Atmospheric Environment 289 (2022) 119323

Contents lists available at ScienceDirect

Atmospheric Environment

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

Prediction of CCN spectra parameters in the North China Plain using a random forest model

Minghua Liang [a,b,1], Jiangchuan Tao [a,b,*,1], Nan Ma [a,b,**], Ye Kuang [a,b], Yanyan Zhang [a,b], Sen Wu [a,b], Xuejuan Jiang [a,b], Yao He [a,b], Chunrong Chen [c], Wenda Yang [d], Yaqing Zhou [a,b], Peng Cheng [d], Wanyun Xu [f], Juan Hong [a,b], Qiaoqiao Wang [a,b], Chunsheng Zhao [e], Guangsheng Zhou [f], Yele Sun [g], Qiang Zhang [c], Hang Su [h], Yafang Cheng [i]

[a] Institute for Environmental and Climate Research, Jinan University, 511443, Guangzhou, China
[b] Guangdong-Hongkong-Macau Joint Laboratory of Collaborative Innovation for Environmental Quality, 511443, Guangzhou, China
[c] Department of Earth System Science, Tsinghua University, 100084, Beijing, China
[d] Institute of Mass Spectrometry and Atmospheric Environment, Jinan University, 510632, Guangzhou, China
[e] Department of Atmospheric and Oceanic Sciences, School of Physics, Peking University, Beijing, China
[f] State Key Laboratory of Severe Weather, Key Laboratory for Atmospheric Chemistry, Institute of Atmospheric Composition, Chinese Academy of Meteorological Sciences, Beijing, 100081, China
[g] State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
[h] Multiphase Chemistry Department, Max Planck Institute for Chemistry, 55128, Mainz, Germany
[i] Minerva Research Group, Max Planck Institute for Chemistry, 55128, Mainz, Germany

HIGHLIGHTS

- Random forest model is used to estimate CCN spectra based on observation data in the North China Plain.
- Mass concentration of black carbon and hemispheric backscattering fraction are the most important input variables.
- Aerosol optical properties data are recommended in estimation of CCN spectra parameters.

ARTICLE INFO

Keywords:
Cloud condensation nuclei
Aerosol
Machine learning

ABSTRACT

Cloud condensation nuclei (CCN) spectra, that indicate the dependence of CCN number concentrations (N_{CCN}) on supersaturation, are important factors in aerosol–cloud interactions. Owing to the lack of direct measurements of N_{CCN}, calculating the CCN spectra based on conventional observational data of aerosols is important to obtain N_{CCN}; however, this is challenging owing to the complex relationship between aerosol properties and CCN activity. Machine learning techniques have recently been applied to estimate N_{CCN} and found to be a promising method for CCN prediction. In this study, the random forest (RF) model was applied to predict CCN spectral parameters using the observation data measured in four campaigns in the North China Plain, and the effects of chemical and optical properties of aerosol on the estimation of CCN spectra were investigated. The results show that the RF model trained with the data of one campaign can be used to estimate CCN spectral parameters in another campaign, with the coefficient of determination between the estimated and measured CCN spectral parameters being approximately 0.5. The deviations of the estimation by the RF model may result from the difference in both the aerosol properties and measurement uncertainties among different campaigns, whose influence on the deviations can be further magnified by overfitting the RF model. Further analysis revealed that the major aerosol properties among the input variables of the RF model were the mass concentration of black carbon and aerosol hemispheric backscattering fraction at a wavelength of 450 nm. In addition, the roles of chemical compositions of aerosol in estimating CCN spectra parameters are different among different campaigns

* Corresponding author. Institute for Environmental and Climate Research, Jinan University, 511443, Guangzhou, China.
** Corresponding author. Institute for Environmental and Climate Research, Jinan University, 511443, Guangzhou, China.
E-mail addresses: taojch@jnu.edu.cn (J. Tao), nan.ma@jnu.edu.cn (N. Ma).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.atmosenv.2022.119323
Received 15 April 2022; Received in revised form 1 August 2022; Accepted 3 August 2022
Available online 12 August 2022

because the poor correlation between aerosol chemical compositions and CCN spectra parameter can affect the performance of the RF model. For estimating CCN spectra, using aerosol optical properties, including the hemispheric backscattering fraction and absorption coefficient, as model inputs is more recommended than using aerosol chemical properties.

1. Introduction

Cloud condensation nuclei (CCN) are aerosol particles that can be activated to form clouds or fog droplets under supersaturated conditions. The variations in CCN can modify micro-physical properties of clouds. Therefore, the number concentration of CCN (N_{CCN}) plays an important role in aerosol–cloud interactions, which determines the indirect radiation effects of aerosols and radiative balance of the atmosphere (Reutter et al., 2009; IPCC, AR5, 2013; Chang et al., 2015; Liu et al., 2020). Numerous studies have revealed that the ability of aerosols to form CCN (aerosol CCN activity) is primarily determined by particle size, chemical composition, and mixing state (Dusek et al., 2006; Kuwata et al., 2008; Petters and Kreidenweis, 2008; Ervens et al., 2010; Rose et al., 2010; Su et al., 2010; Rose et al., 2011). Owing to the limited applications of CCN measurements, information on N_{CCN} and its variations is still lacking for most areas (Deng et al., 2011; Lathem and Nenes, 2011; Liu and Li, 2014; Paramonov et al., 2015; Pöhlker et al., 2018). Therefore, the parameterization of N_{CCN} based on conventional aerosol measurements is important and might play an integral role in the study of aerosol–cloud interactions.

One common approach to parameterize the CCN spectra, which describes N_{CCN} as a function of supersaturated condition (Dusek et al., 2003; Petters and Kreidenweis, 2007; Andreae and Rosenfeld, 2008), is as follows:

$$N_{CCN}(SS) = C \times SS^{k} \quad (1)$$

where SS is the supersaturation ratio, and $N_{CCN}(SS)$ is the N_{CCN} under SS conditions (Twomey, 1959; Martins et al., 2009; Vié et al., 2016; Braga et al., 2017; Fults et al., 2019; Jayachandran et al., 2020; Rejano et al., 2021). The fitting parameter C represents the N_{CCN} at SS = 1%, and the parameter k describes the N_{CCN} variations under different SSs. Previous studies have revealed that C and k are affected by the loading and chemical composition of aerosol particles, respectively (Cohard et al., 1998; Jefferson, 2010; Vié et al., 2016; Rejano et al., 2021). Based on the CCN spectra observed in different regions, C is generally associated with aerosol loading, and k is related to aerosol hygroscopicity and size distributions. Specifically, the C values tend to be high (or low) under polluted (or clean) conditions, and the k values tend to be small (or large) for particles with strong (or weak) hygroscopicity and large (or small) particle sizes (Hegget al., 1991; Martins et al., 2009; Pöhlker et al., 2016; Jayachandran et al., 2020).

CCN spectra have been widely used in aerosol–cloud interaction studies using both field observations and model simulations. The cloud droplet number concentration predicted by weather and climate models based on CCN spectrum was consistent with the observations and can be used to evaluate cloud droplet formation (Twomey, 1959; Khairoutdinov and Kogan, 2000; Fountoukis and Nenes, 2005; Reutter et al., 2009; Lim and Hong, 2010; Pinsky et al., 2012). The variations in CCN spectral parameters can be used to clarify the sources and pollution conditions of aerosols in different regions (Fults et al., 2019; Fanourgakis et al., 2019). In addition, the CCN spectra can be applied in the traditional cloud micro-physical framework to reduce the computational cost and improve the accuracy of cloud prediction (Lim and Hong, 2010; Maronga et al., 2020; Vié et al., 2016; Tsarpalis et al., 2020).

CCN counter measurements at multiple SSs are required for the direct measurements of CCN spectra in the atmosphere. However, direct measurements of CCN spectra are complicated and expensive as a CCN counter requires precise and stable conditions of the flow rate, temperature, and water vapor (CCN counter manual). Therefore, CCN counters are generally used for intensive measurements at one location. Thus, the temporal and spatial resolutions of CCN spectra are poor, and calculating the CCN spectra based on conventional aerosol measurements is important with wide applications.

As aerosol CCN activity is determined by aerosol size and chemical composition, CCN spectra can be calculated using the measurements of micro-physical and chemical properties of the aerosol. For example, the optical properties of aerosol, including scattering coefficient (σ_{sp}), backscattering coefficient (σ_{bsp}), and single scattering albedo (SSA), have been used to estimate CCN spectral parameters (Jefferson, 2010; Shinozuka et al., 2015; Rejano et al., 2021). Rejano et al. (2021) found that the hemispheric backscattering fraction (HBF) and SSA were positively correlated with CCN parameters C and k, respectively, and can be used to calculate them. These positive correlations occurred because HBF and SSA can partially indicate the variations in particle size and aerosol hygroscopicity, respectively (Ghan and Collins, 2004; Jefferson, 2010; Liu and Li, 2014; Shinozuka et al., 2015; Tao et al., 2018). These results demonstrate that the optical properties of aerosol can be effectively used for predicting CCN spectral parameters.

In addition to aerosol optical properties, bulk aerosol chemical composition reportedly has a strong relationship with CCN activity (Gunthe et al., 2009; Ervens et al., 2010; Kamilli et al., 2014; Deng et al., 2018, 2019). However, there is little research on the calculation of CCN activation spectral parameters based on aerosol chemical composition, primarily because chemical composition is more relevant for aerosol hygroscopicity than N_{CCN}. In addition, appropriate methods to evaluate the relationship between CCN activation spectral parameters and aerosol chemical compositions are lacking. In previous studies, the linear relationship between CCN spectral parameters and aerosol properties has been extensively analyzed. However, the relationship between CCN spectral parameters and aerosol chemical compositions can be highly nonlinear (Nair and Yu, 2020) because aerosol CCN activity is affected by the aerosols of different diameters under supersaturated conditions (Ghan et al., 2006; Ervens et al., 2007; Chen et al., 2014; Cai et al., 2018; Zhang et al., 2020). These nonlinear relationships between the CCN spectral parameters and aerosol chemical compositions cannot be comprehensively interpreted by analyzing linear relationships. Machine learning (ML) has recently been widely used in atmospheric science research. The first attempt to apply the ML method in CCN research was by Nair and Yu (2020), followed by Nair et al. (2021). They developed an ML model using aerosol chemical compositions and meteorological data as model inputs to predict N_{CCN} under specific SSs.

