

JGR Atmospheres

RESEARCH ARTICLE

10.1029/2023JD039189

Key Points:

- A new cloud condensation nuclei (CCN) number concentration prediction method was developed
- The results of the CCN number concentration prediction method over the Korean Peninsula were very good
- The newly developed method showed strength in robustness and capability to take into account external mixing

Supporting Information:

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

Correspondence to:

M. Park and S. S. Yum,
ms_park@yonsei.ac.kr;
ssyum@yonsei.ac.kr

Citation:

Park, M., Yum, S. S., Seo, P., Kim, N., & Ahn, C. (2023). A new CCN number concentration prediction method based on multiple linear regression and non-negative matrix factorization: 1. Development, validation, and comparison using the measurement data over the Korean Peninsula. *Journal of Geophysical Research: Atmospheres*, 128, e2023JD039189. <https://doi.org/10.1029/2023JD039189>

Received 1 MAY 2023

Accepted 19 OCT 2023

Author Contributions:

Conceptualization: Minsu Park

Data curation: Minsu Park, Pyosuk Seo

Formal analysis: Minsu Park

Funding acquisition: Minsu Park, Seong Soo Yum

Investigation: Minsu Park, Pyosuk Seo,

Najin Kim, Chanwoo Ahn

Methodology: Minsu Park

Project Administration: Seong Soo Yum

Resources: Minsu Park

Software: Minsu Park

Supervision: Seong Soo Yum




Validation: Minsu Park

Visualization: Minsu Park

Writing – original draft: Minsu Park, Seong Soo Yum

Writing – review & editing: Minsu Park, Seong Soo Yum

A New CCN Number Concentration Prediction Method Based on Multiple Linear Regression and Non-Negative Matrix Factorization: 1. Development, Validation, and Comparison Using the Measurement Data Over the Korean Peninsula

Minsu Park¹ , Seong Soo Yum^{1,2} , Pyosuk Seo¹, Najin Kim³, and Chanwoo Ahn¹ 

¹Department of Atmospheric Sciences, Yonsei University, Seoul, Republic of Korea, ²Climate and Environmental Research Institute, Korea Institute of Science and Technology, Seoul, Republic of Korea, ³Center for Sustainable Environment Research, Korea Institute of Science and Technology, Seoul, Republic of Korea

Abstract To reduce uncertainty in climate change prediction, a large amount of cloud condensation nuclei number concentration (N_{CCN}) data must be obtained. This study aimed to develop a new N_{CCN} prediction method, hereafter MLR_{NMF} method, that applies multiple linear regression (MLR) and non-negative matrix factorization (NMF) to aerosol number size distribution data measured in Seoul and over the Yellow Sea. To verify the reliability of the MLR_{NMF} method, a data set separated from the training data set was used, and sufficient time differences of several years were given between the two data sets to make them as independent as possible from each other. The predicted N_{CCN} was in acceptable agreement with the measured N_{CCN} . The coefficient of determination (R^2) values between measured and predicted N_{CCN} for the Yellow Sea and Seoul were 0.81 and 0.71, respectively. Mean fractional bias (MFB) and mean fractional error (MFE) also met the performance goals ($< \pm 30\%$ and $< +50\%$, respectively). The MLR_{NMF} method had similar accuracy to the backward integration method but showed strength in robustness and the capability to take into account external mixing. The N_{CCN} prediction methods trained using data of a specific season tended to underestimate/overestimate somewhat, but MFB and MFE for all four seasons met the performance goals except for MFB in using June–August (JJA) data, implying that the MLR_{NMF} method trained using only the data of a specific season can predict N_{CCN} for all four seasons to some extent. It is expected that abundant N_{CCN} data can be obtained through the MLR_{NMF} method in future studies.

Plain Language Summary Since cloud condensation nuclei (CCN), the potential cloud-forming particles, exert a significant influence on climate change, reducing the uncertainty of the CCN number concentration (N_{CCN}) can contribute greatly to reducing the uncertainty in climate change prediction. Therefore, efforts have been made worldwide to increase the possibility of predicting the N_{CCN} from aerosol properties as well as verifying such data to reduce the uncertainty of the global N_{CCN} distribution. This study aimed to develop a new N_{CCN} prediction method, hereafter MLR_{NMF} method, that applies multiple linear regression (MLR) and non-negative matrix factorization to aerosol number size distribution data measured in Seoul and over the Yellow Sea. The predicted N_{CCN} of the test data set was in very good agreement with the measured N_{CCN} . The MLR_{NMF} method had similar accuracy to the existing method but showed strength in robustness and the capability to take into account mixing state of aerosols. The results of this study suggest that the N_{CCN} prediction method developed for the Yellow Sea and Seoul can be applied to other regions in and around the Korean Peninsula in future studies, can be expanded on a global scale eventually, and will greatly contribute to reducing the uncertainty of climate change prediction.

1. Introduction

When aerosols are of relevant sizes to be activated as cloud droplets by water vapor supersaturation, they can contribute to cloud formation and affect the dynamic, radiative, and microphysical properties of clouds (Andreae & Rosenfeld, 2008; Lohmann & Feichter, 2005; Seinfeld and Pandis, 2016). These potential cloud-forming particles are called cloud condensation nuclei (CCN). Since cloud properties affect not only water resources, such as precipitation, but also the Earth's radiation budget, CCN exert a heavy influence on climate change (IPCC, 2021). These effective radiative forcing from aerosol-cloud interactions have been considered to be one of the largest factors causing uncertainty in climate change prediction (e.g., Wall et al., 2022).

Sotiropoulou et al. (2006) reported that the error in cloud droplet number concentration prediction was about half of the error in the CCN number concentration (N_{CCN}) prediction. Other previous studies also reported that cloud droplet number concentrations are sensitive to the uncertainties in the N_{CCN} , depending on meteorological and environmental conditions (e.g., Fanourgakis et al., 2019). Therefore, reducing the uncertainty of N_{CCN} can significantly contribute to reducing the uncertainty in climate change prediction. Since N_{CCN} varies greatly depending on the region, global CCN measurement data are required (Schmale, Henning, et al., 2017; Schmale et al., 2018). Measurements of aerosol properties, such as aerosol number concentration and size distribution, are carried out in many places around the world as part of the Global Atmosphere Watch Program of the World Meteorological Organization. However, at present, N_{CCN} measurements are less extensive. Therefore, efforts have been made worldwide to increase the possibility of predicting N_{CCN} from aerosol properties and verifying it (e.g., Bhattu & Tripathi, 2015; Pöhlker et al., 2016; Sotiropoulou et al., 2006; J. Wang et al., 2010; F. Zhang et al., 2014). Moreover, Nair et al. (2021) suggested the possibility of estimating the N_{CCN} based on atmospheric chemistry and meteorological factors using a machine learning/artificial intelligence model. Kammermann et al. (2010) reported that N_{CCN} prediction with the assumption of constant chemistry (no temporal variability of aerosol chemistry) could show sufficiently reliable results. Dusek et al. (2006) compared the effects of the size and chemical composition of aerosols on N_{CCN} prediction, and found that the size effect was much larger than the chemical effect. In other words, the temporal variation of the aerosol number size distribution alone can explain the temporal variation of N_{CCN} reasonably well. Therefore, if N_{CCN} can be significantly obtained from aerosol number size distribution, it can greatly contribute to obtaining abundant N_{CCN} data.

If aerosol hygroscopicity information exists, it would certainly help predict N_{CCN} (e.g., Abdul-Razzak & Ghan, 2000). However, the measurement data on aerosol hygroscopicity are still very limited, compared to those of the aerosol number size distribution, although the hygroscopicity measurements have been carried out intensively recently. In many previous studies, the critical dry diameter for CCN activation was obtained from the backward integration of measured aerosol number size distribution and N_{CCN} at a certain supersaturation (S) (e.g., Cerully et al., 2015; Hung et al., 2014; N. Kim et al., 2018; Moore et al., 2011). Then, the number concentration of aerosols larger than the critical diameter is regarded as N_{CCN} . This N_{CCN} prediction method, hereafter called the backward integration method, is very useful but has some limitations (low robustness and the inability to take into account external mixing, see Sections 3.1 and 4.3 for details). Therefore, in this study, we propose a new N_{CCN} prediction method and present the validation and comparison results. In most previous studies, N_{CCN} prediction was conducted by estimating the critical diameter or aerosol hygroscopicity from certain measurement data and then applying the critical diameter or aerosol hygroscopicity to the aerosol size distribution data measured during the same period at the same observation site. This is called a “CCN closure” study. However, to create an unbiased N_{CCN} prediction method, it is necessary to apply the N_{CCN} prediction method to a separate data set to test the method. Such studies have only recently begun to be conducted (e.g., Liang et al., 2022; Nair et al., 2021). In this study, we intend to test the N_{CCN} prediction methods by utilizing a separate data set that had sufficient time difference of several years from the training data set.

East Asia is one of the most abundant source regions of air pollutants, and indeed, aerosol concentrations in East Asia are very high (e.g., Kaufman et al., 2002; J. H. Kim et al., 2014; Park et al., 2018, 2020, 2021; Ramanathan et al., 2001; Q. Zhang et al., 2009). Furthermore, previous studies have reported that aerosols in East Asia are complex mixtures of marine, volcanic, pollution, and dust chemical components and their spatial inhomogeneity is very large (e.g., Huebert et al., 2003; Li et al., 2019; Quinn et al., 2004). However, it has been known that in recent years the aerosol concentration has been sharply decreasing over East Asia as a result of continued reduction in anthropogenic emissions (e.g., Ramachandran et al., 2020, 2022). In other words, the composition of aerosols in East Asia is very complex and changing rapidly, which makes it very difficult to predict N_{CCN} in this region. Therefore, if N_{CCN} in East Asia can be predicted with high accuracy, it can be said that N_{CCN} in other regions of the world is also likely to be predicted with confidence. In this study, the development, validation, and comparison of N_{CCN} prediction methods were conducted on the data measured over the Yellow Sea, located to the west of the Korean Peninsula, and on the data measured in Seoul, one of the most polluted areas in the Korean Peninsula. We can safely say that the Yellow Sea represents a marine environment that is heavily influenced by continental outflow and Seoul represents a polluted megacity in East Asia.

The CCN distribution data are highly needed for a more accurate assessment of the aerosol indirect effect on climate change but are difficult to obtain due to lack of observation in many parts of the world. This study aimed

to develop a new N_{CCN} prediction method and verify the method using the data over the Yellow Sea and in Seoul, to lay the foundation for obtaining abundant N_{CCN} data.

2. Measurement Data

2.1. Yellow Sea

The Yellow Sea is located between the Chinese mainland and the Korean Peninsula. Although the Yellow Sea is a maritime region, the air mass over the Yellow Sea is affected by continental air masses that are transported from the Chinese mainland and the Korean Peninsula (Park et al., 2018, 2021). Because of the complexity of the air mass over the Yellow Sea, air quality measurements over it have been conducted every spring in recent years (Park et al., 2016, 2018, 2021).

