Environmental Science and Ecotechnology 19 (2024) 100333

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

# Knowledge-guided machine learning reveals pivotal drivers for gas-to-particle conversion of atmospheric nitrate

Bo Xu [a, b], Haofei Yu [c], Zongbo Shi [d], Jinxing Liu [e, f], Yuting Wei [a, b], Zhongcheng Zhang [a, b], Yanqi Huangfu [a, b], Han Xu [a, b], Yue Li [g], Linlin Zhang [h, **], Yinchang Feng [a, b], Guoliang Shi [a, b, *]

[a] State Environmental Protection Key Laboratory of Urban Ambient Air Particulate Matter Pollution Prevention and Control, Tianjin Key Laboratory of Urban Transport Emission Research, College of Environmental Science and Engineering, Nankai University, Tianjin, 300350, China
[b] CMA-NKU Cooperative Laboratory for Atmospheric Environment-Health Research (CLAER), College of Environmental Science and Engineering, Nankai University, Tianjin, 300350, China
[c] Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL, USA
[d] School of Geography Earth and Environment Sciences, University of Birmingham, Birmingham, B15 2TT, UK
[e] State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin Key Laboratory of air Pollutants Monitoring Technology, School of Precision Instrument and Opto-electronics Engineering, Tianjin University, Tianjin, 300072, China
[f] Gigantic Technology (Tianjin) Co., Ltd, Tianjin, 300072, China
[g] College of Computer Science, Nankai University, Tianjin, 300350, China
[h] China National Environmental Monitoring Centre, Beijing, 100012, China

ARTICLE INFO

*Article history:*
Received 7 January 2023
Received in revised form
9 October 2023
Accepted 12 October 2023

*Keywords:*
Machine learning
Data driven
Theoretical approach
Domain knowledge
Guide

ABSTRACT

Particulate nitrate, a key component of fine particles, forms through the intricate gas-to-particle conversion process. This process is regulated by the gas-to-particle conversion coefficient of nitrate ($\varepsilon(NO_3^-)$). The mechanism between $\varepsilon(NO_3^-)$ and its drivers is highly complex and nonlinear, and can be characterized by machine learning methods. However, conventional machine learning often yields results that lack clear physical meaning and may even contradict established physical/chemical mechanisms due to the influence of ambient factors. It urgently needs an alternative approach that possesses transparent physical interpretations and provides deeper insights into the impact of $\varepsilon(NO_3^-)$. Here we introduce a supervised machine learning approach—the multilevel nested random forest guided by theory approaches. Our approach robustly identifies $NH_4^+$, $SO_4^{2-}$, and temperature as pivotal drivers for $\varepsilon(NO_3^-)$. Notably, substantial disparities exist between the outcomes of traditional random forest analysis and the anticipated actual results. Furthermore, our approach underscores the significance of $NH_4^+$ during both daytime (30%) and nighttime (40%) periods, while appropriately downplaying the influence of some less relevant drivers in comparison to conventional random forest analysis. This research underscores the transformative potential of integrating domain knowledge with machine learning in atmospheric studies.



## 1. Introduction

Nowadays, nitrate pollution contributes significantly to haze pollution. The formation and behavior of nitrate in the atmosphere involve various physical and chemical processes, making it a complex issue to address [1–5]. Nitrate is a crucial secondary inorganic component of fine particulate matter ($PM_{2.5}$), accounting for 16–45% of the total $PM_{2.5}$ mass, especially during haze episodes in northern China [6,7]. As a semi-volatile component, nitrate undergoes conversion between the gas phase ($HNO_3$ (g)) and particle phase ($NO_3^-$) through ammonia neutralization in the atmosphere [8]. The gas-to-particle conversion coefficient of nitrate ($\varepsilon(NO_3^-)$) is a crucial parameter that influences nitrate formation [9]. Multiple

* Corresponding author. State Environmental Protection Key Laboratory of Urban Ambient Air Particulate Matter Pollution Prevention and Control, Tianjin Key Laboratory of Urban Transport Emission Research, College of Environmental Science and Engineering, Nankai University, Tianjin, 300350, China.
** Corresponding author.
*E-mail addresses:* zhangll@cnemc.cn (L. Zhang), nksgl@nankai.edu.cn (G. Shi).

https://doi.org/10.1016/j.ese.2023.100333

drivers can affect $\varepsilon(NO_3^-)$, including aerosol acidity (pH), temperature, liquid water content, ionic species in the aerosol, liquid phase, and gaseous pollutants (such as $HNO_3$ (g) and $NH_3$) [4,10–13]. These drivers determine the efficiency of nitrate formation and its partitioning between the gas and particle phases. Despite its significance, there is still a lack of comprehensive and quantitative understanding regarding the highly nonlinear relationship between $\varepsilon(NO_3^-)$ and these drivers.

Machine learning (ML) methods have become increasingly valuable in advancing scientific discovery, attributed to their mature algorithms, rich observation datasets, and powerful computing power [14–16]. This advancement benefits from the capacity that ML methods are highly effective in tackling complex nonlinear problems, making them popular tools in a wide range of fields, including climate change [17,18] and particulate matter pollution [19–22]. Given the complex nonlinear structure of ML methods, the success of ML methods heavily relies on carefully selecting input variables. High-dimensional choices can adversely affect model accuracy and interpretability, requiring the prioritization of relevant variables [23]. In addition, ML methods are inherently data-driven models relying heavily on available labeled data for development. Consequently, the results obtained from these methods may lack clear physical meaning or even contradict physical or chemical mechanisms [24,25]. This is an important consideration when applying ML methods in environmental sciences, where interpretations are crucial for understanding the processes. Besides, high uncertainties are present in air pollution concerns due to ambient factors interfering with observational data, which presents challenges for using ML methods. The outcomes of machine learning models need to be physically meaningful and can be interpreted rationally. The process of identifying appropriate variables and enhancing the physical interpretability of the ML results poses a daunting task.

To overcome these challenges, researchers must strive to enhance the interpretability of ML results. A careful integration of the ML method with existing knowledge of physical mechanisms and chemical processes has been implemented to provide meaningful and interpretable insights into environmental phenomena. Recent studies highlighted the potential of the theoretical approach (TA), which is rooted in physical and chemical mechanisms and can be used as prior knowledge to guide regular ML [26–28]. For this reason, regular ML methods integrated with prior knowledge from theory are suited to solve environmental issues. Therefore, it is desirable to find some ways to make the outcomes of ML methods consistent with the known laws of theory, facilitating the discovery that reveals the gas-to-particle conversion of nitrate in the atmosphere.

The primary goal of this study was to design a method that can improve the performance degradation of the model caused by high dimensional variables and avoid the ambiguous physical meaning of ML's results by introducing prior knowledge. Here, we proposed a knowledge-guided machine learning method using a theoretical approach as prior knowledge to unravel the effect of drivers on $\varepsilon(NO_3^-)$. We first used a multilevel nested random forest model to elucidate the effect of individual drivers on $\varepsilon(NO_3^-)$. We further made a theoretical approach as prior knowledge to guide the multilevel nested random forest by searching for the globally optimal solution to replace the mathematically averaged result of machine learning. The study demonstrates that prior knowledge-guided machine learning can robustly produce quantitative results with physical meaning and highlight the importance of significant drivers on $\varepsilon(NO_3^-)$, which are closer to the actual value. Furthermore, this novel approach exhibits exceptional performance in solving complex environmental issues and yields higher result accuracy than regular ML methods. As a result, this research holds immense potential for environmental and climate research in the future, as it provides a more robust and reliable framework for addressing complex environmental issues.

