JGR Atmospheres

RESEARCH ARTICLE
10.1029/2025JD043926

Using a Machine Learning Approach for Isolating Aerosol Effects on Cloud Droplet Number Concentration in Marine Stratocumulus Clouds

Muhammed Irfan[1], Antti Lipponen[2], Thomas Kühn[3], Sami Romakkaniemi[2], Silvia M. Calderón[2], Eemeli Holopainen[2,4], Annele Virtanen[1], Antti Arola[2], and Harri Kokkola[1,2]

[1]Department of Technical Physics, University of Eastern Finland, Kuopio, Finland, [2]Atmospheric Research Centre of Eastern Finland, Finnish Meteorological Institute, Kuopio, Finland, [3]Weather and Climate Change Research, Finnish Meteorological Institute, Helsinki, Finland, [4]Institute of Chemical Engineering Sciences, Foundation for Research and Technology – Hellas (FORTH/ICE-HT), Patras, Greece

Key Points:

- Elastic Net Regression effectively estimates cloud droplet response to aerosol changes from satellite data, despite multiple factors
- Traditional regression methods can misrepresent aerosol–cloud susceptibility due to measurement uncertainties and oversimplified assumptions
- Variability in cloud activation updrafts conceals the cloud condensation nuclei-cloud droplet number concentration correlation

Supporting Information:

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

Correspondence to:

M. Irfan and H. Kokkola,
mirfan@uef.fi;
harri.kokkola@fmi.fi

Citation:

Irfan, M., Lipponen, A., Kühn, T., Romakkaniemi, S., Calderón, S. M., Holopainen, E., et al. (2025). Using a machine learning approach for isolating aerosol effects on cloud droplet number concentration in marine stratocumulus clouds. *Journal of Geophysical Research: Atmospheres*, *130*, e2025JD043926. https://doi.org/10.1029/2025JD043926

Received 25 MAR 2025
Accepted 24 JUL 2025

Author Contributions:

Conceptualization: Muhammed Irfan, Harri Kokkola
Data curation: Muhammed Irfan, Eemeli Holopainen
Formal analysis: Muhammed Irfan, Thomas Kühn, Antti Arola, Harri Kokkola
Funding acquisition: Annele Virtanen, Harri Kokkola
Investigation: Muhammed Irfan
Methodology: Muhammed Irfan, Antti Lipponen, Harri Kokkola
Project administration: Annele Virtanen, Harri Kokkola
Resources: Annele Virtanen

Abstract Understanding the complex relationship between aerosols and clouds is crucial for accurately estimating anthropogenic aerosol radiative forcing and its impact on weather and climate. However, quantifying the relationship between cloud droplet number concentration (CDNC) and cloud condensation nuclei (CCN) concentration remains challenging due to extensive heterogeneities in cloud properties, processes occurring during cloud lifecycle, and observational uncertainties. This study integrates satellite observations with a global climate model and employs advanced statistical techniques to improve estimates of cloud droplet susceptibility to aerosol perturbations, particularly focusing on the CCN-CDNC slope. A key challenge in determining this relationship is that CDNC is influenced by factors beyond CCN alone, such as atmospheric dynamics and cloud microphysics. These additional factors introduce variability that complicates direct correlation between CCN and CDNC, making it difficult to ascertain their true relationship. In addition, traditional methods like Ordinary Least Squares regression can produce biased slope estimates due to uncertainties in both CCN and CDNC measurements. When applied to CCN-CDNC data, advanced curve fitting methods, such as bivariate least squares, often yield slopes exceeding 1, deviating from expected physical behavior. Our machine learning analysis identifies updraft velocity, among other key predictors, as a major factor leading to this larger-than-one slope between CCN and CDNC. To address this, we utilized Elastic Net Regression to isolate the effect of changes in CCN concentration on CDNC. This method refines the slope estimates by accounting for factors affecting CDNC, resulting in a slope that better captures the susceptibility of CDNC to CCN changes.

Plain Language Summary Clouds form when tiny atmospheric particles, that is, aerosols, take up water vapor and form cloud droplets. An accurate understanding of how many droplets form from a given number of aerosol particles is essential to better predict how emissions of aerosols affect clouds and consequently weather and climate. However, estimation of this relationship is often challenging because multiple processes affect the formation of cloud droplets making it difficult to isolate the effect of aerosols on cloud droplets. In this study, we used statistical techniques and machine learning to improve our understanding of how the number of aerosol particles and number of cloud droplets are related. We found that traditional methods struggled to capture this relationship, leading to large uncertainties in our understanding of interactions between aerosols and clouds. By applying machine learning techniques, we identified key factors affecting cloud formation and refined our estimates of how aerosols influence clouds. This improved understanding matches what we, based on theory, expect to see in the atmosphere. Overall, this study advances our understanding of aerosols and clouds, which is essential for making more accurate weather and climate predictions.

1. Introduction

Aerosol-cloud interactions play a crucial role in Earth's climate system, significantly influencing a range of atmospheric processes and climate dynamics (Bellouin et al., 2019). Aerosols, particularly cloud condensation nuclei (CCN), are essential for cloud formation, influencing cloud droplet number concentration (CDNC), radiative forcing, precipitation patterns, and atmospheric dynamics (Seinfeld & Pandis, 2016; Stier et al., 2024; Wang, Li, et al., 2024; Wang et al., 2023). Therefore, understanding the relationship between CCN and CDNC is

Software: Muhammed Irfan, Sami Romakkaniemi, Silvia M. Calderón, Harri Kokkola
Supervision: Antti Lipponen, Thomas Kühn, Harri Kokkola
Validation: Muhammed Irfan, Antti Lipponen, Thomas Kühn, Sami Romakkaniemi, Annele Virtanen, Antti Arola, Harri Kokkola
Visualization: Muhammed Irfan
Writing – original draft: Muhammed Irfan
Writing – review & editing: Antti Lipponen, Thomas Kühn, Sami Romakkaniemi, Silvia M. Calderón, Eemeli Holopainen, Annele Virtanen, Antti Arola, Harri Kokkola

essential for elucidating the mechanisms affecting aerosol-cloud interactions and their broader implications for climate processes (Jia & Quaas, 2023; Wang, Jia, et al., 2024). The susceptibility of cloud droplets to changes in aerosol concentrations, as reflected in the CCN-CDNC relationship, holds immense significance. This susceptibility is a fundamental aspect of aerosol-cloud interactions, influencing cloud properties and processes (Wang, Li, et al., 2024). Moreover, it remains a key source of uncertainty in estimates of effective radiative forcing due to aerosol-cloud interactions (Quaas et al., 2009). Despite the significance of this relationship, accurately quantifying the effect of changes in CCN concentration on CDNC remains challenging, given the complex interplay of multiple atmospheric processes and retrieval uncertainties associated with remote sensing observations of relevant aerosol and cloud properties (Arola et al., 2022; Gryspeerdt et al., 2023; Kokkola et al., 2025; Quaas et al., 2009).

One notable inherent problem is that analyses using satellite retrievals to estimate CCN-CDNC relationships are typically limited to cloud top measurements of CDNC, while the CCN values used represent atmospheric column-integrated burdens of cloud-free pixels next to the actual cloud fields. In reality, CCN affects CDNC at the cloud base, where the cloud droplets are formed. While the assumptions used to extract meaningful CCN-CDNC relationships usually hold for warm clouds in maritime conditions (Grosvenor et al., 2018), it is to be expected that this also introduces uncertainties into the analysis. With the help of computer models, such assumptions can be tested and the skill of different regression techniques to account for different kinds of uncertainties can be assessed.