Aerosol number size distribution and N_{CCN} measurements were conducted onboard *Gisang 1*, a research vessel of the National Institute of Meteorological Sciences (NIMS) of Korea. Submicron aerosol number size distributions and N_{CCN} were measured continuously with a scanning mobility particle size (SMPS; TSI3936L10) and DMT CCN counter (CCNC; DMT CCN-100), respectively. The CCNC measured N_{CCN} at several supersaturations. In the CCNC, N_{CCN} at a single S is measured every 1 s for 5 min, and a full CCN spectrum was obtained every 30 min. The S range was slightly different depending on the measurement period, but was approximately 0.2%–1.0%. We focused on predicting N_{CCN} at 0.6% supersaturation ($N_{CCN0.6}$) in this study. An SMPS was used to measure the aerosol number size distributions for 10–478 nm mobility diameter range every 3 min, where 2 min is scanning time, and the remaining 1 min is stabilization time. The data collected at the same time span for the two instruments, CCNC and SMPS, were used in this study. Park et al. (2016) reported that the minimum aerosol number concentrations over the Yellow Sea were recorded at a longitude of approximately 124°E, which is the midway point between the mainland China and the Korean Peninsula. Therefore, *Gisang 1* repeatedly cruised from north to south and from south to north at this longitude, on alternating days, to ensure the statistical reliability of the aerosol property data obtained over the Yellow Sea (Park et al., 2018, 2021). Only the measurement data collected when the vessel was in the target area were used for the development and validation of the N_{CCN} prediction methods. To minimize aerosol loss, ambient air was sampled through 0.25-inch conductive tubes from the inlet on the roof of the room. The smokestack of the vessel was located at the rear of the ship, approximately 15 m away from the inlet. Since the vessel continuously emitted pollutants, the data collected when the vessel was stationary were excluded. Moreover, since the speed of the vessel during the cruises was approximately 6–7 m s^{−1}, the data collected when the wind speed in the direction of the vessel's movement was greater than 6 m s^{−1} were also excluded. Because the vessel was mostly stationary at night, most of the nighttime measurement data were discarded.

In this study, at least two data sets are required for the N_{CCN} prediction method. One is a data set for training, and the other is a data set for test. Among the measurement data, the data collected in 2017 were used for training, and the data collected in 2019 and 2021 were used for test. In 2017, N_{CCN} was measured at 0.65% S , therefore, N_{CCN} at 0.65% S was also considered to be approximately the same as $N_{CCN0.6}$. Papers on the aerosol characterization results over the Yellow Sea in 2017 and 2019 have already been published (Park et al., 2018, 2021). However, since SMPS could not measure the concentration of aerosols larger than 250 nm due to instrument problems in June 2019, the data in this period were excluded. The sample air was not dried in 2017 and 2019. On the other hand, the sample air for SMPS in 2021 was dried using a diffusion dryer. However, Park et al. (2018) reported that the difference in particle sizes between ambient sample air and dried sample air over the Yellow Sea was not large (approximately 10%). Therefore, the difference in particle sizes between ambient sample air and dried sample air was ignored for the data in 2021. To correct particle loss in the diffusion dryer, the total aerosol number concentration measured by a condensation particle counter (CPC; TSI 3772) was used. The sample air measured by the CPC and CCNC in 2021 was not dried as in 2017 and 2019. The aerosol number concentration in each size bin was corrected so that the integrated value of all SMPS bins was matched to the CPC concentration. Since there was particle loss due to the diffusion dryer, the difference between the total aerosol number concentration and the integral value of the aerosol number size distribution could become large. Therefore, a case where the difference between the total aerosol number concentration and the integral value of the aerosol number size distribution was large was also included as a valid data set in 2021. Since the particle loss depends on the particle size, in the case of a rapid increase in very small particles, such as new particle formation (NPF), correction of SMPS data by the CPC can change the shape of the aerosol number size distribution. Fortunately, however, no NPF events were

Table 1
Overall Description of Measurement Sites and Data Sets

	Yellow sea		Seoul	
	Training data set	Test data set	Training data set	Test data set
Location	Over the Yellow Sea (32–37.5°N, ~124°E)		Yonsei University campus (37.6°N, 127.0°E)	
Period	April–May 2017	April–May 2019, March–April 2021	2006–2010	2018–2020
Instruments	SMPS, CCN-100	SMPS, CCN-100, CPC (2021 only)	SMPS, CCN-100	SMPS, CCN-100 CPC
Supersaturation	0.6%			
Size range of SMPS	10–478 nm		10–414 nm	
Number of datapoints	489	447	13,565	8,712
Note			Before construction,	After construction,
Used data	Average value during one scan time of SMPS	Average value during one scan time of SMPS	Hourly average value	Average value during one scan time of SMPS
References	Park et al. (2018)	Park et al. (2021) for data in 2019	Schmale, Henning, et al. (2017)	

observed when the vessel cruised in the target area in 2021. Therefore, in 2021, aerosol number size distribution data corrected using CPC data were used.

The aerosol number size distribution in urban and continental areas is known to be dominated by nucleation and Aitken modes (Seinfeld and Pandis, 2016). Park et al. (2020) had a chance to measure the aerosol number size distribution in and around the Korean Peninsula for the size range of 10–1,000 nm, by combining SMPS and laser aerosol spectrometer (LAS) data. They found that the number concentrations of aerosols larger than 400 nm were very low in and around the Korean Peninsula. For such reason, we did not consider the particles larger than the upper limit of SMPS in the N_{CCN} prediction and correction of aerosol number size distribution. Descriptions of the data set used in this study are summarized in Table 1. SMPS, CPC, and LAS are fully specified at www.tsi.com, and CCNC is fully specified at www.dropletmeasurement.com.

2.2. Seoul

Seoul (37.6°N, 127.0°E) is the capital city of South Korea and a megacity with approximately 10 million inhabitants. The Seoul measurement data were presented in the previous studies including the data descriptor paper (N. Kim et al., 2014; Schmale, Henning, et al., 2017). Measurements were made at the Yonsei University campus, which is located in the northwestern part of Seoul, during 2006–2010. The measurement instruments were located on the sixth floor (approximately 24 m above ground) of a building approximately 300 m away from the main traffic roads. The measurement instruments for aerosol number size distribution and N_{CCN} were the same as those in Section 2.1. However, the range of the aerosol number size distribution measured in Seoul was 10–414 nm, unlike the measured over the Yellow Sea. The sample air was neither dried nor diluted, and inlet lines were 0.25-inch conductive tubes approximately 1 m long to minimize loss of submicron particle. Hourly averages of aerosol number size distribution and $N_{CCN,0.6}$ were used in the training data set.

The measurement data in Seoul during 2006–2010 were used as a data set for training. As a data set for test, additional measurement data from the same building in 2018–2020 were used. The temporal difference between the two data sets is approximately 10 years, which is considered sufficient to verify the developed N_{CCN} prediction methods. Although the training data set in Seoul spans four seasons, there are no spring season data in the test data set since the instruments were deployed for ship measurements over the Yellow Sea in spring time. Submicron aerosol number size distributions and N_{CCN} were measured with the SMPS and CCNC as in the training data set. Additionally, the total aerosol number concentration was measured with a CPC. Then the aerosol number concentration in each size bin of SMPS was corrected so that the integral value of the aerosol number size distribution was matched to the total aerosol number concentration. However, if either the total aerosol number concentrations or integral values of the aerosol number size distribution was more than double the other, it was

considered to be an instrument problem and these data were excluded from the test data set. Unlike the training data set, which used hourly average data, the aerosol number size distribution for one scan time (~2 min) of SMPS and $N_{CCN0.6}$ value averaged for the same time span were used as test data sets.

The measurement instruments were located in the same building as presented in Schmale, Henning, et al. (2017), but the room was not the same for the two data sets. The measurement room for the test data set was located on the fifth floor (approximately 20 m above ground) of the building. During 2006–2010, the inlet line was located on the north side of the building, whereas during 2018–2020, the inlet line was located on the southeast side of the building. In addition, large-scale construction was carried out on the Yonsei University campus between the two periods. Before construction, vehicles frequently traveled on campus. However, large construction was carried out to divert vehicle traffic underground, and after the construction, the vehicle traffic on the ground was minimized. Therefore, the aerosol characteristics of the two periods are expected to be somewhat different. If the test results are good for the data set that has somewhat different characteristics from the data set for training, our N_{CCN} prediction method in Seoul would indeed be considered to be reliable.

3. N_{CCN} Prediction Methods

3.1. Backward Integration Method

The backward integration method has been commonly used for N_{CCN} prediction (e.g., Bhattu & Tripathi, 2015; Pöhlker et al., 2016; Sotiropoulou et al., 2006; J. Wang et al., 2010; F. Zhang et al., 2014). The critical dry diameter for CCN activation is obtained from the measured aerosol number size distribution and N_{CCN} at a certain S , as shown in Equation 1.

$$N_{CCN} = \int_{D_{crit}}^{D_{max}} \frac{dN(D_p)}{d \log D_p} d \log D_p \quad (1)$$

The critical diameter can be calculated from the aerosol hygroscopicity or chemical composition data, but this study aimed to use only the aerosol number size distribution and N_{CCN} data to minimize the number of measurement elements. The concept of this method is that N_{CCN} can be predicted by applying the obtained critical diameter from the above equation to other aerosol number size distribution data. Since the manner in which aerosols act as CCN depends on the particle size and hygroscopicity (or chemical composition), as well as ambient S , the backward integration method may be considered scientifically reasonable under the assumption that the aerosol hygroscopicity does not change significantly. If the particles of a specific chemical composition are larger than the critical diameter for CCN activation, the particles can be regarded as CCN. However, the backward integration method has some limitations. First, this method is not robust. That is, it is greatly affected by instrument uncertainty. All measurement instruments have some degree of uncertainty. Typically, instrument calibration is conducted with the goal of reducing uncertainty to within 10%. However, this 10% uncertainty can be critical for the backward integration method, depending on the shape of the aerosol number size distribution. For example, if N_{CCN} is measured to be 10% higher than the true value, while the true value of the aerosol number size distribution is measured, the critical diameter would sometimes deviate significantly from the true value, and in extreme cases it would even be impossible to estimate the critical diameter from the aerosol number size distribution. For example, in some cases, a 10% increase in $N_{CCN0.6}$ resulted in a 50% decrease in critical diameter. Furthermore, for the Yellow Sea data, estimation of the critical diameter was not possible in 7%, 28%, 67%, and 88% of the data when $N_{CCN0.6}$ were increased deliberately by 20%, 30%, 50%, and 100%. Furthermore, if different types of measurement instruments are used or the measurement environment changes, the measured values of the two instruments may be more different. Multiple charging effects in the SMPS also cause uncertainty in the measurement of aerosol number size distribution. In particular, the concentration of large aerosols is greatly affected by multiple charges; therefore, the backward integration method, which integrates from the largest size, may also be affected by multiple charged particles even if the number concentration of large particles is low. This limitation will be described in detail in Section 4.3. A second limitation is that it assumes that all particles are internally mixed. Even if particles have the same size, their chemical composition can be different, so some of the aerosols may act as CCN, while others may not. However, since the backward integration method assumes that all particles are internally mixed, this difference cannot be taken into account. The advantage of the backward integration method is that it can obtain the critical diameter at every measurement cycle time, but in terms of N_{CCN} prediction,

only the average, median, or geometric mean value of the critical diameter is needed, so the advantage is greatly diminished.