In this study, ML method was applied to estimate the CCN spectral parameters based on conventional aerosol measurements, including chemical and optical properties of aerosol, conducted in four field campaigns in the North China Plain (NCP) during autumn and winter. Specifically, the application of aerosol optical properties and prediction of CCN spectral parameters discussed in this study were not conducted in previous CCN studies using ML method. As the measurements of aerosol optical properties are widely used and CCN spectral parameters are important factors in modeling, the present results are significant in CCN prediction and can be used in CCN simulation models. The remainder of the manuscript is organized as follows. The measurements obtained from the four campaigns and theory of the ML method application are described in Section 2. Section 3 presents the prediction of CCN spectral parameters using ML method based on conventional aerosol measurements. Section 4 provides a summary of this study.

2. Methodology

2.1. Observational data

In this study, the ML method was applied using data measured during the four campaigns conducted during fall and winter at the Gucheng site (39.15° N, 115.74° E) in the NCP. The four campaigns were conducted from October 18 to November 13, 2016, January 13 to February 4, 2018, November 16 to December 16, 2018, and October 1 to October 30, 2019. The Gucheng site is located approximately 100 km, 110 km, and 40 km from Beijing, Tianjin, and Baoding, respectively. This site is surrounded by farmlands and small villages. During the campaign, there were no farming activities near the site, and thus, the effects from villages were small. Within 2 km of this site, the road does not have heavy traffic, and thus, the influence of traffic can also be ignored. Thus, observations at this site represent the average pollution conditions and background aerosol properties in the NCP (Kuang et al., 2020; Li et al., 2021). Both PM_{10} and PM_1 inlets were used for aerosol sampling during the second campaign in 2018; however, this study used aerosol data sampled by only the PM_{10} inlet. More details about the four campaigns can be found in Tao et al. (2018), Li et al. (2021), Zhang et al. (2020), Tao et al. (2021), and Zhou et al. (2022).

2.1.1. Aerosol optical property data

During the 2016 and the two 2018 campaigns, aerosol optical properties were measured using a nephelometer (Aurora 3000 or TSI 3563) and an aethalometer (AE-33, Magee Scientific). In detail, σ_{sp} and σ_{bsp} were measured at three wavelengths using the nephelometer, and the absorption coefficient (σ_{ap}) at 637 nm was quantified using the aethalometer. Using the above measurements, the aerosol optical properties, including HBF, Ångström index (Å), and SSA, were calculated. HBF can be defined as the ratio of σ_{bsp} to σ_{sp}:

$$HBF = \frac{\sigma_{bsp}}{\sigma_{sp}} \tag{2}$$

SSA can be defined as the ratio of σ_{sp} and $\sigma_{ap} + \sigma_{sp}$:

$$SSA = \frac{\sigma_{sp}}{\sigma_{ap} + \sigma_{sp}} \tag{3}$$

Å can be calculated from σ_{sp} measured at two wavelengths:

$$\text{Å} = \frac{ln(\sigma_{sp,\lambda 1}) - ln(\sigma_{sp,\lambda 2})}{ln(\lambda_2) - ln(\lambda_1)} \tag{4}$$

where $\sigma_{sp,\lambda 1}$ and $\sigma_{sp,\lambda 2}$ represent the scattering coefficients at different wavelengths, and λ represents the wavelength. The three calculated aerosol optical properties were influenced by both aerosol size and chemical composition and thus can be used to indicate the variations in CCN activity (Rose et al., 2011; Tao et al., 2018; Zhang et al., 2020).

2.1.2. Data of aerosol chemical compositions

The aerosol chemical compositions, including NH_4^+, SO_4^{2-}, NO_3^-, Cl^-, and organic compounds, were measured during three field campaigns since 2018. In the January 2018 campaign, the mass concentrations of water-soluble inorganic ions and organic materials were measured using the Monitor for AeRosols and Gases in Ambient air (MARGA, model ADI, 2080) and the OCEC analyzer, respectively. During the November 2018 campaign, the aerosol chemical composition was measured using a time-of-flight aerosol chemical speciation monitor (TOF-ACSM). In the 2019 campaign, the quadrupole aerosol chemical speciation monitor (Q-ACSM) was used to measure non-refractory particulate matter. More details about the chemical measurements during the three campaigns can be found in Sun et al. (2020), Zhang et al. (2020), and Zhou et al. (2022). In addition, the mass concentration of black carbon (BC) was obtained using the aethalometer.

2.1.3. CCN spectral measurements

The CCN data were measured using a continuous-flow CCN counter (CCNC, Model CCN-200, DMT, USA; Roberts and Nenes, 2005; Lance et al., 2006). The N_{CCN} were measured at five specific SSs from 0.07% to 0.8%, and the CCN spectra were obtained by fitting N_{CCN} at the five SSs. The SS levels of CCNC were calibrated using ammonium sulfate before and after each campaign. The corrected SSs were 0.114%, 0.148%, 0.273%, 0.492%, and 0.864% during the January 2018 campaign; 0.055%, 0.074%, 0.198%, 0.444%, and 0.814% during the November 2018 campaign; and 0.066%, 0.091%, 0.208%, 0.427%, and 0.761% during the October 2019 campaign. In the four campaigns, the N_{CCN} measurement of the five SSs required 1 h for a complete cycle. The CCN spectral parameters were calculated using the measured N_{CCN} at the five SSs with a temporal resolution of 1 h. More information about CCN measurements can be found in Tao et al. (2018), Zhang et al. (2020), and Tao et al. (2021).

2.2. Theory of ML method

2.2.1. Random forest model

In this study, the RF model based on ML method was used to derive the CCN spectra. ML method is a branch of artificial intelligence technology. This method guides the computer to learn from the data, analyzes the potential laws contained in the data, and continuously improves its self-performance in the prediction of new data based on big data learning. The RF model used in this study is a typical supervised learning model using ML method. For supervised learning, the training sample data have the corresponding target results in the training process. The results of the new data can be predicted via supervised learning by establishing a connection between the data sample factors and known target results. The RF model is widely used for classification and nonlinear regression problems (Breiman, 2001). The RF model comprises multiple regression decision trees. A decision tree is a supervised ML algorithm that splits data into two subsets. The RF model constructs many decision trees from the training dataset and uses all the decision trees for predictions. Each decision tree has a random training scheme that is developed based on a certain number of samples drawn from the training dataset and a part of the input features selected for training. The final regression prediction is the average value predicted by multiple regression decision trees. The decision tree may overfit the data owing to data splitting. The RF model offers several advantages. These models are easy to train and can handle high-dimensional data. They can predict complex nonlinear relationships between features compared with linear regression models, and the feature variable importance ranking can be obtained after RF training. In this study, the RF regressor in the Python Scikit-Learn machine learning library (http://scikit-learn.org/stable/index.html) was used.

The RF model must be trained, and its performance must be verified. Therefore, the observation data were split into two parts in a 7:3 ratio, that is, training and testing datasets for the training and verification of the RF model, respectively. In the training process, there were two important parameters: the number of decision trees (n_estimators) and maximum number of features to consider (max_feature). The model prediction generally improves with large values of n_estimator and max _feature. However, when these two parameters become larger than a certain value, the increased n_estimators (or max_feature) value leads to a high computational cost without any improvement in model prediction. Hence, appropriate values of both parameters are required for the optimized performance of ML method (Kuang et al., 2018, 2019; Nair and Yu, 2020; Nair et al., 2021). In this study, the n_estimators was set to 460 and max_feature was set to zero, implying that all input features rather than a random subset were considered.

In this study, the CCN spectral parameters C and k were estimated from the conventional measurements of aerosol properties using ML method. In ML method, various conventional measured aerosol properties can be together used to calculate CCN spectral parameters, which

cannot be achieved by analyzing the linear relationships because of the highly nonlinear relationship between CCN spectral parameters and chemical and optical properties of aerosol (Nair and Yu, 2020). Specifically, because the parameter C varies with the variations in aerosol loading, it can be estimated using aerosol parameters, such as σ_{sp}, σ_{bsp}, and mass concentration of aerosol chemical compositions, in the ML method. Parameter k is primarily influenced by the aerosol size and aerosol hygroscopicity; thus, it is estimated by aerosol parameters such as SSA, HBF, and the mass fraction of aerosol chemical compositions.

2.2.2. Quantification of the importance of input variables in the RF model

To evaluate the influence of the measured micro-physical and chemical properties of aerosol on the estimation of CCN spectra, the variable importance measure (VIM) analysis of the observational data acting as the input variables of the RF model was conducted. The VIM results can quantify the effects of different variables on the results during the prediction process. Analyzing the VIMs by RF can improve the understanding of the input variables, prediction ability, and reliability of the model. In this study, the VIM analysis was performed based on the mean decrease in the accuracy of the ML algorithm to compute the feature importance by permuting out-of-bag (OOB) samples (data that were not used in training the decision tree). The VIM can directly indicate the impact of each input variable on the model prediction accuracy (Breiman, 2001). During the RF training, decision trees were built by sampling repeatedly from the dataset. The OOB data can be used to evaluate the performance of the decision trees and calculate the prediction accuracy of the model, which is called the OOB error (OOB_{err}). The VIM of a certain input variable (i.e., feature importance) can be calculated as follows:

$$VIM_i^{ER} = \frac{1}{N_{tree}} \sum_{t=1}^{N_{tree}} (OOB_{err2} - OOB_{err1}) \quad (5)$$

where N_{tree} is the number of decision trees, OOB_{err1} is the error calculated using the corresponding OOB data, and OOB_{err2} is the OOB data error calculated by adding noise interference to the features of all OOB data samples. If the prediction accuracy of OOB data is significantly reduced by adding noise to a certain input variable, this feature will have a significant impact on the prediction results (i.e., high importance).