Ordinary Least Squares (OLS) regression is one of the widely used methods to estimate the susceptibility of cloud droplets to aerosol perturbations (Bai et al., 2018; Hasekamp et al., 2019; McComiskey & Feingold, 2008; McCoy et al., 2017; S.-Y. Park & Kim, 2021; Ramanathan et al., 2001). However, in its two-dimensional (2D) form, OLS regression can produce spurious results when there is variability or uncertainty in observations of both CCN (x-direction) and CDNC (y-direction), potentially leading to biased slope estimates (Mikkonen et al., 2019; Pitkänen et al., 2016). Such biases often arise from regression dilution, where the uncertainty in the x variable distorts the estimated relationship with y (Pitkänen et al., 2016). This problem is compounded by the fact that OLS primarily minimizes the vertical distance between data points and the fitted line, assuming zero uncertainty for the independent variable on the x-axis. As a result, when OLS is applied in the presence of uncertainty in x, the fitted slope tends to be biased toward zero (Frost & Thompson, 2000; Pitkänen et al., 2016). This problem is particularly pronounced in atmospheric data, which inherently include some degree of error, further complicating the accurate estimation of relationships such as those between CCN and CDNC (Pitkänen et al., 2016). Consequently, relying solely on OLS regression to examine aerosol-cloud interactions may lead to incomplete or misleading conclusions. To address these challenges, in particular to account for the uncertainties in both x and y directions, alternative regression methods such as bivariate least squares (BLS) method suggested by York et al. (2004) may offer a more robust solution than OLS. These methods have shown to provide a better understanding of the complex dynamics governing the relationship between CCN and CDNC, despite the presence of inherent variability within the data sets (Mikkonen et al., 2019).

In addition to the challenges surrounding the quantification of aerosol-cloud interactions, the variability in cloud microphysics and meteorological conditions play a crucial role in shaping the effects of aerosols on cloud formation. Two major factors in cloud droplet activation are the updraft velocity during cloud activation and the number of CCN (Pruppacher et al., 1998; Seinfeld & Pandis, 2016). The updraft velocity affects the water vapor supersaturation reached at cloud base thus affecting which sized aerosol particles are activated to form cloud droplets (Sullivan et al., 2016). Reutter et al. (2009) identified regimes where cloud formation is primarily dependent on either updraft velocity or aerosol particle number concentration. The aerosol-limited regime is associated with high updraft velocities and low aerosol concentrations, where CDNC is directly proportional to the CCN concentration, with CCN-CDNC relations approaching slopes of 1. In this regime, CDNC is insensitive to changes in updraft velocities because the high updraft velocities result in maximum supersaturations (i.e., the maximum relative humidity an air parcel reaches during its ascent through the cloud) sufficient to activate nearly all aerosol particles, except for the very small ones at the lower end of the aerosol size distribution. Conversely, the updraft-limited regime is associated with low updraft velocities and high aerosol concentrations. In this regime, CDNC exhibits a linear dependence on updraft velocity and shows weaker dependence on CCN concentrations. The relatively low updraft velocities and high aerosol concentrations produce maximum supersaturations so small that only large and hygroscopic particles are activated. There is also a transitional regime

characterized by sensitivity to both aerosol concentration and updraft velocity, where CDNC exhibits nonlinear dependencies on both factors. Additionally, factors such as meteorological conditions, aerosol properties, cloud dynamics, and entrainment-mixing can further shape the effects of aerosols on CDNC variability (Gao et al., 2021; Khatri et al., 2023; Liu et al., 2018; Lu et al., 2012). These variations in CDNC can subsequently influence other important cloud properties, such as the liquid water path (LWP), which represents the total amount of liquid water in the atmospheric column.

To address such complexities involved in isolating the effects of aerosols on cloud droplet formation, modern analytical techniques such as machine learning are increasingly being utilized (Gao et al., 2024; Kalbande et al., 2023; Kumar et al., 2022; Li et al., 2024; Miinalainen et al., 2023; Redemann & Gao, 2024). Machine learning approaches are very useful in discerning patterns from large and complex data sets by handling nonlinearity and interdependencies among a multitude of variables. This capability makes machine learning very well-suited for studying aerosol-cloud interactions, where multiple factors like aerosol concentrations, updraft velocity, and meteorological conditions, simultaneously govern cloud microphysical properties. Thus, by applying suitable machine learning methods, it is possible to study how these variables contribute to the variability of the target variable, which in this case, refers to CDNC, providing a nuanced understanding of their relative importance.

This study focuses on marine stratocumulus clouds, one of the most common cloud types prevalent over oceanic regions (Warren et al., 1986). With their high albedo and extensive coverage, stratocumulus clouds are of particularly importance in regulating the Earth's energy balance. They are highly sensitive to variations in the concentration of aerosols, thus making them suitable for studying the impacts induced by anthropogenic emissions (Calderón et al., 2022; Toll et al., 2019; Wilcox, 2010). Unlike continental cloud systems, marine stratocumulus clouds typically form under more stable boundary layer conditions. However, even within these homogenous environments, factors such as the local meteorological variability and aerosol microphysical properties add complexities that must be addressed to fully understand aerosol-cloud interactions. The integration of machine learning methods with observations of marine stratocumulus clouds enables to identify the key predictors of CDNC and helps in the investigation of complex relationships between aerosols and clouds.

In this study, we perform a comprehensive analysis of aerosol-cloud interactions over marine stratocumulous clouds by integrating satellite observations with those derived from the ECHAM-HAMMOZ global aerosol-climate model (Schultz et al., 2018). Our primary objective is to study the relationship between CCN and CDNC, and to identify a regression method that accurately captures this relationship. Deriving a single, aggregate slope between CCN and CDNC is a widely used approach, particularly for estimating the effective radiative forcing due to aerosol–cloud interactions (ERF_{aci}) from satellite observations and global climate models (Bellouin et al., 2019; Gryspeerdt et al., 2017; Ma et al., 2018; Quaas et al., 2009; Virtanen et al., 2025). These susceptibility metrics are often used to scale or constrain model-derived ERF_{aci}, making the robustness and accuracy of the slope estimate critically important. Although the CCN–CDNC relationship varies with meteorological conditions such as boundary layer depth, many large-scale applications rely on a representative mean slope to summarize this sensitivity across heterogeneous cloud regimes. This study aims to estimate this slope using a physically and statistically grounded framework, while systematically evaluating the limitations of widely employed traditional regression approaches. To achieve this, we first evaluate various regression techniques to see if they accurately reflect the true CCN-CDNC relationship as derived from satellite data. We then extend this analysis to climate model data to assess whether similar results are obtained with different regression approaches. Additionally, we explore different machine learning methods to identify key factors influencing the CCN-CDNC relationship and investigate whether these factors are consistent across satellite observations and climate model simulations. Furthermore, we use a machine-learning based regularization method, Elastic Net Regression (ENR) to estimate the slope between CCN and CDNC. This approach allows us to incorporate different forms of regularization, potentially improving the estimation of the CCN-CDNC relationship by considering several factors affecting the CDNC variability.

2. Data and Methodology

In this section, we provide an overview of the satellite-retrieved aerosol and cloud properties, model simulations, and the various techniques used in this study to investigate the CCN-CDNC relationship. Our study focused on two specific regions within the Pacific Ocean known for their high annual coverage of low-level marine

clouds: the Northern Pacific (20–35°N, 150–110°W) and the Southern Pacific (10–30°S, 110–70°W). By selecting these well studied marine regions, we aim to explore aerosol-cloud interactions in environments characterized by distinct meteorological conditions and cloud properties (Bellomo et al., 2014). Our analysis included satellite observations and climate model data corresponding to the year 2015 over the specified regions.

