionment, PMF analyses were conducted separately for each season. Detailed results of the source apportionment are documented in our previous work (Zhang et al., 2019).[20]

To obtain continuous meteorological data, we utilized the ERA5 reanalysis data provided by the European Centre for Medium-Range Weather Forecasts (ECMWF), available through the Copernicus Climate Change Service (CDS) Web site (https://cds.climate.copernicus.eu, last accessed 18th December 2024). The ERA5 data feature a spatial resolution of 0.25° × 0.25° and a temporal resolution of 1 h. We extracted meteorological data for the grid cell corresponding to the SIRTA observation site.

2.2. Ensemble Machine Learning. In this study, we developed regression models to predict OA sources using ensemble machine learning, leveraging OA source factors from the PMF calculation and online measurements of the OA mass spectrometry data. Initially, we evaluated the impact of different OA source prediction features on the performance of machine learning models. As detailed in Table S1, six experimental feature sets were tested using random forest (RF) algorithm: 1) organic mass spectral matrix (m/z 13 to 100) alone; 2) organic mass spectral matrix data combined with time variables such as day of year (DOY), day of week (DOW), and hour of day (HOD); 3) organic mass spectral matrix, the calculated error matrix for corresponding organic ion fragments using the method proposed by previous studies,[32−34] and time variables; 4) organic mass spectral matrix, time variables, three meteorological parameters (relative humidity, air temperature, and pressure); 5) organic mass spectral matrix and three meteorological parameters (relative humidity, air temperature, and pressure); and 6) organic mass spectral matrix, time variables, 18 meteorological parameters (see Table S2). These time variables, i.e., DOY, DOW, and HOD, represent impacts such as primary emissions and secondary formation mechanisms with seasonal, weekly, and diurnal cycles, respectively. Figure S1 presents the model performance evaluated through 10-fold cross-validation across these prediction feature sets. Incorporating time variables in the second experiment set significantly enhanced model performance, especially for SOA factor predictions, compared to the first set. There were no significant differences in prediction performance for POA and SOA factors between the second and third experiment sets. To further test the impact of meteorological parameters for the model prediction performance, we conducted experiments 4, 5, and 6. These sensitivity tests revealed that the models incorporating meteorological parameters as predictors generally performed weaker compared to those without meteorological data (see Figure S1). Finally, due to the offline calculation requirement and the increased computational complexity associated with the error matrix in the third set (which limits its feasibility for real-time source apportionment), we ultimately selected the second experimental set of prediction features for our final model.

Different machine learning algorithms could yield varying prediction performances. In addition to the RF algorithm, we also tested Gradient Boosting Regression Trees (GBRTs) and eXtreme Gradient Boosting (XGBoost). GBRTs is an ensemble method based on decision trees. A decision tree uses a tree structure that starts from the root, branches according to specific conditions, and leads to leaves representing the prediction results.[35] A key disadvantage of decision trees is their tendency to overfit the data if the tree depth is too extensive. RF is a machine learning method that also uses decision trees, but it builds multiple trees by splitting the data set based on random subsets of the data and features. Overfitting is prevented by averaging the predictions from all the individual trees. Compared to RF, GBRTs results are more sensitive to parameter settings during training. However, with optimal parameter tuning, GBRTs can achieve better performance than RF. XGBoost is another machine learning algorithm based on gradient-boosted decision trees. It is a highly scalable and optimized version of gradient boosting machines, improving both speed and predictive performance.[36] XGBoost

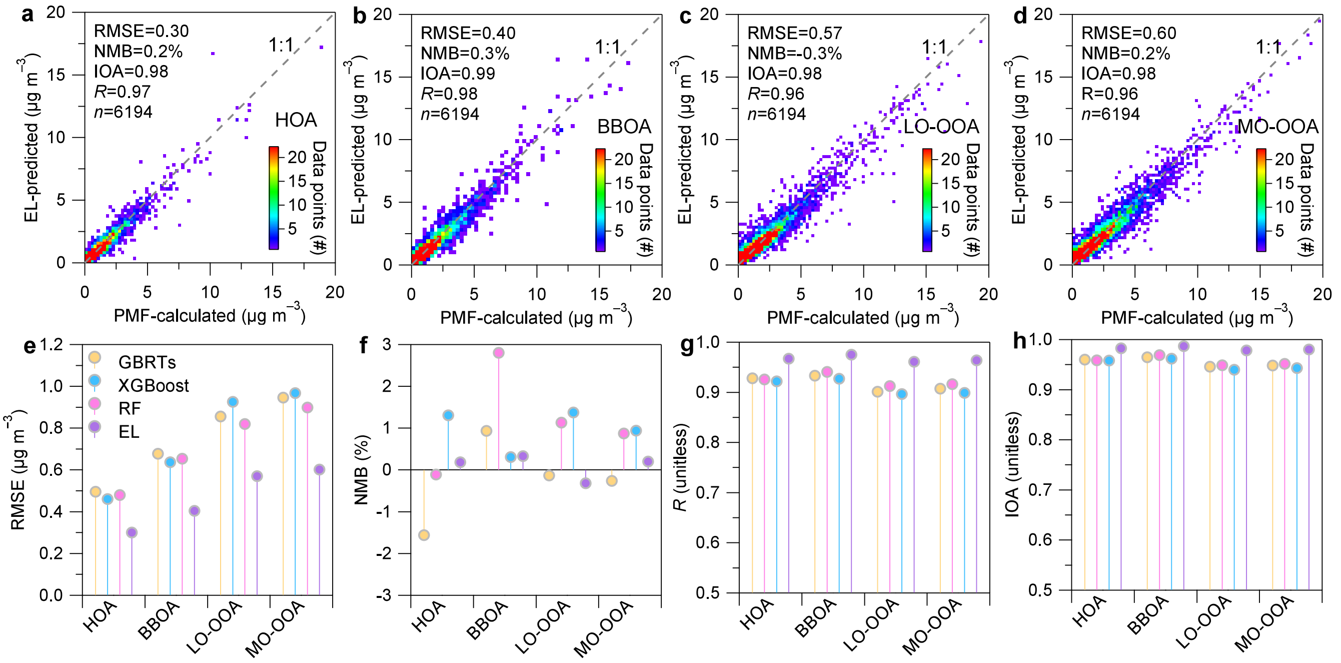


Figure 1. Model performance evaluation. (**a**−**d**) Comparison between the ensemble learning model (EL) predictions and PMF calculated values for OA source factors (HOA, BBOA, LO-OOA, and MO-OOA). (**e**−**h**) Comparison of prediction results based on different machine learning algorithms for the testing data sets.

