Geophysical Research Letters

RESEARCH LETTER
10.1029/2022GL098551

Key Points:

- A physics-guided machine learning (ML) model for particle number concentration (PNC) couples with a global climate model with low computational overhead
- ML PNC are in better agreement with measurements and reduce cloud droplet number concentration changes caused by anthropogenic emissions
- Radiative forcing due to aerosol-cloud interactions indicates weaker cooling (−1.11 vs. −1.46 W·m^{-2}) and is closer to IPCC median value

Correspondence to:
F. Yu,
fyu@albany.edu

Citation:
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*, e2022GL098551. https://doi.org/10.1029/2022GL098551

Received 2 MAR 2022
Accepted 23 JUL 2022

Author Contributions:
Conceptualization: Fangqun Yu, Gan Luo, Arshad Arjunan Nair
Data curation: Fangqun Yu
Formal analysis: Fangqun Yu, Gan Luo
Funding acquisition: Fangqun Yu
Investigation: Fangqun Yu, Gan Luo, Arshad Arjunan Nair
Methodology: Fangqun Yu, Gan Luo, Arshad Arjunan Nair
Project Administration: Fangqun Yu
Resources: Fangqun Yu
Software: Fangqun Yu, Gan Luo, Arshad Arjunan Nair, Kostas Tsigaridis, Susanne E. Bauer
Supervision: Fangqun Yu
Validation: Fangqun Yu, Gan Luo, Arshad Arjunan Nair
Visualization: Fangqun Yu, Gan Luo, Arshad Arjunan Nair
Writing – original draft: Fangqun Yu
Writing – review & editing: Fangqun Yu, Gan Luo, Arshad Arjunan Nair, Kostas Tsigaridis, Susanne E. Bauer

Use of Machine Learning to Reduce Uncertainties in Particle Number Concentration and Aerosol Indirect Radiative Forcing Predicted by Climate Models

Fangqun Yu[1], Gan Luo[1], Arshad Arjunan Nair[1], Kostas Tsigaridis[2,3], and Susanne E. Bauer[2]

[1]Atmospheric Sciences Research Center, State University of New York, Albany, NY, USA, [2]NASA Goddard Institute for Space Studies, New York, NY, USA, [3]Center for Climate Systems Research, New York, NY, USA

Abstract The radiative forcing of anthropogenic aerosols associated with aerosol-cloud interactions (RF_{aci}) remains the largest source of uncertainty in climate prediction. The calculation of particle number concentration (PNC), one of the critical parameters affecting RF_{aci}, is generally simplified in climate models. Here we employ outputs from long-term (30-year) simulations of a global size-resolved (sectional) aerosol microphysics model and a machine-learning tool to develop a Random Forest Regression Model (RFRM) for PNC. We have implemented the PNC RFRM in GISS-ModelE2.1 with a mass-based One-Moment Aerosol module, which is one of CMIP6 models. Compared to the default setting, the GISS-ModelE2.1 simulation based on RFRM reduces the changes of cloud droplet number concentration associated with anthropogenic emissions, and decreases the RF_{aci} from −1.46 to −1.11 W·m^{-2}. This work highlights a promising approach based on machine learning to reduce uncertainties of climate models in predicting PNC and RF_{aci} without compromising their computing efficiency.

Plain Language Summary The largest uncertainty in assessing climate change is due to the interaction of aerosols with clouds and its effect on the Earth's energy budget. To reduce this uncertainty, it is important to accurately quantify aerosols. This is possible by accounting for the physical processes and interactions happening at the scale of the aerosol sizes. However, such an approach would be computationally demanding on climate models, making them impractical to study historical changes. To address this dilemma, we use a machine learning/artificial intelligence (ML/AI) technique that learns aerosol microphysics. When coupled to a climate model, it speedily quantifies aerosols in strong agreement with atmospheric measurements. Compared to the climate model without this implicit physical treatment, the historical changes in cloud droplet numbers and the cooling effect of aerosols are now estimated lower and less uncertain. This highlights the potential of ML/AI in reducing climate model uncertainties without hindering their computational efficiency.

1. Introduction

All cloud droplets formed in the atmosphere start with tiny particles that act as cloud condensation nuclei (CCN). These aerosols can alter cloud properties and precipitation (Albrecht, 1989; Twomey, 1977) and thereby indirectly influence the Earth's radiation budget and climate change. The radiative forcing (RF) associated with aerosol–cloud interactions (aci) remains the largest source of uncertainty in climate prediction. According to the Intergovernmental Panel on Climate Change's Fifth Assessment Report (IPCC AR5, 2013), RF_{aci} of anthropogenic aerosols was estimated to be −0.55 W·m^{-2} with "low" level of confidence. A number of post-IPCC AR5 global climate modeling studies still show large discrepancy in the values of RF_{aci}, ranging from ∼ −0.35 W·m^{-2} (Nazarenko et al., 2017), to −0.7 W·m^{-2} (Rotstayn et al., 2014) to −1.08 W·m^{-2} (Bauer et al., 2020), to −1.28 W·m^{-2} (Tonttila et al., 2015), to −1.54 W·m^{-2} (Bauer et al., 2020), and to −2.19 W·m^{-2} (Zhang et al., 2016). In order to confidently interpret past and accurately project future climate change, it is essential to reduce RF_{aci} discrepancy among different models.

RF_{aci} depends strongly on the response of number concentrations of particles that can act as CCN to anthropogenic emissions (Albrecht, 1989; Twomey, 1977). The increase in cloud drops with particle number concentration (PNC) has been confirmed by many aircraft measurements (e.g., Ramanathan et al., 2001). PNC exhibits significant spatial and temporal variability due to the non-linear dependence of new particle formation and growth

rates on atmospheric conditions and concentrations of gaseous precursors, both subject to changes resulting from climate changes and emission regulatory actions. The calculation of PNC in global climate models (GCMs), including the most recent CMIP6 models, is generally simplified which contributes to the uncertainty in RF_{aci}. For example, among 10 CMIP6 models compared by Zanis et al. (2020), 7 models employ bulk mass-based aerosol schemes not supporting PNC simulation, while 3 models use modal aerosol schemes considering particle size distribution. Large uncertainties exist in the predicted PNC in these models as PNC not only depends on mass concentrations of particles but also on their size distributions which have significant spatiotemporal variations that can only be captured by the use of an aerosol microphysics parameterization. Bellouin et al. (2013) showed substantial differences in the aerosol forcings simulated by the bulk and modal schemes, pointing out that the bulk approach lacked the necessary sophistication to provide realistic aerosol input for aerosol-cloud-radiation calculations.