In the VIM calculation, multiple computations are required, and a series of feature importance must be calculated. The advantage of this method is that it can predict the importance of unknown input features when the model is overfitted. However, the disadvantages are high computational cost and the need of multiple model runs to obtain reliable and stable feature importance values. Furthermore, for large amount of data, the model requires long running time (Breiman, 2001).

3. Results and discussion

Based on the CCN activation spectra and conventional aerosol data observed in the four field campaigns in the NCP, CCN activation spectra were predicted using the ML method. First, the characteristics of the N_{CCN} and CCN spectra measured during the four campaigns were presented. Subsequently, the RF model was trained using the measured CCN spectral parameters and observed aerosol properties and used to calculate the CCN spectral parameters during the two campaigns in 2018. To examine the applicability of ML method, the RF model was trained using observation data from a single campaign and used to estimate the CCN spectral parameters in another campaign. Subsequently, VIM analysis was performed to analyze the role of conventional aerosol data as the input variables in predicting CCN spectral parameters. Finally, in the cases where only aerosol chemical data or aerosol optical data were available, the prediction of CCN spectral parameters was also evaluated based on the ML method using the observation data from the four campaigns.

3.1. Overview of measured CCN spectra

Fig. 1 shows the measured N_{CCN} and CCN spectra obtained by fitting the N_{CCN} measured at the five SSs during the four field campaigns. The CCN spectral curves in Fig. 1 were calculated using the average CCN spectral parameters in each campaign. Owing to the heavy aerosol pollution in the NCP, the N_{CCN} at an SS of 0.8% reached approximately 10^4 cm^{-3}. During the campaign from November to December 2018, the N_{CCN} at a lower supersaturation ratio (less than 0.1%) was significantly higher than 10^3 cm^{-3} because of the large number of accumulation-mode aerosols formed during secondary aerosol formation (Kuang et al., 2020; Tao et al., 2021).

In addition, the CCN spectra greatly varied among different campaigns. For the CCN spectra, the fitting average N_{CCN} of the CCN spectra of the 2019 campaign was relatively low compared to those of other campaigns. This is because, the aerosol concentration level in the NCP has significantly decreased owing to the recent efforts of emission control (Zhang et al., 2019; Lei et al., 2021). Moreover, the average fitting N_{CCN} of the CCN spectra observed in autumn and winter field campaigns were higher than those in other seasons, such as the November–December 2018 campaign, because of secondary aerosol formation and anthropogenic emission enhancement in autumn and winter. During these seasons, the N_{CCN} at different SSs are expected to increase significantly (Zhang et al., 2019; Li et al., 2021). The variations in aerosol pollution in these campaigns were also confirmed by the mass concentrations of aerosol chemical compositions and aerosol optical properties during each period, as shown in Figs. S1 and S2.

Fig. 2 shows the probability distribution function (PDF) of the CCN spectral parameters C and k during the four campaigns. There was little difference in the values of k among the four field campaigns, and the peak k values ranged from 0.3 to 0.5. The k value indicates the variations in N_{CCN} at different SSs, affected by the particle size and hygroscopicity of the dominant aerosol types. The p-value of the PDF of parameter k was approximately 1, indicating similarity among the distributions of k in the different campaigns. Therefore, the variations in k values in the four campaigns indicated that the differences in particle size and hygroscopicity of dominant aerosol types were not significant among the four campaigns. However, there was a small difference in the peak of the PDFs of parameter k in the different campaigns. For example, parameter k was typically larger than 0.5 in the first campaign in 2018, suggesting that more CCN were distributed in small particle size ranges, which may be due to the new particle formation events (Zhang et al., 2020). There was a large difference in parameter C among the four campaigns, with average values of 13400, 13400, 18700, and 7600 cm^3, respectively. As the C value was primarily affected by the bulk aerosol number concentration, the different C values indicated different pollution conditions in the four campaigns.

3.2. Prediction of CCN spectra parameters by applying ML method

The RF model was trained using the measured CCN spectral parameters and conventional aerosol observation data and used to estimate the CCN spectral parameters. Specifically, parameter C was estimated using aerosol parameters σ_{sp} and σ_{bsp} and mass concentration of aerosol chemical compositions, and parameter k was estimated using aerosol parameters SSA and HBF and mass fraction of aerosol chemical compositions (Table 1). Fig. 3 shows a comparison between the CCN activated spectral parameters and their predicted values by the RF model trained with data from the same campaign. Overall, good agreement was obtained for the two campaigns in 2018. For both k and C, the coefficient of determination was approximately 0.9, p-values were less than 0.001, deviations of the predicted values from the measured values were within 30%, and normalized mean bias (NMB) was less than 2%. However, this good agreement between the estimated and measured values of the CCN spectral parameters in the same campaign may result from the overfitting of the RF model. In the following section,

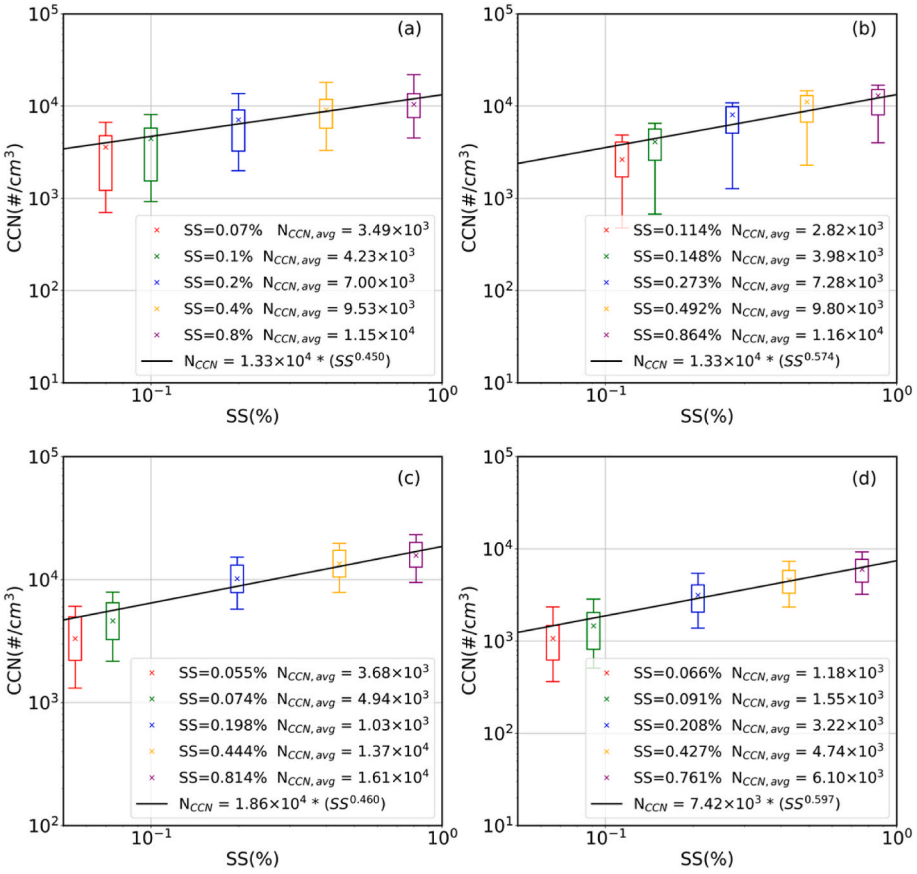

Fig. 1. Cloud condensation nuclei (CCN) spectra obtained based on measured CCN number concentrations (N_{CCN}) under five supersaturation ratios (SSs) in the four field campaigns in the North China Plain: (a) October–November 2016; (b) January–February 2018; (c) November–December 2018; (d) October–November 2019. The boxes and whiskers represent the 10, 25, 75, and 95 percentiles. The cross represent the median value of N_{CCN} under five SSs (the color represents the SSs), and the black solid line is the CCN spectra fitted by the mean value. The average values of N_{CCN} under each SSs are also shown in the legend. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

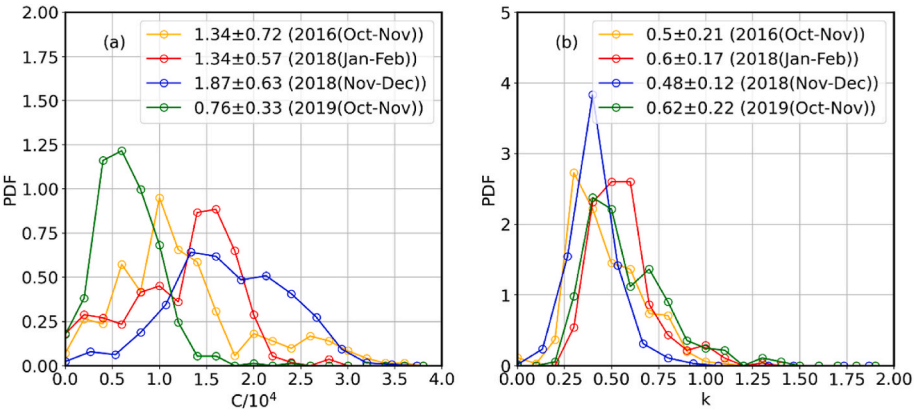

Fig. 2. Probability distribution function (PDF) of cloud condensation nuclei (CCN) spectral parameters. The numbers are mean values and corresponding standard deviations. The colors represent four different field campaigns, yellow: November–December 2016; red: January–February 2018; blue: November–December 2018; and green: October–November 2019. (a) Parameter C; (b) Parameter k. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Training features used in the estimation of cloud condensation nuclei (CCN) spectral parameters.

CCN spectral parameters	Training features
C	m_{BC}; σ_{sp} (450 and 525 nm); σ_{bsp} (450 and 525 nm); and mass concentrations of NO_3^-, SO_4^{2-}, Cl^-, NH_4^+, and Org
k	HBF (450 and 525 nm); SSA; Ångström index; and mass fractions of NO_3^-, SO_4^{2-}, Cl^-, NH_4^+ and Org

m_{BC}: mass concentration of black carbon, σ_{sp}: scattering coefficient, σ_{bsp}: backscattering coefficient, HBF: hemispheric backscattering fraction, SSA: single scattering albedo.

we have discussed this by applying the trained RF model to other campaigns.