2.1. Satellite Observations and Reanalysis Data

We utilized cloud products from Moderate Resolution Imaging Spectroradiometer (MODIS) Level-2 (L2) Collection 6.1 (Platnick et al., 2017) onboard the Aqua and Terra satellites (MOD06_L2 and MYD06_L2) to obtain key cloud properties, including cloud optical depth (COD), cloud effective radius (CER), cloud top temperature (CTT), and cloud top height (CTH). Additionally, we incorporated aerosol products over the ocean (MOD04_L2 and MYD04_L2), which provide aerosol optical depth (AOD) and Ångström exponent (AE) (Levy et al., 2013). For the satellite-derived CCN, we relied on the MODIS CCN data, PSML003_Ocean, which provides the column-integrated burden of fine-mode particles over the ocean (Gassó & Hegg, 2003; Levy et al., 2013). This data set was chosen over other widely used CCN proxies, such as AOD and aerosol index, which are used as inputs in the MODIS CCN retrieval, as it facilitates better comparison between the model and satellite data. However, aerosol index, defined as the product of AOD and AE, was incorporated into this study as a supplementary metric to support the analysis (Hasekamp et al., 2019). By utilizing aerosol index, which accounts for the spectral dependence of AOD, we aimed to enhance the robustness of our analysis and further strengthen our argument regarding aerosol-cloud interaction mechanisms. Both cloud and aerosol data were aggregated to a spatial resolution of 1° × 1° to reduce the pixel-level variability, thereby improving the robustness of our analysis (Grosvenor et al., 2018).

We then calculated CDNC and LWP from satellite-derived COD and CER as suggested by Quaas et al. (2006) using the following equations:

$$\mathrm{CDNC} = \alpha \times \tau_c^{0.5} \times r_e^{-2.5} \tag{1}$$

$$\mathrm{LWP} = \frac{5}{9}\rho_w \times r_e \times \tau_c, \tag{2}$$

where $\alpha = 1.37 \times 10^{-5}\mathrm{m}^{-0.5}$ is a constant, ρ_w is the density of liquid water, typically assumed as 1,000 kg m^{-3}, and τ_c and r_e are the COD and CER, respectively. In our calculations of CDNC and LWP, we followed the adiabatic assumptions as outlined in Quaas et al. (2006) and Gryspeerdt et al. (2019), where the cloud liquid water content profile is a constant fraction of its adiabatic value and CDNC is constant over the whole cloud depth. This method also assumes that the CER is proportional to the volume mean radius, implying a constant cloud droplet size distribution within the area of the pixel and vertically across the cloud. Although, in reality, the width of the droplet size distribution may vary particularly under extreme aerosol conditions (Grosvenor et al., 2018; Lu et al., 2020; Wang et al., 2023), our approach assumes a constant width for droplet size distribution. Previous studies have shown that the uncertainties arising from this assumption are likely partially compensated by other parameters choices (Gryspeerdt et al., 2022; Merk et al., 2016; Painemal & Zuidema, 2010). It is also worth noting the filtering criteria used may introduce a sampling bias toward scenes with high cloud fraction and high optical thickness. As noted by Jia et al. (2022), retrievals in broken or partially cloudy scenes can violate key assumptions (e.g., 1D plane-parallel geometry), which can result in overestimation of CER and, consequently, underestimation of CDNC. While this limits the representativeness of our data set across more heterogeneous cloud regimes, it helps ensure the physical consistency of the retrieved CDNC values.

Along with the satellite data, we also used ERA-5 reanalysis data, which is generated from the European Center for Medium-Range Weather Forecasts Integrated Forecast System (Hersbach et al., 2020). This data set provided essential meteorological variables, such as sea surface temperature, sea surface pressure, 10-m winds, air temperature, and relative humidity.

2.2. Aerosol—Climate Model

The 3-hourly outputs from aerosol-climate model ECHAM-HAMMOZ (ECHAM6.3-HAM2.3) (Schultz et al., 2018) are used to support our investigation into the satellite-derived aerosol-cloud interactions. ECHAM is

a global atmospheric general circulation model that solves the equations of motion and continuity for the atmosphere (Stevens et al., 2013). For our simulations, we employed the T63 spectral truncation in horizontal resolution, corresponding to a grid spacing of approximately 1.9° × 1.9°, and utilized 47 hybrid sigma-pressure levels for vertical resolution. The aerosol life-cycle processes in ECHAM were simulated using HAM (Hamburg Aerosol Module) (Tegen et al., 2019), which is responsible for simulating aerosol formation, growth, and removal within the atmosphere. Furthermore, ECHAM-HAM is coupled with the aerosol microphysical model SALSA (Sectional Aerosol Module for Large Scale Applications) (Kokkola et al., 2018). The sectional approach used in SALSA was preferred over the modal approach due to its ability to accurately capture the variability in CDNC in varying conditions for cloud activation, as demonstrated by Korhola et al. (2014). SALSA divides the aerosol size distribution into 10 discrete size classes ranging from 3 nm to 10 μm. Cloud droplet activation in SALSA was solved using the parameterization described by Abdul-Razzak and Ghan (2002). In ECHAM-HAMMOZ, we use N_{100}, that is, the sum of all aerosol particles with diameters larger than 100 nm, as proxy for the CCN concentration. CDNC is calculated explicitly. To maintain consistency with satellite-retrieved variables, we also define CCN as column-burden CCN and CDNC as cloud-top CDNC in ECHAM, unless stated otherwise.

2.3. Regression Methods Used in This Study

To analyze the relationship between the logarithms of CCN and CDNC, we utilized two regression techniques: OLS and BLS. We specifically applied OLS regression in its 2D form, where two variables are considered in the relationship, with one as the independent variable and the other as the dependent variable. OLS is a commonly used method that estimates the relationship between two variables by minimizing the sum of squared differences between observed (y_i) and predicted ($\hat{y}_i$) values of the dependent variable (Dismuke & Lindrooth, 2006). The OLS regression model is given by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{3}$$

where y_i represents the log-transformed CDNC values, x_i represents the log-transformed CCN values, β_0 is the intercept, β_1 is the slope, ϵ_i is the error term. The parameters β_0 and β_1 are estimated by minimizing the sum of squared residuals.

$$C_{\mathrm{OLS}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{4}$$

However, OLS assumes that the independent variable (x_i) is measured without error, which can introduce inaccuracies when there are uncertainties in the x variable.

To address the limitations of OLS, we also used the BLS method following (York et al., 2004), which accounts for uncertainties in both x and y variables. This method minimizes a weighted cost function based on the deviations between the observed and adjusted values, ensuring that the adjusted points lie exactly on the best-fit regression line. The model is represented as:

$$C_{\mathrm{york}} = \sum_{i=1}^{n} \frac{1}{1 - r_i^2} \left\{ w_{x_i} (x_{i,\mathrm{adj}} - x_i)^2 - 2 r_i \sqrt{w_{x_i} w_{y_i}} (x_{i,\mathrm{adj}} - x_i)(y_{i,\mathrm{adj}} - y_i) + w_{y_i} (y_{i,\mathrm{adj}} - y_i)^2 \right\}, \tag{5}$$

subject to

$$y_{i,\mathrm{adj}} = \alpha_{\mathrm{york}} + \beta_{\mathrm{york}} x_{i,\mathrm{adj}}. \tag{6}$$

where, x_i and y_i are the observed data points, while $x_{i,\mathrm{adj}}$ and $y_{i,\mathrm{adj}}$ are their adjusted counterparts (Mikkonen et al., 2019). The intercept and slope of this line are denoted by α_{york} and β_{york}, respectively. The weight coefficients are $w_{x_i} = 1/\sigma_x^2$ and $w_{y_i} = 1/\sigma_y^2$, where σ_x^2 and σ_y^2 represent the error variances in the x and y directions, respectively The term r_i represents the correlation coefficient between x_i and y_i. The parameters α_{york} and β_{york} are estimated iteratively, beginning with an initial slope estimate, often derived from ordinary least squares, and

refined until convergence. Further details on the OLS and York regression methods can be found in Mikkonen et al. (2019) and York et al. (2004).

2.4. Feature Importance Analysis

We employed Random Forest (RF), a powerful ensemble learning technique known for its robustness in feature selection (Breiman, 2001), to identify the most influential variables that explain the CDNC variability. RF builds a multitude of decision trees during training and aggregates their predictions to provide insights into the importance of each feature. For our analysis, we utilized the scikit-learn Python library (Pedregosa et al., 2011), which offers an extensive range of tools for machine learning tasks. The RF method allowed us to efficiently handle large data sets and extract relevant features that significantly influence the variability in CDNC. The RF model has been widely utilized in aerosol-cloud interaction studies, showing its robustness in predicting various parameters, such as aerosol number concentration, and revealing the underlying dependence of different feature variables on the target variable (Y. Chen et al., 2022; Miinalainen et al., 2023; Nair & Yu, 2020).