achieves these improvements by employing a novel tree-learning algorithm and leveraging parallel and distributed computing to accelerate model training.[36,37]

For the RF model, we used Gini importance to rank the key prediction features affecting OA sources. The Gini importance is calculated by summing the decrease in Gini impurity for all nodes in the forest when a split on a variable is made.[38,39] GBRTs can also be used for feature importance ranking, typically aggregating the importance across all base learners.[35] For the XGBoost model, we used the Gain parameter to explore the importance of predictive features. This parameter reflects the contribution of each feature to the improvement in model performance, with higher Gain values indicating a greater impact on reducing model error.[40,41] Figures S2 and S3 compare the 10-fold cross-validation and training data set predictions against PMF calculated values for these three algorithms, including metrics such as root-mean-square error (RMSE), normalized mean bias (NMB), index of agreement (IOA), and correlation coefficient (R). Results show that the algorithms differ in their performance for predicting OA source factors, with XGBoost demonstrating the best performance. Minimal differences between 10-fold cross-validation and training data set prediction performances indicate that all three models effectively avoided overfitting. To mitigate algorithmic discrepancies and further improve prediction accuracy, we employed an ensemble machine learning model that integrates the predictions from the three algorithms, including RF, GBRTs, and XGBoost. The ensemble model combines these predictions using the regression model, as represented in eq 1. In this equation, the dependent variable *Y* represents the OA source factors (i.e., HOA, BBOA, LO-OOA, and MO-OOA), while the independent variables *X* correspond to the predictions from the three machine learning algorithms. The regression model's coefficients, denoted by *m*, are used to optimize the contributions of each algorithm's predictions. For regression analysis, multiple linear regression is commonly used to analyze the influence of multiple independent variables on a single dependent variable, making it suitable for exploring interactions among predictors. Compared to multiple linear regression, ridge regression is better suited for handling multicollinearity among predictor variables,[42] effectively addressing this issue through a regularization term.[43] Therefore, we applied the ridge regression to construct the regression model (eq 1) in this study, effectively addressing multicollinearity among predictions from the different machine learning models and ensuring reliable and accurate predictions.

$$Y = m_1 \cdot X_{RF} + m_2 \cdot X_{GBRTs} + m_3 \cdot X_{XGBoost} \qquad (1)$$

3. RESULTS AND DISCUSSION

3.1. Model Evaluation and Comparison.

Figure 1 presents the comparison of predicted values from the ensemble machine learning model versus individual machine learning models and PMF-calculated values, as well as a performance comparison of different algorithmic models. The results indicate that despite very good performance of individual models, the ensemble learning model outperforms single machine learning models in predicting OA source factors, including HOA, BBOA, LO-OOA, and MO-OOA. Specifically, the ensemble model shows superior performance with RMSE and NMB values ranging from 0.3 to 0.6 $\mu g/m^3$ and 0.2% to 0.3%, respectively, while *R* and IOA values range from 0.96 to 0.98 and 0.98 to 0.99, respectively (Figure 1a-d). Compared to individual machine learning models such as GBRTs, XGBoost, and RF, the ensemble model achieves reductions in RMSE and NMB values by approximately 20−37% and 74−250% and improvements in *R* and IOA values by 3−7% and 2−4%, respectively (Figure 1e-h). These results demonstrate that the ensemble machine learning approach provides more accurate predictions of OA factor sources compared to any individual machine learning model.

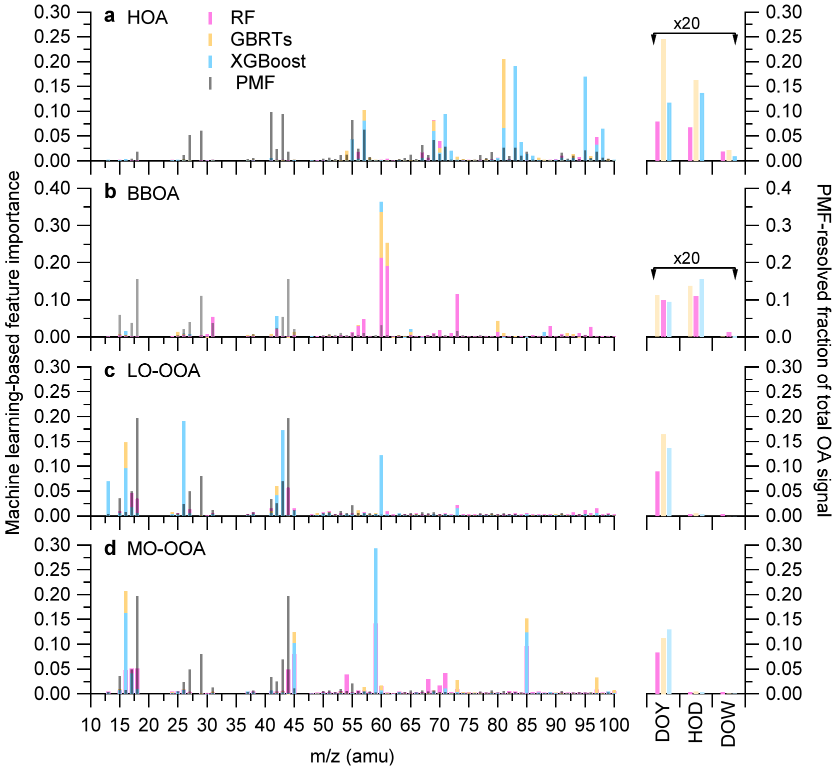


Figure 2. Importance of machine learning prediction features. Distributions of the relative importance of prediction features (left axis) derived from various machine learning algorithms and the fraction of total OA signals in PMF-resolved profiles (right axis) for different OA factors. The comparison highlights the OA source profiles calculated using the PMF model. Note: The relative importance of time variables for POA factors (HOA and BBOA) has been magnified 20 times for improved visual clarity.