The trained RF models were used to estimate the CCN spectral parameters in another campaign in 2018 (Fig. 4). The differences between the measured CCN spectral parameters and their estimated values were larger than those in Fig. 3. The coefficient of determination and NMB were ~0.7 and ~40% for parameter C and ~0.5 and ~5% for parameter k, respectively. The corresponding p-values were less than 0.001, indicating significant correlations between the measured CCN spectra and predicted CCN spectra using the RF model trained by another campaign. Large differences may result from the differences in the aerosol properties and measurement uncertainties among different campaigns, which can be magnified by the overfitting in the training of the RF model. As shown in Figs. 6, S3, and S4, the correlations between the CCN

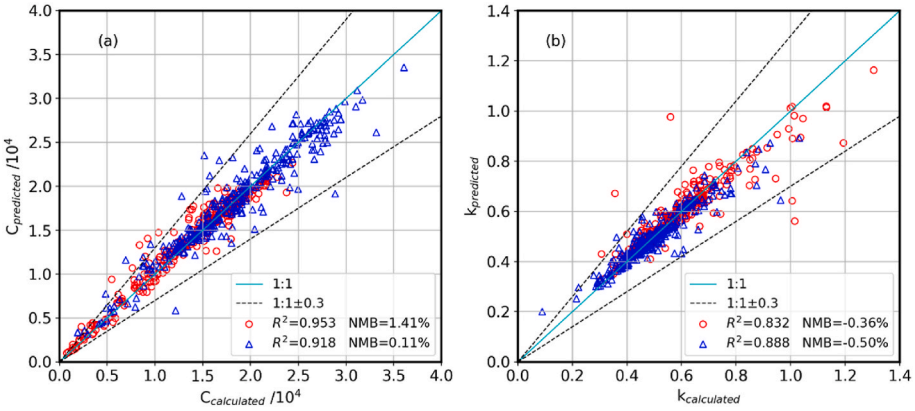

Fig. 3. Comparison between cloud condensation nuclei (CCN) spectral parameters fitted from CCN number concentrations (N_{CCN}) measurement and estimated by the random forest (RF) model trained by observation data in the same campaign: (a) C; (b) k. The colors represent the two field campaigns in 2018, red: January–February 2018; blue: November–December 2018. The black dashed lines are the 30% error line. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

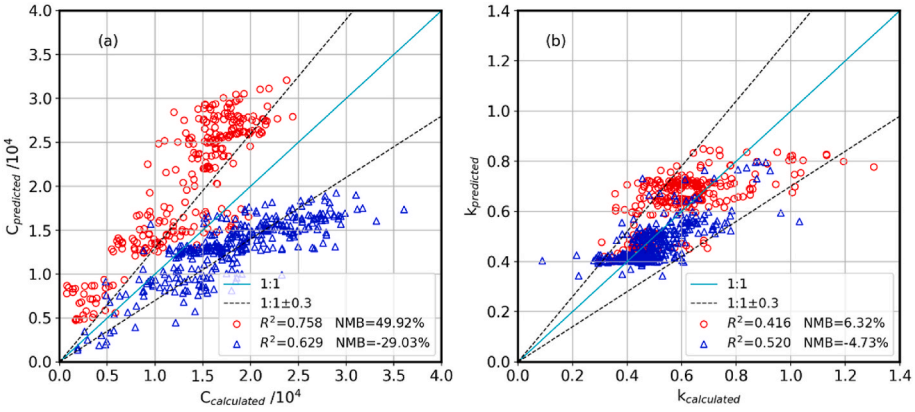

Fig. 4. Comparison between cloud condensation nuclei (CCN) spectral parameters calculated from CCN number concentrations (N_{CCN}) measurement and predicted by the random forest (RF) model trained by observation data: (a) C; (b) k. The colors represent two different field campaigns, red: comparison in the January–February 2018 campaign with the RF model trained in the November–December 2018 campaign; blue: comparison in the November–December 2018 campaign with the RF model trained in the January–February 2018 campaign. The black dashed lines are the 30% error line. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

spectral parameters and corresponding input parameters were different in the two campaigns, primarily because of the differences in microphysical and chemical properties of aerosol as well as the measurement uncertainty. These correlations are the basis of the RF model training, which comprises multiple regression decision trees. This may lead to different prediction results for the RF model trained in different campaigns. In addition, if the RF model is trained with the dataset of both campaigns, the estimated CCN spectral parameters correspond well with their calculated values, as shown in Fig. S5. However, the good performance of the trained model (similar to that shown in Fig. 3) may also result from the overfitting of the RF algorithm. Thus, long-term measurement data rather than multiple campaign data are required for the training of the RF model for better application of ML method in estimating CCN spectral parameters. Even if there are large deviations in the estimated CCN spectral parameters by applying the RF model, it can be used to evaluate the roles of each measured aerosol property on the estimation of CCN spectral parameters, which are important for further studies.

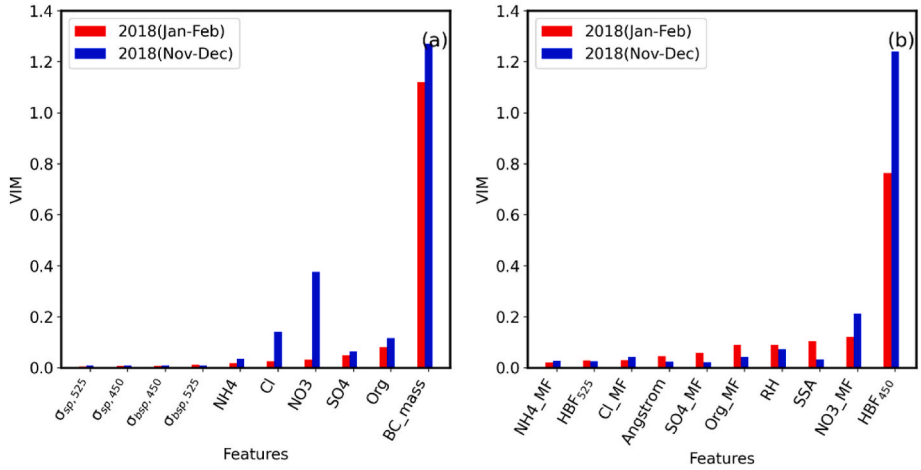

Fig. 5. Importance of input variables in the machine learning (ML) method based on variable importance measure (VIM) analysis for predicting cloud condensation nuclei (CCN) spectral parameters: (a) C; (b) k. The colors represent two different field campaigns. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

By calculating the VIM of each input parameter, their impact on the prediction of the CCN spectral parameters was analyzed using the ML method (Fig. 5). For predicting C and k, the VIMs of the mass concentration of BC (m_{BC}) and HBF at 450 nm (HBF_{450}) were the highest in both campaigns, indicating that m_{BC} and HBF_{450} are the major input variables in the estimation of CCN parameters in the NCP. The significant influence of m_{BC} on parameter C may be the result of the dominant particle size range of BC, 150–500 nm in the NCP (Zhao et al., 2019). Particles in this size range have a stronger influence on CCN activity than particles beyond this size range. Compared with BC particles, organic and inorganic compounds are formed with considerably larger particle size in the NCP (Kuang et al., 2020), whose influences on CCN activity are smaller, resulting in weaker influences of their mass concentrations on the prediction of parameter C. These organic and inorganic compounds dominate the scattering ability of the bulk aerosol; thus, the influence of σ_{sp} on the prediction of parameter C is also weak. The HBF is affected by the size and chemical composition of aerosols (Jefferson, 2010; Ma et al., 2011; Kuang et al., 2018), which are also important factors affecting the CCN activity of aerosols (Köhler, 1936; Dusek et al., 2006). The results of this study highlight the importance of HBF in the prediction of CCN spectra. For predicting parameters C and k, the mass concentration and mass fraction of NO_3^- had the second highest VIMs and can be important in the RF model. Parameter k is significantly affected by aerosol hygroscopicity. Among the aerosol chemical compositions observed in the NCP, both sulfate and nitrate were important components with strong hygroscopicity. In recent years, air pollution control has drastically reduced the concentrations of SO_2 and sulfate aerosols in the NCP (Lei et al., 2021). However, nitrate is a major chemical compound in aerosol particles in the NCP. Therefore, the effect of sulfate on CCN activation becomes weaker than that of nitrate, and the VIM of sulfate in predicting CCN spectral parameters becomes smaller than that of nitrate.

Furthermore, considering the high importance of m_{BC} and HBF_{450} in the estimation of CCN spectra parameters, the relationships between m_{BC}, HBF_{450}, and CCN spectral parameters were analyzed. Fig. 6 shows the correlation of C–m_{BC} and k–HBF_{450} levels. For the correlation between C and m_{BC}, the coefficient of determination was ~0.6 in the two campaigns. For the variations in parameter k, except when HBF was greater than 0.2, most of the data exhibited a linear relationship between k and HBF, with a coefficient of determination of ~0.6. The correlations between two CCN spectral parameters and each input parameter other than m_{BC} and HBF are shown in Fig. S3 (for parameter C) and S4 (for parameter k). They had a lower coefficient of determination compared with those of C–m_{BC} and k–HBF correlations. Compared with the coefficients of determination between the estimated and measured CCN spectral parameters shown in Fig. 3, the coefficient of determination between C–m_{BC} and k–HBF_{450} was lower. In previous CCN studies, relationships similar to those in Fig. 6 were used to estimate CCN spectral parameters from conventional observation data (Jefferson, 2010; Shinozuka et al., 2015). Thus, the present results show that the estimation of the CCN spectral parameters using only m_{BC} or HBF_{450} is less applicable than the trained RF model, highlighting the application of ML method in estimating CCN spectral parameters.