## 2. Materials and methods

### 2.1. Site description and instrumentation

Real-time pollution measurements were conducted on the campus of Nankai University (38°59′ N, 117°20’ E), in the Jinnan district of Tianjin, China, from November 2017 to October 2018. The observation site is situated in a typical suburban area surrounded by a wetland park (Fig. S1). Measurement instruments were stationed within the Air Quality Research Supersite, with sampling inlets installed on the rooftop. Throughout the study campaign, hourly online monitoring was conducted on particulate matter, particle chemical species, trace gas pollutants, and meteorological conditions. Details of the observed species can be found in the Supporting Information (Table S1). Mass concentrations of $PM_{2.5}$ were measured using beta ray particulate matter automatic monitors (BPM-200, Focused Photonics Inc), while concentrations of gases ($SO_2$, $NO_x$, and $O_3$) were measured by gas analyzers (T101, T201, and T400, Teledyne API Inc.). Water-soluble ions ($NO_3^-$, $SO_4^{2-}$, $NH_4^+$, $Cl^-$, $K^+$, $Mg^{2+}$, and $Ca^{2+}$) and semi-volatile species in the gas phase ($HNO_3$ (g), $NH_3$) were measured using ion chromatography (URG9000 Thermo Fisher Scientific Inc, USA). Metallic elements (K, Na, Mg, Zn, Fe, etc.) were measured using an atmospheric heavy metal analyzer (AMMS-100, Focused Photonics Inc). Organic carbon (OC) and element carbon (EC) were measured hourly with a semi-continuous thermal-optical carbon analyzer (OCEC-100, Focused Photonics Inc). Meteorological conditions, including relative humidity and temperature, were measured using an automatic meteorological observation system (WS600-UMB, LUFFT).

### 2.2. Mechanism-driven theoretical approach

The gas-to-particle conversion process of nitrate can be understood through classic thermodynamic principles, which provide a theoretical foundation for analysis. Previous studies have revealed an S-shaped curve between $\varepsilon(NO_3^-)$ and aerosol pH, providing a conceptual foundation for exploring the individual effect of drivers on $\varepsilon(NO_3^-)$ [27]. We first calculated the theoretical value of $\varepsilon(NO_3^-)$ (defined as $\varepsilon(NO_3^-)$*) for each sample based on theoretical calculation (mechanism-driven) in equation (1) [8], which was determined by $H^+$, aerosol water content, and ambient temperature. To estimate liquid water content (LWC) and $H^+$ in aerosol, we employed the thermodynamic model (ISORROPIA-II), which considers corresponding gases for the $Na^+$-$K^+$-$Ca^{2+}$-$Mg^{2+}$-$NH_4^+$-$SO_4^{2-}$-$NO_3^-$-$Cl^-$-$H_2O$ aerosol system [29].

$$\varepsilon\left(NO_3^-\right)^* \approx \frac{3.2RTL\exp\left(8700\left(\frac{1}{T}-\frac{1}{273.15}\right)\right)}{[H^+]+3.2RTL\exp\left(8700\left(\frac{1}{T}-\frac{1}{273.15}\right)\right)} \quad (1)$$

In this equation, R is the ideal gas constant equal (0.08205 atm L $mol^{-1}$ $K^{-1}$); $T$ (K) is the temperature; and $L$ (g $m^{-3}$) is the aerosol liquid water content (as estimated by ISORROPIA-II).

We then calculated the effect of drivers on $\varepsilon(NO_3^-)$* using the empirical thermodynamic model. Further details on the calculation process are presented in Text S1.

### 2.3. Data-driven machine learning method

#### 2.3.1. Random forest model (RF)

RF is a widely used machine learning model with satisfactory performance that contains multiple decision trees constructed randomly [30]. Each decision tree serves as a basic unit in the RF framework, resembling a tree structure where the leaves represent independent variables. In contrast to a single decision tree, RF iteratively selects random samples from the original datasets to generate multiple decision trees, thus forming a random forest [31]. In this study, the dependent variable is $\varepsilon(NO_3^-)$, while the independent variables include anions, cations, temperature, relative humidity, etc. The construction method of the model adopts a ten-fold cross-validation [32]. Here, the datasets are divided into two parts: 90% of the samples are used as training datasets to fit the RF model, and the remaining 10% are designated as test datasets for evaluating the model's performance [31]. Evaluation metrics, including the coefficient of determination ($R^2$), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), are selected for assessing the model's performance. Here, model building is carried out using the Python package scikit-learn in Python 3.7. Further details about RF are presented in Text S2(1).

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} g_i(x) \tag{2}$$

In equation (2), $f(x)$ is the predicted averaged value of an RF, such as $\overline{\varepsilon(NO_3^-)}$ in this study, $g_i(x)$ is the predicted value of each decision tree $i$, such as $\varepsilon(NO_3^-)$ in this study; $n$ is the number of decision trees in RF.

#### 2.3.2. Multilevel nested random forest model (MNRF)

To address the issue of reduced model performance due to the high dimensionality of independent variables (i.e., drivers), we propose a multilevel nested random forest model with a total of K layers (Text S2(2)). The flow chart of the multilevel nested random forest model is shown in Fig. S7, where K represents the number of layers. In the new method, the initial K-1 layers of RF establish a nonlinear response among initial drivers (e.g., some ionic species and gaseous pollutants) and pH, allowing for the automatic selection of important drivers using four filter indicators. The final K of layer RF further explores the effect of screened drivers on $\varepsilon(NO_3^-)$.

#### 2.3.3. Multilevel nested random forest-permutation importance (MNRF-PI) and multilevel nested random forest-partial dependence plot (MNRF-PDP)

While the MNRF model can technically be created with numerous parameter combinations, their interpretations can pose challenges. To address this, the MNRF model incorporates two integral components: MNRF-PI and MNRF-PDP. The MNRF-PI algorithm is designed to assess the importance of variables by measuring variations in the prediction accuracy of the RF model after permuting values of a single input variable (Text S2(3)) [33]. The MNRF-PDP algorithm analyzes the average effect of one or two variables on prediction value. Specially, MNRF-PDP is a function of $X_k$ (driver) and model prediction values (equation (3)). The individual effect of drivers on $\varepsilon(NO_3^-)$ is calculated using MNRF-PDP (Text S2(4)).

$$f(X_k) = \frac{1}{n}\sum_{i=1}^{n} RF(Y, X_k, X_m) \tag{3}$$

In equation (3), $Y$ is the independent values input into the model, such as $\varepsilon(NO_3^-)$ in this study; $RF$ is a trained RF model; $X_k$ is the $k$th driving factor introduced into the machine learning model (such as a certain driving factor, relative humidity); $X_m$ is the $m$th driving factor that was introduced into the machine learning model with a fixed value, such as the other driving factors; n is the number of samples, totaling 4175.

### 2.4. Supervised MNRF guided by theoretical approach

The MNRF-PDP method calculates the average effect of drivers on $\varepsilon(NO_3^-)$ by averaging the results across all samples. However, this approach is not constrained by physical or chemical mechanisms, which may result in an incomplete characterization of the drivers' impact on $\varepsilon(NO_3^-)$. To address this limitation, we propose a method called supervised MNRF guided by a theoretical approach.

The MNRF-PDP algorithm can calculate the average effect of drivers on $\varepsilon(NO_3^-)$ by averaging the results of all samples. The average effect of drivers on $\varepsilon(NO_3^-)$ can be expressed as the outcomes of the RF model. To make the results of RF more consistent with theories, a theoretical approach is adopted to guide the results of MNRF by searching for the globally optimal solution to replace mathematical averages. Two indicators, $VDS_{\varepsilon(NO_3^-)_j^n}$ (equation (4)) and $DDS_{\varepsilon(NO_3^-)_j^n}$ (equation (5)), are selected to screen for globally optimal results.

$$VDS_{\varepsilon(NO_3^-)_j^n} = \sum_{i}^{m}\left(Effe\ k_{the_{ij}} - Effe\ k_{obs_{ij}}\right)^2 \tag{4}$$

In equation (4), $DR_{\varepsilon(NO_3^-)_j^n}$ represents the square sum of the differences between the value of $\varepsilon(NO_3^-)$* (estimated by TA) and $\varepsilon(NO_3^-)$ (estimated by MNRF) in the $j$th sample under the gradient change of drivers; $Effe\ k_{the_{ij}}$ represents the effect of the $k$th driving factor ($X_k$) on $\varepsilon(NO_3^-)$* in theoretical data; $Effe\ k_{obs_{ij}}$ represents the effect of $k$th driving factor ($X_k$) on $\varepsilon(NO_3^-)$ in observational data; n is the number of samples, totaling 4175; and $i$, $m$ represent the gradient change of drivers.

