https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements

Arshad Arjunan Nair and Fangqun Yu

Atmospheric Sciences Research Center, State University of New York, Albany, New York 12203, USA

Correspondence: Arshad Arjunan Nair (aanair@albany.edu)

Abstract. Cloud condensation nuclei (CCN) number concentrations are an important aspect of aerosol–cloud interactions and the subsequent climate effects; however, their measurements are very limited. We use a machine learning tool, random decision forests, to develop a Random Forest Regression Model (RFRM) to derive CCN at 0.4% supersaturation ([CCN0.4]) from commonly available measurements. The RFRM is trained on the long-term simulations in a global size-resolved particle microphysics model. Using atmospheric state and composition variables as predictors, through associations of their variabilities, the RFRM is able to learn the underlying dependence of [CCN0.4] on these predictors, which are: 8 fractions of PM2.5 (NH_4, SO_4, NO_3, secondary organic aerosol (SOA), black carbon (BC), primary organic carbon (POC), dust, and salt), 7 gaseous species (NO_x, NH_3, O_3, SO_2, OH, isoprene, and monoterpene), and 4 meteorological variables (temperature (T), relative humidity (RH), precipitation, and solar radiation). The RFRM is highly robust: median mean fractional bias (MFB) of 4.4% with $\sim 96.33\%$ of the derived [CCN0.4] within a good agreement range of $-60\% <$ MFB $< +60\%$ and strong correlation of Kendall's τ coefficient ≈ 0.88. The RFRM demonstrates its robustness over 4 orders of magnitude of [CCN0.4] over varying spatial (such as continental to oceanic, clean to polluted, and near surface to upper troposphere) and temporal (from the hourly to the decadal) scales. At the Atmospheric Radiation Measurement Southern Great Plains observatory (ARM SGP) in Lamont, Oklahoma, United States, long-term measurements for PM2.5 speciation (NH_4, SO_4, NO_3, and organic carbon (OC)), NO_x, O_3, SO_2, T, and RH, as well as [CCN0.4] are available. We modify, optimise, and retrain the developed RFRM to make predictions from $19 \rightarrow 9$ of these available predictors. This retrained RFRM (RFRM-ShortVars) shows a reduction in performance due to the unavailability and sparsity of measurements (predictors); it captures the [CCN0.4] variability and magnitude at SGP with $\sim 67.02\%$ of the derived values in the good agreement range. This work shows the potential of using the more commonly available measurements of PM2.5 speciation to alleviate the sparsity of CCN number concentrations' measurements.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

1 Introduction

Minute particles suspended in the atmosphere prove to be the most non-trivial sources of uncertainty (variability across modeling efforts) in our understanding of climate change (IPCC AR5, 2013). These particles or aerosols, or rather CCN (cloud condensation nuclei: aerosols capable of being imbibed in clouds and modifying their properties) have direct and indirect sources. They can directly get into the atmosphere as sea salt, primary inorganic particulates such as dust and carbon, or primary organic particulates. Aqueous chemistry can modify the chemical species in the cloud droplet, which on evaporation will result in an aerosol size distribution capable of acting as CCN at lower RH (Hoppel et al., 1994). In the air, emissions of SO_2/DMS, NO_x, and organics can undergo gas-phase (photo-)chemistry to form condensible vapors that may take part in new particle formation (nucleation) that converts gas to particle; this process is an important source of aerosols, especially from anthropogenic sources, and subsequent growth contributes up to 50% or more of global CCN (e.g., Merikanto et al., 2009; Yu and Luo, 2009).

Of the uncertainty in the effective radiative forcing (ERF) in global climate models associated with aerosols, the aerosol indirect effect primarily through aerosol–cloud interactions (*aci*) is predominant (IPCC AR5, 2013): $\mathrm{ERF}_{aci} = -1.2$ to 0 Wm^{-2}. These aerosol–cloud interactions are through CCN that affect cloud micro- and macro-physics primarily through their interaction with water vapor to modify cloud droplet size and number. There are various such indirect effects: First Indirect Effect (Twomey, 1977), Second Indirect Effect (Albrecht, 1989), and others such as effects on cloud formation and precipitation dynamics (Seinfeld et al., 2016, and the references therein), which affect the atmospheric energy balance.

The numerous physical and chemical effects of and on CCN and their non-linear interactions as detailed above provide a glimpse into the complexities and challenges associated with developing a valid physical description of aerosol processes in global climate models. A major problem is the accurate characterisation of CCN number concentrations ([CCN]) in the atmosphere and quantification of their effects on Earth's radiative budget. Extensive measurements of [CCN] would help in this regard, towards reducing the uncertainties in modeling aerosol–cloud interactions. Unfortunately, these are sparse; there are some in situ measurements available during short campaigns and for a few sites from networks such as Atmospheric Radiation Measurement Climate Research Facility (ARM), Aerosol, Clouds and Trace Gases Research Infrastructure (ACTRIS), and Global Atmosphere Watch (GAW), while satellite inference of [CCN] is not yet robust, suffering from missing data and coarse resolution. In contrast, particle mass concentration and speciation have been routinely measured in a large number of networks, such as the Interagency Monitoring of Protected Visual Environments (IMPROVE), Chemical Speciation Network (CSN/STN), and Clean Air Status and Trends Network (CASTNET) in the United States, Campaign on Atmospheric Aerosol Research network of China (CARE-China), National Air Quality Monitoring Programme (NAMP) in India, and AirBase and the EMEP (European Monitoring and Evaluation Programme) networks in the European Union, with some of the earliest measurements from the 1980s.

We investigate the possibility of using machine learning techniques to obtain CCN number concentrations from these measurements. Machine learning is a statistical learning branch of artificial intelligence where computers learn without being explicitly programmed to generalise from knowledge acquired by being trained on a huge number of specific scenarios. The

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020


Atmospheric Chemistry and Physics Discussions

development and use of machine learning to develop predictive models has been burgeoning over the last couple of decades, with recent applications in the atmospheric sciences such as in atmospheric new particle formation (e.g., Joutsensaari et al., 2018; Zaidan et al., 2018), mixing-state (e.g., Christopoulos et al., 2018; Hughes et al., 2018), air quality (e.g., Huttunen et al., 2016; Grange et al., 2018), remote-sensing (e.g., Fuchs et al., 2018; Mauceri et al., 2019; Okamura et al., 2017), and other aspects (e.g., Dou and Yang, 2018; Jin et al., 2019). These and other studies show the power of machine learning as a tool to account for the high non-linearities in the associations between atmospheric states and compositions towards predictive modeling.

The goal of the present study is to explore the possibility of deriving the number concentrations of CCN at 0.4% supersaturation ([CCN0.4]) through more ubiquitous measurements of atmospheric state and composition. For this, we develop a Random Forest Regression Model (RFRM). The machine learning model is trained on 30 years' (1989–2018) simulations from a chemical transport model incorporating a size-resolved particle microphysics model. The expectation is that the RFRM is able to learn the associations between atmospheric state and composition predictors and the [CCN0.4] (outcome).

The remainder of this paper is organised as follows: Section 2 details the data, techniques, and statistical performance metrics used in this study; Section 3 details the development of the RFRM, validation of its performance, and evaluation with empirical data; Section 4.1 summarises the major findings of this study; and Section 4.2 discusses some of the implications of this work and the avenues it opens up.

2 Data and methods

2.1 GEOS-Chem-APM model (GCAPM)

GEOS-Chem is a global 3D chemical transport model (CTM) driven by assimilated meteorological observations from the Goddard Earth Observing System (GEOS) of the NASA Global Modeling and Assimilation Office (GMAO). Several research groups develop and use this model, which contains numerous state-of-the-art modules treating emissions (van Donkelaar et al., 2008; Keller et al., 2014) and various chemical and aerosol processes (e.g., Bey et al., 2001; Evans and Jacob, 2005; Martin et al., 2003; Murray et al., 2012; Park, 2004; Pye and Seinfeld, 2010) for solving a variety of atmospheric composition research problems. The ISORROPIA II scheme (Fountoukis and Nenes, 2007) is used to calculate the thermodynamic equilibrium of inorganic aerosols. Secondary organic aerosol formation and aging are based on the mechanisms developed by Pye and Seinfeld (2010) and Yu (2011). MEGAN v2.1 (Model of Emissions of Gases and Aerosols from Nature; Guenther et al. (2012)) implements biogenic emissions and GFED4 (Global Fire Emissions Database; Giglio et al. (2013)) implements biomass burning emissions in GEOS-Chem.

The present study uses GEOS-Chem version 10-01 with the implementation of the Advanced Particle Microphysics (APM) package (Yu and Luo, 2009), henceforth referred to as GCAPM. The APM model has the following features of relevance towards accurate simulation of CCN number concentrations: (1) 40 bins to represent secondary particles with high size resolution for the size range important for growth of nucleated particles to CCN sizes (Yu and Luo, 2009) (2) state-of-the-art Ternary Ion mediated Nucleation (TIMN) mechanism (Yu et al., 2018) and temperature-dependent organic nucleation parameterisation

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020


Atmospheric Chemistry and Physics Discussions

(Yu et al., 2017); (3) calculation of H_2SO_4 condensation and the successive oxidation aging of secondary organic gases (SOG) and explicit kinetic condensation of low volatile SOG onto particles (Yu, 2011); (4) contributions of nitrate and ammonium via equilibrium uptake and semi-volatile organics through partitioning to particle growth considered (Yu, 2011). CCN number concentrations simulated by GCAPM has previously been shown to agree well with measurements (Yu and Luo, 2009; Yu, 2011; Yu et al., 2013).

The horizontal resolution of GCAPM in this study is $2° \times 2.5°$, with 47 vertical layers (14 layers from surface to 2 km above the surface). The period of global simulation is 30-years from 1989–2018. For 47 sites spread across the globe, co-located GCAPM data is output at the half-hourly time-step for all model layers in the troposphere. In the present application, we use those at 6 selected vertical heights: surface, ∼1 km, ∼2 km, ∼4 km, ∼6 km, and ∼8 km.

2.2 Random Forest Regression Modeling (RFRM)

There are various machine learning techniques: Linear Regression, Logistic Regression, Decision Trees, Support Vector Machines, Naive Bayesian Classifier, k- Nearest Neighbors, K-Means clustering, Random Forest, Dimensionality Reduction, and Gradient Boosting are the most commonly developed and applied. These techniques can be broadly categorised into (a) Supervised Learning (b) Unsupervised Learning, and (c) Reinforcement Learning.

In this study, we use the Random Forest due to our objective of predicting values of [CCN0.4], their ease-of-physical-interpretability, ease-of-implementation, and the ability to tune the supervised machine learning.

