Supplement File

Machine learning significantly improves the simulation of hourly-to-yearly scale cloud nuclei concentration and radiative forcing in polluted atmosphere

Jingye Ren[1,2], Songjian Zou[3], Honghao Xu[3], Guiquan Liu[3], Zhe Wang[3], Anran Zhang[3], Chuanfeng Zhao[4], Min Hu[5], Dongjie Shang[5], Lizi Tang[5], Ru-Jin Huang[1], Yele Sun[6], Fang Zhang[3]*

[1]State Key Laboratory of Loess Science, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an, 710061, China

[2]Xi'an Institute for Innovative Earth Environment Research, Xi'an, 710061, China

[3]School of Civil and Environmental Engineering, Harbin Institute of Technology, Shenzhen, 518005, China

[4]Department of Atmospheric and Oceanic Sciences, School of Physics, Peking University, Beijing, 100871, China

[5]State Key Joint Laboratory of Environmental Simulation and Pollution Control, College of Environmental Sciences and Engineering, Peking University, Beijing, 100871, China

[6]State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, 100029, China

Corresponding author: Fang Zhang, *zhangfang2021@hit.edu.cn*

Data and Methods

1. Study area

In the present study, a region over North China Plain (NCP) with latitudes ranging from 32° to 40°N and longitudes ranging from 114° to 121°E is chosen as the study area for the following reasons. Firstly, being one of the most polluted areas in China, properties of aerosol particles are more complex and diverse, which further contribute to potentiality cloud concentration nuclei (CCN) concentrations with high uncertain. Especially in recent years, emissions of gas pollutants and fine particles have shown a significant downward trend year by year due to the implementation of the vigorous emission reduction. There is a pressing need to quantitatively understand the characteristics of CCN activity in the study area from the point of view of assessment of the climate effect of aerosols. Secondarily, more observation data covering several stations in NCP before would help to validate the predictability of random forest (RF) model. The campaigns covered three summer and three winter months, and involved years from 2014 to 2018. The spatial-temporal scale is diverse, and the pollution degree (eg., $PM_{2.5}$ mass concentration) is also different, which would be more conducive to verify the accuracy of the model.

2. WRF-Chem model

WRF-Chem version 4.1.5 is used to simulate N_{CCN} in this study (Grell et al., 2005), which nested a domain in 10 km×10 km covering the entire North China Plain (Fig. S1) and contained 181×170 grids. The simulation period is from January 1, 2014 to

December 31, 2018, and the simulation was conducted month by month with the time resolution of hourly data, with the first 48 hours as spin-up. The physical parameterizations employed in this study include Morrison 2-monment scheme (Morrison et al., 2009), RRTMG short-wave and long-wave radiation scheme (Iacono et al., 2008), Grell 3-D cumulus scheme (Grell et al., 2002), Yonsei University boundary layer scheme and Noah land surface model (Hong et al., 2006). The Carbon Bond Mechanism Z photochemical mechanism and 4-bin sectional Model for Simulating Aerosol Interactions and Chemistry (MOSAIC) aerosol model with aqueous chemistry are chosen in this study (Zaveri et al., 1999, 2008). 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 applied to simulate the long-term CCN concentration. The simulated N_{CCN} is as one of the input parameters of the machine learning model which has been trained and validated with field observational dataset obtained at the sites in this region.

The initial meteorological variables are provided from the National Center for Environmental Prediction's Final Operational Global (NCEP/FNL) datasets. The horizontal resolution of the reanalysis data was 0.25°×0.25°, and the temporal resolution was 6 h. The initial and boundary chemical conditions are from the Community Atmosphere Model with Chemistry model (Buchholz et al., 2019). The anthropogenic emissions are from Multi resolution Emission Inventory for China with a horizontal resolution of 0.25°×0.25° and a temporal resolution of monthly data (Zheng

et al., 2018), corresponding to the years 2014-2018. The biological and biomass emissions are taken from the Model of Emissions of Gases and Aerosols from Nature and the Fire Inventory from NCAR, respectively.

3. Auxiliary data

The auxiliary data used here includes the chemical compositions of $PM_{2.5}$, gaseous pollutants, and meteorological elements. Detailed information was listed in Table S1. Chemical components were adopted from Tsinghua University TAP (Tracking Air Pollution in China) dataset (Geng et al., 2021). Gas pollutants were collected from the China National Environmental Monitoring Centre network, including NO_2, SO_2, CO, O_3 and $PM_{2.5}$ observed up to 378 monitoring sites in NCP. The fifth generation European Centre for Medium-Range Weather Forecasts reanalysis (ERA-5) data is used for the meteorological inputs with the temporal resolution of 1 h and spatial resolution of 0.25°×0.25°, including temperature, relative humidity (RH), total precipitation (TP), wind speed (WS), wind direction (WD), boundary layer height (BL), surface pressure (SP) and surface net solar radiation (SNSR).

Cartesian coordinates are added to the characteristic parameters due to the information of space-time scale involved. Time information is usually expressed as the day of the year (DOY). This representation does not reflect seasonality because the first and last days of the year are in winter, but DOY values vary widely. To solve this problem, convert the days of the year DOY into a Cartesian coordinate system:

$$t_x = \cos\left(2\pi \frac{DOY}{T}\right),\ t_y = \sin\left(2\pi \frac{DOY}{T}\right); \quad (1)$$

here t_x, t_y are Cartesian coordinates and T is the total number of days in a year. Based on the equation, DOY can be normalized to the interval [0 2π] and further converted to polar coordinates. Similar methods have also been widely used to process time information in machine learning model construction (Yang et al., 2022; Wei et al., 2023). Spatial information is generally given in polar coordinates of latitude and longitude, which is not applicable to the representation of distance due to uneven changes in polar coordinates (Gates et al., 1962). The Cartesian coordinate system can realize uniform changes in values, so the longitude and latitude information of the state control station is converted to Cartesian coordinate system as the characteristic parameter input model. The specific transformation formula is as follows:

$$g_x = Rsin\theta cos\varphi,\ g_y = Rsin\varphi,\ g_z = Rcos\theta cos\varphi; \quad (2)$$

where φ is the latitude, θ is the longitude, and R is the radius of the earth.

To construct the training test, we defined spatiotemporal match criteria that match the auxiliary variables with simulated CCN values into daily resolution and resampled into monitoring sites for consistency.

