JGR Atmospheres

RESEARCH ARTICLE
10.1029/2024JD042477

Key Points:

- The proposed ACTE model overcomes the limitations of traditional techniques, delivering high-resolution VOC data
- High-resolution VOC data enhances ozone mechanism simulations and deepens the understanding of chemical processes
- Machine learning techniques exhibit strong adaptability and practical potential in constrained monitoring scenarios

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

X.-F. Huang,
huangxf@pku.edu.cn

Citation:

Cheng, Y., Huang, X.-F., Peng, Y., Cao, L.-M., Peng, X., Wu, J., & He, L.-Y. (2025). Enhancing time resolution of ambient VOC measurement data by machine learning: From one-hour to five minutes. *Journal of Geophysical Research: Atmospheres*, *130*, e2024JD042477. https://doi.org/10.1029/2024JD042477

Received 13 SEP 2024
Accepted 15 FEB 2025

Author Contributions:

Conceptualization: Yong Cheng, Xiao-Feng Huang
Data curation: Yong Cheng, Yan Peng
Formal analysis: Yong Cheng, Xiao-Feng Huang
Funding acquisition: Xiao-Feng Huang, Ling-Yan He
Investigation: Yan Peng, Li-Ming Cao, Xing Peng, Jiang Wu
Methodology: Yong Cheng
Project administration: Ling-Yan He
Resources: Xiao-Feng Huang
Software: Yong Cheng
Supervision: Ling-Yan He
Validation: Yong Cheng
Visualization: Yong Cheng
Writing – original draft: Yong Cheng
Writing – review & editing: Yong Cheng, Xiao-Feng Huang, Yan Peng,

Enhancing Time Resolution of Ambient VOC Measurement Data by Machine Learning: From One-Hour to Five Minutes

Yong Cheng[1], Xiao-Feng Huang[1], Yan Peng[1], Li-Ming Cao[1], Xing Peng[1], Jiang Wu[1], and Ling-Yan He[1]

[1]Laboratory of Atmospheric Observation Supersite, School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen, China

Abstract Atmospheric volatile organic compounds (VOCs) significantly impact the environment and public health, necessitating precise, continuous online monitoring. Currently, VOCs monitoring primarily uses Gas Chromatography-Mass Spectrometry (GC-MS) and Proton Transfer Reaction-Time of Flight Mass Spectrometry (PTR-ToF-MS). GC-MS is favored for its accurate compound identification capabilities but is limited by its lower temporal resolution. Conversely, PTR-ToF-MS, while achieving minute-scale resolution by directly ionizing samples, struggles to detect low-proton-affinity compounds. Here, based on 5 years of long-term online monitoring data, we propose the Adaptive Convolutional Tree Ensemble (ACTE) model to surpass current instruments limitations and accurately obtain high-resolution (5-min) VOCs. Our results indicate that this model consistently achieves robust predictive accuracy across different major species categories, notably achieving R^2 values of 0.92 and 0.89 for alkanes and alkenes, respectively, which mostly have low-proton-affinity. Furthermore, by comparing simulations using different VOCs temporal resolutions in ozone mechanism modeling, we found that models with higher temporal resolution more accurately and comprehensively capture the rapidly occurring photochemical reactions, whereas hourly models tend to overlook many details, potentially leading to inaccuracies in understanding the related mechanisms. This study underscores the potential of machine learning to improve monitoring of atmospheric pollutants and enhance our understanding of atmospheric chemical processes.

Plain Language Summary Atmospheric volatile organic compounds (VOCs) significantly impact the environment and public health, requiring precise, continuous monitoring. Traditional methods like Gas Chromatography-Mass Spectrometry (GC-MS) and Proton Transfer Reaction-Time of Flight Mass Spectrometry (PTR-ToF-MS) have limitations: GC-MS is accurate but slow, while PTR-ToF-MS updates data quickly but struggles with compounds of low proton affinity, leading to data gaps. We introduce the Adaptive Convolutional Tree Ensemble (ACTE) model, built from 5 years of monitoring data, to overcome these limitations. The ACTE model provides high-resolution data every few minutes with high predictive accuracy across various VOC categories, notably achieving R^2 values of 0.92 and 0.89 for alkanes and alkenes. Comparisons of ozone mechanism modeling using different VOC data resolutions reveal that hourly scale models often miss many details, whereas high-resolution data significantly enhances accuracy. This study highlights the potential of advanced machine learning to improve both monitoring and our understanding of atmospheric chemical processes.

1. Introduction

VOCs in the atmosphere have significant environmental and health impacts (Ehn et al., 2014; Mellouki et al., 2015). Not only are some of these compounds inherently toxic, such as formaldehyde and benzene, which are known carcinogens, but most also directly participate in atmospheric chemical reactions (Ferracci et al., 2024; Lamkaddam et al., 2021; Xiong et al., 2024). These reactions lead to the formation of pollutants like ozone (O_3) and fine particulate matter ($PM_{2.5}$) (T. Wang et al., 2017; Zhang et al., 2014), thereby exacerbating air pollution and posing direct threats to human health (Dedoussi et al., 2020; He et al., 2024; Huang et al., 2023). Due to significant variations in the reaction rates of different VOC species in the atmosphere, ranging from seconds to longer time scales (Seinfeld & Pandis, 1998), and with most VOCs involved in photochemical reactions exhibiting rapid reaction rates, conventional monitoring at hourly resolution scales often overlooks short-term episodic events and difficult to meet the high-quality simulation requirements of relevant mechanistic models (Lee et al., 2021; Wolfe et al., 2016). Therefore, continuous monitoring of VOC concentrations and their dynamic

Li-Ming Cao, Xing Peng, Jiang Wu, Ling-Yan He

variations at higher temporal resolutions in the atmosphere is of paramount importance for effectively assessing and mitigating their environmental and health impacts (Huang et al., 2020; Z. Li et al., 2020; W. Wang, Li, et al., 2024).

Currently, the field of online atmospheric environmental monitoring primarily utilizes two types of instruments to comprehensively monitor the concentration changes of VOCs in the atmosphere: GC-MS and PTR-ToF-MS (Coggon et al., 2021; Gkatzelis et al., 2021; W. Wang et al., 2022). Due to the different monitoring principles of these instruments, there are significant differences in their operation, performance, and application scopes (Chang et al., 2022; Kajos et al., 2015; W. Wang, Yuan, et al., 2024). GC-MS separates VOCs based on their volatility and interaction with a stationary phase, followed by mass spectral analysis for identification and quantification (Nozière et al., 2015). It provides high precision, especially for C_2–C_{12} non-methane hydrocarbons (NMHCs) (Tan et al., 2018; Zhang et al., 2022; Zhou et al., 2023). However, its temporal resolution is limited (∼1 hr), as sample collection is typically done in short intervals, with the remaining time spent on system cleaning and analysis (Cheng, Huang, et al., 2023; Zhang et al., 2024). In contrast, PTR-ToF-MS is based on the principle of proton transfer reactions, allowing for the direct ionization of VOCs during sampling without the need for complex sample preparation (Graus et al., 2010; Jordan et al., 2009; Lindinger et al., 1998). This method enables real-time monitoring of VOC concentration changes in the atmospheric environment with a temporal resolution of minutes or better, making it particularly suitable for rapid detection applications, especially for compounds with high-proton-affinity, such as alcohols, esters, ethers, and others (Cappellin et al., 2012; Park et al., 2013; Yuan et al., 2017). However, the proton transfer ionization mechanism of PTR-ToF-MS has lower detection efficiency for compounds with low-proton-affinity, such as alkenes and alkanes, and faces challenges in differentiating isomers of compounds (Pfannerstill et al., 2024; Yuan et al., 2017). These limitations significantly impact its widespread application in practical uses and theoretical studies (Nozière et al., 2015; Zhu et al., 2023). With the rapid growth of environmental data and advances in machine learning, these techniques are now highly effective in managing large high-dimensional data sets and identifying complex nonlinear patterns in atmospheric science (Bi et al., 2023; Wei et al., 2023; Zhu et al., 2024), positioning them as promising tools for air pollution prediction. However, most previous studies have focused on predicting $PM_{2.5}$ or ozone concentrations (Cheng, Zhu, et al., 2021; Dimri et al., 2024; Wen et al., 2024), while VOCs, which are more challenging to obtain accurate observational data for, have rarely been used as prediction targets (Ye et al., 2022), especially under real atmospheric conditions. Furthermore, there is a notable lack of predictive frameworks for multi-species atmospheric VOC concentrations, particularly in high-temporal-resolution prediction scenarios.

Here, we conducted long-term online atmospheric observations over 5 years (2019–2023) using instruments such as online GC-MS and PTR-ToF-MS. Based on these observational data, we have proposed a machine learning-based model called the ACTE. This model aims to surpass current instrument limitations by enabling high temporal resolution, effective, and accurate computation of a more comprehensive range of VOC species, particularly alkenes, alkanes, and aromatic hydrocarbons, most of which exhibit low proton affinity. Our research results show that this model not only significantly enhances the temporal resolution of multiple VOC species, achieving a level of five-minute, but also greatly improves the ability to capture short-term variations and peak values of different VOC species. Furthermore, it can provide strong support for more refined pollution source analysis and high-resolution calculations of atmospheric chemical model, further enhancing our understanding of the chemistry of atmospheric VOCs and related secondary pollution mechanisms.

2. Materials and Methods

2.1. Data Sources and Processing

This study is based on comprehensive observation data from multiple instruments over various periods at two different sites. The first site is the Peking University Atmospheric Observation Supersite (PKU-AOSS), located in Nanshan District, Shenzhen, China, which captures urban pollution characteristics. The second site is the Yang Mei Keng (YMK) station, also located in Shenzhen City, downwind of the Dapeng Peninsula in southern China, reflecting regional characteristics. The exact geographic coordinates and elevations of the two sites are as follows: PKU-AOSS (113.98°E, 22.60°N, altitude: 14 m) and YMK (114.60°E, 22.55°N, altitude: 35 m), as shown in Figure S1 in Supporting Information S1. The sampling period at PKU-AOSS was primarily divided into two consecutive phases: from 12 September 2020, to 18 November 2020, and from 6 September 2021, to 13 December 2023, totaling 897 days. The observation period at the YMK station mainly spanned from 22 September 2019, to 30

October 2019, totaling 38 days. The overall data collection timeframe extended across 5 years from 2019 to 2023, with the actual effective continuous sampling and data acquisition period exceeding 2 years.