3.2. Multiple Linear Regression Using Non-Negative Matrix Factorization (MLR_{NMF}) Method

Machine learning is the study of computer algorithms that can improve themselves through experience and by the use of abundant data (Mitchell & Mitchell, 1997). Machine learning can also be applied to N_{CCN} prediction. In this study, we used the Python scikit-learn library, which provides a standard interface for implementing machine learning algorithms (Pedregosa et al., 2011). The number concentration of each bin of the aerosol number size distributions positively correlates with the N_{CCN} except for very small particles. Multiple linear regression (MLR), which is commonly used in machine learning, can be applied in this case. To predict the N_{CCN} , the number concentrations of individual bins of the aerosol number size distributions can be used as predictors. If the coefficients of the bins whose particle size is larger than the critical diameter are 1 and the remainders are 0, it is the same as the backward integration method. However, it is difficult to draw scientific meaning and interpretation from the linear regression method. This is due to large collinearity between aerosol size bins. That is, even if the coefficients of the two adjacent bins are significantly different, it cannot be said with certainty which of the two bins is more related to the N_{CCN} .

As described above, the backward integration method and simple MLR method have some limitations. In this study, a new method was used for CCN prediction to overcome the limitations of both methods. The method used in this study is non-negative matrix factorization (NMF), which is one of the dimensionality reduction methods. NMF can be applied when all elements in the matrix are non-negative. Since the values of each bin of aerosol number size distribution are non-negative, NMF can be applied to aerosol number size distribution data. NMF decomposes a non-negative matrix V into two non-negative matrix Factors W and H to extract meaningful information in certain data sets (Gillis, 2020; Lee & Seung, 1999, 2000). This can be expressed simply as $V \approx WH$, or

$$V_{ij} \approx (WH)_{ij} = \sum_{k=1}^r W_{ik} H_{kj}, \quad (2)$$

where V denotes the aerosol number size distribution data and r is the number of reduced dimensions. The notations i, j , and k in this equation are simple indices to indicate elements of the matrix. It finds a decomposition of sample V into two matrices W and H of non-negative elements by optimizing the distance between V and WH . There are several ways to express the distance between two matrices. In this study, the squared Frobenius norm, which is an obvious extension of the Euclidean norm to matrices, was used as the distance function. For an $n \times m$ matrix V , the dimensions of the matrices W and H are $n \times r$ and $r \times m$, respectively. A non-negative matrix W is a basis matrix and a non-negative matrix H is weighting matrix. Therefore, the matrix W contains information about distinct features of aerosol number size distribution, and the matrix H shows the time dependent variation of each basis. Since the aerosol number size distribution in the atmosphere usually consists of several distinct modes, this dimensionality reduction method can be applied.

After this dimensionality reduction process, we can look at only a few representative bases and their time-dependent weightings, not all the bins of the aerosol number size distribution. The key point of dimensionality reduction method, such as NMF, is to make the dimensionally reduced data to retain significantly meaningful characteristics of the original data. Since time-dependent aerosol number size distribution data were used in this study, the crucial property of the aerosol number size distribution data is the temporal variation. Therefore, the data within the same basis of NMF can be said to have similar temporal variability. In other words, the data within the same basis tended to increase or decrease simultaneously. The concentrations of each aerosol size bin would be influenced by sources, transformation, or sink. Therefore, from the fact that aerosols within the same basis had similar increasing/decreasing trends, it can be inferred that these aerosols were emitted from similar sources and underwent similar transformation processes. Therefore, each basis of NMF typically represent a particular source and potentially capture the aerosol composition characteristics. Furthermore, since whether aerosols act as CCN or not depends on the particle size, hygroscopicity (or chemical composition), and ambient S , the weighting of each basis of the aerosol number size distributions except for the bases indicating small aerosols has a positive linear relationship with N_{CCN} , and the weighting values of each basis are all meaningful. If the weighting value of each basis is used as the predictor, the limitation of the MLR method can be reduced. Furthermore, since the number of predictors was reduced by grouping them using NMF, the overfitting problem can also be expected to

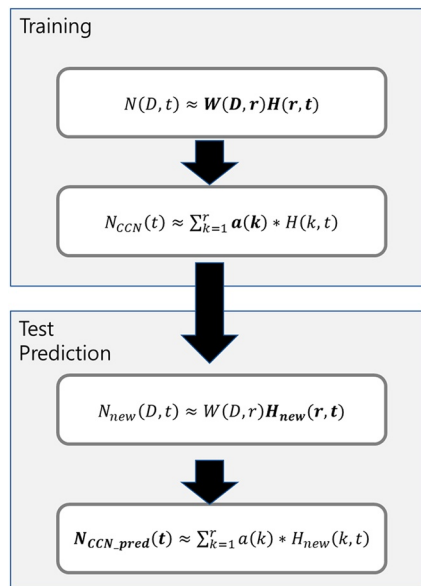


Figure 1. A flowchart of the non-negative matrix factorization (MLR_{NMF}) method to predict N_{CCN} . Variables written in bold type indicate that they are obtained in that step. $N(D, t)$, $N_{CCN}(t)$, and $N_{new}(D, t)$ were obtained from measurements.

be reduced. Therefore, in this study, we intended to predict N_{CCN} by applying the MLR method to the results of dimension reduction using the NMF method. Hereafter, this method will be called the MLR with non-negative matrix factorization (MLR_{NMF}) method.

The flow chart of this method is shown in Figure 1. In the training stage, first, aerosol number size distribution data are decomposed into two non-negative matrices W and H . Each matrix W can indicate an aerosol size distribution shape having multiple modes, but in this case, it difficult to say that each W represents aerosols of a particular size. Therefore, the NMF analysis was performed again by removing sub-modes from the W matrix. That is, each W represents an aerosol size distribution shape having a single mode. Then, coefficients “ a ” are calculated through a MLR method using rows of H as predictors and N_{CCN} as a predictand. Since the predictors either positively correlate with N_{CCN} or have no correlation with N_{CCN} , each coefficient of MLR was constrained to a non-negative value, and non-negative least squares were applied. Since we did not allow negative coefficients, the y-intercept must be zero to make N_{CCN} to be zero in the absence of aerosols (all predictors are zero). Through this process, we can determine the relationship between each basis of aerosol number size distribution and N_{CCN} . In the test or prediction stage, N_{CCN} of a separate new data set can be predicted from the aerosol number size distribution based on this relationship. New aerosol number size distribution data are decomposed into the same matrix W as in the training stage and a new matrix H . In this process, it is assumed that the aerosol number size distribution data sets in the training and prediction stages

have the same distinct modes. Then, N_{CCN} is predicted by multiplying the new matrix H by the coefficients “ a ” obtained in the training stage. Since aerosol number size distribution data are decomposed into two non-negative matrices W and H , which can have multiple solutions (e.g., $W \cdot H = 2 \cdot W - 0.5H = 0.5 \cdot W \cdot 2H$). We put a restriction to normalize W and H . Here, we set the area of each basis of the matrix W to be one. Then, the value of H indicates the number concentration of the aerosols (N_{CN}) of the corresponding basis at that time. The coefficient “ a ” has a range of 0 to approximately 1 and represents the N_{CCN}/N_{CN} ratio for a given mode.

A critical issue in the dimensionality reduction method, including NMF, is to select the reduced number of components (r). Here, the number of components is the number of dimensions. As the number of components increases, the difference between the original matrix (V) and the NMF-estimated matrix (WH) decreases. However, since the advantage of dimensionality reduction fades if too many dimensions are used for NMF analysis, it is important to set an appropriate number of dimensions. The advantage of the dimensionality reduction is that large data can be expressed with fewer variables and it improves the understanding of the correlation between each variable. Many studies have been conducted on a method for selecting an appropriate number of dimensions, and have varying opinions (e.g., Brunet et al., 2004; Hutchins et al., 2008; Lee, 2020). In this study, we selected the appropriate number of dimensions using the method proposed by Hutchins et al. (2008). First, we created a purely random matrix of the same size as the training data set. To have the same average value as the measured data set, the range of values in the random matrix was set from zero to twice the average value. The difference between the original matrix and the NMF-estimated matrix for the purely random matrix will show a roughly linear decrease with an increased number of dimensions. On the other hand, for a matrix with a specific pattern, the difference will show an inflection point approximately at the number of patterns. Therefore, an appropriate number of dimensions could be selected based on these gradients. Here, we used the mean squared error (MSE) to select the appropriate number of dimensions similar to that shown in Hutchins et al. (2008). However, unlike Hutchins et al. (2008), we used MSE instead of the residual sum of squares (RSS). Since the random matrix and the training data set have the same size, it is considered that comparing MSE and comparing RSS have the same meaning. Both the results for the Yellow Sea and Seoul showed a distinct decreasing pattern of MSE. It can be seen that the slope of the MSE of the training data set gradually decreases, but when the number of dimensions exceeds 6, the MSE of the training data set does not decrease significantly compared to that of the random matrix (Figure 2). This pattern appeared both over the Yellow Sea and in Seoul. Therefore, we selected 6 as the appropriate number of dimensions for both regions. That is, aerosol number size distributions for both regions were expressed as 6 representative basis vectors (columns of W) and their corresponding weightings (rows of H).

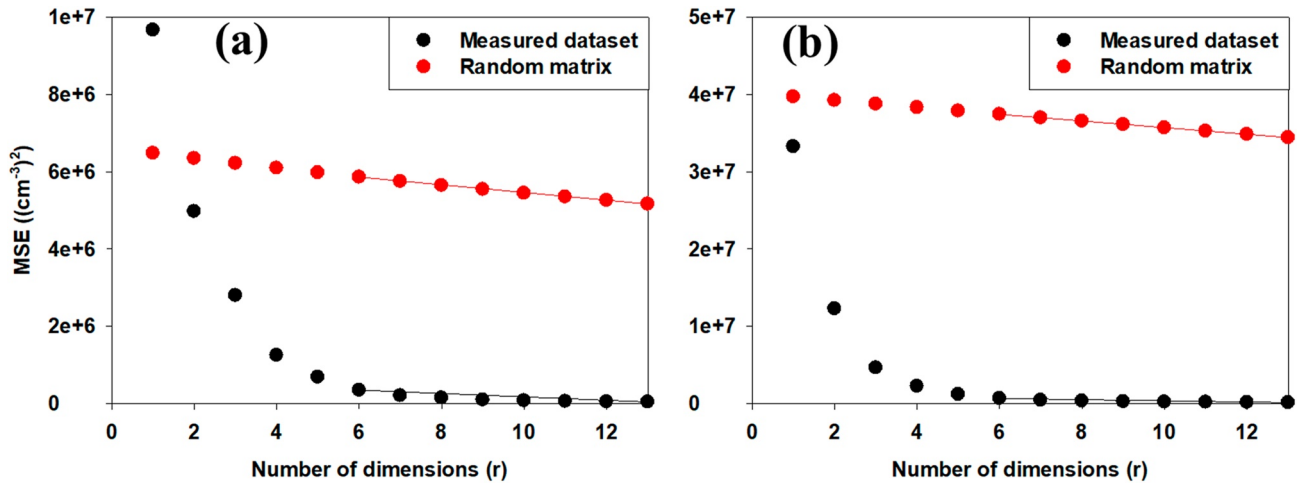


Figure 2. The mean squared error of the reconstructed data calculated from the training data set (black) and that calculated from a purely random matrix (red) (a) for the Yellow Sea and (b) Seoul. The solid lines are for showing the slope.

