Referee 1

Ren et al. used the Random Forest Regression (RFR) model to predict cloud condensation nuclei (CCN) concentrations in the North China Plain. Their results showed a reduced prediction bias compared to WRF-Chem. Moreover, by incorporating observational data such as $PM_{2.5}$, NO_2, SO_2, the model captured the long-term decreasing trend in aerosol concentration from 2014 to 2018.

While the approach and the results are interesting, I have several concerns regarding the current manuscript:

1. The methodology particularly and the manuscript generally lacks critical information for readers to follow. Why the authors decided to put information such as study domain, details about WRF-Chem, RFR configurations/training and validating and more in the Supplemental Information (SI)? I find it's very hard to understand and follow the methodology session.

Re: Thank you for your efforts and time on handling the paper. We have put more information in the main text of the paper and updated the section of **Methods** and see as follows or **Lines 106-225**: "

2. Methods

2.1 Study area

In this work, we select the North China Plain (NCP) (32°-40°N and 114°-121°E) as the study area. Being one of the most polluted areas in China, the aerosol particles in NCP are with more complex composition and mixing state, which leads to great challenge in accurate prediction of cloud concentration nuclei (CCN) concentrations. In recent years, emissions of gas pollutants and fine particles have shown a significant downward trend year by year (Wei et al., 2023) due to the implementation of the vigorous emission reduction in China (Zheng et al., 2018). This also makes changes in aerosols CCN activity in the study area from the point of view in assessment of the climate effect of aerosols.

2.2 Model construction and validation

Here we develop the ML-based N_{CCN} prediction model by employing the Random Forest Regression method (RFRM) that has been demonstrated and can solve multivariate and nonlinear regression problems (Nair and Yu, 2020; Liang et al., 2022). The diagram of the model construction and the N_{CCN} prediction is shown in Figure 1. Due to lack of a large spatial scale observed N_{CCN} data, we use simulated N_{CCN} by WRF-Chem model as the targeted variable that can basically capture the ambient temporal variability of CCN concentration despite a deviation of ~40% by comparing with our six field observations (Figure 4). The input parameters include the chemical components of $PM_{2.5}$ (organic, sulfate, nitrate, ammonium, black carbon) from the Tsinghua University Tracking Air Pollution in China dataset (Liu et al., 2022) and gas and particulate pollutants (nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3) and $PM_{2.5}$) collected from the China National Environmental Monitoring Centre network. Meteorological parameters are from the European Centre for Medium-range Weather Forecasts Reanalysis version 5 (ERA-5) and include temperature, relative humidity (RH), total precipitation (TP), wind speed

(WS), wind direction (WD), planetary boundary layer height (BLH), surface pressure (SP) and surface net solar radiation (SNSR). These datasets have undergone validation and been proven highly suitable for developing machine learning models for atmospheric applications (Nair and Yu, 2020; Wei et al., 2023). Cartesian coordinates were also added as input due to the spatiotemporal nature of the input data (Yang et al., 2022). Supplemental Table 1 provides more details about the input parameters.

When constructing the model, all the aforementioned species data (predictor and target variables) were processed to match the same spatiotemporal resolution, and then split into 7:3 ratio for model training and testing respectively. As a result, a total of 274365 samples is included in the training datasets. In order to assure a stronger generalization ability of the N_{CCN} prediction model, the 10-fold cross-validation is adopted (Wei et al., 2023). The optimization parameters of RF model were examined by varying hyperparameters (Fig. S1). In addition, cross-validation (CV) is applied to select the hyperparameters during the data preprocessing (Yang et al., 2022). The CV results showed that when the number of trees (n_estimators) was less than 200, the prediction accuracy increased rapidly with the increase of the number of trees, and then gradually stabilized. According to the CV score and the number of data sample, the number of trees was set to 500 in this study. The impact of max depth on the CV score showed that, with the increase of depth, the complexity of the model increases. Thus, the max depth is set to 28. Also, the model generalization error was larger when the minimum sample number of the leaf and branch node are large, indicating that the model itself is close to the optimal model complexity level. Therefore, a higher value was set given the large sample size in this case. The influence of the maximum selection feature number on CV score showed a trend of increasing first and then decreasing, so the maximum value of CV curve was set to 16.

