


# Geophysical Research Letters®



## RESEARCH LETTER

10.1029/2024GL110757

## On the Prediction of Aerosol-Cloud Interactions Within a Data-Driven Framework

Xiang-Yu Li<sup>1</sup> , Hailong Wang<sup>1</sup> , TC Chakraborty<sup>1</sup>, Armin Sorooshian<sup>2</sup>, Luke D. Ziemba<sup>3</sup> , Christiane Voigt<sup>4</sup> , Kenneth Lee Thornhill<sup>3</sup> , and Emma Yuan<sup>1,5</sup>

<sup>1</sup>Pacific Northwest National Laboratory, Richland, WA, USA, <sup>2</sup>Department of Chemical and Environmental Engineering and Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, USA, <sup>3</sup>NASA Langley Research Center, Hampton, VA, USA, <sup>4</sup>Institut für Physik der Atmosphäre, Deutsches Zentrum für Luft- und Raumfahrt (DLR), Oberpfaffenhofen, Germany, and Institute for Physics of the Atmosphere, Johannes Gutenberg-University Mainz, Wessling, Germany, <sup>5</sup>Hanford High School, Richland, WA, USA

### Key Points:

- Three-year in situ measurements (179 flights) provide adequate data to train and validate a random forest model (RFM) to study aerosol-cloud interactions
- The RFM can successfully predict cloud droplet number concentration  $N_c$  and identify importance of key predictors
- Data-driven  $N_c$  prediction in individual cases shows strong dependency on sampling strategy

### Supporting Information:

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

### Correspondence to:

X.-Y. Li and H. Wang,  
xiangyu.li@pnnl.gov;  
hailong.wang@pnnl.gov

### Citation:

Li, X.-Y., Wang, H., Chakraborty, T.C., Sorooshian, A., Ziemba, L. D., Voigt, C., et al. (2024). On the prediction of aerosol-cloud interactions within a data-driven framework. *Geophysical Research Letters*, 51, e2024GL110757. <https://doi.org/10.1029/2024GL110757>

Received 13 JUN 2024

Accepted 14 OCT 2024

**Abstract** Aerosol-cloud interactions (ACI) pose the largest uncertainty for climate projection. Among many challenges of understanding ACI, the question of whether ACI can be deterministically predicted has not been explicitly answered. Here we attempt to answer this question by predicting cloud droplet number concentration  $N_c$  from aerosol number concentration  $N_a$  and ambient conditions using a data-driven framework. We use aerosol properties, vertical velocity fluctuations, and meteorological states from the ACTIVATE field observations (2020–2022) as predictors to estimate  $N_c$ . We show that the campaign-wide  $N_c$  can be successfully predicted using machine learning models despite the strongly nonlinear and multi-scale nature of ACI. However, the observation-trained machine learning model fails to predict  $N_c$  in individual cases while it successfully predicts  $N_c$  of randomly selected data points that cover a broad spatiotemporal scale. This suggests that, within a data-driven framework, the  $N_c$  prediction is uncertain at fine spatiotemporal scales.

**Plain Language Summary** Ambient aerosol particles act as seeds for ice crystals and cloud droplets that form clouds. Both aerosols and clouds regulate the energy and water budgets of the Earth via radiative and cloud micro/macro-processes. This is the so-called aerosol-cloud interactions (ACI). ACI remains the source of the largest uncertainty for accurate climate projections, due to incomplete understanding of nonlinear multi-scale processes, limited observations across various cloud regimes, and insufficient computational power to resolve them in models. Quantifying the relation between the cloud droplet ( $N_c$ ) and aerosol ( $N_a$ ) number concentration has been a central challenge of understanding and representing ACI. In this work, we tackle this challenge by predicting  $N_c$  from observations made during the Aerosol Cloud meTeorology Interactions oVer the western ATlantic Experiment (ACTIVATE) using machine learning models. We show that the climatological  $N_c$  can be successfully predicted despite the strongly nonlinear and multi-scale nature of ACI. However, the observation-trained machine learning model fails to predict  $N_c$  at fine spatiotemporal scales.

## 1. Introduction

Atmospheric aerosols regulate Earth's energy budget directly via scattering or absorbing solar radiation (Bellouin et al., 2020; Seinfeld et al., 2016) and indirectly via acting as the seeds of cloud droplets, through which aerosols can alter cloud microphysical and macrophysical properties (Albrecht, 1989; Twomey, 1974). Clouds modulate Earth's energy budget and water cycle (including precipitation), which, in turn, affect the sink, source, and transport processes of aerosols. This interplay between aerosols and clouds is the so-called aerosol-cloud interactions (ACI). ACIs remain the largest source of uncertainties in numerical models for accurate climate projections (Bellouin et al., 2020; Bock et al., 2020; Ghan et al., 2016; Seinfeld et al., 2016) due to poor understanding of the governing processes, scarce observations, and limited computational power to resolve ACIs at native scales in numerical models. ACIs are strongly nonlinear and involve multi-scale processes with nm-sized aerosols, km-sized clouds, and hundreds of km-sized weather systems and large-scale circulations. Simulating ACIs over such a wide scale range at the native scales of all the physical processes is intractable. In addition, our current physical understanding of these physical processes is incomplete. For example, the physical mechanism of the collision-coalescence of particles that is critical for precipitation and aerosol budget is still not fully understood (Grabowski & Wang, 2013; Li et al., 2020).

© 2024 Battelle Memorial Institute and The Author(s). This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs License](https://creativecommons.org/licenses/by/4.0/), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Among many challenges of understanding and quantifying ACI, the question of whether ACI can be deterministically predicted has *not* been explicitly formulated or asked. It is possible that, in the context of our state-of-the-art measurement techniques and modeling tools, the intrinsically chaotic and stochastic characteristics of the atmospheric systems dominate the quantification of ACI. Here, stochasticity is a relative concept that largely depends on the entropy of a system, that is, how much we know about a system (Boltzmann, 2022). Ideally, ACI is deterministic in nature, that is, for a given ambient condition, if we can predict the spatiotemporal evolution of aerosol and droplet size distributions, including cloud-nucleating attributes of aerosols, we would be able to determine the macroscopic response of clouds (e.g., liquid water path and cloud fraction adjustment) to microscopic perturbations (e.g., aerosol number concentration). Unfortunately, predicting the spatiotemporal evolution of aerosol and droplet size distributions at the native process-level scales from first principles is beyond the horizon. An intermediate approach is to determine if ACI is deterministic or stochastic at certain spatiotemporal scales within our current understanding of ACI and available physics-based or data-driven models. Both the condensation and the collision-coalescence processes of cloud droplets are shown to be stochastic (Li, 2018; Li, Mehlig, et al., 2022) at their native spatiotemporal scales. However, these studies only focus on part of ACI pathways and their stochasticity cannot be generalized to ACI for the following reasons: 1. Interactions among these cloud processes and aerosol processes are missing; 2. The mean state of ACI is not represented because of limited spatiotemporal scales.