Critical toward more accurate modeling of aerosols' effect on clouds is to have a robust representation of aerosol processes key for quantifying PNC. These include non-linear processes of secondary particle formation and growth as well as interactions among particles of different sizes and compositions (e.g., Yu & Luo, 2009). The main challenge is the high computational expense of simulating size- and composition-resolved particle microphysics in climate models. Here we show that this dilemma of the need of more accurate aerosol properties important for RF_{aci} and consideration of computing efficiency can be resolved by using machine learning. Machine learning is a branch of artificial intelligence, where systems trained on a large number of scenarios learn to build a statistical predictive model without explicitly programming. Over the last two decades, there has been rapid development and application of machine learning, with recent applications in the atmospheric sciences such as in atmospheric new particle formation (e.g., Zaidan et al., 2018), mixing-state (e.g., Hughes et al., 2018), air quality (e.g., Grange et al., 2018), remote-sensing (e.g., Mauceri et al., 2019), tropical cyclone intensity change forecast (Su et al., 2020), and other aspects (e.g., Jin et al., 2019; D. J. Miller et al., 2020). These studies demonstrate the strong utility of machine learning in the development of predictive models considerate of the non-linear associations between atmospheric states and compositions. In a recent study, Nair et al. (2021) showed that machine learning can extract aerosol size information from aerosol composition and additionally from atmospheric chemical and meteorological variables. In this study, we employ outputs from long-term (30-year) simulations of a global size-resolved (sectional) aerosol microphysics model and a machine-learning tool to develop a Random Forest Regression Model (RFRM) for PNC. We have implemented the PNC RFRM in the version of GISS-ModelE2.1 with a mass-based One-Moment Aerosol (OMA) module, which is one of the models participating in CMIP6 (GISS-E2.1). We want to note, that the GISS model includes an aerosol microphysical model, MATRIX (Multiconfiguration Aerosol TRacker of mIXing state) (Bauer et al., 2008). However here we choose to use the mass-based aerosol scheme to demonstrate that the calculation of PNC in GCMs with bulk aerosol schemes can be improved with RFRM. To the best of our knowledge, this is the first application of machine learning to improve PNC simulations and tackle the persistent uncertainty in RF_{aci} in climate models.

2. Materials and Methods

2.1. GEOS-Chem-APM Model (GCAPM)

The GEOS-Chem model is a global 3-D model of atmospheric composition (e.g., Bey et al., 2001) and is continuously being improved (e.g., Evans & Jacob, 2005; Holmes et al., 2019; Keller et al., 2014; Luo et al., 2020; Martin et al., 2003; Murray et al., 2012; Pye & Seinfeld, 2010; van Donkelaar et al., 2008). The present study uses GEOS-Chem version 10-01 with the incorporation of the size-resolved (sectional) Advanced Particle Microphysics (APM) package (Yu & Luo, 2009), henceforth referred to as GCAPM. The APM model has the following features of relevance toward accurate simulation of (PNCs) that are important for aerosol-cloud interactions: (a) 40 bins to represent secondary particles with 30 of these bins (and thus quite high resolution) for the size range (diameter 1.2–120 nm) important for the growth of nucleated particles to CCN sizes (Yu & Luo, 2009); (b) state-of-the-art Ternary Ion mediated Nucleation (TIMN) mechanism (Yu et al., 2018, 2020) and temperature-dependent organics-mediated nucleation (Yu et al., 2017); (c) explicit kinetic condensation of both H_2SO_4 and low volatile organic gases onto particles as well as consideration of contributions to particle growth of nitrate and ammonium via equilibrium uptake and semi-volatile organics through partitioning (Yu, 2011); and (d) explicit resolution of the coating of secondary species on primary particles (Yu & Luo, 2009). Particle size

distributions and PNC simulated by GCAPM have previously been evaluated against a number of measurements (Luo & Yu, 2010; Williamson et al., 2019; Yu, 2011; Yu et al., 2010, 2013, 2017; Yu & Luo, 2009).

In this study, detailed outputs of GCAPM long-term (1989–2018) global simulations (2° × 2.5° horizontal resolution) at each time step (half-hour) for all model layers in the troposphere at 47 sites across the globe are used for training the machine-learning model. We assume that parts of the world during parts of time periods even at present-day emissions are clean enough to be representative for the preindustrial atmosphere, and the RFRM can capture non-linear dependence of PNC on key variables.

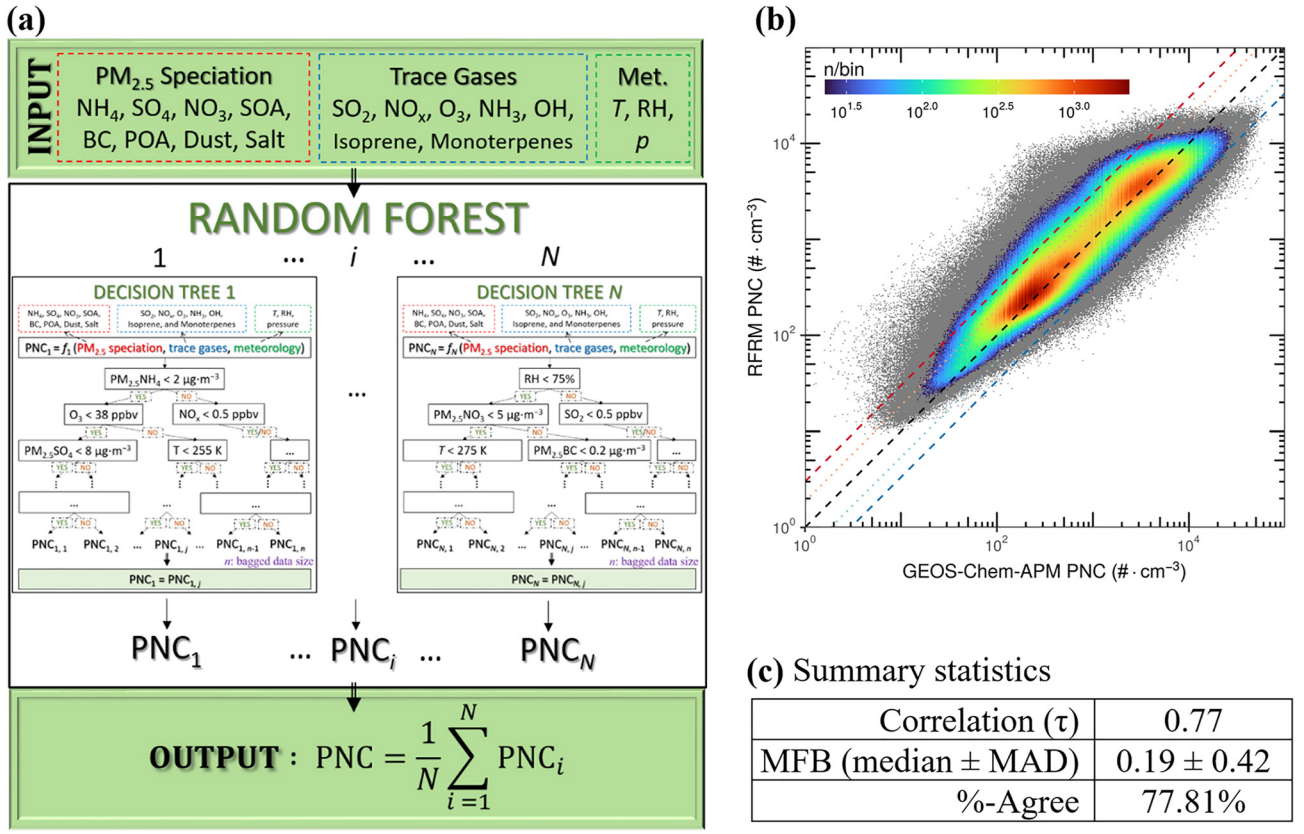
2.2. Machine-Learning Tool and Random Forest Regression Modeling (RFRM)