A Random Forest is an ensemble of decision trees. A decision tree is a supervised machine learning algorithm that recursively splits the data into subsets based on the input variables that best split the data into homogeneous sets. This is a top-down 'greedy' approach called recursive binary splitting. Decision trees are easy to visualise, are not influenced by missing data or outliers, and are non-parametric. They can, however, overfit on the data. Random Forest modeling is an ensemble technique of growing numerous decision trees from subsets (bags) of the training data and then using all the decision trees to make an aggregated (typically mean) prediction. This approach corrects for the overfitting of single decision trees. Additionally, the bootstrap aggregating (bagging) allows for model validation during training, by evaluating each component tree of the Random Forest with the out-of-bag training examples (training data that was not subsetted in growing the decision tree). Random Forest models are advantageous due to the component decision trees being able to resolve complex non-linear relationships between predictor variables regardless of their inter-dependencies or cross-correlations and the outcome to be predicted. Further, they are relatively easier to visualise and interpret as compared to black-box neural network or deep learning methods. For the purpose of predictions, Random Forest models are one of the most accurate machine learning models with the ability to be trained fast due to the parallelisability of the growth of decision trees. For these reasons, Random Forest is our chosen machine learning tool.

In this study, we use a fast implementation (Wright and Ziegler, 2017) of Random Forest models (Breiman, 2001) in R (R Core Team, 2020) trained on the above detailed GCAPM modeled [CCN0.4]. Further details of model development and applications are in section 3.1.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

2.3 Statistical estimators of model performance

In this study, we use the Kendall rank correlation coefficient (τ) and Mean Fractional Bias (MFB) as statistical estimators of correlation and deviation, respectively. These statistical parameters are more robust (as discussed later in this section) than the conventionally used Pearson product-moment correlation coefficient (r) and Mean Normalised Error (MNE) (or similar parameters).

Pearson product-moment correlation coefficient (r) is:

$$r = \sum_{i=1}^{n} \frac{(C_i^m - \overline{C^m})(C_i^o - \overline{C^o})}{\sqrt{\sum_{i=1}^{n}(C_i^m - \overline{C^m})^2}\sqrt{\sum_{i=1}^{n}(C_i^o - \overline{C^o})^2}}$$

and Kendall rank correlation coefficient (τ):

$$\tau = \frac{\sum_{i=2}^{n}(sign(C_i^m - C_{i-1}^m))(sign(C_i^o - C_{i-1}^o))}{\sqrt{\binom{n}{2} - \frac{1}{2}\sum_{i=1}^{n} t_i^m(t_i^m - 1)}\sqrt{\binom{n}{2} - \frac{1}{2}\sum_{i=1}^{n} t_i^o(t_i^o - 1)}}$$

where n is the sample size, C is the value, t is the number of ties in the i^{th} group of ties, and $[\,]^m$ denotes modeled and $[\,]^o$ denotes observed values.

As noted in Nair et al. (2019):

> In the use of Pearson's r are the following assumptions: (1) continuous measurements with pairwise complete observations for the two samples being compared (2) absence of outliers (3) Gaussian distribution of values (4) linearity between the two distributions, with minimal and homogenous variation about the linear fit (homoscedasticity).
>
> Kendall's τ is a nonparametric rank correlation coefficient that is not constrained by the assumptions in the use of Pearson's r. This parameter is also intuitive and simpler to interpret due to (a) the maximum possible value of $+1$ indicative of complete concordance and the minimum possible value of -1 indicative of complete discordance and (b) the ratio of concordance to discordance being $(1+\tau)/(1-\tau)$ (Kendall, 1970; Noether, 1981).

Mean Normalised Error (MNE) and Mean Fractional Bias (MFB) are defined as:

$$\text{MNE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{C_i^m - C_i^o}{C_i^o}\right| \quad \text{and} \quad \text{MFB} = \frac{1}{n}\sum_{i=1}^{n}\frac{(C_i^m - C_i^o)}{(\frac{C_i^m + C_i^o}{2})}$$

where n is the sample size, C is the value, $[\,]^m$ denotes modeled and $[\,]^o$ denotes observed values, and i is the i^{th} of their n pairs.

As noted in Nair et al. (2019):

> In the use of MNE, it is assumed that observed values are true values and not just estimates. MNE can easily blow up to ∞ when observed values are very small. Further, positive bias is weighted more than negative bias. Related parameters such as Normalised Mean Bias and Error (NMB and NME) also suffer from these deficiencies.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020


Atmospheric Chemistry and Physics Discussions

Mean Fractional Bias (MFB) is not limited by the issues in the use of MNE. As a measure of deviation, Mean Fractional Bias (MFB), ranging from $[-2, +2]$, is symmetric about 0 and also not skewed by extreme differences in the compared values.

The following quantitative ranges are provided (arbitrarily) to qualitatively describe the degree of correlation: (1) Poor agreement: $\tau \leq 0.2$, (2) Fair agreement: $0.2 < \tau \leq 0.4$, (3) Moderate agreement: $0.4 < \tau \leq 0.6$, (4) Good agreement: $0.6 < \tau \leq 0.8$), and (5) Excellent agreement: $0.8 < \tau \leq 1.0$. Additionally, it is defined (arbitrarily; factor of 1.86 deviation) that there is good agreement between derived and expectated values when the MFB is within $[-0.6, +0.6]$.

Statistical analyses are performed using R: a freely available language and environment for statistical computing and graphics (R Core Team, 2020) and with the aid of the 'Kendall'(McLeod, 2011) and 'pcaPP'(Filzmoser et al., 2018) packages.

2.4 Observational data

For validation of the developed RFRM, we use in situ measurements of atmospheric state and composition as inputs to the RFRM and compare the output [CCN0.4] with its measurements. The U.S. Department of Energy's (DOE) Atmospheric Radiation Measurement (ARM) Southern Great Plains (SGP) Central Facility located in Lamont, Oklahoma ($36°36'18''$ N, $97°29'6''$ W, 318 m; Fig. 1) was established with the mission statement of "provid[ing] the climate research community with strategically located in situ and remote-sensing observatories designed to improve the understanding and representation, in climate and earth system models, of clouds and aerosols as well as their interactions and coupling with the Earth's surface." We use data from this facility, which has the longest record of [CCN0.4] at the hourly resolution.

2.4.1 [CCN0.4] measurements

[CCN0.4] measurements at this SGP site have been made from 2007–present using a Cloud Condensation Nuclei Particle Counter (CCNc) developed by Roberts and Nenes (2005) with technical details in Uin (2016). The CCNc is a continuous-flow thermal-gradient diffusion chamber for measuring aerosols that can act as CCN. To measure these, aerosol is drawn into a column, where well-controlled and quasi-uniform centerline supersaturation created. Through software controls, the temperature gradient and flow rate are modified to vary (0.1–3%) supersaturations and obtain CCN spectra. Water vapor condenses on CCN in the sampled air to form droplets, just as cloud drops form in the atmosphere; these activated droplets are counted and sized by an Optical Particle Counter (OPC).

We integrate the quality-checked data (Shi and Flynn, 2007; Smith et al., 2011a, b; Hageman et al., 2017) made publicly available through ARM from co-located instruments at the SGP site to form a long-term record of [CCN0.4] as in Fig. 2 and for later analysis and validation of the RFRM.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

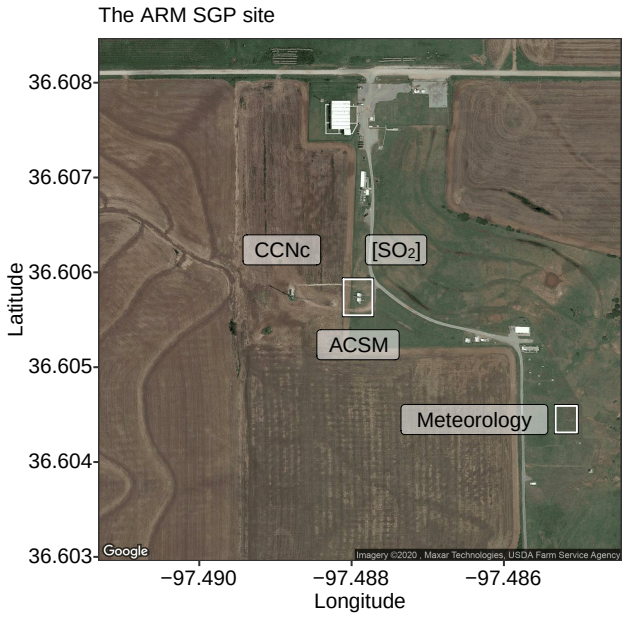

Figure 1. The ARM SGP site in Lamont, Oklahoma, U.S.A with marked locations of the instruments. The image is adapted from satellite imagery © 2020 Maxar Technologies, USDA Farm Service Agency obtained through the Google Maps Static API.

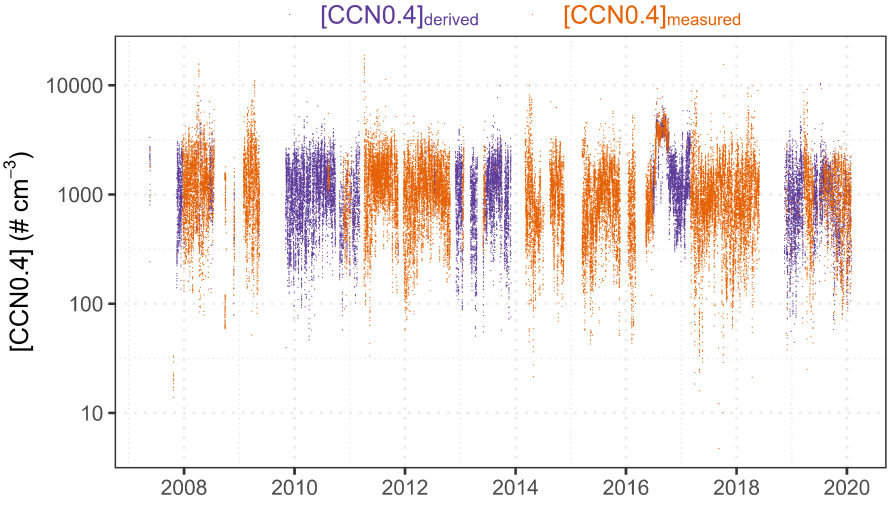

Figure 2. Hourly [CCN0.4] measurements at Lamont, Oklahoma from 2007–2020. The orange circles show the data for [CCN0.4] measured here. The purple circles show the data for [CCN0.4] derived from [CCN0.2–0.6] according to the method described in Section 2.4.2.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020


Atmospheric Chemistry and Physics Discussions

Table 1. Evaluation metrics of the Random Forest models developed to derive [CCN0.4] from other supersaturations.

RF–Model	IQR(MFB)	τ
rf_2356	0.05	0.93
rf_235	0.06	0.92
rf_256	0.08	0.91
rf_236	0.09	0.91
rf_356	0.10	0.90
rf_25	0.11	0.88
rf_26	0.11	0.89
rf_35	0.11	0.88
rf_56	0.11	0.89
rf_36	0.15	0.87
rf_23	0.25	0.74

2.4.2 Filling the [CCN0.4] measurement gaps

The temporal range of available observations for [CCN0.4] at the SGP site is 2007-05-19 17:00:00 to 2020-01-29 23:00:00. For this period of 111319 hours, there is only $\sim 42\%$ data completeness. To improve the data coverage, we examine the possibility of using [CCN] for other reported supersaturation ratios $(0.2 - 0.6\%)$.