The physics behind ACI is far from being fully explored and understood, so existing physics-based models established upon governing equations for known physics are unexpected to capture all observational characteristics of ACI, not even mentioning the unresolved part of known physics due to computational constraints. We now face the longstanding dilemma that considering all the known physical interactions across scales from nm-sized aerosols to hundreds of km-sized circulations is not feasible and focusing on part of the ACI pathways is incomplete. To tackle this dilemma and examine the seemingly stochastic prediction (or indeterminable part in physics-based models) of ACI, instead of pursuing the causality behind ACI at limited scales, here we focus on a phenomenological description of ACI based on observations and data-driven models. Specifically, we use the models to predict the droplet number concentration  $N_c$  from observed aerosol number concentration  $N_a$ , chemical components of aerosol particles, ambient thermodynamics, and turbulence, as these factors contribute the most to ACI metrics according to our current scientific understanding of ACI. This is done by predicting the  $N_c$  from the opportune data set afforded by Aerosol Cloud meteorology Interactions over the western ATLantic Experiment (ACTIVATE) field measurements (Sorooshian et al., 2019) using a data-driven random forest regression model (RFM) (Breiman, 2001).

For context, we first provide a brief summary of relevant ACI results in terms of  $N_a$ - $N_c$  relation for the ACTIVATE region. Dadashazar et al. (2021) showed that  $N_c$  peaks in winter in contrast to  $N_a$  that peaks in summer, due to stronger ACI in winter such that for a given number of aerosol particles more can activate into cloud droplets. Subsequent studies showed that seasonally, updraft velocities and turbulence (i.e., dynamics driving  $N_a$  activation) are generally stronger in winter (Brunke et al., 2022), but that microphysical/chemical attributes may be more important within a season. Also, the susceptibility of  $N_c$  to  $N_a$  is suspected to be stronger farther offshore of the U.S. East Coast over more remote oceans (Sorooshian et al., 2019). Processes-level modeling studies of ACI for individual cases using large-eddy simulations have shown case-dependent challenges and uncertainties in predicting  $N_c$  (Li et al., 2023, 2024) from available  $N_a$  measurements along with information of aerosol chemical attributes and meteorological conditions, which motivates us to study ACI over a wider spatiotemporal range than individual cases.

## 2. Methods

### 2.1. Observation as Training and Validation Data for the RFM

To study ACIs in marine boundary-layer clouds, 179 research flights were carried out between 2020 and 2022 using a dual-aircraft approach during the ACTIVATE campaign over the Western North Atlantic Ocean (WNAO) region (25° – 50°N, 60° – 85°W). The WNAO region is characterized by large natural and anthropogenic aerosol variability, diverse meteorological conditions, and different low-cloud regimes, which is ideal for studying ACI and for collecting unprecedented observations of aerosols, clouds, and meteorological states (Corral et al., 2021; Painemal et al., 2021). ACTIVATE's low-flying Falcon HU25 sampled vertical profiles by performing below cloud-base (BCB), above cloud-base (ACB), below cloud-top (BCT), and above cloud-top (ACT) flight legs, that

is, using either a stairstepping flight strategy or “wall” strategy involving stacked level legs (Sorooshian et al., 2023). In situ aerosol properties and cloud microphysical properties were measured during the BCB and ACB/BCT flight legs, respectively. In this study, we use all the in situ measurements of aerosol and cloud properties, turbulence, and thermodynamics during the ACTIVATE campaign.

Aerosol particles with diameter between 3 – 100 nm and larger than 100 nm BCB were measured by a Scanning Mobility Particle Sizer (SMPS) and a Laser Aerosol Spectrometer (LAS), respectively (Moore et al., 2021). Mass concentrations of major aerosol chemical components (e.g., sulfate, nitrate, organics, ammonium, chloride) are measured by an Aerodyne High Resolution Time-of-Flight Aerosol Mass Spectrometer (HR-ToF-AMS) (DeCarlo et al., 2008). The cloud microphysical properties (e.g.,  $N_c$ ) were measured by the fast cloud droplet probe (FCDP) for cloud droplets with diameter ranging from 3 – 50  $\mu\text{m}$  (Kirschler et al., 2022, 2023). A list of data and the corresponding instruments are shown in Table S1 in Supporting Information S1. Clouds are defined using the threshold of  $N_c \geq 20 \text{ cm}^{-3}$ ,  $\text{LWC} \geq 0.02 \text{ g m}^{-3}$ , and effective diameter  $d_{\text{eff}} \geq 3.5 \mu\text{m}$ . A more comprehensive description and discussion on ACTIVATE measurements and instrument details are provided by Sorooshian et al. (2023).

As we aim to predict  $N_c$ , all measurements are synced to FCDP- $N_c$  measurements. The wind speed measurements are synced to FCDP- $N_c$  measurements spatiotemporally by averaging them around each FCDP- $N_c$  data point at a given time over a window size of 20 (data points) as the sampling rate of the wind speed measurements and reported FCDP data is 20 and 1 Hz, respectively. Syncing  $N_a$  and non-refractory mass concentration ( $m_\chi$ ) of aerosol chemical components to FCDP- $N_c$  measurements are more challenging because they were not strictly collocated at the native sampling frequency (i.e., aerosols sampled during BCB flight legs and cloud microphysics measured during ACB or BCT flight legs). We overcome this challenge by averaging  $N_a$  and  $m_\chi$  over all BCB legs of each flight, assuming that aerosols BCB are representative of sampling region. Other than the collocation reasoning, this syncing strategy is justified by the fact that aerosol measurements (AMS, SMPS, and LAS) at each BCB flight leg only lasted for  $\sim 3$  minutes for a flight ( $\sim 30$  minutes for a entire flight with  $\sim 10$  BCB sampling legs; see Figure S1 in Supporting Information S1). Even though this syncing strategy ignores aerosol transport during BCB-leg sampling, the statistical properties of aerosols over the measurement domain are well represented. The vertical velocity fluctuation  $w' = w - \overline{w}$  (the same for  $u'$  and  $v'$ ) is calculated from the native 20 Hz data for each flight leg, where  $\overline{w}$  is obtained. Overall, the training and validation data for the machine learning (ML) model are based on subsets of a sample size of 69,159 with a sampling rate of 1 Hz.

Another challenge of using the in situ aircraft measurements as the training and validation data for the ML model is to find a physical spatial scale that can represent the characteristic scales of aerosols and cloud droplets of the targeted cloud systems. This is because the aircraft performs measurements of aerosol and cloud droplet properties with a speed of about  $100 \text{ m s}^{-1}$ . The 1-Hz data corresponds to a length scale of 100 m, which is too small to represent typical length scales of boundary-layer cloud systems. To work around this issue, we perform a running-average of the input data and examine the  $r^2$  validation score as a function of the running-average window size. The  $r^2$  saturates (hits 0.99) at a window size of 20 data points (from 1-Hz data) as listed in Table S2 in Supporting Information S1 (the corresponding comparison of  $N_c^{\text{predicted}}$  and  $N_c^{\text{obs}}$  is shown in Figure S4 in Supporting Information S1). Therefore, we apply a running-average window of 20 (data points) to all the 1-Hz observational data to obtain smoothed data sets for training and validating the RFM model. More importantly, this window size corresponds to a length scale of 2 km that is more representative of the length scale of cloud systems being studied here. The 20-s running average could lead to data leakage and the resulting overfitting of RFM.