4. Random forest regression model

In this study, the random forest regression model (RFRM) is trained to derive N_{CCN} at typical atmospheric supersaturations 0.2% and 0.4%. It has been proved to understand the dependence of the outputs on its predictors well, which is due to the randomly in the set of decision trees as well as in the subset of training data for decision trees (Liang et al., 2022). Detailed feature variables and target parameters are described below. Eighteen variables of atmospheric state and chemical composition are selected

as the input features, including mass concentrations of particle matter smaller than 2.5 μm, organic matter, sulfate, nitrate, ammonium, black carbon, 2 m temperature, relative humidity, wind speed, wind direction, boundary layer, surface pressure, total precipitation, surface net solar radiation, and concentrations of NO_2, SO_2, CO, O_3. The spatiotemporal information is provided to the training model based on the cartesian coordinates (Yang et al., 2022). The target variable is provided by the simulation from WRF-Chem (Section 2). All the aforementioned species data (feature and target variables) were processed to match the same spatiotemporal resolution, and then split into 7:3 ratio for model training and testing.

Based on the previous studies reported (Nair et al., 2020), the optimization parameters of RF model were preliminary examination by varying hyperparameters (Fig. S2). The general 10-fold cross-validation is adopted (Wei et al., 2023). Cross-validation (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 selection parameter of 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 is shown as that with the increase of depth, the accuracy of the model moves to the right, that is, the complexity of the model increases. Thus, the value of the max depth here is set to 28. The larger the minimum sample number of the leaf and the minimum sample number of the branch node, the larger the model generalization error, indicating that the model itself is close to the optimal model complexity level. Here a higher value was set based on the default value given the large

sample size. The influence of the selection 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.

In order to assess the predictability for N_{CCN} based on machine learning technique, six field campaigns of N_{CCN} was used to compare with the outputs from retrieval models. Further details about in situ observations are in Sect. 5. The quality metrics for model performance are based on the correlation coefficient (R^2), root mean square error (RMSE) and the slope of the predicted and the observed CCN concentrations. Based on the features and parameters selected above, CCN concentration retrieval model is constructed.

5. Ground-level CCN measurements

Six campaigns were used to validate the predictability accuracy, from November 8 to November 30, 2014, August 20 to October 6, 2015, November 16 to December 20, 2016, May 28 to May 27, 2017 at Beijing (BJ) site, from May 17 to June 14 at Xingtai (XT) site, from January 23 to February 3, 2018 at Gucheng (GC) site. The BJ site (Lon, 116.37° E; Lat, 39.97°N) is located at the meteorological tower station of Institute of Atmospheric Physics, Chinese Academy of Sciences. It can represent the general emission situation in urban areas of the northern part of the NCP. The main pollution sources are the surrounding traffic and residential emissions here. The observation at XT site (Lon, 114.37° E; Lat, 37.18°N) is located at a national weather station. It is mainly affected by the emission of surrounding towns and factories (such as coal power, coking, steel, cement, chemical industry, etc.), thus can reflect the polluted suburban

158 conditions of the southern part of the NCP. The GC site (Lon, 115.74° E, Lat, 39.15° N)
159 is located at the integrated ecological-meteorological observation and experiment
160 station of the Chinese Academy of Meteorological Sciences. This site is mainly
161 surrounded by the nearby villages, farmland, and transportation network, which can
162 reflect the regional background pollution of the northern part of the NCP. Details about
163 the observations could be found in Fan et al. (2020), Ren et al. (2018), and Zhang et al.
164 (2019). Overall, observations at these sites could represent the average polluted and
165 background conditions in the NCP.

166 The CCN number concentrations were measured by using the Droplet
167 Measurement Technologies CCN counter (model CCNC-100, DMT Inc. Lance et al.,
168 2006) at BJ and XT site. The supersaturation (S) levels set for each CCN measurement
169 cycle were 0.07%, 0.1%, 0.2%, 0.4%, and 0.8%, respectively. Another measurement at
170 GC site was referred from Zhang et al. (2020). In this study, the comparisons between
171 the measured and predicted N_{CCN} were mostly based on the value at S=0.2% and
172 S=0.4%.

173 **6. Evaluation of the aerosol indirect effects**

174 Atmospheric particles acting as cloud condensation nuclei could alter the cloud-
175 rain interactions and further have impact on the climate change. The methods for
176 evaluating the cloud and radiative properties caused by the deviations of CCN
177 concentrations output from WRF-Chem simulation and random forest technique was
178 refereed from Wang et al. 2019 and Wang et al. 2008.

As for the first indirect effect, the cloud optical thickness (τ) and single scatter albedo coefficient (ω_0) are the crucial estimators and can be calculated as:

$$\tau \approx \frac{3}{2} W r_e^{-1} \quad (3)$$

$$1 - \omega_0 = 1.7 k_w r_e \quad (4)$$

where W means the liquid water path, k_w is the complex part of the refractive index of water, r_e is the effective radius of cloud droplets and can be expressed by:

$$r_e = \beta \left(\frac{3LWC}{4\pi \rho_w N_C}\right)^{1/3} \quad (5)$$

where LWC is the liquid water content in cloud, β is the scaling factor. Here, we mainly focused on the response of cloud droplet effective radius to the change of cloud droplet number concentration N_c, so β is assumed to be constant (Quaas et al., 2004). Cloud number concentration can be calculated based on the N_{CCN}-SS parameterizations in many climate models (Fan et al., 2012). Note that the power relationship between the r_e and N_c referred from the previous observations is only an approximate relationship (Garrett et al., 2004).

For evaluating second aerosol indirect forcing, it is critical to estimate the relevant variables associated with the cloud-rain process. The parameterizations of the cloud-to-rain conversion process can be expressed by conversion threshold function (TA):

$$TA = \left[\frac{\int_{r_c}^{\infty} r^6 n(r) dr}{\int_0^{\infty} r^6 n(r) dr}\right] \left[\frac{\int_{r_c}^{\infty} r^3 n(r) dr}{\int_0^{\infty} r^3 n(r) dr}\right] \quad (6)$$

where r is the cloud droplet radius, $n(r)$ is the size distribution of the cloud droplet (here refereed from Wang et al. 2019 to evaluate). According to the Liu et al. (2004), the critical radius of the auto-conversion process r_c can be expressed by

$$r_c \approx 4.09 \times 10^{-4} \beta_{con}^{1/6} \frac{{N_C}^{1/6}}{LWC^{1/3}} \quad (7)$$

where β_{con} is a fixed parameter of 1.15×10^{23} s^{-1} (Liu et al., 2004). Here, the focus is mainly on the effect of the ΔN_{CCN} between WRF-Chem and RF model on the aerosol indirect effects in our research.