In a sneak preview of the efficacy of Random Forest for regression, we train Random Forest models that output [CCN0.4] from [CCN0.2–0.6]. These models are listed in Table 1, with the notation rf_n, where n denotes the supersaturations used as input (for instance, rf_356 indicates [CCN0.3], [CCN0.5], and [CCN0.6] were used as inputs to derive [CCN0.4]). The approach works exceptionally well and shows the potential for application with other datasets to fill in such gaps as well as to perform sanity checks on available data. As reported in Table 1, there is high correlation (Kendall's τ) and minimal deviation (interquartile range of mean fractional bias (IQR(MFB)) between Random Forest derived [CCN0.4] and measured [CCN0.4], when simultaneous data for [CCN0.4] is available.

Using this approach, we fill in the missing observations and improve data completeness from $42\% \rightarrow 66\%$, a 54% increase, for [CCN0.4] during this period (see Fig. 2).

2.4.3 Atmospheric state and composition measurements

Meteorological data is sourced from the ARM Surface Meteorology Systems (MET) (Holdridge and Kyrouac, 1993) with technical details in Ritsche (2011). We use the ARM Best Estimate Data Products (ARMBEATM; Chen and Xie (1994)) derived from (Holdridge and Kyrouac, 1993) when available (1994–2016).

Trace gas concentrations are obtained from the ARM Aerosol Observing System (AOS; Hageman et al. (1996)) with technical details in Jefferson (2011). Unfortunately, at the SGP site, measurements (Springston, 2012) are available only for $[SO_2]$

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020


Atmospheric Chemistry and Physics Discussions

Table 2. Selected atmospheric state and composition variables as RFRM predictors for [CCN0.4].

Meteorological			
Temperature	Precipitation	RH	Solar Radiation
Chemical Species			
$[NO_x]$	$[NH_3]$	[OH]	[Isoprene]
$[SO_2]$	$[O_3]$		[Monoterpene]
PM2.5 Speciated Mass Fraction			
SO_4	NO_3	NH_4	BC
POC	SOA	Dust	Salt

from 2016–present. To compensate for this, we use data from the United States Environmental Protection Agency (EPA) Air Quality System (AQS) made publicly available at https://www.epa.gov/air-data from monitors in the vicinity (< 100 km) of the SGP site.

Real-time aerosol mass loadings and their chemical composition measurements have been made from 2010–present using an Aerodyne Aerosol Chemical Speciation Monitor (ACSM; Ng et al. (2011)) with technical details in (Watson, 2017). We use the aerosol chemical speciation data (Watson et al., 2018; Kulkarni, 2019; Behrens et al., 1990), which is publicly available through ARM.

3 Results and Discussion

3.1 RFRM: training, testing, and optimising

The GEOS-Chem-APM (GCAPM) output for 47 sites across the globe, for 6 selected vertical levels from surface to ~ 8 km, and for 30 years (1989–2018) at the half-hour time-step ($\sim$150 million rows, i.e. sets of predictors and [CCN0.4]) is considered in training the random forest regression model (RFRM). The predictors of importance in controlling [CCN0.4] are listed in Table 2. The RFRM is trained on a subset of this data. First, the ARM SGP site is ignored, which will be discussed separately with available observational data in a later section. The remaining data for 46 sites and 6 vertical levels each are partitioned into training ($\sim$101 million rows) and testing sets ($\sim$44 million rows) in a 7:3 ratio. Due to the large number of training and testing examples, these sets are reduced to a 1% random subset; it is ensured to be representative, with almost identical statistical properties, of the training datasets (Fig. 3).

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

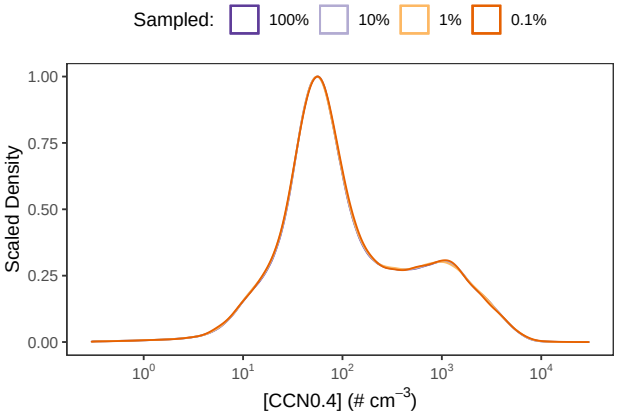

Figure 3. Scaled Gaussian kernel density estimate for [CCN0.4] for the training set (dark purple) and each of its subsets (10%: light purple; 1%: light orange; 0.1%: dark orange). The distributions are almost identical. A 1% randomly sampled subset is used to train the RFRM.

Once the data has been selected to train the RFRM, we tune the parameters of the model. The default implementation of Wright and Ziegler (2017) comes with reasonable (balancing speed and accuracy) choices for the number of trees in the forest ($numtrees$), the minimum number of variables to consider for each split ($mtry$), and the minimum node size ($min.node.size$), i.e. the minimum size of homogeneous data to prevent overfitting. These defaults are: $numtrees$ = 500, $mtry$ = rounded-down square root of the number of variables, and $min.node.size$ = 5. We verify if these hyperparameter choices are optimal by performing a grid-search of the hyperparameters and training multiple Random Forest models and not just examining their performance with the training set, but also additionally with the test set. Figs. 4 & 5 show the results of this exercise.

Fig. 4 has 9 panels; from top to bottom:

- the top 3 are for varying number of trees in the Random Forest: $numtrees$ from 1 (a single decision tree) to 1800, through the default choice of 500 (marked with vertical black dotted line);
- the middle 3 are for varying number of variables considered at each split point: $mtry$ from 1 to 19, through the default choice of 4 (marked with vertical black dotted line); and
- the bottom 3 are for varying minimum size of the terminal nodes: $min.node.size$ from 1 to 10, through the default choice of 5 (marked with vertical black dotted line).

from left to right:

- the left 3 (green) show the overall out-of-bag error, i.e. the mean square error for the entire Random Forest computed using the complement of the bootstrapped data used to train each tree;
- the middle 3 (orange) show the R^2 values indicating the explained variance by the Random Forest; and

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

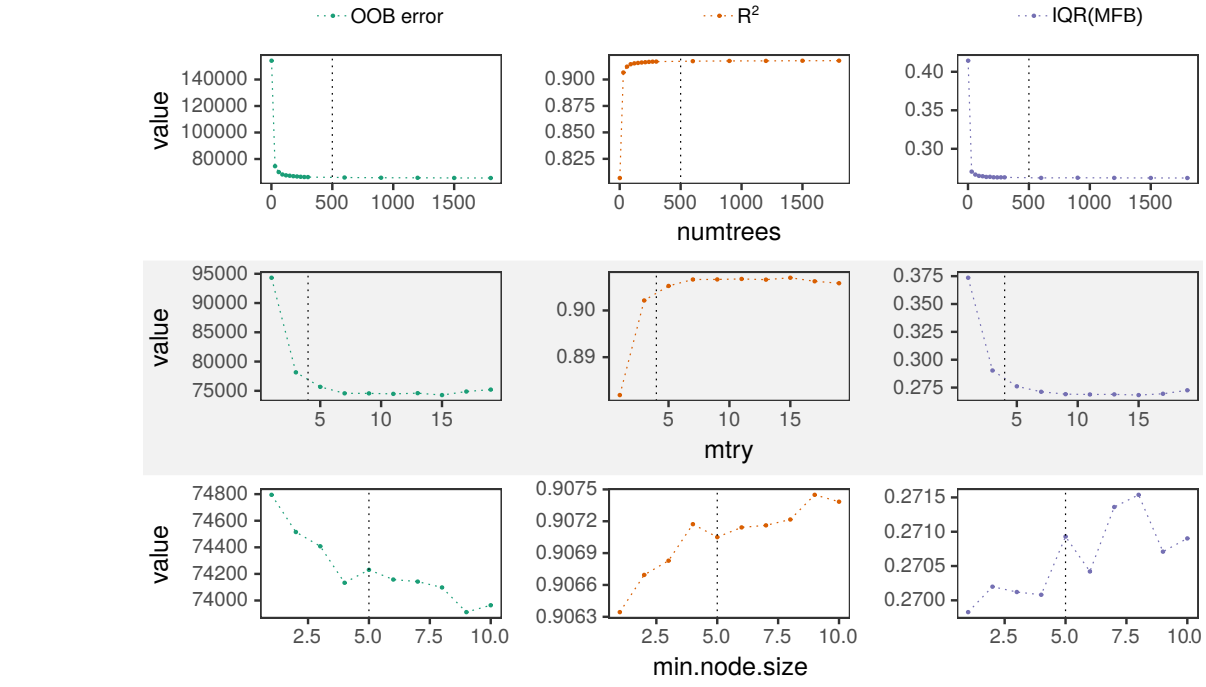

Figure 4. RFRM evaluation through OOB error (green), R^2 (orange), and IQR(MFB) (purple) with varying hyperparameters. The hyperparameters are (from top to bottom): $numtrees$, $mtry$, and $min.node.size$. Default hyperparameters for each trained model are: $numtrees$ = 500, $mtry$ = 4, and $min.node.size$ = 5, also shown with the vertical black dotted lines when the corresponding default parameter has been varied.

– the right 3 (purple) show the interquartile range of the mean fractional bias of the Random Forest model when applied to the test set.

A single decision tree (left-most point in each of the top 3 panels of Fig. 4) is able to explain the variance ($R^2 \approx 0.81$) in [CCN0.4] through the predictors' and has an interquartile range in the MFB (which has a median of ~ 0.004: not pictured as the symmetry of MFB means the median $\rightarrow 0$) of 0.41, which corresponds to a deviation of $\sim \pm 20\%$. However, this is drastically improved when moving beyond a simple decision tree to even a small ensemble of 30 trees ($R^2 \approx 0.91$; IQR(MFB) ≈ 0.27), which plateaus (within a range of ± 0.0005) after $\sim$500 trees. The out-of-bag (OOB) error shows a similar trend. Growing a Random Forest of 500 trees with a $min.node.size$ of 5, we see the effect of varying $mtry$ in the middle 3 panels. Instead, keeping the $mtry$ fixed at the default value of 4: the rounded down square root of the 19 predictor variables, the Random Forest performance metrics with varying $min.node.size$ are shown in the bottom 3 panels. Fig. 4 thus shows the possibility of improving the Random Forest derivation of [CCN0.4] by changing the default choices of Wright and Ziegler (2017) for this specific work. It must be kept in mind that a reasonable cutoff, beyond which there is imperceptible gain in performance at increased computational cost, should be considered.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

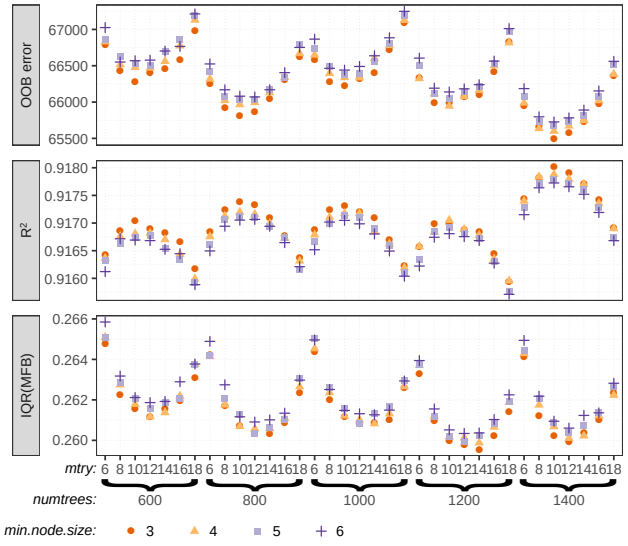

Figure 5. RFRM evaluation through OOB error (top), R^2 (middle), and IQR(MFB) (bottom) with varying hyperparameters selected from the best balanced (accuracy and computational expense) ranges in Fig. 4.