In a previous study, we use random decision forests to develop a Random Forest Regression Model (RFRM) to derive CCN concentration at 0.4% supersaturation [CCN0.4] from commonly available variables (Nair & Yu, 2020). A Random Forest (Breiman, 2001) is a supervised machine-learning algorithm; the "forest" is an ensemble of decision trees (Breiman et al., 1984) that aggregates (mean) all the component decision trees' regression of a variable against input predictor variables. Each decision tree splits the training data into homogenous subsets using recursive binary splitting by choosing the input variables ("predictors") that minimize the variance of the outcome. Decision trees are not influenced by missing data or outliers and are non-parametric. Random Forest models are advantageous due to their 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 comparatively easier to visualize and interpret as opposed to black-box neural network or deep learning methods. RFRM is one of the most accurate predictive machine-learning models with the ability to be trained fast due to the parallelizability of building the decision tree. For these reasons, Random Forest is our chosen machine-learning tool in this study.

The approach for training, testing, validating, and optimizing the RFRM is the same as those detailed in Nair and Yu (2020) except that this study focuses on PNC, instead of [CCN0.4]. Figure 1a illustrates the random forest machine-learning technique. 247.2 million rows (sets of 18 predictor inputs as listed in Figure 1a and PNC) of GCAPM output data are used to train the RFRM. Here PNC represents the number concentrations of all particles with diameter larger than 10 nm. We select 18 atmospheric state and composition predictor variables that are available in GISS-E2-1-OMA (see Section 2.3) as predictors, including mass concentrations of sulfate, nitrate, ammonium, secondary organic aerosol (SOA), black carbon (BC), primary organic carbon (POC), sea salt, and dust in particles smaller than 2.5 μm, concentrations of SO_2, NH_3, NO_x, O_3, OH, isoprene, and monoterpenes, temperature (T), relative humidity (RH), and pressure. No explicit spatiotemporal information is provided to the RFRM in its training so that the RFRM remains generalizable and agnostic to the host model. The speed-optimized RFRM with negligible change in performance is trained and tested (training data:testing data: 7:3) on a statistically representative 1% random subset of GCAPM simulations for 32 randomly selected sites (out of 47) and further validated with data for the remaining sites that the RFRM has not been exposed to. For the present case, we consider 30 trees with other hyperparameters being the defaults prescribed by Wright and Ziegler (2017).

2.3. GISS ModelE

NASA Goddard Institute for Space Studies (GISS) climate modeling has a long development history (Hansen et al., 1983, 1997, 2002) and contributed to CMIPs (Kelley et al., 2020; Schmidt et al., 2006, 2014). Three versions of GISS-E2.1 simulations have been submitted to CMIP6: (a) NINT (non-interactive), (b) OMA model (Bauer & Koch, 2005; Koch et al., 2006; R. L. Miller et al., 2006; Tsigaridis et al., 2013), and (c) MATRIX (Multiconfiguration Aerosol TRacker of mIXing state) model (Bauer et al., 2008). The two aerosol schemes differ by degree of complexity (Bauer et al., 2020), with OMA having more detailed chemistry regarding secondary organic aerosol formation while MATRIX having resolved aerosol microphysical processes, including mixing state. The mass-based OMA module is faster than MATRIX but both models are computationally efficient, which is critical for long-term climate simulations. The OMA transient monthly mean output is used as offline fields to drive GISS-E2-1 in all climatological (control), historical, and future NINT atmospheric composition simulations for CMIP6 (Bauer et al., 2020).



(c) Summary statistics

Correlation (τ)	0.77
MFB (median ± MAD)	0.19 ± 0.42
%-Agree	77.81%

Figure 1. (a) Schematic of the random forest machine learning technique. (b) Binned scatterplot of Random Forest Regression Model (RFRM)-derived versus GEOS-Chem-APM Model (GCAPM) simulated values for particle number concentration (PNC). Color bar shows the number of points in each bin; total number of points = 7,363,160. Bins with low counts (<1% of maximum count: ~2.36% of the data) are shaded gray. The lines indicate MFB of 0 (black: perfect agreement), +1 (dark red), −1 (dark blue), +0.6 (light red), and −0.6 (light blue). (c) Summary statistics quantifying RFRM agreement with GCAPM. Listed are Kendall's rank correlation coefficient (τ), MFB: median ± median absolute deviation of the fractional bias, and %-Agree: the percentage of RFRM-derived PNC in good agreement (−0.6 < fractional bias <0.6).

In OMA, the PNC_{OMA} that impacts clouds is obtained from the aerosol mass as described in Menon and Rotstayn (2006). GISS-E2.1 only includes the first indirect effect, as such aerosols only influence cloud optical depth, the Twomey effect. PNC_{OMA} depends only on the mass concentrations of various aerosol components. Similar to other CMIP6 models with bulk mass-based aerosol schemes (Zanis et al., 2020), the mass-to-number conversion coefficients are derived from prescribed particle size distributions (Menon and Rotstayn, 2006), which is undoubtedly a large source of uncertainty in aerosol radiative forcing calculations. One approach to address this issue is to resolve aerosol microphysics and particle sizes (e.g., Fanourgakis et al., 2019). However, this is generally associated with a large increase in the computing cost, especially for the sectional aerosol microphysics models. In addition, the inclusion of more complex particle microphysics can induce additional uncertainties if relevant processes and parameterizations are not well constrained. In this study, we employ outputs from GCAPM simulations and a machine-learning tool to develop a RFRM for PNC, solving the challenge of the need of more accurate aerosol properties important for RF_{aci}, while maintaining the computing efficiency of GISS-E2.1-OMA (and, potentially, other mass-based CMIP6 models).

This work is based on GISS-E2-1 downloaded from the GISS website (https://www.giss.nasa.gov/tools/modelE/). We run the updated GISS-E2-1 at 2° × 2.5° horizontal resolution from surface to 0.1 hPa with 40 vertical layers. Model configurations were following the settings recommended by R. L. Miller et al. (2021) and Kelley et al. (2020) for CMIP6 simulations. Anthropogenic emissions of gases and aerosols were from the Community Emissions Data System (CEDS; Hoesly et al., 2018). Dust emission was simulated following the approach of R. L. Miller et al. (2006), while sea salt emission was simulated as per Tsigaridis et al. (2013). Sea surface temperatures and sea ice covers were prescribed by the Met Office Hadley Center's sea ice and sea surface temperature

data set (Rayner et al., 2003). We run the model nudged with NCEP reanalysis horizontal winds, which are available from 1948 to present.