## 2.2. Random Forest Model

One of the objectives of this study is to predict  $N_c$  from in situ measurements of aerosol ( $N_a$  and  $m_\chi$ ), turbulence ( $u'$ ,  $v'$ , and  $w'$ ), and thermodynamics ( $T$  and  $q_v$ ), that is, to construct a function representing

$$N_c = \mathcal{G}(m_\chi, N_a, w', u', v', T, q_v, \mathbf{x}, \theta_z). \quad (1)$$

here  $\mathbf{x}$  and  $\theta_z$  denote the location (latitude, longitude, and altitude) and zenith angle, respectively. The random forest model (RFM) is chosen to achieve this because it is effective in prediction (Breiman, 2001). In addition, the RFM has the advantage of accurate prediction of nonlinear complex systems and easy implementation and

physical interpretation of the prediction compared to other machine learning methods (e.g., neural network or deep learning techniques) (Breiman, 2001). It is also fast due to the parallelizability in building decision trees.

To determine the feature importance of predictors for  $N_c$ , we adopt the permutation feature importance (PFI) technique. PFI measures the contribution of each feature to a fitted model's statistical performance on a given tabular data set. This technique is particularly useful for nonlinear estimators, and involves randomly shuffling the values of a single feature and observing the resulting degradation of the model's score. We note PFI *does not* reflect the physical importance of a feature to a complex system but reflects how important this feature is for a particular data-driven model (Breiman, 2001; Pedregosa et al., 2011). It is, however, important to place the PFI based on physical understanding of a complex system, to better understand what the model is really capturing, as we will discuss for the ACI studied here. We cross validate the permutation importance using the K-fold (10 fold) and Monte-Carlo (50 random sampling) validation method. The relative root mean square error (RRMSE) and mean normalized error (MNE) are adopted to evaluate the predictions.

RFM has been widely used in Earth Systems and shown to be a robust tool for prediction and inference (Arjunan Nair & Yu, 2020; Chakraborty & Lee, 2021; Chen et al., 2022; Dadashazar et al., 2021; Michel et al., 2022). Here we limit our discussion on the application of RFM to ACI. Arjunan Nair and Yu (2020) showed that RFM is highly robust in predicting number concentration of cloud condensation nuclei (CCN) with the atmospheric state and composition variables as predictors from a global chemical transport model and can learn the underlying dependence of CCN on these predictors. Subsequent works further show that RFM can learn aerosol size information (Nair et al., 2021) and can reduce uncertainties of climate models in predicting particle number concentration and radiative forcing associated with ACI (Yu et al., 2022). We use the RFM implemented in open-source scikit-learn (Pedregosa et al., 2011).

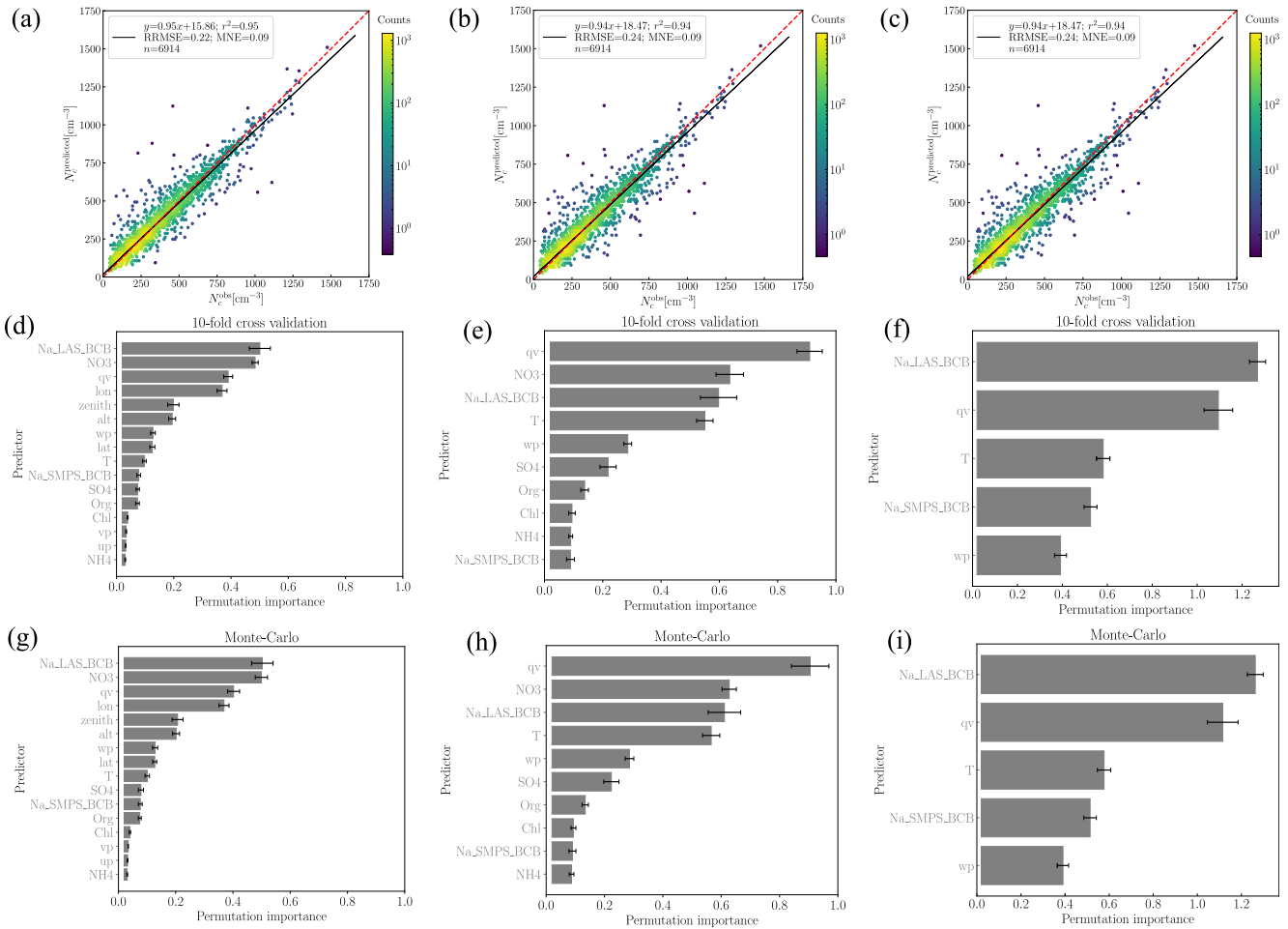
The coefficient of determination,  $r^2$  score, converges at hyperparameters maximum tree depth = 30, number of trees = 98, and test size = 0.1 (i.e., 10% data for the validation and 90% data for the training), which are used for all the training in this study.