$$DDS_{\varepsilon(NO_3^-)_j^n} = \sum_{i}^{m}\left(\frac{dEffe\ k_{the_{i+1}} - dEffe\ k_{the_i}}{dx_{i+1} - dx_i} - \left(\frac{dEffe\ k_{obs_{i+1}} - dEffe\ k_{obs_i}}{dx_{i+1} - dx_i}\right)_j\right)^2 = \sum_{i}^{m}\left(\frac{dEffe\ k_{the}}{dx} - \left(\frac{dEffe\ k_{obs}}{dx}\right)_j\right)^2 \tag{5}$$

In equation (5), $DD_{\varepsilon(NO_3^-)_j^n}$ represents the square sum of the differences between the differentiation of $\varepsilon(NO_3^-)$* (estimated by TA) and $\varepsilon(NO_3^-)$ (estimated by MNRF) in the $j$th sample under the gradient change of drivers.

In this academic article, we employed the following procedures to search for globally optimal results: (1) Calculation of $VDS_{\varepsilon(NO_3^-)_j^n}$ and $DDS_{\varepsilon(NO_3^-)_j^n}$ for each sample. (2) Selection of the top 10% of

samples with the lowest $VDS_{\varepsilon(NO_3^-)_j^n}$ and the lowest $DDS_{\varepsilon(NO_3^-)_j^n}$. (3) Utilization of the intersection of these two datasets to calculate *GOR* (equation (6)). *GOR* demonstrates greater consistency with the theoretical approach compared with regular RF model. (4) Assessment of the importance of drivers based on *GOR*. Further details of MNRF-TA can be found in Text S3.

$$GOR = \sum_{i}^{p} Effe\ k_{obs_i} \tag{6}$$

In equation (6), *GOR* is the globally optimal searched result; $Effe\ k_{obs_i}$ represents screened samples; and $p$ is the number of screened samples.

## 3. Results

### 3.1. Gas-to-particle conversion of nitrate-based on observational data

Online observations of $PM_{2.5}$ and chemical species during the campaign period were plotted to illustrate characteristics of nitrate pollution (Fig. S2). The average $PM_{2.5}$ concentration was 49 ± 49 μg $m^{-3}$ (mean ± standard deviation). Secondary inorganic ions ($NO_3^-$, $NH_4^+$, and $SO_4^{2-}$) accounted for 47% of $PM_{2.5}$, highlighting their dominant roles in the aerosol [34–36]. $NO_3^-$ emerged as the key component with an average concentration of 10.3 ± 12.1 μg $m^{-3}$ in $PM_{2.5}$ (Fig. S2a), significantly higher than that of $SO_4^{2-}$ (6.4 ± 7.5 μg $m^{-3}$) [37]. The contribution of $NO_3^-$ to $PM_{2.5}$ increased from 16.2% in non-haze periods ($PM_{2.5}$ < 35 μg $m^{-3}$) to 22.1% in the polluted periods ($PM_{2.5}$ > 150 μg $m^{-3}$), underscoring its crucial roles in the formation of $PM_{2.5}$ haze pollution [6]. The average concentration of $HNO_3$ (g) was 3.3 ± 11.8 μg $m^{-3}$. Here, $\varepsilon(NO_3^-)$ was calculated as $NO_3^-/(HNO_3\,(g) + NO_3^-)$, which reflects the conversion of $HNO_3$ to $NO_3^-$. In this work, the mean value of $\varepsilon(NO_3^-)$ was 0.77 ± 0.22, demonstrating that nitrate prefers the particle phase.

Moreover, nitric acid gas-to-particle conversion exhibits distinct patterns during both daytime and nighttime, contingent on the nitrate formation mechanism. At daytime, $HNO_3$ in the gas phase, originates from the photochemical oxidation of $NO_2$ and subsequently combines with alkaline gas ($NH_3$) to form $NH_4NO_3$ in the particle phase. Conversely, $HNO_3$ in the particulate phase, generated from hydrolysis of $N_2O_5$, converts into $HNO_3$ (g) at nighttime (specific mechanisms see Text S4) [38–40]. Significant diurnal variations were also found for the observed species ($NO_3^-$, $HNO_3$ (g), etc., Fig. S2c). $\varepsilon(NO_3^-)$ displays a gradual decline during daytime (with a mean value of 0.74) and a subsequent increase during nighttime (with a mean value of 0.80). The fluctuations in $\varepsilon(NO_3^-)$ are closely linked to gaseous pollutants, anion-cation balance, and meteorological conditions [41]. Notably, the diurnal variation of $HNO_3$ (g) reveals a single peak pattern, with higher values during daytime (4.0 μg $m^{-3}$) than nighttime (2.7 μg $m^{-3}$). Similarly, the diurnal variation of $NO_3^-$ exhibits bimodal peaks, occurring at 2:00 a.m. and 10:00 a.m., with elevated values observed during nighttime (11.1 μg $m^{-3}$) in contrast to daytime (9.7 μg $m^{-3}$). As for alkaline gas, the diurnal variation of $NH_3$ was highly consistent with that of $\varepsilon(NO_3^-)$. It is noteworthy that meteorological conditions, particularly temperature and relative humidity, significantly influence $\varepsilon(NO_3^-)$ by impacting the nitrate equilibrium constant and facilitating nitrate formation or volatilization [42]. The temperature exhibits a distinct pattern, rising from 6:00 a.m., peaking around 3:00 p.m., and gradually subsiding thereafter. This pattern contrasts with that of relative humidity. The aforementioned elucidation underscores the nonlinear relationship between $\varepsilon(NO_3^-)$ and drivers in the ambient atmosphere, thus emphasizing the need to delve deeper into the impact of these drivers on $\varepsilon(NO_3^-)$.

### 3.2. The mechanism-driven model ascertains drivers and ambient factors

The mechanism-driven theoretical approach was adopted to investigate drivers for $\varepsilon(NO_3^-)$ and validate the existence of ambient factors with the observed data. As a function of pH, $\varepsilon(NO_3^-)$ can be estimated by leveraging a thermodynamic equation based on the properties of $HNO_3$ (g) and $NO_3^-$ and assuming an ideal solution condition [43]. Several previous studies have highlighted an S-shaped curve between $\varepsilon(NO_3^-)$ and aerosol pH, the roles of multiple drivers on $NO_3^-$ [27,28,38]. However, the anticipated S-shaped relationship between observed $\varepsilon(NO_3^-)$ and pH was less evident in this study (Fig. 1a). Building on our prior works, we captured the sensitive band of the S-shaped curve after filtering the observed data with the high weighting of sulfate (when $SO_4^{2-}$ concentrations are more significant than 5 μg $m^{-3}$, Fig. 1b). The observed data only covers part of the S-shaped curve, instead of the entire curve [42]. The involvement of complex ambient processes may be responsible for this phenomenon. Additionally, we calculated estimators ($\varepsilon(NO_3^-)$*) using an entirely thermodynamic formula (equation (1)). A clear S-shaped curve was found between $\varepsilon(NO_3^-)$* and pH (Fig. 1c). Overlaying observed data ($\varepsilon(NO_3^-)$) with theoretical estimators ($\varepsilon(NO_3^-)$*) revealed a noteworthy dispersion of outlier points (Fig. S3). The theoretically estimated $\varepsilon(NO_3^-)$* can serve as the baseline relationship between $\varepsilon(NO_3^-)$* and drivers, while $\varepsilon(NO_3^-)$ computed from observed data was also affected by additional processes in the complex ambient environment (ambient factors), in addition to drivers. Consequently, relationships estimated based on observed data do not necessarily reproduce theoretical predictions due mainly to the interference of ambient factors in the complex ambient atmospheric environment. Moreover, outcomes derived from the theoretical approach exhibit an obvious S-shaped curve, aligning with prior knowledge of the actual scenario. This underscores that the result obtained through the theoretical approach can represent the actual result and be considered as expected.