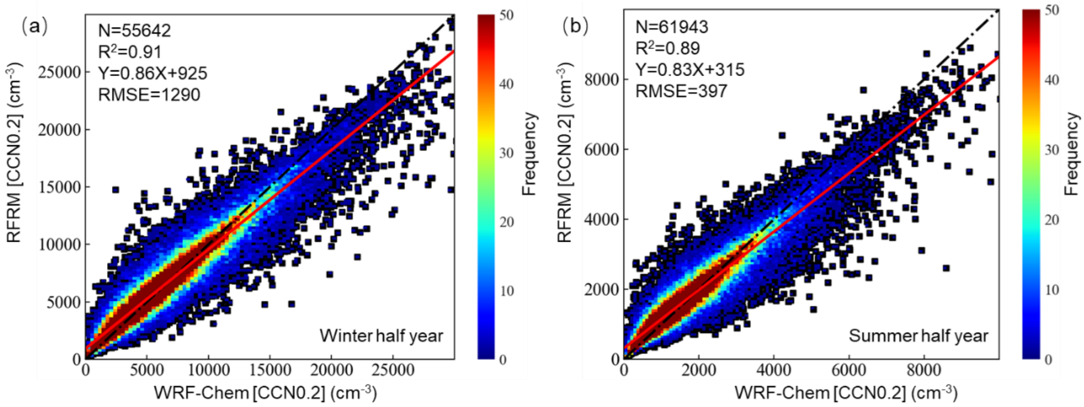

Figure 2. Comparison of RFRM retrieval and WRF-Chem simulated N_{CCN} at S=0.2%. (a and b) Density plots of retrieval N_{CCN} at S=0.2% as a function of the simulations from WRF-Chem on the testing dataset.

Fig. 2 shows the performance of our developed RFRM models by comparing the testing dataset (sample size: N=117585) of CCN simulated by WRF-Chem with the predicted CCN. Here the quality metrics for model performance are based on the correlation coefficient (R^2), root mean square error (RMSE) and the slope of the RFRM predicted and the WRF-Chem simulated CCN concentrations. It showed that the estimated N_{CCN} at S=0.2% are highly correlated with the values from WRF-Chem, with the correlation R^2 of ~0.89-0.91 and slopes of 0.83 and 0.86 in our developed RFRM model. This suggests that our model works well in estimating N_{CCN} with a high aerosol loading environment. We also found that the accuracy of the CCN prediction will deteriorate slightly if not including the information of chemical compositions (Fig. S2), or if using XGBoost algorithm (Fig.S3) when constructing the model.

Given that the ultimate goal of model development is to more accurately predict the actual atmospheric concentration of CCN, we further conducted a comparative analysis of the prediction results against the in-situ observation data in this region, spanning from the hourly scale to the interannual scale (see Figure 3-6), which will be presented and discussed in detail in Section 3.2 and 3.3.

2.3 Data and other details in the model construction

2.3.1 N_{CCN} simulated by WRF-Chem model

The WRF-Chem version 4.1.5 is used to simulate N_{CCN} in this study, which nested a domain in 10 km×10 km covering the entire NCP (Fig. S4) and contained 181×170 grids. The simulation in WRF-Chem is conducted from 1 January 2014 to 31 December 2018 with an hourly resolution. In the WRF-Chem modeling system, the sectional Model for Simulating Aerosol Interactions and Chemistry (MOSAIC), the Morrison two-moment scheme (Morrison et al., 2009) and the Carbon Bond Mechanism Z photochemical mechanism (Zaveri et al., 1999) are employed. We also compared the simulation using the Regional Acid Deposition Model (Stockwell et al., 1990) and the Lin microphysics scheme (Lin et al., 1983). Considering the calculation efficiency and accuracy with the measurements, the CBMZ-MOSAIC and Morrison 2-monment scheme were finally applied to simulate the long-term CCN concentration. More details about the other parameterizations used for WRF-Chem simulation were given in SI.

2.3.2 Gound-based measurements and datasets

Ground measurements of atmospheric gaseous precursors, fine particles chemical compositions, and CCN number concentration (at supersaturations of 0.2% and 0.4%) were collected during six field campaigns at three sites in the NCP (Fig. 4), used to assess the performance of the developed ML-based model in predicting N_{CCN}. The six campaigns were conducted as follows: at the Beijing (BJ) site from 8–30 November 2014, 20 August to 6 October 2015, 16 November to 20 December 2016, and 28 May to 27 June 2017; at the Xingtai (XT) site from 17 May to 14 June 2016; and at the Gucheng (GC) site from 23 January to 3 February 2018. They are accordingly named BJ2014_WIN, BJ2015_AUT, BJ2016_WIN, BJ2017_SUM, XT2016_SUM, and GC2018_WIN (Fig. 4a).