We also evaluate the perturbation of cloud radiative forcing due to the uncertainty in CCN concentration estimated from machine learning (RFRM) and numerical model (WRF-Chem). Here under the assumption of the fixed liquid water volume, an approximate analytical expression for estimating the cloud radiative of the change in N_{CCN} as follows (Charlson et al, 1992; Wang et al., 2008):

$$\Delta F_C \approx -\frac{F_T}{4} A_{mst} T_r^2 \Delta R_C \approx -2.2 \times \Delta \ln N_{CCN} \quad (8)$$

$$\Delta R_C = \frac{1}{3}[R_C(1-R_C)] \times \Delta ln N_d \approx 0.075 \times \Delta ln N_d \quad (9)$$

$$\Delta ln N_d \approx 0.5 \times \Delta \ln N_{CCN} \quad (10)$$

where ΔF_c is the uncertainty in cloud radiative forcing, F_T is the solar constant, A_{mst} is the cloud fractional coverage, T_r is the fractional transmission of shortwave radiation above the cloud layer, ΔR_c is the uncertainty in cloud albedo, N_d is the cloud droplet number concentration, N_{CCN} is the cloud condensation nuclei number concentration. In equation (8), A_{mst} and T_r were set as 0.3 and 0.76, respectively (Charlson et al, 1987; Schwartz et al., 1996), which represented an average fraction of nonoverlapped marine stratus and stratocumulus. Here the uncertainty in ΔR_c in equation (9) for the typical range of cloud albedo was calculated according to the Wang et al. 2008. The uncertainty N_d associated with change in N_{CCN} was referred from Sotriopoulou et al. (2006) with a value of 0.5 in equation (10).

References

Buchholz, R. R., Emmons, L. K., Tilmes, S., & The CESM2 Development Team.: CESM2.1/CAM-chem Instantaneous Output for Boundary Conditions. UCAR/NCAR - Atmospheric Chemistry Observations and Modeling Laboratory. e.g. Lat: -10 to 10, Lon: 100 to 150, September 2015 - February 2016, 2019.

Charlson, R. et al.: Climate Forcing by Anthropogenic Aerosols, Science, 255, 423–430, https://doi.org/10.1126/science.255.5043.423, 1992.

Charlson, R. et al.: Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate, Nature, 326(6114), 655–661, https://doi.org/10.1038/326655a0, 1987.

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.

Fan, J., Leung, L. R., Li, Z. et al.: Aerosol impacts on clouds and precipitation in eastern China: Results from bin and bulk microphysics, Journal of Geophysical Research, 117, D00K36, https://doi.org/10.1029/2011JD016537, 2012.

Garrett, T. J., Zhao, C., Dong, X., Mace, G.G., Hobbs, P.V.: Effects of varying aerosol regimes on low-level arctic stratus, Geophysical Research Letters, 31, https://doi.org/10.1029/2004GL019928, 2004.

Grell, G. A., Peckham, S. E., Schmitz, R. et al.: Fully coupled "online" chemistry within the WRF model, Atmos. Environ. 39(37), 6957–6975, https://doi.org/10.1016/j.atmosenv.2005.04.027, 2005.

Grell, G. A., Dezső, D.: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques, Geophysical Research Letters, 29, 1693, https://doi.org/10.1029/2002GL015311, 2002.

Geng, G., Xiao, Q., Liu, S. et al.: Tracking air pollution in China: near real-time PM2.5 retrievals from multisource data fusion, Environmental Science & Technology, 55(17), 12106–12115, https://doi.org/10.1021/acs.est.1c01863, 2021.

Gates, W., Riegel, C.: A study of numerical errors in the integration of barotropic flow on a spherical grid, Journal of geophysical research, 67(2), 773–784, 1962.

Hong, S. Y., Noh, Y., Dudhia, J.: A new vertical diffusion package with an explicit treatment of entrainment processes, Mon. Weather Rev, 134(9), 2318, https://doi.org/10.1175/MWR3199.1, 2006.

Iacono, M. J., Delamere, J. S., Mlawer, E. J. et al.: Radiative forcing by long-lived greenhouse gases: calculations with the AER radiative transfer models, Journal of Geophysical Research: Atmospheres, 113 (D13), https://doi.org/10.1029/2008JD009944, 2008.

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.

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.

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.

Liu, Y., Daum, P. H., & McGraw, R.: An analytical expression for predicting the critical radius in the autoconversion parameterization, Geophysical Research Letters, 31, L06121, https://doi.org/10.1029/2003GL019117, 2004.

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.

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.

Quaas, J., Boucher, O., & Breon, F. M.: Aerosol indirect effects in POLDER satellite data and the Laboratoire de Meteorologie Dynamique Zoom (LMDZ) general circulation model, Journal of Geophysical Research, 109, D08205, https://doi.org/10.1029/2003JD004317, 2004.

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.

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.

Schwartz, S. E. and Slingo, A.: Enhanced shortwave cloud radiative forcing due to anthropogenic aerosols., in: Clouds, Chemistry, and Climate., edited by: Crutzen, P., and Ramanathan, V., Springer, Heidelberg. 191–236, 1996.

Sotiropoulou, R. E. P., Medina, J., and Nenes, A.: CCN predictions: Is theory sufficient for assessments of the indirect effect? Geophysical Research Letters, 33, L05816, https://doi.org/10.1029/2005GL025148, 2006.

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.

Wang, Y., Niu, S., Lv, J. et al.: A new method for distinguishing unactivated particles in cloud condensation nuclei measurements: implications for aerosol indirect effect evaluation, Geophysical Research Letters, 46, 14,185–14,194, https://doi.org/10.1029/2019GL085379, 2019.

Wang, J., Lee, Y.-N., Daum, P. H., Jayne, J., and Alexander, M. L.: Effects of aerosol organics on cloud condensation nucleus (CCN) concentration and first indirect aerosol effect, Atmos. Chem. Phys., 8, 6325–6339, https://doi.org/10.5194/acp-8-6325-2008, 2008.

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.

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.

Zaveri, R. A., Easter, R. C., Fast, J. D. et al.: Model for simulating aerosol interactions and chemistry (MOSAIC), Journal of Geophysical Research: Atmospheres, 113(D13), https://doi.org/10.1029/2007JD008782, 2008.

Zheng, B., Tong, D., Li, M. 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.

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.

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.