The thermodynamic equation (equation (1)) underscores the relevance of $HNO_3$ solubility and dissociation in shaping $\varepsilon(NO_3^-)$ dynamics during daytime and nighttime. Multiple drivers also contribute to this intricate process, including temperature, particle liquid water content, etc. (Fig. S6) [44,45] Temperature greatly influences the nitrate equilibrium constant, and relative humidity governs aerosol liquid water. Generally, certain anion-cation species strongly impact aerosol pH. Concurrently, gaseous pollutants such as $NH_3$ operate as pH buffers. $NH_3$ affects the conversion gas-to-particle of nitrate by consuming $H^+$, which elevates pH and converts $HNO_3$ (g) into $NO_3^-$. Overall, ambient temperature, relative humidity, $NH_4^+$, $SO_4^{2-}$, $NH_3$, $Ca^{2+}$, $Cl^-$, $K^+$, $Mg^{2+}$, and $Na^+$ were regarded as the drivers of $\varepsilon(NO_3^-)$, and the intricate manner by which these drivers influence $\varepsilon(NO_3^-)$ necessitates thorough exploration.

### 3.3. Quantifying the effect of drivers on $\varepsilon(NO_3^-)$ using MNRF

Numerous machine learning algorithms are available for utilization, including RF, XGBoost, Gradient Boosting (GB), support vector machine (SVM), Long short-term memory (LSTM), and deep neural network (DNN) [20,46]. To find a suitable model for dealing with the $\varepsilon(NO_3^-)$-drivers system in this study, machine learning methods (including RF, XGBoost, GB, and SVM) and deep learning methods (including LSTM and DNN) were applied with the same input data, and their performances were cross-compared. Model

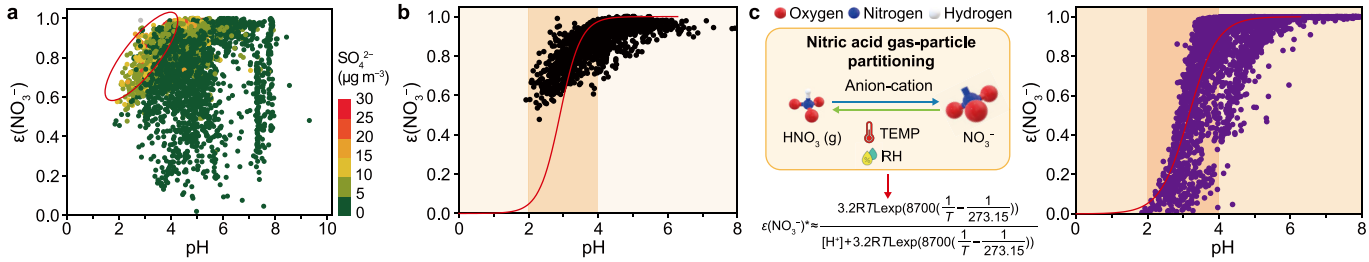


**Fig. 1.** Relationship between pH and ε($NO_3^-$). **a**, All ambient data: the S-shaped curve between ε($NO_3^-$) and pH is less obvious in an ambient atmospheric environment. **b**, Selected data in the ε($NO_3^-$)–pH-sensitive band of the S-shaped curve when $SO_4^{2-}$ concentrations were greater than 5 μg $m^{-3}$. The subsamples were shown as black points and readjusted by the Boltzmann function. **c**, An obvious S-shaped curve was found between ε($NO_3^-$)* and pH. Observational data cannot represent the actual result ("S-shaped curve") as a result of interference of ambient factors (panel **a**). The result from the theoretical approach can reflect the expected actual result, and a clear S-shaped curve can be obtained.

performances were evaluated using four metrics: $R^2$, RMSE, MAE, and MAPE (for more information, see Text S5), and the results were presented in Figs. S8–S10. The evaluation of model performance revealed that RF ($R^2$ = 0.82, RMSE = 0.09, MAE = 0.06, MAPE = 0.13) outperformed XGBoost ($R^2$ = 0.78, RMSE = 0.10, MAE = 0.07, MAPE = 0.14), GB ($R^2$ = 0.77, RMSE = 0.10, MAE = 0.06, MAPE = 0.13), DNN ($R^2$ = 0.75, RMSE = 0.10, MAE = 0.07, MAPE = 0.14), LSTM ($R^2$ = 0.58, RMSE = 0.12, MAE = 0.08, MAPE = 0.13), and SVM ($R^2$ = 0.53, RMSE = 0.13, MAE = 0.06, MAPE = 0.20). Based on these performance metrics, RF has emerged as the optimal choice for modeling the ε($NO_3^-$)-drivers system. The parameters of the model directly affect the model performance. The key parameters of the RF model primarily include the number of decision trees, the maximum depth of decision trees, and the maximum number of features (Fig. S11).

Here, we proposed the MNRF model to iteratively screen for pivotal drivers via four filter indicators and further quantify the effect of drivers on ε($NO_3^-$). The high dimensions of independent variables elevate the complexity of the model, resulting in overfitting. Judicious selection of features will improve model performance, expedite training, and enhance the interpretability of results. Hence, feature selection stands as a crucial data preprocessing process, delineating the upper limit of model performance. Here, the MNRF model, comprising K layers, follows a distinct architecture. The initial K−1 layer established a nonlinear response to all base variables (potential influence factors such as ions, gaseous, etc.). The aim is to sieve out indispensable drivers that substantially influence model performances via recursive feature elimination. Unlike conventional applications of RF that rely on MSE as the sole metric for evaluating variable importance, our model operates under the constraints set by multiple indicators. Four indicators are chosen to quantify the importance of each variable in the initial K-1 layer of MNRF: (1) increase in MSE, (2) increase in node purity (NP), (3) *P* value, and (4) mean minimal depth (MMD) of variable (for more information, see Text S6) [23]. These four indicators are harnessed to pinpoint noteworthy variables from diverse perspectives that contributed to comprehensively choosing prominent variables. This holistic method circumvents potential biases that could emerge from reliance on a single filter indicator (Fig. 2a and b). The findings spotlight certain variables, $NH_4^+$, $SO_4^{2-}$, $NH_3$, $Cl^-$, and $Ca^{2+}$ exhibited a significant increase in MSE and NP. The associated statistically significant *P* values also confirmed their importance (Fig. 2a). Moreover, the distribution of MMD suggested that $NH_4^+$, $SO_4^{2-}$, and $NH_3$ appeared more frequently toward the roots of the tree than other variables (Fig. 2b). Overall, it can be concluded that $NH_4^+$, $SO_4^{2-}$, $NH_3$, $Cl^-$, and $Ca^{2+}$ were found to be essential variables, they were selected for next step in the estimation of ε($NO_3^-$). After features screening, the model's performance has also been improved, with $R^2$ ranging from 0.75 to 0.81 and RMSE ranging

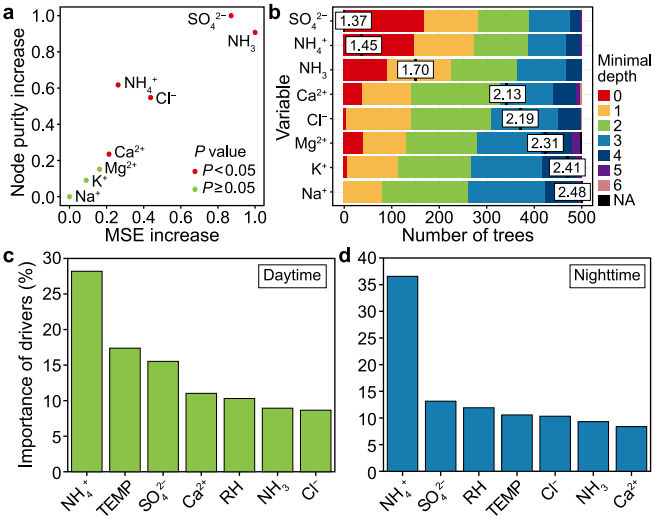


**Fig. 2.** **a**, Three of the four filter indicators on the importance of variables, including the mean squared error (MSE), node purity (NP), and *P* values. The importance of $K^+$, $Mg^{2+}$, and $Na^+$ were not evident. **b**, Each variable's mean minimal depth. $SO_4^{2-}$, $NH_4^+$, and $NH_3$ were closer to the root node than other variables. **c**–**d**, The importance of screened drivers during daytime (**c**) and nighttime (**d**). The multilevel nested random forest (MNRF) model can iteratively screen for important drivers via four filter indicators and further quantify the importance of drivers on ε($NO_3^-$).