To further understand the role of different predictor variables (including OA-specific fragment ions and time variables) in the machine learning models, we analyzed the relative importance of predictor variables in different algorithms and compared these findings with the OA source profiles obtained from traditional PMF calculations (see Figure 2). In PMF analysis, the source profiles of different OA factors typically show distinct characteristics and specific tracer ion fragments. For instance, HOA, mainly associated with vehicle emissions, is characterized by ions such as m/z 55, m/z 57, and m/z 73.[2,25,44] These ions are prominently displayed in the PMF source profiles and also show significant importance in all three machine learning algorithms. The BBOA factor, originating from biomass burning, is most notably traced by the m/z 60 ion, which is widely recognized in various biomass burning sources such as wood burning,[45,46] wildfires,[24,47] and agricultural residue burning.[48] Figure 2b shows that m/z 60 is most important in the three BBOA machine learning models, indicating that this tracer ion has the highest weight among all predictor variables in the models, further validating its significant contribution in the PMF profile of BBOA. In addition, for the two POA factors, time-related predictor variables, such as DOY and HOD, exhibited relative importance in the models, which aligns with expectations. For example, vehicle emissions, a key contributor to HOA, are influenced by rush hours, while emissions from biomass burning heating (e.g., wood burning) show both diurnal peaks and seasonal variations.[20] These time-dependent patterns further validate the predictive models, supporting the relevance of incorporating temporal variables in machine learning-based source apportionment.

Unlike primary OA factors, OOA factors arise from complex atmospheric chemical processes and exhibits seasonal variation. Figure 2c shows that in the three LO-OOA machine learning models, m/z 43 consistently demonstrates significant importance, aligning with PMF results.[19,49] Additionally, m/z 60 also holds notable importance in the LO-OOA models, likely due to the influence of biomass burning on LO-OOA sources during winter,[20] which is reflected in the corresponding ion signals. For MO-OOA, m/z 44 is the most representative tracer fragment ion.[49] Figure 2d highlights that m/z 44 is consistently a key feature across different MO-OOA machine learning models, with m/z 16, m/z 45, m/z 59, and m/z 85 also showing significant importance. Some of these features (except m/z 44), however, are less prominent in traditional PMF models, reflecting methodological differences between machine learning and PMF approaches. Although the feature importance derived from machine learning models may not directly correspond to PMF source profiles, it likely indicates the different influence of feature variables on machine learning prediction outcomes. Such indirect filtering of tracer fragment ions could provide additional evidence for source identification and could be overall acceptable for such applications. Additionally, the strong influence of the seasonal feature variable DOY in the machine learning models for both OOA factors (LO-OOA and MO-OOA) underscores the seasonal variability in SOA formation.

3.2. Prediction of OA Sources.

Figure 3 illustrates the performance of the ensemble machine learning model in predicting different OA source factors across various time scales. We evaluated five different training sample sizes (i.e., 1 year, 2 years, 3 years, 4 years, and 5 years) for the ensemble

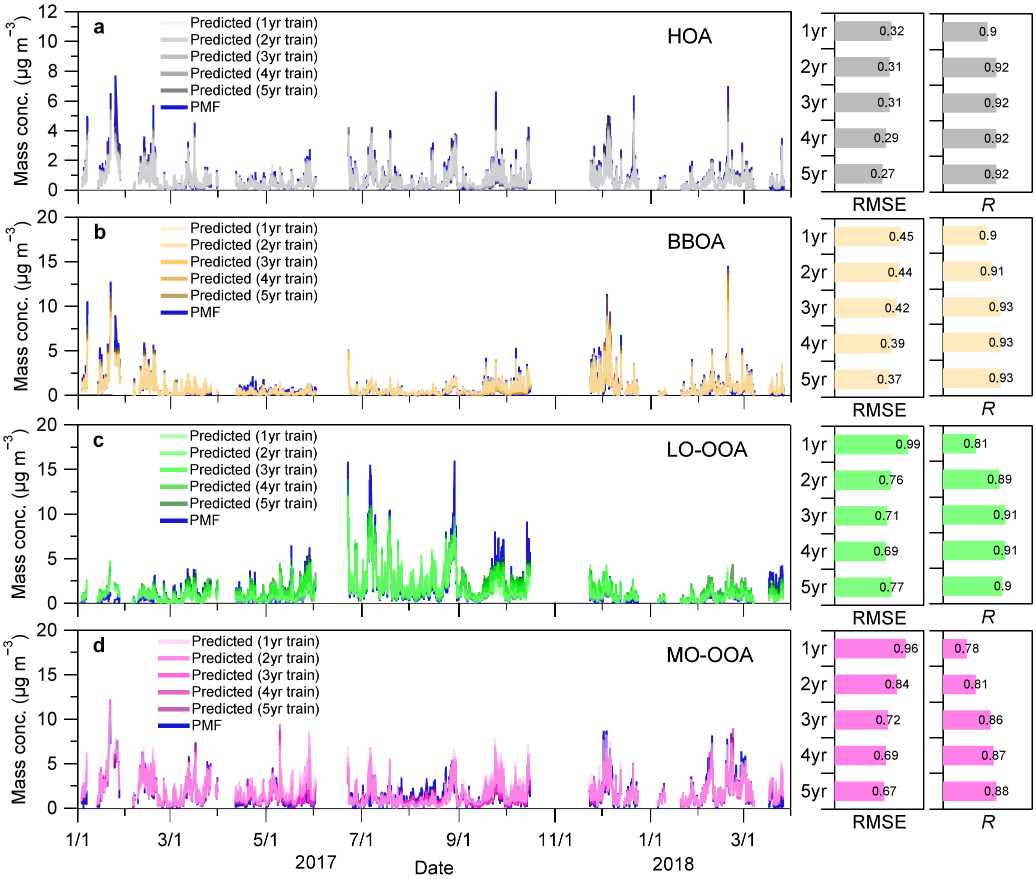


Figure 3. Sensitivity test of prediction performance based on ensemble machine learning models. (a−d) Comparison of the performance of the ensemble learning model in predicting OA sources for the same period (January 2017 to March 2018) using training data sets of different data duration (1 year, 2 years, 3 years, 4 years and 5 years, respectively). The corresponding differences between ensemble machine learning-predicted values and PMF-calculated values for OA factors are provided in Figure S4.