The feature selection process using the RF algorithm was applied separately to two data sets: satellite observations and climate model outputs. Each data set was randomly divided into training and test subsets to ensure unbiased evaluation of model performance. Specifically, two-thirds of the data from each data set were allocated to the training subset, while the remaining one-third were reserved for testing. We performed hyperparameter tuning to optimize the performance of the RF model separately for both data sets. The grid-search approach was used to find the best combination of hyperparameters, while model performance was assessed with K-fold cross-validation (10-fold). Key hyperparameters, such as max_depth (the maximum depth of each decision tree) and n_estimators (the number of trees in the forest), were tuned for each of the data sets. For the climate model data, max_depth was equal to 50, while n_estimators was fixed at 200. The relative root mean squared error was used as the primary metric to evaluate the predictive accuracy of the RF model. This choice allowed for a normalized comparison of prediction errors across data sets with differing scales. Thus, by tuning the hyperparameters and validating through cross-validation, the RF model was optimized to effectively capture the distinct relationships in both data sets, ensuring reliable feature selection and prediction performance.

In addition, SHapley Additive exPlanations (SHAP) analysis was used for validating the feature importance rankings derived from the RF model (Lundberg & Lee, 2017). SHAP provides an interpretable framework to assess the contribution of individual features to model predictions, ensuring the consistency and robustness of the selected predictors across both data sets. To further validate the results, permutation feature importance (Altmann et al., 2010) was also applied as a complementary approach. This method involves shuffling feature values and measuring the resulting decrease in model performance, providing an additional layer of confidence in the identified feature importance rankings. The combination of SHAP and permutation feature importance ensured that the selected features were both statistically robust and aligned with model interpretability.

The input features selected for the feature importance analysis include LWP, CCN burden, updraft velocity, cloud base height, temperature, wind speed, relative humidity, surface temperature, and pressure, with CDNC as the target variable. Note that we included LWP as predictor variable, despite its non-deterministic influence on CDNC, because LWP can be considered as a proxy for clouds at different meteorological conditions, such as different updrafts and different geometric thicknesses (Kokkola et al., 2025). An overview of the variables used for feature importance analysis is given in Table 1.

2.5. Machine Learning Model Selection

Following the feature selection step, we evaluated a range of supervised machine learning models to identify the most suitable approach for predicting CDNC and estimating the CCN-CDNC relationship. The models included ENR (J. Friedman et al., 2010), RF (Breiman, 2001), XGBoost (T. Chen & Guestrin, 2016), Support Vector Regression (SVR) (Drucker et al., 1999), Gradient Boosting (J. H. Friedman, 2001), LightGBM (Ke et al., 2017), and a Neural Network using a Multi-Layer Perceptron (MLP) (Pinkus, 1999). These models were chosen due to their robust performance in regression tasks and their ability to handle complex data sets.

To compare the performance of these models, we used three evaluation metrics: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). These metrics provide complementary information about the accuracy and reliability of the models. MAE quantifies the average magnitude of prediction

Table 1
Overview of the Variables Used in the Machine Learning Model

Variable name	Abbreviation
Predictors	
Liquid water path	LWP
CCN burden	CCN
Updraft velocity at cloud base	W_b
Temperature at cloud top	T_{CT}
Temperature at cloud base	T_{CB}
10 m wind speed	U_{10}/V_{10}
Relative humidity at cloud base	RH_{CB}
Surface pressure	P_{surf}
Surface temperature	T_{surf}
Cloud base height	H_{CB}
Predictant	
Cloud droplet number concentration	CDNC

errors, RMSE gives higher weight to larger errors by squaring them, and R^2 describes how much variance in the target variable was explained by the fitted model.

The evaluation process involved splitting the data sets into training (two-thirds) and testing (one-third) subsets. All models were optimized with their best performing hyperparameters. Among all the models, ENR demonstrated the best performance, achieving the highest R^2 value and the lowest MAE and RMSE for both satellite and climate model data sets. RF and XGBoost showed relatively close performance to ENR in terms of these metrics. However, we selected ENR as the best model primarily because it directly provides a slope, which is crucial for interpreting the CCN-CDNC relationship. Neural networks (MLP), despite their capacity to model complex relationships, exhibited higher variability and required a larger number of hyperparameter tuning to achieve competitive performance. See Figure S1 in Supporting Information S1 for a detailed comparison of different machine learning models predicting CDNC.

The superior performance of ENR can be attributed to its regularization properties, which mitigate overfitting and enhance robustness, particularly when dealing with multicollinear predictors such as LWP, CCN, updraft velocity, and temperature. The effectiveness of ENR in handling complex and multivariate data sets has been demonstrated in prior studies such as Zou and Hastie (2005) and J. Friedman et al. (2010), supporting its application in this analysis. ENR focuses on relevant features and adjusts coefficient magnitudes to stabilize CDNC predictions and reduce variability. ENR aims to minimize the following loss function:

$$\text{minimize}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2+\alpha\rho\sum_{j=1}^{p}|\beta_j|+\frac{\alpha(1-\rho)}{2}\sum_{j=1}^{p}\beta_j^2\right), \tag{7}$$

where N represents the total number of observations in the data set, p denotes the number of features or predictor variables used in the regression model, y_i signifies the observed CDNC for the i-th observation, and $\hat{y}_i$ denotes the predicted CDNC for the i-th observation as estimated by the regression model. The coefficients β_j, corresponding to each feature j, directly influence $\hat{y}_i$ by determining the weight that each feature contributes to the prediction (Rosa, 2010). The regularization parameter α controls the strength of regularization applied to the model, while the mixing parameter ρ balances between L1 and L2 penalties.

The first term in the loss function represents the residual sum of squares, ensuring that the model fits the data well. The second and third terms are the regularization penalties, where the first term applies L1 regularization to encourage sparsity in the coefficient estimates (Lasso regression), and the second term applies L2 regularization to shrink the coefficients toward zero (Ridge regression) (Zou & Hastie, 2005). The Elastic Net algorithm minimizes this loss function by adjusting the coefficients (β_j) to find the optimal balance between fitting the data and preventing overfitting.

We implemented ENR using the scikit-learn Python library (Pedregosa et al., 2011), which includes hyperparameters α and $l1\_ratio$ (denoted as ρ in Equation 7). These hyperparameters play a crucial role in fine-tuning the regularization strength and determining the sparsity of the model. By adjusting these hyperparameters appropriately, we can modify the ENR model to our specific data set and optimize its performance in estimating the CCN-CDNC relationship. In this work, we selected the optimal hyperparameters by considering their effect on the RMSE between the observed CDNC and predicted CDNC. We conducted a grid search over a range of α and $l1\_ratio$ values, systematically evaluating their performance with respect to minimizing RMSE. By testing different combinations of α and $l1\_ratio$, we identified the values that yielded the lowest RMSE, indicating the best fit of the ENR model to the data. This was done separately for satellite observations and climate model data due to their distinct characteristics and complexities of each data set. Through this approach, we found that setting alpha to 0.05 for satellite data and 0.01 for climate model data, and $l1\_ratio$ to 0.1 for both data sets resulted the best fit for our ENR model as shown in Figures S2 and S3 in Supporting Information S1. Although slightly lower

*l*1_*ratio* values yielded marginally lower RMSE, values ≤0.1 are known to be less stable in Elastic Net implementations. Therefore, we selected 0.1 as a conservative and robust choice, consistent with best practices (Pedregosa et al., 2011).

