Environmental Science & Technology Letters

pubs.acs.org/journal/estlcu

Letter

Machine Learning Reveals the Parameters Affecting the Gaseous Sulfuric Acid Distribution in a Coastal City: Model Construction and Interpretation

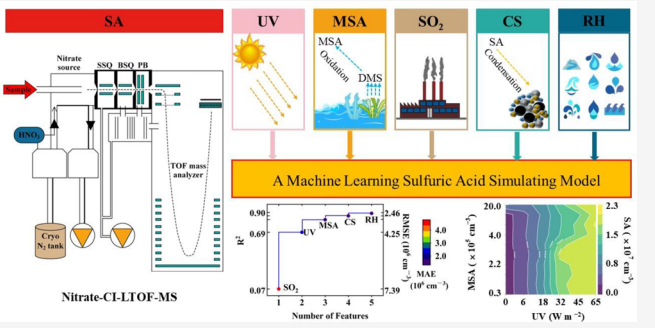
Chen Yang, Hesong Dong, Yuping Chen, Yonghong Wang, Xiaolong Fan,* Yee Jun Tham, Gaojie Chen, Lingling Xu,* Ziyi Lin, Mengren Li, Youwei Hong, and Jinsheng Chen*

Cite This: *Environ. Sci. Technol. Lett.* 2023, 10, 1045−1051

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Although the fundamental mechanisms of atmospheric new particle formation events are largely associated with gaseous sulfuric acid monomer (SA), the parameters affecting SA generation and elimination remain unclear, especially in coastal areas where certain sulfur-containing precursors are abundant. In this study, we utilized machine learning (ML) in combination with field observations to map the link between SA and the influencing parameters. The developed random forest (RF) model performed well in creating simulations with an R^2 of 0.90, and the significant factors were ultraviolet, methanesulfonic acid (MSA), SO_2, condensation sink, and relative humidity in descending order. Among the five factors, MSA served as an indicator for sulfur-containing species from marine emissions. The black box of ML was broken to determine the marginal contribution of these five parameters to the model output using partial dependence plots and centered-individual conditional expectation plots. These results indicated that MSA had a positive impact on the performance of the RF model, and a co-occurring relationship was observed between MSA and SA during the nocturnal period. Our findings reveal that sulfur-containing species emitted from the marine environment have an impact on the formation of SA and should be considered in coastal areas.



KEYWORDS: *machine learning, random forest, new particle formation, sulfuric acid, methanesulfonic acid, coastal city*

1. INTRODUCTION

The frequent occurrence of new particle formation (NPF) events promotes the formation of fine atmospheric particles, which in turn deteriorates air quality, affects the climate globally, and even harms human health.[1−5] The gaseous sulfuric acid dimer, formed from the sulfuric acid monomer (SA), is frequently involved in atmospheric cluster formation, which is also a critical initial step in NPF in various mechanisms.[6−14] Therefore, a thorough understanding of the sources and sinks of SA is essential to understanding the basic mechanisms governing atmospheric NPF.

SA proxies have the potential to explain the mechanisms of SA sources and sinks. Petäjä et al.[15] proposed the first proxy using the SO_2-OH radical process as the only source of SA and a condensation sink (CS) as the only sink in a boreal forest station. Mikkonen et al.[16] discovered that the proxy was enhanced by including relative humidity (RH) in the sink term. As an additional source of SA for the proxy, Dada et al.[17] incorporated the reaction of SO_2 with stabilized Criegee intermediates (sCIs). On the basis of a budget analysis of SA, Yang et al.[18] applied primary emission and dry deposition of SA to the proxy. These initiatives have greatly improved our understanding of SA mechanisms in diverse situations. Although current proxy methods have been shown to estimate SA concentrations, they cannot examine the marginal effects of individual (or the intersection of multiple) parameters on SA. Nevertheless, previous studies have reported that 95% of the natural sulfur emissions from the ocean can be attributed to dimethyl sulfide (DMS) emissions from seawater, and DMS can be oxidized to produce SA and methanesulfonic acid (MSA).[19−21] Although sulfur emissions from the ocean must be significant in coastal regions, no proxies or models have yet accounted for this specific process of SA formation.