328 **Table S1.** Summary of the datasets used in this study from multiple sources

Data category	Content (Abbreviation)	Unit	Spatial resolution	Temporal resolution	Data source
N_{CCN}	CCN concentration (N_{CCN}) at S=0.20% and S=0.40%	cm^{-3}	10 km	Hourly	WRF-Chem model
Chemical compositions	Organic matter (OM) Sulfate (SO_4^{2-}) Nitrate (NO_3^-) Ammonium ($NH4^+$) Black carbon (BC)	$\mu g\ m^{-3}$ $\mu g\ m^{-3}$ $\mu g\ m^{-3}$ $\mu g\ m^{-3}$ $\mu g\ m^{-3}$	10 km - - - -	Daily - - - -	TAP (http://tapdata.org.cn/?page_id=59&item=pm25)
Meteorological data	2m temperature (TEM) Relative humidity (RH) Wind speed (WS) Wind direction (WD) Boundary layer (BL) Surface pressure (SP) Total precipitation (TP) Surface net solar radiation (SNSR)	K % $m\ s^{-1}$ Degree M hPa mm $W\ m^{-2}$	$0.25^o \times 0.25^o$ - - - - - - -	Hourly - - - - - - -	ERA-5 (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form)
Gas pollutants	$PM_{2.5}$ NO_2 SO_2 CO O_3	$\mu g\ m^{-3}$ - - - -	- - - - -	Hourly - - - -	CEMC (https://quotsoft.net/air/)

329

Figures

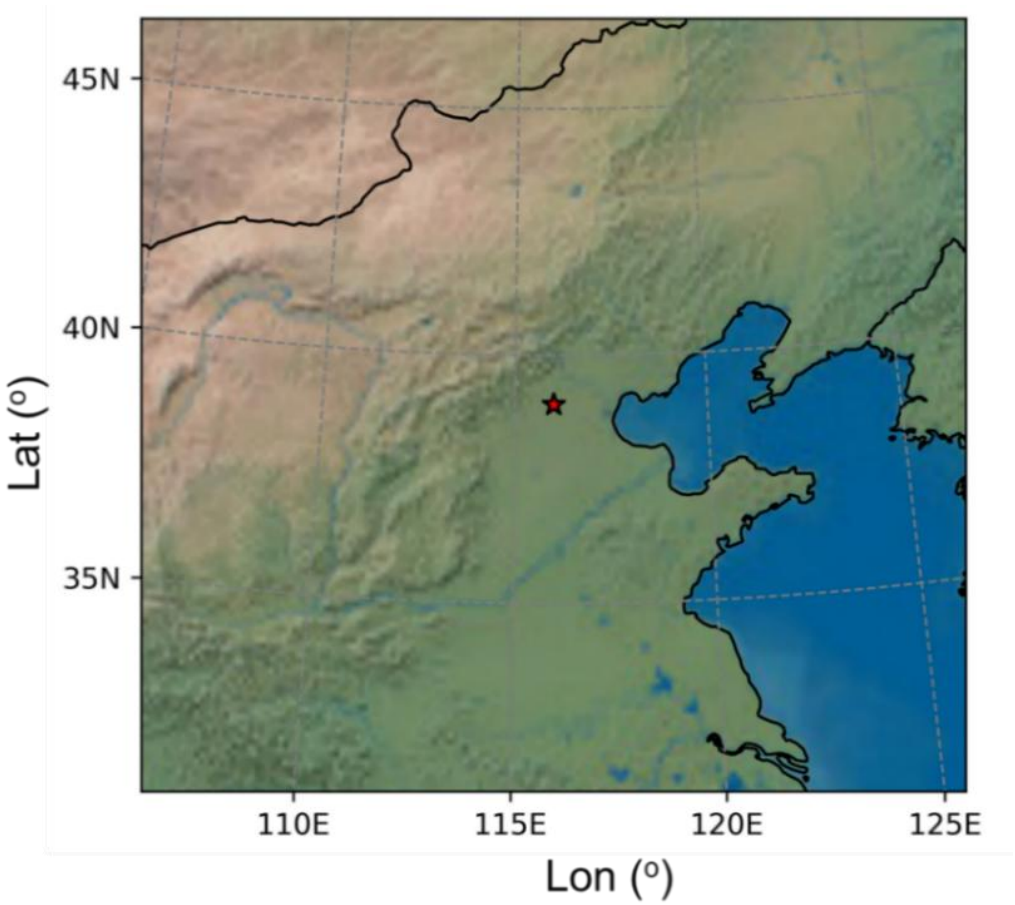

Figure S1. WRF-Chem model simulation domain diagram drawn by using the Cartopy library. The center point (red star) of the simulation domain is located at 39 ° N and 116 ° E, with a grid spacing of 10 km. The number of grids in the east-west and north-south directions is 181 and 170, respectively.