2.6. Cloud Base Updrafts

For the satellite data, the cloud base updraft velocity (W_b) was calculated using the linear-fit method based on the correlation between cloud top radiative cooling (CTRC) and W_b as found by Zheng et al. (2016). We utilized the libRadtran, library for radiative transfer, version 2.0.5 (Emde et al., 2016) to simulate radiative transfer processes and thereby to estimate CTRC. The radiative transfer equation in libRadtran was solved for a 1-D plane-parallel atmosphere using the DISORT method (discrete ordinate radiative transfer solver). Absorptive cross sections for atmospheric species in the solar spectral range were obtained using the k-correlated method with Kato et al. (1999) parameterization, while Fu and Liou (1992) parameterization was used for the thermal or terrestrial longwave range. Absorptive and scattering properties of aerosols were estimated using the maritime mode of the parameterization by Shettle (1990).

We used the parameterization proposed by Zheng et al. (2016) to calculate W_b from CTRC:

$$W_b = -0.44 \times \mathrm{CTRC} + 22.30 \pm 13, \tag{8}$$

where W_b and CTRC have units of cm s^{-1} and W m^{-2}, respectively. This equation was found to perform well when compared to large eddy simulations of marine stratocumulus (Ahola et al., 2022). Using libRadtran, we simulated the radiative properties of the atmosphere and cloud layers for all the cloudy pixels from the satellite data, enabling us to estimate W_b. Each cloudy pixel was analyzed separately using local characteristics that include time, longitude and latitude, surface albedo, cloud cover and the vertical profile of cloud liquid water content. We followed the similar setup for the simulations as described by Zheng et al. (2019) to model CTRC.

The input variables for libRadtran were obtained from MODIS cloud products and ERA-5 reanalysis data. The inputs from MODIS retrieved cloud properties include CTT, CTH, CER, and LWP. Cloud geometric thickness (CGT) was derived from LWP, assuming adiabatic clouds (Brenguier et al., 2000). The cloud base height was then approximated as the difference between CTH and CGT. Vertical profile of cloud liquid water content was calculated for each model vertical layer following the method described by Wood et al. (2009). Subsequently, we calculated CTRC from the vertical profiles of shortwave and longwave radiative fluxes. A detailed theoretical explanation on CTRC and its calculation is provided in Zheng et al. (2021).

This setup, integrating the capabilities of libRadtran RTM with satellite data, enabled us to estimate the W_b from CTRC. Thus, we bridge the gap between climate model simulations and observational data, facilitating a comprehensive comparison of the responses of CDNC to updraft velocities between both data sets.

2.7. Cloud Parcel Model

We used a cloud parcel model to demonstrate the theoretical variability in CDNC due to CCN and varying updraft velocities within an aerosol particle population. A detailed description of the model framework is given in Romakkaniemi et al. (2009). The model simulates the adiabatic ascent, expansion, and cooling of a homogeneous air parcel from a sub-saturated state to a supersaturated cloudy environment at a uniform vertical velocity. A sectional method is employed in the model to define the dry aerosol size distribution, with log-normally spaced size bins evolving freely during condensation and evaporation. The initial conditions for the model simulations are derived from the Northern Pacific Ocean region to align with the satellite observations. The model is initialized with the below cloud size distribution and meteorological parameters to simulate cloud droplet formation and their subsequent condensational growth within the air parcel. Specifically, we used the trimodal aerosol distribution described by J. Park et al. (2020), which includes Aitken mode, accumulation mode, and coarse mode particles. This distribution is characterized in the model by the total number concentration, geometric mean diameter, and geometric standard deviation of the log-normal distribution, as reported by J. Park et al. (2020). In our simulations, we used the Aitken mode with a total number concentration of 150 cm^{-3}, a geometric mean diameter of 80 nm, and a geometric standard deviation of 1.4. The accumulation mode has a total number concentration of 110 cm^{-3}, a geometric mean diameter of 160 nm, and a geometric standard deviation of 1.4. The coarse mode has a total number concentration of 15 cm^{-3}, a geometric mean diameter of 300 nm, and a

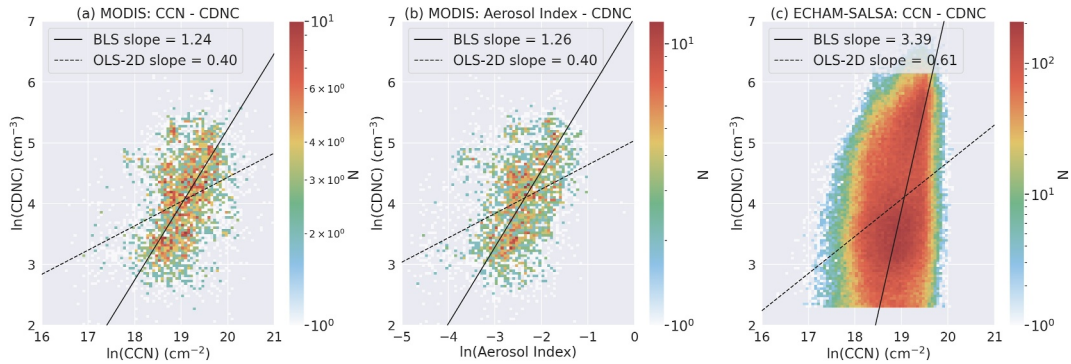

Figure 1. Susceptibility of cloud-top CDNC to changes in (a) CCN from satellite observations, (b) aerosol Index from satellite observations and (c) CCN from climate model simulation, using BLS and OLS regression methods, with colors representing the number of data points.

geometric standard deviation of 1.4. Additionally, the initial meteorological parameters include a temperature of 298 K, a pressure of 9,000 hPa, and a relative humidity of 95%. The vertical velocity is varied between 0.1 and 0.6 m s^{-1} to capture a range of dynamic scenarios. To explore how CDNC evolves with changes in aerosol concentrations under varying updraft conditions, we adjusted the total number concentrations in the Aitken and accumulation modes around the baseline values provided by J. Park et al. (2020). By incrementally increasing and decreasing these concentrations, we examined how different levels of aerosol particle populations influence CDNC when subjected to different updraft velocities. The initial meteorological parameters included in the model are temperature, pressure, relative humidity and vertical velocity. The simulations continue until the air parcel reaches the user-defined cloud base height, set at 1,000 m for this study. We considered particles with diameters larger than 100 nm as CCN and 2 μm as cloud droplets.

3. Results

3.1. Comparison Between OLS and BLS Fit for CCN and CDNC

We examined the results of OLS and BLS regression methods for assessing the relationship between CCN and CDNC in both satellite observations and climate model data. To better understand this relationship, we focused our analysis on the Northern and Southern Pacific Oceans. The results from the Northern Pacific region are depicted in Figure 1, while Figure S4 in Supporting Information S1 presents the corresponding analysis for the Southern Pacific region. In the climate model data, CCN refers to the column burden of CCN, and CDNC represents the CDNC at cloud top, ensuring alignment with satellite retrievals. In addition to the CCN-CDNC analysis from satellite data, we also considered aerosol index, a commonly used CCN proxy to support our findings (Hasekamp et al., 2019).

From Figures 1a and 1c, it is evident that the OLS regression does not visually fit well with the data points depicting the relationship between CCN and CDNC whether in satellite observations or climate model simulations. This finding corroborates previous research by Pitkänen et al. (2016), which highlights the limitations of OLS regression in accurately capturing the relationship between these variables. A key issue with OLS is its assumption that the independent variable (CCN) is measured without error. In reality, both CCN and CDNC contain observational and model uncertainties, and failing to account for these can lead to biased slope estimates (Mikkonen et al., 2019). In contrast, BLS regression produces a visually better fit to the data compared to OLS, with the regression line more closely aligning with the spread of CDNC and CCN values. However, despite this improved visual agreement, the slope estimated using BLS consistently exceeds 1 for both satellite retrievals and climate model data. This result is further supported by the aerosol index-CDNC relationships from satellite data, as shown in Figure 1b. As cloud droplets can only form around CCN, CDNC should never be able to exceed CCN concentrations. In the case that if CCN is the only factor which influences CDNC concentrations, a slope larger than 1 in the CDNC-CCN correlation is unphysical. This discrepancy indicates that CDNC is significantly influenced by additional confounding factors beyond CCN, such as meteorological conditions, cloud microphysical processes, and retrieval uncertainties.