The results shown in Fig. 4 motivate a zoomed-in hyperparameter grid search to choose the optimal (accurate and fast) RFRM. Fig. 5 shows this for the best performing Random Forest models with $numtrees$ ranging from 600 to 1400, $mtry$ from 6 to 18, and $min.node.size$ from 3 to 6. While there is variability, it must be noted that the y-axes range over 2%, 0.2%, and 2% of the values of OOB error, R^2, and IQR(MFB), respectively. While, indeed, considering a larger ($numtrees$) forest is beneficial, considering the cost-to-benefit ratio, the hyperparameters we choose are a maximum number of 800 trees in the forest, 12 ($mtry$) variables are randomly chosen at each split, minimum node size of 3 as the only control on tree depth, and the splitting rule is the minimisation of variance. With these parameters, the Random Forest has IQR(MFB), R^2, and OOB error of $\sim 100.34\%$, $\sim 99.93\%$, and $\sim 100.56\%$ of the best-performing model for each parameter, respectively.

3.2 What the model learns

We train the optimised RFRM using the 19 predictors listed in Table 2 as predictors of [CCN0.4]; these are: 8 fractions of PM2.5 (ammonium, sulfate, nitrate, secondary organic aerosol (SOA), black carbon (BC), primary organic carbon (POC), dust, and salt), 7 gaseous species (nitrogen oxides (NO_x), ammonia (NH_3), ozone (O_3), sulfur dioxide (SO_2), hydroxyl radical (OH), isoprene, and monoterpene), and 4 meteorological variables (temperature, relative humidity (RH), precipitation, and solar radiation).

Fig. 6 shows the importance of each predictor in determining the CCN0.4 number concentration in the above-trained RFRM. This importance measure is obtained by randomly permuting values of each predictor to break the association with CCN0.4. Also, the model is fed a pseudo-predictor of randomly generated white noise, labelled 'Random' in Fig. 6. Most important are

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

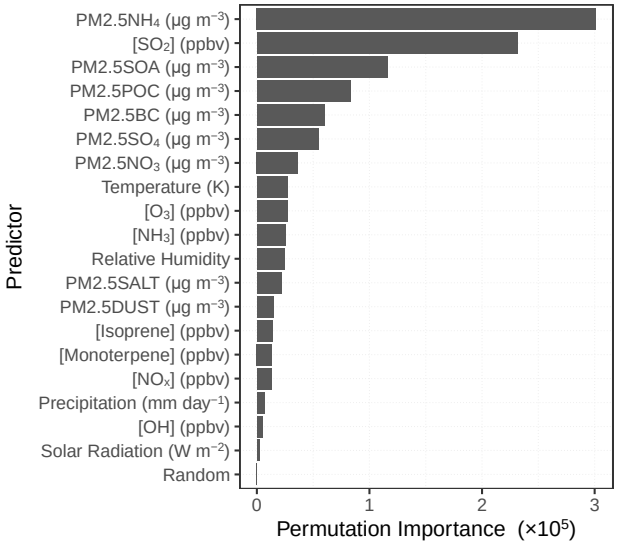

Figure 6. Importance (in decreasing order) of each predictor in the RFRM derivation of [CCN0.4].

component mass fractions of PM2.5, especially its inorganic fraction (ammonium, sulfate, and nitrate), $[SO_2]$, the other PM2.5 fractions excluding the salt and dust fractions, $[NO_x]$, and $[NH_3]$. The 'Random' predictor is least important; contributing imperceptibly to the [CCN0.4] prediction.

To quantify the importance of each of the 19 predictors, we do the following: (1) RF-Blind: train RFRMs without considering one variable at a time and (2) RF-Random: randomise each predictor variable and input into RFRM-AllVars. These exercises will provide a more robust, as well as more intuitive measure of the importance of each predictor by analyzing and comparing the deviation in predicted [CCN0.4] compared to the baseline optimised RFRM (hereafter, RFRM-AllVars). Table 3 shows the median $\pm$ median-absolute-deviation of mean fractional bias (MFB) for each such trained RF-Blind and RF-Random evaluated with the test dataset, with the values for the baseline RF-AllVars at the bottom. For the MFB, its median-absolute-deviation (and even IQR(MFB)) is a stronger indicator of the performance than the median, due to the symmetry of MFB. Examining the median-absolute-deviation of the MFB, the 'Blind' approach shows that the ammonium ($PM2.5NH_4$), sulfate ($PM2.5SO_4$), and secondary organic (PM2.5SOA) fractions of PM2.5 are most important (in increasing order) in determining [CCN0.4]. The 'Blind' approach may however underestimate the importance of a predictor due to possible correlations with other predictors, as seen in Fig. 7. These cross-correlations could mean the implicit participation of a predictor despite its absence in training the RFRM. To overcome this limitation, in the 'Random' approach, the trained RFRM-AllVars is then input randomised predictors (one at a time) from the testing dataset. This breaks the association of each predictor with the outcome ([CCN0.4]) as well as with other predictors. The resulting change (if any) in the RFRM-AllVars would show the importance of the specific predictor. RF-Random for each predictor shows that all predictors (except solar radiation) are important, with the most important being PM2.5SOA, $PM2.5SO_4$, and $PM2.5NH_4$.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

Table 3. Mean fractional bias (MFB) of each Random Forest Regression Model (RFRM). RF-Blind refers to the RFRM trained ignoring the particular predictor. RF-Random refers to the randomisation of the particular predictor before input into RF-AllVars. RF-AllVars (at the bottom of the table) is the baseline model where no variable is omitted/randomised. Values are median ± median-absolute-deviation of MFB.

Predictor	**RF-Blind**	**RF-Random**
PM2.5SOA	0.056 ± 0.261	0.120 ± 0.888
PM2.5SO_4	0.051 ± 0.238	0.133 ± 0.493
PM2.5NH_4	0.044 ± 0.214	0.079 ± 0.345
Temperature	0.050 ± 0.221	0.109 ± 0.333
$[SO_2]$	0.049 ± 0.215	0.059 ± 0.296
PM2.5BC	0.044 ± 0.211	0.094 ± 0.263
PM2.5SALT	0.044 ± 0.224	0.054 ± 0.256
PM2.5POC	0.044 ± 0.211	0.082 ± 0.246
$[O_3]$	0.044 ± 0.212	0.056 ± 0.239
$[NO_x]$	0.043 ± 0.209	0.058 ± 0.223
PM2.5NO_3	0.044 ± 0.208	0.059 ± 0.223
PM2.5DUST	0.044 ± 0.211	0.056 ± 0.221
Relative Humidity	0.040 ± 0.208	0.047 ± 0.221
$[NH_3]$	0.046 ± 0.209	0.052 ± 0.218
[Monoterpene]	0.045 ± 0.209	0.048 ± 0.213
[Isoprene]	0.044 ± 0.209	0.046 ± 0.211
[OH]	0.044 ± 0.208	0.048 ± 0.211
Precipitation	0.044 ± 0.209	0.046 ± 0.210
Solar Radiation	0.044 ± 0.208	0.045 ± 0.209
RF-AllVars	0.044 ± 0.209	

Table 3 thus shows the importance of each predictor towards determining [CCN0.4] in decreasing order, which complements the results in Fig. 6. Fig. 6 shows the out-of-bag increase in mean square error upon permutation of a specific predictor. We modify the approach in Wright and Ziegler (2017) (a) to leverage the advantages of the Mean Fractional Bias (MFB); (b) to account for any implicit correlations between the data used to train (bagged) and evaluate (out-of-bag) the RFRM by using the unseen testing dataset; and (c) to dissociate the effects of cross-correlations between predictors. The results of this modified evaluation of the RFRM are in Table 3. These exercises to probe into the working of the RFRM show that all of the predictors were deemed necessary to capture [CCN0.4] magnitude and variability. The most important predictors are the PM2.5 speciated components, gases including $[SO_2]$, $[O_3]$, and $[NO_x]$, and temperature and relative humidity.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

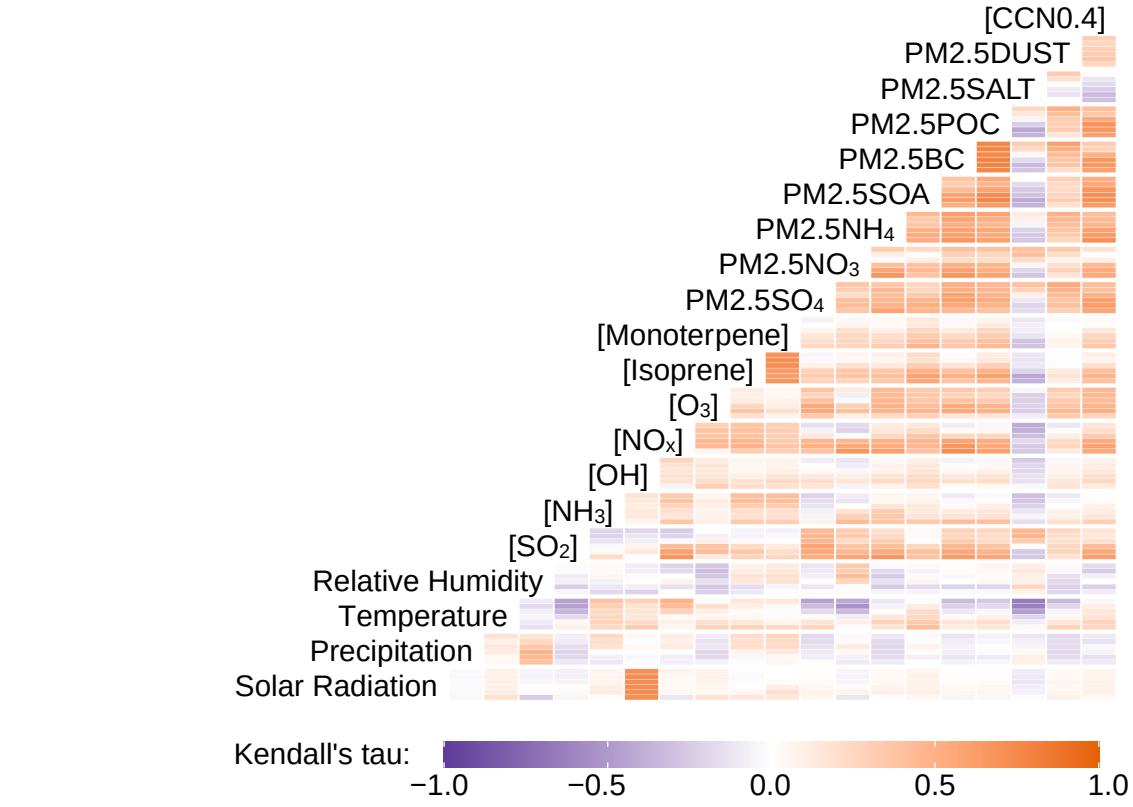

Figure 7. Kendall rank correlation (τ) for the 19 predictors and [CCN0.4] in the RFRM training dataset ($\sim$ 101 million rows). The boxes show corresponding τ value according to color-scale. Boxes are divided vertically to represent τ for each selected model height: surface, $\sim$1 km, $\sim$2 km, $\sim$4 km, $\sim$6 km, and $\sim$8 km.