3. Results

The Random Forest Regression Model (RFRM) as described in Section 2.1 is used to predict PNC from atmospheric state and composition variables as illustrated in Figure 1a. Figure 1b shows the binned scatterplot comparison of RFRM-derived versus GCAPM-simulated PNC values. RFRM PNC is highly correlated ($\tau \approx 0.77$) and in good agreement (summary statistics in Figure 1c) with GCAPM values. Overall, the RFRM is robust in its derivation of PNC for various locations around the globe, at various altitudes, and across a varied range of PNC magnitudes.

We have implemented the PNC RFRM in GISS ModelE2.1-OMA. It increases the computing cost by ∼5%, non-trivial but acceptable even for centennial-long simulations. Figure 2 compares PNC estimated in OMA with prescribed mass-to-number coefficients (PNC_{OMA}) based on Menon and Rotstayn (2006) and those based on the machine-learning RFRM (PNC_{ML}). The measured annual mean PNC values at 35 sites across the globe and daily mean time series at Pinnacle State Park (PSP), New York, are also shown. The sources of PNC data are those described in Yu (2011) plus some additional sites from EBAS database (http://ebas.nilu.no/) and the PSP data is from the Atmospheric Science Research Center's Air Quality Monitoring Products (http://atmoschem.asrc.cestm.albany.edu/∼aqm). In the surface layer, the spatial distributions (or patterns) of PNC_{OMA} (Figure 2a) and PNC_{ML} (Figure 2b) are similar as particles associated with anthropogenic emissions dominate the number concentrations. Both PNC_{OMA} and PNC_{ML} capture the observed spatial variations of annual mean PNC at the 35 sites, with correlation coefficients of 0.82 and 0.88 (Figure 2c), respectively. Nevertheless, PNC_{OMA} is generally higher than PNC_{ML} in the surface layer, but is lower over Australia, polar regions, and areas with high topography. The global average PNC_{OMA} in the model surface layer is 1866 #·cm^{-3}, which is about 70% higher than that of PNC_{ML}. In addition to the spatial variation, temporal variation is also critical for aerosol–cloud interactions. To explore this, we examine the monthly mean values at the Hyytiala, Finland site (Figure 2d) and the daily mean values at the PSP site (Figure 2e) where PNC measurements are available and the annual mean PNC_{OMA} and PNC_{ML} are almost identical (Sites "F" and "O," Figure 2c). For the Hyytiala site, the seasonal variation of PNC_{ML} is in much better agreement with observations than that of PNC_{OMA} ($r = 0.51$ vs. 0.16). Large differences can be seen between daily mean time series of PNC_{OMA} and PNC_{ML} at the PSP site, with PNC_{OMA} peaking in the summer and winter seasons while PNC_{ML} shows higher values in the spring. PNC_{ML} agrees much better with observed daily mean values ($r = 0.52$) than that for PNC_{OMA} ($r = 0.07$). Our comparisons show that RFRM-derived PNC is in much better agreement with relevant observations at PSP and Hyytiala, Finland. The improvement is associated with more complex dependence of PNC not only on particle mass but also on atmospheric weather (T, RH, pressure) and atmospheric composition (SO_2, NH_3, NO_x, O_3, OH, isoprene, monoterpenes).

In GISS ModelE2.1-OMA, cloud droplet number concentration (CDNC) is related to the PNC via empirical parameterizations described in Menon and Rotstayn (2006). Figure 3 shows cloud-cover weighted 10-year (2005–2014) mean CDNC for warm large-scale clouds based on PNC_{OMA} and PNC_{ML} under the pre-industrial (PI) and present-day (PD) emission scenarios. While both $CDNC_{OMA}$ and $CDNC_{ML}$ show generally higher values over main continents where PNC is larger, substantial differences in their spatial distributions can be clearly seen. For example, compared to PD $CDNC_{OMA}$, PD $CDNC_{ML}$ is lower in the southern part of Eurasia, Australia, and the polar regions. More importantly, compared to $CDNC_{OMA}$, global mean $CDNC_{ML}$ is larger (by 18%) under PI emission but is smaller (by 6%) under PD emission. The difference is caused by the change in particle size distributions from PI to PD scenarios, which is not considered in PNC_{OMA} but is taken into account in PNC_{ML}. As a result, the relative change of global mean CDNC from PI to PD decreases from 61% for $CDNC_{OMA}$ to 28% for $CDNC_{ML}$.

Same as in Bauer et al. (2020), and following the method described in Ghan (2013), we calculate RF_{aci} using GISS ModelE based on both PNC_{OMA} and PNC_{ML}. Figure 4 shows that the application of machine-learning RFRM in predicting total aerosol numbers used for CDNC calculations in OMA reduces the RF_{aci} from -1.46 W·m^{-2} to -1.11 W·m^{-2}. The reduction is due to a decrease in relative change of CDNC (Figure 3) associated with anthropogenic emissions (from pre-industrial (PI, 1850) to present-day (PD, 2010) emissions). RF_{aci} of -1.11 W·m^{-2} is closer to that based on GISS-E2-1-MATRIX (Bauer et al., 2020) and is closer to the median value given in IPCC