learning models. Generally, the overall performance of the model improves as the duration of training data increases. For POA factors, such as HOA and BBOA, model performance remains relatively stable across different training sample sizes. For instance, *R* values for HOA and BBOA predicted by different ensemble learning models range from 0.9 to 0.92 and 0.9 to 0.93, respectively, while the RMSE values range from 0.27 to 0.32 μg/m^3 and 0.37 to 0.45 μg/m^3, respectively. In contrast, the model performance for SOA factors, such as LO-OOA and MO-OOA, exhibits a stronger dependence on the sample size for model training. For example, with a 1-year training period, the RMSE values for LO-OOA and MO-OOA are 0.99 μg/m^3 and 0.96 μg/m^3, respectively, with *R* values of 0.81 and 0.78. When the model training time scale is extended to 2 years, the RMSE values decrease substantially, and *R* values increase substantially. Specifically, the RMSE and *R* values for LO-OOA and MO-OOA decreased by approximately 9−20% and increased by 2−9%, respectively, with a 2-year training period. Further extending the training time scale results in only marginal improvements in the model performance. This indicates that, to achieve robust performance in predicting SOA factors, we propose a minimum of two years of sample data is required for effective model training. Notably, the ensemble machine learning model achieves stronger predictive performance for POA factors compared to SOA factors. This discrepancy could partly arise from differences in source apportionment treatment using the ME-2 framework. POA factors, such as HOA and BBOA, are constrained with reference spectra, reducing uncertainties in their identification. In contrast, SOA factors are resolved without such constraints, leading to higher uncertainty in their apportionment. The inherent stability of constrained POA factors likely contributes to the more consistent model performance across varying training sample sizes. Future work would need to focus on addressing the uncertainties associated with SOA source apportionment to further enhance prediction accuracy.

3.3. Applications and Limitations.

OA is the most critical chemical components in fine particulate matter in the atmosphere. Reducing its concentration is key to further improving air quality worldwide. Accurately identifying and quantifying the sources of OA and their trends is fundamental to effective OA pollution management. Therefore, the development of efficient OA source apportionment techniques is of significant importance. AMS technology, particularly aerodyne AMS, has become one of the primary methods for OA source apportionment. With the global proliferation and application of AMS technology, there is an increasing demand for OA source apportionment, especially for long-term and high-time-resolution analyses. Traditional source apportionment methods often face challenges related to efficiency and the complexity of data handling, limiting their application in real-time and long-term observations.

The ensemble machine learning method proposed in this study offers a new approach that overcomes the limitations of traditional techniques. By integrating multiple machine learning models, this method not only significantly enhances the efficiency of OA source apportionment but also enables near-real-time online analysis. The application of this method will greatly improve the real-time monitoring capabilities of OA sources and trend changes, allowing scientific research and policy-making to address air quality issues more swiftly.

However, it is important to recognize the limitations of the ensemble learning approach. One notable drawback is that supervised training relies on the accuracy of the training data. If there are uncertainties or biases present in the PMF results used for training, these issues will be reproduced in the model's predictions. Consequently, any inaccuracies in the training data set will persist in the predictions. Additionally, while this method excels at improving efficiency and handling large data sets, it is not designed to identify new or previously unrecognized OA sources. The model can only predict sources that have been defined in the training data, limiting its applicability in scenarios where novel sources may emerge or where the source profiles are not well-represented in the training data.

Looking forward, as data acquisition technologies and computational capabilities advance, the ensemble machine learning method is expected to be further optimized and refined. Combining larger-scale observational data with more sophisticated algorithms could lead to more detailed source apportionment of OA. Additionally, integrating ensemble learning methods with other advanced technologies, such as remote sensing and high-resolution aerosol composition analysis, will provide more comprehensive support for global air quality management and pollution control.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestair.4c00262.

Machine learning model experiment sets and corresponding model performance tests (Table S1 and Figure S1), detailed information for meteorological parameters (Table S2) and additional model performance tests (Figures S2–S4) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Yunjiang Zhang − *Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Atmospheric Environment Monitoring and Pollution Control, School of Environmental Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; Institut National de l'Environnement Industriel et des Risques, Verneuil-en-Halatte 60550, France;* orcid.org/0009-0005-8777-2082; Email: yjzhang@nuist.edu.cn

Jean-Eudes Petit − *Laboratoire des Sciences du Climat et de l'Environnement, CEA/Orme des Merisiers, Gif sur Yvette 91191, France;* orcid.org/0000-0003-1516-5927; Email: jean-eudes.petit@lsce.ipsl.fr

Authors

Jie Fang − *Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Atmospheric Environment Monitoring and Pollution Control, School of Environmental Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China*

Qingxiao Meng − *Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Atmospheric Environment Monitoring and Pollution Control, School of Environmental Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China*

Xinlei Ge − *Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Atmospheric Environment Monitoring and Pollution Control, School of Environmental Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China;* orcid.org/0000-0001-9531-6478

Hasna Chebaicheb − *Institut National de l'Environnement Industriel et des Risques, Verneuil-en-Halatte 60550, France*

Olivier Favez − *Institut National de l'Environnement Industriel et des Risques, Verneuil-en-Halatte 60550, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestair.4c00262

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 42207124), Natural Science Foundation of Jiangsu Province (grant no. BK20210663), and this research has been supported by the EU-FP7 and H2020 ACTRIS projects (grant nos. 262254 and 654109), by the National Center for Scientific Research (CNRS), by the French alternatives energies and Atomic Energy Commission (CEA), and by the French ministry of Environment through its funding to the reference laboratory for air quality monitoring (LCSQA).