The observations at the two sites in this study primarily utilized three categories of instruments. The first category is VOCs online monitoring instruments, including GC-MS and PTR-ToF-MS. At both sites, the same model PTR-ToF-MS (6000X2, Ionicon Analytik GmbH, Innsbruck, Austria) was used. However, the GC-MS instruments differed between the sites. At the PKU-AOSS site, the TH-300B online GC-MS (7820A-5977E, Agilent Technologies, Inc., USA) was used, while at the YMK site, the ZF-PKU-VOC1007 VOCs online monitoring system was employed. The ZF-PKU-VOC1007 operates on similar principles and sampling analysis processes as the TH-300B, with the main difference being the use of the Shimadzu GCMS-QP2010 SE analysis system. Both systems use DB-624 chromatography columns and employ FID and MS methods for VOC analysis.

GC-MS and PTR-ToF-MS differ significantly in their underlying principles, usage, performance, and advantages. Both the TH-300B and ZF-PKU-VOC1007 are equipped with an ultra-low temperature capture system developed by Peking University, providing good recognition capabilities for different VOC species and isomers, though with relatively low temporal resolution (approximately 1 hr). In this study, the TH-300B online GC-MS monitored 94 VOC species, including 14 aromatic hydrocarbons, nine oxygenated volatile organic compounds (OVOCs), 32 halogenated hydrocarbons, 27 alkanes, and 12 alkenes and alkynes, while the ZF-PKU-VOC1007 monitored 90 VOC species, four fewer than the TH-300B. In contrast, the 6000X2 PTR-ToF-MS is capable of detecting hundreds of VOC species at minute-level resolution. However, in this study, we focused on 21 VOC species observed with this instrument. These species were selected based on the availability and representativeness of their corresponding calibration standard gases, ensuring accurate quantification. The selected VOCs primarily include oxygenated volatile organic compounds (OVOCs) and some aromatic hydrocarbons, which generally have high proton affinities, such as CH_3OH (Methanol), CH_3CHO (Acetaldehyde), C_2H_5OH (Ethanol), and C_6H_6 (Benzene), among others. Calibration was performed using the standard gas for each species. A complete list of the 21 VOC species can be found in Table S1 in Supporting Information S1, with further details provided therein.

The second category comprises conventional pollutant monitoring instruments, including O_3, nitrogen oxides (NOx, NO, NO_2), sulfur dioxide (SO_2), and $PM_{2.5}$, monitored using various instruments from Thermo Fisher Scientific. The third category includes photochemical and meteorological data monitoring instruments, measuring the photolysis rate constants for O_3 and NO_2 using a Photolysis Spectrometer (PFS-100, Focused Photonics, Inc., China), with other surface meteorological data monitored using a weather monitoring instrument (WXT520, Vaisala, Inc., Finland).

For a detailed list of different online monitoring data, please refer to Table S1 in Supporting Information S1. Additionally, detailed introductions to the analysis methods for online observation equipment and information on quality control and assurance can be found in our previous studies (Z.-J. Li et al., 2024; Xia et al., 2023; Zhu et al., 2021).

The computational workflow is illustrated in Figure 1, where part 1a describes the data collection and preprocessing phases. Initially, we collected VOCs data, conventional pollutant data, photolysis rate constants, and meteorological data, which serve as the foundation for subsequent model training and validation. To ensure high data quality, we conducted rigorous checks and cleaning during the preprocessing stage, primarily addressing missing values and outliers. For handling missing values, to ensure the model learns from the most accurate and complete data, we deleted the rows containing missing values in the relevant feature data. This step was taken to prevent missing data from negatively affecting the model training process while retaining as much complete data as possible to ensure the model's representativeness. Regarding outliers in the observed data, we implemented stricter management measures. Specifically, during routine instrument observations, we conduct daily reports and monitor the data, promptly recording any potential anomalies. All anomalous data are discussed and addressed in bi-weekly data review meetings. This process includes analyzing the source of the outliers, such as whether they were caused by instrument malfunctions or environmental factors, and making corrections or exclusions when necessary. Through this approach, we ensure high standards of data quality, while also enhancing the accuracy and reliability of the data. These data quality control measures are not only critical for model training but also provide robust data support for subsequent air quality decision-making analyses.

As a result of these preprocessing steps, we created two data sets with different temporal resolutions. The first data set is an hourly scale data set, which includes VOC species data required for the model, conventional pollutant

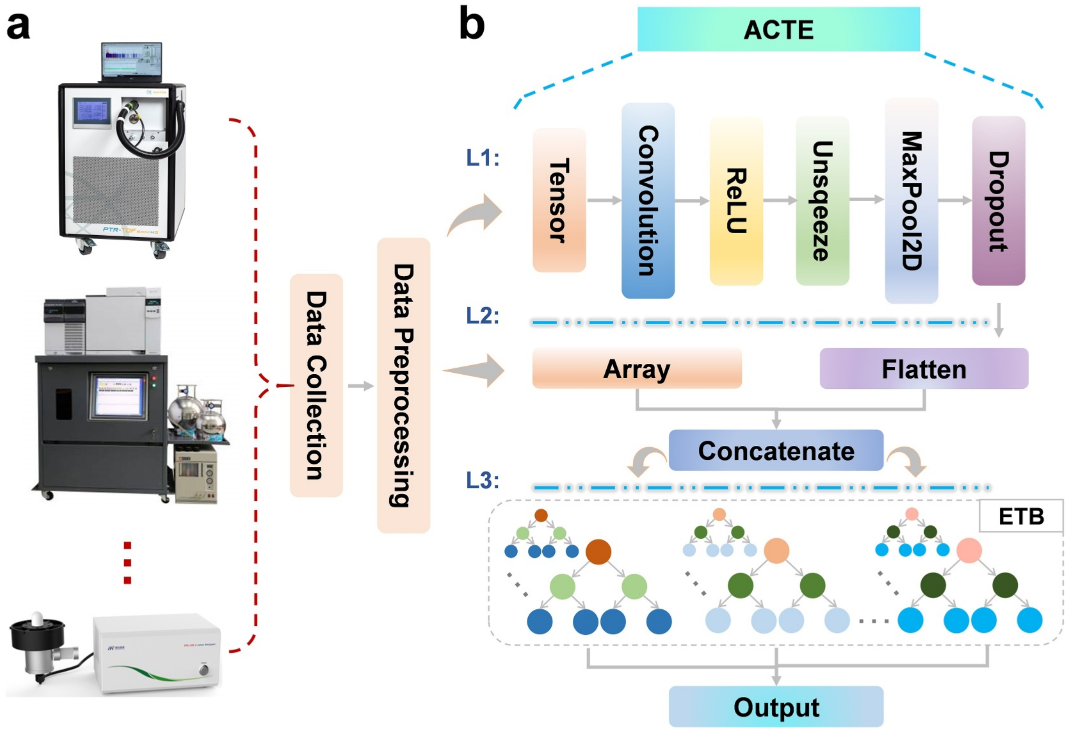

Figure 1. Computational workflow diagram. (a) Data collection and preprocessing. (b) The ACTE model architecture diagram.

data, photolysis rate constants, and meteorological data. It is important to note that due to the limitations of the GC-MS instrument, although most studies consider its resolution to be 1 hr, this is not entirely accurate. The actual sampling time is only a few minutes, with the remaining time required for processes such as sample analysis and system cleaning. In this study, the GC-MS sampling settings at the YMK site in 2019 and the PKU-AOSS site in 2020 were set to collect atmospheric samples for 5 min every hour. From 2021 to 2023, the GC-MS sampling settings at the PKU-AOSS site were adjusted to collect atmospheric samples for 10 min every hour. Consequently, we used half of the 2020 GC-MS data and the continuous data from the PKU-AOSS site from 2021 to 2023 for model training and validation. The other half of the 2020 GC-MS data and the data from the YMK site were used for model testing. This strategy ensures that the models can effectively learn the data characteristics at both 10- and 5-min scales and provides a robust evaluation of the model's accuracy and generalization ability on two independent data sets.

Additionally, since the algorithm proposed in this study requires high-quality training data, any hour with missing parameter data was excluded from the data set. Due to maintenance, calibration, and malfunctions, discrepancies among different instruments led to the exclusion of a considerable portion of data. The second data set is a 5-min-scale data set, primarily used for actual model predictions. This includes minute-level VOC species data monitored by PTR-ToF-MS, minute-level conventional pollutant data, photolysis data, and meteorological data.

2.2. Overview of the ACTE Model

In the field of data prediction, commonly used statistical models can be broadly categorized into classical regression models and machine learning models. A key representative of classical regression models is multiple linear regression, which is known for its simplicity, ease of use, and computational efficiency. It performs well in scenarios where linear relationships are prominent (Burke et al., 2023). However, linear regression tends to perform poorly when dealing with complex nonlinear relationships, as it fails to capture the intricate patterns within the data, which in turn impacts prediction accuracy (Wei et al., 2019). In contrast, machine learning models, such as decision trees, random forests, and deep neural networks, are particularly well-suited for handling complex nonlinear relationships (Bi et al., 2023; Cheng, Huang, et al., 2023). These models are especially effective in capturing the intricate interactions between air pollutants and other environmental factors (Wei et al., 2023; Zhu et al., 2024). Although these models offer significant advantages in various applications, a well-established framework for quickly selecting the best model and optimizing features for optimal performance, particularly in high-temporal-resolution air pollution prediction, remains lacking.