The BJ site (Longitude: 116.37° E; Latitude: 39.97° N) is located at the meteorological tower station of the Institute of Atmospheric Physics, Chinese Academy of Sciences. It is representative of the general emission conditions in urban areas of the northern NCP. The primary pollution sources here are surrounding traffic and residential emissions. The XT site (Longitude: 114.37° E; Latitude: 37.18° N) is situated at a national weather station. It is primarily influenced by emissions from surrounding towns and factories (e.g., coal-fired power plants, coking, steel, cement, and chemical industries) and thus reflects polluted suburban conditions in the southern NCP. The GC site (Longitude: 115.74° E; Latitude: 39.15° N) is located at the

Integrated Ecological-Meteorological Observation and Experiment Station of the Chinese Academy of Meteorological Sciences. Surrounded mainly by nearby villages, farmland, and transportation networks, this site represents the regional background pollution in the northern NCP.

The CCN number concentrations were measured by using the Droplet Measurement Technologies CCN counter (model CCNC-100, DMT Inc. Lance et al., 2006) at BJ and XT site. The supersaturation (*S*) levels set for each CCN measurement cycle were 0.1%, 0.2%, 0.4%, and 0.8%, respectively. Another measurement at GC site was referred from Zhang et al. (2020). In this study, the comparisons between the measured and predicted N_{CCN} were mostly based on the value at *S*=0.2% and *S*=0.4%. The observed N_{CCN} varies from a few hundred to tens of thousands at these sites, and the campaign mean mass concentration of $PM_{2.5}$ ranges from 35.6 to 160 μg m^{-3} (Fig. 4b), indicating that the observations can represent various atmospheric conditions, spanning from clean to polluted in the region. More details about the observations could be found in Fan et al. (2020), Ren et al. (2018), and Zhang et al. (2019). In addition, the long-term measurement of particle number size distribution (PNSD) at a field site in Beijing (Fig. S5, Shang et al., 2022) is also used for deriving the long-term trend of yearly averaged N_{CCN}.

For example, it is unclear whether nitrate was the output of the WRF-Chem?

Re: The nitrate was from the Tsinghua University Tracking Air Pollution in China dataset (Liu et al., 2022). We have provided a detailed introduction to the data in the methodology section and see as follows or **Lines 125-127**:

"…The input parameters include the chemical components of $PM_{2.5}$ (organic, sulfate, nitrate, ammonium, black carbon) from the Tsinghua University Tracking Air Pollution in China dataset (Liu et al., 2022) …"

A minor point: "new model" is not a preferred terminology in the result figures.

Re: The "new model" has been replaced with "RFRM" in the result figures.

While the relative importance of the input parameters is informative, it does not adequately explain the substantial improvement in CCN predictions. With a large bias in WRF-Chem simulations (~39%), it is unclear why the authors included WRF-Chem outputs as predictors?

Re: Currently, N_{CCN} observation is still very lacking, mostly consisting of single field observations (Ren et al., 2018; Zhang et al., 2019). The N_{CCN} from satellite-retrieved method can be as high as there are usually large deviations of -30%–+90% in the estimation accuracy (Shen et al., 2019). Numerical model has been demonstrated that it can capture the relative amplitude of the variability of the aerosol particle number concentration and CCN number concentration (Fanourgakis et al., 2019; Nair and Yu, 2020). To develop a spatiotemporal-scale model for predicting CCN concentrations, the CCN concentration output by WRF-Chem is used as the target variable. We have added some explanations as follows or see **Lines 122-125:**

"…Due to lack of a large spatial scale observed N_{CCN} data, we use simulated N_{CCN} by WRF-Chem model as the targeted variable that can basically capture the ambient temporal variability of CCN concentration despite a deviation of ~40% by comparing with our six field observations (Figure 4) …"

On the other hand, if the authors only used observational data to train the ML model, then the current results would not be apple to apple comparison. A more appropriate benchmark would be a comparison between the RFRM and other machine learning methods used in previous studies.