Despite the multiple SA formation mechanisms, machine learning (ML) is a data-driven strategy that could minimize the relationships among complex data.[22,23] A decision tree-based integrated supervised ML system called random forest (RF)

Special Issue: Data Science for Advancing Environmental Science, Engineering, and Technology

Received: March 7, 2023
Revised: April 3, 2023
Accepted: April 3, 2023
Published: April 12, 2023

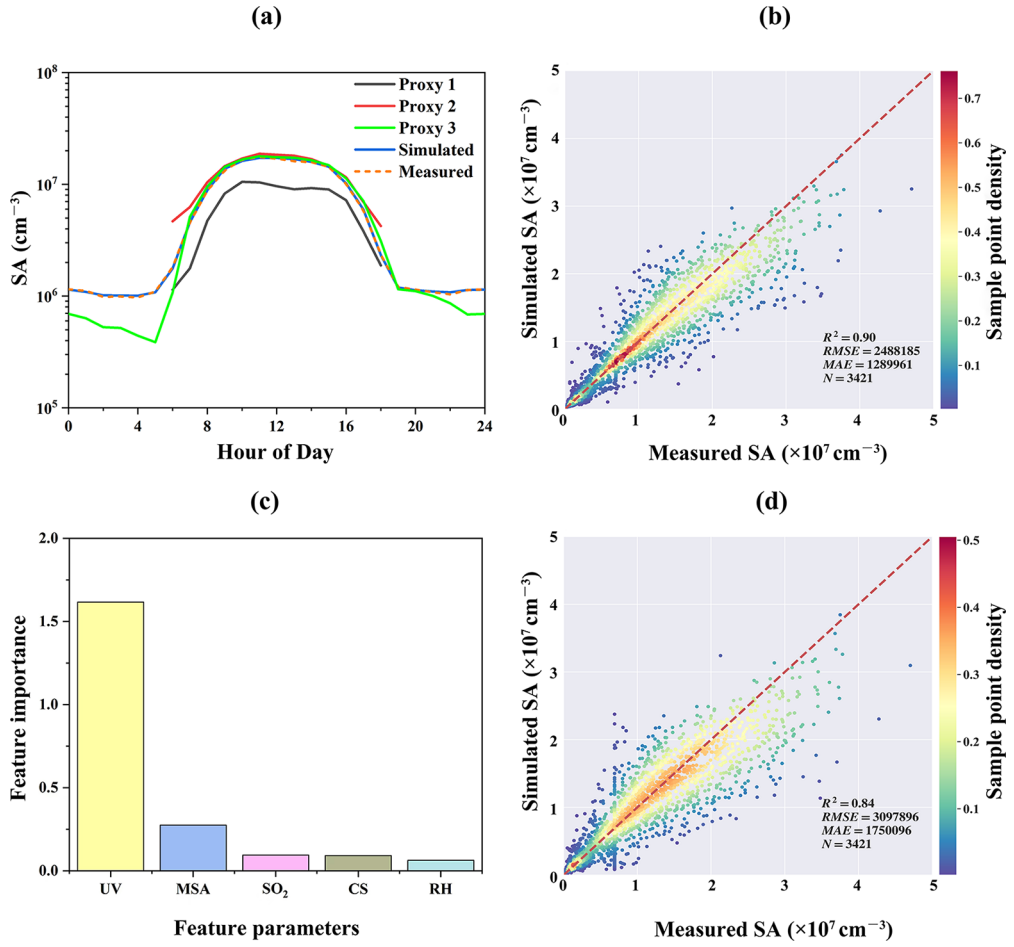


Figure 1. (a) Diurnal variation of SA (median) calculated by the three proxies, simulated by the RF model, and measured. The specific details of the three proxies are described in Text S9. (b) Simulation performance of SA in the test data set using the RF model with five feature parameters. The slope of the red dash line equals 1. There are 3421 valid data points in total, each of which is colored on the basis of sample point density. (c) Importance ranking of the five selected feature parameters. (d) Performance of the RF model obtained by applying the four feature parameters trained in addition to the MSA, where the interpretation of each metric is the same as in panel b. The slope of the red dash line equals 1.

helps identify nonlinear correlations between parameters.[24,25] RF models have been successfully used in forecasting studies of environmental pollutants, such as estimating the emission of brake wear in $PM_{2.5}$ and predicting the OH radicals near the surface in North American cities.[26,27] In this study, an RF model was applied to simulate the SA concentration and to examine the impacts of parameters. To clarify the influences of single-input and multi-input feature parameters on the model outcomes, two black-box visualization tools, partial dependence plots (PDPs) and centered-individual conditional expectation (c-ICE), were carefully scrutinized.[28−32] The development of ML-based models provides new perspectives for comprehension of SA mechanisms.

2. MATERIALS AND METHODS

2.1. Field Observation and Data Collection. The observations were carried out at the Atmospheric Environment Observation Supersite at the Institute of Urban Environment, Chinese Academy of Sciences, in the southeastern coastal city of Xiamen, China (118°03′E, 24°36′N) (Figure S1 and Text S1), during summer (from July 15 to August 23, 2022).[33,34] On the basis of the 24 h backward trajectory analysis of the HYSPLIT4 model, it can be found that the air masses during the sampling period all originated from the ocean (Figure S2).[35]

A chemical ionization atmospheric-pressure interface long time-of-flight spectrometer (CI-APi-LTOF, Aerodyne Research, Inc.) with a nitrate source was used to measure SA and MSA. The instrument configuration details and quantification of MSA and SA can be found in Text S2, and this method was previously described.[36] The identified peaks for SA and MSA are shown in Figure S3. Detailed descriptions of additional auxiliary measurements of meteorological parameters, ultraviolet (UV) radiation, trace gases, and the particle number size distribution are described in Text S3. Additionally, Text S4 details the CS calculation method.