Each method described in Section 3 is briefly schematized in Figure 3. In the backward integration method, if the size of the particles is larger than the critical diameter for CCN activation, all particles are regarded as CCN. In the MLR_{NMF} method, aerosol number size distribution data are decomposed into two non-negative matrices W (basis matrix) and H (weighting matrix). Then, coefficients “a” are calculated through the MLR method using H as the predictors and N_{CCN} as a predictand. The basis representing a large particle size is likely to act as CCN, so the coefficient of MLR is also high accordingly. The matrix H indicates the weighting of the corresponding basis at that time. That is, the product of coefficients “a” and H indicates the N_{CCN} of the corresponding basis at that time. Therefore, the sum of the products of each “a” value and each component of H , or the product of coefficients “a” and matrix H , represents N_{CCN} .

3.3. Performance Metrics

There are several performance metrics that can be used to examine performance of the N_{CCN} prediction methods. In this study, coefficient of determination (R^2) and Root Mean Square Error (RMSE), mean fractional bias (MFB) and mean fractional error (MFE) were used to examine performance of the N_{CCN} prediction methods. Mean fractional bias (MFB) and MFE can be written as

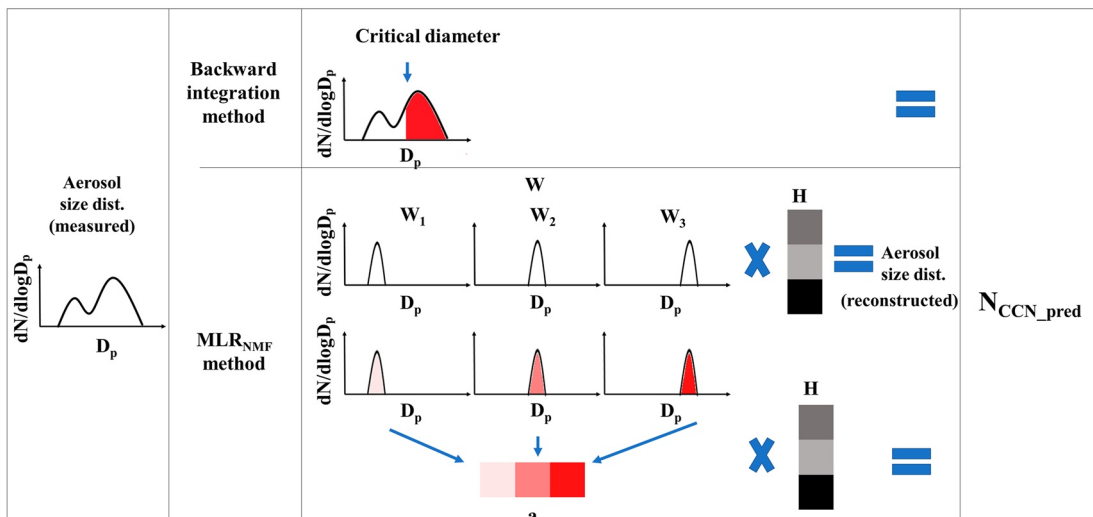


Figure 3. Schematic of the different N_{CCN} prediction approaches. The shade indicates the magnitude of the coefficient “a” and H of the corresponding basis, respectively.

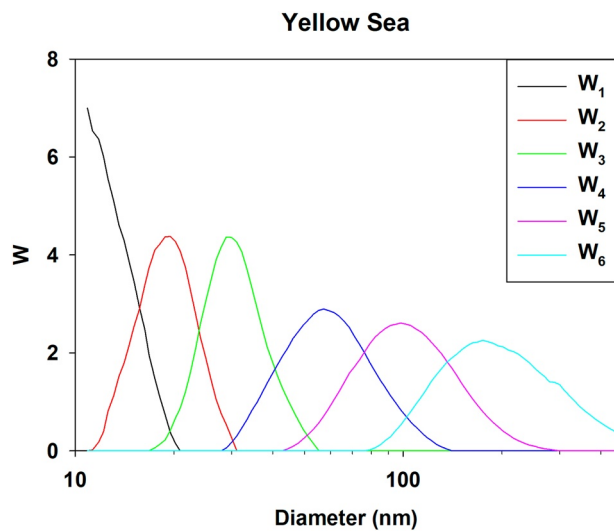


Figure 4. The matrix W decomposed by non-negative matrix factorization (MLR_{NMF}) for the Yellow Sea data. The number in the subscript indicates the column number of W .

For the MLR_{NMF} method, the aerosol number size distribution over the Yellow Sea was decomposed into six distinct bases, as shown in Figure 4. The matrices W and H were aligned in the order of aerosol size represented by the base vector. The numbers in the subscript of W and H indicate the column number of W and the row number of H , respectively. That is, W_1 and H_1 respectively indicate the basis size distribution of smallest aerosols and its time-dependent weightings, and W_6 and H_6 respectively indicate the basis size distribution of largest aerosols and its time-dependent weightings. Except for the two adjacent bases, there was low collinearity between weightings of different bases (Figure S1 in Supporting Information S1). Since the concentrations of individual aerosol size bins correlate with each other and the degree of freedom of aerosol size distribution is not exactly the same as the number of bases, it is unrealistic for all the bases to be completely uncorrelated. Nevertheless, there was very low collinearity between weightings of different bases except for the two adjacent bases. Therefore, we think that the MLR method is a right method to use in this study. Low values of coefficient “ a ” for the weightings of the first three bases (i.e., H_1 – H_3) indicated that the aerosols in these bases cannot easily act as CCN whereas the high values of “ a ” for the H_4 – H_6 indicated that the aerosols in these bases are likely to act as CCN at 0.6% S (Table 2). The fact that the last three bases showed a positive correlation with $N_{CCN,0.6}$ indicates the same result (Figure S1 in Supporting Information S1). However, even the weighting of the last basis (i.e., H_6) does not have a coefficient of 1, which will be covered in detail in Section 4.3. For the fourth basis, the coefficient and the mode diameter were 0.50 and 57.3 nm, respectively. So, it seems that half of these aerosols can act as CCN at 0.6% S and it depends on whether the diameter of aerosols is larger or smaller than 57.3 nm over the Yellow Sea. Of course, there are additional factors to consider, such as if the aerosols are externally mixed, some aerosols may not act as CCN even if they are larger than the threshold size. In general, aging processes move the aerosol population toward a more internally mixed state (Riemer et al., 2019). However, despite the fact that the Yellow Sea is a sea,

there were various types of aerosols over the Yellow Sea, and a significant portion of them was fresh aerosols that have not yet been internally mixed (Kwak et al., 2022). It is thought that this is because the distance from the pollutant source is not far, so the atmosphere over the Yellow Sea is greatly affected by the aerosols directly emitted from the pollutant sources. This diameter was smaller than the critical diameter obtained from the backward integration method (70.6 nm). Figures 5a and 5b show comparison results between the measured $N_{CCN,0.6}$ and the predicted $N_{CCN,0.6}$ using the two N_{CCN}

Table 2

The Mode Diameters of Matrix W Decomposed by Non-Negative Matrix Factorization and the Corresponding Coefficients in the Multiple Linear Regression for the Yellow Sea Data

Mode diameter (nm)	10.9	19.5	28.9	57.3	98.2	174.7
Coefficient	0.00	0.06	0.02	0.50	0.85	0.78

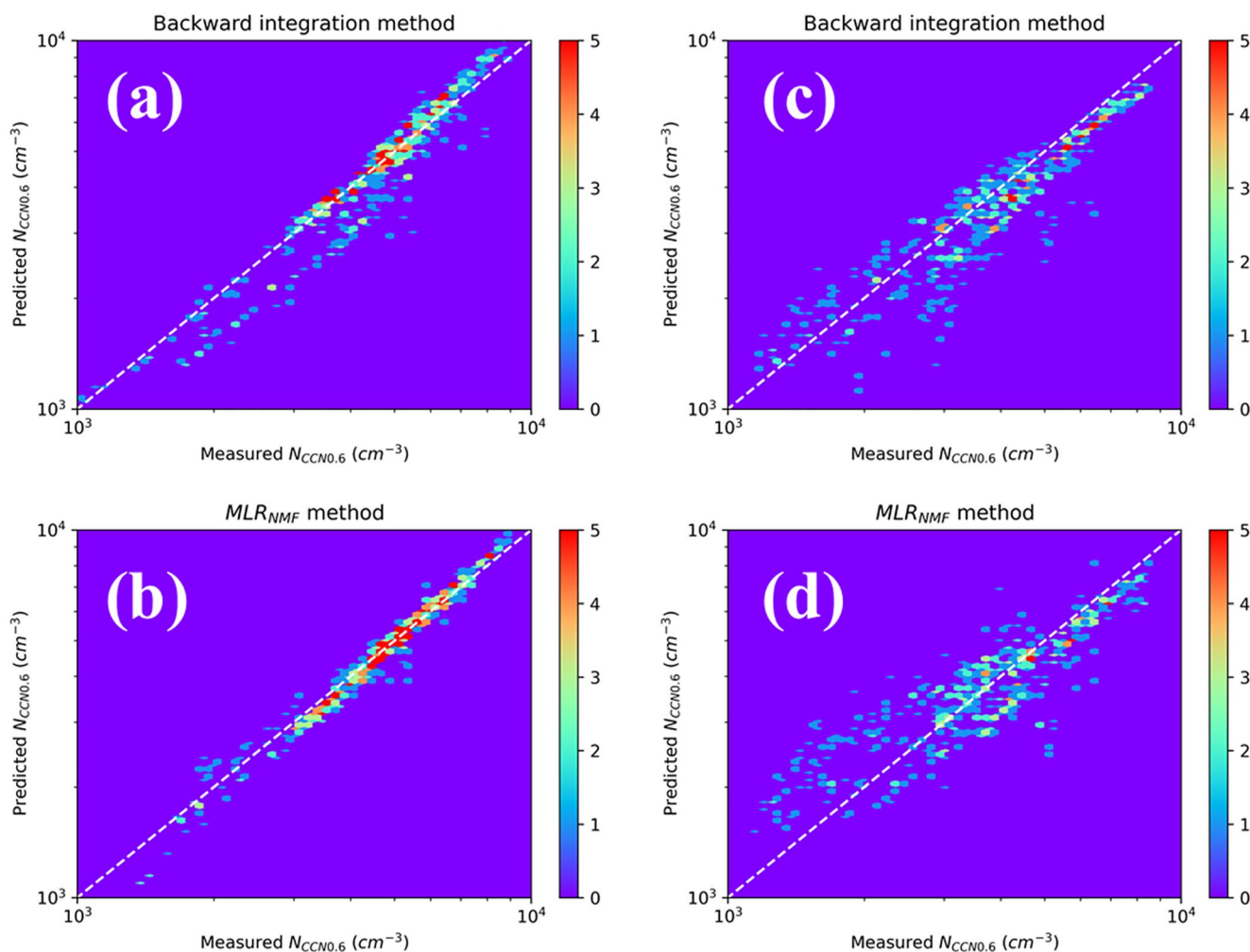


Figure 5. Hexbin plots of measured versus predicted $N_{CCN0.6}$ using the backward integration method (top), and the MLR_{NMF} method (bottom) for the training data set (left) and the test data set (right) for the Yellow Sea. There are 50×50 bins, and the color scale indicates the number of datapoints in the bin. The white dashed lines indicate the one-to-one line.