■ REFERENCES

(1) Zhang, Q.; Jimenez, J. L.; Canagaratna, M. R.; Allan, J. D.; Coe, H.; Ulbrich, I.; Alfarra, M. R.; Takami, A.; Middlebrook, A. M.; Sun, Y. L.; Dzepina, K.; Dunlea, E.; Docherty, K.; DeCarlo, P. F.; Salcedo, D.; Onasch, T.; Jayne, J. T.; Miyoshi, T.; Shimono, A.; Hatakeyama, S.; Takegawa, N.; Kondo, Y.; Schneider, J.; Drewnick, F.; Borrmann, S.; Weimer, S.; Demerjian, K.; Williams, P.; Bower, K.; Bahreini, R.; Cottrell, L.; Griffin, R. J.; Rautiainen, J.; Sun, J. Y.; Zhang, Y. M.; Worsnop, D. R. Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes. *Geophys. Res. Lett.* **2007**, *34* (13), No. L13801.

(2) Jimenez, J. L.; Canagaratna, M. R.; Donahue, N. M.; Prevot, A. S. H.; Zhang, Q.; Kroll, J. H.; DeCarlo, P. F.; Allan, J. D.; Coe, H.; Ng, N. L.; Aiken, A. C.; Docherty, K. S.; Ulbrich, I. M.; Grieshop, A. P.; Robinson, A. L.; Duplissy, J.; Smith, J. D.; Wilson, K. R.; Lanz, V. A.; Hueglin, C.; Sun, Y. L.; Tian, J.; Laaksonen, A.; Raatikainen, T.; Rautiainen, J.; Vaattovaara, P.; Ehn, M.; Kulmala, M.; Tomlinson, J. M.; Collins, D. R.; Cubison, M. J.; Dunlea, J.; Huffman, J. A.; Onasch, T. B.; Alfarra, M. R.; Williams, P. I.; Bower, K.; Kondo, Y.; Schneider, J.; Drewnick, F.; Borrmann, S.; Weimer, S.; Demerjian, K.; Salcedo, D.; Cottrell, L.; Griffin, R.; Takami, A.; Miyoshi, T.; Hatakeyama, S.; Shimono, A.; Sun, J. Y.; Zhang, Y. M.; Dzepina, K.; Kimmel, J. R.; Sueper, D.; Jayne, J. T.; Herndon, S. C.; Trimborn, A. M.; Williams, L. R.; Wood, E. C.; Middlebrook, A. M.; Kolb, C. E.; Baltensperger, U.; Worsnop, D. R. Evolution of Organic Aerosols in the Atmosphere. *Science* **2009**, *326* (5959), 1525–1529.

(3) Daellenbach, K. R.; Uzu, G.; Jiang, J.; Cassagnes, L.-E.; Leni, Z.; Vlachou, A.; Stefenelli, G.; Canonaco, F.; Weber, S.; Segers, A.; Kuenen, J. J. P.; Schaap, M.; Favez, O.; Albinet, A.; Aksoyoglu, S.; Dommen, J.; Baltensperger, U.; Geiser, M.; El Haddad, I.; Jaffrezo, J.-L.; Prévôt, A. S. H. Sources of particulate-matter air pollution and its oxidative potential in Europe. *Nature* **2020**, *587* (7834), 414−419.

(4) Huang, R. J.; Zhang, Y.; Bozzetti, C.; Ho, K. F.; Cao, J. J.; Han, Y.; Daellenbach, K. R.; Slowik, J. G.; Platt, S. M.; Canonaco, F.; Zotter, P.; Wolf, R.; Pieber, S. M.; Bruns, E. A.; Crippa, M.; Ciarelli, G.; Piazzalunga, A.; Schwikowski, M.; Abbaszade, G.; Schnelle-Kreis, J.; Zimmermann, R.; An, Z.; Szidat, S.; Baltensperger, U.; El Haddad, I.; Prevot, A. S. High secondary aerosol contribution to particulate pollution during haze events in China. *Nature* **2014**, *514* (7521), 218−222.

(5) IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, United Kingdom and New York, NY, USA, 2021; In Press.

(6) Shrivastava, M.; Cappa, C. D.; Fan, J.; Goldstein, A. H.; Guenther, A. B.; Jimenez, J. L.; Kuang, C.; Laskin, A.; Martin, S. T.; Ng, N. L.; Petaja, T.; Pierce, J. R.; Rasch, P. J.; Roldin, P.; Seinfeld, J. H.; Shilling, J.; Smith, J. N.; Thornton, J. A.; Volkamer, R.; Wang, J.; Worsnop, D. R.; Zaveri, R. A.; Zelenyuk, A.; Zhang, Q. Recent advances in understanding secondary organic aerosol: Implications for global climate forcing. *Reviews of Geophysics* **2017**, *55* (2), 509−559.

(7) Hallquist, M.; Wenger, J. C.; Baltensperger, U.; Rudich, Y.; Simpson, D.; Claeys, M.; Dommen, J.; Donahue, N. M.; George, C.; Goldstein, A. H.; Hamilton, J. F.; Herrmann, H.; Hoffmann, T.; Iinuma, Y.; Jang, M.; Jenkin, M. E.; Jimenez, J. L.; Kiendler-Scharr, A.; Maenhaut, W.; McFiggans, G.; Mentel, T. F.; Monod, A.; Prévôt, A. S. H.; Seinfeld, J. H.; Surratt, J. D.; Szmigielski, R.; Wildt, J. The formation, properties and impact of secondary organic aerosol: current and emerging issues. *Atmos. Chem. Phys.* **2009**, *9* (14), 5155−5236.