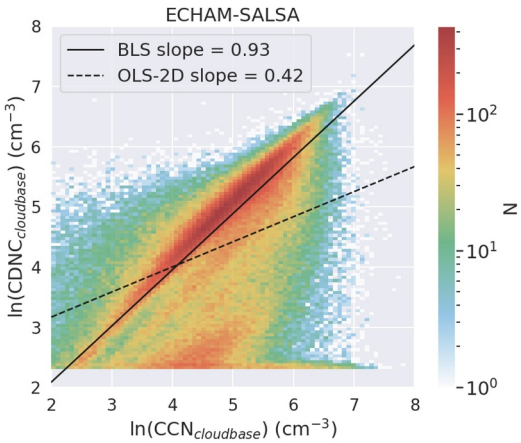

Figure 2. Susceptibility of cloud base CDNC to changes in cloud base CCN using BLS and OLS from climate model data, with colors representing the number of data points.

To better understand the limitations of OLS and BLS, we performed a supplementary analysis using synthetic data with a prescribed slope and controlled levels of variability. As detailed in Text 1 in Supporting Information S1, this analysis revealed that OLS systematically fails to recover the prescribed slope due to its inability to account for multivariate dependencies and noise interactions. Similarly, BLS tends to overestimate the slope when there is significant variability in dependent variables or when additional influencing factors are present. These findings underscore the limitations of accurately capturing the CCN-CDNC relationship in the presence of realistic noise and complexity and highlights the limitations of both OLS and BLS under such conditions.

We also examined the relationship between CCN at cloud base and CDNC at cloud base, where the influence of CCN on cloud droplet formation is most significant. This analysis, based on climate model data shown in Figure 2, indeed exhibits a slope less than 1, indicating a realistic relationship between the two variables. This result points to the possibility that the column burden of CCN averages out the real variability in the cloud base CCN, resulting in an artificially inflated slope between columnar CCN and CDNC. On the other hand, when performing BLS regression between the column burden of CCN and cloud-top CDNC, the estimated slope may be influenced by additional physical factors besides CCN, such as updraft velocity, cloud dynamics, aerosol composition, and other environmental factors. These factors can affect the slope obtained through regression considering only CCN and CDNC, which implicitly assumes a much simpler relationship between the two. This discrepancy underscores the complexity of aerosol-cloud interactions and highlights the need to elucidate the specific mechanisms driving the observed relationship between CCN and CDNC.

3.2. Feature Selection

To identify the underlying factors responsible for the higher-than-one slope in the BLS fitting between CCN burden and cloud-top CDNC, we utilized RF model for feature selection. We began by assessing the ability of RF model in predicting CDNC using feature variables derived from both satellite observations and climate model data. Figures 3a and 3b show that the RF model can successfully predict CDNC, achieving an R^2 value of 0.82 for satellite data and 0.62 for climate model data. These R^2 values align with relative root mean square errors of 0.31 for satellite data and 0.65 for climate model data, providing confidence in the RF model's predictive capability. Based on these results, we then performed a feature importance analysis using the same model setup to investigate the influence of each parameter on CDNC variability. The resulting feature importance scores reflect the relative contribution of each parameter to CDNC variability, with higher scores indicating greater influence. This analysis was performed only for the Northern Pacific Ocean, selected as a representative study region. It is important to note that the feature importance analysis can be biased when features are strongly correlated. In such cases, RF may prioritize one correlated feature over others, leading to an underestimation of the true importance of the redundant features. However, in our analysis, the features are only weakly correlated, which minimizes the potential bias in feature importances.

In order to ensure consistency of influential factors driving CDNC variability, we conducted the feature importance analysis separately for both satellite observations and climate model data. The analysis involved a range of atmospheric and meteorological parameters as given in Table 1.

Figures 3c and 3d display the normalized feature importance scores of various parameters from satellite observation and climate model data, respectively. Our analysis reveals a remarkable similarity in the importance of features across both data sets, indicating a convergence in the most important factors influencing CDNC variability. Notably, the three most important features remain consistent across both data sets: LWP, CCN burden, and W_b.

To further validate these findings, we employed SHAP analysis and permutation feature importance, providing an additional layer of interpretability to our results. As shown in Figures S5 and S6 in Supporting Information S1, both methods confirm that CCN, W_b, and LWP are the primary drivers of CDNC variability. This consistency strengthens confidence in the robustness of our conclusions across different machine learning models. This

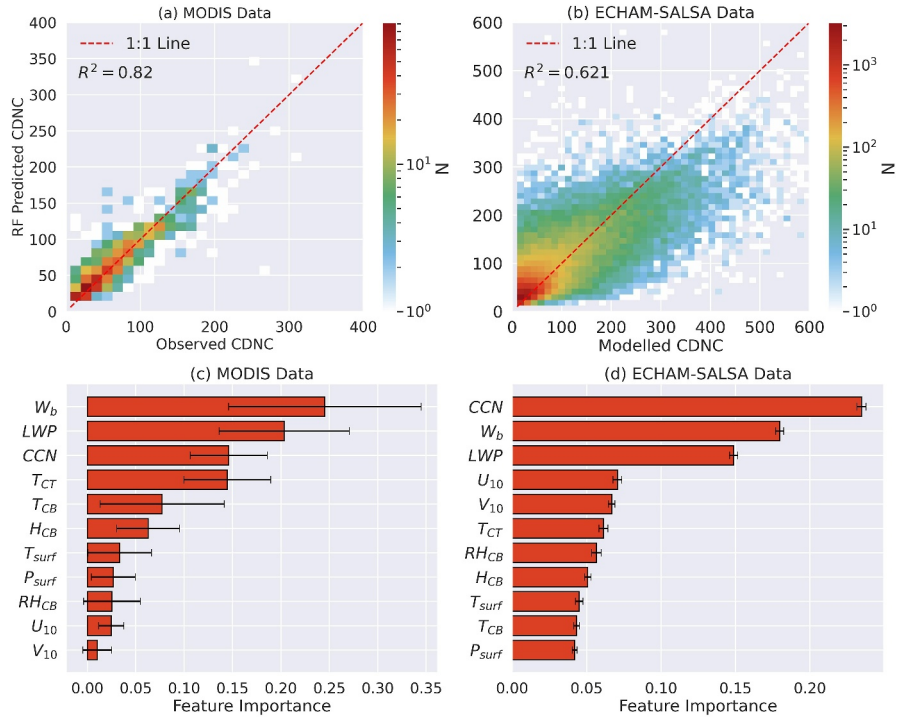

Figure 3. (a) Satellite derived CDNC against RF predicted CDNC, (b) modeled CDNC against RF predicted CDNC, with colors representing the number of data points. (c) Normalized feature importance scores for satellite-based observations. (d) Normalized feature importance scores for climate model data. Error bars (c) and (d) represent the standard deviation of the feature importance distribution across all trees in the forest.

alignment of feature importance across observed and modeled data sets underscores the critical role these factors play in modulating CDNC. W_b, in particular, is crucial in aerosol activation processes, directly influencing CDNC and indirectly affecting factors like LWP. The implications of these factors will be discussed in more detail in the subsequent sections.

3.3. Impact of Updraft Velocity on CDNC Variability

In this section, we investigated the specific dynamics and influence of W_b on CDNC in both satellite observations and climate model data. To complement these analyses, we performed cloud parcel model simulations to simulate droplet activation under varying updrafts and CCN concentrations, which helped us understand the CDNC distribution in these scenarios. Figure 4 shows the CDNC as a function of CCN concentration for both MODIS retrievals (left) and model data (right). The data points are grouped into equally-sized bins and each bin is colored according to the mean updrafts averaged over all data points falling into the bin. For the MODIS retrievals (left panel), the figure also shows the results of our cloud parcel model simulations. In these cloud parcel model simulations, CCN concentrations are integrated over the cloud base height to calculate CCN burden, assuming the boundary layer is well-mixed, in order to allow a direct comparison between the simulated and observed data.