2.2. Selection of Feature Parameters and Model Construction. On the basis of similar studies, the workflow of the RF model employed in this research is elaborately described in Text S5.[37] First, we employed the parameters used in previous SA proxy research and additional parameters observed simultaneously to build the RF model. An in-depth discussion of the potential impact of the selectable parameters on SA and the guiding rules for selecting feature parameters can be found in Text S6. In developing the RF model, we selected UV, RH, SO_2, CS, and MSA.

https://doi.org/10.1021/acs.estlett.3c00170
Environ. Sci. Technol. Lett. 2023, 10, 1045−1051

The Python scikit-learn package was used to train the RF model. There were around 11 400 valid data points with a temporal precision of 5 min. The data set was arbitrarily split into two sections, with 30% being used to test and 70% being used as a training set for establishing the RF model.[38] We input the test data into the RF model to obtain the simulated SA concentration, and linear regression was employed to evaluate the matching between the simulated and measured SA concentration. Three metrics were used to assess the model's performance. A higher coefficient of determination (R^2), a lower root-mean-square error (RMSE), and a lower mean absolute value error (MAE) imply better model performance.[39] A detailed explanation of these metrics can be found in Text S7. In addition, 10-fold cross-validation was used to confirm the accuracy of the RF model.

2.3. Model Interpretation. We applied two distinct approaches to perform feature importance analysis, including the built-in interpreter of the RF model and the widely used feature analysis technique called Shapley additive explanation (SHAP).[38,40] PDPs were used to estimate the average marginal influence of two feature parameters on the simulated outcome.[31,41] Then, c-ICE plots were employed to obtain a deeper understanding of the heterogeneity across observations.[42] A c-ICE plot was drawn for each unique synthetic instance of a feature parameter of interest, focusing on a specific forecast point while keeping the values of the other feature parameters constant. Text S8 presents the PDPs and c-ICE plots in detail.

3. RESULTS AND DISCUSSION

3.1. General Characteristics of SA and Model Performance. A time profile of SA and the five parameters during the sampling period, which were utilized to build the model, is shown in Figure S6. Similar diurnal variation characteristics were shown in the SA concentration and UV intensity (Figure S7), which agrees with earlier research indicating that photochemical reactions are the primary source of SA.[15−17,43] Table S1 shows the SA and associated feature parameter data collected during sampling in this study, as well as data from previous works. In this study, the median SA concentration (2.3×10^6 molecules cm^{-3}) was higher than those in forested and rural areas like Hyytiälä, closer to those of some urban areas like Atlanta, and lower than those in significantly polluted megacities like Beijing.[16,17,43]

The performance of the RF model established by the five feature parameters is shown in Figure 1b and Figure S8. In the regression results of the model test set, the simulated values matched well with the measured values, with an R^2 of 0.90, an RMSE of 2.5×10^6 molecules cm^{-3}, and an MAE of 1.3×10^6 molecules cm^{-3}. In comparison, we constructed three proxies based on the data of this study with reference to earlier research. The specific details of these three proxies are described in Text S9. Figure 1a displays the diurnal fluctuation of the measured SA, the SA calculated by the three proxies, and the SA simulated by the RF model. The performance of proxy 3 is the best among the three proxies. Compared to the three proxies, the diurnal variation of the SA simulated by the RF model is more similar to that of the measured values. Interestingly, proxy 3, which considered the pathway of SO_2 oxidation by sCI to SA, clearly underestimated the SA concentrations at night. In contrast, the RF model involving MSA reproduced well the nighttime SA concentrations. For a comprehensive analysis of the benefits and drawbacks associated with the three conventional proxies, along with the RF model and the implementation of scenario analysis, see Text S10. In summary, the performance of the ML-based SA simulation model constructed in this study is better than that of the traditional SA proxies. Our results also indicate that the oxidation of sulfur-containing species from marine emissions might play an important role in SA distribution.