To address this gap, this study proposes a novel approach: the ACTE model, which enables rapid feature selection for VOC prediction and ensures high-quality prediction stability. The proposed ACTE model combines the advantages of convolutional neural networks and tree-based models, effectively integrating prior empirical feature parameters to achieve high-precision predictions for different VOCs with maximum adaptability. The model primarily consists of three layers: (a) the convolutional pooling layer, (b) the data reconstruction layer, and (c) the prediction output layer. Each layer serves distinct functions and responsibilities, outlined as follows: The input data are first normalized and converted into tensor format. The convolutional layer processes this data using filters with specified sizes, strides, and padding strategies, sliding across the data to extract features from various positions. This layer employs the Rectified Linear Unit (ReLU) activation function, followed by a max pooling operation that samples these features, selecting the maximum value within each local region primarily to reduce data dimensions and enhance the model's insensitivity to input variations. After pooling, a dropout layer is introduced to randomly discard some neuronal outputs, thereby mitigating the risk of overfitting. The features processed by the pooling layer are then flattened into a two-dimensional vector, aimed at transforming the convolutional and pooling layer-extracted features into a format suitable for fully connected layers or other data processing formats.

In this study, the feature vectors extracted by the convolutional layer are reconstructed with the original optimal feature set to create a reshaped feature data set, which is then used for result prediction output through an Ensemble Tree-Based (ETB) model. The original optimal feature set consists of the top-ranked features in terms of importance, determined through pre-training validation with the ETB model. The model in this study is configured to randomly discard 3–5 unimportant features based on coefficient of determination (R^2). Additionally, the ETB model is an optimized version of a tree ensemble model previously developed in our research, distinguished by its ability to leverage different tree models' predictive advantages based on the data features of various data sets and to determine the best model for prediction. For more detailed information about the model, please refer to our previous studies (Cheng, Huang, et al., 2023; Cheng, Peng, et al., 2024) and Text S1.2 in Supporting Information S1.

To more accurately assess the performance and generalization ability of the ACTE model, this study uses half of the valid data from the Peking University Atmospheric Observation Supersite (PKU-AOSS) site in 2020 and all the valid data from the Yang Mei Keng (YMK) site in 2019 as two independent test sets. The remaining data are used for the model's training and validation. Detailed descriptions of the observation sites, instruments, and data sets can be found in Section 2.1. Moreover, to minimize the impact of model randomness on the results, grid search (see Text S1.1 in Supporting Information S1) and 10-fold cross-validation are employed for each VOC species to determine the optimal model parameters. Furthermore, various input data sources were used in this study, including conventional air pollutant observation data (e.g., O_3 and NOx), meteorological data (e.g., temperature and humidity), and photolysis data (e.g., j[O^1D] and j[NO_2]), all of which play significant roles in the variation and dynamics of atmospheric pollution concentrations. More importantly, the model also utilized high-resolution data of 21 VOC species observed by the calibrated PTR-ToF-MS instrument, which provided richer information to support the model.

All the aforementioned species data were processed to match the same time period and resolution format as the 94 VOC species data monitored by GC-MS, and then used for model training, validation, and testing. Ultimately, this study established 94 ACTE models, each corresponding to a different VOC species. To evaluate the performance of the ACTE model, it was compared with other mainstream models, including Extreme Gradient Boosting (XGB), LightGBM model (LGBM), CatBoost (CB) model, Random Forest (RF) model, and Deep Neural Networks (DNN) model (X. Li et al., 2024; Wen et al., 2024; Zhu et al., 2024). All these comparison models were trained using the same strategies as the ACTE model, and each model had 94 corresponding VOC models for comparison. Finally, the optimal model will perform high-resolution, continuous predictions for the 94 VOC species (at 5-min intervals) using input data at this continuous resolution, thereby overcoming the limitations of traditional monitoring.

For a detailed description of the comparison models, the basic feature variables (input data), and key parameter settings, please refer to Texts S1.2–S1.4 and Tables S1 and S2 in Supporting Information S1. The machine

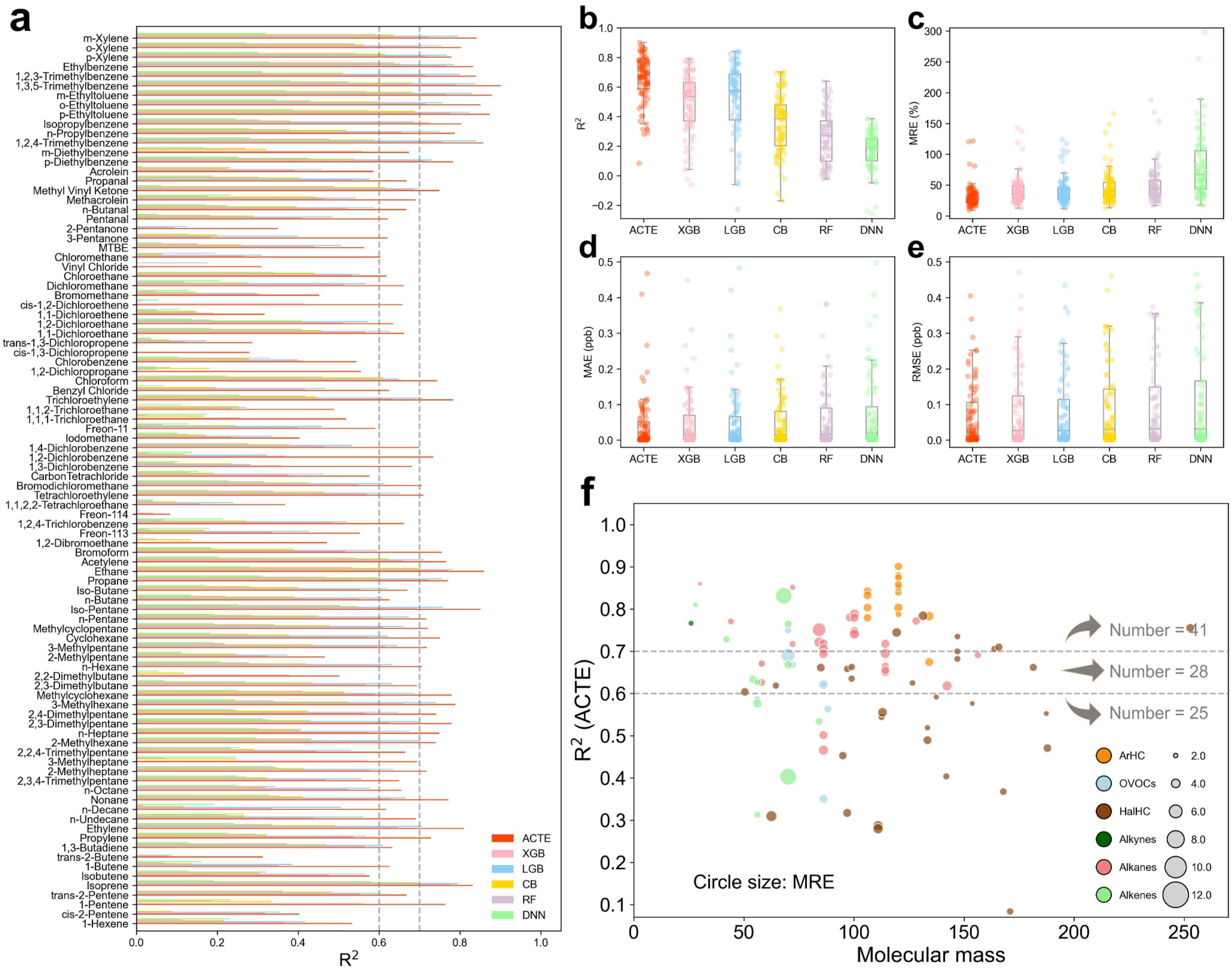

Figure 2. Model performance evaluation at the PKU-AOSS site. (a) Comparison of R^2 metrics for different VOC species across various models. (b) Box scatter plot distribution of R^2 values across different models. (c) Box scatter plot distribution of *MRE* values across different models. (d) Box scatter plot distribution of *MAE* values across different models. (e) Box scatter plot distribution of *RMSE* values across different models. (f) Bubble plot illustrating ACTE model performance for different VOC species, with bubble size representing MRE. The chart shows R^2 values plotted against VOC molecular mass, with circles colored according to VOC type, including alkanes, alkenes, aromatic hydrocarbons (ArHC), OVOCs, and halogenated hydrocarbons (HalHC).

learning models are primarily implemented using Python 3.8 on the Anaconda 5.2 platform, with the deep learning processes computed primarily using the PyTorch (CUDA) framework.

3. Results

3.1. Comparative Evaluation of Multiple Model Performances

To comprehensively evaluate the performance of the ACTE model, the assessment is based on four key metrics: coefficient of determination (R^2), root mean square error (*RMSE*), mean absolute error (*MAE*), and mean relative error (*MRE*) (Cheng, Zhu, et al., 2021). For detailed descriptions of these metrics, please refer to Text S1.5 in Supporting Information S1. Figure 2 shows the evaluation of model performance at the PKU-AOSS site, including R^2 metrics for different VOC species (a total of 94 species) across various models; other metrics are shown in Figures S2–S4 in Supporting Information S1. Additionally, Figures 2b–2e present a comprehensive view of these models' performance across various metrics in the form of box scatter plots. Combining the analysis from the above figures, it can be observed that the ACTE model consistently ranks among the top performers in predicting individual VOC species, and on the box scatter plots, it exhibits a tighter and higher distribution of values in the R^2 metric, indicating better fitting and consistency in predicting VOC concentrations. Furthermore, the distribution of *MAE* and *RMSE* metrics in Figures 2d and 2e further demonstrates that the ACTE model surpasses others in prediction accuracy, with relatively lower *MRE* (Figure 2c).