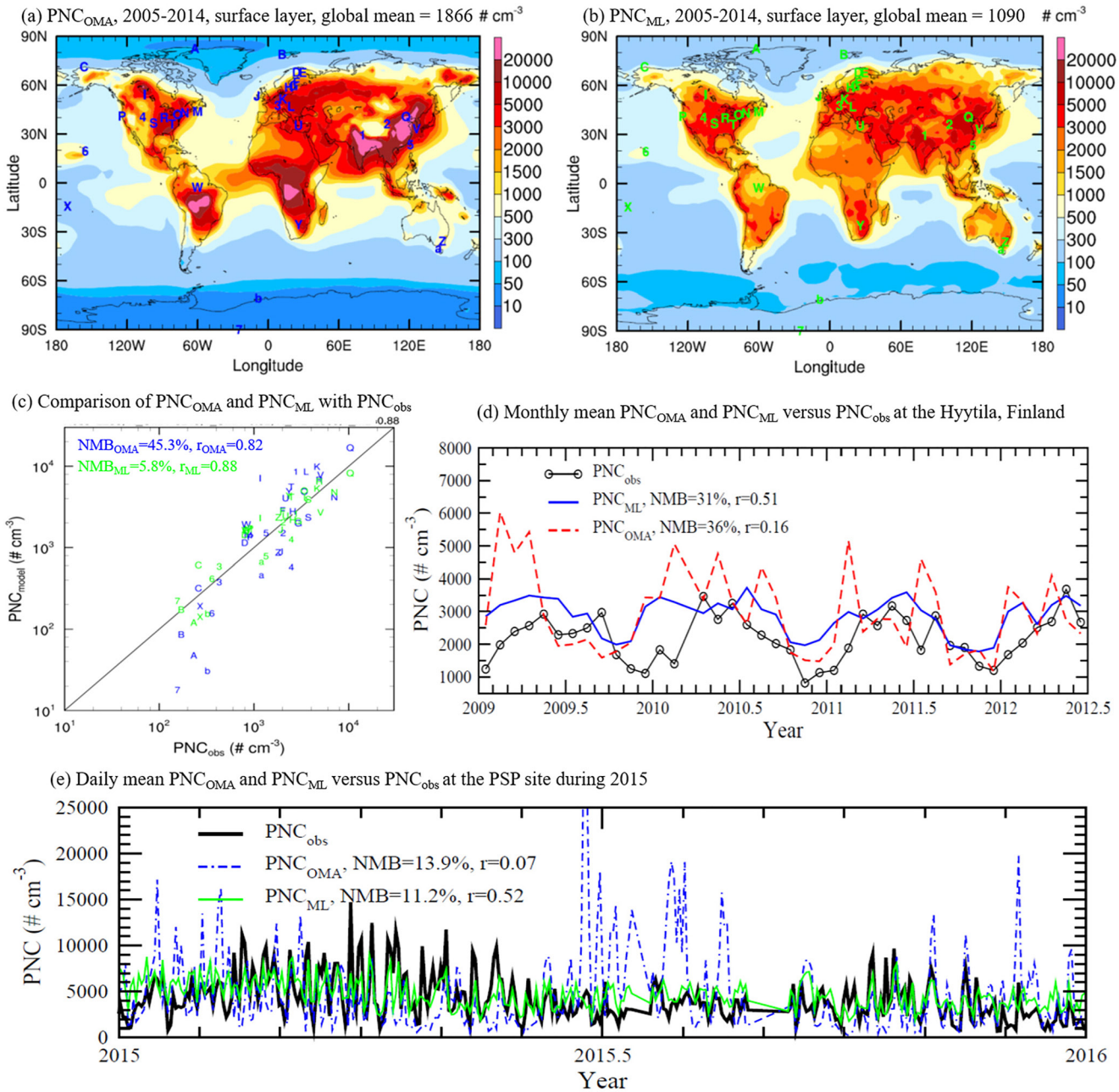


Figure 2. Decadal mean (2005–2014) surface layer particle number concentration (PNC) (a) estimated in One-Moment Aerosol (OMA) with prescribed mass to number coefficients (PNC$_{OMA}$), and (b) calculated based on the Random Forest Regression Model given in Figure 1 (PNC$_{ML}$). (c) Comparisons with observed annual mean PNC (PNC$_{obs}$) at 35 sites across the globe (marked on (a and b)) of PNC$_{OMA}$ and PNC$_{ML}$. (d) Multiple-year monthly mean time series of PNC$_{obs}$, PNC$_{OMA}$, and PNC$_{ML}$ at Hyytiala, Finland (site "F"). (e) Daily mean time series of PNC$_{obs}$, PNC$_{OMA}$, and PNC$_{ML}$ at Pinnacle State Park, New York (site "O") in the year 2015.

report (IPCC AR5, 2013). It should be noted that RF$_{aci}$ based on PNC$_{ML}$ is generally weaker over continents but is stronger over North Pacific and North Atlantic. The large difference highlights the need to account for the particle size changes from PI to PD in PNC and CDNC calculations.

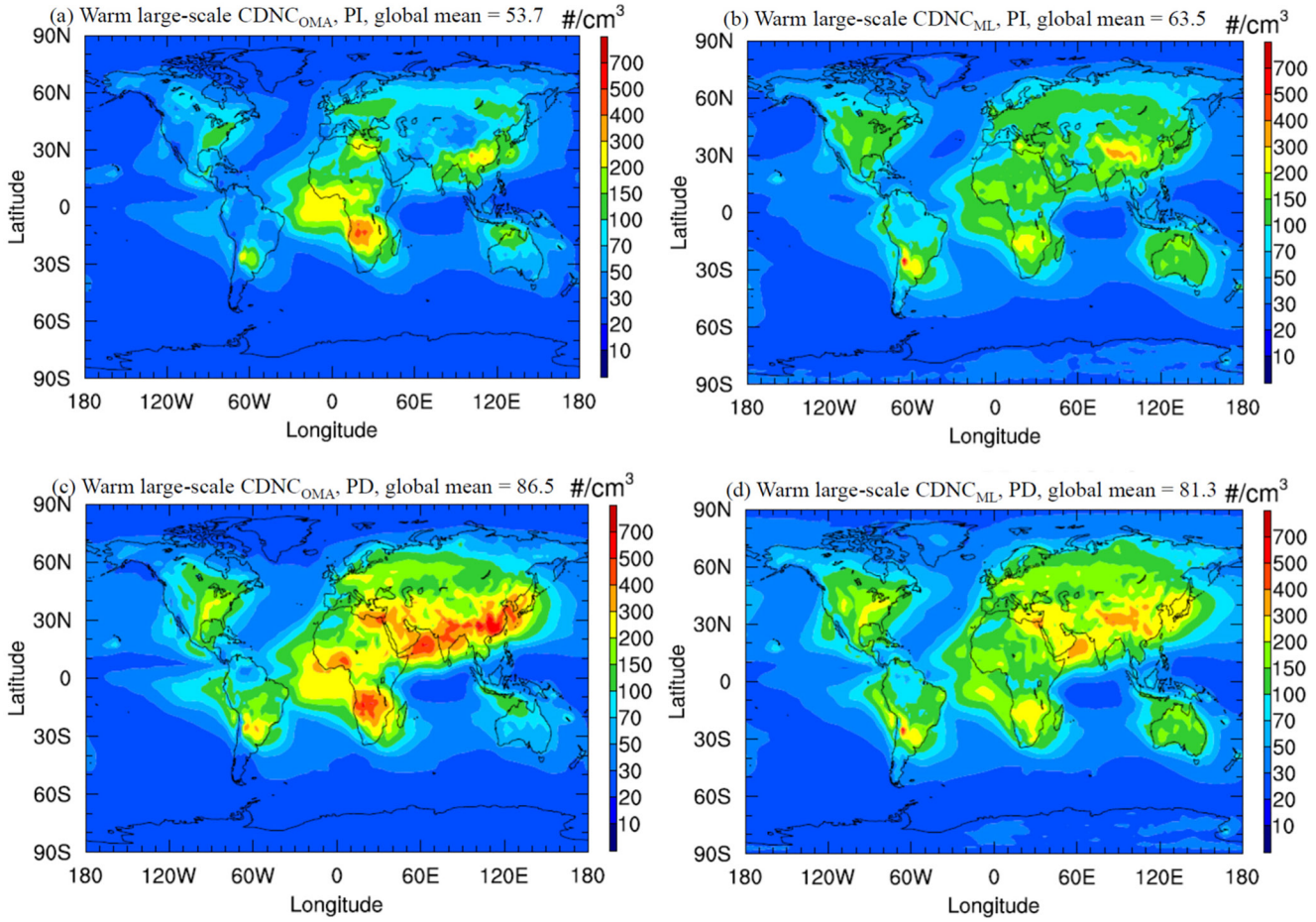


Figure 3. Cloud-cover weighted decadal mean (2005–2014) cloud droplet number concentration based on GISS ModelE2.1—One-Moment Aerosol with particle number concentration calculated with prescribed mass to number coefficients (a and c), and with the Random Forest Regression Model module (b and d) under the pre-industrial (a and b) and present-day (c and d) emission scenarios.