from 0.63 to 0.55 (Fig. S13). The importance of each driver, including screened drivers and meteorological conditions, was estimated separately for daytime and nighttime by MNRF-PI (Fig. 2c and d). The results revealed that $NH_4^+$ played a primary role for ε($NO_3^-$) for both daytime (PI = 28.2%) and nighttime (PI = 36.6%), followed by the temperature at daytime (17.4%) and $SO_4^{2-}$ at nighttime (13.2%). The importance of RH at nighttime (PI = 11.8%) was significantly higher than that during daytime (PI = 10.3%) owing to the dependence of heterogeneous reaction rates on RH, which is attributed to the role of water film in the aerosol surface during gas uptake [45].

MNRF further revealed the individual effect of each driver on ε($NO_3^-$) during daytime and nighttime, which provided insights into relevant processes governing the conversion gas-to-particle of nitrate. The MNRF-PDP algorithm was used here to analyze the sensitivity of ε($NO_3^-$) to these drivers. The individual effect of drivers is shown in Fig. 3a–f and Figs. S14–S15. $NH_4^+$ positively impacted the variations in ε($NO_3^-$), consistent with theoretical knowledge that $NH_4^+$ can significantly alter the aerosol pH (Fig. 3a and b). When $NH_4^+$ concentrations were between 0 and 30 μg $m^{-3}$, ε($NO_3^-$) at nighttime was between 0.53 and 0.91, slightly higher than in daytime (0.51–0.85). When $NH_4^+$ concentrations exceeded

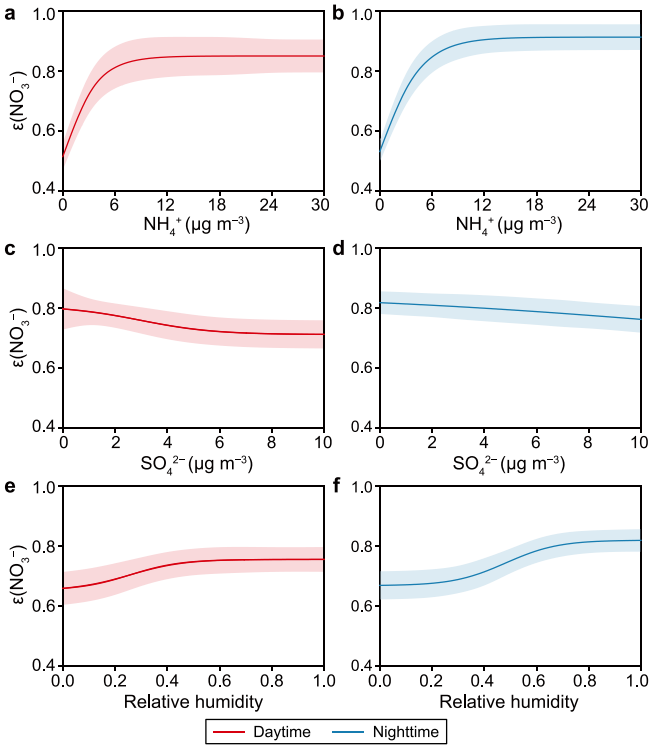


**Fig. 3.** Sensitivity curves of each effect of the driver on $\varepsilon(NO_3^-)$ during daytime and nighttime, as estimated by MNRF-PDP. **a**–**b**, The individual effect of $NH_4^+$ on $\varepsilon(NO_3^-)$ during daytime (**a**) and nighttime (**b**). **c**–**d**, The individual effect of $SO_4^{2-}$ on $\varepsilon(NO_3^-)$ during daytime (**c**) and nighttime (**d**). **e**–**f**, The individual effect of RH on $\varepsilon(NO_3^-)$ during daytime (**e**) and nighttime (**f**).

12 μg $m^{-3}$, $\varepsilon(NO_3^-)$ remained stable at 0.85 during daytime and 0.91 at nighttime. This stability hints at the occurrence of thermodynamic equilibrium, indicating that the system reached a balanced state in terms of nitrate conversion. The above MNRF-PDP results agreed with our previous works [44]: moderate $NH_4^+$ levels can enhance the aerosol pH, while rich $NH_4^+$ results in a stable pH level. Additionally, enhanced aerosol acidity also inhibits the transformation of $HNO_3$ (g) to the particle phase. An inverse relationship was observed between $SO_4^{2-}$ and $\varepsilon(NO_3^-)$, where higher $SO_4^{2-}$ concentrations corresponded to lower values of $\varepsilon(NO_3^-)$ (Fig. 3c and d). Our previous work showed that $SO_4^{2-}$ is negatively associated with aerosol pH [47]. The enhanced aerosol acidity limits the conversion of $HNO_3$ (g) into the particle phase. In the range of 0–10 μg $m^{-3}$ of $SO_4^{2-}$ concentrations, $\varepsilon(NO_3^-)$ varied in the range of 0.71–0.80 at daytime and 0.76–0.82 at nighttime. RH was positively correlated with $\varepsilon(NO_3^-)$ (Fig. 3e and f). Such a relationship was likely caused by the highly hydrophilic properties of secondary inorganic aerosols; that is, RH can promote the hygroscopic growth of nitrate mainly by increasing liquid water content in aerosol [45,48,49]. $\varepsilon(NO_3^-)$ was higher at nighttime (0.69–0.82) than at daytime (0.66–0.76) with the same levels of relative humidity.

The relationship between $NH_3$ and $\varepsilon(NO_3^-)$ was considerably more complex. As an abundant alkaline gas in the atmosphere, $NH_3$ can regulate aerosol acidity, thus affecting the gas-to-particle conversion of nitrate. However, its impact on aerosol pH was relatively weak when pH was above 4, attributed to the low $\varepsilon(NH_4^+)$ levels observed at higher aerosol pH [50]. The temperature negatively impacted the changes of $\varepsilon(NO_3^-)$. Higher temperature favored retention of nitric acid in the gas phase, and the range of variation in $\varepsilon(NO_3^-)$ was slightly higher during daytime (0.81–0.72) than at nighttime (0.80–0.75) with the same level of temperature. Increasing $K^+$ and decreasing $Cl^-$ favored the partition of $HNO_3$ (g) into the particle phase, though the impacts of these ions were relatively weak. In summary, MNRF-PDP provided quantitative information on the individual effects of drivers on $\varepsilon(NO_3^-)$ and boosted the interpretability of the machine learning model.

The atmospheric environment is a complex system. The inconsistencies between observed $\varepsilon(NO_3^-)$ and theoretical $\varepsilon(NO_3^-)$* were most likely caused by the interference of ambient factors. Here, we also quantified the impacts of ambient factors (herein referred to as $Effe_{amb}$) using MNRF. The estimated $Effe_{amb}$ for different drivers during daytime and nighttime were shown in Figs. S16–S17. Detailed descriptions were provided in Text S7. As discussed above, quantitative evidence on ambient factors further indicated that there were still high differences between the results from MNRF and those calculated by TA, reaffirming the need to constrain MNRF models with theoretical mechanisms to enhance their physical representativeness.