### 3. Results

#### 3.1. A Successful Data-Driven Prediction of $N_c$

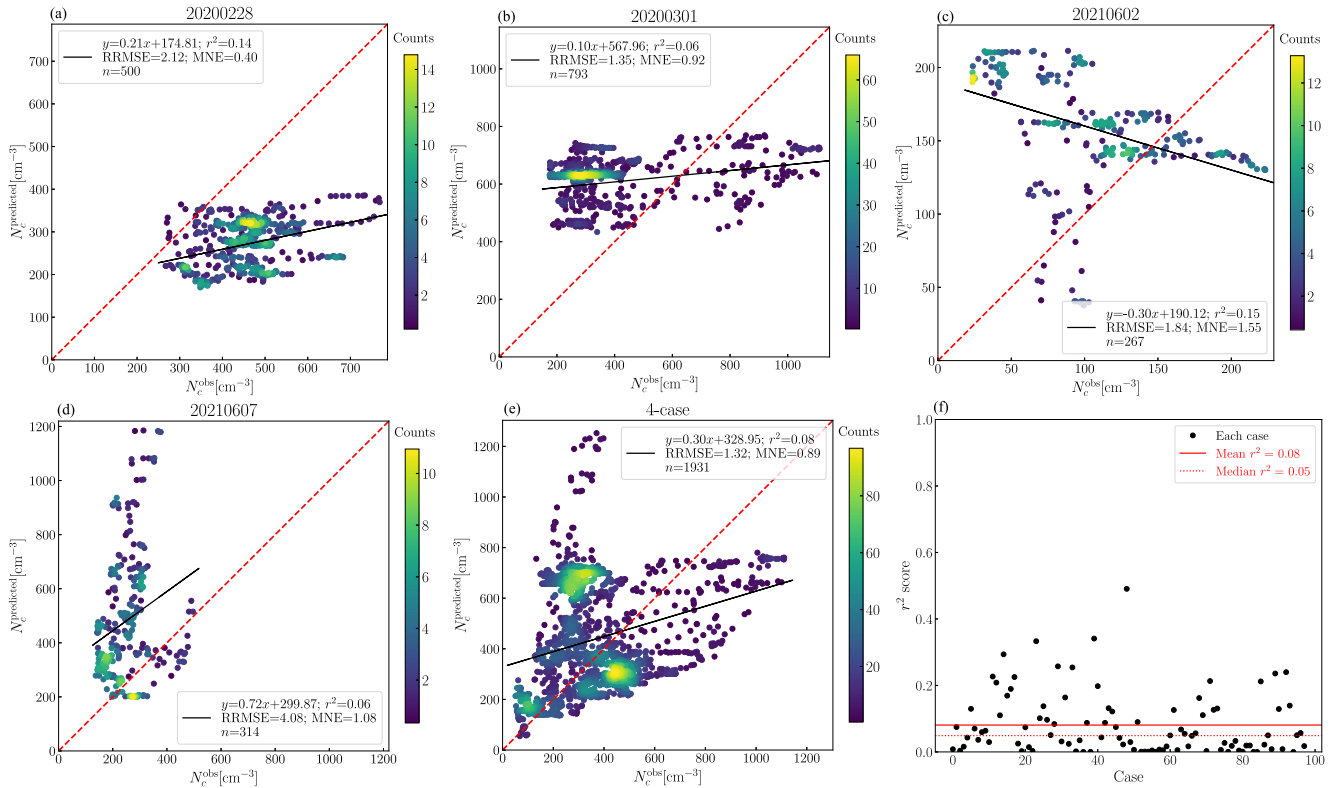
To quantify the ACI over the WNAO region climatologically, we start with predicting  $N_c$  using all physically related measurements from the ACTIVATE campaign. The RFM can successfully predict  $N_c$  from the predictors  $m_x, N_a, w', u', v', T, q_v, x, \theta_z$  with  $r^2 = 0.99$  using a 20-point running-average window (see detailed discussion in Section 2.1) as shown in Figure 1a. This is remarkable considering that the  $N_c$  depends on highly nonlinear and multiscale processes. A natural question arises as to what is the relative importance of each predictor for estimating  $N_c$ . As shown in Figures 1d and 1g, the dominant predictors are number concentration of large-size ( $\geq 100$  nm) mode aerosol particles  $N_{a,LAS}$ , mass fraction of nitrate  $m_{NO_3}$ , and water vapor mixing ratio  $q_v$ . The importance of all three quantities to the  $N_c$  prediction is consistent with our physical understanding. Less straightforward is the importance of  $m_{NO_3}$ , as the dominant anthropogenic chemical component contributing to aerosol activation is  $SO_4$  conventionally. In the ACTIVATE study region, sulfate and organics are the most dominant submicron species, with the latter having more of an offshore gradient as compared to sulfate which has strong influence from ocean biogenic emissions such as dimethylsulfide even over the remote ocean. However, we note that the seasonal variation of the  $m_{NO_3}$  follows that of  $N_c$  over the WNAO region (Dadashazar et al., 2022). Even though nitrate may not be as abundant by mass as sulfate and organics in any given season over the WNAO (Dadashazar et al., 2022), it thermodynamically favors colder conditions, which is why it is the only species (vs. sulfate and organics) exhibiting higher absolute concentrations in winter as compared to other seasons (Corral et al., 2022), which may explain its strong association with  $N_c$ . We note that  $N_c$  is successfully predicted from observations at different locations of clouds where highly different physical processes affecting  $N_c$  budget are at play, in part because the vertical structure of  $N_c$  does not vary strongly within clouds. On the other hand, this also speaks to the point that the well-trained RFM has the ability to predict  $N_c$  based on spatially separated but physically related variables.

So far, we have showed that the RFM model is able to successfully predict  $N_c$  from all the available measured predictors and is able to identify top contributors to the  $N_c$  prediction, which motivates us to explore whether this holds with fewer physically motivated predictors in the spirit of dimension reduction. We first predict  $N_c$  by



**Figure 1.** Binned scatter-plot of  $N_c^{\text{obs}}$  and  $N_c^{\text{predicted}}$  from different predictors (a):  $N_c = \mathcal{G}(m_{\mathcal{X}}, N_a, w', u', v', T, q_v, \mathbf{x}, \theta_z)$ , that is, all available measurements; (b)  $N_c = \mathcal{G}(m_{\mathcal{X}}, N_a, w', T, q_v)$ , and (c)  $N_c = \mathcal{G}(N_a, w', T, q_v)$ . Color bar shows the counts of data points in each hexagonal bin. The red dashed line represents the one-to-one line. The solid black line represents the linear regression relation  $y = ax + b$  with  $a$  and  $b$  being the regression coefficients. The test size of the validation data is represented by  $n$  in the legend of the scatter-plots. Error bars in the average permutation feature importance (PFI) plots represent  $\sigma$  deviation of PFI from the 10-fold and Monte-Carlo cross validations.

removing the physically less important or covariant quantities, the horizontal wind speed  $u$  &  $v$ , geo-coordinate  $\mathbf{x}$ , and the zenith angle  $\theta_z$ , from the predictors pool. This again yields a successful prediction of  $N_c$  with  $r^2 = 0.95$  (Figure 1b) and retains  $N_{a,LAS}$ ,  $m_{NO_3}$ , and  $q_v$  as the main contributors (Figures 1e and 1h). Note that even though the latitude is an important predictor (Figure 1d) and related to the cloud development, we drop it here to focus on the well-established Köhler theory (Köhler, 1936). Ideally, the  $N_c$  would be determined by chemical components of aerosols  $m_{\mathcal{X}}$ , aerosol number concentration  $N_a$ , the vertical component of turbulence  $w'$ , and the thermodynamics ( $T$  and  $q_v$ ) according to the Köhler theory (Köhler, 1936). However, knowing the mass fraction of  $\mathcal{X}$  is challenging for numerical models and observations. We therefore drop  $m_{\mathcal{X}}$  from the predictor pool and predict  $N_c$  only from  $N_a$ ,  $T$ ,  $q_v$ , and  $w'$ . The prediction of  $N_c$  and variable importance is again remarkably successful (Figures 1c, 1f, and 1i). We further examine whether the successful prediction of  $N_c$  and the feature importance are ML model dependent by applying XGBoost to the data set. The  $N_c$  prediction and the corresponding feature importance from the XGBoost are nearly identical (Figures S10c, S10f, and S10i in Supporting Information S1) to the RFM, suggesting that our results are independent of state-of-the-art tree-based ML models. In addition, we cross validate the feature importance by adopting the Shapley additive explanations (SHAP) analysis that offers local explainability and Shapley additive global explanation (SAGE) that offers global explainability. Both SHAP and SAGE provide consistent feature importance from both RFM (Figure S14 in Supporting Information S1) and