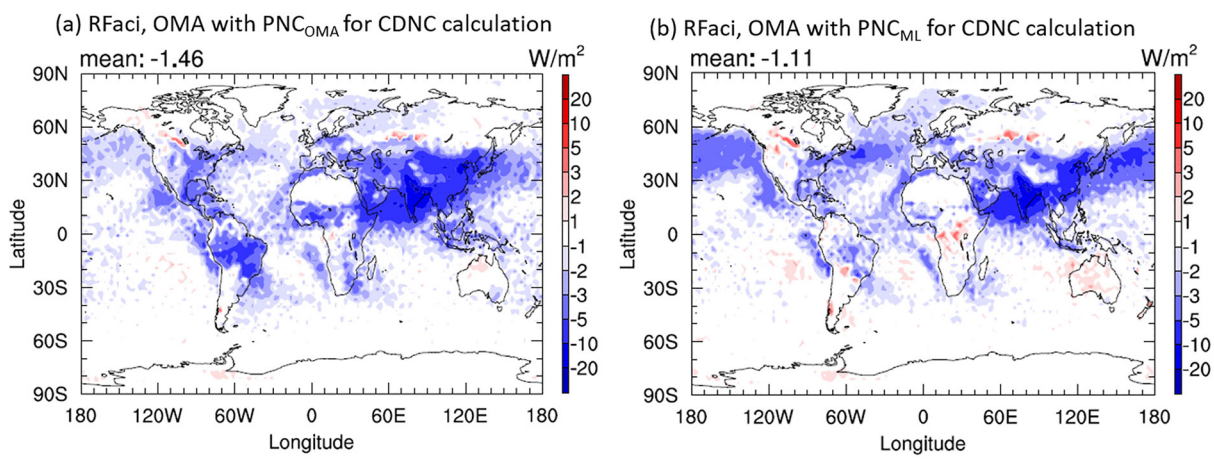


Figure 4. Decadal mean (2005–2014) RF$_{aci}$ based on GISS ModelE2.1—One-Moment Aerosol with particle number concentration calculated with prescribed mass to number coefficients (a), and with the Random Forest Regression Model module (b).

4. Summary and Discussion

PNC is one of the key parameters determining RF_{aci} but its calculation in climate models is generally simplified because of the lack of simulating size- and composition-resolved particle microphysics due to the associated high computing cost. In this study, we propose and demonstrate the ability for a more accurate representation of PNC important for RF_{aci}, while maintaining computing efficiency, by the development of a RFRM using a machine-learning approach trained with long-term simulations of a global size-resolved (sectional) aerosol microphysics model. The RFRM takes into account the complex dependence of PNC not only on mass concentrations of various aerosol components but also on other key variables representing meteorological and chemical conditions. We have implemented the PNC RFRM in GISS-E2-1-OMA. The PNC RFRM significantly improves the agreement of PNC predicted by GISS-E2-1-OMA with measurements across the globe (in terms of both spatial distributions and temporal variations), reduces the relative changes of cloud droplet number concentration associated with changes of emissions from pre-industrial to present-day, and decreases the RF_{aci} from -1.46 to -1.11 $W{\cdot}m^{-2}$. In addition, the simulations based on PNC RFRM also show different spatial distributions of CDNC and RF_{aci}.

This study highlights the sensitivity of RF_{aci} in GCMs to PNC calculation and the necessity to improve it. Our exploratory work shows that the RFRM, trained on outputs from a global model with full size-resolved particle microphysics, can be used to reduce uncertainties of climate models in predicting PNC and RF_{aci} without compromising their computing efficiency. Compared to the fully size-resolved microphysics model, which generally increases the computing cost by a factor of two or more, the RFRM only increases the computing cost by ∼5%. A number of factors affect the speed and accuracy of the RFRM, including the percentage of GCAPM outputs used for training, the number and specifics of selected predictor variables, the number of trees in the forest, the minimum number of variables to consider for each split, and the minimum node size (Nair & Yu, 2020). Further improvement of GCAPM and optimization may improve the accuracy yet reduce the computing cost of the RFRM. The number and specifics of selected predictor variables can be varied based on what is available in most models and observations (for validation). It should be noted that similar RFRMs for CCN, CDNC, and aerosol optical properties, which are all important for aerosol radiative forcing and depend on particle size distributions and compositions, can also be derived. These machine-learning algorithms will enable GCMs used for climate change studies to calculate parameters key for aerosol radiative forcing more robustly using commonly available and widely observed variables (and thus can be well-constrained), and thus can help reduce the diversity or uncertainties in climate change projections.

Data Availability Statement

The GISS-ModelE2.1 is available to the public at https://www.giss.nasa.gov/tools/modelE/. The GEOS-Chem model is available to the public at https://geos-chem.seas.harvard.edu/. The observation data used in this study (Figure 2), which was averaged using the raw data from the EBAS database and University at Albany Atmospheric Science Research Center's Air Quality Monitoring Products, is archived at https://doi.org/10.5281/zenodo.6960013.

Acknowledgments
This research has been supported by NASA (Grant Nos. 80NSSC19K1275 and NNX17AG35G) and NYSERDA (contract no. 137487). The authors thank the EBAS database teams (http://ebas.nilu.no/) and University at Albany Atmospheric Science Research Center's Air Quality Monitoring Products team (http://atmoschem.asrc.cestm.albany.edu/∼aqm/) for making their measurement data publicly available.