(8) Xu, L.; Guo, H.; Boyd, C. M.; Klein, M.; Bougiatioti, A.; Cerully, K. M.; Hite, J. R.; Isaacman-VanWertz, G.; Kreisberg, N. M.; Knote, C.; Olson, K.; Koss, A.; Goldstein, A. H.; Hering, S. V.; de Gouw, J.; Baumann, K.; Lee, S.-H.; Nenes, A.; Weber, R. J.; Ng, N. L. Effects of anthropogenic emissions on aerosol formation from isoprene and monoterpenes in the southeastern United States. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112* (1), 37−42.

(9) Hu, W. W.; Campuzano-Jost, P.; Palm, B. B.; Day, D. A.; Ortega, A. M.; Hayes, P. L.; Krechmer, J. E.; Chen, Q.; Kuwata, M.; Liu, Y. J.; de Sá, S. S.; McKinney, K.; Martin, S. T.; Hu, M.; Budisulistiorini, S. H.; Riva, M.; Surratt, J. D.; St. Clair, J. M.; Isaacman-Van Wertz, G.; Yee, L. D.; Goldstein, A. H.; Carbone, S.; Brito, J.; Artaxo, P.; de Gouw, J. A.; Koss, A.; Wisthaler, A.; Mikoviny, T.; Karl, T.; Kaser, L.; Jud, W.; Hansel, A.; Docherty, K. S.; Alexander, M. L.; Robinson, N. H.; Coe, H.; Allan, J. D.; Canagaratna, M. R.; Paulot, F.; Jimenez, J. L. Characterization of a real-time tracer for isoprene epoxydiols-derived secondary organic aerosol (IEPOX-SOA) from aerosol mass spectrometer measurements. *Atmos. Chem. Phys.* **2015**, *15* (20), 11807−11833.

(10) Chen, G.; Canonaco, F.; Tobler, A.; Aas, W.; Alastuey, A.; Allan, J.; Atabakhsh, S.; Aurela, M.; Baltensperger, U.; Bougiatioti, A.; De Brito, J. F.; Ceburnis, D.; Chazeau, B.; Chebaicheb, H.; Daellenbach, K. R.; Ehn, M.; El Haddad, I.; Eleftheriadis, K.; Favez, O.; Flentje, H.; Font, A.; Fossum, K.; Freney, E.; Gini, M.; Green, D. C.; Heikkinen, L.; Herrmann, H.; Kalogridis, A.-C.; Keernik, H.; Lhotka, R.; Lin, C.; Lunder, C.; Maasikmets, M.; Manousakas, M. I.; Marchand, N.; Marin, C.; Marmureanu, L.; Mihalopoulos, N.; Močnik, G.; Nęcki, J.; O'Dowd, C.; Ovadnevaite, J.; Peter, T.; Petit, J.-E.; Pikridas, M.; Matthew Platt, S.; Pokorná, P.; Poulain, L.; Priestman, M.; Riffault, V.; Rinaldi, M.; Różański, K.; Schwarz, J.; Sciare, J.; Simon, L.; Skiba, A.; Slowik, J. G.; Sosedova, Y.; Stavroulas, I.; Styszko, K.; Teinemaa, E.; Timonen, H.; Tremper, A.; Vasilescu, J.; Via, M.; Vodička, P.; Wiedensohler, A.; Zografou, O.; Cruz Minguillón, M.; Prévôt, A. S. H. European aerosol phenomenology−8: Harmonised source apportionment of organic aerosol using 22 Year-long ACSM/AMS datasets. *Environ. Int.* **2022**, *166*, 107325.

(11) Zhou, W.; Xu, W.; Kim, H.; Zhang, Q.; Fu, P.; Worsnop, D. R.; Sun, Y. A review of aerosol chemistry in Asia: insights from aerosol mass spectrometer measurements. *Environ. Sci.: Processes Impacts* **2020**, *22*, 1616−1653.

(12) Li, Y. J.; Sun, Y.; Zhang, Q.; Li, X.; Li, M.; Zhou, Z.; Chan, C. K. Real-time chemical characterization of atmospheric particulate matter in China: A review. *Atmos. Environ.* **2017**, *158*, 270−304.

(13) Gunthe, S. S.; Liu, P.; Panda, U.; Raj, S. S.; Sharma, A.; Darbyshire, E.; Reyes-Villegas, E.; Allan, J.; Chen, Y.; Wang, X.; Song, S.; Pöhlker, M. L.; Shi, L.; Wang, Y.; Kommula, S. M.; Liu, T.; Ravikrishna, R.; McFiggans, G.; Mickley, L. J.; Martin, S. T.; Pöschl, U.; Andreae, M. O.; Coe, H. Enhanced aerosol particle growth sustained by high continental chlorine emission in India. *Nat. Geosci.* **2021**, *14* (2), 77−84.

(14) Fu, P.; Kawamura, K.; Seki, O.; Izawa, Y.; Shiraiwa, T.; Ashworth, K. Historical Trends of Biogenic SOA Tracers in an Ice Core from Kamchatka Peninsula. *Environ. Sci. Technol. Lett.* **2016**, *3* (10), 351−358.

(15) Lanz, V. A.; Alfarra, M. R.; Baltensperger, U.; Buchmann, B.; Hueglin, C.; Prévôt, A. S. H. Source apportionment of submicron organic aerosols at an urban site by factor analytical modelling of aerosol mass spectra. *Atmos. Chem. Phys.* **2007**, *7* (6), 1503−1522.

(16) Zhang, Q.; Jimenez, J. L.; Canagaratna, M. R.; Ulbrich, I. M.; Ng, N. L.; Worsnop, D. R.; Sun, Y. Understanding atmospheric organic aerosols via factor analysis of aerosol mass spectrometry: a review. *Anal Bioanal Chem.* **2011**, *401* (10), 3045−67.

(17) Srivastava, D.; Favez, O.; Perraudin, E.; Villenave, E.; Albinet, A. Comparison of Measurement-Based Methodologies to Apportion Secondary Organic Carbon (SOC) in PM2.5: A Review of Recent Studies. *Atmosphere* **2018**, *9* (11), 452.