Our analysis shows a positive correlation between W_b and CDNC in both satellite observations and climate model data. However, there are notable differences between the updrafts derived from satellite data and those simulated by the climate model ECHAM-SALSA, with the model generally simulating higher updrafts. Despite this discrepancy, the general pattern of W_b remains consistent between satellite observations and model simulations. From Figure 4, we can also see that the distribution of updrafts between panels (a) and (b) vary significantly. Simulated updrafts have much higher variability and this also results in high variability in CDNC. ECHAM-SALSA appears to overestimate updrafts compared to satellite-derived values. Virtanen et al. (2025) have previously shown that the distribution of updrafts in ECHAM-SALSA are much wider than those observed at Puijo, Pallas, and Zeppelin measurement stations. They also show that the mean of the updrafts is overestimated in ECHAM-SALSA by more than a factor of two. This overestimation of updrafts in the model affects the simulated

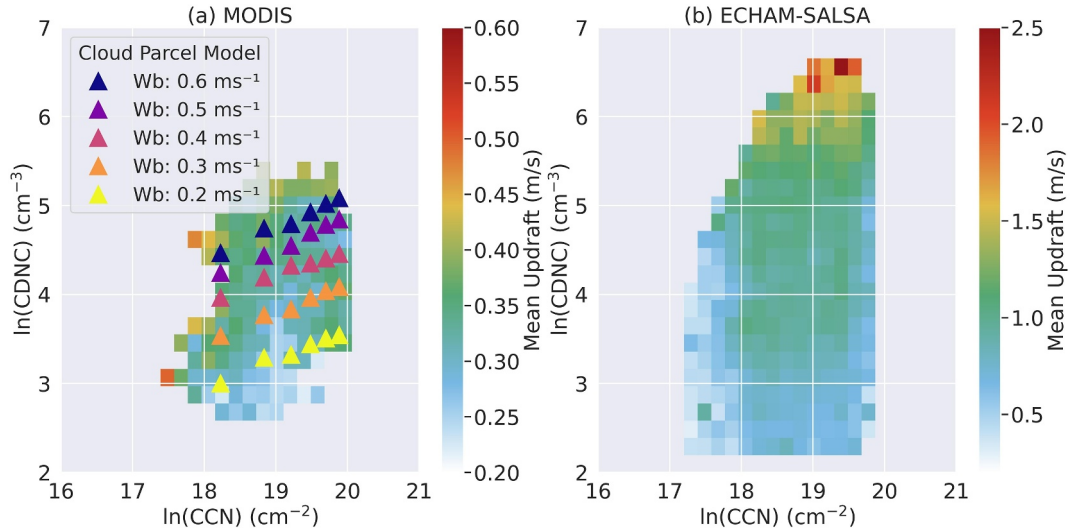

Figure 4. Variability of mean updraft velocity with cloud-top CDNC as a function of CCN burden in (a) satellite and (b) climate model data. Note different scales for updraft velocity. Markings in (a) are CDNC simulated by cloud parcel model for different updraft velocities.

CDNC, as updraft strength directly influences cloud droplet activation. It is known that CDNC increases with stronger updraft velocities, indicating that higher updrafts correlate with elevated CDNC levels. Stronger updrafts lead to higher supersaturation of the rising air parcels, activating more CCN and increasing CDNC even at the same CCN levels (Malavelle et al., 2014; Seinfeld & Pandis, 2016). This highlights the importance of accounting for variability in W_b when analyzing the CCN-CDNC relationship.

The cloud parcel model simulations, which specifically varied W_b and CCN concentration, further supported these findings by illustrating the detailed dynamics of CDNC under different conditions as depicted in Figure 3a. In these simulations, we varied W_b from 0.2 to 0.6 m s^{-2}, consistent with values derived from satellite observations, while keeping CCN concentrations constant. Additionally, we varied CCN concentrations while maintaining a constant W_b. The simulations demonstrated that with increasing W_b at fixed CCN, the maximum supersaturation within the cloud parcel increased, leading to the activation of more CCN and a corresponding rise in CDNC.

These simulations provided a clear understanding of how updrafts and CCN interact to influence CDNC, reinforcing the positive correlation observed in both satellite and climate model data. However, it is essential to recognize that the data are inflated in the y-direction due to the variability in updrafts, as seen in Figure 4a, which would probably lead to an upwards tilt of the regression line if only BLS fitting was used. BLS considers only two variables at a time, in this case CCN and CDNC, and fails to account for the additional variability introduced by W_b. Since updrafts play a significant role in cloud droplet formation, ignoring this variable oversimplifies the complex processes involved. This can lead to an overestimation of the direct correlation between CCN and CDNC.

In addition to the influence of updrafts on CDNC, the interaction between W_b and CDNC can have further implications on cloud dynamics and atmospheric processes (Hsieh et al., 2009; Sanchez et al., 2016; West et al., 2014). One significant effect is on the mixing of different cloud types within the atmosphere. Variability in updrafts can influence the vertical distribution and mixing of clouds with different characteristics. In regions with stronger updrafts, where CDNC exhibit a more pronounced sensitivity to changes in updraft strength, the vertical mixing of clouds can be faster (Stull, 2012). Strong updrafts can entrain air from surrounding regions, leading to vigorous vertical mixing within clouds and redistribution of cloud droplets. This process can result in the mixing of different cloud types, such as stratiform and convective clouds, and can contribute to the development of complex cloud structures. On the other hand, in regions with weaker updrafts, the vertical mixing of clouds may be less vigorous. This can lead to more distinct cloud layers with less mixing between different cloud types.

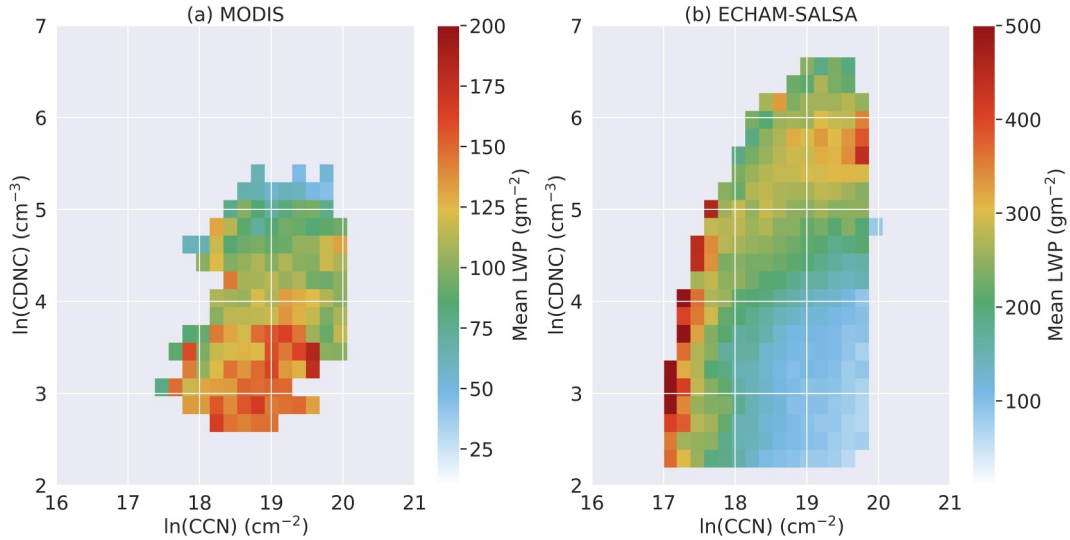

Figure 5. Variability of mean LWP with cloud-top CDNC as a function of CCN burden in (a) satellite and (b) climate model data. Note different scales for LWP.