### 3.4. MNRF-TA method enhanced interpretability of machine learning results

The outcomes of MNRF-PDP were mathematically averaged values of all samples, which contain considerable uncertainties from ambient factors. They may deviate away from the expected actual results. To address this issue, a novel method called MNRF guided by a theoretical approach was developed. This method combined prior knowledge with data-driven models to further search for the globally optimal physical result that replaces the mathematical result in the complex atmosphere. The two previously described filter indicators, $VDS_{\varepsilon(NO_3^-)_j^n}$ in equation (4) and $DDS_{\varepsilon(NO_3^-)_j^n}$ in equation (5), were utilized here.

The effect of drivers on $\varepsilon(NO_3^-)$ calculated by RF (herein referred to as effect from RF) and the expected actual effect of drivers do share similarities. However, significant gaps still existed, which may be due to the interference of ambient factors. After applying the MNRF-TA method (results referred to as the effect from MNRF-TA), the gap was significantly reduced and was more consistent with the expected actual effect. The screened variables were analyzed separately, as shown in Fig. 4. For $NH_4^+$ (Fig. 4a,c), both effects from RF and the expected actual effect exhibited similar growth trends, although discrepancies were still observed. On the other hand, the effect from MNRF-TA was considerably closer to the expected actual effect (Fig. 4b,d). The positive effect of MNRF-TA for $NH_4^+$ may have originated from the semi-volatility of $NH_4NO_3$. Under low pH conditions, $NO_3^-$ prefers the gas phase; nitrate tends to convert into the particle phase as pH increases. Theoretically, $NH_4^+$ contributed to the formation of $NH_4NO_3$ aerosol by neutralizing $HNO_3$ (g). From the perspective of ion conversion, aerosol acidity can be enhanced by acquiring one additional $H^+$, leading to the conversion of $HNO_3$ (g) to $NO_3^-$ and ultimately resulting in the loss of one $H^+$ from the gas phase [50]. Hence, in acidic environments, aerosols are formed [44,48,49]. For RH (Fig. S18), the gap is greatly narrowed between the effect from RF and the expected actual effect, which was eliminated mainly by MNRF-TA. From the effect from MNRF-TA, in the range of 30–80% of RH, increasing relative humidity promoted nitrate formation, a finding consistent with past studies [51]. However, $\varepsilon(NO_3^-)$ decreased under high humidity conditions (RH > 90%), here water content in the aerosol increased, which competed for available reaction sites on the particle surface, resulting in the inhibition of nitrate heterogeneous reaction (Fig. S19) [52]. As for $SO_4^{2-}$, the gap was relatively more significant between the effect from RF and the expected actual effect: the expected actual effect was greater than that of RF before the guidance of the theoretical approach (Fig. S20).

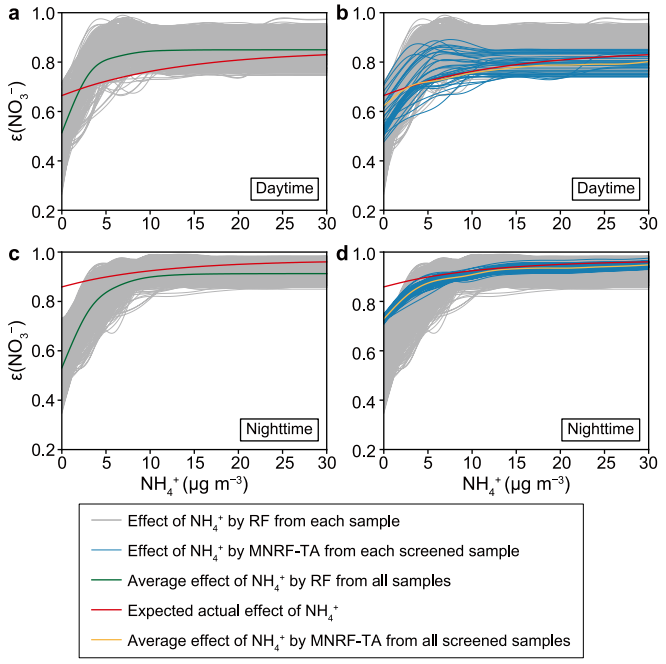


**Fig. 4.** Results of MNRF-TA (supervised MNRF guided by theoretical approach). The gray lines were the effect of drivers for all samples (during daytime and nighttime). The green lines were the outcome of regular RF (MNRF in this case), which were simple mathematically averaged results for all samples. The red lines were the expected actual effect of drivers on $\varepsilon(NO_3^-)$, as calculated by the theoretical approach. The blue lines were the effect of drivers for some samples screened by the MNRF-TA method. The yellow lines were the outcome of MNRF-TA, which were averaged results of all the blue lines. **a**, **c**, The effect of $NH_4^+$ calculated by RF and expected actual effect of $NH_4^+$ during daytime (**a**) and nighttime (**c**). **b**, **d**, The effect of $NH_4^+$ calculated by MNRF-TA and expected actual effect of $NH_4^+$ during daytime (**b**) and nighttime (**d**). Significant differences existed between regular RF results and expected actual results, while MNRF-TA can significantly reduce discrepancies, indicating that the results calculated by MNRF-TA are much more consistent with expected actual results.

Moreover, agreements were substantially improved between the effect of MNRF-TA and the expected actual effect. The trends of MNRF-TA plots for $SO_4^{2-}$ were also reasonable: the negative effect of $SO_4^{2-}$ on $\varepsilon(NO_3^-)$ may be due to the hygroscopicity and acidity of aerosol that was affected by $SO_4^{2-}$. In addition, $SO_4^{2-}$ had a higher priority to combine with $NH_4^+$ than $NO_3^-$, which inhibits the partitioning of nitrate into the particle phase. Temperature curves also had an apparent gap between the effect from RF and the expected actual effect (Fig. S21), which MNRF-TA helped to reduce. The effect of temperature on $\varepsilon(NO_3^-)$ had more significant changes during daytime than at nighttime, possibly due to the involvement of photochemical reactions. The effect of temperature from MNRF-TA agreed with the theoretical principle: higher temperatures were expected to enhance nitrate volatilization into the gas phase, leading to lower $\varepsilon(NO_3^-)$. In contrast, the low temperature would enhance nitrate formation by shifting gas-particle equilibrium. Moreover, we applied MNRF-TA to calculate the importance of drivers for $\varepsilon(NO_3^-)$, and its results were compared with RF, as shown in Fig. 5. $NH_4^+$ still played a primary role for $\varepsilon(NO_3^-)$ for both daytime (PI = 30%) and nighttime (PI = 40%); while the importance of temperature (PI = 18%) and $NH_3$ (PI = 14%) was enhanced by MNRF-TA at daytime, the importance of RH (PI = 15%) and $NH_3$ (PI = 10%) was enhanced by MNRF-TA at nighttime. Interestingly, some unimportant drivers, such as $Cl^-$ and $Ca^{2+}$, were downweighed by MNRF-TA. In summary, while regular RF can identify drivers' importance, MNRF-TA can further highlight the importance of critical drivers.

In conclusion, these findings demonstrate that MNRF-TA yielded

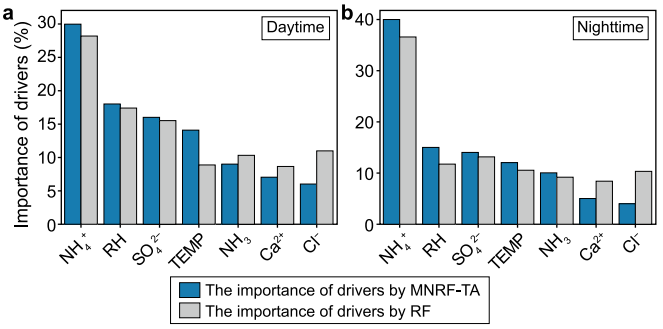


**Fig. 5.** The comparison of the importance of drivers between MNRF-TA and RF during daytime (**a**) and nighttime (**b**). MNRF-TA further emphasizes the importance of significant drivers and downweighs some irrelevant drivers compared with regular RF.

significantly improved result compared to traditional machine learning method. This enhancement increased physical representativeness, meaning the model's predictions closely aligned with real-world observations. This study further enriched the application of machine learning methods and guided a better understanding of nitrate formation in the ambient environment.