(18) Canagaratna, M. R.; Jayne, J. T.; Jimenez, J. L.; Allan, J. D.; Alfarra, M. R.; Zhang, Q.; Onasch, T. B.; Drewnick, F.; Coe, H.; Middlebrook, A.; Delia, A.; Williams, L. R.; Trimborn, A. M.; Northway, M. J.; DeCarlo, P. F.; Kolb, C. E.; Davidovits, P.; Worsnop, D. R. Chemical and microphysical characterization of ambient aerosols with the aerodyne aerosol mass spectrometer. *Mass Spectrom. Rev.* **2007**, *26* (2), 185−222.

(19) Sun, Y.; Xu, W.; Zhang, Q.; Jiang, Q.; Canonaco, F.; Prévôt, A. S. H.; Fu, P.; Li, J.; Jayne, J.; Worsnop, D. R.; Wang, Z. Source apportionment of organic aerosol from 2-year highly time-resolved measurements by an aerosol chemical speciation monitor in Beijing, China. *Atmos. Chem. Phys.* **2018**, *18* (12), 8469−8489.

(20) Zhang, Y.; Favez, O.; Petit, J. E.; Canonaco, F.; Truong, F.; Bonnaire, N.; Crenn, V.; Amodeo, T.; Prévôt, A. S. H.; Sciare, J.; Gros, V.; Albinet, A. Six-year source apportionment of submicron organic aerosols from near-continuous highly time-resolved measurements at SIRTA (Paris area, France). *Atmos. Chem. Phys.* **2019**, *19* (23), 14755−14776.

(21) Paatero, P.; Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **1994**, *5* (2), 111−126.

(22) Canonaco, F.; Crippa, M.; Slowik, J. G.; Baltensperger, U.; Prévôt, A. S. H. SoFi, an IGOR-based interface for the efficient use of the generalized multilinear engine (ME-2) for the source apportionment: ME-2 application to aerosol mass spectrometer data. *Atmos. Meas. Technol.* **2013**, *6* (12), 3649−3661.

(23) Ng, N. L.; Canagaratna, M. R.; Jimenez, J. L.; Zhang, Q.; Ulbrich, I. M.; Worsnop, D. R. Real-Time Methods for Estimating Organic Component Mass Concentrations from Aerosol Mass Spectrometer Data. *Environ. Sci. Technol.* **2011**, *45* (3), 910−916.

(24) Cubison, M. J.; Ortega, A. M.; Hayes, P. L.; Farmer, D. K.; Day, D.; Lechner, M. J.; Brune, W. H.; Apel, E.; Diskin, G. S.; Fisher, J. A.; Fuelberg, H. E.; Hecobian, A.; Knapp, D. J.; Mikoviny, T.; Riemer, D.; Sachse, G. W.; Sessions, W.; Weber, R. J.; Weinheimer, A. J.; Wisthaler, A.; Jimenez, J. L. Effects of aging on organic aerosol from open biomass burning smoke in aircraft and laboratory studies. *Atmos. Chem. Phys.* **2011**, *11* (23), 12049−12064.

(25) Zhang, Q.; Worsnop, D. R.; Canagaratna, M. R.; Jimenez, J. L. Hydrocarbon-like and oxygenated organic aerosols in Pittsburgh: insights into sources and processes of organic aerosols. *Atmos. Chem. Phys.* **2005**, *5* (12), 3289−3311.

(26) Chen, G.; Canonaco, F.; Slowik, J. G.; Daellenbach, K. R.; Tobler, A.; Petit, J.-E.; Favez, O.; Stavroulas, I.; Mihalopoulos, N.; Gerasopoulos, E.; El Haddad, I.; Baltensperger, U.; Prévôt, A. S. H. Real-Time Source Apportionment of Organic Aerosols in Three European Cities. *Environ. Sci. Technol.* **2022**, *56* (22), 15290−15297.

(27) Via, M.; Chen, G.; Canonaco, F.; Daellenbach, K. R.; Chazeau, B.; Chebaicheb, H.; Jiang, J.; Keernik, H.; Lin, C.; Marchand, N.; Marin, C.; O'Dowd, C.; Ovadnevaite, J.; Petit, J. E.; Pikridas, M.; Riffault, V.; Sciare, J.; Slowik, J. G.; Simon, L.; Vasilescu, J.; Zhang, Y.; Favez, O.; Prévôt, A. S. H.; Alastuey, A.; Minguillón, M. C. Rolling vs. Seasonal PMF: Real-world multi-site and synthetic dataset comparison. *Atmos. Meas. Tech.* **2022**, *15* (18), 5479−5495.

(28) Canonaco, F.; Tobler, A.; Chen, G.; Sosedova, Y.; Slowik, J. G.; Bozzetti, C.; Daellenbach, K. R.; El Haddad, I.; Crippa, M.; Huang, R. J.; Furger, M.; Baltensperger, U.; Prévôt, A. S. H. A new method for long-term source apportionment with time-dependent factor profiles and uncertainty assessment using SoFi Pro: application to 1 year of organic aerosol data. *Atmos. Meas. Technol.* **2021**, *14* (2), 923−943.

(29) Via, M.; Chen, G.; Canonaco, F.; Daellenbach, K. R.; Chazeau, B.; Chebaicheb, H.; Jiang, J.; Keernik, H.; Lin, C.; Marchand, N.; Marin, C.; O'Dowd, C.; Ovadnevaite, J.; Petit, J. E.; Pikridas, M.; Riffault, V.; Sciare, J.; Slowik, J. G.; Simon, L.; Vasilescu, J.; Zhang, Y.; Favez, O.; Prévôt, A. S. H.; Alastuey, A.; Cruz Minguillón, M. Rolling vs. seasonal PMF: real-world multi-site and synthetic dataset comparison. *Atmos. Meas. Technol.* **2022**, *15* (18), 5479−5495.