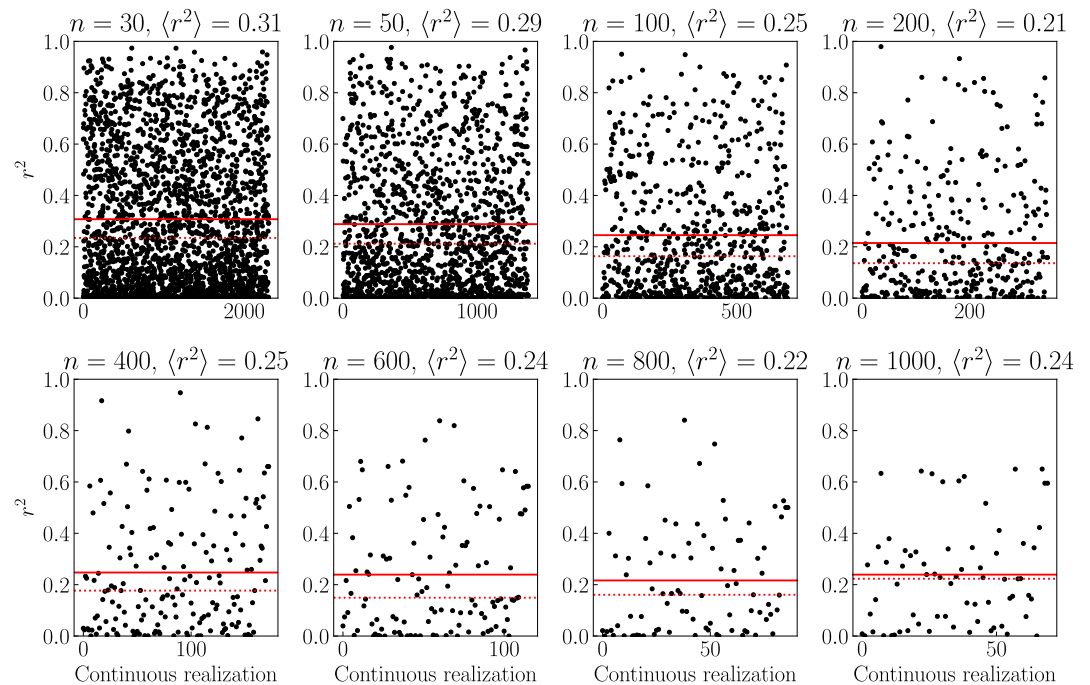
**Figure 2.**  $N_c$  predicted from the individual observed cases (01 Mar and 28 Feb 2020 and 02/07 June 2021) using the observation-RFM (no corresponding observed case in the training data set) is compared to the FCDP- $N_c$ . The same averaging strategy is applied to the validation data for each case.

XGBoost (Figure S15 in Supporting Information S1) as compared to the PFI, suggesting that  $N_a$  and  $q_v$  indeed determine the  $N_c$  prediction within our data-driven framework.

Even though the top predictors are consistent for different predictor pools and ML models, we cannot conclude the causal relation between  $N_c$  and the top predictors  $N_a$  and  $q_v$ , because the feature importance may reflect the behavior of ML models instead of the behavior of the training data set (Silva & Keller, 2023).

### 3.2. Dependence of $N_c$ Predictability on Sampling Strategy of Predictors

With the successful prediction of  $N_c$  from the observational data in hand, we can now examine how the  $N_c$  prediction is scale-dependent within the data-driven framework and its implications on our understanding of multiscale ACI processes. Recall that the observation-RFM is trained and validated using 3-year in situ measurements, which represents a statistical  $N_c$  prediction for the WNAO domain over multiple years. The scale-dependency of ACI metrics can be pursued by examining the predictive ability of the observation-RFM for individual flights. This is achieved by predicting  $N_c$  for single-day events using observation-RFM trained and validated from the 3-year in situ measurements excluding the targeted flights. Such observation-RFM fails to predict  $N_c$  for all four specific events we choose to examine, including two wintertime cold-air outbreak cases observed on 01 Mar and 28 Feb 2020, respectively, and two summertime cumulus cases on 02 and 07 June 2021, respectively, as shown in Figures 2a–2d. We further predict  $N_c$  for the combined 4 cases using the observation-RFM trained and validated excluding these 4 cases to consider the seasonal variations in aerosol and clouds to some extent. The observation-RFM again fails to reproduce the observed  $N_c$  from the observational inputs (Figure 2e), as the same by using XGBoost (Figure S11 in Supporting Information S1). The same for all each individual cases with sufficient data points (Figure 2f). To study whether this failure is case-dependent or generic, we predict  $N_c$  for the continuous data set with different sample sizes using observation-RFM trained *without* the corresponding subsample. Excluding the target subsample in the training data can mitigate the potential overfitting. Data sets with different sample size  $n$  represent either a single event or a few randomly selected events. The samples are sequential series of  $n$  data points. The  $N_c$  predictability is quite low for the  $n = 30$  data sets as the