3.3. Prediction of CCN spectral parameters by different data combinations

In aerosol observations, there may be situations where the measurements of only chemical properties or optical properties of aerosol are available. For example, there were no aerosol chemical data in the 2016 campaign and no aerosol optical data in the 2019 campaign. In these cases, the RF model trained above cannot be used to estimate the CCN spectral parameters. Therefore, when limited number of aerosol instrument is available, the RF model must be trained based on different data combinations to estimate the CCN spectral parameters. In this study, two data combinations were considered based on the data combinations provided by different instruments: aerosol optical properties (AOP) and aerosol chemical composition (ACC) data (Table 2). To estimate parameter C, the AOP dataset included σ_{sp}, σ_{bsp}, and m_{BC}, and the ACC dataset included the mass concentrations of inorganic and organic components. To estimate parameter k, the AOP dataset included HBF, SSA, and Ångström, and the ACC dataset included the mass fraction of the inorganic and organic components. Different datasets were then used to train the RF model to predict the CCN spectral parameters. Using different datasets, the VIM of each input variable and applicability of the ML method to predict the CCN spectra were evaluated.

In this study, the RF model trained based on both the AOP and ACC datasets was compared with only one campaign, while the RF model

Table 2
Combination of datasets provided by different instrument.

Dataset	Training features for C	Training features for k	Campaign
Aerosol optical properties (AOP)	m_{BC}, σ_{sp} (450 and 525 nm), and σ_{bsp} (450 and 525 nm)	HBF (450 and 525 nm), SSA, Ångström	2016 (Oct–Nov), 2018 (Jan–Feb), 2018 (Nov–Dec)
Aerosol chemical compositions (ACC)	Mass concentrations of NO_3^-, SO_4^{2-}, Cl^-, NH_4^+, and Org	Mass fractions of NO_3^-, SO_4^{2-}, Cl^-, NH_4^+, and Org	2018 (Jan–Feb), 2018 (Nov–Dec), 2019 (Oct–Nov)

m_{BC}: mass concentration of black carbon, σ_{sp}: scattering coefficient, σ_{bsp}: backscattering coefficient, HBF: hemispheric backscattering fraction, SSA: single scattering albedo.

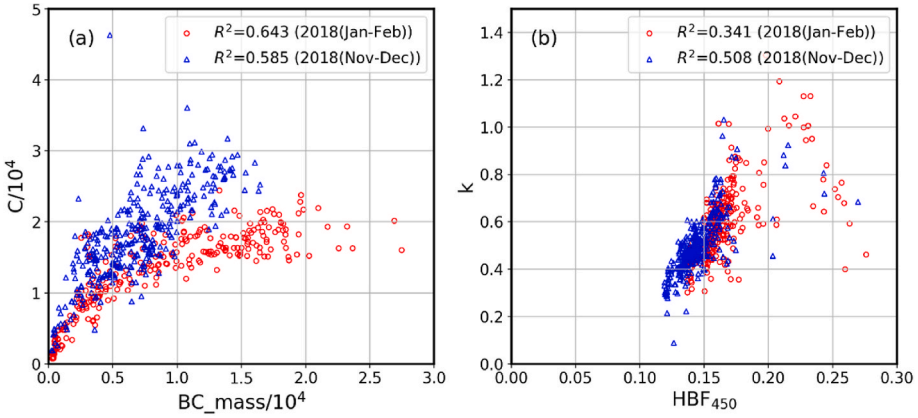

Fig. 6. Correlation between (a) CCN spectral parameter C and mass concentration of black carbon (m_{BC}) and (b) CCN spectra parameter k and hemispheric backscattering fraction (HBF_{450}). The colors represent two different field campaigns. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

trained based on only the AOP or ACC dataset was compared with two campaigns. This is because, both AOP and ACC data were available only in two campaigns in 2018, and the training of the RF model based on only the AOP or ACC dataset was available in three campaigns. As overfitting can affect the evaluation of the RF model, we focused on the training based on only one campaign rather than multiple campaigns. The RF model likely performed well by training using datasets (ACC, AOP, or both datasets) from multiple campaigns (Figs. S5, S6, and S7). However, this good performance primarily resulted from the overfitting of the RF model, similar to the RF model shown in Fig. 3. Therefore, comparing the RF models trained by datasets from different campaigns is necessary to evaluate the performance of the RF model.

The CCN spectral parameters were compared with the values estimated by the RF model trained with the AOP or ACC dataset in other campaigns (Table 3). Compared with the RF model trained by both aerosol optical and chemical data (Fig. 4), the RF model trained using only AOP or ACC dataset led to large deviations in the estimated CCN spectral parameters. While comparing parameter C using AOP dataset from each campaign, the coefficient of determination was approximately 0.5, and the NMB was in the range of 10–60%. While comparing parameter k using AOP dataset, the coefficient of determination was approximately 0.4, and the NMB was lower than 33%. In all the cases, the p-values were lower than 0.001, indicating significant correlations. For the ACC dataset, the coefficient of determination and NMB were similar to those for the AOP dataset in the comparison of parameter C. However, in the comparison of parameter k, the coefficient of determination and NMB were considerably worse. For example, a low coefficient of determination of approximately 0.01, high p-value considerably larger than 0.05, and high NMB of approximately 100% were obtained when the RF model trained by the observation data in 2019 was applied to the second campaign in 2018. The large deviations for the ACC dataset suggest the less application of the RF model trained by the ACC data than that trained by the AOP data.

In addition, the accuracy of the RF model significantly varied with the measurement period. This may be because the RF model predictions are primarily driven by the correlations between the conventional observation variables and CCN spectral parameters. As mentioned in the discussion of Fig. 4, the accuracy of the RF model was determined by the difference in the aerosol properties, measurement uncertainty among different campaigns, and magnified influence of these uncertainties on the prediction of CCN spectral parameters due to overfitting in the RF model training. The correlations between the CCN spectral parameters and corresponding input parameters were different in the two campaigns (Figs. 6, S3, and S4), primarily because of the differences in micro-physical and chemical properties of aerosol. The aerosol variables that are crucial for model prediction and have a high correlation with the CCN spectral parameter in one campaign may have a poor correlation in other campaigns, leading to significant uncertainties in the prediction accuracy of the RF model. For example, HBF_{450} and the mass fraction of NO_3 in the conventional observation dataset played an important role in the prediction of parameter k according to the VIM shown in Fig. 5; however, their correlation with k was different in the two campaigns, leading to inconsistency and bias in the prediction of k in Fig. 4.

Fig. 7 shows the importance of input variables in the prediction of parameter C. For the AOP dataset, m_{BC} was the key input variable, even in the 2016 campaign, similar to the training results using all datasets as shown in Fig. 5. For the ACC dataset, the mass fraction of organic compounds was the key input variable in the RF model with the highest VIM; however, the differences in the VIM among different chemical compositions were smaller than those among different AOP (Fig. 7a). A high correlation between C and BC mass concentration was observed during the two campaigns, which can be attributed to the predominant size of BC particles. In the NCP, BC particles were primarily distributed with sizes in the range of 150–500 nm (Zhao et al., 2019; Yang et al., 2022). In this particle size range, the variations in aerosol property number concentration can lead to large variations in N_{CCN} in the NCP (Deng et al., 2011; Ma et al., 2016; Tao et al., 2021). As parameter C represents CCN loading, the high correlation between the BC mass concentration and parameter C was expected. A similarly high correlation between N_{CCN} and BC mass concentration was observed in other polluted regions in China under the conditions of high BC mass concentration (Leng et al., 2014).

Fig. 8 shows the importance of input variables in the prediction of parameter k. For the AOP dataset, the importance of HBF_{450} was the highest, which is similar to the training results based on all datasets as shown in Fig. 5. For the ACC dataset, the VIMs of different aerosol chemical components acting as input variables were significantly different among the three campaigns. The key input variable in the RF model was Cl^- mass fraction in the January–February 2018 campaign but NO_3^- mass fraction during November–December 2018 and October–November 2019 campaigns. In addition, in the winter campaign from January to February 2018, the differences among the VIMs of mass fractions of different chemical species were generally small.

The different VIMs of aerosol chemical components in different seasons may result from the variations in the emissions and secondary formation of different aerosol chemical components, which can vary with atmospheric conditions, such as ambient relative humidity (Cheng et al., 2016; Kuang et al., 2020). In previous ML studies of CCN, there was a large bias of chemical compositions in the prediction of N_{CCN} (Nair and Yu, 2020). In this study, the VIMs of nitrate were particularly low in the first campaign in 2018, during which the correlation between ACCs and CCN spectral parameters was considerably less than those in the other campaigns. Sensitivity analysis revealed that when the RF model is

Table 3
Application of the random forest (RF) model trained by different datasets.

Dataset	Training data	Testing data	C			k		
			R^2	NMB	P	R^2	NMB	P
AOP	2016 (Oct–Nov)	2018 (Jan–Feb)	0.625	6.37%	<0.01	0.377	4.13%	<0.01
		2018 (Nov–Dec)	0.524	−25.7%	<0.01	0.523	−1.07%	<0.01
	2018 (Jan–Feb)	2016 (Oct–Nov)	0.467	2.21%	<0.01	0.342	1.03%	<0.01
		2018 (Nov–Dec)	0.575	−33.1%	<0.01	0.575	−33.0%	<0.01
	2018 (Nov–Dec)	2016 (Oct–Nov)	0.433	52.8%	<0.01	0.344	−0.80%	<0.01
		2018 (Jan–Feb)	0.795	53.8%	<0.01	0.449	10.1%	<0.01
ACC	2018 (Jan–Feb)	2018 (Nov–Dec)	0.587	−33.4%	<0.01	0.210	11.7%	<0.01
		2019 (Oct–Nov)	0.428	−47.7%	<0.01	0.0005	−7.47%	0.802
	2018 (Nov–Dec)	2018 (Jan–Feb)	0.537	47.8%	<0.01	0.267	−0.073%	<0.01
		2019 (Oct–Nov)	0.461	13.1%	<0.01	0.213	−44.6%	<0.01
	2019 (Oct–Nov)	2018 (Jan–Feb)	0.748	10.4%	<0.01	0.0737	68.7%	<0.01
		2018 (Nov–Dec)	0.522	−21.5%	<0.01	0.0153	98.5%	0.003
AOP + ACC	2018 (Jan–Feb)	2018 (Nov–Dec)	0.763	49.6%	<0.01	0.421	6.62%	<0.01
	2018 (Nov–Dec)	2018 (Jan–Feb)	0.632	−28.8%	<0.01	0.509	−5.14%	<0.01

AOP: Aerosol optical properties, ACC: Aerosol chemical compositions, NMB: normalized mean bias.