(30) Pande, P.; Shrivastava, M.; Shilling, J. E.; Zelenyuk, A.; Zhang, Q.; Chen, Q.; Ng, N. L.; Zhang, Y.; Takeuchi, M.; Nah, T.; Rasool, Q. Z.; Zhang, Y.; Zhao, B.; Liu, Y. Novel Application of Machine Learning Techniques for Rapid Source Apportionment of Aerosol Mass Spectrometer Datasets. *ACS Earth and Space Chemistry* **2022**, *6* (4), 932−942.

(31) Qin, Y.; Ye, J.; Ohno, P.; Liu, P.; Wang, J.; Fu, P.; Zhou, L.; Li, Y. J.; Martin, S. T.; Chan, C. K. Assessing the Nonlinear Effect of Atmospheric Variables on Primary and Oxygenated Organic Aerosol Concentration Using Machine Learning. *ACS Earth and Space Chemistry* **2022**, *6* (4), 1059−1066.

(32) Ng, N. L.; Herndon, S. C.; Trimborn, A.; Canagaratna, M. R.; Croteau, P. L.; Onasch, T. B.; Sueper, D.; Worsnop, D. R.; Zhang, Q.; Sun, Y. L.; Jayne, J. T. An Aerosol Chemical Speciation Monitor (ACSM) for Routine Monitoring of the Composition and Mass Concentrations of Ambient Aerosol. *Aerosol Sci. Technol.* **2011**, *45* (7), 780−794.

(33) Ulbrich, I. M.; Canagaratna, M. R.; Zhang, Q.; Worsnop, D. R.; Jimenez, J. L. Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data. *Atmos. Chem. Phys.* **2009**, *9* (9), 2891−2918.

(34) Sun, Y.; Wang, Z.; Dong, H.; Yang, T.; Li, J.; Pan, X.; Chen, P.; Jayne, J. T. Characterization of summer organic and inorganic aerosols in Beijing, China with an Aerosol Chemical Speciation Monitor. *Atmos. Environ.* **2012**, *51*, 250−259.

(35) Konstantinov, A. V.; Utkin, L. V. Interpretable machine learning with an ensemble of gradient boosting machines. *Knowledge-Based Systems* **2021**, *222*, 106993.

(36) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: San Francisco, California, USA, 2016; pp 785−794.

(37) Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54* (3), 1937−1967.

(38) Wright, M. N.; Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, *77* (1), 1−17.

(39) Nembrini, S.; König, I. R.; Wright, M. N. The revival of the Gini importance? *Bioinformatics* **2018**, *34* (21), 3711−3718.

(40) Sagi, O.; Rokach, L. Approximating XGBoost with an interpretable decision tree. *Information Sciences* **2021**, *572*, 522−542.

(41) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure−Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56* (12), 2353−2360.

(42) Kidwell, J. S.; Brown, L. H. Ridge Regression as a Technique for Analyzing Models with Multicollinearity. *Journal of Marriage and Family* **1982**, *44* (2), 287−299.

(43) Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12* (1), 55−67.

(44) Robinson, A. L.; Donahue, N. M.; Shrivastava, M. K.; Weitkamp, E. A.; Sage, A. M.; Grieshop, A. P.; Lane, T. E.; Pierce, J. R.; Pandis, S. N. Rethinking organic aerosols: semivolatile emissions and photochemical aging. *Science* **2007**, *315* (5816), 1259−62.

(45) Elsasser, M.; Crippa, M.; Orasche, J.; DeCarlo, P. F.; Oster, M.; Pitz, M.; Cyrys, J.; Gustafson, T. L.; Pettersson, J. B. C.; Schnelle-Kreis, J.; Prévôt, A. S. H.; Zimmermann, R. Organic molecular markers and signature from wood combustion particles in winter ambient aerosols: aerosol mass spectrometer (AMS) and high time-resolved GC-MS measurements in Augsburg, Germany. *Atmos. Chem. Phys.* **2012**, *12* (14), 6113−6128.

(46) Gilardoni, S.; Massoli, P.; Paglione, M.; Giulianelli, L.; Carbone, C.; Rinaldi, M.; Decesari, S.; Sandrini, S.; Costabile, F.; Gobbi, G. P.; Pietrogrande, M. C.; Visentin, M.; Scotto, F.; Fuzzi, S.; Facchini, M. C. Direct observation of aqueous secondary organic aerosol from biomass-burning emissions. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113* (36), 10013−10018.

(47) Forrister, H.; Liu, J.; Scheuer, E.; Dibb, J.; Ziemba, L.; Thornhill, K. L.; Anderson, B.; Diskin, G.; Perring, A. E.; Schwarz, J. P.; Campuzano-Jost, P.; Day, D. A.; Palm, B. B.; Jimenez, J. L.; Nenes, A.; Weber, R. J. Evolution of brown carbon in wildfire plumes. *Geophys. Res. Lett.* **2015**, *42* (11), 4623−4630.

(48) Zhang, Y. J.; Tang, L. L.; Wang, Z.; Yu, H. X.; Sun, Y. L.; Liu, D.; Qin, W.; Canonaco, F.; Prévôt, A. S. H.; Zhang, H. L.; Zhou, H. C. Insights into characteristics, sources, and evolution of submicron aerosols during harvest seasons in the Yangtze River delta region, China. *Atmos. Chem. Phys.* **2015**, *15* (3), 1331−1349.

(49) Ng, N. L.; Canagaratna, M. R.; Zhang, Q.; Jimenez, J. L.; Tian, J.; Ulbrich, I. M.; Kroll, J. H.; Docherty, K. S.; Chhabra, P. S.; Bahreini, R.; Murphy, S. M.; Seinfeld, J. H.; Hildebrandt, L.; Donahue, N. M.; DeCarlo, P. F.; Lanz, V. A.; Prévôt, A. S. H.; Dinar, E.; Rudich, Y.; Worsnop, D. R. Organic aerosol components observed in Northern Hemispheric datasets from Aerosol Mass Spectrometry. *Atmos. Chem. Phys.* **2010**, *10* (10), 4625−4641.