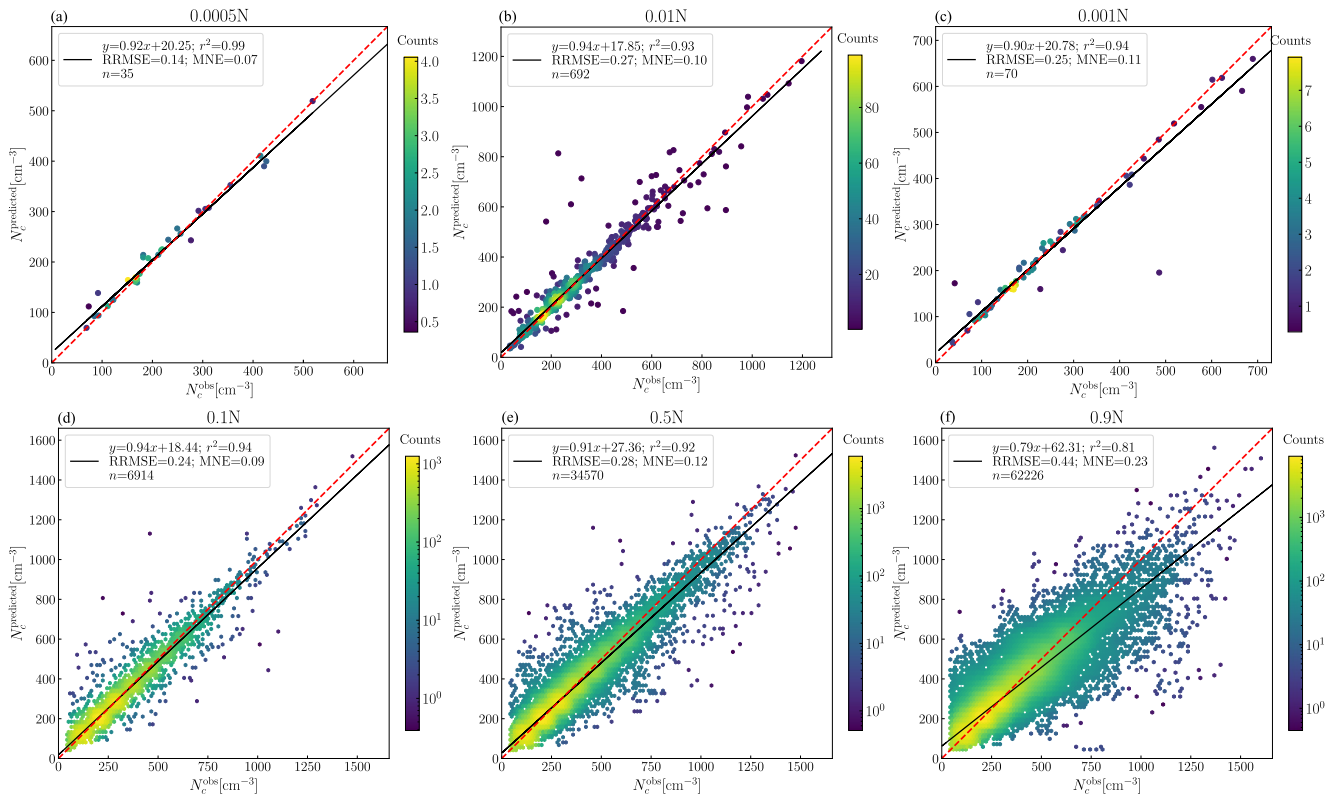


**Figure 3.**  $r^2$  of  $N_c = \mathcal{G}(N_a, w', T, q_n)$  predictions from continuous sampling with different sampling size  $n$  using the observation-RFM trained without the subsample. The solid and dashed lines represent mean (denoted as  $\langle r^2 \rangle$  in the title of the subplots) and median values of  $r^2$ , respectively. Note that the samples are drawn from running-averaged data set with a window size of 20 data points.

$r^2$  values vary randomly from  $\sim 0.0$  to  $\sim 1.0$  with a mean value of  $\langle r^2 \rangle = 0.31$  (Figure 3 and Figure S12 in Supporting Information S1), which echoes the low-performance  $N_c$  prediction for the specific cases in Figure 2. The  $\langle r^2 \rangle$  does *not* improve for  $n \geq 200$ , that is, continuous multiple events (i.e., consecutive data points), and surprisingly stays almost the same for different durations ( $n$ ). Physically, the failure of predicting the local  $N_c$  could be due to the underrepresentation of cloud regimes in the RFM as the corresponding data (i.e., continuous subsamples) are excluded from the training data. In other words, the individual cases, consisting of localized or concentrated data points, have different determining factors for the  $N_c$  prediction from the one derived from the entire data set. This motivates us to explore the predictability of RFM for global (i.e., the entire data range)  $N_c$  of our data set. In contrary to the continuous sampling for a sub-dataset with a fixed sample size, we randomly sample a sub-dataset from the 3-year observational data set. The predicted  $N_c$  from these randomly sampled sub-datasets reproduce the observed  $N_c$  remarkably well with  $r^2 > 0.9$  (Figure 4 and Figure S13 in Supporting Information S1).  $r^2 = 0.99$  even for the sub-dataset with a sample size of  $n = 0.0005N \approx 35$  with  $N$  being the total number of 3-year data points. The score for sub-dataset with this sample size is statistically significant as can be seen from  $r^2$  for 1,000 realizations (Figure S5 in Supporting Information S1). The successful prediction of  $N_c$  globally compared to the locally failed prediction regardless of the sample size shows that the  $N_c$  prediction is uncertain for short timescales and is only physically and statistically meaningful for long timescales. Namely, the cloud droplet response to aerosols, as one of the ACI metrics, appears to be more uncertain at the shorter timescales (or within a limited sampling area). We note that it is the spatiotemporal range of predictors that determines the  $N_c$  prediction instead of the number of events.

### 3.3. Observation-RFM as an Emulator for the LES Microphysics

Cloud microphysical processes for ACI are challenging to represent even in Large-eddy simulations (LES; see section SII for details of LES configuration) (Li et al., 2023; Li, Wang, et al., 2022). In this section, we tackle this challenge by using the observation-RFM as an emulator for the LES microphysics. Important turbulence scales for cloud formation can be resolved in the LES. In addition, we have shown that the observation-RFM can successfully predict observed  $N_c$  in Section 3.1. It is natural to ask whether the observation-RFM as an emulator



**Figure 4.** Test of the input sample size for the observation-RFM trained without the corresponding subsamples. The same features  $N_a, w', T,$  and  $q_v$  are used as in Figures 2 and 3. The data points for each subsample are randomly selected from the averaged full data set and are excluded from the training data set.  $N = 69,159$  is the size of the full data set.

can mitigate the uncertainties of cloud-microphysics representation in LES. We use LES from two cold-air outbreak cases (Li et al., 2023; Li, Wang, et al., 2022) and two summertime marine cumuli cases (Li et al., 2024) to represent different aerosol conditions, meteorological states, and cloud regimes over the WNAO region.  $w', T, q_v,$  and  $N_a$  from LES are taken as predictors to predict  $N_c$  using the observation-RFM. The observation-RFM fails to predict  $N_c$  for the four LES cases (Figure S8a in Supporting Information S1), which is consistent with the scale-dependent prediction of ACI within this data-driven framework, as discussed in Section 3.2. We further evaluate the LES- $N_c$  against  $N_c^{\text{obs}}$  for completeness. The LES fails to reproduce the observed  $N_c$  at the same flight-leg levels, as indicated in the four LES cases with  $r^2 \approx 0$  (Figure S8b in Supporting Information S1). This can be attributed to the following three main reasons: 1. LES with the periodic boundary conditions in horizontal directions cannot simulate the observed lateral variation of  $N_c$ , which makes the comparison of  $N_c$  (spatiotemporally point-to-point comparison) between LES and observations challenging; 2. Prescribed aerosols in the LES cannot represent the fine-scale spatiotemporal variability of aerosol size distribution; and 3. The single-value bulk hygroscopicity for all size modes is not representative of reality. The resulting  $N_c$  is expected to deviate from observations. Despite the poor performance of LES microphysics and the observation-RFM in reproducing the observed  $N_c$  for individual cases, it is still informative to compare  $N_c^{\text{LES}}$  and  $N_c^{\text{predicted}}$  directly. We find that they are nearly uncorrelated to each other (Figure S8c in Supporting Information S1).