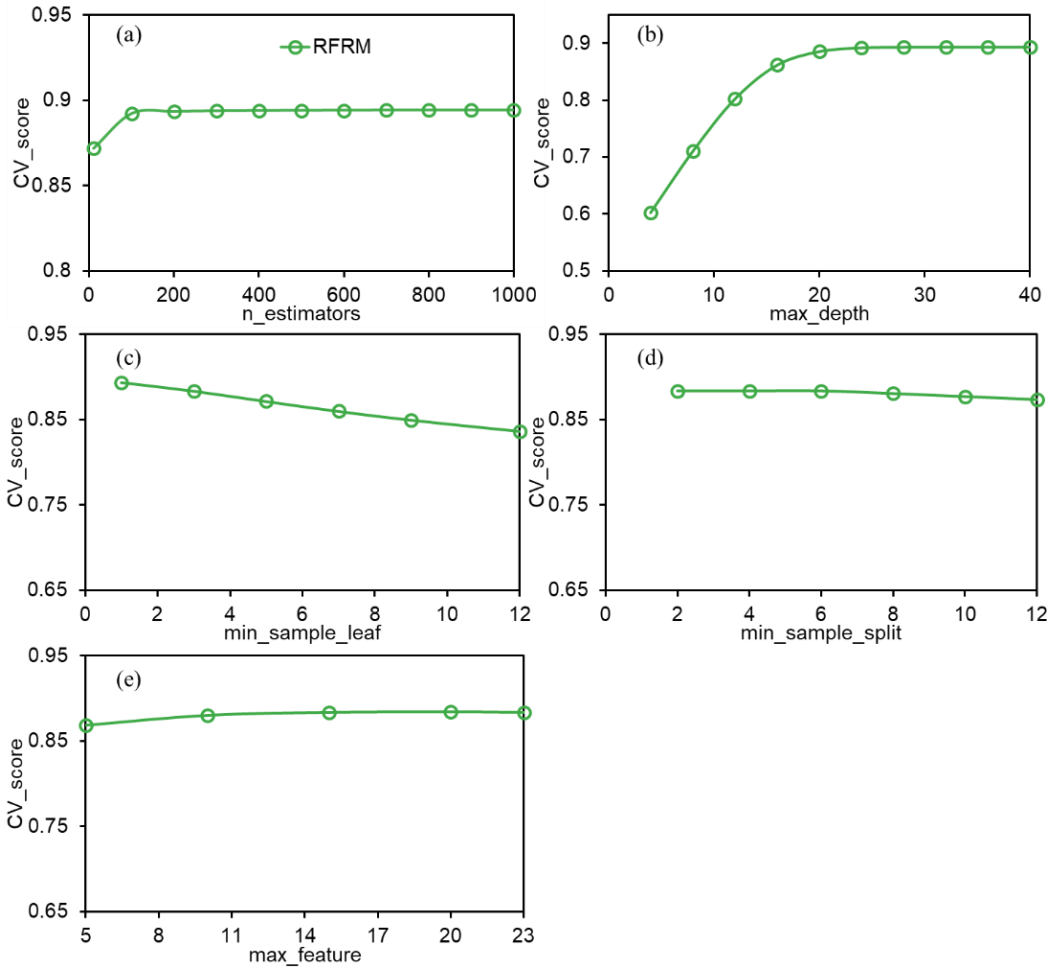

Figure S2. RFRM model evaluation through cross-validation (CV) score with varying hyperparameters.

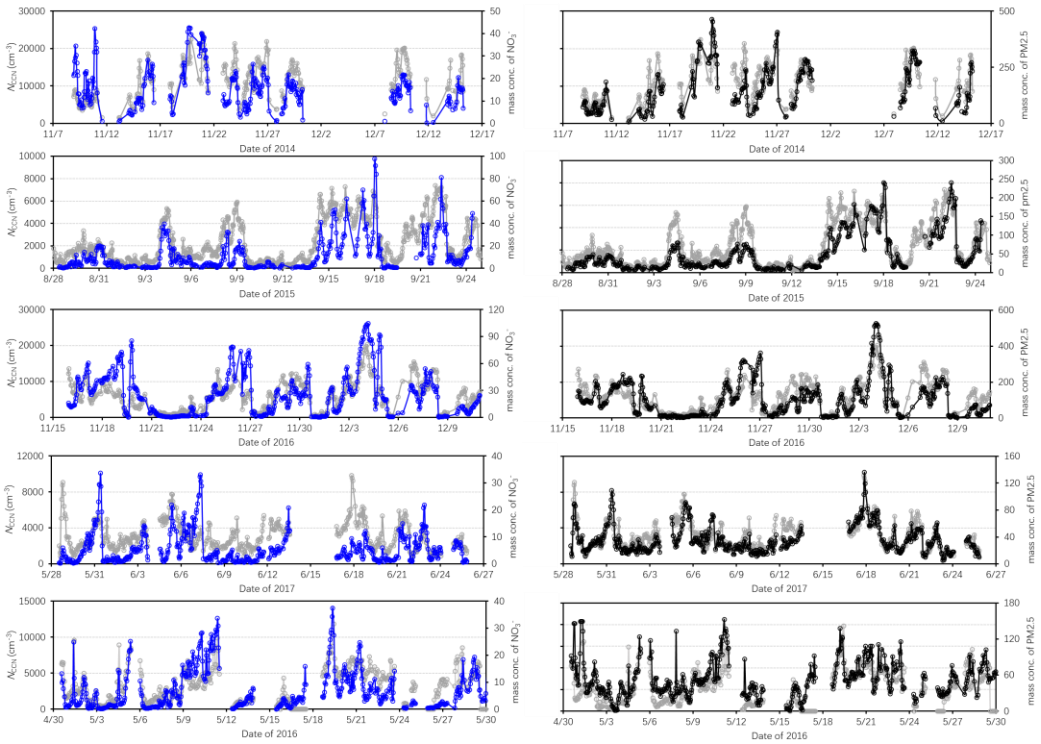

Figure S3. Time series of CCN number concentration (grey line) and mass concentration of NO_3^- (blue line), mass concentration of $PM_{2.5}$ (black line) during the field campaign in North China Plain.

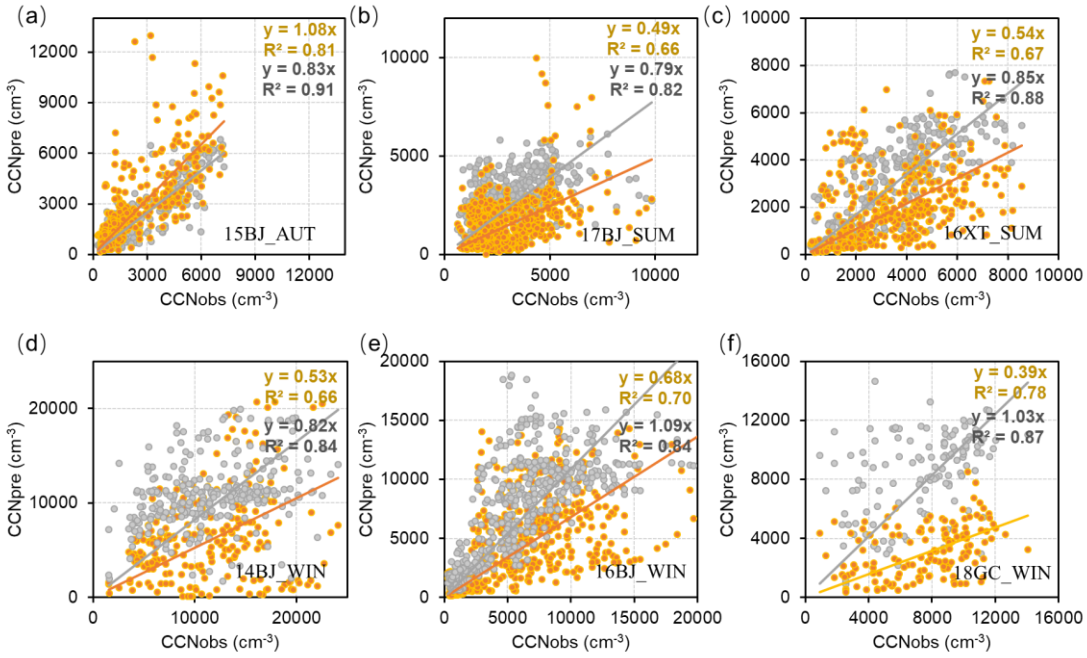

Figure S4. Scatter plots of the observed N_{CCN} with the RFRM-predicted and WRF-Chem-simulated for each campaign at S=0.2% (a-f).