## 4. Conclusions

The gas-to-particle conversion of nitrate ($\varepsilon(NO_3^-)$) was known to be influenced by multiple drivers within the intricate ambient atmospheres. Machine learning methods offer a promising avenue to address the highly nonlinear relationship between $\varepsilon(NO_3^-)$ and its drivers. A regular random forest model was first proposed to screen for essential variables automatically via recursive feature elimination. The result showed that $NH_4^+$, $SO_4^{2-}$, RH, temperature, $NH_3$, $Ca^{2+}$, and $Cl^-$ played primary roles for $\varepsilon(NO_3^-)$ during daytime and nighttime. It emerged that $NH_4^+$, RH, $NH_3$, and $Ca^{2+}$ generally exhibit positive correlations with $\varepsilon(NO_3^-)$, whereas $SO_4^{2-}$, temperature, and $Cl^-$ demonstrate negative associations with $\varepsilon(NO_3^-)$. These findings corroborate established theoretical insights into the process. While the regular RF can adeptly identify the pivotal drivers, it falls short in bridging the gap between regular RF's results and the expected actual result. Since regular RF is data-driven, together with interference from ambient factors, the physical representativeness of results can be concealed. Here, we propose a novel approach (a multilevel nested random forest guided by a theoretical approach (MNRF-TA)), aimed at elucidating physically meaningful results by pursuing globally optimal solutions. The gaps mentioned above were significantly ameliorated using MNRF-TA, thus enhancing alignment with theoretical knowledge. Meanwhile, MNRF-TA can further highlight the importance of crucial drivers and downweigh some irrelevant ones compared with the regular RF.

This study extends the applications of machine learning methods, focusing on the MNRF-TA model, to unravel the intricacies of nitrate gas-to-particle conversion. The outcomes demonstrated a robust alignment with theoretical principles and provided enhanced elucidation of nitrate gas-to-particle conversion in the atmosphere. Notably, this augmentation in clarity improves the interpretability of the findings. The pivotal conclusion drawn from this study is that machine learning guided by theoretical knowledge holds substantial promise in handling complex environmental issues. By infusing pertinent theoretical knowledge into the machine learning method, the outcomes gleaned exhibit heightened reliability and coherence. This strategy stands as a testament to the potency of merging theoretical knowledge with advanced computational methodologies. Furthermore, this study

underscores the critical importance of further inquiries into the theoretical mechanisms stemming from the laboratory. Such investigations can refine and validate the theoretical underpinnings, ultimately bolstering the robustness of the MNRF-TA model. These endeavors are indispensable for establishing a comprehensive grasp of the potential and limitations of the methodology. Moving forward, the advancements in knowledge-guided ML methods will be pivotal for gaining deeper insights into the underlying process and enable significant progress towards solving pressing environmental challenges and promoting sustainable practices.

## CRediT authorship contribution statement

**Bo Xu:** Conceptualization, Investigation, Data Curation, Formal Analysis, Writing - Original Draft. **Haofei Yu:** Formal Analysis, Writing - Review & Editing. **Zongbo Shi:** Formal Analysis, Writing - Review & Editing. **Jinxing Liu:** Formal Analysis, Writing - Review & Editing. **Yuting Wei:** Investigation, Data Curation. **Zhongcheng Zhang:** Investigation, Code. **Yanqi Huangfu:** Formal Analysis, Writing - Review & Editing. **Han Xu:** Investigation, Code. **Yue Li:** Code, Support Machine Learning Methods. **Linlin Zhang:** Writing - Review & Editing, Resources, Supervision. **Yinchang Feng:** Resources, Funding Acquisition. **Guoliang Shi:** Design the Study, Project Administration, Resources, Funding Acquisition, Supervision.

## Declaration of competing interest

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

## Acknowledgement

This study was financially supported by the National Natural Science Foundation of China (42077191), the National Key Research and Development Program of China (2022YFC3703400), the Blue Sky Foundation, Tianjin Science and Technology Plan Project (18PTZWHZ00120), and Fundamental Research Funds for the Central Universities (63213072 and 63213074).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at. https://doi.org/10.1016/j.ese.2023.100333.