3.2. Importance of Feature Parameters. Figure 1c illustrates the importance of the five selected feature parameters, as obtained from the built-in interpreter of the RF model, while Figure S9 shows the importance of the same five parameters as determined by the SHAP values; both methods indicate the same order of importance for the parameters: UV, MSA, SO_2, CS, and RH. Via the repeated addition of feature parameters, the consequent increase in R^2 and the corresponding decrease in RMSE and MAE confirmed the validity of the features selected for model improvement in Figure S10. Previous studies have confirmed that solar radiation intensity, and other parameters can indicate the concentration of OH radicals, which is a critical parameter of the SA proxy.[15−17,43] Unexpectedly, MSA was more significant in the RF model than SO_2, the most important precursor acknowledged for SA production in inland areas. To illustrate the superiority of these two feature parameters, two new RF models were built: one without SO_2 and the other without MSA, each trained with four feature parameters. As the model without MSA [R^2 = 0.84 (Figure 1d)] performed worse than the model without SO_2 [R^2 = 0.88 (Figure S11)], we can conclude that the former was more crucial than the latter for SA simulation in this study. Compared to the model without MSA, adding MSA to the feature parameters enhanced the model's simulative performance, with decreased MAE and RMSE values and R^2 increased from 0.84 to 0.90.

As shown in Figure S12, the SA:MSA ratio varied throughout the day but remained stable from 0.34 to 0.46 at night (20:00−04:00). The high SA:MSA ratio during the day was consistent with the diurnal variation of UV, suggesting an additional pathway of SA generation (likely the reaction of SO_2 and OH radicals) during daytime compared to MSA. In addition, the stable SA:MSA ratio at night indicates that SA and MSA may have similar sources. The potential pathways for the oxidation of DMS by OH or nitrate radicals to produce MSA and SA are depicted in Figure S13.[19] DMS undergoes several processes to produce CH_3SO_2, which can then generate SA by first decomposing to SO_2 and then being oxidized to SO_3. Additionally, MSA can be formed by the reaction of CH_3SO_2 with O_3 or NO_2 to produce CH_3SO_3, largely depending on the humidity level. The scatter plot of the SA and MSA data at night showed that high-SA events were associated with high MSA concentrations (Figures S14 and S15). To further probe the influence of MSA on nighttime SA concentration, data between 20:00 and 4:00 the subsequent day were selectively screened to construct a nocturnal RF model. Details regarding the construction and analysis of the nocturnal RF model are provided in Text S11. The performance of the nocturnal RF model and the parameter importance of MSA are elucidated in Figure S16. MSA was the most parametrically important, and SA concentration increased with MSA concentration. The results indicated that sulfur-containing species emitted from the ocean might play an important role in SA formation, which cannot be neglected in coastal areas. However, this study was conducted at only one coastal urban site, which is restricted in terms of observation

https://doi.org/10.1021/acs.estlett.3c00170
Environ. Sci. Technol. Lett. 2023, 10, 1045−1051

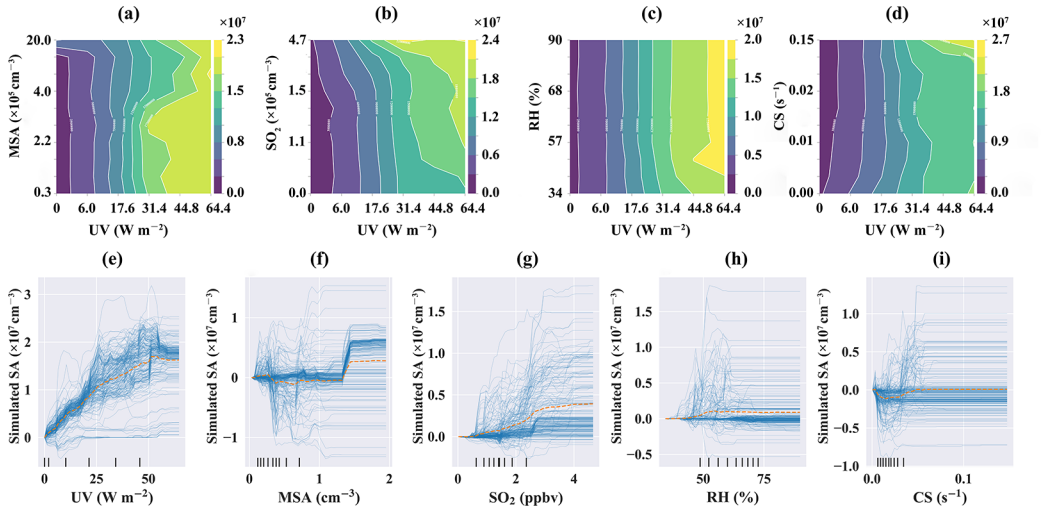


Figure 2. Marginal effects of feature parameters on SA concentration. (a–d) Combined impact of UV and the other four feature parameters on SA presented using two-dimensional PDPs. The color bar represents the concentration of SA, while the axes are on a logarithmic scale. (e–i) c-ICE curves (blue) and their averages (dashed orange) depicting the relationships among UV, MSA, SO_2, RH, CS, and SA.

site type and geographical scale. Future study regions should be broadened to include numerous coastal urban and rural sites with diverse meteorological circumstances.