prediction methods for the training data set for the Yellow Sea. Information for each data set is described in Table 1. For both methods, most datapoints were very close to the one-to-one line. Approximately 71% and 90% of datapoints were located in $\pm 10\%$ error range for the backward integration method and the MLR_{NMF} method, respectively. Next, the N_{CCN} prediction methods were applied to the test data set over the Yellow Sea. Figures 5c and 5d compare the results of the application of both methods to the test data set for the Yellow Sea. Even for test data set, most of the datapoints were near one-to-one line (Figure 5). Approximately 89% and 80% of datapoints were located in $\pm 30\%$ error range for the backward integration method and the MLR_{NMF} method, respectively. There were two main cases of relatively poor N_{CCN} prediction over the Yellow Sea. The first was a sudden rise of aerosol concentrations of tens of nm in diameter on 22 April 2019. At this time, N_{CCN} and the sulfate mass concentration increased, while the number concentration of large aerosols did not increase much (Park et al., 2021). The second case was when NPF occurred and the nucleated aerosols were growing. In the backward integration method, it was considered that these aerosols were smaller than the critical diameter for CCN activation, so N_{CCN} was underpredicted. On the other hand, in the MLR_{NMF} method, the concentration of small aerosols increased very significantly and some of these particles did act as CCN, so N_{CCN} was overpredicted. That is, N_{CCN} can be overpredicted or underpredicted when a large portion of aerosols has the aerosol hygroscopicity that is significantly different from that is typically exhibited in the training data set.

Similarly, the same methods were applied to the data measured in Seoul. The aerosol number size distribution in Seoul was also decomposed into six distinct bases and their weightings, as shown in Figure 6. Except

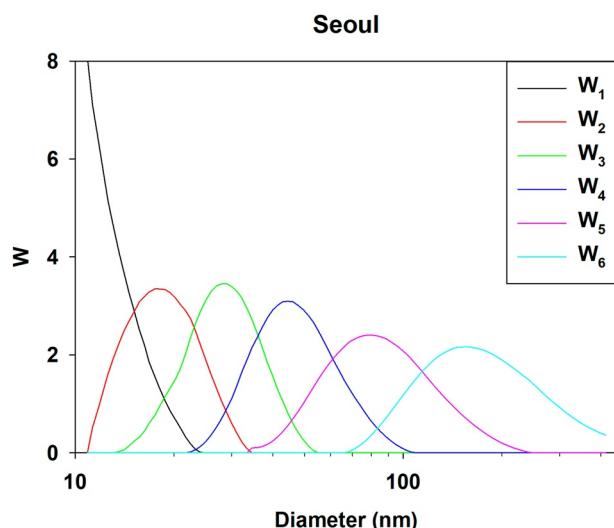


Figure 6. Same as Figure 4, except for Seoul.

for the two adjacent bases, there was low collinearity between weightings of different bases for Seoul data (Figure S2 in Supporting Information S1). Similarly to the results of the Yellow Sea, the coefficient values of H_1 – H_6 indicated that the aerosols in the first three bases (W_1 – W_3) cannot easily act as CCN whereas the aerosols in the last basis (W_6) are likely to act as CCN and approximately 30%–45% of aerosols in the fourth and the fifth basis (W_4 and W_5) can act as CCN at 0.6% S (Table 3; Figure S2 in Supporting Information S1). That is, whether aerosols can act as CCN or not depends on whether their diameter is larger or smaller than what value slightly larger than 79.1 nm (mode diameter of W_5) in Seoul. This diameter was similar to the critical diameter obtained from the backward integration method (74.5 nm) in Seoul. But these diameters were larger than those over the Yellow Sea, which indicates that aerosols in Seoul have to become larger to act as CCN at the same supersaturation than those over the Yellow Sea. Figures 7a and 7b compare the results of the application of both methods to the training data set for Seoul. Although the accuracy was not as good as that over the Yellow Sea, we consider that the N_{CCN} prediction methods for Seoul is still valid. So, both methods were applied to the test data set for Seoul (Figures 7c and 7d). For the test data set, most of the data points were near one-to-one line (Figure 7). Approximately 68% and 66% of datapoints were located in $\pm 30\%$ error range for the backward integration method and the MLR_{NMF} method, respectively.

4.2. Performance Examination

Prior to examining the N_{CCN} prediction method, to justify whether it is correct to decompose the training data set and the test data set using the same bases of NMF, RMSE between reconstructed and original aerosol size distributions ($dN/d\log D_p$) were calculated for both the training data set and the test data. For the Yellow Sea data, the RMSE for the training data set and test data set were 609 and 1,051 cm^{-3} , respectively, and for Seoul data, the RMSE for the training data set and test data set were 847 and 759 cm^{-3} , respectively. For both regions, the RMSE for the test data set did not increase significantly compared to that for the training data set. Therefore, the assumption, that the aerosol number size distribution data sets in the training and prediction stages have the same distinct modes, seems acceptable. The N_{CCN} prediction methods were examined using several performance metrics described in Section 3.3. Table 4 shows the result over the Yellow Sea and in Seoul. Both N_{CCN} prediction methods showed very low RMSE and very high R^2 values over the Yellow Sea. Absolute value of MFB and MFE were also well below performance goal values suggested by Boylan and Russell (2006). That is, both N_{CCN} prediction methods performed well for predicting $N_{CCN,0.6}$ over the Yellow Sea. For the test data set, although R^2 were slightly lower and RMSE, absolute values of MFB, and MFE were slightly higher than when applied to the training data set, they were still acceptable (Table 4). This is impressive in the sense that N_{CCN} predictions were conducted using only the aerosol number size distribution data over the Yellow Sea.

For the data set in Seoul, performance metrics showed slightly worse results than when applied to the data set over the Yellow Sea (Table 4). Since Seoul is an urban area, the types of pollutants are much more diverse than those over the Yellow Sea. Therefore, the N_{CCN} in Seoul should have been determined by more various factors and these make it more difficult to predict. The data set in Seoul is large and spans four seasons, and the time interval between the training data set and the test data set in Seoul was approximately 10 years. It is assumed that these factors contributed to lowering the accuracy of N_{CCN} prediction in Seoul. Furthermore, as described in Section 2.2, the environmental conditions of the observation site between the training data set and test data set in Seoul were quite different. The location of the inlet in the building was changed, and the operation of vehicles on the ground inside the measurement site was minimized. Furthermore, recent studies have

reported that the concentration of pollutants in East Asia is decreasing (e.g., Lin et al., 2019; S. X. Wang et al., 2014; Zheng et al., 2018). These condition changes significantly reduced aerosol and CCN number concentrations at the measurement site during the measurement period of the test data set. Despite these changes, the results seemed acceptable. It even showed much smaller RMSE and MFE and higher R^2 than those when applied to the training data

Table 3
Same as Table 2, Except for Seoul

Mode diameter (nm)	10.9	17.5	28.9	44.5	79.1	156.8
Coefficient	0.00	0.00	0.04	0.30	0.45	0.78

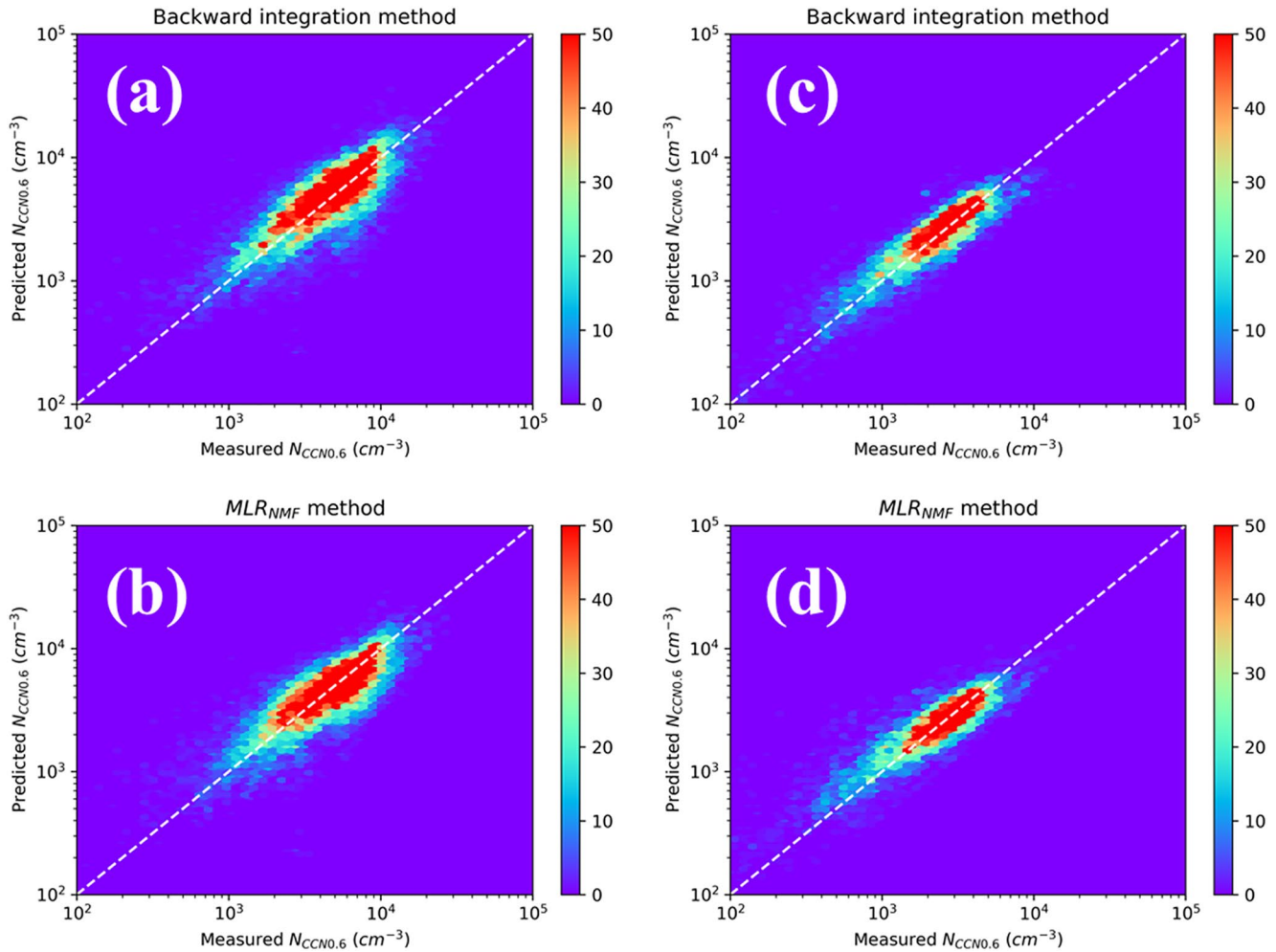


Figure 7. Same as Figure 5, except for Seoul.