3.2.1 Comparison with GCAPM [CCN0.4]

We examine the RFRM in further detail with the subset of data excluded for testing, i.e. for the sites that RF-AllVars has not been exposed to during its training. Fig. 8 (a) shows RF-AllVars derived [CCN0.4] against that simulated by GCAPM for all of the testing dataset. RF-AllVars predicted [CCN0.4] values are highly correlated with expected values from GCAPM, with a correlation of $\tau \approx 0.88$ and highest density along the dashed black line in Fig. 8 (a) denoting MFB = 0 or complete agreement. The dashed blue and red lines denote $-1 < MFB < +1$; $\sim 99.69\%$ of the values are within this factor of $3\times$ deviation. The dotted lighter blue and red lines denote $-0.6 < MFB < +0.6$; $\sim 96.33\%$ of the values are within this range of good agreement between derived and expected [CCN0.4] values.

Overall, the RFRM is able to derive [CCN0.4] with a median (median-absolute-deviation) MFB of $4.4(21)\%$. A comparison to expected [CCN0.4] values from GCAPM in Fig. 9 (a) by means of the MFB shows the robustness of RF-AllVars in greater detail. The highest density of MFBs is on or around 0, reiterating how well RF-AllVars predictions of [CCN0.4] compares to GCAPM simulated values.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

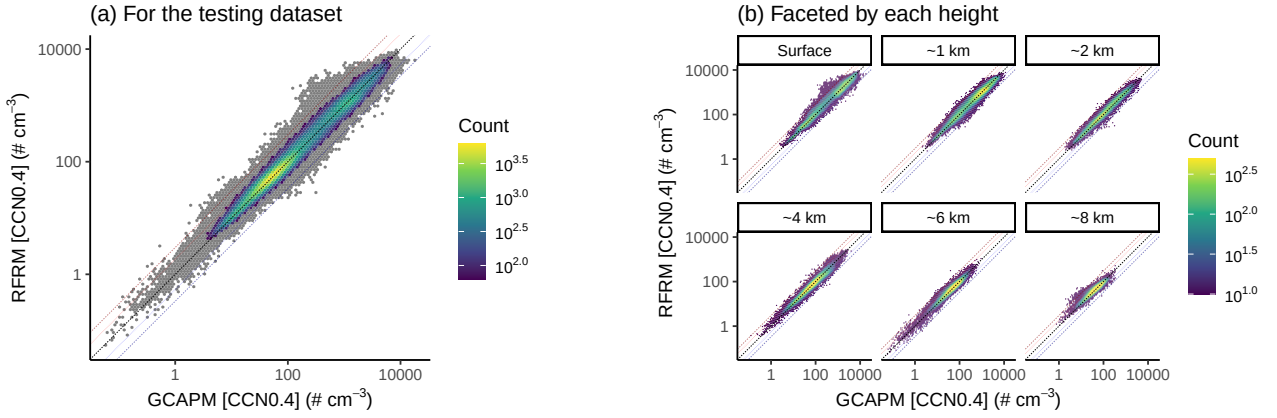

Figure 8. Binned scatterplot of RFRM versus GCAPM simulated [CCN0.4]. Color bar shows the counts of points in each hexagonal bin. Bins with low counts ($< 1\%$ of maximum count: $\sim 3\%$ of the data) are shaded grey. The lines indicate MFB of 0 (black; perfect agreement), $+1$ (darker red), -1 (darker blue), $+0.6$ (lighter red), and -0.6 (lighter blue).

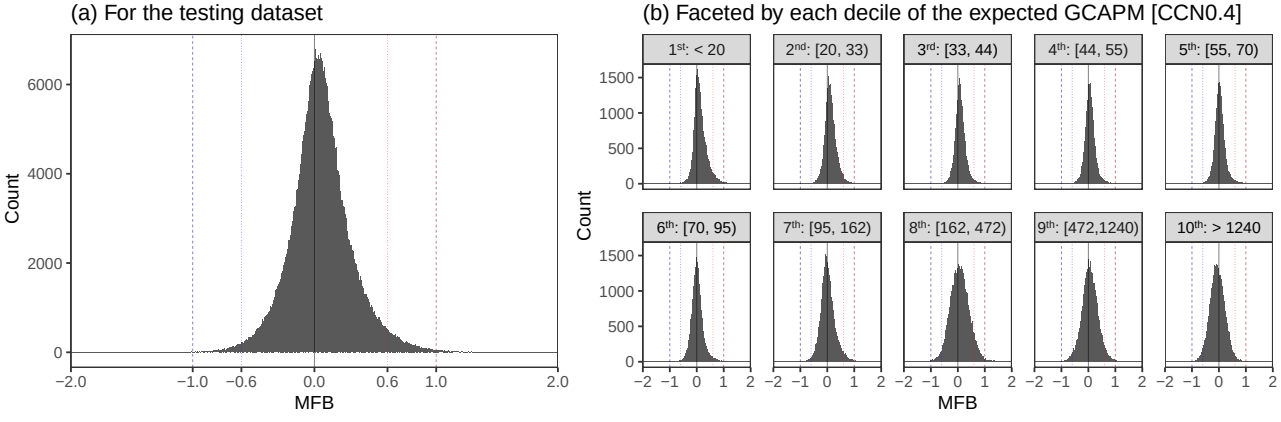

Figure 9. Mean fractional bias (MFB) of RFRM derived [CCN0.4] compared to expected values from GCAPM in the testing dataset (441756 values). Histogram shows the counts of the pairs by MFB. The lines indicate perfect agreement (black), MFB of $+1$ (dashed red), and MFB of -1 (dashed blue). The dotted lines indicate MFB of $+0.6$ (dotted red) and -0.6 (dotted blue).

300 Fig. 8 (a) is faceted by height in Fig. 8 (b). Across the various heights, with varied [CCN0.4] ranges, RF-AllVars performs robustly with $|\text{MFB}| < 2$. Its robustness across varied [CCN0.4] ranges is further shown in Fig. 9 (b), where Fig. 9 (a) is faceted by the deciles of the GCAPM simulated [CCN0.4]. RF-AllVars performs well over 4 orders of magnitude of [CCN0.4] from 10^0 to 2.7×10^4 cm^{-3}.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

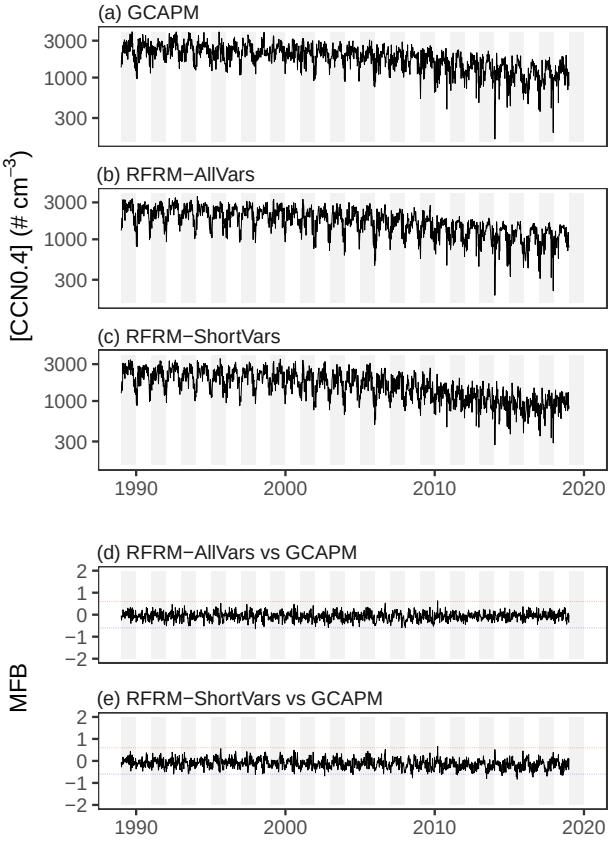

Figure 10. Time series (weekly-aggregated) for the surface $2 \times 2.5^{\circ}$ GCAPM gridbox containing the SGP site: **(a)** GCAPM simulated, **(b)** RFRM-AllVars, and **(c)** RFRM-ShortVars derived [CCN0.4]; and the Mean Fractional Bias (MFB) of the RFRM with **(b)** AllVars and **(e)** ShortVars. Dotted lines show the good agreement range of $|\text{MFB}| < 0.6$.

3.2.2 Comparisons for the SGP site

We also examine the temporal trends of [CCN0.4] for the SGP site (the surface $2 \times 2.5^{\circ}$ model gridbox), which was completely excluded from the RFRM training. Fig. 10 shows the weekly-aggregated time series from 1989–2018 for (a) GCAPM simulated and (b) RFRM-AllVars derived [CCN0.4]. Also shown in Fig. 10(d) is the comparison, using MFB, of the derived [CCN0.4] of RFRM-AllVars versus the GCAPM [CCN0.4] values. The RFRM performs well, being able to capture weekly variations with good correlation ($\tau \approx 0.68$) and low deviation ($\sim 99.87\%$ within the good agreement range of $|\text{MFB}| < 0.6$).

As detailed in Section 2.4, there are numerous observations of atmospheric state and composition for the SGP site that can be utilised to validate the RFRM with empirical data. In an ideal situation, continuous, long-term, high quality measurements for all the inputs to the RFRM (the predictor variables listed in Table 2) would have aided in this analysis. However, due to the sparsity and absence of measurements of certain predictors, we are limited to the available factors listed in Table 4. These

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

Table 4. RFRM predictors for [CCN0.4] as listed in Table 2. Italicised text shows those predictors determined (in Section 3.2) to not strongly impact RFRM prediction of [CCN0.4]. Strike-through text shows the absence of their hourly measurements.