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

Bauer, S. E., & Koch, D. (2005). Impact of heterogeneous sulfate formation at mineral dust surfaces on aerosol loads and radiative forcing in the Goddard Institute for Space Studies general circulation model. *Journal of Geophysical Research*, *110*(D17), D17202. https://doi.org/10.1029/2005JD005870

Bauer, S. E., Tsigaridis, K., Faluvegi, G., Kelley, M., Lo, K. K., Miller, R. L., et al. (2020). Historical (1850–2014) aerosol evolution and role on climate forcing using the GISS ModelE2.1 contribution to CMIP6. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS001978. https://doi.org/10.1029/2019MS001978

Bauer, S. E., Wright, D. L., Koch, D., Lewis, E. R., McGraw, R., Chang, L. S., et al. (2008). MATRIX (Multiconfiguration aerosol TRacker of mIXing state): An aerosol microphysical module for global atmospheric models. *Atmospheric Chemistry and Physics*, *8*(20), 6003–6035. https://doi.org/10.5194/acp-8-6003-2008

Bellouin, N., Mann, G. W., Woodhouse, M. T., Johnson, C., Carslaw, K. S., & Dalvi, M. (2013). Impact of the modal aerosol scheme GLOMAP-mode on aerosol forcing in the Hadley Centre Global Environmental Model. *Atmospheric Chemistry and Physics*, *13*(6), 3027–3044. https://doi.org/10.5194/acp-13-3027-2013

Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., et al. (2001). Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research*, *106*(D19), 23073–23095. https://doi.org/10.1029/2001JD000807

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. https://doi.org/10.1201/9781315139470

Evans, M. J., & Jacob, D. J. (2005). Impact of new laboratory studies of N_2O_5 hydrolysis on global model budgets of tropospheric nitrogen oxides, ozone, and OH. *Geophysical Research Letters*, *32*(9), L09813. https://doi.org/10.1029/2005gl022469

Fanourgakis, G. S., Kanakidou, M., Nenes, A., Bauer, S. E., Bergman, T., Carslaw, K. S., et al. (2019). Evaluation of global simulations of aerosol particle and cloud condensation nuclei number, with implications for cloud droplet formation. *Atmospheric Chemistry and Physics*, *19*(13), 8591–8617. https://doi.org/10.5194/acp-19-8591-2019

Ghan, S. J. (2013). Estimating aerosol effects on cloud radiative forcing. *Atmospheric Chemistry and Physics*, *13*(19), 9971–9974. https://doi.org/10.5194/acp-13-9971-2013

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., & Hueglin, C. (2018). Random forest meteorological normalisation models for Swiss PM 10 trend analysis. *Atmospheric Chemistry and Physics*, *18*(9), 6223–6239. https://doi.org/10.5194/acp-18-6223-2018

Hansen, J., Russell, G., Rind, D., Stone, P., Lacis, A., Lebedeff, S., et al. (1983). Efficient three-dimensional global models for climate studies: Models I and II. *Monthly Weather Review*, *111*(4), 609–662. https://doi.org/10.1175/1520-0493(1983)111%3C0609:ETDGMF%3E2.0.CO;2

Hansen, J., Sato, M., Nazarenko, L., Ruedy, R., Lacis, A., Koch, D., et al. (2002). Climate forcings in Goddard Institute for space studies SI2000 simulations. *Journal of Geophysical Research*, *107*(D18), 4347. https://doi.org/10.1029/2001JD001143

Hansen, J., Sato, M., & Ruedy, R. (1997). Radiative forcing and climate response. *Journal of Geophysical Research*, *102*(D6), 6831–6864. https://doi.org/10.1029/96JD03436

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., et al. (2018). Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS). *Geoscientific Model Development*, *11*(1), 369–408. https://doi.org/10.5194/gmd-11-369-2018

Holmes, C. D., Bertram, T. H., Confer, K. L., Graham, K. A., Ronan, A. C., Wirks, C. K., & Shah, V. (2019). The role of clouds in the tropospheric NO_x cycle: A new modeling approach for cloud chemistry and its global implications. *Geophysical Research Letters*, *46*(9), 4980–4990. https://doi.org/10.1029/2019GL081990

Hughes, M., Kodros, J. K., Pierce, J. R., West, M., & Riemer, N. (2018). Machine learning to predict the global distribution of aerosol mixing state metrics. *Atmosphere*, *9*(1), 15. https://doi.org/10.3390/atmos9010015

Jin, J., Lin, H. X., Segers, A., Xie, Y., & Heemink, A. (2019). Machine learning for observation bias correction with application to dust storm data assimilation. *Atmospheric Chemistry and Physics*, *19*(15), 10009–10026. https://doi.org/10.5194/acp-19-10009-2019

Keller, C. A., Long, M. S., Yantosca, R. M., Da Silva, A. M., Pawson, S., & Jacob, D. J. (2014). HEMCO v1. 0: A versatile, ESMF-compliant component for calculating emissions in atmospheric models. *Geoscientific Model Development*, *7*(4), 1409–1417. https://doi.org/10.5194/gmd-7-1409-2014

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., et al. (2020). GISS-E2.1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS002025. https://doi.org/10.1029/2019MS002025

Koch, D., Schmidt, G. A., & Field, C. V. (2006). Sulfur, sea salt, and radionuclide aerosols in GISS ModelE. *Journal of Geophysical Research*, *111*(D6), D06206. https://doi.org/10.1029/2004JD005550

Luo, G., & Yu, F. (2010). A numerical evaluation of global oceanic emissions of alpha-pinene and isoprene. *Atmospheric Chemistry and Physics*, *10*(4), 2007–2015. https://doi.org/10.5194/acp-10-2007-2010

Luo, G., Yu, F., & Moch, J. M. (2020). Further improvement of wet process treatments in GEOS-chem v12. 6.0: Impact on global distributions of aerosols and aerosol precursors. *Geoscientific Model Development*, *13*(6), 2879–2903. https://doi.org/10.5194/gmd-13-2879-2020

Martin, R. V., Jacob, D. J., Chance, K., Kurosu, T. P., Palmer, P. I., & Evans, M. J. (2003). Global inventory of nitrogen oxide emissions constrained by space-based observations of NO_2 columns. *Journal of Geophysical Research*, *108*(D17), 4537. https://doi.org/10.1029/2003JD003453

Mauceri, S., Kindel, B., Massie, S., & Pilewskie, P. (2019). Neural network for aerosol retrieval from hyperspectral imagery. *Atmospheric Measurement Techniques*, *12*(11), 6017–6036. https://doi.org/10.5194/amt-12-6017-2019

Menon, S., & Rotstayn, L. (2006). The radiative influence of aerosol effects on liquid-phase cumulus and stratus clouds based on sensitivity studies with two climate models. *Climate Dynamics*, *27*, 345–356.