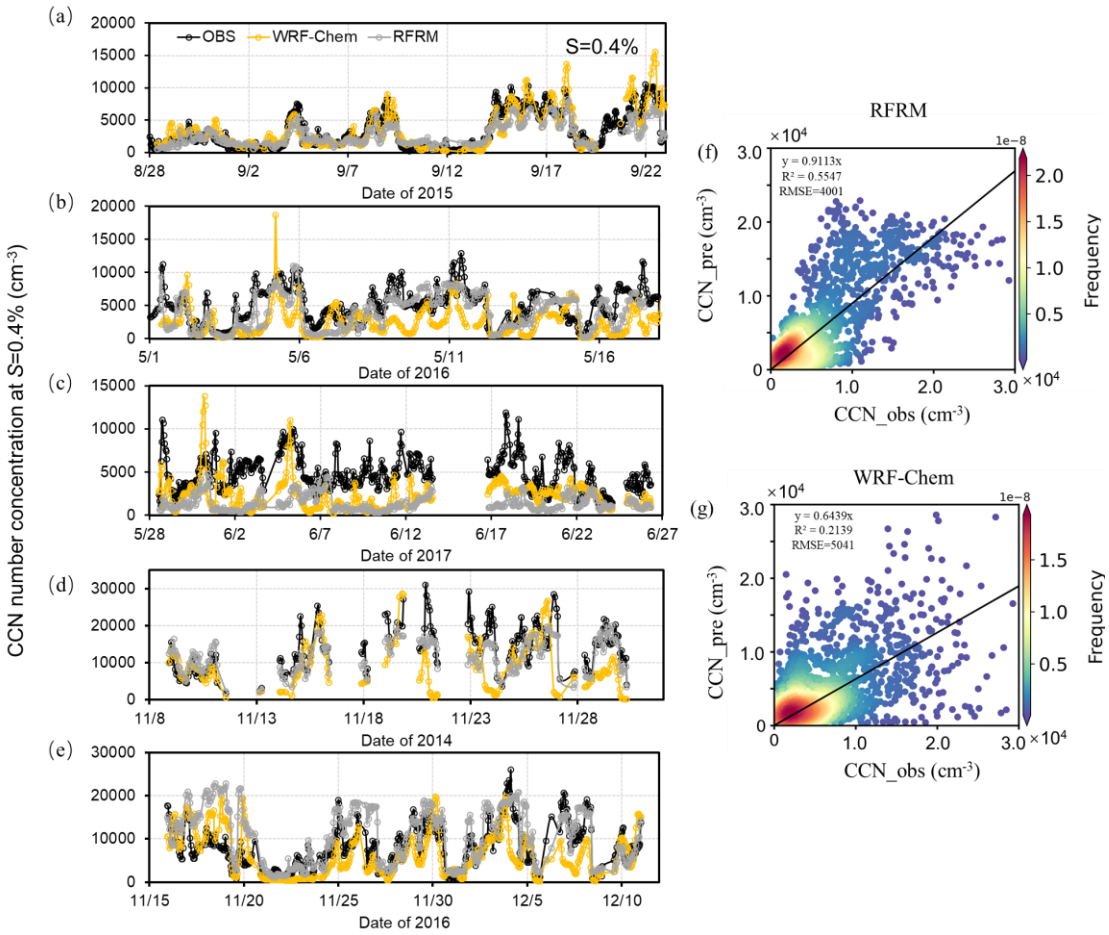

Figure S5. Time series of the comparison for each campaign in the North China Plain, (a) in 2015 summer at Beijing site (BJ2015_AUT), (b) in 2016 summer at Xingtai site (XY2016_SUM), (c) in 2017 summer at BJ site (BJ2017_SUM), (d) in 2014 winter at BJ site (BJ2014_WIN), (e) in 2016 winter at BJ site (BJ2014_WIN). Scatter plots of the observed N_{CCN} at S=0.4% with the RFRM-predicted (f), WRF-Chem-simulated (g) (yellow line: simulations from WRF-Chem, grey line: predictions from random forest model, and black line: the observed N_{CCN}).

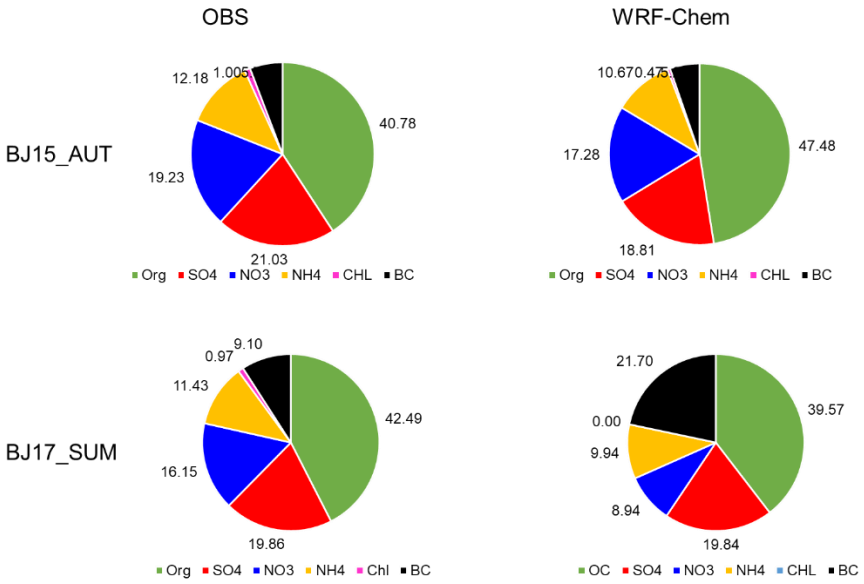

Figure S6. Pie charts of the mass fraction of the chemical composition between the observation and simulation from WRF-Chem, taken the BJ15_AUT and BJ17_SUM as the examples.

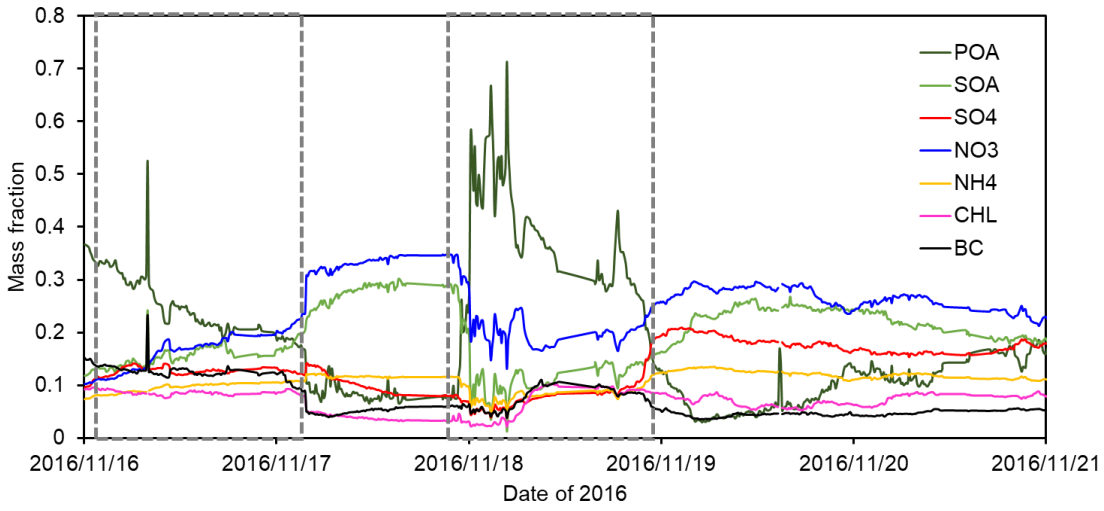

Figure S7. Timeseries of chemical composition during the case in the year of 2016 at Beijing site.