3.3. Marginal Effects of Feature Parameters on SA Concentration. UV was undoubtedly the most crucial factor of SA in this study. The effects of the interaction between UV and the other four feature parameters on SA were evaluated using PDPs (Figure 2a–d). When the UV intensity was close to 0, the concentration of SA was always $<3.0 \times 10^6$ molecules cm^{-3} regardless of the variation of the feature parameters other than MSA (Figure 2b–d). However, the nighttime SA showed a significant increase as the MSA concentration increased. In particular, when the MSA concentration reached a peak at night, the concentration of SA reached 8.0×10^6 molecules cm^{-3}, which might indicate a co-occurrence relationship between MSA and SA (Figure 2a). This is yet more evidence that oceanic emissions of sulfur-containing species had an important impact on SA at night. When the UV intensity was high ($\geq$45 W m^{-2}), the dramatic fluctuations of SA with SO_2 concentration indicated that SO_2 was the most important feature parameter limiting SA concentration under a high UV intensity (Figure 2b). Therefore, during the day, especially at midday when solar radiation was intense, the interaction between SO_2 and OH radicals was the principal source of SA.

The application of PDPs is constrained because they do not offer sufficient insight into the heterogeneity of the data resulting from interactions between feature parameters. Therefore, for each unique sample, c-ICE plots are created for each relevant feature parameter, highlighting a specific forecast point while holding the values of unimportant feature parameters constant (Figure 2e–i). SA displayed a substantial marginal increase effect with an increase in UV when the UV intensity was <55 W m^{-2} (Figure 2e). Figure 2g indicates that the concentration of SA increased with SO_2 concentration up to 4.0 ppbv. These phenomena gradually diminished thereafter, indicating that the increase in SA was positively correlated with UV intensity or SO_2 concentration until they reached a peak. This is due to the fact that the reaction between SO_2 and OH radicals dominated SA production, but only when one parameter increased; the other parameter limited the SA concentration, leading to a weaker marginal increase effect. The effect on SA dramatically increased when the MSA concentration was $>1.4 \times 10^6$ molecules cm^{-3} (Figure 2f). The high concentration of MSA implied the presence of high concentrations of sulfur-containing precursors, leading to a significant promotion of SA. When the RH was low, the SA concentration increased as the RH increased (Figure 2h). CS was the only feature parameter in this analysis that had a negative correlation with SA (Figure 2i). This is because, in addition to self-recommendation to form sulfuric acid dimers, condensation by preexisting aerosol is the main sink of SA. More SA was adsorbed on the surface of the particulate matter as the quantity of fine particulate matter in the atmosphere increased, increasing the contact area of SA with the particulate matter. The tiny thresholds and marginal effects imply that the SA concentration was minimally impacted by CS and RH (Figure 2h, i).

To investigate the parameters affecting the distribution of SA in a coastal city, we developed a simulation model for SA concentration using ML based on field observations in the coastal city of Xiamen. Compared with the traditional SA proxies, the SA simulation model constructed on the basis of ML had better performance, with R^2 reaching 0.90. The five feature parameters of the constructed model were UV, MSA, SO_2, CS, and RH in order of feature importance (Figure 1c). Compared to the model without MSA (Figure 1d), the model performance showed a significant improvement, demonstrating sulfur-containing species from marine emissions were important sources of SA in this study. PDPs indicated a significant co-occurrence relationship between MSA and SA during the night, whereas SO_2 and UV intensity play a dominant role during the day (Figure 2a–d). A persistent marginal increasing effect of SO_2 and UV on SA before reaching a peak was revealed by the c-ICE analysis (Figure 2e–i). Thus, sulfur-containing species originating from marine emissions may play a role in determining nighttime SA concentration and should be taken into account in coastal regions. Future research should make simultaneous observations of MSA, SA, and oceanic emissions of sulfur-containing precursors (e.g., DMS and CH_3SO_2) to clarify the special SA mechanisms in coastal areas.

https://doi.org/10.1021/acs.estlett.3c00170
Environ. Sci. Technol. Lett. 2023, 10, 1045–1051

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.estlett.3c00170.