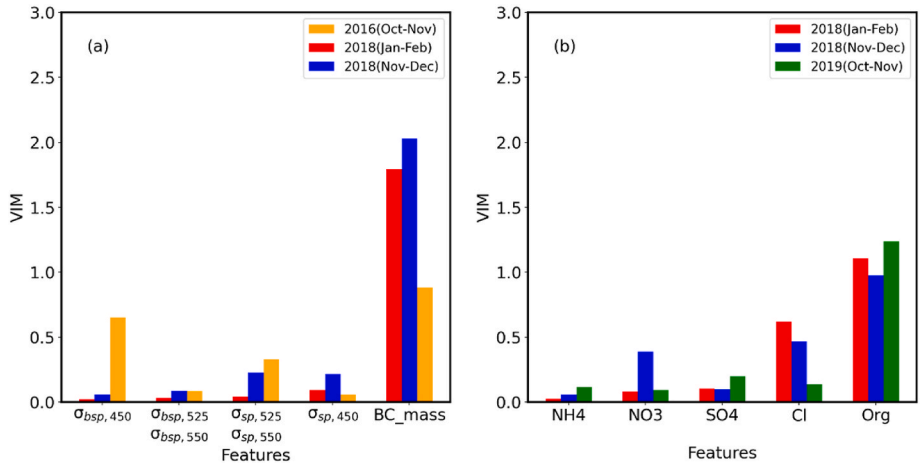

Fig. 7. Variable importance measure (VIM) of the two data combinations in the cloud condensation nuclei (CCN) spectra normalization parameter prediction: (a) aerosol optical property (AOP) dataset and (b) aerosol chemical composition (ACC) dataset. The colors of bars represent the four campaigns. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

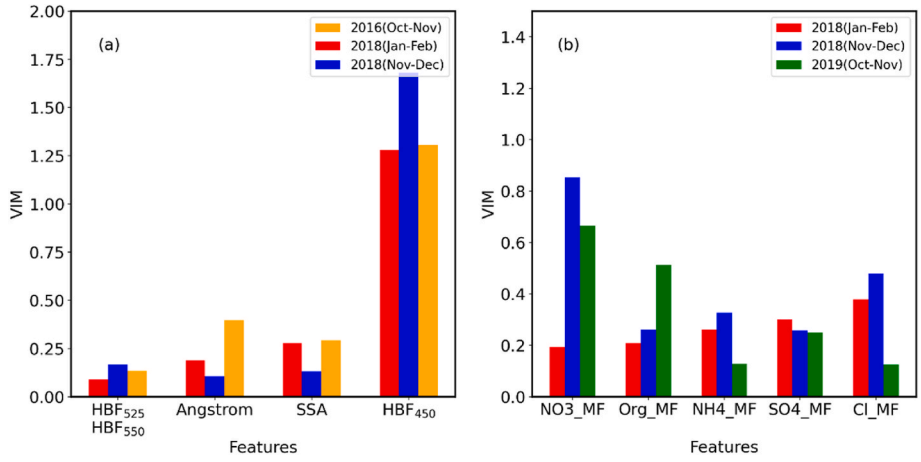

Fig. 8. Variable importance measure (VIM) of the two data combinations in the prediction of CCN spectra parameter k: (a) aerosol optical property (AOP) dataset and (b) aerosol chemical composition (ACC) dataset. The colors of bars represent the four campaigns. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

trained using a dataset with a poor correlation between the input parameters and target parameters, the VIMs of the input parameters may be unreasonable. Therefore, whether nitrate can act as a good predictor in the RF model is determined by the correlation between ACCs (not just nitrate) and parameter k in the training dataset. The variations in the correlation between ACCs and parameter k in the different campaigns were primarily due to the variations in the predominant particle size ranges of CCN in these campaigns, which were dominated by different particle formation mechanisms.

As shown in Fig. S4, the correlation between the mass fraction of NO_3^- and parameter k in the first campaign in 2018 was considerably less than those in the other two campaigns. Although the correlation of nitrate with k was better than that of other chemical compositions in the first campaign in 2018, it may still be insufficient for the training of a reasonable RF model. Therefore, the role of nitrate in the RF model trained in the first campaign in 2018 may be unreasonable and thus different from that in the RF models trained in other campaigns. Consequently, nitrate had the lowest VIM among the different ACCs in the first campaign in 2018, although nitrate had highest correlation coefficient with parameter k. Furthermore, the VIM of nitrate significantly decreased in the first campaign in 2018 compared to those of the other two campaigns, explaining the varied roles of nitrate in different campaigns (Fig. 8).

To investigate the influence of poor correlation between the chemical composition and parameter k in the training of the RF model, HBF was added to the ACC dataset to train the RF model. The VIMs of these input parameters indicated that nitrate was the major chemical component in the RF model in the first campaign in 2018 (Fig. S8). However, in this campaign, nitrate had low VIM in the RF model trained using the ACC dataset (Fig. 8). The varied role of nitrate in the two RF models (trained by the ACC dataset or the ACC + HBF dataset) suggests that the above-mentioned RF model trained by ACC datasets in the first campaign in 2018 may be unreasonable, likely because of the poor correlation between ACCs and CCN spectral parameters. In addition, there were significant variations in the VIMs of nitrate in different campaigns, and nitrate was not a good predictor in the 2019 campaign. These variations also resulted from the variations in the correlation between nitrate and parameter k, primarily because of the variations in the predominant particle size ranges of CCN in different campaigns.

Particle size is the key parameter in determining particle CCN activity (Dusek et al., 2006). However, bulk aerosol chemical composition is more affected by the accumulation mode particles in the NCP (Liu et al., 2014; Tao et al., 2018). Therefore, the effect of bulk aerosol chemical composition on particle CCN activity depends on the variations in the predominant particle size ranges of CCN (Farmer et al., 2015; Cheng et al., 2016; Kuang et al., 2020; Tao et al., 2021). During the four

campaigns conducted, the dominant particle formation processes were different, which led to the variations in the predominant particle size range of the CCN. During the winter campaign (January–February 2018), there were new particle formation events, and CCN were primarily distributed in the Aiken mode (Zhang et al., 2020). During the autumn campaigns (November–December 2018 and October–November 2019), there was evident secondary aerosol formation owing to multiple-phase chemical reactions, and CCN were primarily distributed in the accumulation mode (Tao et al., 2021; Wu et al., 2022). Thus, in the first campaign in 2018, the correlation between NO_3^- mass fraction and parameter k of CCN spectra was weaker than those in other campaigns. However, interpreting the correlation between chemical composition and CCN spectral parameters is not the topic of this study and will be discussed in future. The significantly different VIMs among the different campaigns suggest the weak applicability of the RF model trained by the ACC dataset in the estimation of parameter k.

In summary, in the estimation of CCN spectral parameters in different campaigns, the roles of each ACC are less consistent than those of each AOP. Thus, the ACC data are less applicable in the estimation of CCN spectral parameters compared with AOP data. Furthermore, the results highlight the application of AOP data in the acquisition of CCN data.

4. Conclusions

In this study, using the conventional aerosol observations obtained from the NCP, the prediction of the CCN activation spectral parameters was investigated based on the ML method using the RF algorithm.

To examine the applicability of the ML method, the RF model was trained using measured data from one campaign and used to estimate the CCN spectral parameters in another campaign. And the RF model is trained with data from just one campaign rather than multiple campaigns because the performance of RF model can be overestimated by the overfitting. The results show that the RF model trained in one campaign can be used to estimate the CCN spectral parameters in another campaign. The coefficient of determination (which is verified by significance test) and NMB were ~0.7 and ~40% for parameter C and ~0.5 and ~5% for parameter k, respectively. The deviations of the estimated CCN spectral parameters from their values by applying the RF model trained in another campaign may result from the differences in the aerosol properties and measurement uncertainties among different campaigns, whose influence on the deviations can be further magnified owing to overfitting in the training of the RF model. In addition, these cross-validation results can be used as a reference to evaluate the difference between the N_{CCN} estimated by the RF models and measured N_{CCN}. As indicated by the cross-validation results, the uncertainty in the estimated NCCN can be ~50%. If the systematic deviations in different campaigns due to the overfitting of the RF mode are ignored, the uncertainty in the estimated NCCN can be ~30%. The analysis of the importance of different aerosol data acting as input variables in the RF model reveals that, in all campaigns, m_{BC} and HBF at 450 nm are the key variables for predicting CCN spectral parameters C and k, respectively. Although the accuracy of the CCN spectral parameters estimated by the trained RF model in different campaigns in this study was not sufficient, the roles of the optical and chemical properties of aerosol in estimating CCN spectral parameters were confirmed among different campaigns. In addition, based on different combinations of conventional aerosol observation data, the application of ML method to estimate the CCN activation spectral parameters was studied. The results show that, compared with AOP data, ACC data are less applicable in the estimation of CCN spectral parameters, owing to the inconsistent roles of each ACC in different campaigns. The role of each ACC in the RF model is determined by the correlation between ACCs and parameter k in the training dataset, which can be affected by predominant particle size ranges of CCN in these campaigns. Sensitivity analysis revealed that when the RF model is trained using a dataset with a poor correlation between the input parameters and target parameters, the VIMs of the input parameters may be unreasonable. Thus the poor correlation between ACCs and parameter k limits the application of ACC data in the RF model, the application of AOP data in the estimation of CCN spectral parameters is recommended.

This study provides suggestions for the observation of CCN spectra and N_{CCN}, particularly for detecting N_{CCN} over large temporal and spatial ranges. In addition, because the CCN spectral parameters investigated in this study have wide applications in cloud models, the present results are helpful in the simulation models of CCN.