Re: Thank you for your suggestion. We have added detailed description regarding the RFRM model construction, and also compared with XGBoost model as follows or see **Lines 117-173:**

"… **2.2 Model construction and validation**

Here we develop the ML-based N_{CCN} prediction model by employing the Random Forest Regression method (RFRM) that has been demonstrated and can solve multivariate and nonlinear regression problems (Nair and Yu, 2020; Liang et al., 2022). The diagram of the model construction and the N_{CCN} prediction is shown in Figure 1. Due to lack of a large spatial scale observed N_{CCN} data, we use simulated N_{CCN} by WRF-Chem model as the targeted variable that can basically capture the ambient temporal variability of CCN concentration despite a deviation of ~40% by comparing with our six field observations (Figure 4). The input parameters include the chemical components of $PM_{2.5}$ (organic, sulfate, nitrate, ammonium, black carbon) from the Tsinghua University Tracking Air Pollution in China dataset (Liu et al., 2022) and gas and particulate pollutants (nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3) and $PM_{2.5}$) collected from the China National Environmental Monitoring Centre network. Meteorological parameters are from the European Centre for Medium-range Weather Forecasts Reanalysis version 5 (ERA-5) and include temperature, relative humidity (RH), total precipitation (TP), wind speed (WS), wind direction (WD), planetary boundary layer height (BLH), surface pressure (SP) and surface net solar radiation (SNSR). These datasets have undergone validation and been proven highly suitable for developing machine learning models for atmospheric applications (Nair and Yu, 2020; Wei et al., 2023). Cartesian coordinates were also added as input due to the spatiotemporal nature of the input data (Yang et al., 2022). Supplemental Table 1 provides more details about the input parameters.

When constructing the model, all the aforementioned species data (predictor and target variables) were processed to match the same spatiotemporal resolution, and then split into 7:3 ratio for model training and testing respectively. As a result, a total of 274365 samples is included in the training datasets. In order to assure a stronger generalization ability of the N_{CCN} prediction model, the 10-fold cross-validation is adopted (Wei et al., 2023). The optimization parameters of RF model were examined by varying hyperparameters (Fig. S1). In addition, cross-validation (CV) is applied to select the hyperparameters during the data preprocessing (Yang et al., 2022). The CV results showed that when the number of trees (n_estimators) was less than 200, the prediction accuracy increased rapidly with the increase of the number of trees, and then gradually stabilized. According to the CV score and the number of data sample, the number of trees was set to 500 in this study. The impact of max depth on the CV score showed that, with the increase of depth, the complexity of the model increases. Thus, the max depth is set to 28. Also, the model generalization error was larger when the minimum sample number of the leaf and branch node are large, indicating that the model itself is close to the optimal model complexity level. Therefore, a higher value was set given the large sample size in this case. The influence of the maximum selection feature number on CV score showed a trend of increasing first and then decreasing, so the maximum value of CV curve was set to 16.

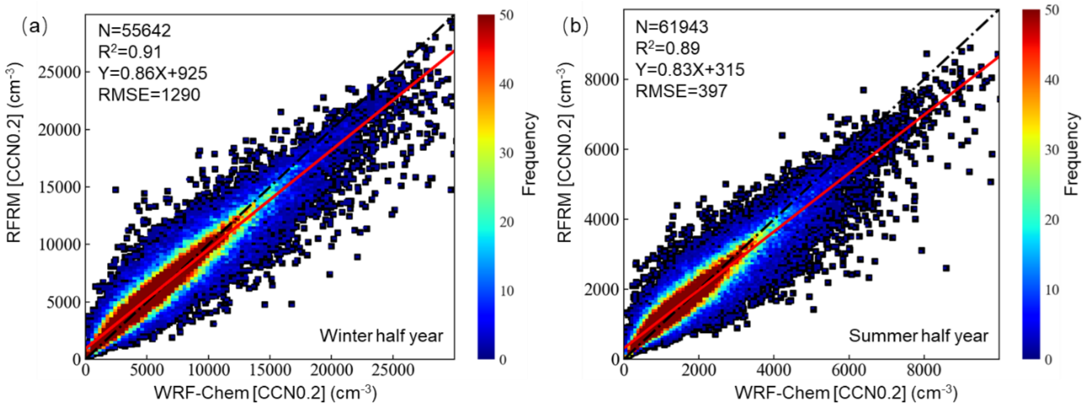

Figure 2. Comparison of RFRM retrieval and WRF-Chem simulated N_{CCN} at S=0.2%. (a and b) Density plots of retrieval N_{CCN} at S=0.2% as a function of the simulations from WRF-Chem on the testing dataset.