Meteorological			
Temperature	*Precipitation*	RH	*Solar Radiation*
Chemical Species			
$[NO_x]$	~~$[NH_3]$~~	~~*[OH]*~~	~~*[Isoprene]*~~
$[SO_2]$	$[O_3]$		~~*[Monoterpene]*~~
PM2.5 Speciated Mass Fraction			
SO_4	NO_3	NH_4	~~*BC*~~
POC†	SOA†	~~*Dust*~~	~~*Salt**~~

Note: † Measurements for PM2.5 POC and SOA are reported as PM2.5 OC (total organic carbon). * For PM2.5 Salt, PM2.5 Chloride measurements are available, but subject to a large percent of missing data.

are shown in Fig. 11, with the speciated PM2.5 predictors in Fig. 11(a), the trace gas measurements in Fig. 11(b), and the meteorological variables in Fig. 11(c).

Thus, with a reduction from $19 \rightarrow 9$ predictor variables, we retrain the RFRM to use only these as inputs to derive [CCN0.4]. The RFRM optimisation is carried out as described in Sec. 2.2; the RFRM parameters are $numtrees = 1000$, $mtry = 3$, and $min.node.size = 3$ and uses the 9 predictors listed in Table 4 to derive [CCN0.4]. This retrained RFRM (henceforth, RF-ShortVars) is evaluated using the testing dataset; with median (median-absolute-deviation) MFB deteriorating from $0.044(0.209) \rightarrow -0.184(0.382)$, $96.33 \rightarrow 80.30\%$ of the derived [CCN0.4] values in the good agreement range, and correlation reducing from $\tau \approx 0.88 \rightarrow 0.79$. While RFRM-ShortVars is less robust compared to RFRM-AllVars, the statistical estimators of model performance are still high for RFRM-ShortVars.

Specifically for the SGP site, Fig. 10 (c) & (e) show the weekly-aggregated RFRM-ShortVars derived [CCN0.4] and its comparison with the GCAPM [CCN0.4] values, respectively. RFRM-ShortVars performs well, being able to capture these variations with good correlation ($\tau \approx 0.66$) and low deviation ($\sim 98.72\%$ within the good agreement range).

Using measurements of the 9 predictor variables at the SGP site for 2010–2020, we use the developed RF-ShortVars to derive [CCN0.4]. Compared to [CCN0.4] measurements, the RFRM performs well, with $\tau \approx 0.36$ and $\sim 67\%$ with $|\text{MFB}| < 0.6$. Apart from the expected deteriorated performance due to the reduction of necessary predictors to the available ones, the uncertainties associated with the measurements themselves may compound this. Regardless, the variability from diurnal to decadal scales are captured by the RFRM (top-left panel in Fig. 12) when compared to the measurements (bottom-left panel

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

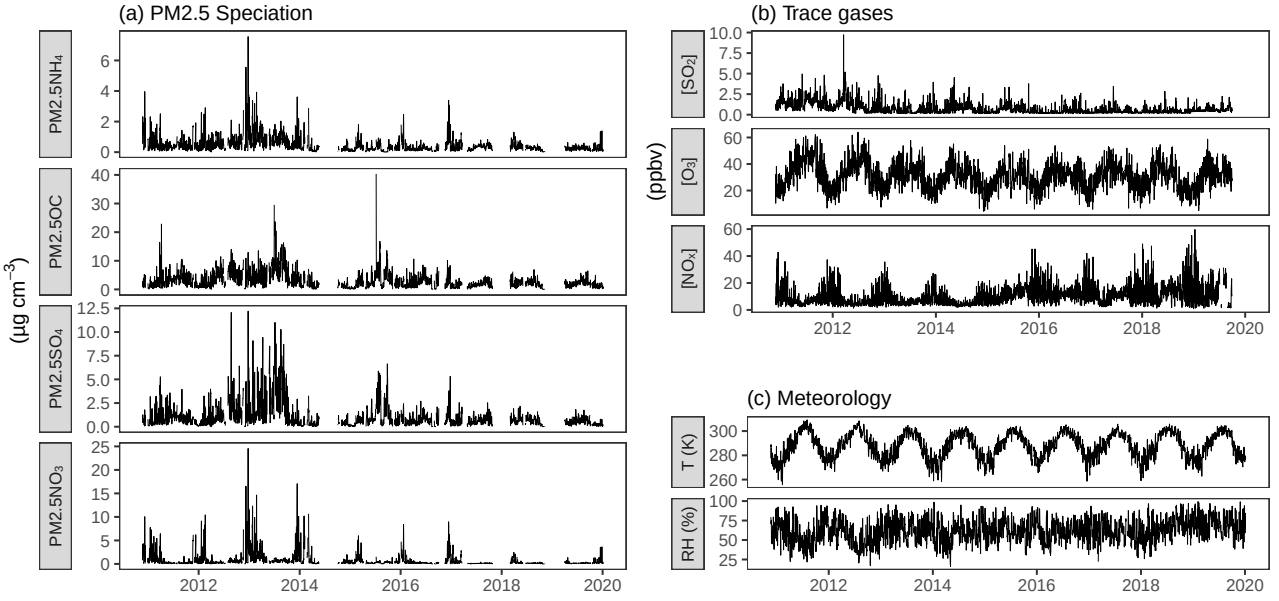

Figure 11. Time series (daily-aggregated) for predictors at the ARM SGP site: (a) PM2.5 speciation; (b) trace gas; and (c) meteorological measurements.

in Fig. 12) of [CCN0.4]. For reference, in Fig. 12 the top-right and bottom-right panels are the same as in Fig. 10 (c) & (a), respectively, but for this period of measurements.

Development of the machine learning model generally requires a large number of training examples, however, we also investigate the possibility of developing an RFRM with the measurement data alone. The RFRM optimisation is carried out as described in Sec. 2.2; this RFRM trained on actual measurements at SGP has $numtrees = 1000$, $mtry = 3$, and $min.node.size = 5$. Compared to the RFRM trained on GCAPM simulated data (∼150 million training examples), such an RFRM has only ∼34,000 training examples for the SGP site. We train such an RFRM, henceforth denoted as ORF (observation-based random forest regression model), and examine its performance.

In comparison with SGP measured [CCN0.4], ORF shows correlation of $\tau \approx 0.53$ and good agreement with $\sim 81.22\%$ of its derived [CCN0.4] values. The time series (daily-aggregated) of ORF-derived [CCN0.4] is shown in Fig. 12, with measured predictors as input to the RFRM in the middle-left panel and GCAPM predictors as input to the RFRM in the middle-right panel. In Fig. 13(b) ORF-derived [CCN0.4] is compared to hourly measurements. ORF appears to perform better than RF-ShortVars from the summary statistics. However, it is unable to capture the range of variations in magnitude (middle-left panel in Fig. 12). Similar results are observed (Fig. 13(d) and Fig. 12 middle-right) when ORF is applied to GCAPM simulated data. This exercise of developing an observation-based RFRM is, however, not technically justifiable due to the small number of training examples and applicability to only the SGP site. Regardless, this exercise provides insight into whether considering only 9 out of the 19 required predictor variables is sufficient for RF-ShortVars, examination of which suggests the affirmative.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

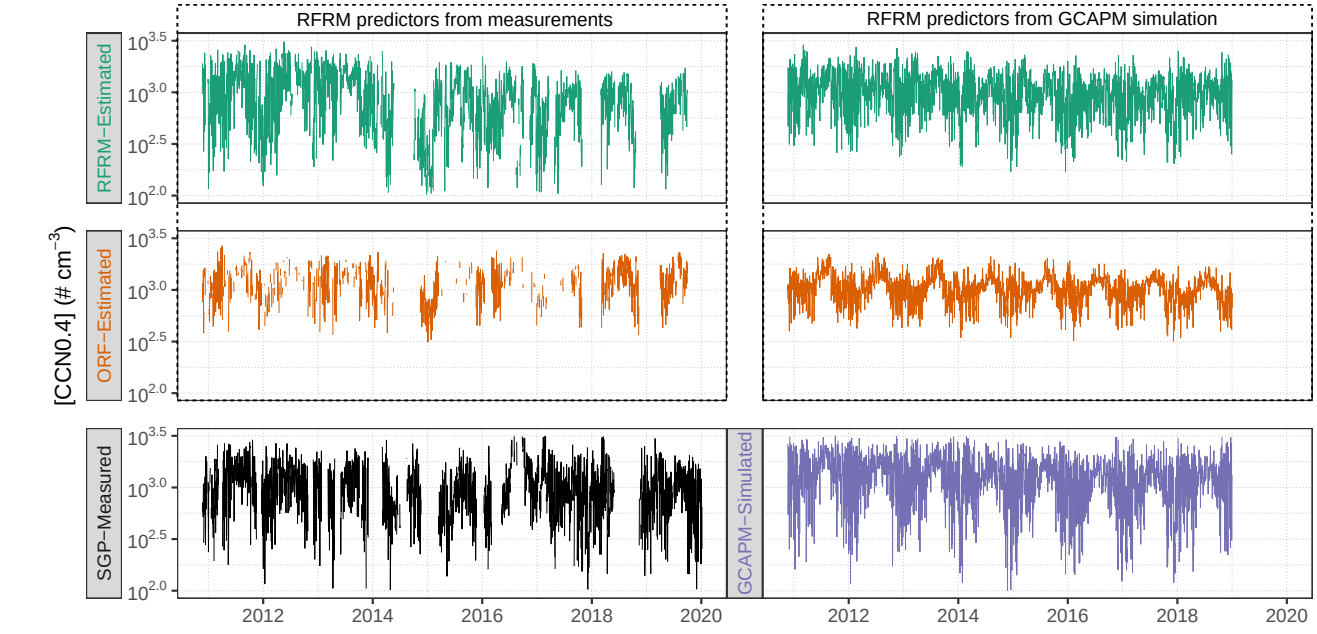

Figure 12. Time series (daily-aggregated) for [CCN0.4] derived by RF-ShortVars (top two) and ORF (middle two), compared to SGP measured (bottom-left) and GCAPM-simulated (bottom-right).