Detailed information about the study domain and period (Text S1), nitrate−CIMS settings (Text S2), auxiliary measurements (Text S3), calculation of the condensation sink (Text S4), workflow of the RF model (Text S5), selection of parameters (Text S6), model evaluation metrics (Text S7), PDPs and c-ICE plots (Text S8), development of three SA proxies (Text S9), analysis of three traditional proxies and the RF model (Text S10), nighttime RF model (Text S11), study site location and cluster analyses (Figures S1 and S2), peak fitting of selected species (Figure S3), selection of hyperparameters (Figure S4), importance ranking of all feature parameters (Figure S5), time series of the different parameters (Figure S6), diurnal trend of SA and MSA (Figure S7), comparation of measured and simulated SA (Figure S8), feature performance gained from the Shapley additive explanation method (Figure S9), model performance (Figure S10), four-parameter RF model performance (Figure S11), diurnal trend of the SA:MSA ratio (Figure S12), DMS oxidation scheme (Figure S13), scatter plot of nighttime SA and MSA concentrations (Figures S14 and S15), analysis of the nighttime RF model (Figure S16), SA calibration (Figure S17), and performance of three SA proxies (Figure S18) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Jinsheng Chen − *Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China;* orcid.org/0000-0001-9255-4323; Email: jschen@iue.ac.cn

Xiaolong Fan − *Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China*; Email: xlfan@iue.ac.cn

Lingling Xu − *Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China*; Email: linglingxu@iue.ac.cn

Authors

Chen Yang − *Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Hesong Dong − *Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Yuping Chen − *Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Yonghong Wang − *State Key Joint Laboratory of Environment Simulation and Pollution Control, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China;* orcid.org/0000-0003-2498-9143

Yee Jun Tham − *School of Marine Sciences, Sun Yat-sen University, Zhuhai 519082, China*

Gaojie Chen − *Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Ziyi Lin − *Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Mengren Li − *Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China*

Youwei Hong − *Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China; Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.estlett.3c00170

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (U22A20578, 42277091, and 42175118), the National Key Research and Development Program (2022YFC3700304), the Science and Technology Department of Fujian Province (2022L3025), the Xiamen Atmospheric Environment Observation and Research Station of Fujian Province, and the Fujian Key Laboratory of Atmospheric Ozone Pollution Prevention (Institute of Urban Environment, Chinese Academy of Sciences).

■ REFERENCES

(1) McMeeking, G.; Morgan, W.; Flynn, M.; Highwood, E.; Turnbull, K.; Haywood, J.; Coe, H. Black carbon aerosol mixing