CRediT authorship contribution statement

Minghua Liang: Formal analysis, Investigation, Visualization, Writing – original draft. **Jiangchuan Tao:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Nan Ma:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – review & editing. **Ye Kuang:** Conceptualization, Data curation, Investigation, Resources, Supervision, Writing – review & editing. **Yanyan Zhang:** Data curation, Investigation. **Sen Wu:** Data curation, Investigation. **Xuejuan Jiang:** Data curation, Investigation. **Yao He:** Data curation, Investigation. **Chunrong Chen:** Data curation, Investigation. **Wenda Yang:** Data curation, Investigation. **Yaqing Zhou:** Investigation. **Peng Cheng:** Data curation, Investigation, Resources. **Wanyun Xu:** Data curation, Investigation, Resources. **Juan Hong:** Investigation. **Qiaoqiao Wang:** Funding acquisition. **Chunsheng Zhao:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision. **Guangsheng Zhou:** Project administration. **Yele Sun:** Project administration, Resources, Writing – review & editing. **Qiang Zhang:** Resources, Project administration. **Hang Su:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision. **Yafang Cheng:** Conceptualization, Funding acquisition, Project administration, Resources, 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.

Data availability

Data will be made available on request.

Acknowledgements

This research has been supported by the Fundamental Research Funds for the Central Universities (grant no. 21620420), the National Natural Science Foundation of China (grant nos. 41805110 and 91644218), the Ministry of Science and Technology of the People's Republic of China (grant no. 2017YFC0210104), the Guangdong Innovative and Entrepreneurial Research Team Program (Research team on atmospheric environmental roles and effects of carbonaceous species, grant no. 2016ZT06N263), the Special Fund Project for Science and Technology Innovation Strategy of Guangdong Province (grant no. 2019B121205004), and the Basic Research Fund of CAMS (grant no. 2020Z002).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2022.119323.

References

Andreae, M.O., Rosenfeld, D., 2008. Aerosol–cloud–precipitation interactions. Part 1. The nature and sources of cloud-active aerosols[J]. Earth Sci. Rev. 89 (1–2), 13–41.

Braga, R.C., Rosenfeld, D., Weigel, R., et al., 2017. Comparing parameterized versus measured microphysical properties of tropical convective cloud bases during the ACRIDICON–CHUVA campaign[J]. Atmos. Chem. Phys. 17 (12), 7365–7386.

Breiman, L., 2001. Random forests[J]. Mach. Learn. 45 (1), 5–32.

Cai, M., Tan, H., Chan, C.K., et al., 2018. The size-resolved cloud condensation nuclei (CCN) activity and its prediction based on aerosol hygroscopicity and composition in the Pearl Delta River (PRD) region during wintertime 2014[J]. Atmos. Chem. Phys. 18 (22), 16419–16437.

Chang, D., Cheng, Y., Reutter, P., Trentmann, J., Burrows, S.M., Spichtinger, P., Nordmann, S., Andreae, M.O., Pöschl, U., Su, H., 2015. Comprehensive mapping and characteristic regimes of aerosol effects on the formation and evolution of pyro-convective clouds. Atmos. Chem. Phys. 15, 10325–10348. https://doi.org/10.5194/acp-15-10325-2015.

Chen, J., Zhao, C.S., Ma, N., et al., 2014. Aerosol hygroscopicity parameter derived from the light scattering enhancement factor measurements in the North China Plain[J]. Atmos. Chem. Phys. 14 (15), 8105–8118.

Cheng, Y.F., Zheng, G., Wei, C., Mu, Q., Zheng, B., Wang, Z., Gao, M., Zhang, Q., He, K., Carmichael, G., 2016. Reac- tive nitrogen chemistry in aerosol water as a source of sul- fate during haze events in China. Sci. Adv. 2, e1601530 https://doi.org/10.1126/sciadv.1601530.

Cohard, J.M., Pinty, J.P., Bedos, C., 1998. Extending Twomey's analytical estimate of nucleated cloud droplet concentrations from CCN spectra[J]. J. Atmos. Sci. 55 (22), 3348–3357.

Deng, Y., Kagami, S., Ogawa, S., et al., 2018. Hygroscopicity of organic aerosols and their contributions to CCN concentrations over a midlatitude forest in Japan[J]. J. Geophys. Res. Atmos. 123 (17), 9703–9723.

Deng, Y., Yai, H., Fujinari, H., et al., 2019. Diurnal variation and size dependence of the hygroscopicity of organic aerosol at a forest site in Wakayama, Japan: their relationship to CCN concentrations[J]. Atmos. Chem. Phys. 19 (9), 5889–5903.

Deng, Z.Z., Zhao, C.S., Ma, N., et al., 2011. Size-resolved and bulk activation properties of aerosols in the North China Plain[J]. Atmos. Chem. Phys. 11 (8), 3835–3846.

Dusek, U., Covert, D.S., Wiedensohler, A., et al., 2003. Cloud condensation nuclei spectra derived from size distributions and hygroscopic properties of the aerosol in coastal south-west Portugal during ACE-2[J]. Tellus B 55 (1), 35–53.

Dusek, U., Frank, G.P., Hildebrandt, L., et al., 2006. Size matters more than chemistry for cloud-nucleating ability of aerosol particles[J]. Science 312 (5778), 1375–1378.

Ervens, B., Cubison, M., Andrews, E., et al., 2007. Prediction of cloud condensation nucleus number concentration using measurements of aerosol size distributions and composition and light scattering enhancement due to humidity[J]. J. Geophys. Res. Atmos. 112 (D10).

Ervens, B., Cubison, M.J., Andrews, E., et al., 2010. CCN predictions using simplified assumptions of organic aerosol composition and mixing state: a synthesis from six different locations[J]. Atmos. Chem. Phys. 10 (10), 4795–4807.

Farmer, D.K., Cappa, C.D., Kreidenweis, S.M., 2015. Atmospheric processes and their controlling influence on cloud condensation nuclei activity[J]. Chem. Rev. 115 (10), 4199–4217.

Fanourgakis, G.S., Kanakidou, M., Nenes, A., et al., 2019. Evaluation of global simulations of aerosol particle and cloud condensation nuclei number, with implications for cloud droplet formation[J]. Atmos. Chem. Phys. 19 (13), 8591–8617.

Fountoukis, C., Nenes, A., 2005. Continued development of a cloud droplet formation parameterization for global climate models[J]. J. Geophys. Res. Atmos. 110 (D11).

Fults, S.L., Massmann, A.K., Montecinos, A., et al., 2019. Wintertime aerosol measurements during the Chilean coastal orographic precipitation experiment[J]. Atmos. Chem. Phys. 19 (19), 12377–12396.

Ghan, S.J., Collins, D.R., 2004. Use of in situ data to test a Raman lidar–based cloud condensation nuclei remote sensing method[J]. J. Atmos. Ocean. Technol. 21 (2), 387–394.

Ghan, S.J., Rissman, T.A., Elleman, R., et al., 2006. Use of in situ cloud condensation nuclei, extinction, and aerosol size distribution measurements to test a method for retrieving cloud condensation nuclei profiles from surface measurements[J]. J. Geophys. Res. Atmos. 111 (D5).

Gunthe, S.S., King, S.M., Rose, D., et al., 2009. Cloud condensation nuclei in pristine tropical rainforest air of Amazonia: size-resolved measurements and modeling of atmospheric aerosol composition and CCN activity[J]. Atmos. Chem. Phys. 9 (19), 7551–7575.

Jayachandran, V.N., Suresh Babu, S.N., Vaishya, A., et al., 2020. Altitude profiles of cloud condensation nuclei characteristics across the Indo-Gangetic Plain prior to the onset of the Indian summer monsoon[J]. Atmos. Chem. Phys. 20 (1), 561–576.

Jefferson, A., 2010. Empirical estimates of CCN from aerosol optical properties at four remote sites[J]. Atmos. Chem. Phys. 10 (14), 6855–6861.

Kamilli, K.A., Poulain, L., Held, A., et al., 2014. Hygroscopic properties of the Paris urban aerosol in relation to its chemical composition[J]. Atmos. Chem. Phys. 14 (2), 737–749.

Khairoutdinov, M., Kogan, Y., 2000. A new cloud physics parameterization in a large-eddy simulation model of marine stratocumulus[J]. Mon. Weather Rev. 128 (1), 229–243.

Köhler, H., 1936. The nucleus in and the growth of hygroscopic droplets[J]. Trans. Faraday Soc. 32, 1152–1161.

Kuang, Y., Tao, J.C., Xu, W.Y., et al., 2019. Calculating ambient aerosol surface area concentrations using aerosol light scattering enhancement measurements. [J]. Atmospheric environment 216, 116919.

Kuang, Y., Zhao, C.S., Zhao, G., et al., 2018. A novel method for calculating ambient aerosol liquid water content based on measurements of a humidified nephelometer system[J]. Atmos. Meas. Tech. 11 (5), 2967–2982.

Kuang, Ye, Yao, He, Xu, Wanyun, Yuan, Bin, Zhang, Gen, Ma, Zhiqiang, Wu, Caihong, et al., 2020. Photochemical aqueous-phase reactions induce rapid daytime formation of oxygenated organic aerosol on the North China plain. Environ. Sci. Technol. 54 (7), 3849–3860. https://doi.org/10.1021/acs.est.9b06836.

Kuwata, M., Kondo, Y., Miyazaki, Y., et al., 2008. Cloud condensation nuclei activity at Jeju Island, Korea in spring 2005[J]. Atmos. Chem. Phys. 8 (11), 2933–2948.

Lance, S., Nenes, A., Medina, J., et al., 2006. Mapping the operation of the DMT continuous flow CCN counter. Aerosol. Sci. Technol. 40 (4), 242–254.

Lathem, T.L., Nenes, A., 2011. Water vapor depletion in the DMT continuous-flow CCN chamber: effects on supersaturation and droplet growth[J]. Aerosol. Sci. Technol. 45 (5), 604–615.

Lei, L., Zhou, W., Chen, C., et al., 2021. Long-term characterization of aerosol chemistry in cold season from 2013 to 2020 in Beijing, China[J]. Environ. Pollut. 268, 115952.