Table 4

The Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Bmean Fractional Bias (MFB), and Mean Fractional Error (MFE) Between the Measured $N_{CCN0.6}$ and Predicted $N_{CCN0.6}$ Using the Two N_{CCN} Prediction Methods for the Yellow Sea and in Seoul

		Yellow sea		Seoul	
		Training data set	Test data set	Training data set	Test data set
Backward integration	R^2	0.92	0.90	0.60	0.73
	RMSE	543	655	2,047	786
	MFB	-0.032	-0.088	0.032	0.015
	MFE	0.086	0.144	0.295	0.208
MLR _{NMF}	R^2	0.98	0.81	0.63	0.71
	RMSE	262	749	1,874	824
	MFB	-0.021	0.002	-0.019	0.033
	MFE	0.050	0.165	0.279	0.235

Note. The unit of RMSE is cm^{-3} .

set (Table 4). This improvement in prediction accuracy is thought to be due to the reduction of the influence of direct local emissions at the observation site by the renovation of the Yonsei University campus. Direct influence of local emissions near the measurement site would have made N_{CCN} prediction difficult. In other words, the N_{CCN} after the changes in the environmental conditions of the observation site can be predicted with confidence using the method developed using the data collected before the changes. Furthermore, this suggests that the N_{CCN} of other observation sites with different measurement environments are also predictable.

To note is that the MLR_{NMF} method performed better than the backward integration method for the Yellow Sea training data set (i.e., lower RMSE, absolute value of MFB, and MFE and higher R^2 , Table 4). However, for the test data set, the MLR_{NMF} method showed slightly worse results than the backward integration method. Regression methods show generally better predictive result for training data set than for test data set because of overfitting. In this study, however, this difference was minimal, and as discussed in Section 3.2, the MLR_{NMF} method has intrinsic advantages over the backward integration method, so the MLR_{NMF} method is still considered highly valid. On the other hand, the results for the test data set were better than those for the training data set in Seoul. This indicates that the overfitting problem was negligible in

Table 5

The Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Fractional Bias (MFB), and Mean Fractional Error (MFE) Between the Measured $N_{CCN0.6}$ and Predicted $N_{CCN0.6}$ for Seoul When the $N_{CCN0.6}$ in the Training Data Set Were Doubled or Halved

		$2*N_{CCN}$		$0.5*N_{CCN}$	
		Training data set	Test data set	Training data set	Test data set
Backward integration	R^2	0.61	0.66	0.48	0.57
	RMSE	5,654	2,893	3,340	1,755
	MFB	0.598	0.667	−0.641	−0.713
	MFE	0.617	0.669	0.651	0.716
MLR _{NMF}	R^2	0.63	0.71	0.63	0.71
	RMSE	5,872	2,905	3,355	1,689
	MFB	0.630	0.682	−0.664	−0.622
	MFE	0.648	0.684	0.675	0.640

Note. The unit of RMSE is cm^{-3} .

the N_{CCN} prediction in Seoul. In the training data set, the hourly averaged data were used, despite the fact that the aerosol number size distribution and $N_{CCN0.6}$ can change over an hour. Meanwhile, SMPS data for each scan time (~ 2 min) and the average $N_{CCN0.6}$ values during the corresponding SMPS scanning time were used in the test data set, and maybe that was the reason why the N_{CCN} prediction for the test data set showed better results.

4.3. Advantages of the MLR_{NMF} Method

As shown in Sections 4.1 and 4.2, the accuracy of N_{CCN} predictions by the MLR_{NMF} method was similar to that by the backward integration method. Nevertheless, the reason for emphasizing the validity of the MLR_{NMF} method in this study is the robustness of the method and the capability to take into account external mixing. As described in Section 3.1, the backward integration method has been known to be sensitively affected by measurement errors. Ideally, the observed values should always represent the true values, but in practice, the errors that arise from various factors always exist. Ideally, the integral value of the aerosol number size distribution measured by the SMPS, and the value of the total aerosol number concentration measured by the CPC should be approximately the same. However, due to various factors, such as instrument error and differences in inlet systems, the two values sometimes showed a large difference. All measurement instruments have some degree of uncertainty. Moreover, for different instruments, additional errors arise because the measuring principles of the instruments are different. Instrument calibration is performed to reduce these errors, but even the calibration result is not perfect, and sometimes the instrument calibration process creates new errors. There were even cases where one value was twice or more higher than the other value. Therefore, in this study, the value of each bin of the aerosol number size distribution was adjusted by the ratio of the total aerosol number concentration from CPC and the integral value of the aerosol number size distribution from SMPS.

To investigate the sensitivity due to such errors, the N_{CCN} prediction methods were developed by deliberately changing the $N_{CCN0.6}$ in the training data set in Seoul by 0.5 and 2 times, while the aerosol number size distribution remained the same. Then, the newly developed methods were applied to the test data set to evaluate the accuracy of N_{CCN} prediction. Table 5 shows the results. For the backward integration method, not only the RMSE, MFB, and MFE increased but also the R^2 decreased as the $N_{CCN0.6}$ in the training data set changed. That is, when the $N_{CCN0.6}$ in the training data set were doubled, the predicted $N_{CCN0.6}$ were not uniformly doubled, but additional error arose. This is because the estimated critical diameter changed significantly. It is sometimes impossible to estimate the critical diameter, depending on the shape of the aerosol number size distribution. On the other hand, for the MLR_{NMF} method, the RMSE, MFB, and MFE increased but R^2 was maintained although the $N_{CCN0.6}$ in the training data set changed. This is because only the coefficients “a” in the MLR_{NMF} method changed. When the $N_{CCN0.6}$ in training data set were doubled, the coefficients “a” were doubled, and then the predicted $N_{CCN0.6}$ were also uniformly doubled. The result showed the same pattern when the $N_{CCN0.6}$ in the training data set were halved. That is, when the $N_{CCN0.6}$ in the training data set were halved, the coefficients “a” were halved, and then

the predicted $N_{CCN0.6}$ were also uniformly halved. In short, in the MLR_{NMF} method, if the $N_{CCN0.6}$ in the training data set are overestimated, the $N_{CCN0.6}$ in the prediction stage are also overpredicted at the same rate, and likewise if the $N_{CCN0.6}$ in the training data set are underestimated, the $N_{CCN0.6}$ in the prediction stage are also underpredicted at the same rate. Therefore, the MLR_{NMF} method developed in this study is much more robust than the backward integration method. Of course, multiplying 0.5 or 2 times creates an extreme case, so in general, the difference may be less than that presented in this section. However, since small uncertainty of instruments can be critical for the backward integration method depending on the shape of the aerosol number size distribution as described in Section 3.1, there is no doubt that the MLR_{NMF} method is more robust than the backward integration method.

The MLR_{NMF} method has the capability to take into account external mixing, unlike the methods proposed in other studies. Ervens et al. (2010) reported that inappropriate assumptions about organic composition and mixing state could lead to the high uncertainties of the N_{CCN} prediction especially in polluted urban environment with fresh aerosols sources. Therefore, it is a great advantage to have the capability to take into account mixing state of aerosols. N. Kim et al. (2017, 2018, 2020) reported the results of aerosol hygroscopicity measurements at the campus of the Korea Institute of Science and Technology, located approximately 10 km from the measurement site in this study, during the Megacity Air pollution Studies-Seoul (MAPS Seoul) campaign (May–June 2015), and at the Olympic Park, located approximately 17 km from the measurement site in this study, during the Korea-US Air Quality (KORUS-AQ) campaign (May–June 2016). These studies also found that external mixing was prevalently observed during the campaign. The kappa value (κ), which indicates aerosol hygroscopicity (Petters & Kreidenweis, 2007), ranged from 0.11 to 0.27 for the aerosols in 30–150 nm diameter range (N. Kim et al., 2017, 2018). When this kappa value range is converted to the critical diameter for activation at 0.6% S, it ranges from 52 to 70 nm. In this study, the averaged critical diameter in Seoul obtained from backward integration method was 74.5 nm, and therefore this value was slightly larger than the critical diameter calculated from the kappa value reported by N. Kim et al. (2017, 2018). The backward integration method assumes that all aerosols are internally mixed. It is presumed that this limitation caused the difference in the critical diameter. In the results obtained with the MLR_{NMF} method, on the other hand, even the largest basis does not have a coefficient of 1, which is speculated to be because the range of aerosol sizes of the basis is wide and due to some externally mixed aerosols. When aerosols with different hygroscopic properties are externally mixed, some aerosols may act as CCN, while some others may not even if their sizes are the same, depending on their hygroscopicity. Therefore, the coefficients have values between 0 and 1, allowing some consideration of these externally mixed aerosols. In the results of the MLR_{NMF} method for Seoul, 30% of aerosols in the basis with the mode diameter of 44.5 nm, 45% of aerosols in the basis with the mode diameter of 79.1 nm, and 78% of aerosols in the basis with the mode diameter of 156.8 nm were likely to act as CCN at 0.6% S, respectively (Table 3). The proportions of more hygroscopic aerosols among the aerosols with dry diameters of 50, 100, and 150 nm were 31%, 65% and 78% from the aerosol hygroscopicity measurements at the Olympic Park, respectively, and the rest were less hygroscopic aerosols ($\kappa \leq 0.05$) (N. Kim et al., 2020). That is, the MLR_{NMF} method showed similar results to those obtained from the in-situ aerosol hygroscopicity measurements. Since the basis is wider than the monodisperse particles classified from a differential mobility analyzer, it is difficult to say that the basis fully represents the particles of a particular size. Nevertheless, the fact that the MLR_{NMF} method showed similar results to those obtained from the in-situ aerosol hygroscopicity measurements indicates that the MLR_{NMF} method has the capability to take into account external mixing to some extent.

4.4. Seasonal Variability

In general, most meteorological variables have seasonal variability. In addition, human activities also show differences according to the seasons. As a result, natural and anthropogenic emissions vary by season, which means that the characteristics of aerosols and CCN may also have seasonal variability. In particular, the Korean Peninsula is a region where seasonal changes are evident due to the East Asian summer monsoon (Wu and Wang, 2002). J. H. Kim et al. (2014) and Park et al. (2015) reported clear seasonal variations in aerosol and CCN concentrations in Seoul. The aerosol and CCN number concentrations in Seoul during the winter (December–February) were approximately double those during the summer (June–July). Furthermore, frequency of NPF events also varied seasonally, with the maximum in spring.