state, organic aerosols and aerosol optical properties over the United Kingdom. *Atmospheric Chemistry and Physics* **2011**, *11* (17), 9037−9052.
(2) Yue, D.; Hu, M.; Wu, Z.; Wang, Z.; Guo, S.; Wehner, B.; Nowak, A.; Achtert, P.; Wiedensohler, A.; Jung, J.; et al. Characteristics of aerosol size distributions and new particle formation in the summer in Beijing. *J. Geophys. Res.: Atmos.* **2009**, *114* (D2), ZZZ.
(3) Pope, C. A., III; Dockery, D. W. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air Waste Management Association* **2006**, *56* (6), 709−742.
(4) Bréon, F.-M. How do aerosols affect cloudiness and climate? *Science* **2006**, *313* (5787), 623−624.
(5) Guo, S.; Hu, M.; Zamora, M. L.; Peng, J.; Shang, D.; Zheng, J.; Du, Z.; Wu, Z.; Shao, M.; Zeng, L.; et al. Elucidating severe urban haze formation in China. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (49), 17373−17378.
(6) Nieminen, T.; Manninen, H. E.; Sihto, S.-L.; Yli-Juuti, T.; Mauldin, R. L., III; Petaja, T.; Riipinen, I.; Kerminen, V.-M.; Kulmala, M. Connection of sulfuric acid to atmospheric nucleation in boreal forest. *Environ. Sci. Technol.* **2009**, *43* (13), 4715−4721.
(7) Weber, R.; Marti, J.; McMurry, P.; Eisele, F.; Tanner, D.; Jefferson, A. Measurements of new particle formation and ultrafine particle growth rates at a clean continental site. *Journal of Geophysical Research: Atmospheres* **1997**, *102* (D4), 4375−4385.
(8) Yao, L.; Garmash, O.; Bianchi, F.; Zheng, J.; Yan, C.; Kontkanen, J.; Junninen, H.; Mazon, S. B.; Ehn, M.; Paasonen, P.; et al. Atmospheric new particle formation from sulfuric acid and amines in a Chinese megacity. *Science* **2018**, *361* (6399), 278−281.
(9) Benson, D. R.; Young, L. H.; Kameel, F. R.; Lee, S. H. Laboratory-measured nucleation rates of sulfuric acid and water binary homogeneous nucleation from the SO_2 + OH reaction. *Geophys. Res. Lett.* **2008**, *35* (11), n/a DOI: 10.1029/2008GL033387.
(10) Almeida, J.; Schobesberger, S.; Kürten, A.; Ortega, I. K.; Kupiainen-Määttä, O.; Praplan, A. P.; Adamov, A.; Amorim, A.; Bianchi, F.; Breitenlechner, M.; et al. Molecular understanding of sulphuric acid-amine particle nucleation in the atmosphere. *Nature* **2013**, *502* (7471), 359−363.
(11) Lehtipalo, K.; Yan, C.; Dada, L.; Bianchi, F.; Xiao, M.; Wagner, R.; Stolzenburg, D.; Ahonen, L. R.; Amorim, A.; Baccarini, A.; et al. Multicomponent new particle formation from sulfuric acid, ammonia, and biogenic vapors. *Sci. Adv.* **2018**, *4* (12), No. eaau5363.
(12) Kirkby, J.; Curtius, J.; Almeida, J.; Dunne, E.; Duplissy, J.; Ehrhart, S.; Franchin, A.; Gagné, S.; Ickes, L.; Kürten, A.; et al. Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation. *Nature* **2011**, *476* (7361), 429−433.
(13) Yan, C.; Yin, R.; Lu, Y.; Dada, L.; Yang, D.; Fu, Y.; Kontkanen, J.; Deng, C.; Garmash, O.; Ruan, J.; et al. The Synergistic Role of Sulfuric Acid, Bases, and Oxidized Organics Governing New-Particle Formation in Beijing. *Geophys. Res. Lett.* **2021**, *48* (7), No. e2020GL091944.
(14) Wang, Y.; Ma, Y.; Yan, C.; Yao, L.; Cai, R.; Li, S.; Lin, Z.; Zhao, X.; Yin, R.; Deng, C.; et al. Sulfur dioxide transported from the residual layer drives atmospheric nucleation during haze periods in Beijing. *Geophys. Res. Lett.* **2023**, *50* (6), No. e2022GL100514.
(15) Petäjä, T.; Mauldin, R. L., III; Kosciuch, E.; McGrath, J.; Nieminen, T.; Paasonen, P.; Boy, M.; Adamov, A.; Kotiaho, T.; Kulmala, M. Sulfuric acid and OH concentrations in a boreal forest site. *Atmos. Chem. Phys.* **2009**, *9* (19), 7435−7448.
(16) Mikkonen, S.; Romakkaniemi, S.; Smith, J.; Korhonen, H.; Petäjä, T.; Plass-Duelmer, C.; Boy, M.; McMurry, P.; Lehtinen, K.; Joutsensaari, J.; et al. A statistical proxy for sulphuric acid concentration. *Atmos. Chem. Phys.* **2011**, *11* (21), 11319−11334.
(17) Dada, L.; Ylivinkka, I.; Baalbaki, R.; Li, C.; Guo, Y.; Yan, C.; Yao, L.; Sarnela, N.; Jokinen, T.; Daellenbach, K. R.; et al. Sources and sinks driving sulfuric acid concentrations in contrasting environments: implications on proxy calculations. *Atmos. Chem. Phys.* **2020**, *20* (20), 11747−11766.
(18) Yang, L.; Nie, W.; Liu, Y.; Xu, Z.; Xiao, M.; Qi, X.; Li, Y.; Wang, R.; Zou, J.; Paasonen, P.; et al. Toward building a physical proxy for gas-phase sulfuric acid concentration based on its budget analysis in polluted Yangtze River Delta, East China. *Environ. Sci. Technol.* **2021**, *55* (10), 6665−6676.
(19) Barnes, I.; Hjorth, J.; Mihalopoulos, N. Dimethyl sulfide and dimethyl sulfoxide and their oxidation in the atmosphere. *Chem. Rev.* **2006**, *106* (3), 940−975.
(20) Eisele, F.; Tanner, D. Measurement of the gas phase concentration of H2SO4 and methane sulfonic acid and estimates of H2SO4 production and loss in the atmosphere. *Journal of Geophysical Research: Atmospheres* **1993**, *98* (D5), 9001−9010.
(21) Kettle, A.; Andreae, M. Flux of dimethylsulfide from the oceans: A comparison of updated data sets and flux models. *Journal of Geophysical Research: Atmospheres* **2000**, *105* (D22), 26793−26808.
(22) Cai, Q.; Luo, X.; Wang, P.; Gao, C.; Zhao, P. Hybrid model-driven and data-driven control method based on machine learning algorithm in energy hub and application. *Applied Energy* **2022**, *305*, 117913.
(23) Ma, J.; Cheng, J. C.; Lin, C.; Tan, Y.; Zhang, J. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmos. Environ.* **2019**, *214*, 116885.
(24) Eze, I. C.; Hemkens, L. G.; Bucher, H. C.; Hoffmann, B.; Schindler, C.; Künzli, N.; Schikowski, T.; Probst-Hensch, N. M. Association between ambient air pollution and diabetes mellitus in Europe and North America: systematic review and meta-analysis. *Environ. Health Perspect.* **2015**, *123* (5), 381−389.
(25) Oh, E.; Liu, R.; Nel, A.; Gemill, K. B.; Bilal, M.; Cohen, Y.; Medintz, I. L. Meta-analysis of cellular toxicity for cadmium-containing quantum dots. *Nature Nanotechnol.* **2016**, *11* (5), 479−486.
(26) Zhu, Q.; Laughner, J. L.; Cohen, R. C. Combining Machine Learning and Satellite Observations to Predict Spatial and Temporal Variation of near Surface OH in North American Cities. *Environ. Sci. Technol.* **2022**, *56* (56), 7362−7371.
(27) Wei, N.; Jia, Z.; Men, Z.; Ren, C.; Zhang, Y.; Peng, J.; Wu, L.; Wang, T.; Zhang, Q.; Mao, H. Machine Learning Predicts Emissions of Brake Wear PM2. 5: Model Construction and Interpretation. *Environmental Science & Technology Letters* **2022**, *9* (5), 352−358.
(28) Lewis, J. E.; Kemp, M. L. Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance. *Nat. Commun.* **2021**, *12* (1), 2700.
(29) Ilany, A.; Akcay, E. Social inheritance can explain the structure of animal social networks. *Nat. Commun.* **2016**, *7* (1), 12084.
(30) Khoda Bakhshi, A.; Ahmed, M. M. Utilizing black-box visualization tools to interpret non-parametric real-time risk assessment models. *Transportmetrica A: Transport Science* **2021**, *17* (4), 739−765.
(31) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **2001**, 1189−1232.
(32) Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. J. j. o. C. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational Graphical Statistics* **2015**, *24* (1), 44−65.
(33) Zhang, Y.; Xu, L.; Zhuang, M.; Zhao, G.; Chen, Y.; Tong, L.; Yang, C.; Xiao, H.; Chen, J.; Wu, X.; et al. Chemical composition and sources of submicron aerosol in a coastal city of China: Results from the 2017 BRICS summit study. *Sci. Total Environ.* **2020**, *741*, 140470.
(34) Hu, B.; Duan, J.; Hong, Y.; Xu, L.; Li, M.; Bian, Y.; Qin, M.; Fang, W.; Xie, P.; Chen, J. Exploration of the atmospheric chemistry of nitrous acid in a coastal city of southeastern China: results from measurements across four seasons. *Atmospheric Chemistry and Physics* **2022**, *22* (1), 371−393.
(35) Stein, A.; Draxler, R. R.; Rolph, G. D.; Stunder, B. J.; Cohen, M.; Ngan, F. NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bulletin of the American Meteorological Society* **2015**, *96* (12), 2059−2077.
(36) Jokinen, T.; Sipilä, M.; Junninen, H.; Ehn, M.; Lönn, G.; Hakala, J.; Petäjä, T.; Mauldin, R. L., III; Kulmala, M.; Worsnop, D.