Leng, C., et al., 2014. Variations of cloud condensation nuclei (CCN) and aerosol activity during fog-haze episode: a case study from shanghai. Atmos. Chem. Phys. 14 (22), 12499–12512. https://doi.org/10.5194/acp-14-12499-2014.

Li, G., Su, H., Ma, N., Tao, J., Kuang, Y., Wang, Q., Hong, J., Zhang, Y., Kuhn, U., Zhang, S., 2021. Multiphase Chemistry Experiment in Fogs and Aerosols in the North China Plain (McFAN): Integrated Analysis and Intensive Winter Campaign 2018. Faraday Discussions.

Lim, K.S.S., Hong, S.Y., 2010. Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (CCN) for weather and climate models[J]. Mon. Weather Rev. 138 (5), 1587–1612.

Liu, H.J., Zhao, C.S., Nekat, B., Ma, N., Wiedensohler, A., van Pinxteren, D., Spindler, G., Müller, K., Herrmann, H., 2014. Aerosol hygroscopicity derived from size-segregated chemical composition and its parameterization in the North China Plain. Atmos. Chem. Phys. 14, 2525–2539. https://doi.org/10.5194/acp-14-2525-2014.

Liu, J., Li, Z., 2014. Estimation of cloud condensation nuclei concentration from aerosol optical quantities: influential factors and uncertainties[J]. Atmos. Chem. Phys. 14 (1), 471–483.

Liu, L., Cheng, Y., Wang, S., Wei, C., Pöhlker, M.L., Pöhlker, C., Artaxo, P., Shrivastava, M., Andreae, M.O., Pöschl, U., Su, H., 2020. Impact of biomass burning aerosols on radiation, clouds, and precipitation over the Amazon: relative importance of aerosol–cloud and aerosol–radiation interactions. Atmos. Chem. Phys. 20, 13283–13301. https://doi.org/10.5194/acp-20-13283-2020.

Ma, N., Zhao, C.S., Nowak, A., et al., 2011. Aerosol optical properties in the North China Plain during HaChi campaign: an in-situ optical closure study[J]. Atmos. Chem. Phys. 11 (12), 5959–5973.

Ma, N., Zhao, C., Tao, J., Wu, Z., Kecorius, S., Wang, Z., Groß, J., Liu, H., Bian, Y., Kuang, Y., Teich, M., Spindler, G., Muller, K., van Pinxteren, D., Herrmann, H., Hu, M., Wiedensohler, A., 2016. Variation of CCN activity during new particle formation events in the North China Plain. Atmos. Chem. Phys. 16, 8593–8607. https://doi.org/10.5194/acp-16-8593-2016.

Maronga, B., Banzhaf, S., Burmeister, C., et al., 2020. Overview of the PALM model system 6.0[J]. Geosci. Model Dev. (GMD) 13 (3), 1335–1372.

Martins, J.A., Gonçalves, F.L.T., Morales, C.A., et al., 2009. Cloud condensation nuclei from biomass burning during the Amazonian dry-to-wet transition season[J]. Meteorol. Atmos. Phys. 104 (1), 83–93.

Nair, A.A., Yu, F., 2020. Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements[J]. Atmos. Chem. Phys. 20 (21), 12853–12869.

Nair, A., Yu, F., Jost, P.C., et al., 2021. Machine Learning Uncovers Aerosol Size Information from Chemistry and Meteorology to Quantify Potential Cloud-Forming particles[J].

Paramonov, M., Kerminen, V.M., Gysel, M., et al., 2015. A synthesis of cloud condensation nuclei counter (CCNC) measurements within the EUCAARI network [J]. Atmos. Chem. Phys. 15 (21), 12211–12229.

Petters, M.D., Kreidenweis, S.M., 2007. A single parameter representation of hygroscopic growth and cloud condensation nucleus activity[J]. Atmos. Chem. Phys. 7 (8), 1961–1971.

Petters, M.D., Kreidenweis, S.M., 2008. A single parameter representation of hygroscopic growth and cloud condensation nucleus activity–Part 2: including solubility[J]. Atmos. Chem. Phys. 8 (20), 6273–6279.

Pinsky, M., Khain, A., Mazin, I., et al., 2012. Analytical estimation of droplet concentration at cloud base[J]. J. Geophys. Res. Atmos. 117 (D18).

Pöhlker, M.L., Ditas, F., Saturno, J., et al., 2018. Long-term observations of cloud condensation nuclei over the Amazon rain forest–Part 2: variability and characteristics of biomass burning, long-range transport, and pristine rain forest aerosols[J]. Atmos. Chem. Phys. 18 (14), 10289–10331.

Pöhlker, M.L., Pöhlker, C., Ditas, F., et al., 2016. Long-term observations of cloud condensation nuclei in the Amazon rain forest–Part 1: aerosol size distribution, hygroscopicity, and new model parametrizations for CCN prediction[J]. Atmos. Chem. Phys. 16 (24), 15709–15740.

Rejano, F., Titos, G., Casquero-Vera, J.A., et al., 2021. Activation properties of aerosol particles as cloud condensation nuclei at urban and high-altitude remote sites in southern Europe[J], 762 Sci. Total Environ., 143100

Reutter, P., Su, H., Trentmann, J., et al., 2009. Aerosol-and updraft-limited regimes of cloud droplet formation: influence of particle number, size and hygroscopicity on the activation of cloud condensation nuclei (CCN)[J]. Atmos. Chem. Phys. 9 (18), 7067–7080.

Rose, D., Gunthe, S.S., Su, H., et al., 2011. Cloud condensation nuclei in polluted air and biomass burning smoke near the mega-city Guangzhou, China–Part 2: size-resolved

aerosol chemical composition, diurnal cycles, and externally mixed weakly CCN-active soot particles[J]. Atmos. Chem. Phys. 11 (6), 2817–2836.
Rose, D., Nowak, A., Achtert, P., et al., 2010. Cloud condensation nuclei in polluted air and biomass burning smoke near the mega-city Guangzhou, China–Part 1: size-resolved measurements and implications for the modeling of aerosol particle hygroscopicity and CCN activity[J]. Atmos. Chem. Phys. 10 (7), 3365–3383.
Shinozuka, Y., Clarke, A.D., Nenes, A., et al., 2015. The relationship between cloud condensation nuclei (CCN) concentration and light extinction of dried particles: indications of underlying aerosol processes and implications for satellite-based CCN estimates[J]. Atmos. Chem. Phys. 15 (13), 7585–7604.
Su, H., Rose, D., Cheng, Y.F., et al., 2010. Hygroscopicity distribution concept for measurement data analysis and modeling of aerosol particle mixing state with regard to hygroscopic growth and CCN activation[J]. Atmos. Chem. Phys. 10 (15), 7489–7503.
Tao, J., Zhao, C., Kuang, Y., et al., 2018. A new method for calculating number concentrations of cloud condensation nuclei based on measurements of a three-wavelength humidified nephelometer system[J]. Atmos. Meas. Tech. 11 (2), 895–906.
Sun, Y., He, Y., Kuang, Y., et al., 2020. Chemical differences between PM1 and PM2.5 in highly polluted environment and implications in air pollution studies. Geophys. Res. Lett. 47 (5) e2019GL086288.
Tao, J., Kuang, Y., Ma, N., et al., 2021. Secondary aerosol formation alters CCN activity in the North China Plain[J]. Atmos. Chem. Phys. 21 (9), 7409–7427.
Tsarpalis, K., Katsafados, P., Papadopoulos, A., et al., 2020. Assessing desert dust indirect effects on cloud microphysics through a cloud nucleation scheme: a case study over the Western Mediterranean[J]. Rem. Sens. 12 (21), 3473.
Twomey, S., 1959. The nuclei of natural cloud formation part II: the supersaturation in natural clouds and the variation of cloud droplet concentration[J]. Geofisica pura e applicata 43 (1), 243–249.
Vié, B., Pinty, J.P., Berthet, S., et al., 2016. LIMA (v1. 0): a quasi two-moment microphysical scheme driven by a multimodal population of cloud condensation and ice freezing nuclei[J]. Geosci. Model Dev. (GMD) 9 (2), 567–586.
Wu, S., Tao, J., Ma, N., Kuang, Y., Zhang, Y., He, Y., Sun, Y., Xu, W., Hong, J., Xie, L., Wang, Q., Su, H., Cheng, Y., 2022. Particle number size distribution of PM1 and PM10 in fogs and implications on fog droplet evolutions. Atmos. Environ. 277 https://doi.org/10.1016/j.atmosenv.2022.119086, 119086.
Yang, Z., Ma, N., Wang, Q., Li, G., Pan, X., Dong, W., Zhu, S., Zhang, S., Gao, W., He, Y., Xie, L., Zhang, Y., Kuhn, U., Xu, W., Kuang, Y., Tao, J., Hong, J., Zhou, G., Sun, Y., Su, H., Cheng, Y., 2022. Characteristics and source apportionment of black carbon aerosol in the North China Plain. Atmos. Res. 276, 106246 https://doi.org/10.1016/j.atmosres.2022.106246.
Zhao, G., Tao, J., Kuang, Y., Shen, C., Yu, Y., Zhao, C., 2019. Role of black carbon mass size distribution in the direct aerosol radiative forcing. Atmos. Chem. Phys. 19, 13175–13188. https://doi.org/10.5194/acp-19-13175-2019.
Zhang, F., Ren, J., Fan, T., et al., 2019. Significantly enhanced aerosol CCN activity and number concentrations by nucleation-initiated haze events: a case study in urban Beijing [J]. J. Geophys. Res. Atmos. 124 (24), 14102–14113.
Zhang, Y., Tao, J., Ma, N., et al., 2020. Predicting cloud condensation nuclei number concentration based on conventional measurements of aerosol properties in the North China Plain[J]. Sci. Total Environ. 719, 137473.
Zhou, Y., Ma, N., Wang, Q., et al., 2022. Bimodal distribution of size-resolved particle effective density: results from a short campaign in a rural environment over the North China Plain[J]. Atmos. Chem. Phys. 22, 2029–2047. https://doi.org/10.5194/acp-22-2029-2022.