In Figure 2f, we employed two metrics, R^2 and *MRE*, to further investigate the capability of the ACTE model in predicting concentrations of various VOC species. The results show that the ACTE model maintains a commendable level of predictive accuracy across VOCs of varying molecular masses. Specifically, for 41 VOCs, the R^2 value exceeds 0.7, with approximately 77% having an R^2 greater than 0.6, significantly outperforming several other models (see Figure S5 in Supporting Information S1). Additionally, most species exhibit low errors in terms of the *MRE* metric. Notably, the ACTE model excels in predicting key categories such as alkanes, alkenes, and aromatic hydrocarbons. These categories are typically difficult to observe at high resolution (minute-scale) due to either low proton affinity or the presence of isomers, yet the ACTE model demonstrates good predictive capability for most of these species. Among these, the top five performing species include 1, 3, 5-Trimethylbenzene ($R^2 = 0.90$), m-Ethyltoluene ($R^2 = 0.88$), p-Ethyltoluene ($R^2 = 0.87$), Ethane ($R^2 = 0.86$), and 1, 2, 4-Trimethylbenzene ($R^2 = 0.86$), all of which exhibit the highest R^2 values, showcasing the model's good performance in accurately predicting VOC concentrations. It is also worth noting that the ACTE model is less effectiveness in predicting certain VOC species, particularly some halogenated hydrocarbons. We speculate that there may be two main reasons for this: first, the current model training parameters may not effectively cover the prediction of all VOC species, which could result in significant prediction errors for some VOCs; second, the inherent characteristics of certain VOC species, such as the strong stability and unique photochemical behavior of halogenated hydrocarbons, may reduce the model's accuracy in predicting their concentration changes (Orkin et al., 2020; Orkin & Khamaganov, 1993).

In addition to the local evaluation at the PKU-AOSS site, we also conducted off-site testing at the YMK site. The results from the YMK site are shown in Figures S6–S10 in Supporting Information S1. By comparing the predictive performance of different models on these two independent test sets, we further validated the spatial generalization capability of the ACTE model. Although the ACTE model's prediction accuracy and spatial generalization remain superior to several other models, a noticeable decrease in prediction accuracy for most VOC species was observed compared to the local PKU-AOSS site. While approximately half of the VOC species still met the prediction standards, the significant challenges posed by differing geographical environments cannot be ignored. Therefore, in practical predictive applications, incorporating off-site data into the model for incremental training can be an effective strategy. However, achieving satisfactory predictive results requires substantial baseline model training to be conducted in advance.

3.2. Performance of the ACTE Model Across Different Major VOC Categories

Apart from performance on individual VOC species, we are also concerned about the overall predictive performance of the model across different major VOC categories. Figure 3 and Figure S11 in Supporting Information S1 provide detailed insights into the predictive capabilities of the ACTE model for various major VOC categories: aromatic hydrocarbons (ArHC), oxygenated VOCs (OVOCs), halogenated hydrocarbons (HalHC), alkynes, alkanes, and alkenes. The results show that the top three performing species are alkanes, alkenes, and aromatic hydrocarbons, with high R^2 values (0.92, 0.89, and 0.86, respectively), and low *RMSE* (0.23, 0.10, and 0.10 ppb), *MRE* (34.35%, 42.71%, and 29.44%), and *MAE* (0.08, 0.04, and 0.03 ppb), indicating the ACTE model's high accuracy and consistency in handling these VOC categories, especially in predicting alkanes and alkenes, where the ACTE model has almost reached an ideal prediction level. The next best-performing categories are halogenated hydrocarbons, OVOCs, and alkynes, with an average R^2 of 0.80, *RMSE* of 0.19 ppb, *MRE* of 24.90%, and *MAE* of 0.08 ppb. Although their performance is slightly lower than that of the top three categories, they still showing a high level of predictive performance. Additionally, Figure 3g illustrates that the ACTE model demonstrates considerable stability and accuracy in predicting these major categories of species, with the box plot distributions of predicted values closely aligning with actual values.

At the YMK site, the performance of the ACTE model in predicting various major VOC categories exhibited some differences compared to the PKU-AOSS site (see Figures S12 and S13 in Supporting Information S1). The model performed relatively better for alkanes, alkynes, and aromatic hydrocarbons at the YMK site, achieving R^2 values of 0.81, 0.78, and 0.75, respectively, along with *RMSE* values of 0.23, 0.19, and 0.05 ppb. However, the other three categories, particularly OVOCs and halogenated hydrocarbons, performed slightly worse. Overall, in

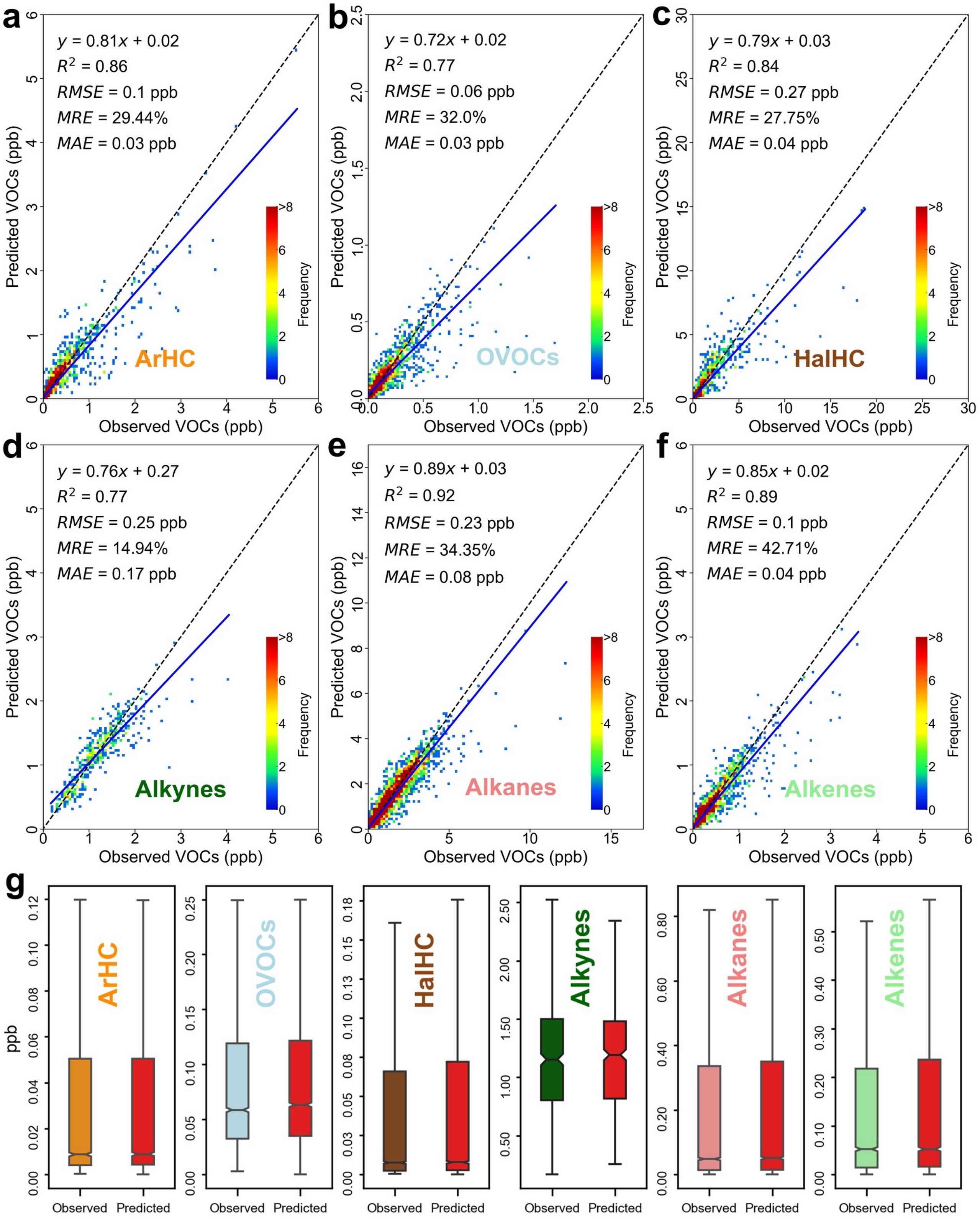
a
y = 0.81x + 0.02
R^2 = 0.86
RMSE = 0.1 ppb
MRE = 29.44%
MAE = 0.03 ppb
ArHC
b
y = 0.72x + 0.02
R^2 = 0.77
RMSE = 0.06 ppb
MRE = 32.0%
MAE = 0.03 ppb
OVOCs
c
y = 0.79x + 0.03
R^2 = 0.84
RMSE = 0.27 ppb
MRE = 27.75%
MAE = 0.04 ppb
HalHC
d
y = 0.76x + 0.27
R^2 = 0.77
RMSE = 0.25 ppb
MRE = 14.94%
MAE = 0.17 ppb
Alkynes
e
y = 0.89x + 0.03
R^2 = 0.92
RMSE = 0.23 ppb
MRE = 34.35%
MAE = 0.08 ppb
Alkanes
f
y = 0.85x + 0.02
R^2 = 0.89
RMSE = 0.1 ppb
MRE = 42.71%
MAE = 0.04 ppb
Alkenes
Predicted VOCs (ppb)
Observed VOCs (ppb)
Frequency
>8
g
ppb
ArHC
OVOCs
HalHC
Alkynes
Alkanes
Alkenes
Observed
Predicted

Figure 3.

comparison with the PKU-AOSS site, except for an improvement in alkynes prediction accuracy, the model's accuracy for other major categories decreased, highlighting the reduced spatial generalization issue due to differences in environmental conditions and source characteristics between the two locations.

In comparison to predictions for individual VOC species, the ACTE model demonstrated relatively higher accuracy and stability in predicting major categories. The model not only maintained prediction accuracy for alkanes, alkenes, and aromatic hydrocarbons but also showed improvements in other categories, especially at the more thoroughly trained PKU-AOSS site. Nevertheless, it is important to emphasize that the model's spatial generalization still has significant room for improvement, particularly in environments with markedly different geographical conditions.

3.3. Presentation of High Temporal Resolution Predictive Outcomes and Their Application in Mechanistic Model

In the preceding analysis, the ACTE model demonstrated its performance in predicting concentrations of various VOCs. The following section will focus on showcasing the model's capability in high temporal resolution VOC concentration prediction within real-world scenarios and its potential for practical application. The actual observed data from the first 5 min of each hour (commonly referred to as hourly resolution data), collected from PKU-AOSS, covers the period from 12 September to 18 November 2020. Our independent test set is primarily derived from this period. Therefore, it is appropriate to compare the observed data from this period with the predicted 5-min resolution data. However, due to factors such as instrument malfunctions or adverse weather conditions, some periods have significant data gaps, particularly in the first half of this period, where data continuity is poorer. Given that the focus of this study is on high-precision prediction using existing data segments, the missing data segments were excluded from the prediction. Regarding the zoomed-in period from October 29 to October 31, this timeframe was chosen as it includes significant concentration peaks for various major VOC categories. Considering the inherent challenges of capturing instantaneous high values, focusing on this period allows for a direct evaluation of the model's ability to predict rapid changes and extremes. This approach also highlights the strengths of the model as well as areas that require improvement, providing a more comprehensive and objective performance evaluation. Therefore, we selected data segments with better continuity and higher representativeness for comparison and presentation.