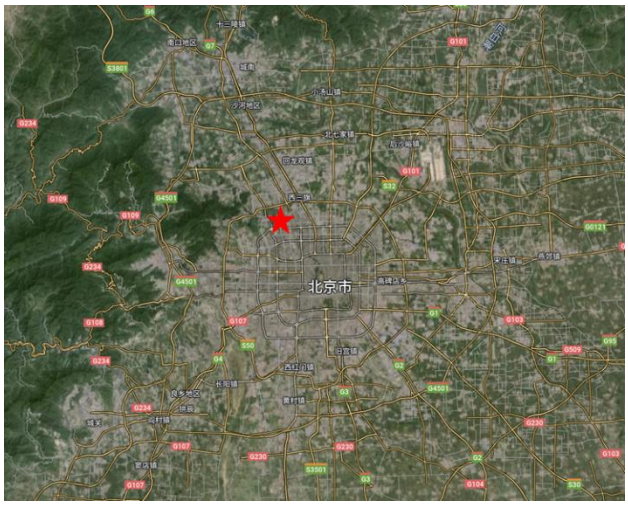

Figure S8. Map location of the sampling site (red star represents the Peking University). (© Google Earth https://maps.google.com/, last access: 8 September 2024).

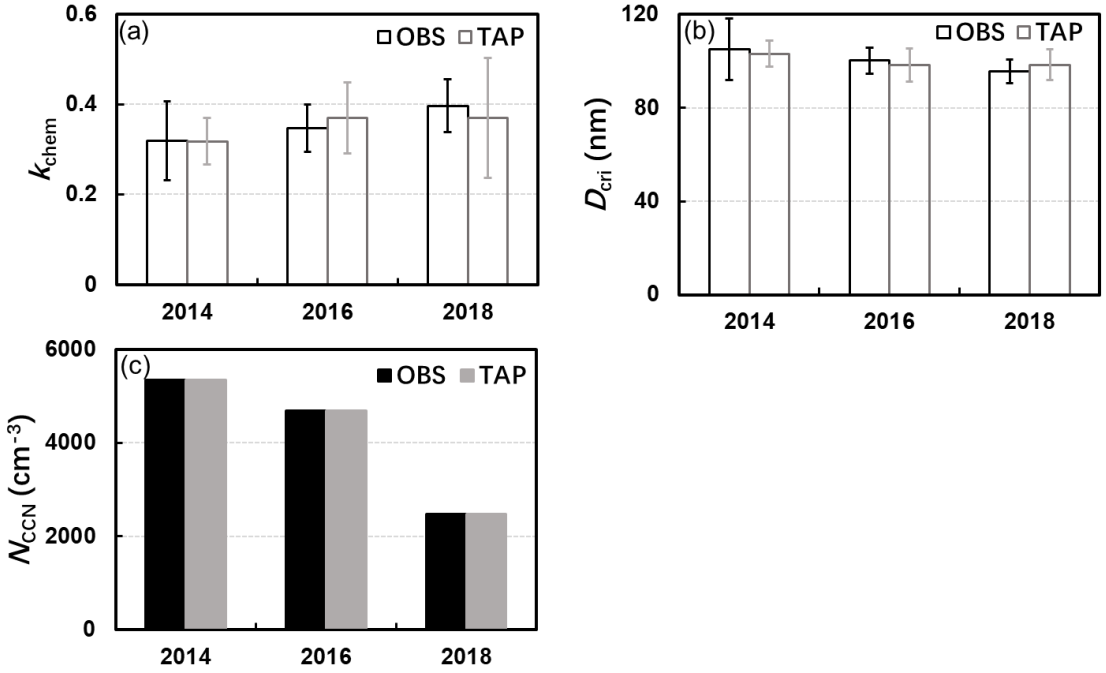

Figure S9. Average annual value of the k_{chem} calculated from chemical composition (a), the critical diameter at S=0.2% (b), N_{CCN} at S=0.2% (c) between the observed and TAP dataset in the winter of 2014, 2016, 2018.

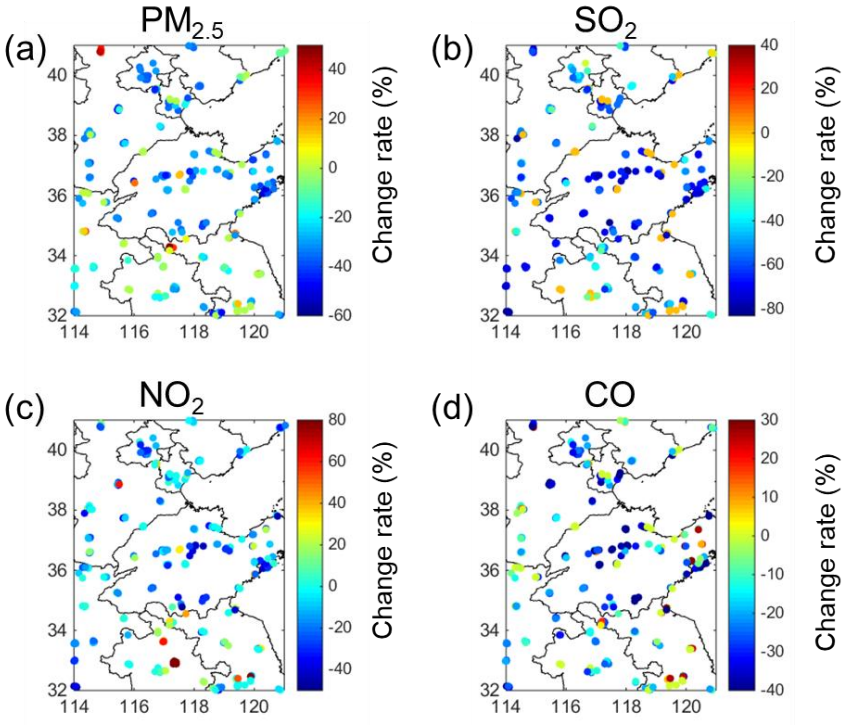

Figure S10. Spatial pattern of the change rates of the $PM_{2.5}$, SO_2, NO_2 and CO from 2014 to 2018 in NCP.