Fig. 2 shows the performance of our developed RFRM models by comparing the testing dataset (sample size: N=117585) of CCN simulated by WRF-Chem with the predicted CCN. Here the quality metrics for model performance are based on the correlation coefficient (R^2), root mean square error (RMSE) and the slope of the RFRM predicted and the WRF-Chem simulated CCN concentrations. It showed that the estimated N_{CCN} at S=0.2% are highly correlated with the values from WRF-Chem, with the correlation R^2 of ~0.89-0.91 and slopes of 0.83 and 0.86 in our developed RFRM model. This suggests that our model works well in estimating N_{CCN} with a high aerosol loading environment. We also found that the accuracy of the CCN prediction will deteriorate slightly if not including the information of chemical compositions (Fig. S2), or if using XGBoost algorithm (Fig. S3) when constructing the model.

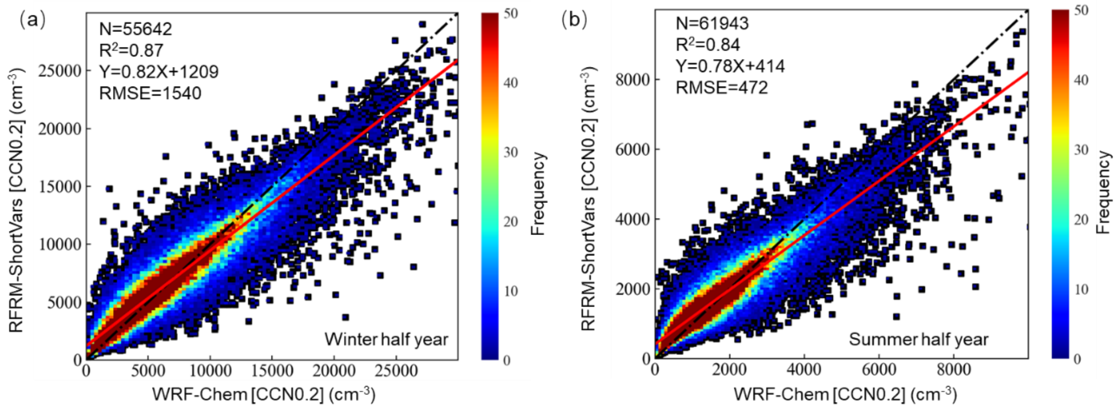

Figure S2. Comparison of RFRM-ShortVars model retrieval and WRF-Chem simulated NCCN at S=0.2%. (a and b) Density plots of retrieval NCCN at S=0.2% as a function of the simulations from WRF-Chem on the testing dataset.

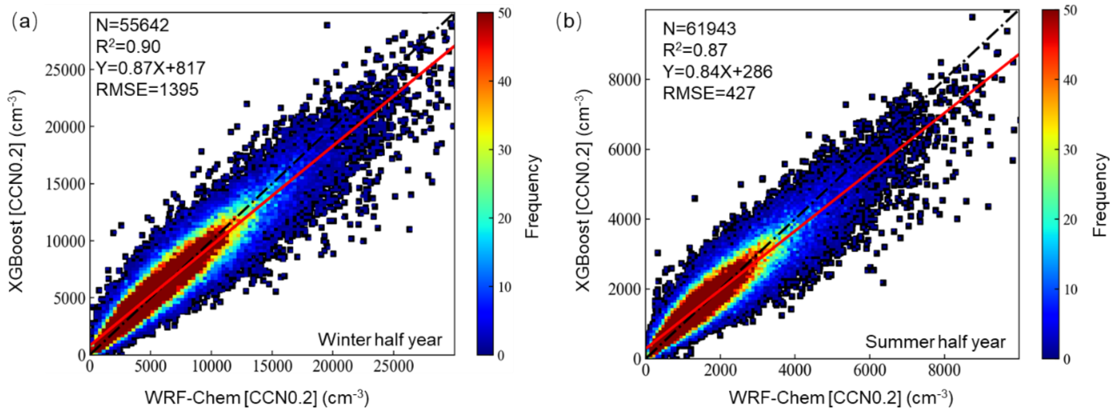

Figure S3. Same as Figure S2 but from XGBoost model.

Given that the ultimate goal of model development is to more accurately predict the actual atmospheric concentration of CCN, we further conducted a comparative analysis of the prediction results against the in-situ observation data in this region, spanning from the hourly scale to the interannual scale (see Figure 3-6), which will be presented and discussed in detail in Section 3.2 and 3.3 …"

References:

Morrison, H., Thompson, G., Tatarskii, V.: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one-and two-moment schemes, Monthly Weather Review, 137(3), 991–1007, https://doi.org/10.1175/2008MWR2556.1, 2009.

Zaveri, R. A., Peters, L. K.: A new lumped structure photochemical mechanism for large-scale applications, Journal of Geophysical Research: Atmospheres, 104, 30387–30415, https://doi.org/10.1029/1999JD900876, 1999.