As shown in Figure 4, the ACTE model not only captures the diurnal trends in the concentrations of major species categories but also more accurately portrays concentration dynamics over shorter timescales. Figure 4a displays the results based on 1-hr actual observation data and 5-min resolution predictions, clearly demonstrating the model's responsiveness to rapid changes. The transition from 1-hr to 5-min resolution significantly increases the data volume, revealing more detailed information. However, when comparing the actual results, particularly in the magnified concentration changes (Figure 4b), the model performed relatively better in capturing the instantaneous peaks and troughs of alkanes, alkenes, and alkynes, while it was somewhat less effective for halogenated hydrocarbons and OVOCs. Nonetheless, the model's overall performance was commendable and met expectations. It is important to note that achieving highly accurate predictions of extreme values in nonlinear relationships remains a significant challenge across various fields, even with today's advances in computational power (Bi et al., 2023).

We selected typical VOC species from various major categories, including ethane, m-xylene, ethylene, acetylene, acrolein, and chloroethane, each characterized by relatively low proton affinity, for high temporal resolution predictive demonstrations. As shown in Figure S14 in Supporting Information S1, the model demonstrates good predictive accuracy for most typical individual VOC species, effectively capturing the daily fluctuation patterns of VOC concentrations, particularly for low proton affinity species such as alkanes and alkenes. However, challenges remain in predicting the extreme values for certain species, especially halogenated hydrocarbons.

Figure 3. Performance comparison of the ACTE model for different major categories of VOCs at the PKU-AOSS site. (a) Aromatic hydrocarbons (ArHC). (b) OVOCs. (c) Halogenated hydrocarbons (HalHC). (d) Alkynes. (e) Alkanes. (f) Alkenes. Each data point represents the correspondence between observed VOC concentrations and those predicted by the ACTE model. The color depth reflects the density of data points, with more intense colors indicating higher frequency. The black dashed line represents the line of perfect prediction consistency, while the blue trend line shows the general pattern of predictions by the ACTE model. (g) Box plot comparison of actual versus predicted values for different categories of VOCs. The box represents the interquartile range (IQR), spanning from the 25th to the 75th percentile, while the whiskers indicate the range within 1.5 times the IQR. The central line denotes the median.

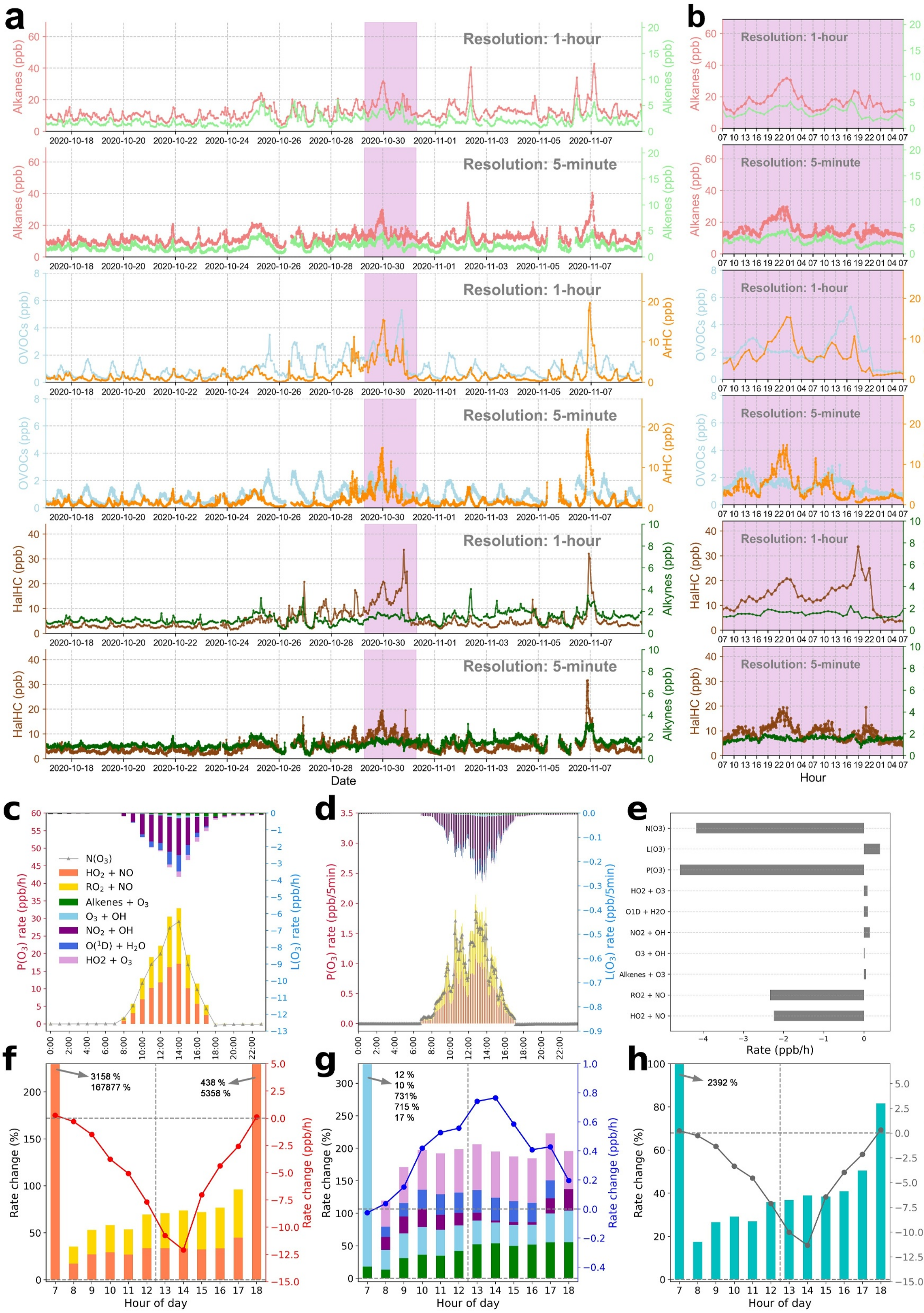
a
Resolution: 1-hour
Alkanes (ppb)
Alkenes (ppb)
Resolution: 5-minute
OVOCs (ppb)
ArHC (ppb)
HalHC (ppb)
Alkynes (ppb)
2020-10-18
2020-10-20
2020-10-22
2020-10-24
2020-10-26
2020-10-28
2020-10-30
2020-11-01
2020-11-03
2020-11-05
2020-11-07
Date
b
07 10 13 16 19 22 01 04 07 10 13 16 19 22 01 04 07
Hour
c
N(O_3)
HO_2 + NO
RO_2 + NO
Alkenes + O_3
O_3 + OH
NO_2 + OH
O(^{1}D) + H_2O
HO2 + O_3
P(O_3) rate (ppb/h)
L(O_3) rate (ppb/h)
d
P(O_3) rate (ppb/5min)
L(O_3) rate (ppb/5min)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
e
N(O3)
L(O3)
P(O3)
HO2 + O3
O1D + H2O
NO2 + OH
O3 + OH
Alkenes + O3
RO2 + NO
HO2 + NO
Rate (ppb/h)
f
3158 %
167877 %
438 %
5358 %
Rate change (%)
Rate change (ppb/h)
Hour of day
g
12 %
10 %
731%
715 %
17 %
Rate change (%)
Rate change (ppb/h)
Hour of day
h
2392 %
Rate change (%)
Rate change (ppb/h)
Hour of day

Figure 4.

Given the current limitations of instrumental detection technology in providing real-time VOC species monitoring at a 5-min resolution (particularly for species with low-proton-affinity or isomers), the ACTE model is capable of achieving continuous high-resolution monitoring of VOC species under more challenging environmental conditions, providing a more continuous and detailed data stream. This capability provides scientists and policymakers with greater opportunities to capture transient pollution events that might be overlooked at an hourly resolution, particularly those involving instantaneous emissions primarily composed of alkanes, thereby offering a basis for implementing more targeted control measures.

Additionally, the availability of more comprehensive high-resolution VOCs data allows for more accurate mechanistic modeling of related secondary pollution. In previous studies, due to instrumental limitations, researchers typically modeled the ozone mechanism model based on hourly scales (Tan et al., 2018; Xiong et al., 2023). However, daytime photochemical reactions occur very rapidly, and hourly scales clearly cannot meet more accurate simulation requirements (Lyu et al., 2023). In this study, we used both hourly and 5-min VOCs data, with the same parameter settings, to model ozone for 30 October 2020, based on the Master Chemical Mechanism (MCM v.3.3.1) (Jenkin et al., 2003; Wolfe et al., 2016). The main reason for choosing this day for modeling was its relatively stable wind speed and meteorological conditions, as well as the higher VOCs concentrations (Figures 4a and 4b). To account for the physical loss of ozone, we adjusted a first-order dilution constant (kdill) to fit the ozone concentration measurements for the simulated days (Rickly et al., 2023; Womack et al., 2023). The dilution constant derived from the simulations is approximately 1.9×10^{-5} s^{-1}, with a ±20% fluctuation, which was included as the uncertainty range in the model (Rickly et al., 2023; Womack et al., 2023). After applying the same parameters and settings, we compared the simulation results with the observed ozone concentrations at both hourly and 5-min resolutions (Figure S15 in Supporting Information S1). The comparison revealed that while the model generally captured the ozone concentration trends, notable differences were observed in peak concentrations and the timing of ozone formation. At the 5-min resolution, the model predictions aligned more closely with observed rapid fluctuations, especially during peak hours. In contrast, the hourly resolution overestimated ozone concentrations, likely due to the smoothing of high-frequency fluctuations in the ozone levels. Additionally, we compared the key reaction rates for ozone production and loss at both resolutions. Figures 4c and 4d show the variations in these rates, providing further insights into how the model's performance varies with different time scales (For more details on the mechanistic model, see Text S1.6 in Supporting Information S1).