Since the training data set in Seoul is large and spans four seasons, it can be said that the N_{CCN} prediction methods developed in this study were based on information about the average properties of aerosols for all seasons. Unfortunately, however, many of the measurements carried out around the world are not long-term measurements across

Table 6

The Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Fractional Bias (MFB), and Mean Fractional Error (MFE) Between the Measured $N_{CCN0.6}$ and Predicted $N_{CCN0.6}$ Segregated for Each Season for Seoul

		MAM	JJA	SON	DJF
Backward integration	R^2	0.73	0.69	0.72	0.75
	RMSE	808	1,028	862	1,047
	MFB	−0.034	−0.226	−0.097	0.252
	MFE	0.211	0.286	0.225	0.300
MLR _{NMF}	R^2	0.74	0.74	0.74	0.69
	RMSE	816	1,180	909	908
	MFB	−0.045	−0.320	−0.107	0.159
	MFE	0.217	0.352	0.249	0.276

Note. The unit of RMSE is cm^{-3} .

all seasons. Therefore, it is necessary to check the performance even when long-term N_{CCN} is predicted using the N_{CCN} prediction methods trained using measurement data concentrated at specific periods. In this section, we developed the N_{CCN} prediction methods using only data from each season of the training data set and applied them to predict the $N_{CCN0.6}$ of the entire test data set. The data in the training data set were divided by season, simply treating 3 months as one season (i.e., March–May; MAM, June–August; JJA, September–November; SON, and December–February; DJF). For each seasonal data, the coefficient “a” and the critical diameter of N_{CCN} prediction slightly changed. Despite these differences, the accuracy of N_{CCN} prediction did not decrease significantly (Table 6). Rather, slightly better results were obtained when only the data from the MAM data set were used. This is because the N_{CCN} prediction methods trained from only the MAM data set were very similar to that obtained using the entire training data set. The critical diameters, bases of NMF, and coefficients of MLR for the MAM data set were very similar to those obtained for the entire training data set, which implies that aerosol hygroscopicity in MAM was moderate among the four seasons. Therefore, performance metrics for the MAM data set were also very similar to those obtained for the entire training data set, even slightly higher on some metrics.

Furthermore, despite the tendency to underestimate/overestimate in JJA/DJF, MFB and MFE for all four seasons also met the performance goals except for MFB in JJA (−0.314, slightly lower than the performance goal). These results suggest that once the N_{CCN} prediction method of a specific region is developed using a considerable amount of data in a certain season, the N_{CCN} of other seasons can also be predicted to some extent. Of course, judging from the fact that the N_{CCN} prediction methods trained from JJA data underestimated, having data for all seasons would be better for improving the accuracy of N_{CCN} prediction. Since these results were obtained in the region of significant seasonal variation, such as Seoul, the results would be better for other regions with relatively smaller seasonal variations.

5. Conclusions

This study aimed to develop a new N_{CCN} prediction method based on the measured aerosol number size distribution data using MLR and NMF. The newly developed method showed similar accuracy to the existing method (backward integration method) but has advantages such as robustness and the capability to take into account external mixing compared to the existing method. Therefore, it is expected that the new method (MLR_{NMF} method) can be used to replace the existing method (backward integration method).

To increase and verify the reliability of the N_{CCN} prediction method, separate data sets were used, and sufficient time differences were given between the training data set and the test data set. The results of N_{CCN} prediction in the target regions, the Yellow Sea and Seoul, were generally acceptable. In particular, for Seoul, the predicted $N_{CCN0.6}$ was in very good agreement with the measured $N_{CCN0.6}$ for the test data set despite the large variability, such as seasonal variation and reduced traffics, in the environmental conditions of the measurement site. This means that the N_{CCN} after the changes in the conditions of the observation region can be predicted with confidence using the method developed before the changes. Furthermore, this suggests that the N_{CCN} of other urban sites whose environment is similar to Seoul can also be predictable. Of course, the

N_{CCN} prediction method developed over the Yellow Sea is also thought to be applicable to regions similar to the Yellow Sea. However, since the N_{CCN} prediction methods use average aerosol hygroscopicity information obtained in the training stage and time-varying aerosol number size distribution to predict N_{CCN} , N_{CCN} can be overpredicted or underpredicted when a large portion of aerosols has the aerosol hygroscopicity that is significantly different from that is typically exhibited in the training data set (e.g., NPF events). This reveals the limitation of the N_{CCN} prediction method that uses the average aerosol hygroscopicity information obtained in the training stage and the time-varying aerosol number size distribution. However, in long-term data, the effect of these outliers would be small, so the N_{CCN} prediction method to secure abundant N_{CCN} data is still legitimate and valuable.

Through this study, we saw the possibility of applying the developed N_{CCN} prediction method to other regions. The MLR_{NMF} method is applicable to regions where the aerosol hygroscopicity does not differ greatly from that of the training data set. In other words, this method can be applied after grouping regions with similar aerosol hygroscopicity and training the method for each group. Therefore, by classifying regions into several groups and securing additional training data set representing each of these groups, it would become possible to predict N_{CCN} using the aerosol number size distribution data in many regions around the world. First of all, in the companion paper (Park, Yum, & Seo, 2023; Part 2), we will apply the N_{CCN} prediction method developed in this study to the expanded regions in and around the Korean Peninsula. Based on this method, an N_{CCN} distribution map in and around the Korean Peninsula can also be created. Furthermore, using the aerosol number size distribution data in many regions around the world, we plan to verify the MLR_{NMF} method, and it is expected that the N_{CCN} distribution map can be expanded on a global scale. Such global N_{CCN} distribution can be used as an initial input data for global climate models, and may greatly contribute to reducing uncertainties of climate change prediction.

Data Availability Statement

The data in Seoul in 2006–2010 are available in this citation reference: Schmale, Petäjä, et al. (2017). The data in Seoul in 2018–2020 and over the Yellow Sea in 2017, 2019, and 2021 are available in this citation reference: Park, Yum, Seo, Ahn, et al. (2023). The python code for the MLR_{NMF} method is available in the form of a Jupyter notebook (Park, 2023).

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A2B5B02002458), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00246513), and the Yonsei Signature Research Cluster Program of 2023-22-0009. The authors are grateful to the National Institute of Meteorological Sciences for the support of the Gisang 1 research vessel and its crew members.

References

- Abdul-Razzak, H., & Ghan, S. J. (2000). A parameterization of aerosol activation: 2. Multiple aerosol types. *Journal of Geophysical Research*, 105(D5), 6837–6844. <https://doi.org/10.1029/1999JD901161>
- Andreae, M. O., & Rosenfeld, D. (2008). Aerosol–cloud–precipitation interactions. Part 1. The nature and sources of cloud-active aerosols. *Earth-Science Reviews*, 89(1), 13–41. <https://doi.org/10.1016/j.earscirev.2008.03.001>
- Bhattach, D., & Tripathi, S. N. (2015). CCN closure study: Effects of aerosol chemical composition and mixing state. *Journal of Geophysical Research: Atmospheres*, 120(2), 766–783. <https://doi.org/10.1002/2014JD021978>
- Boylan, J. W., & Russell, A. G. (2006). PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. *Atmospheric Environment*, 40(26), 4946–4959. <https://doi.org/10.1016/j.atmosenv.2005.09.087>
- Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(12), 4164–4169. <https://doi.org/10.1073/pnas.0308531101>
- Cerully, K. M., Bougiatioti, A., Hite, J. R., Jr., Guo, H., Xu, L., Ng, N. L., et al. (2015). On the link between hygroscopicity, volatility, and oxidation state of ambient and water-soluble aerosols in the southeastern United States. *Atmospheric Chemistry and Physics*, 15(15), 8679–8694. <https://doi.org/10.5194/acp-15-8679-2015>
- Dusek, U., Frank, G. P., Hildebrandt, L., Curtius, J., Schneider, J., Walter, S., et al. (2006). Size matters more than chemistry for cloud-nucleating ability of aerosol particles. *Science*, 312(5778), 1375–1378. <https://doi.org/10.1126/science.1125261>
- Ervens, B., Cubison, M. J., Andrews, E., Feingold, G., Ogren, J. A., Jimenez, J. L., et al. (2010). CCN predictions using simplified assumptions of organic aerosol composition and mixing state: A synthesis from six different locations. *Atmospheric Chemistry and Physics*, 10(10), 4795–4807. <https://doi.org/10.5194/acp-10-4795-2010>
- Fanourgakis, G. S., Kanakidou, M., Nenes, A., Bauer, S. E., Bergman, T., Carslaw, K. S., et al. (2019). Evaluation of global simulations of aerosol particle and cloud condensation nuclei number, with implications for cloud droplet formation. *Atmospheric Chemistry and Physics*, 19(13), 8591–8617. <https://doi.org/10.5194/acp-19-8591-2019>
- Gillis, N. (2020). *Nonnegative matrix factorization*. SIAM. <https://doi.org/10.1137/1.9781611976410>
- Huebert, B. J., Bates, T., Russell, P. B., Shi, G., Kim, Y. J., Kawamura, K., et al. (2003). An overview of ACE-Asia: Strategies for quantifying the relationships between Asian aerosols and their climatic impacts. *Journal of Geophysical Research*, 108(D23), 8633. <https://doi.org/10.1029/2003JD003550>
- Hung, H.-M., Lu, W.-J., Chen, W.-N., Chang, C.-C., Chou, C. C.-K., & Lin, P.-H. (2014). Enhancement of the hygroscopicity parameter kappa of rural aerosols in northern Taiwan by anthropogenic emissions. *Atmospheric Environment*, 84, 78–87. <https://doi.org/10.1016/j.atmosenv.2013.11.032>
- Hutchins, L. N., Murphy, S. M., Singh, P., & Graber, J. H. (2008). Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, 24(23), 2684–2690. <https://doi.org/10.1093/bioinformatics/btn526>