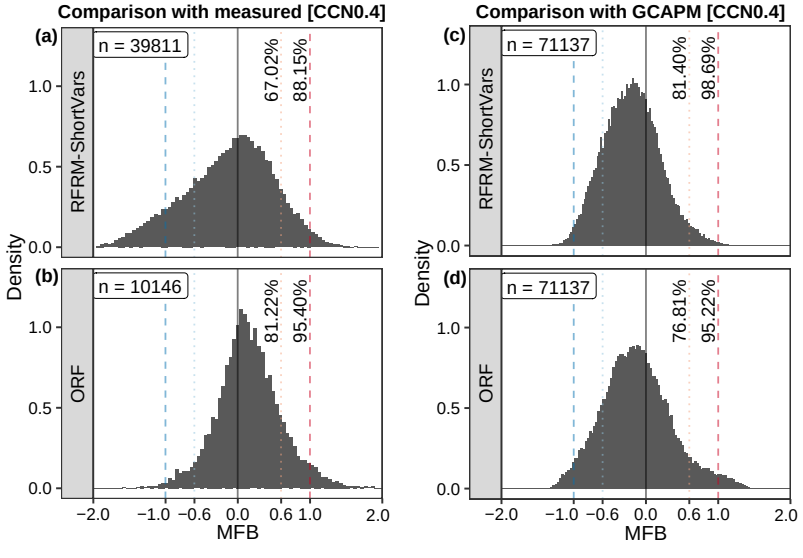

Figure 13. Mean fractional bias (MFB) of the random forest derived [CCN0.4] compared to expected values. *Lefthand panels:* comparison with measured [CCN0.4]. *Righthand panels:* comparison with GCAPM simulated [CCN0.4]. The histograms show the pairwise counts (total is inset top-left in each panel) by MFB. The lines indicate MFB of 0 (black), $+1$ (dashed red), -1 (dashed blue), $+0.6$ (dotted red), and -0.6 (dotted blue). The percentage of RFRM derived values in good ($|\text{MFB}| < 0.6$) and fair ($|\text{MFB}| < 1$) agreement are shown close to the +0.6 and +1.0 MFB lines, respectively.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020


Atmospheric Chemistry and Physics Discussions

4 Conclusions

4.1 Summary

We develop a Random Forest Regression Model (RFRM) to predict the number concentrations of cloud condensation nuclei at 0.4% supersaturation ([CCN0.4]) from atmospheric state and composition variables. This RFRM, trained on 30-year simulations by a chemical transport model (GEOS-Chem) with a detailed microphysics scheme (APM), is able to predict [CCN0.4] values. The RFRM learns that the PM2.5 fractions (except salt and dust) and gases such as SO_2 and NO_x are the most important determinants of [CCN0.4]. The RFRM is robust in its derivation of [CCN0.4], with median (median-absolute-deviation) mean fractional bias of $4.4(21)\%$ with 96.33% of the derived values within the good agreement range ($|\text{MFB}| < 0.6$) and strong correlation of Kendall's $\tau \approx 0.88$ for various locations around the globe, at various altitudes in the troposphere, and across a varied range of [CCN0.4] magnitudes. We also demonstrate application of this technique for deriving [CCN0.4] from measurements of [CCN] at other supersaturations. For a location in the Southern Great Plains region of the United States, using real measurements as input to the RFRM demonstrates its applicability. To use the measurement data as input to the RFRM required its tweaking to account for unavailable measurements for certain predictors. The truncated RFRM performs robustly despite these adjustments: median (median-absolute-deviation) mean fractional bias (MFB) of $-18(38)\%$ with 80.30% of the derived [CCN0.4] in good agreement and strong correlation of Kendall's $\tau \approx 0.79$. Specifically for the ARM SGP site, using measured predictors as input to the RFRM and comparison with measured [CCN0.4], the median (median-absolute-deviation) mean fractional bias (MFB) is $-6(61)\%$ with 67.02% of the derived [CCN0.4] in good agreement and Kendall's correlation coefficient $\tau \approx 0.36$.

4.2 Further discussion and outlook

There are a number of limitations in the application of the present study to augment empirical measurements of [CCN]. These are as follows:

- The RFRM is trained on GCAPM, with the assumption that the physical and chemical processes that relate the 'predictor' variables to the [CCN0.4] outcome are accurate. Previous studies show that GCAPM performs reasonably when compared to observations, but uncertainties in both model and observation may contribute to uncertainties in the RFRM derivation.

- Co-located and simultaneous (with [CCN0.4] measurements) measurements of the required 'predictor' variables were available only for certain predictors. We had to retrain the RFRM to account for these constraints, which sacrificed its accuracy in deriving [CCN0.4]. Other issues with measurements arise from the limitations of their accuracy, precision, and detection limits. Further sources of error are the utilisation of 'predictor' measurements from nearby (but not co-located) monitors to fill in significant gaps in the required data for the SGP site.

- Random Forest was the machine learning tool of choice due to its parallelisability and high degree of accuracy. There are, however, other tools such as XGBoost (a choice for many winners of machine learning competitions) or regression

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020


Atmospheric Chemistry and Physics Discussions

neural networks. These, among others, could offer improved [CCN0.4] derivation. A cursory examination for the present study, however, showed no significant improvement at the cost of much higher computational expense.

- To overcome some of the gaps in the [CCN0.4] measurement data from the SGP site, we proposed and implemented a derivation of [CCN0.4] from [CCN] measurements at other supersaturations using the Random Forest technique. While this derivation was confirmed to be exceptionally good, this is an approximation.
- Development of an observation-based RFRM is presented in this study. However, it can be significantly improved with more observations for the SGP site, and generalisable if trained with observations from numerous other sites. This was presently not possible.

Despite the caveats associated with this work, this proof-of-concept shows promise for wide-ranging development and deployment. This machine learning approach can provide improved representation of cloud condensation nuclei numbers:

- in locations where their direct measurements are limited, but measurements of other atmospheric state and composition variables are available. Typically, since measurements of PM2.5 speciation, trace gases, and meteorology are easier than those of [CCN0.4], there are longer and more widely (spatially) distributed in situ measurement records, especially through air quality monitoring networks. The developed RFRM can derive [CCN0.4] from these ubiquitous measurements to complement [CCN0.4] measurements when available and fill in the gaps in their absence.
- in earth system models:
  - by providing a more accurate alternative to bulk microphysical parameterisations
  - by providing a computationally less intensive alternative to explicit bin-resolving microphysics models

This work is an initial step towards fast and accurate derivation of [CCN0.4], in the absence of their measurements, constrained by empirical data for other measurements of atmospheric state and composition. This work demonstrates the possible applications of machine learning tools in handling the complex, non-linear, ordinal, and large amounts of data in the atmospheric sciences.

Code and data availability. Data for training the RFRM are generated by GEOS-Chem, a grass-roots open-access model available at http://acmg.seas.harvard.edu/geos/, last accessed on 20/05/2020. Model training and analysis of predictions in this study are though the free software environment for statistical computing and graphics: R (https://www.r-project.org/) and facilitated mainly by the 'ranger' package (https://cran.r-project.org/web/packages/ranger/index.html). Data were obtained from the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science User Facility managed by the Biological and Environmental Research program, which is publicly available at the ARM Discovery Data Portal at https://www.archive.arm.gov/discovery/, last accessed on 20/05/2020 and the EPA AirData Portal at https://www.epa.gov/air-data/, last accessed on 20/05/2020

.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

Author contributions. The authors contributed equally to this manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. This study was supported by the NSF under grant 1550816, NASA under grant NNX17AG35G, and NYSERDA under contract 137487. We thank the DOE ARM SGP Research Facility and EPA Air Quality Analysis Group data teams for the operation and maintenance of instruments, quality checks, and making this measurement data publicly available.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Atmospheric Chemistry and Physics Discussions

References

Albrecht, B. A.: Aerosols, Cloud Microphysics, and Fractional Cloudiness, Science, 245, 1227–1230, https://doi.org/10.1126/science.245.4923.1227, 1989.

Behrens, B., Salwen, C., Springston, S., and Watson, T.: ARM: AOS: aerosol chemical speciation monitor, https://doi.org/10.5439/1046180, 1990.

Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., Li, Q., Liu, H. Y., Mickley, L. J., and Schultz, M. G.: Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation, Journal of Geophysical Research: Atmospheres, 106, 23 073–23 095, https://doi.org/10.1029/2001JD000807, 2001.

Breiman, L.: Random forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Chen, X. and Xie, S.: ARM: ARMBE: Atmospheric measurements, https://doi.org/10.5439/1095313, 1994.

Christopoulos, C. D., Garimella, S., Zawadowicz, M. A., Möhler, O., and Cziczo, D. J.: A machine learning approach to aerosol classification for single-particle mass spectrometry, Atmospheric Measurement Techniques, 11, 5687–5699, https://doi.org/10.5194/amt-11-5687-2018, 2018.

Dou, X. and Yang, Y.: Comprehensive evaluation of machine learning techniques for estimating the responses of carbon fluxes to climatic forces in different terrestrial ecosystems, Atmosphere, 9, 83, https://doi.org/10.3390/atmos9030083, 2018.

Evans, M. and Jacob, D. J.: Impact of new laboratory studies of N2O5 hydrolysis on global model budgets of tropospheric nitrogen oxides, ozone, and OH, Geophysical Research Letters, 32, https://doi.org/10.1029/2005GL022469, 2005.

Filzmoser, P., Fritz, H., and Kalcher, K.: pcaPP: Robust PCA by Projection Pursuit, https://CRAN.R-project.org/package=pcaPP, r package version 1.9-73, 2018.

Fountoukis, C. and Nenes, A.: ISORROPIA II: a computationally efficient thermodynamic equilibrium model for $K+–Ca2+–Mg2+–NH_4+–Na+–SO_4 2−–NO_3−–Cl−–H_2O$ aerosols, Atmospheric Chemistry and Physics, 7, 4639–4659, https://doi.org/10.5194/acp-7-4639-2007, 2007.

Fuchs, J., Cermak, J., and Andersen, H.: Building a cloud in the southeast Atlantic: understanding low-cloud controls based on satellite observations with machine learning, Atmospheric Chemistry and Physics, 18, 16 537–16 552, https://doi.org/10.5194/acp-18-16537-2018, 2018.

Giglio, L., Randerson, J. T., and van der Werf, G. R.: Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4), Journal of Geophysical Research: Biogeosciences, 118, 317–328, https://doi.org/10.1002/jgrg.20042, 2013.

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., and Hueglin, C.: Random forest meteorological normalisation models for Swiss PM_{10} trend analysis, Atmospheric Chemistry and Physics, 18, 6223–6239, https://doi.org/10.5194/acp-18-6223-2018, 2018.

Guenther, A. B., Jiang, X., Heald, C. L., Sakulyanontvittaya, T., Duhl, T., Emmons, L. K., and Wang, X.: The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions, Geoscientific Model Development, 5, 1471–1492, https://doi.org/10.5194/gmd-5-1471-2012, 2012.

Hageman, D., Behrens, B., Smith, S., Uin, J., Salwen, C., Koontz, A., Jefferson, A., Watson, T., Sedlacek, A., Kuang, C., Dubey, M., Springston, S., and Senum, G.: ARM: Aerosol Observing System (AOS): aerosol data, 1-min, mentor-QC applied, https://doi.org/10.5439/1025259, 1996.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Hageman, D., Behrens, B., Smith, S., Uin, J., Salwen, C., Koontz, A., Jefferson, A., Watson, T., Sedlacek, A., Kuang, C., Dubey, M., Springston, S., and Senum, G.: ARM: Aerosol Observing System (AOS): cloud condensation nuclei data, https://doi.org/10.5439/1150249, 2017.

Holdridge, D. and Kyrouac, J.: ARM: ARM-standard Meteorological Instrumentation at Surface, https://doi.org/10.5439/1025220, 1993.

Hoppel, W. A., Frick, G. M., Fitzgerald, J. W., and Wattle, B. J.: A Cloud Chamber Study of the Effect That Nonprecipitating Water Clouds Have on the Aerosol Size Distribution, Aerosol Science and Technology, 20, 1–30, https://doi.org/10.1080/02786829408959660, 1994.