Data at the 5-min resolution revealed more complex and dynamic changes in reaction rates, offering a detailed reflection of the rapidly occurring photochemical processes. These intricacies are significantly simplified in 1-hr resolution data, potentially leading to misunderstandings of these critical processes. By calculating the differences in ozone production and loss rates at two different resolutions (where 5-min scale simulation results are summed to obtain hourly results, which are then compared with hourly scale simulations), as shown in Figures 4e–4h, we found that traditional hourly scale mechanism modeling in the area often overestimates daytime ozone production and loss reaction rates. Specifically, the average overestimations during the day were approximately 4.6 and 0.4 ppb/h, respectively. Furthermore, we observed that the peak and trough of the differences between the two resolutions occurred around 14:00, when the difference was greatest, with the differences in ozone production and loss rates reaching approximately 12.1 and 0.8 ppb/h, respectively. Since the ozone production rate is a good measure of atmospheric oxidizing capacity (Tan et al., 2024), these significant differences suggest that hourly resolution mechanism modeling may significantly overestimate the actual atmospheric oxidizing capacity, especially around 14:00. It can be inferred that, in the investigation of the discrepancy between the OH radical model simulated concentrations and actual observations, in addition to potential missing sources or unclear

Figure 4. High temporal resolution prediction results for different categories of VOCs and their application in zero-dimensional box model based on the master chemical mechanism (MCM). (a) Time series comparison of VOC predictions at 1-hr and 5-min resolutions from 17 October to 8 November 2020. (b) Zoomed-in comparison of different resolution results from 29 October to 31 October 2020. (c) Changes in key chemical reaction rates for ozone production and loss at an hourly scale. (d) Changes in key chemical reaction rates for ozone production and loss at a 5-min resolution. (e) Total daily differences in chemical reaction rates under two different resolutions in the mechanistic model. (f) The 5-min scale simulation results of key ozone production reaction rates at different times of the day, which are then summed to obtain hourly results, and the changes relative to the hourly scale simulation. (g) The 5-min scale simulation results of key ozone loss reaction rates at different times of the day, which are then summed to obtain hourly results, and the changes relative to the hourly scale simulation. (h) The 5-min scale simulation results of net ozone production reaction rates at different times of the day, which are then summed to obtain hourly results, and the changes relative to the hourly scale simulation. Please note that the hourly resolution in panels (a) and (b) actually represents the average concentration sampled by the GC-MS during the first 5 min of each hour.

mechanistic factors (Peeters et al., 2014; Yang et al., 2024), this discrepancy may also be partially attributed to the relatively low data resolution used in the model simulations. Of course, as high-resolution modeling progresses, the gap between observed data and model simulations may become more pronounced. This further underscores the importance of high-resolution VOCs data.

Overall, this study not only demonstrates the ACTE model's effectiveness in enhancing the temporal resolution of VOCs data but also, through its application in the mechanistic model, highlights the crucial role of high-resolution environmental data in improving the accuracy and reliability of model predictions. These findings have important implications for environmental science research and practical applications.

4. Discussion and Conclusion

Our research has made significant contributions to achieving more comprehensive monitoring of high-resolution (5-min) VOCs data and to further understanding the formation of related secondary pollution after considering the current limitations of different ambient VOCs online monitoring instruments and the complexity of data for prediction. We recognize that obtaining comprehensive VOC data at the minute level, covering a wide range of species, is a challenging and complex task. Firstly, due to the inherent differences in the properties of various VOC species, the sensitivity and effective measurement methods of different instruments vary widely. A single instrument struggles to obtain broad coverage and high-resolution VOC data simultaneously, and the long-term, high-resolution monitoring of low proton affinity species remains a limitation of current instruments. Secondly, acquiring long-term online atmospheric VOC data is inherently challenging, not only in terms of raw data processing but also due to the demands on the instruments themselves, particularly when dealing with multiple instruments, which require more refined daily maintenance. Furthermore, there are currently no mature algorithmic strategies or models supporting high-precision VOC prediction, and research in this area is significantly lacking. Therefore, effectively addressing the current issues is a challenge.

Based on such considerations, we conducted a long-term online monitoring spanning 5 years (2019–2023) using multiple standardized instruments, and proposed a machine learning-based model, ACTE. The model is used to compute high temporal resolution VOC data in rapidly changing environments, thereby capturing more detailed and comprehensive changes of VOCs in the atmospheric environment. Such high-resolution data are crucial for scientists to understand atmospheric chemical reactions occurring within short periods in the actual atmosphere, especially when assessing the formation processes of O_3 and other related secondary pollutants. In environments with limited conditions, such as remote areas or during aerial monitoring tasks, we are not always able to deploy complete, high-precision instruments. Under such constraints, machine learning technology demonstrates immense potential by effectively utilizing available partial data to compensate for information gaps, showcasing its strong adaptability and application value in extreme or special monitoring conditions.

Furthermore, this study emphasizes that although our understanding of some complex atmospheric chemical reaction mechanisms remains incomplete, the ACTE model can still effectively utilize limited data to predict the continuous changes of other atmospheric chemical components at higher temporal resolution. This broader information on VOC component variations will be highly beneficial for researching atmospheric pollution mechanisms. Such capabilities not only improve the precision of our atmospheric pollution event simulations but also open up new avenues for exploring the mechanisms underlying pollution formation. The application of this prediction strategy and method extends beyond enhancing existing monitoring networks and could help in uncovering new potential reaction pathways or key reactive species in atmospheric chemical reactions that have not yet been fully decoded. However, we fully recognize that, despite the good performance of the ACTE model in the aforementioned aspects, there are still areas that require further improvement and refinement, such as measurement uncertainties, particularly those related to instrument calibration and variations in VOC species (Coggon et al., 2024), may affect the model's performance. While we have considered and optimized uncertainty issues in data preprocessing, model prediction strategy formulation, and model architecture design, we acknowledge that the current model has not fully accounted for the potential impacts of these measurement uncertainties. Therefore, we plan to improve the model in future work by incorporating uncertainty as an additional feature or introducing a weighted loss function, and systematically assess the potential impact of uncertainty on model performance.

Moreover, the model still faces challenges in predicting the behavior of VOCs under extreme weather conditions, which can significantly affect the reaction pathways and environmental fate of the compounds (N. Wang et al., 2022). In terms of application scenarios, although we used long-term measurement data to train the model, these training data were still obtained from a single site. During testing at the remote YMK site, it was found that when the model was applied to areas with significant geographic differences, the prediction errors for certain species increased substantially. This indicates that a model trained on data from a single site has limitations in its spatial generalization ability. Furthermore, the measurement periods at the two sites do not overlap, which also affects the model's prediction results. Generally, using data from the same time period for training helps reduce inconsistencies caused by temporal differences, thereby improving prediction accuracy. However, to maintain the consistency of the core instrument PTR-ToF-MS and ensure the model's applicability across different regions, we chose data from different locations and time periods for testing. This approach helps more accurately reflect the model's performance in real-world applications, particularly in terms of its suitability and generalization ability under different environmental conditions. Therefore, when making predictions in other regions, especially those with significant environmental differences, it is recommended to incorporate data from different environmental conditions and time periods during training, particularly by combining local data from the target region for training and validation. This method can effectively enhance the model's adaptability, reduce regional or seasonal biases, and improve its performance and prediction accuracy in practical applications. Otherwise, the model's predictive ability is likely to be constrained (Cheng, He, & Huang, 2021; Han et al., 2023; Zhong et al., 2021).

In conclusion, this research emphasizes that machine learning technology not only improves the current capability for high-resolution observation of atmospheric pollutants but also deepen our understanding of atmospheric chemical mechanisms, thereby providing a stronger scientific basis for predicting and managing atmospheric pollution. As these technologies continue to advance and be applied, the role of machine learning in atmospheric pollution monitoring and chemical mechanism research is expected to grow increasingly important, with substantial potential for practical applications.

Data Availability Statement

The primary data supporting the findings of this study, along with the core code, are available in the Zenodo open data repository (Cheng, Huang, et al., 2024). The code for the MCM box model can also be accessed from the same repository (Wolfe & Haskins, 2023). Additional detailed data are available from the corresponding author upon reasonable request.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2023YFC3705101) and the Science and Technology Plan of Shenzhen Municipality (JCYJ20220818100812028).

References

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Burke, M., Childs, M. L., dela Cuesta, B., Qiu, M., Li, J., Gould, C. F., et al. (2023). The contribution of wildfire to $PM_{2.5}$ trends in the USA. *Nature*, *622*(7984), 761–766. https://doi.org/10.1038/s41586-023-06522-6

Cappellin, L., Karl, T., Probst, M., Ismailova, O., Winkler, P. M., Soukoulis, C., et al. (2012). On quantitative determination of volatile organic compound concentrations using proton transfer reaction time-of-flight mass spectrometry. *Environmental Science & Technology*, *46*(4), 2283–2290. https://doi.org/10.1021/es203985t

Chang, X., Zhao, B., Zheng, H., Wang, S., Cai, S., Guo, F., et al. (2022). Full-volatility emission framework corrects missing and underestimated secondary organic aerosol sources. *One Earth*, *5*(4), 403–412. https://doi.org/10.1016/j.oneear.2022.03.015

Cheng, Y., He, L.-Y., & Huang, X.-F. (2021). Development of a high-performance machine learning model to predict ground ozone pollution in typical cities of China. *Journal of Environmental Management*, *299*, 113670. https://doi.org/10.1016/j.jenvman.2021.113670

Cheng, Y., Huang, X.-F., Peng, Y., Cao, L.-M., Peng, X., Wu, J., & He, L.-Y. (2024). Enhancing time resolution of ambient VOC measurement data by machine learning: From one-hour to five minutes [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.14533796

Cheng, Y., Huang, X.-F., Peng, Y., Tang, M.-X., Zhu, B., Xia, S.-Y., & He, L.-Y. (2023). A novel machine learning method for evaluating the impact of emission sources on ozone formation. *Environmental Pollution*, *316*(Pt 2), 120685. https://doi.org/10.1016/j.envpol.2022.120685