#### 4. Discussion and Conclusion

Quantifying aerosol-cloud interactions (ACI) is very challenging due to the nonlinear multiscale nature and the incomplete understanding of related physical processes. Coarse-resolution Earth System Models only model a mean state of ACI at the model grid-scale that lacks accurate representation of physical processes (Morrison et al., 2020), while the small-scale LES simulates incomplete physical processes of ACI and only offers partial representation of ACI (Li et al., 2023, 2024). An imminent question would be whether the ACI is stochastic or deterministic within our current modeling and observation capability. Many studies (Bellouin et al., 2020;



Seinfeld et al., 2016) have alluded to this question but fallen short in explicitly formulating and answering it. In this study, we explore the stochastic characteristics of ACI by applying a widely used machine learning (ML) technique, Random Forest Model (RFM), to unprecedented 3-year in situ aircraft measurements to predict  $N_c$  over the western North Atlantic. Here, we focus on the response of  $N_c$  to aerosols properties, thermodynamics, and turbulence. The RFM can successfully predict the climatological  $N_c$  using the measured aerosol number concentration  $N_a$ ,  $w'$ , temperature  $T$ , and water vapor mixing ratio  $q_i$ , despite the strongly nonlinear aerosol and cloud microphysical processes, for example, aerosol activation and condensation and collision-coalescence of cloud droplets. However, the observation-trained RFM (observation-RFM) fails to predict  $N_c$  at shorter timescales that only cover a limited number of flights. This suggests that within this data-driven framework, the ACI is more challenging to predict at the shorter timescales. In addition, case studies of ACI (Li et al., 2023, 2024) may only represent a single realization of ACI for specific cloud regimes. Nevertheless, case studies are still useful for testing new physical processes that are important to case-dependent ACI metrics. We remark that the stochasticity of ACI discussed here is based on inferred distribution of reality based on a set of observations by the data-driven RFM instead of on first principles. The successful global prediction versus the failed local prediction regardless of the sample size likely relates to the fundamental nature of machine learning models. This includes how machine learning models learn from the data distribution of the training sample and what that might mean for cases that are usually not random sampling from those distributions.

Moving forward, one may study the stochasticity of ACI by applying data-driven algorithms to multi-field campaigns covering a wider range of different aerosol conditions and cloud regimes. Other than the stochasticity of  $N_c$  prediction, one may explore the stochasticity of cloud macrophysical responses to aerosols (e.g., liquid water and cloud fraction adjustments to aerosol perturbations) within similar data-driven modeling frameworks.

## Data Availability Statement

The source code used for the simulations of this study, the Weather Research and Forecasting (WRF) model, is freely available at Li (2023). The source code for the random forest model is publicly available at <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>. ACTIVATE observational data (Team, 2020) are publicly available at <https://asdc.larc.nasa.gov/project/ACTIVATE>.

## Acknowledgments

This work was supported through the ACTIVATE Earth Venture Suborbital-3 (EVS-3) investigation, which is funded by NASA's Earth Science Division under project no. NNL19OB081, NNL24OB07A, and 80NSSC19K0442 and managed through the Earth System Science Pathfinder Program Office. C.V. is funded by DFG SPP-1294 HALO under project no 522359172. The Pacific Northwest National Laboratory (PNNL) is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC05-76RLO1830. The simulations were performed using resources available through Research Computing at PNNL.

## References

- Albrecht, B. A. (1989). Aerosols, cloud microphysics, and fractional cloudiness. *Science*, 245(4923), 1227–1230. <https://doi.org/10.1126/science.245.4923.1227>
- Arjunan Nair, A., & Yu, F. (2020). Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements. *Atmospheric Chemistry and Physics*, 20(21), 12853–12869. <https://doi.org/10.5194/acp-20-12853-2020>
- Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., et al. (2020). Bounding global aerosol radiative forcing of climate change. *Reviews of Geophysics*, 58(1), e2019RG000660. <https://doi.org/10.1029/2019RG000660>
- Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., et al. (2020). Quantifying progress across different CMIP phases with the ESMValTool. *Journal of Geophysical Research: Atmospheres*, 125(21). <https://doi.org/10.1029/2019JD032321>
- Boltzmann, L. (2022). *Lectures on gas theory*. Univ of California Press.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12343 LNCS.
- Brunke, M. A., Cutler, L., Urzua, R. D., Corral, A. F., Crosbie, E., Hair, J., et al. (2022). Aircraft observations of turbulence in cloudy and cloud-free boundary layers over the Western North Atlantic Ocean from ACTIVATE and implications for the Earth system model evaluation and development. *Journal of Geophysical Research: Atmospheres*, 127(19), e2022JD036480. <https://doi.org/10.1029/2022JD036480>
- Chakraborty, T. C., & Lee, X. (2021). Using supervised learning to develop BaRAD, a 40-year monthly bias-adjusted global gridded radiation dataset. *Scientific Data*, 8(1), 238. <https://doi.org/10.1038/s41597-021-01016-4>
- Chen, Y., Hayward, J., Wang, Y., Malavelle, F., Jordan, G., Partridge, D., et al. (2022). Machine learning reveals climate forcing from aerosols is dominated by increased cloud cover. *Nature Geoscience*, 15(8), 609–614. <https://doi.org/10.1038/s41561-022-00991-6>
- Corral, A. F., Braun, R. A., Cairns, B., Goroooh, V. A., Liu, H., Ma, L., et al. (2021). An overview of atmospheric features over the western North Atlantic Ocean and North American East Coast – Part 1: Analysis of aerosols, gases, and wet deposition chemistry. *Journal of Geophysical Research: Atmospheres*, 126(4), e2020JD032592. <https://doi.org/10.1029/2020JD032592>
- Corral, A. F., Choi, Y., Collister, B. L., Crosbie, E., Dadashazar, H., DiGangi, J. P., et al. (2022). Dimethylamine in cloud water: A case study over the northwest Atlantic Ocean. *Environmental Sciences: Atmosphere*, 2(6), 1534–1550. <https://doi.org/10.1039/d2ea00117a>
- Dadashazar, H., Corral, A. F., Crosbie, E., Dmitrovic, S., Kirschler, S., McCauley, K., et al. (2022). Organic enrichment in droplet residual particles relative to out of cloud over the northwestern Atlantic: Analysis of airborne ACTIVATE data. *Atmospheric Chemistry and Physics*, 22(20), 13897–13913. <https://doi.org/10.5194/acp-22-13897-2022>
- Dadashazar, H., Painemal, D., Alipanah, M., Brunke, M., Chellappan, S., Corral, A. F., et al. (2021). Cloud drop number concentrations over the western North Atlantic Ocean: Seasonal cycle, aerosol interrelationships, and other influential factors. *Atmospheric Chemistry and Physics*, 21(13), 10499–10526. <https://doi.org/10.5194/acp-21-10499-2021>