Miller, D. J., Segal-Rozenhaimer, M., Knobelspiesse, K., Redemann, J., Cairns, B., Alexandrov, M., et al. (2020). Low-level liquid cloud properties during ORACLES retrieved using airborne polarimetric measurements and a neural network algorithm. *Atmospheric Measurement Techniques*, *13*(6), 3447–3470. https://doi.org/10.5194/amt-13-3447-2020

Miller, R. L., Cakmur, R. V., Perlwitz, J., Geogdzhayev, I. V., Ginoux, P., Koch, D., et al. (2006). Mineral dust aerosols in the NASA Goddard Institute for Space Sciences ModelE atmospheric general circulation model. *Journal of Geophysical Research*, *111*(D6), D06208. https://doi.org/10.1029/2005JD005796

Miller, R. L., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Kelley, M., Ruedy, R., et al. (2021). CMIP6 historical simulations (1850–2014) with GISS-E2. 1. *Journal of Advances in Modeling Earth Systems*, *13*(1), e2019MS002034. https://doi.org/10.1029/2019MS002034

Murray, L. T., Jacob, D. J., Logan, J. A., Hudman, R. C., & Koshak, W. J. (2012). Optimized regional and interannual variability of lightning in a global chemical transport model constrained by LIS/OTD satellite data. *Journal of Geophysical Research*, *117*(D20). https://doi.org/10.1029/2012JD017934

Nair, A. 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-2020-509

Nair, A. A., Yu, F., Campuzano-Jost, P., DeMott, P. J., Levin, E. J. T., 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

Nazarenko, L., Rind, D., Tsigaridis, K., Del Genio, A. D., Kelley, M., & Tausnev, N. (2017). Interactive nature of climate change and aerosol forcing. *Journal of Geophysical Research: Atmospheres*, *122*(6), 3457–3480. https://doi.org/10.1002/2016JD025809

Pye, H. O., & Seinfeld, J. H. (2010). A global perspective on aerosol from low-volatility organic compounds. *Atmospheric Chemistry and Physics*, *10*(9), 4377–4401. https://doi.org/10.5194/acp-10-4377-2010

Ramanathan, V. C. P. J., Crutzen, P. J., Kiehl, J. T., & Rosenfeld, D. (2001). Aerosols, climate, and the hydrological cycle. *Science*, *294*(5549), 2119–2124. https://doi.org/10.1126/science.1064034

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., et al. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, *108*(D14), 4407. https://doi.org/10.1029/2002JD002670., .

Rotstayn, L. D., Plymin, E. L., Collier, M. A., Boucher, O., Dufresne, J. L., Luo, J. J., et al. (2014). Declining aerosols in CMIP5 projections: Effects on atmospheric temperature structure and midlatitude jets. *Journal of Climate*, *27*(18), 6960–6977. https://doi.org/10.1175/jcli-d-14-00258.1

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., et al. (2014). Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, *6*(1), 141–184. https://doi.org/10.1002/2013MS000265

Schmidt, G. A., Ruedy, R., Hansen, J. E., Aleinov, I., Bell, N., Bauer, M., et al. (2006). Present-day atmospheric simulations using GISS ModelE: Comparison to in situ, satellite, and reanalysis data. *Journal of Climate*, *19*(2), 153–192. https://doi.org/10.1175/JCLI3612.1

Su, H., Wu, L., Jiang, J. H., Pai, R., Liu, A., Zhai, A. J., et al. (2020). Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophysical Research Letters*, *47*(17), e2020GL089102. https://doi.org/10.1029/2020GL089102

Tonttila, J., Järvinen, H., & Räisänen, P. (2015). Explicit representation of subgrid variability in cloud microphysics yields weaker aerosol indirect effect in the ECHAM5-HAM2 climate model. *Atmospheric Chemistry and Physics*, *15*(2), 703–714. https://doi.org/10.5194/acp-15-703-2015

Tsigaridis, K., Koch, D., & Menon, S. (2013). Uncertainties and importance of sea spray composition on aerosol direct and indirect effects. *Journal of Geophysical Research: Atmospheres*, *118*(1), 220–235. https://doi.org/10.1029/2012JD018165

Twomey, S. (1977). The influence of pollution on the shortwave albedo of clouds. *Journal of the Atmospheric Sciences*, *34*(7), 1149–1152. https://doi.org/10.1175/1520-0469(1977)034%3C1149:tiopot%3E2.0.co;2

van Donkelaar, A., Martin, R. V., Leaitch, W. R., Macdonald, A. M., Walker, T. W., Streets, D. G., et al. (2008). Analysis of aircraft and satellite measurements from the Intercontinental Chemical Transport Experiment (INTEX-B) to quantify long-range transport of East Asian sulfur to Canada. *Atmospheric Chemistry and Physics*, *8*(11), 2999–3014. https://doi.org/10.5194/acp-8-2999-2008

Williamson, C. J., Kupc, A., Axisa, D., Bilsback, K. R., Bui, T., Campuzano-Jost, P., et al. (2019). A large source of cloud condensation nuclei from new particle formation in the tropics. *Nature*, *574*(7778), 399–403. https://doi.org/10.1038/s41586-019-1638-9

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C^{++} and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

Yu, F. (2011). A secondary organic aerosol formation model considering successive oxidation aging and kinetic condensation of organic compounds: Global scale implications. *Atmospheric Chemistry and Physics*, *11*(3), 1083–1099. https://doi.org/10.5194/acp-11-1083-2011

Yu, F., & Luo, G. (2009). Simulation of particle size distribution with a global aerosol model: Contribution of nucleation to aerosol and CCN number concentrations. *Atmospheric Chemistry and Physics*, *9*(20), 7691–7710. https://doi.org/10.5194/acp-9-7691-2009

Yu, F., Luo, G., Bates, T. S., Anderson, B., Clarke, A., Kapustin, V., et al. (2010). Spatial distributions of particle number concentrations in the global troposphere: Simulations, observations, and implications for nucleation mechanisms. *Journal of Geophysical Research*, *115*(D17), D17205. https://doi.org/10.1029/2009JD013473

Yu, F., Luo, G., Nadykto, A. B., & Herb, J. (2017). Impact of temperature dependence on the possible contribution of organics to new particle formation in the atmosphere. *Atmospheric Chemistry and Physics*, *17*(8), 4997–5005. https://doi.org/10.5194/acp-17-4997-2017

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

Yu, F., Nadykto, A. B., Herb, J., Luo, G., Nazarenko, K. M., & Uvarova, L. A. (2018). 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*(23), 17451–17474. https://doi.org/10.5194/acp-18-17451-2018

Yu, F., Nadykto, A. B., Luo, G., & Herb, J. (2020). H 2 SO 4–H 2 O binary and H_2SO_4–H_2O–NH_3 ternary homogeneous and ion-mediated nucleation: Lookup tables version 1.0 for 3-D modeling application. *Geoscientific Model Development*, *13*(6), 2663–2670. https://doi.org/10.5194/gmd-13-2663-2020

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

Zanis, P., Akritidis, D., Georgoulias, A. K., Allen, R. J., Bauer, S. E., Boucher, O., et al. (2020). Fast responses on pre-industrial climate from present-day aerosols in a CMIP6 multi-model study. *Atmospheric Chemistry and Physics*, *20*(14), 8381–8404. https://doi.org/10.5194/acp-20-8381-2020

Zhang, H., Zhao, S., Wang, Z., Zhang, X., & Song, L. (2016). The updated effective radiative forcing of major anthropogenic aerosols and their effects on global climate at present and in the future. *International Journal of Climatology*, *36*(12), 4029–4044. https://doi.org/10.1002/joc.4613