Stockwell, W. R., Middleton, P., Chang, J. S., et al.: The second-generation regional acid deposition model chemical mechanism for regional air quality modeling, Journal of Geophysical Research: Atmospheres, 95(D10): 16343-16367, https://doi.org/10.1029/JD095iD10p16343, 1990.

Lin, Y., Farley, R., Orville, H.: Bulk parameterization of the snow field in a cloud model, Journal of Applied Meteorology and Climatology, 22(6): 1065-1092, https://doi.org/10.1175/1520-0450(1983)022<1065:BPOTSF>2.0.CO;2, 1983.

Lance, S., Nenes, A., Medina, J. et al.: Mapping the operation of the DMT continuous flow CCN counter, Aerosol Science and Technology, 40(4), 242–254, https://doi.org/10.1080/02786820500543290, 2006.

Liang, M., Tao, J., Ma, N. et al.: 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, 2022.

Liu, S., Geng, G., Xiao, Q., Zheng, Y., Liu, X., Cheng, J., & Zhang, Q.: Tracking daily concentrations of PM2.5 chemical composition in China since 2000, Environ Sci Technol, 56, 16517–16527, https://doi.org/10.1021/acs.est.2c06510, 2022.

Fan, X., Liu, J., Zhang, F. et al.: Contrasting size-resolved hygroscopicity of fine particles derived by HTDMA and HR-ToF-AMS measurements between summer and winter in Beijing: the impacts of aerosol aging and local emissions, Atmos. Chem. Phys., 20, 915–929, https://doi.org/10.5194/acp-20-915-2020, 2020.

Ren, J., Zhang, F., Wang, Y. et al.: Using different assumptions of aerosol mixing state and chemical composition to predict CCN concentrations based on field measurements in urban Beijing, Atmos. Chem. Phys., 18(9), 6907–6921, https://doi.org/10.5194/acp-18-6907-2018, 2018.

Yang, N., Shi, H., Tang, H. et al.: Geographical and temporal encoding for improving the estimation of $PM_{2.5}$ concentrations in China using end-to-end gradient boosting, Remote Sensing of Environment, 269, 112828, https://doi.org/10.1016/j.rse.2021.112828, 2022.

Zhang, Y., Tao, J., Ma, N. et al.: Predicting cloud condensation nuclei number concentration based on conventional measurements of aerosol properties in the North China Plain, Science of The Total Environment, 719, 137473, https://doi.org/10.1016/j.scitotenv.2020.137473, 2020.

Wei, J., Li, Z., Wang, J. et al.: Ground-level gaseous pollutants (NO_2, SO_2, and CO) in China: daily seamless mapping and spatiotemporal variations, Atmos. Chem. Phys., 23, 1511–1532, https://doi.org/10.5194/acp-23-1511-2023, 2023.

Zhang, F., Ren, J., Fan, T. et al.: Significantly enhanced aerosol CCN activity and number concentrations by nucleation-initiated haze events: A case study in urban Beijing, Journal of Geophysical Research: Atmospheres, 124(24), 14102–14113, https://doi.org/10.1029/2019JD031457, 2019.

Shen, Y., Virkkula, A., Ding, A. et al.: Estimating cloud condensation nuclei number concentrations using aerosol optical properties: role of particle number size distribution and parameterization, Atmos. Chem. Phys., 19(24), 15483–15502, https://doi.org/10.5194/acp-19-15483-2019, 2019.

Nair, A. A., Yu, F.: Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements, Atmos. Chem. Phys., 20(21), 12853–12869, https://doi.org/10.5194/acp-20-12853-2020, 2020.

Fanourgakis, G., Kanakidou, M., Nenes, A. et al.: Evaluation of global simulations of aerosol particle and cloud condensation nuclei number, with implications for cloud droplet formation, Atmos. Chem. Phys., 19(13), 8591–8617, https://doi.org/ 10.5194/acp-19-8591-2019, 2019.

Zheng, B., Tong, D., Li, M., Liu, F. et al.: Trends in China's anthropogenic emissions since 2010 as the consequence of clean air actions, Atmos. Chem. Phys., 18, 14095-14111, https://doi.org/10.5194/acp-18-14095-2018, 2018.

Shang, D., Tang, L., Fang, X. et al.: Variations in source contributions of particle number concentration under long-term emission control in winter of urban Beijing, Environmental Pollution, 304, 119072, https://doi.org/10.1016/j.envpol.2022.119072, 2022.