- IPCC. (2021). In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al. (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. <https://doi.org/10.1017/9781009157896>
- Kammermann, L., Gysel, M., Weingartner, E., Herich, H., Cziczó, D. J., Holst, T., et al. (2010). Subarctic atmospheric aerosol composition: 3. Measured and modeled properties of cloud condensation nuclei. *Journal of Geophysical Research*, 115(D4), D04202. <https://doi.org/10.1029/2009JD012447>
- Kaufman, Y. J., Tanré, D., & Boucher, O. (2002). A satellite view of aerosols in the climate system. *Nature*, 419(6903), 215–223. <https://doi.org/10.1038/nature01091>
- Kim, J. H., Yum, S. S., Shim, S., Kim, W. J., Park, M., Kim, J.-H., et al. (2014). On the submicron aerosol distributions and CCN number concentrations in and around the Korean Peninsula. *Atmospheric Chemistry and Physics*, 14(16), 8763–8779. <https://doi.org/10.5194/acp-14-8763-2014>
- Kim, N., Park, M., Yum, S. S., Park, J. S., Shin, H. J., & Ahn, J. Y. (2018). Impact of urban aerosol properties on cloud condensation nuclei (CCN) activity during the KORUS-AQ field campaign. *Atmospheric Environment*, 185, 221–236. <https://doi.org/10.1016/j.atmosenv.2018.05.019>
- Kim, N., Park, M., Yum, S. S., Park, J. S., Song, I. H., Shin, H. J., et al. (2017). Hygroscopic properties of urban aerosols and their cloud condensation nuclei activities measured in Seoul during the MAPS-Seoul campaign. *Atmospheric Environment*, 153, 217–232. <https://doi.org/10.1016/j.atmosenv.2017.01.034>
- Kim, N., Yum, S. S., Park, M., Park, J. S., Shin, H. J., & Ahn, J. Y. (2020). Hygroscopicity of urban aerosols and its link to size-resolved chemical composition during spring and summer in Seoul, Korea. *Atmospheric Chemistry and Physics*, 20(19), 11245–11262. <https://doi.org/10.5194/acp-20-11245-2020>
- Kwak, N., Lee, H., Maeng, H., Seo, A., Lee, K., Kim, S., et al. (2022). Morphological and chemical classification of fine particles over the Yellow Sea during spring. *Environmental Pollution*, 305, 119286. <https://doi.org/10.1016/j.envpol.2022.119286>
- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791. <https://doi.org/10.1038/44565>
- Lee, D. D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems* (Vol. 13).
- Lee, S. (2020). Estimating the rank of a nonnegative matrix factorization model for automatic music transcription based on stein's unbiased risk estimator. *Applied Sciences*, 10(8), 2911. <https://doi.org/10.3390/app10082911>
- Li, Z., Wang, Y., Guo, J., Zhao, C., Cribb, M. C., Dong, X., et al. (2019). East Asian study of tropospheric aerosols and their impact on regional clouds, precipitation, and climate (EAST-AIRCPC). *Journal of Geophysical Research: Atmospheres*, 124(23), 13026–13054. <https://doi.org/10.1029/2019JD030758>
- Liang, M., Tao, J., Ma, N., Kuang, Y., Zhang, Y., Wu, S., et al. (2022). Prediction of CCN spectra parameters in the North China Plain using a random forest model. *Atmospheric Environment*, 289, 119323. <https://doi.org/10.1016/j.atmosenv.2022.119323>
- Lin, C.-A., Chen, Y.-C., Liu, C.-Y., Chen, W.-T., Seinfeld, J. H., & Chou, C. C.-K. (2019). Satellite-derived correlation of SO₂, NO₂, and aerosol optical depth with meteorological conditions over East Asia from 2005 to 2015. *Remote Sensing*, 11(15), 1738. <https://doi.org/10.3390/rs11151738>
- Lohmann, U., & Feichter, J. (2005). Global indirect aerosol effects: A review. *Atmospheric Chemistry and Physics*, 5(3), 715–737. <https://doi.org/10.5194/acp-5-715-2005>
- Mitchell, T. M., & Mitchell, T. M. (1997). *Machine learning* (Vol. 1). McGraw-hill.
- Moore, R. H., Bahreini, R., Brock, C. A., Froyd, K. D., Cozic, J., Holloway, J. S., et al. (2011). Hygroscopicity and composition of Alaskan Arctic CCN during April 2008. *Atmospheric Chemistry and Physics*, 11(22), 11807–11825. <https://doi.org/10.5194/acp-11-11807-2011>
- Nair, A., Yu, F., Jost, P. C., DeMott, P., Levin, E., Jimenez, J., et al. (2021). Machine learning uncovers aerosol size information from chemistry and meteorology to quantify potential cloud-forming particles. *Geophysical Research Letters*, 48(21), e2021GL094133. <https://doi.org/10.1029/2021GL094133>
- Park, M. (2023). Python code of the multiple linear regression using non-negative matrix factorization method to predict CCN number concentration [Software]. Figshare. <https://doi.org/10.6084/m9.figshare.24296770>
- Park, M., Yum, S. S., Seo, P., Ahn, C., Kim, N., Anderson, B. E., & Thornhill, K. L. (2023). A new CCN number concentration prediction method based on multiple linear regression and non-negative matrix factorization: 2. Application to secure CCN spectra in and around the Korean Peninsula. *Journal of Geophysical Research: Atmospheres*. <https://doi.org/10.1029/2023JD039234>
- Park, M., Yum, S. S., & Kim, J. H. (2015). Characteristics of submicron aerosol number size distribution and new particle formation events measured in Seoul, Korea, during 2004–2012. *Asia-Pacific Journal of Atmospheric Sciences*, 51(1), 1–10. <https://doi.org/10.1007/s13143-014-0055-0>
- Park, M., Yum, S. S., Kim, N., Anderson, B. E., Beyersdorf, A., & Thornhill, K. L. (2020). On the submicron aerosol distributions and CCN activity in and around the Korean Peninsula measured onboard the NASA DC-8 research aircraft during the KORUS-AQ field campaign. *Atmospheric Research*, 243, 105004. <https://doi.org/10.1016/j.atmosres.2020.105004>
- Park, M., Yum, S. S., Kim, N., Cha, J. W., & Ryoo, S. B. (2016). Characteristics of aerosol and cloud condensation nuclei concentrations measured over the Yellow Sea on a meteorological research vessel, GISANG 1. *Atmosphere*, 26(2), 243–256. <https://doi.org/10.14191/Atmos.2016.26.2.243>
- Park, M., Yum, S. S., Kim, N., Cha, J. W., Shin, B., & Ryoo, S.-B. (2018). Characterization of submicron aerosols and CCN over the Yellow Sea measured onboard the Gisang 1 research vessel using the positive matrix factorization analysis method. *Atmospheric Research*, 214, 430–441. <https://doi.org/10.1016/j.atmosres.2018.08.015>
- Park, M., Yum, S. S., Kim, N., Jeong, M., Yoo, H.-J., Kim, J. E., et al. (2021). Characterization of submicron aerosols over the Yellow Sea measured onboard the Gisang 1 research vessel in the spring of 2018 and 2019. *Environmental Pollution*, 284, 117180. <https://doi.org/10.1016/j.envpol.2021.117180>
- Park, M., Yum, S. S., & Seo, P. (2023). Aerosol size distribution and CCN number concentration data_Seoul and Yellow Sea [Dataset]. Figshare. <https://doi.org/10.6084/m9.figshare.22723570.v1>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Petters, M. D., & Kreidenweis, S. M. (2007). A single parameter representation of hygroscopic growth and cloud condensation nucleus activity. *Atmospheric Chemistry and Physics*, 7(8), 1961–1971. <https://doi.org/10.5194/acp-7-1961-2007>
- Pöhlker, M. L., Pöhlker, C., Ditas, F., Klimach, T., Hrabě de Angelis, I., Araújo, A., 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. *Atmospheric Chemistry and Physics*, 16(24), 15709–15740. <https://doi.org/10.5194/acp-16-15709-2016>

- Quinn, P. K., Coffman, D. J., Bates, T. S., Welton, E. J., Covert, D. S., Miller, T. L., et al. (2004). Aerosol optical properties measured on board the Ronald H. Brown during ACE-Asia as a function of aerosol chemical composition and source region. *Journal of Geophysical Research*, 109(D19), D19S01. <https://doi.org/10.1029/2003JD004010>
- Ramachandran, S., Rupakheti, M., & Lawrence, M. G. (2020). Aerosol-induced atmospheric heating rate decreases over South and East Asia as a result of changing content and composition. *Scientific Reports*, 10(1), 1–17. <https://doi.org/10.1038/s41598-020-76936-z>
- Ramachandran, S., Rupakheti, M., & Cherian, R. (2022). Insights into recent aerosol trends over Asia from observations and CMIP6 simulations. *Science of the Total Environment*, 807, 150756. <https://doi.org/10.1016/j.scitotenv.2021.150756>
- Ramanathan, V., Crutzen, P. J., Lelieveld, J., Mitra, A. P., Althausen, D., Anderson, J., et al. (2001). Indian Ocean Experiment: An integrated analysis of the climate forcing and effects of the great Indo-Asian haze. *Journal of Geophysical Research*, 106(D22), 28371–28398. <https://doi.org/10.1029/2001JD900133>
- Riemer, N., Ault, A. P., West, M., Craig, R. L., & Curtis, J. H. (2019). Aerosol mixing state: Measurements, modeling, and impacts. *Reviews of Geophysics*, 57(2), 187–249. <https://doi.org/10.1029/2018RG000615>
- Schmale, J., Henning, S., Decesari, S., Henzing, B., Keskinen, H., Sellegri, K., et al. (2018). Long-term cloud condensation nuclei number concentration, particle number size distribution and chemical composition measurements at regionally representative observatories. *Atmospheric Chemistry and Physics*, 18(4), 2853–2881. <https://doi.org/10.5194/acp-18-2853-2018>
- Schmale, J., Henning, S., Henzing, B., Keskinen, H., Sellegri, K., Ovadnevaite, J., et al. (2017). Collocated observations of cloud condensation nuclei, particle size distributions, and chemical composition. *Scientific Data*, 4(1), 170003. <https://doi.org/10.1038/sdata.2017.3>
- Schmale, J., Petäjä, T., Kulmala, M., Ovadnevaite, J., O'Dowd, C. D., Matsuki, A., et al. (2017). Data from collocated observations of cloud condensation nuclei, particle size distributions, and chemical composition (Version 1) [Dataset]. Figshare. <https://doi.org/10.6084/m9.figshare.c.3471585.v1>
- Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric chemistry and physics: From air pollution to climate change*. John Wiley & Sons.
- Sotiropoulou, R. P., Medina, J., & Nenes, A. (2006). CCN predictions: Is theory sufficient for assessments of the indirect effect? *Geophysical Research Letters*, 33(5), L05816. <https://doi.org/10.1029/2005GL025148>
- Wall, C. J., Norris, J. R., Possner, A., McCoy, D. T., McCoy, I. L., & Lutsko, N. J. (2022). Assessing effective radiative forcing from aerosol–cloud interactions over the global ocean. *Proceedings of the National Academy of Sciences of the United States of America*, 119(46), e2210481119. <https://doi.org/10.1073/pnas.2210481119>
- Wang, J., Cubison, M. J., Aiken, A. C., Jimenez, J. L., & Collins, D. R. (2010). The importance of aerosol mixing state and size-resolved composition on CCN concentration and the variation of the importance with atmospheric aging of aerosols. *Atmospheric Chemistry and Physics*, 10(15), 7267–7283. <https://doi.org/10.5194/acp-10-7267-2010>
- Wang, S. X., Zhao, B., Cai, S. Y., Klimont, Z., Nielsen, C. P., Morikawa, T., et al. (2014). Emission trends and mitigation options for air pollutants in East Asia. *Atmospheric Chemistry and Physics*, 14(13), 6571–6603. <https://doi.org/10.5194/acp-14-6571-2014>
- Wu, R., & Wang, B. (2002). A contrast of the East Asian summer monsoon–ENSO relationship between 1962–77 and 1978–93. *Journal of Climate*, 15(22), 3266–3279. [https://doi.org/10.1175/1520-0442\(2002\)015<3266:ACOTEA>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<3266:ACOTEA>2.0.CO;2)
- Zhang, F., Li, Y., Li, Z., Sun, L., Li, R., Zhao, C., et al. (2014). Aerosol hygroscopicity and cloud condensation nuclei activity during the AC 3 Exp campaign: Implications for cloud condensation nuclei parameterization. *Atmospheric Chemistry and Physics*, 14(24), 13423–13437. <https://doi.org/10.5194/acp-14-13423-2014>
- Zhang, Q., Streets, D. G., Carmichael, G. R., He, K. B., Huo, H., Kannari, A., et al. (2009). Asian emissions in 2006 for the NASA INTEX-B mission. *Atmospheric Chemistry and Physics*, 9(14), 5131–5153. <https://doi.org/10.5194/acp-9-5131-2009>
- Zheng, B., Chevallier, F., Ciais, P., Yin, Y., Deeter, M. N., Worden, H. M., et al. (2018). Rapid decline in carbon monoxide emissions and export from East Asia between years 2005 and 2016. *Environmental Research Letters*, 13(4), 44007. <https://doi.org/10.1088/1748-9326/aab2b3>