Atmospheric sulphuric acid and neutral cluster measurements using CI-APi-TOF. *Atmos. Chem. Phys.* **2012**, *12* (9), 4117−4125.
(37) Zhan, J.; Liu, Y.; Ma, W.; Zhang, X.; Wang, X.; Bi, F.; Zhang, Y.; Wu, Z.; Li, H. Ozone formation sensitivity study using machine learning coupled with the reactivity of volatile organic compound species. *Atmospheric Measurement Techniques* **2022**, *15* (5), 1511−1520.
(38) Li, J.; Pan, L.; Suvarna, M.; Tong, Y. W.; Wang, X. Fuel properties of hydrochar and pyrochar: Prediction and exploration with machine learning. *Applied Energy* **2020**, *269*, 115166.
(39) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547−555.
(40) Li, J.; Zhang, L.; Li, C.; Tian, H.; Ning, J.; Zhang, J.; Tong, Y. W.; Wang, X. Data-driven based in-depth interpretation and inverse design of anaerobic digestion for CH4-rich biogas production. *ACS ES&T Eng.* **2022**, *2* (4), 642−652.
(41) Zhu, X.; Wang, X.; Ok, Y. S. The application of machine learning methods for prediction of metal sorption onto biochars. *Journal of Hazardous Materials* **2019**, *378*, 120727.
(42) Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational Graphical Statistics* **2015**, *24* (1), 44−65.
(43) Lu, Y.; Yan, C.; Fu, Y.; Chen, Y.; Liu, Y.; Yang, G.; Wang, Y.; Bianchi, F.; Chu, B.; Zhou, Y.; et al. A proxy for atmospheric daytime gaseous sulfuric acid concentration in urban Beijing. *Atmos. Chem. Phys.* **2019**, *19* (3), 1971−1983.

https://doi.org/10.1021/acs.estlett.3c00170
Environ. Sci. Technol. Lett. 2023, 10, 1045−1051