- DeCarlo, P. F., Dunlea, E. J., Kimmel, J. R., Aiken, A. C., Sueper, D., Crouse, J., et al. (2008). Fast airborne aerosol size and chemistry measurements above Mexico City and Central Mexico during the MILAGRO campaign. *Atmospheric Chemistry and Physics*, 8(14), 4027–4048. <https://doi.org/10.5194/acp-8-4027-2008>
- Ghan, S., Wang, M., Zhang, S., Ferrachat, S., Gettelman, A., Griesfeller, J., et al. (2016). Challenges in constraining anthropogenic aerosol effects on cloud radiative forcing using present-day spatiotemporal variability. *Proceedings of the National Academy of Sciences*, 113(21), 5804–5811. <https://doi.org/10.1073/pnas.1514036113>
- Grabowski, W. W., & Wang, L. P. (2013). Growth of cloud droplets in a turbulent environment. *Annual Review of Fluid Mechanics*, 45(1), 293–324. <https://doi.org/10.1146/annurev-fluid-011212-140750>
- Kirschler, S., Voigt, C., Anderson, B., Campos Braga, R., Chen, G., Corral, A. F., et al. (2022). Seasonal updraft speeds change cloud droplet number concentrations in low-level clouds over the western North Atlantic. *Atmospheric Chemistry and Physics*, 22(12), 8299–8319. <https://doi.org/10.5194/acp-22-8299-2022>
- Kirschler, S., Voigt, C., Anderson, B. E., Chen, G., Crosbie, E. C., Ferrare, R. A., et al. (2023). Overview and statistical analysis of boundary layer clouds and precipitation over the western North Atlantic Ocean. *Atmospheric Chemistry and Physics*, 23(18), 10731–10750. <https://doi.org/10.5194/acp-23-10731-2023>
- Köhler, H. (1936). The nucleus in and the growth of hygroscopic droplets. *Transactions of the Faraday Society*, 32(0), 1152–1161. <https://doi.org/10.1039/TF9363201152>
- Li, X.-Y. (2018). Droplet growth in atmospheric turbulence: A direct numerical simulation study (doctoral dissertation, Stockholm). Retrieved from <http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-158537>; <http://su.diva-portal.org/smash/get/diva2:1237428/FULLTEXT01.pdf>; <http://su.diva-portal.org/smash/get/diva2:1237428/PREVIEW01.jpg>
- Li, X.-Y. (2023). Xiang-yu/WRF-LASSO: WRF-LASSO [Software]. *Zenodo*. <https://doi.org/10.5281/zenodo.10421287>
- Li, X.-Y., Brandenburg, A., Svensson, G., Haugen, N. E. L., Mehlig, B., & Rogachevskii, I. (2020). Condensational and collisional growth of cloud droplets in a turbulent environment. *Journal of the Atmospheric Sciences*, 77(1), 337–353. <https://doi.org/10.1175/JAS-D-19-0107.1>
- Li, X.-Y., Mehlig, B., Svensson, G., Brandenburg, A., & Haugen, N. E. L. (2022a). Collision fluctuations of lucky droplets with superdroplets. *Journal of the Atmospheric Sciences*, 79(7), 1821–1835. <https://doi.org/10.1175/JAS-D-20-0371.1>
- Li, X.-Y., Wang, H., Chen, J., Endo, S., George, G., Cairns, B., et al. (2022b). Large-eddy simulations of marine boundary layer clouds associated with cold-air outbreaks during the ACTIVATE campaign. Part I: Case setup and sensitivities to large-scale forcings. *Journal of the Atmospheric Sciences*, 79(1), 73–100. <https://doi.org/10.1175/JAS-D-21-0123.1>
- Li, X.-Y., Wang, H., Chen, J., Endo, S., Kirschler, S., Voigt, C., et al. (2023). 1). Large-eddy simulations of marine boundary-layer clouds associated with cold-air outbreaks during the ACTIVATE campaign. Part II: Aerosol–Meteorology–Cloud interaction. *Journal of the Atmospheric Sciences*, 80(4), 1025–1045. <https://doi.org/10.1175/JAS-D-21-0324.1>
- Li, X.-Y., Wang, H., Christensen, M. W., Chen, J., Tang, S., Kirschler, S., et al. (2024). Process modeling of aerosol-cloud interaction in summertime precipitating shallow cumulus over the Western North Atlantic. *Journal of Geophysical Research: Atmospheres*, 129(7), e2023JD039489. <https://doi.org/10.1029/2023JD039489>
- Michel, S. L., Swingedouw, D., Ortega, P., Gastineau, G., Mignot, J., McCarthy, G., & Khodri, M. (2022). Early warning signal for a tipping point suggested by a millennial Atlantic Multidecadal Variability reconstruction. *Nature Communications*, 13(1), 5176. <https://doi.org/10.1038/s41467-022-32704-3>
- Moore, R. H., Wiggins, E. B., Ahern, A. T., Zimmerman, S., Montgomery, L., Campuzano Jost, P., et al. (2021). Sizing response of the ultra-high sensitivity aerosol spectrometer (UHSAS) and laser aerosol spectrometer (LAS) to changes in submicron aerosol composition and refractive index. *Atmospheric Measurement Techniques*, 14(6), 4517–4542. <https://doi.org/10.5194/amt-14-4517-2021>
- Morrison, H., van Lier-Walqui, M., Fridlind, A. M., Grabowski, W. W., Harrington, J. Y., Hoose, C., et al. (2020). Confronting the challenge of modeling cloud and precipitation microphysics. *Journal of Advances in Modeling Earth Systems*, 12(8), e2019MS001689. <https://doi.org/10.1029/2019MS001689>
- Nair, A. A., Yu, F., Campuzano-Jost, P., DeMott, P. J., Levin, E. J., Jimenez, J. L., et al. (2021). Machine learning uncovers aerosol size information from chemistry and meteorology to quantify potential cloud-forming particles. *Geophysical Research Letters*, 48(21), e2021GL094133. <https://doi.org/10.1029/2021GL094133>
- Painemal, D., Spangenberg, D., Smith, W. L., Jr., Minnis, P., Cairns, B., Moore, R. H., et al. (2021). Evaluation of satellite retrievals of liquid clouds from the GOES-13 imager and MODIS over the midlatitude North Atlantic during the NAAMES campaign. *Atmospheric Measurement Techniques*, 14(10), 6633–6646. <https://doi.org/10.5194/amt-14-6633-2021>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830.
- Seinfeld, J. H., Bretherton, C., Carslaw, K. S., Coe, H., DeMott, P. J., Dunlea, E. J., et al. (2016). Improving our fundamental understanding of the role of aerosol-cloud interactions in the climate system. *Proceedings of the National Academy of Sciences*, 113(21), 5781–5790. <https://doi.org/10.1073/pnas.1514043113>
- Silva, S. J., & Keller, C. A. (2023). Limitations of XAI methods for process-level understanding in the atmospheric sciences. *Artificial Intelligence for the Earth Systems*, 3(1). <https://doi.org/10.1175/aies-d-23-0045.1>
- Sorooshian, A., Alexandrov, M. D., Bell, A. D., Bennett, R., Betito, G., Burton, S. P., et al., (2023). Spatially coordinated airborne data and complementary products for aerosol, gas, cloud, and meteorological studies: The NASA ACTIVATE dataset. *Earth System Science Data*, 15(8), 3419–3472. <https://doi.org/10.5194/essd-15-3419-2023>
- Sorooshian, A., Anderson, B., Bauer, S. E., Braun, R. A., Cairns, B., Crosbie, E., et al. (2019). Aerosol–cloud–meteorology interaction airborne field investigations: Using lessons learned from the U.S. West coast in the design of activate off the U.S. East Coast. *Bulletin of the American Meteorological Society*, 100(8), 1511–1528. <https://doi.org/10.1175/BAMS-D-18-0100.1>
- Team, A. S. (2020). Aerosol cloud meTeorology interactions oVer the western ATlantic experiment [Dataset]. <https://doi.org/10.5067/SUBORBITAL/ACTIVATE/DATA001>
- Twomey, S. (1974). Pollution and the planetary albedo. *Atmospheric Environment*, 8(12), 1251–1256. [https://doi.org/10.1016/0004-6981\(74\)90004-3](https://doi.org/10.1016/0004-6981(74)90004-3)
- Yu, F., Luo, G., Nair, A. A., Tsigaridis, K., & Bauer, S. E. (2022). Use of machine learning to reduce uncertainties in particle number concentration and aerosol indirect radiative forcing predicted by climate models. *Geophysical Research Letters*, 49(16), e2022GL098551. <https://doi.org/10.1029/2022GL098551>