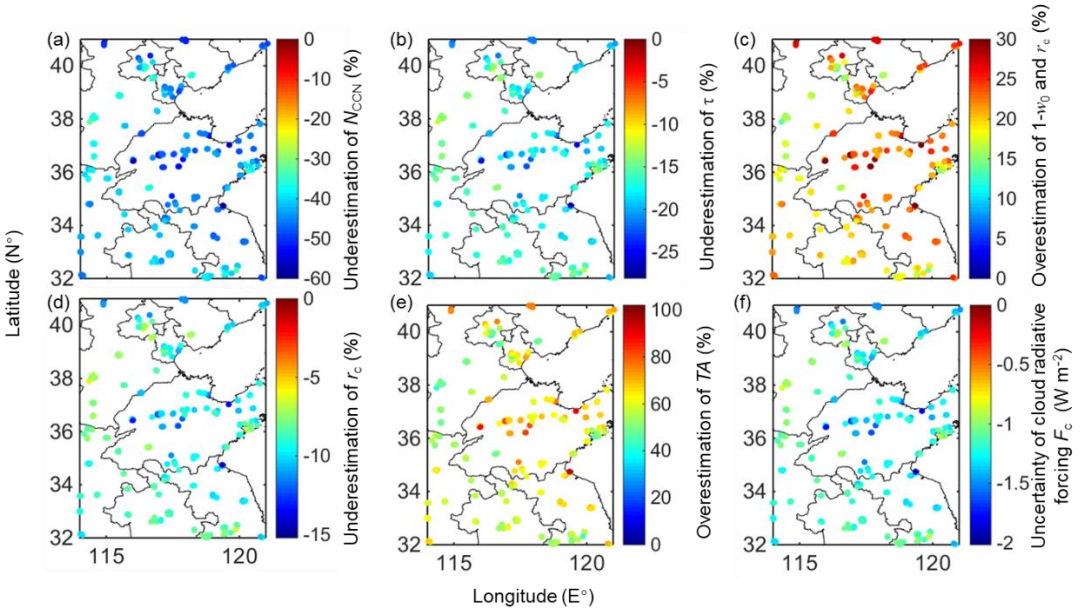

Figure S11. Spatial pattern of the differences of CCN number concentration between the new model and WRF-Chem (a), effects of the discrepancy from CCN number concentration on cloud optical thickness (τ) (b), on absorption coefficient ($1-\omega_0$) and effective radius (r_e) (c), on critical radius (r_c) (d), the cloud-to-rain conversion threshold function (TA) (e), cloud radiative forcing (F_c) (f).

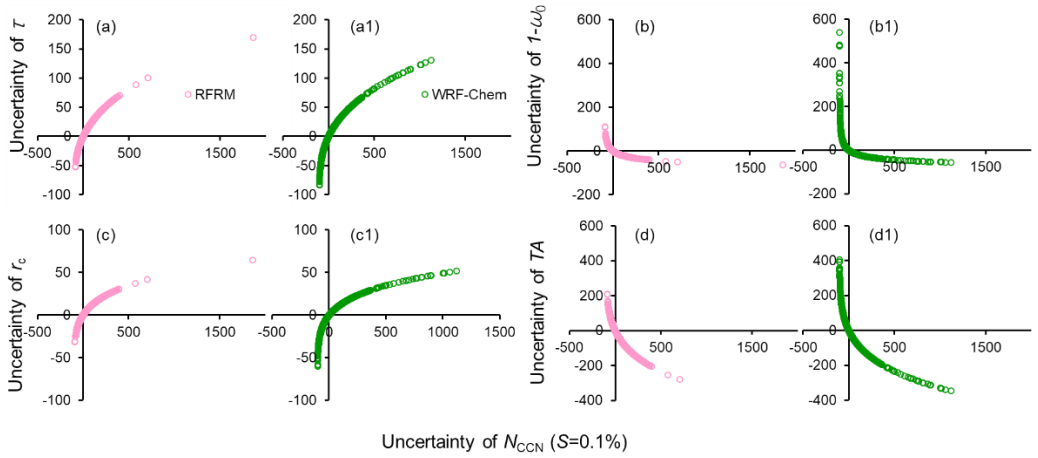

Figure S12. Sensitivity of the cloud properties on the CCN concentration at S=0.1%, uncertainty of the CCN number concentration on cloud optical thickness (τ) between the observed with the new model (a) and with the WRF-Chem model (a1), on absorption coefficient (1-ω_0) and effective radius (r_e) (b and b1), on critical radius (r_c) (c and c1), the cloud-to-rain conversion threshold function (TA) (d and d1).

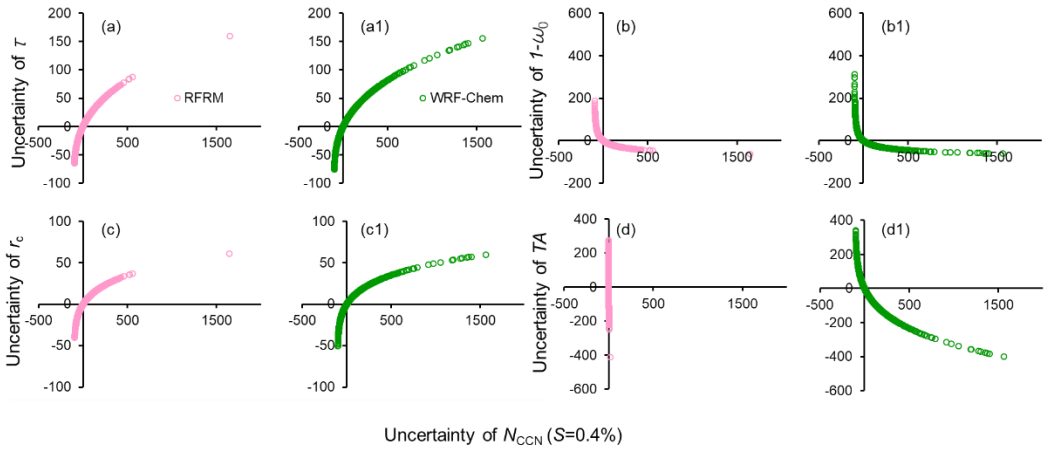

Figure S13. Same as Supplementary Figure 12 but at S=0.4%.