Hughes, M., Kodros, J., Pierce, J., West, M., and Riemer, N.: Machine Learning to Predict the Global Distribution of Aerosol Mixing State Metrics, Atmosphere, 9, 15, https://doi.org/10.3390/atmos9010015, 2018.

Huttunen, J., Kokkola, H., Mielonen, T., Mononen, M. E. J., Lipponen, A., Reunanen, J., Lindfors, A. V., Mikkonen, S., Lehtinen, K. E. J., Kouremeti, N., Bais, A., Niska, H., and Arola, A.: Retrieval of aerosol optical depth from surface solar radiation measurements using machine learning algorithms, non-linear regression and a radiative transfer-based look-up table, Atmospheric Chemistry and Physics, 16, 8181–8191, https://doi.org/10.5194/acp-16-8181-2016, 2016.

IPCC AR5: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Jefferson, A.: Aerosol Observing System (AOS) Handbook, Tech. rep., DOE Office of Science Atmospheric Radiation Measurement (ARM) Program, https://doi.org/10.2172/1020729, 2011.

Jin, J., Lin, H. X., Segers, A., Xie, Y., and Heemink, A.: Machine learning for observation bias correction with application to dust storm data assimilation, Atmospheric Chemistry and Physics, 19, 10 009–10 026, https://doi.org/10.5194/acp-19-10009-2019, 2019.

Joutsensaari, J., Ozon, M., Nieminen, T., Mikkonen, S., Lähivaara, T., Decesari, S., Facchini, M. C., Laaksonen, A., and Lehtinen, K. E. J.: Identification of new particle formation events with deep learning, Atmospheric Chemistry and Physics, 18, 9597–9615, https://doi.org/10.5194/acp-18-9597-2018, 2018.

Keller, C. A., Long, M. S., Yantosca, R. M., Da Silva, A. M., Pawson, S., and Jacob, D. J.: HEMCO v1.0: a versatile, ESMF-compliant component for calculating emissions in atmospheric models, Geoscientific Model Development, 7, 1409–1417, https://doi.org/10.5194/gmd-7-1409-2014, 2014.

Kendall, M.: Rank Correlation Methods, Theory and applications of rank order-statistics, Griffin, 1970.

Kulkarni, G.: aosacsm.b1, https://doi.org/10.5439/1558768, 2019.

Martin, R. V., Jacob, D. J., Yantosca, R. M., Chin, M., and Ginoux, P.: Global and regional decreases in tropospheric oxidants from photochemical effects of aerosols, Journal of Geophysical Research: Atmospheres, 108, n/a–n/a, https://doi.org/10.1029/2002jd002622, 2003.

Mauceri, S., Kindel, B., Massie, S., and Pilewskie, P.: Neural network for aerosol retrieval from hyperspectral imagery, Atmospheric Measurement Techniques, 12, 6017–6036, https://doi.org/10.5194/amt-12-6017-2019, 2019.

McLeod, A.: Kendall: Kendall rank correlation and Mann-Kendall trend test, https://CRAN.R-project.org/package=Kendall, r package version 2.2, 2011.

Merikanto, J., Spracklen, D. V., Mann, G. W., Pickering, S. J., and Carslaw, K. S.: Impact of nucleation on global CCN, Atmospheric Chemistry and Physics, 9, 8601–8616, https://doi.org/10.5194/acp-9-8601-2009, 2009.

Murray, L. T., Jacob, D. J., Logan, J. A., Hudman, R. C., and Koshak, W. J.: Optimized regional and interannual variability of lightning in a global chemical transport model constrained by LIS/OTD satellite data, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2012jd017934, 2012.

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Nair, A. A., Yu, F., and Luo, G.: Spatioseasonal Variations of Atmospheric Ammonia Concentrations Over the United States: Comprehensive Model-Observation Comparison, Journal of Geophysical Research: Atmospheres, 124, 6571–6582, https://doi.org/10.1029/2018JD030057, 2019.

Ng, N. L., Herndon, S. C., Trimborn, A., Canagaratna, M. R., Croteau, P. L., Onasch, T. B., Sueper, D., Worsnop, D. R., Zhang, Q., Sun, Y. L., and Jayne, J. T.: An Aerosol Chemical Speciation Monitor (ACSM) for Routine Monitoring of the Composition and Mass Concentrations of Ambient Aerosol, Aerosol Science and Technology, 45, 780–794, https://doi.org/10.1080/02786826.2011.560211, 2011.

Noether, G. E.: Why Kendall Tau?, Teaching Statistics, 3, 41–43, https://doi.org/10.1111/j.1467-9639.1981.tb00422.x, 1981.

Okamura, R., Iwabuchi, H., and Schmidt, K. S.: Feasibility study of multi-pixel retrieval of optical thickness and droplet effective radius of inhomogeneous clouds using deep learning, Atmospheric Measurement Techniques, 10, 4747–4759, https://doi.org/10.5194/amt-10-4747-2017, 2017.

Park, R. J.: Natural and transboundary pollution influences on sulfate-nitrate-ammonium aerosols in the United States: Implications for policy, Journal of Geophysical Research, 109, https://doi.org/10.1029/2003jd004473, 2004.

Pye, H. O. T. and Seinfeld, J. H.: A global perspective on aerosol from low-volatility organic compounds, Atmospheric Chemistry and Physics, 10, 4377–4401, https://doi.org/10.5194/acp-10-4377-2010, 2010.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2020.

Ritsche, M.: ARM Surface Meteorology Systems Instrument Handbook, Tech. rep., Office of Scientific and Technical Information (OSTI), https://doi.org/10.2172/1007926, 2011.

Roberts, G. C. and Nenes, A.: A Continuous-Flow Streamwise Thermal-Gradient CCN Chamber for Atmospheric Measurements, Aerosol Science and Technology, 39, 206–221, https://doi.org/10.1080/027868290913988, 2005.

Seinfeld, J. H., Bretherton, C., Carslaw, K. S., Coe, H., DeMott, P. J., Dunlea, E. J., Feingold, G., Ghan, S., Guenther, A. B., Kahn, R., Kraucunas, I., Kreidenweis, S. M., Molina, M. J., Nenes, A., Penner, J. E., Prather, K. A., Ramanathan, V., Ramaswamy, V., Rasch, P. J., Ravishankara, A. R., Rosenfeld, D., Stephens, G., and Wood, R.: Improving our fundamental understanding of the role of aerosol-cloud interactions in the climate system, Proceedings of the National Academy of Sciences, 113, 5781–5790, https://doi.org/10.1073/pnas.1514043113, 2016.

Shi, Y. and Flynn, C.: ARM: Aerosol Observing System (AOS): cloud condensation nuclei data, averaged, https://doi.org/10.5439/1095312, 2007.

Smith, S., Salwen, C., Uin, J., Senum, G., Springston, S., and Jefferson, A.: ARM: AOS: Cloud Condensation Nuclei Counter, https://doi.org/10.5439/1256093, 2011a.

Smith, S., Salwen, C., Uin, J., Senum, G., Springston, S., and Jefferson, A.: ARM: AOS: Cloud Condensation Nuclei Counter (Single Column), averaged, https://doi.org/10.5439/1342133, 2011b.

Springston, S.: ARM: AOS: Sulfur Dioxide Analyzer, https://doi.org/10.5439/1095586, 2012.

Twomey, S. A.: The Influence of Pollution on the Shortwave Albedo of Clouds, Journal of the Atmospheric Sciences, 34, 1149–1152, https://doi.org/10.1175/1520-0469(1977)034<1149:TIOPOT>2.0.CO;2, 1977.

Uin, J.: Cloud Condensation Nuclei Particle Counter (CCN) Instrument Handbook, Tech. rep., DOE Office of Science Atmospheric Radiation Measurement (ARM) Program, https://doi.org/10.2172/1251411, 2016.

van Donkelaar, A., Martin, R. V., Leaitch, W. R., Macdonald, A. M., Walker, T. W., Streets, D. G., Zhang, Q., Dunlea, E. J., Jimenez, J. L., Dibb, J. E., Huey, L. G., Weber, R., and Andreae, M. O.: Analysis of aircraft and satellite measurements from the Intercontinental

https://doi.org/10.5194/acp-2020-509
Preprint. Discussion started: 10 June 2020

Chemical Transport Experiment (INTEX-B) to quantify long-range transport of East Asian sulfur to Canada, Atmospheric Chemistry and Physics, 8, 2999–3014, https://doi.org/10.5194/acp-8-2999-2008, 2008.

Watson, T., , Aiken, A., Zhang, Q., Croteau, P., Onasch, T., Williams, L., and and, C. F.: First ARM Aerosol Chemical Speciation Monitor Users' Meeting Report, Tech. rep., DOE Office of Science Atmospheric Radiation Measurement (ARM) Program, https://doi.org/10.2172/1455055, 2018.

Watson, T. B.: Aerosol Chemical Speciation Monitor (ACSM) Instrument Handbook, Tech. rep., DOE Office of Science Atmospheric Radiation Measurement (ARM) Program, https://doi.org/10.2172/1375336, 2017.

Wright, M. N. and Ziegler, A.: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, Journal of Statistical Software, 77, https://doi.org/10.18637/jss.v077.i01, 2017.

Yu, F.: A secondary organic aerosol formation model considering successive oxidation aging and kinetic condensation of organic compounds: global scale implications, Atmospheric Chemistry and Physics, 11, 1083–1099, https://doi.org/10.5194/acp-11-1083-2011, 2011.

Yu, F. and Luo, G.: Simulation of particle size distribution with a global aerosol model: contribution of nucleation to aerosol and CCN number concentrations, Atmospheric Chemistry and Physics, 9, 7691–7710, https://doi.org/10.5194/acp-9-7691-2009, 2009.

Yu, F., Ma, X., and Luo, G.: Anthropogenic contribution to cloud condensation nuclei and the first aerosol indirect climate effect, Environmental Research Letters, 8, 024 029, https://doi.org/10.1088/1748-9326/8/2/024029, 2013.

Yu, F., Luo, G., Nadykto, A. B., and Herb, J.: Impact of temperature dependence on the possible contribution of organics to new particle formation in the atmosphere, Atmospheric Chemistry and Physics, 17, 4997–5005, https://doi.org/10.5194/acp-17-4997-2017, 2017.

Yu, F., Nadykto, A. B., Herb, J., Luo, G., Nazarenko, K. M., and Uvarova, L. A.: H_2SO_4-H_2O-NH_3 ternary ion-mediated nucleation (TIMN): Kinetic-based model and comparison with CLOUD measurements, Atmospheric Chemistry and Physics, 18, 17 451–17 474, https://doi.org/10.5194/acp-2018-396, 2018.

Zaidan, M. A., Haapasilta, V., Relan, R., Junninen, H., Aalto, P. P., Kulmala, M., Laurson, L., and Foster, A. S.: Predicting atmospheric particle formation days by Bayesian classification of the time series features, Tellus B: Chemical and Physical Meteorology, 70, 1–10, https://doi.org/10.1080/16000889.2018.1530031, 2018.