## References

[1] F. Wang, W. Wang, Z. Wang, Z. Zhang, Y. Feng, A.G. Russell, et al., Drivers of PM2.5-O3 co-pollution: from the perspective of reactive nitrogen conversion pathways in atmospheric nitrogen cycling, Sci. Bull. 67 (2022) 1833–1836.
[2] Z. An, R.J. Huang, R. Zhang, X. Tie, G. Li, J. Cao, et al., Severe haze in northern China: a synergy of anthropogenic emissions and atmospheric processes, P Natl Acad. Sci. USA 116 (2019) 8657–8666.
[3] J. Peng, M. Hu, D. Shang, Z. Wu, Z. Du, T. Tan, et al., Explosive secondary aerosol formation during severe haze in the north China plain, Environ. Sci. Technol. 55 (2021) 2189–2207.
[4] Y. Cheng, G. Zheng, C. Wei, Q. Mu, B. Zheng, Z. Wang, et al., Reactive nitrogen chemistry in aerosol water as a source of sulfate during haze events in China, Sci. Adv. 2 (2016) e1601530.
[5] Y. Wang, Y. Sun, Z. Zhang, Y. Cheng, Spatiotemporal variation and source analysis of air pollutants in the Harbin-Changchun (HC) region of China during 2014–2020, Environ. Sci. Ecotechnol. 8 (2021) 100126.
[6] Y. Xie, G. Wang, X. Wang, J. Chen, J. Gao, Observation of nitrate dominant PM2.5 and particle pH elevation in urban Beijing during the winter of 2017, Atmos. Chem. Phys. (2019) 1–25.
[7] H. Li, Q. Zhang, B. Zheng, C. Chen, N. Wu, H. Guo, et al., Nitrate-driven urban haze pollution during summertime over the North China Plain, Atmos. Chem. Phys. 18 (2018) 5293–5306.
[8] J.H. Seinfeld, S.N. Pandis, Atmospheric Chemistry and Physics: from Air Pollution to Climate Change, Wiley-Interscience, 1998.
[9] Y. Liu, J. Zhan, F. Zheng, B. Song, Y. Zhang, W. Ma, et al., Dust emission reduction enhanced gas-to-particle conversion of ammonia in the North China Plain, Nat. Commun. 13 (2022) 6887.
[10] X. Fu, T. Wang, J. Gao, P. Wang, L. Xue, Persistent heavy winter nitrate pollution driven by increased photochemical oxidants in northern China, Environ. Sci. Technol. 7 (2020) 3881–3889.
[11] R.J. Weber, H. Guo, A.G. Russell, A. Nenes, High aerosol acidity despite declining atmospheric sulfate concentrations over the past 15 years, Nat. Geosci. 9 (2016) 282–285.
[12] P. Vasilakos, A. Russell, R. Weber, A. Nenes, Understanding nitrate formation in a world with less sulfate, Atmos. Chem. Phys. 18 (2018) 12765–12775.
[13] Q. Mu, M. Shiraiwa, M. Octaviani, N. Ma, A. Ding, H. Su, et al., Temperature effect on phase state and reactivity controls atmospheric multiphase chemistry and transport of PAHs, Sci. Adv. 4 (2018) p7314.
[14] A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, et al., Advancing mathematics by guiding human intuition with AI, Nature 600 (2021) 70–74.
[15] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, et al., Deep learning and process understanding for data-driven Earth system science, Nature 566 (2019) 195–204.
[16] Z. Li, H. Liu, C. Zhang, G. Fu, Generative adversarial networks for detecting contamination events in water distribution systems using multi-parameter, multi-site water quality monitoring, Environ. Sci. Ecotechnol. 14 (2023) 100231.
[17] M. Callaghan, C. Schleussner, S. Nath, Q. Lejeune, T.R. Knutson, M. Reichstein, et al., Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies, Nat. Clim. Change 11 (2021) 966–972.
[18] G. Chen, Q. Cheng, T.W. Lyons, J. Shen, F. Agterberg, N. Huang, et al., Reconstructing Earth's atmospheric oxygenation history using machine learning, Nat. Commun. 13 (2022) 5862.
[19] Z. Shi, C. Song, B. Liu, G. Lu, R.M. Harrison, Abrupt but smaller than expected changes in surface air quality attributable to COVID-19 lockdowns, Sci. Adv. 7 (2021) d6696.
[20] L. Hou, Q. Dai, C. Song, B. Liu, F. Guo, T. Dai, et al., Revealing drivers of haze pollution by explainable machine learning, Environ. Sci. Technol. Lett. 9 (2022) 112–119.
[21] Z. Zhang, B. Xu, W. Xu, F. Wang, J. Gao, Y. Li, et al., Machine learning combined with the PMF model reveal the synergistic effects of sources and meteorological factors on PM2.5 pollution, Environ. Res. 212 (2022) 113322.
[22] Z. Sun, A.T. Archibald, Multi-stage ensemble-learning-based model fusion for surface ozone simulations: a focus on CMIP6 models, Environ. Sci. Ecotechnol. 8 (2021) 100124.
[23] F. Yu, C. Wei, P. Deng, T. Peng, X. Hu, Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles, Sci. Adv. 7 (2021) f4130.
[24] P.C. Hanson, A.B. Stillman, X. Jia, A. Karpatne, H.A. Dugan, C.C. Carey, et al., Predicting lake surface water phosphorus dynamics using process-guided machine learning, Ecol. Model. 430 (2020) 109136.
[25] O.P. Abimbola, G.E. Meyer, A.R. Mittelstet, D.R. Rudnick, T.E. Franz, Knowledge-guided machine learning for improving daily soil temperature prediction across the United States, Vadose Zone J. 20 (2021) e20151.
[26] C. Irrgang, N. Boers, M. Sonnewald, E.A. Barnes, C. Kadow, J. Staneva, et al., Towards neural Earth system modelling by integrating artificial intelligence in Earth system science, Nat. Mach. Intell. 3 (2021) 667–674.
[27] H. Guo, R. Otjes, P. Schlag, A. Kiendler-Scharr, A. Nenes, R.J. Weber, Effectiveness of ammonia reduction on control of fine particle nitrate, Atmos. Chem. Phys. 18 (2018) 12241–12256.
[28] H. Guo, J. Liu, K.D. Froyd, J.M. Roberts, P.R. Veres, P.L. Hayes, et al., Fine particle pH and gas–particle phase partitioning of inorganic species in Pasadena, California, during the 2010 CalNex campaign, Atmos. Chem. Phys. 17 (2017) 5703–5719.
[29] A. Nenes, S.N. Pandis, C. Pilinis, ISORROPIA: a new thermodynamic equilibrium model for multiphase multicomponent inorganic aerosols, Aquat. Geochem. 4 (1998) 123–152.
[30] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[31] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[32] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Springer, Berlin, 2001.
[33] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (2010) 1340–1347.
[34] D. Shang, J. Peng, S. Guo, Z. Wu, M. Hu, Secondary aerosol formation in winter haze over the Beijing-Tianjin-Hebei Region, China, Front. Environ. Sci. Eng. 15 (2020) 34.
[35] Y. Cheng, Q. Yu, J. Liu, Y. Sun, L. Liang, Z. Du, et al., Formation of secondary inorganic aerosol in a frigid urban atmosphere, Front. Environ. Sci. Eng. 16 (2021) 18.
[36] S. Guo, M. Hu, M.L. Zamora, J. Peng, D. Shang, J. Zheng, et al., Elucidating severe urban haze formation in China, P Natl Acad. Sci. USA 111 (2014) 17373–17378.
[37] Q. Xu, S. Wang, J. Jiang, N. Bhattarai, X. Li, X. Chang, et al., Nitrate dominates the chemical composition of PM2.5 during haze event in Beijing, China, Sci. Total Environ. 689 (2019) 1293–1303.
[38] H. Guo, A.P. Sullivan, P. Campuzano-Jost, J.C. Schroder, F.D. Lopez-Hilfiker, J.E. Dibb, et al., Fine particle pH and the partitioning of nitric acid during

winter in the northeastern United States, J. Geophys. Res. 121 (10) (2016) 310–355, 376.
[39] X. Wang, Y. Zhang, H. Chen, X. Yang, J. Chen, F. Geng, Particulate nitrate formation in a highly polluted urban area: a case study by single-particle mass spectrometry in Shanghai, Environ. Sci. Technol. 43 (2009) 3061.
[40] Z. Zhang, H. Guan, L. Luo, N. Zheng, H. Xiao, Response of fine aerosol nitrate chemistry to Clean Air Action in winter Beijing: Insights from the oxygen isotope signatures, Sci. Total Environ. 746 (2020) 141210.
[41] M. Li, Z. Zhang, T. Wang, M. Xie, Y. Han, Nonlinear responses of particulate nitrate to NOx emission controls in the megalopolises of China, Atmos. Chem. Phys. 21 (2021) 15135–15152.
[42] X. Shi, A. Nenes, Z. Xiao, S. Song, H. Yu, G. Shi, et al., High-resolution data sets unravel the effects of sources and meteorological conditions on nitrate and its gas-particle partitioning, Environ. Sci. Technol. 53 (2019) 3048–3057.
[43] Y. Tao, J.G. Murphy, The sensitivity of PM2.5 acidity to meteorological parameters and chemical composition changes: 10-year records from six Canadian monitoring sites, Atmos. Chem. Phys. 19 (2019) 9309–9320.
[44] G. Shi, J. Xu, X. Shi, B. Liu, X. Bi, Z. Xiao, et al., Aerosol pH dynamics during haze periods in an urban environment in China: use of detailed, hourly, speciated observations to study the role of ammonia availability and secondary aerosol formation and urban environment, J. Geophys. Res. Atmos. 124 (2019) 9730–9742.
[45] J. Sun, L. Liu, L. Xu, Y. Wang, Z. Wu, M. Hu, et al., Key role of nitrate in phase transitions of urban particles: implications of important reactive surfaces for secondary aerosol formation, J. Geophys. Res. Atmos. 123 (2018) 1234–1243.
[46] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: Proceedings of the 23rd International Conference on Machine Learning, Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, 2006, pp. 161–168.
[47] G. Shi, J. Xu, X. Peng, Z. Xiao, K. Chen, Y. Tian, et al., pH of aerosols in a polluted atmosphere: source contributions to highly acidic aerosol, Environ. Sci. Technol. 51 (2017) 4289–4296.
[48] J. Gao, Y. Wei, G. Shi, H. Yu, Z. Zhang, S. Song, et al., Roles of RH, aerosol pH and sources in concentrations of secondary inorganic aerosols during different pollution periods, Atmos. Environ. 241 (2020) 117770.
[49] X. Peng, P. Vasilakos, A. Nenes, G. Shi, Y. Qian, X. Shi, et al., Detailed analysis of estimated pH, activity coefficients, and ion concentrations between the three aerosol thermodynamic models, Environ. Sci. Technol. 53 (2019) 8903–8913.
[50] Q. Zhao, A. Nenes, H. Yu, S. Song, Z. Xiao, K. Chen, et al., Using high-temporal-resolution ambient data to investigate gas-particle partitioning of ammonium over different seasons, Environ. Sci. Technol. 54 (2020) 9834–9843.
[51] J. Stutz, B. Alicke, R. Ackermann, A. Geyer, S. Wang, A.B. White, et al., Relative humidity dependence of HONO chemistry in urban areas, J. Geophys. Res. Atmos. 109 (2004).
[52] X. Wang, W. Wang, L. Yang, X. Gao, W. Nie, Y. Yu, et al., The secondary formation of inorganic aerosols in the droplet mode through heterogeneous aqueous reactions under haze conditions, Atmos. Environ. 63 (2012) 68–76.