Cheng, Y., Peng, Y., Cao, L.-M., Huang, X.-F., & He, L.-Y. (2024). Identifying the geospatial relationship of surface ozone pollution in China: Implications for key pollution control regions. *Science of the Total Environment*, *930*, 172763. https://doi.org/10.1016/j.scitotenv.2024.172763

Cheng, Y., Zhu, Q., Peng, Y., Huang, X.-F., & He, L.-Y. (2021). Multiple strategies for a novel hybrid forecasting algorithm of ozone based on data-driven models. *Journal of Cleaner Production*, *326*, 129451. https://doi.org/10.1016/j.jclepro.2021.129451

Coggon, M. M., Gkatzelis, G. I., McDonald, B. C., Gilman, J. B., Schwantes, R. H., Abuhassan, N., et al. (2021). Volatile chemical product emissions enhance ozone and modulate urban chemistry. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(32), e2026653118. https://doi.org/10.1073/pnas.2026653118

Coggon, M. M., Stockwell, C. E., Claflin, M. S., Pfannerstill, E. Y., Xu, L., Gilman, J. B., et al. (2024). Identifying and correcting interferences to PTR-ToF-MS measurements of isoprene and other urban volatile organic compounds. *Atmospheric Measurement Techniques*, *17*(2), 801–825. https://doi.org/10.5194/amt-17-801-2024

Dedoussi, I. C., Eastham, S. D., Monier, E., & Barrett, S. R. H. (2020). Premature mortality related to United States cross-state air pollution. *Nature*, *578*(7794), 261–265. https://doi.org/10.1038/s41586-020-1983-8

Dimri, R., Choi, Y., Salman, A. K., Park, J., & Singh, D. (2024). AGATNet: An adaptive graph attention network for bias correction of CMAQ-forecasted $PM_{2.5}$ concentrations over South Korea. *Journal of Geophysical Research: Machine Learning and Computation*, *1*(3), e2024JH000244. https://doi.org/10.1029/2024JH000244

Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., et al. (2014). A large source of low-volatility secondary organic aerosol. *Nature*, *506*(7489), 476–479. https://doi.org/10.1038/nature13032

Ferracci, V., Weber, J., Bolas, C. G., Robinson, A. D., Tummon, F., Rodríguez-Ros, P., et al. (2024). Atmospheric isoprene measurements reveal larger-than-expected Southern Ocean emissions. *Nature Communications*, *15*(1), 2571. https://doi.org/10.1038/s41467-024-46744-4

Gkatzelis, G. I., Coggon, M. M., McDonald, B. C., Peischl, J., Aikin, K. C., Gilman, J. B., et al. (2021). Identifying volatile chemical product tracer compounds in U.S. Cities. *Environmental Science & Technology*, *55*(1), 188–199. https://doi.org/10.1021/acs.est.0c05467

Graus, M., Müller, M., & Hansel, A. (2010). High resolution PTR-TOF: Quantification and formula confirmation of VOC in real time. *Journal of the American Society for Mass Spectrometry*, *21*(6), 1037–1044. https://doi.org/10.1016/j.jasms.2010.02.006

Han, W., He, T.-L., Jiang, Z., Zhu, R., Jones, D., Miyazaki, K., & Shen, Y. (2023). The capability of deep learning model to predict ozone across continents in China, the United States and Europe. *Geophysical Research Letters*, *50*(24), e2023GL104928. https://doi.org/10.1029/2023GL104928

He, Y., Zhao, B., Wang, S., Valorso, R., Chang, X., Yin, D., et al. (2024). Formation of secondary organic aerosol from wildfire emissions enhanced by long-time ageing. *Nature Geoscience*, *17*(2), 124–129. https://doi.org/10.1038/s41561-023-01355-4

Huang, X.-F., Peng, Y., Wei, J., Peng, J., Lin, X.-Y., Tang, M.-X., et al. (2023). Microphysical complexity of black carbon particles restricts their warming potential. *One Earth*, *7*(1), 136–145. https://doi.org/10.1016/j.oneear.2023.12.004

Huang, X.-F., Zhang, B., Xia, S.-Y., Han, Y., Wang, C., Yu, G.-H., & Feng, N. (2020). Sources of oxygenated volatile organic compounds (OVOCs) in urban atmospheres in North and South China. *Environmental Pollution*, *261*, 114152. https://doi.org/10.1016/j.envpol.2020.114152

Jenkin, M. E., Saunders, S. M., Wagner, V., & Pilling, M. J. (2003). Protocol for the development of the master chemical mechanism, MCM v3 (Part B): Tropospheric degradation of aromatic volatile organic compounds. *Atmospheric Chemistry and Physics*, *3*(1), 181–193. https://doi.org/10.5194/acp-3-181-2003

Jordan, A., Haidacher, S., Hanel, G., Hartungen, E., Märk, L., Seehauser, H., et al. (2009). A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS). *International Journal of Mass Spectrometry*, *286*(2), 122–128. https://doi.org/10.1016/j.ijms.2009.07.005

Kajos, M. K., Rantala, P., Hill, M., Hellén, H., Aalto, J., Patokoski, J., et al. (2015). Ambient measurements of aromatic and oxidized VOCs by PTR-MS and GC-MS: Intercomparison between four instruments in a boreal forest in Finland. *Atmospheric Measurement Techniques*, *8*(10), 4453–4473. https://doi.org/10.5194/amt-8-4453-2015

Lamkaddam, H., Dommen, J., Ranjithkumar, A., Gordon, H., Wehrle, G., Krechmer, J., et al. (2021). Large contribution to secondary organic aerosol from isoprene cloud chemistry. *Science Advances*, *7*(13), eabe2952. https://doi.org/10.1126/sciadv.abe2952

Lee, Y., Huey, L. G., Wang, Y., Qu, H., Zhang, R., Ji, Y., et al. (2021). Photochemistry of volatile organic compounds in the Yellow River Delta, China: Formation of O_3 and peroxyacyl nitrates. *Journal of Geophysical Research: Atmospheres*, *126*(23), e2021JD035296. https://doi.org/10.1029/2021JD035296

Li, X., Ye, C., Lu, K., Xue, C., Li, X., & Zhang, Y. (2024). Accurately predicting spatiotemporal variations of near-surface nitrous acid (HONO) based on a deep learning approach. *Environmental Science & Technology*, *58*(29), 13035–13046. https://doi.org/10.1021/acs.est.4c02221

Li, Z., Ho, K.-F., & Yim, S. H. L. (2020). Source apportionment of hourly-resolved ambient volatile organic compounds: Influence of temporal resolution. *Science of the Total Environment*, *725*, 138243. https://doi.org/10.1016/j.scitotenv.2020.138243

Li, Z.-J., He, L.-Y., Ma, H.-N., Peng, X., Tang, M.-X., Du, K., & Huang, X.-F. (2024). Sources of atmospheric oxygenated volatile organic compounds in different air masses in Shenzhen, China. *Environmental Pollution*, *340*, 122871. https://doi.org/10.1016/j.envpol.2023.122871

Lindinger, W., Hansel, A., & Jordan, A. (1998). On-line monitoring of volatile organic compounds at pptv levels by means of proton-transfer-reaction mass spectrometry (PTR-MS) medical applications, food control and environmental research. *International Journal of Mass Spectrometry and Ion Processes*, *173*(3), 191–241. https://doi.org/10.1016/S0168-1176(97)00281-4

Lyu, X., Li, K., Guo, H., Morawska, L., Zhou, B., Zeren, Y., et al. (2023). A synergistic ozone-climate control to address emerging ozone pollution challenges. *One Earth*, *6*(8), 964–977. https://doi.org/10.1016/j.oneear.2023.07.004

Mellouki, A., Wallington, T. J., & Chen, J. (2015). Atmospheric chemistry of oxygenated volatile organic compounds: Impacts on air quality and climate. *Chemical Reviews*, *115*(10), 3984–4014. https://doi.org/10.1021/cr500549n

Nozière, B., Kalberer, M., Claeys, M., Allan, J., D'Anna, B., Decesari, S., et al. (2015). The molecular identification of organic compounds in the atmosphere: State of the art and challenges. *Chemical Reviews*, *115*(10), 3919–3983. https://doi.org/10.1021/cr5003485

Orkin, V. L., & Khamaganov, V. G. (1993). Determination of rate constants for reactions of some hydrohaloalkanes with OH radicals and their atmospheric lifetimes. *Journal of Atmospheric Chemistry*, *16*(2), 157–167. https://doi.org/10.1007/BF00702785

Orkin, V. L., Kurylo, M. J., & Fleming, E. L. (2020). Atmospheric lifetimes of halogenated hydrocarbons: Improved estimations from an analysis of modeling results. *Journal of Geophysical Research: Atmospheres*, *125*(16), e2019JD032243. https://doi.org/10.1029/2019JD032243

Park, J.-H., Goldstein, A. H., Timkovsky, J., Fares, S., Weber, R., Karlik, J., & Holzinger, R. (2013). Active atmosphere-ecosystem exchange of the vast majority of detected volatile organic compounds. *Science*, *341*(6146), 643–647. https://doi.org/10.1126/science.1235053

Peeters, J., Müller, J.-F., Stavrakou, T., & Nguyen, V. S. (2014). Hydroxyl radical recycling in isoprene oxidation driven by hydrogen bonding and hydrogen tunneling: The upgraded LIM1 mechanism. *The Journal of Physical Chemistry A*, *118*(38), 8625–8643. https://doi.org/10.1021/jp5033146

Pfannerstill, E. Y., Arata, C., Zhu, Q., Schulze, B. C., Ward, R., Woods, R., et al. (2024). Temperature-dependent emissions dominate aerosol and ozone formation in Los Angeles. *Science*, *384*(6702), 1324–1329. https://doi.org/10.1126/science.adg8204

Rickly, P. S., Coggon, M. M., Aikin, K. C., Alvarez, R. J., II., Baidar, S., Gilman, J. B., et al. (2023). Influence of wildfire on urban ozone: An observationally constrained box modeling study at a site in the Colorado front range. *Environmental Science & Technology*, *57*(3), 1257–1267. https://doi.org/10.1021/acs.est.2c06157

Seinfeld, J. H., & Pandis, S. N. (1998). *Atmospheric chemistry and physics: From air pollution to climate change*. Wiley.

Tan, Z., Feng, M., Liu, H., Luo, Y., Li, W., Song, D., et al. (2024). Atmospheric oxidation capacity elevated during 2020 spring lockdown in Chengdu, China: Lessons for future secondary pollution control. *Environmental Science & Technology*, *58*(20), 8815–8824. https://doi.org/10.1021/acs.est.3c08761

Tan, Z., Lu, K., Jiang, M., Su, R., Dong, H., Zeng, L., et al. (2018). Exploring ozone pollution in Chengdu, southwestern China: A case study from radical chemistry to O_3-VOC-NOx sensitivity. *Science of the Total Environment*, *636*, 775–786. https://doi.org/10.1016/j.scitotenv.2018.04.286

Wang, N., Huang, X., Xu, J., Wang, T., Tan, Z.-M., & Ding, A. (2022). Typhoon-boosted biogenic emission aggravates cross-regional ozone pollution in China. *Science Advances*, *8*(2), eabl6166. https://doi.org/10.1126/sciadv.abl6166

Wang, T., Xue, L., Brimblecombe, P., Lam, Y. F., Li, L., & Zhang, L. (2017). Ozone pollution in China: A review of concentrations, meteorological influences, chemical precursors, and effects. *Science of the Total Environment*, *575*, 1582–1596. https://doi.org/10.1016/j.scitotenv.2016.10.081

Wang, W., Li, X., Cheng, Y., Parrish, D. D., Ni, R., Tan, Z., et al. (2024). Ozone pollution mitigation strategy informed by long-term trends of atmospheric oxidation capacity. *Nature Geoscience*, *17*(1), 20–25. https://doi.org/10.1038/s41561-023-01334-9

Wang, W., Yuan, B., Peng, Y., Su, H., Cheng, Y., Yang, S., et al. (2022). Direct observations indicate photodegradable oxygenated volatile organic compounds (OVOCs) as larger contributors to radicals and ozone production in the atmosphere. *Atmospheric Chemistry and Physics*, *22*(6), 4117–4128. https://doi.org/10.5194/acp-22-4117-2022

Wang, W., Yuan, B., Su, H., Cheng, Y., Qi, J., Wang, S., et al. (2024). A large role of missing volatile organic compound reactivity from anthropogenic emissions in ozone pollution regulation. *Atmospheric Chemistry and Physics*, *24*(7), 4017–4027. https://doi.org/10.5194/acp-24-4017-2024

Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., & Cribb, M. (2019). Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach. *Remote Sensing of Environment*, *231*, 111221. https://doi.org/10.1016/j.rse.2019.111221

Wei, J., Li, Z., Chen, X., Li, C., Sun, Y., Wang, J., et al. (2023). Separating daily 1 km $PM_{2.5}$ inorganic chemical composition in China since 2000 via deep learning integrating ground, satellite, and model data. *Environmental Science & Technology*, *57*(46), 18282–18295. https://doi.org/10.1021/acs.est.3c00272

Wen, Y., Zhang, S., Wang, Y., Yang, J., He, L., Wu, Y., & Hao, J. (2024). Dynamic traffic data in machine-learning air quality mapping improves environmental justice assessment. *Environmental Science & Technology*, *58*(7), 3118–3128. https://doi.org/10.1021/acs.est.3c07545

Wolfe, G. M., & Haskins, J. (2023). AirChem/F0AM: V4.3 (v4.3) [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.8305950

Wolfe, G. M., Marvin, M. R., Roberts, S. J., Travis, K. R., & Liao, J. (2016). The framework for 0-D atmospheric modeling (F0AM) v3.1. *Geoscientific Model Development*, *9*(9), 3309–3319. https://doi.org/10.5194/gmd-9-3309-2016

Womack, C. C., Chace, W. S., Wang, S., Baasandorj, M., Fibiger, D. L., Franchin, A., et al. (2023). Midlatitude ozone depletion and air quality impacts from industrial halogen emissions in the Great Salt Lake basin. *Environmental Science & Technology*, *57*(5), 1870–1881. https://doi.org/10.1021/acs.est.2c05376

Xia, S.-Y., Huang, X.-F., Li, Z.-J., Fu, N., Jiang, Z., Cao, L.-M., et al. (2023). Seasonal variation characteristics of atmospheric peroxyacetyl nitrate (PAN) and its source apportionment in a megacity in southern China. *Science of the Total Environment*, *892*, 164662. https://doi.org/10.1016/j.scitotenv.2023.164662

Xiong, Y., Chai, J., Mao, H., Mariscal, N., Yacovitch, T., Lerner, B., et al. (2023). Examining the summertime ozone formation regime in southeast Michigan using MOOSE ground-based $HCHO/NO_2$ measurements and F0AM box model. *Journal of Geophysical Research: Atmospheres*, *128*(19), e2023JD038943. https://doi.org/10.1029/2023JD038943

Xiong, Y., Du, K., & Huang, Y. (2024). One-third of global population at cancer risk due to elevated volatile organic compounds levels. *npj Climate and Atmospheric Science*, *7*(1), 54. https://doi.org/10.1038/s41612-024-00598-1

Yang, X., Wang, H., Lu, K., Ma, X., Tan, Z., Long, B., et al. (2024). Reactive aldehyde chemistry explains the missing source of hydroxyl radicals. *Nature Communications*, *15*(1), 1648. https://doi.org/10.1038/s41467-024-45885-w

Ye, H., Du, Z., Lu, H., Tian, J., Chen, L., & Lin, W. (2022). Using machine learning methods to predict VOC emissions in chemical production with hourly process parameters. *Journal of Cleaner Production*, *369*, 133406. https://doi.org/10.1016/j.jclepro.2022.133406

Yuan, B., Koss, A. R., Warneke, C., Coggon, M., Sekimoto, K., & de Gouw, J. A. (2017). Proton-transfer-reaction mass spectrometry: Applications in atmospheric sciences. *Chemical Reviews*, *117*(21), 13187–13229. https://doi.org/10.1021/acs.chemrev.7b00325

Zhang, G., Hu, R., Xie, P., Hu, C., Liu, X., Zhong, L., et al. (2024). Intensive photochemical oxidation in the marine atmosphere: Evidence from direct radical measurements. *Atmospheric Chemistry and Physics*, *24*(3), 1825–1839. https://doi.org/10.5194/acp-24-1825-2024

Zhang, X., Schwantes, R. H., Coggon, M. M., Loza, C. L., Schilling, K. A., Flagan, R. C., & Seinfeld, J. H. (2014). Role of ozone in SOA formation from alkane photooxidation. *Atmospheric Chemistry and Physics*, *14*(3), 1733–1753. https://doi.org/10.5194/acp-14-1733-2014

Zhang, Y., Xue, L., Mu, J., Chen, T., Li, H., Gao, J., & Wang, W. (2022). Developing the maximum incremental reactivity for volatile organic compounds in major cities of central-eastern China. *Journal of Geophysical Research: Atmospheres*, *127*(22), e2022JD037296. https://doi.org/10.1029/2022JD037296

Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., et al. (2021). Machine learning: New ideas and tools in environmental science and engineering. *Environmental Science & Technology*, *55*(19), 12741–12754. https://doi.org/10.1021/acs.est.1c01339

Zhou, B., Guo, H., Zeren, Y., Wang, Y., Lyu, X., Wang, B., & Wang, H. (2023). An observational constraint of VOC emissions for air quality modeling study in the Pearl River Delta region. *Journal of Geophysical Research: Atmospheres*, *128*(11), e2022JD038122. https://doi.org/10.1029/2022JD038122

Zhu, B., Cao, L.-M., Xia, S.-Y., Niu, Y.-B., Man, H.-Y., Du, K., et al. (2023). Identifying the airport as a key urban VOC source in the Pearl River Delta, China. *Atmospheric Environment*, *301*, 119721. https://doi.org/10.1016/j.atmosenv.2023.119721

Zhu, B., Huang, X.-F., Xia, S.-Y., Lin, L.-L., Cheng, Y., & He, L.-Y. (2021). Biomass-burning emissions could significantly enhance the atmospheric oxidizing capacity in continental air pollution. *Environmental Pollution*, *285*, 117523. https://doi.org/10.1016/j.envpol.2021.117523

Zhu, L., Wang, Y., Chavas, D., Johncox, M., & Yung, Y. L. (2024). Leading role of Saharan dust on tropical cyclone rainfall in the Atlantic Basin. *Science Advances*, *10*(30), eadn6106. https://doi.org/10.1126/sciadv.adn6106

References From the Supporting Information

Liu, X., Lu, D., Zhang, A., Liu, Q., & Jiang, G. (2022). Data-driven machine learning in environmental pollution: Gains and problems. *Environmental Science & Technology*, *56*(4), 2124–2133. https://doi.org/10.1021/acs.est.1c06157

Prokhorenkova, L., Gusev, G., Vorobev, A., Veronika Dorogush, A., & Gulin, A. (2017). CatBoost: Unbiased boosting with categorical features. arXiv:1706.09516.

Tan, Z., Lu, K., Ma, X., Chen, S., He, L., Huang, X., et al. (2022). Multiple impacts of aerosols on O_3 production are largely compensated: A case study Shenzhen, China. *Environmental Science & Technology*, *56*(24), 17569–17580. https://doi.org/10.1021/acs.est.2c06217

Xing, J., Zheng, S., Ding, D., Kelly, J. T., Wang, S., Li, S., et al. (2020). Deep learning for prediction of the air quality response to emission changes. *Environmental Science & Technology*, *54*(14), 8589–8600. https://doi.org/10.1021/acs.est.0c02923

Xue, L., Wang, T., Louie, P. K. K., Luk, C. W. Y., Blake, D. R., & Xu, Z. (2014). Increasing external effects negate local efforts to control ozone air pollution: A case study of Hong Kong and implications for other Chinese cities. *Environmental Science & Technology*, *48*(18), 10769–10775. https://doi.org/10.1021/es503278g