Weaker updrafts may also limit the vertical extent of clouds, resulting in shallower cloud layers and reduced vertical mixing.

3.4. Impact of Cloud Mixing on LWP

The mixing of clouds in different meteorological conditions and their phase within the cloud structure can significantly influence the variability in LWP. LWP is a key parameter in understanding cloud properties and microphysics. We analyzed how LWP responds to changes in CDNC in both satellite and climate model data. Interestingly, we observed distinct patterns in the response of LWP to CDNC as a function of CCN levels. While W_b exhibited similar patterns in both data sets, there were significant differences in the LWP patterns between satellite observations and climate model simulations. Figure 5 shows the results of this analysis. The figure was generated in the same way as Figure 4, but here the color scale visualizes the mean LWP of the data points in each bin.

In the satellite data (Figure 5a), we noted a tendency of LWP to decrease as CDNC increases, indicating a negative correlation between these two variables. Specifically, for lower CDNC values, LWP tends to be higher, suggesting that clouds with fewer droplets have higher liquid water content. As CDNC increases, LWP decreases, indicating a reduction in liquid water content with the cloud. A recent study by Arola et al. (2022) attributed this decrease in LWP at higher CDNC ranges to satellite retrieval errors, particularly in the variability in CER. The results obtained for the satellite retrievals may therefore not be representative of actual behavior in the atmosphere (Arola et al., 2022; Kokkola et al., 2025).

In contrast, the response of LWP to CDNC in the climate model shows a more complex pattern. At very low CCN and CDNC concentrations, LWP tends to be higher, indicating a higher liquid water content for clouds formed under conditions of low aerosol abundance and fewer cloud droplets. However, as these clouds do precipitate and the pattern aligns well with the modeled large-scale precipitation patterns (Figure S8b in Supporting Information S1), it is likely that the wet scavenging efficiently reduces the aerosol concentration and thus affects the observed LWP-CCN-CDNC relationship. On the other hand, we can see the increase in LWP as a function of CDNC for most of the data. Comparing this to the mean updraft in Figure 4, we can see that the increase in CDNC is correlated with higher updrafts. These updrafts are high enough to allow particles smaller than 100 nm to be activated as cloud droplets. Another interesting region of the data is around the CDNC values of 300 per cc, where LWP increases with increasing CCN concentration. This could be due to a correlation between large-scale meteorology and CCN advection from source areas.

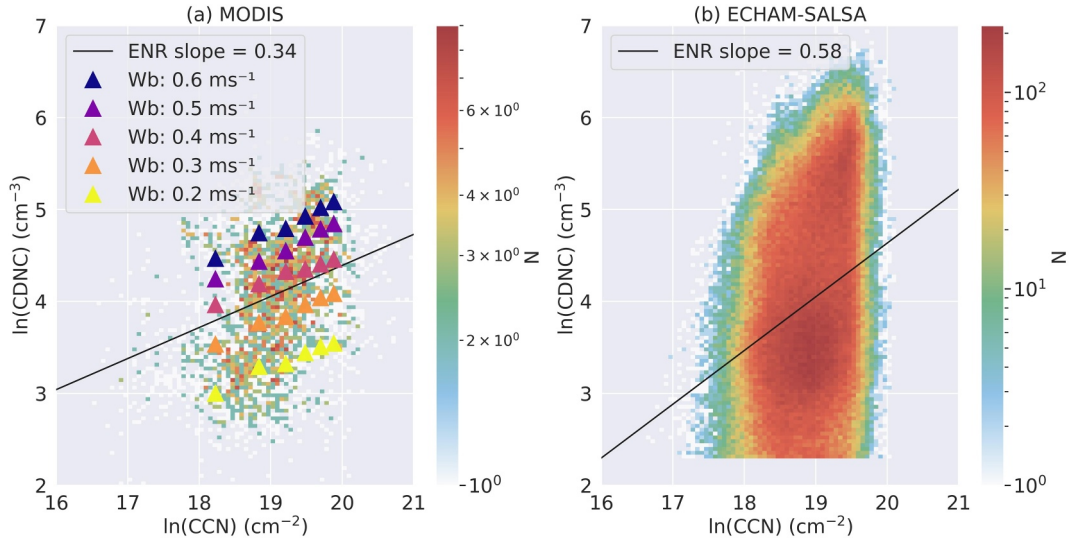

Figure 6. Susceptibility of cloud-top CDNC to changes in CCN burden from (a) satellite observations and (b) climate model data using Elastic Net Regression. The data points represent actual CDNC and CCN values from MODIS and the model, respectively, while the regression lines show the ENR-derived relationships.

3.5. Slope Estimation Using Elastic Net Regression

Markings in **a** are CDNC simulated by cloud parcel model for different updraft velocities.

As we observe that there are multiple factors influencing CDNC beyond CCN alone, and bivariate regression results in an unphysical higher-than-one slope estimate between CCN and CDNC, it is imperative to explore alternative approaches capable of accurately capturing the role of aerosols in cloud formation. To address this challenge, we evaluate the efficacy of using ENR, a novel machine learning-based regularization approach, to extract the CCN-CDNC correlation. ENR is widely used to model complex relationships while reducing issues such as multicollinearity and overfitting. ENR allows incorporation of a set of feature variables relevant for the prediction of the target variable, including various atmospheric and environmental factors influencing cloud formation and behavior. We used the same set of features for ENR as those used for feature selection using RF (see Table 1). ENR can account for both linear and non-linear dependencies by fitting an improved version of the linear regression model by imposing L1 and L2 regression penalties as given in Section 2.4.

Our results, as illustrated in Figure 6, show that the slope estimates obtained through ENR are much lower values than BLS fitting. Notably, the ENR slope estimates for both satellite observations and climate model data are less than 1, contrasting with the steeper, unrealistic slopes produced by BLS. This difference in slope estimates is particularly significant given the physical constraints of aerosol-cloud interactions.

The ENR derived slope closely aligns with the cloud parcel model results, which is based on well-established physical principles of droplet activation and supersaturation. The cloud parcel model was run with observed initial conditions to mimic real-world scenarios, providing an accurate theoretical representation of how CCN concentrations influence CDNC under varied updraft conditions. This alignment suggests that ENR accurately captures a physically plausible CCN-CDNC slope under the observed conditions, effectively mitigating the impact of noise and other confounding factors in the observational data. As a result, ENR produces a slope that aligns more closely with the theoretical expectations. This consistency between ENR estimates and the cloud parcel model underscores ENR's ability to provide a more reliable estimate of cloud droplet susceptibility to aerosol perturbations. ENR achieves this by incorporating regularization techniques and feature selection, providing more accurate slope estimates and reinforcing confidence in its effectiveness for capturing the complex CCN-CDNC relationship. The coefficients of different features derived from the ENR are provided in Tables S1 and S2 in Supporting Information S1 for further reference.

Furthermore, in our study, the slope estimates derived from ENR closely match those obtained through OLS regression. However, it is important to exercise caution when using OLS regression, as it may produce good

results for the wrong reasons, particularly in more complex or multicollinear data sets where its assumptions may not hold. Additionally, the supplementary analysis, which compares various regression methods using synthetic CCN and CDNC data, further adds confidence to the efficacy of ENR. This analysis demonstrates how ENR outperforms other approaches by accounting for the multivariate relationships in the data and providing more physically consistent slope estimates. This comparison is detailed in Text S1 in Supporting Information S1, which offers additional evidence supporting the robustness of ENR in capturing CCN-CDNC relationship under diverse conditions.

4. Conclusions

In conclusion, our study employed a comprehensive approach to investigate the intricate relationships between CCN, CDNC, and associated atmospheric dynamics using a combination of regression methodologies. We utilized both OLS and BLS fitting methods to assess the CDNC's susceptibility to aerosol perturbations. While OLS regression consistently struggled to fit the data points in both satellite observations and climate model simulations, BLS yielded unphysical relationships. Notably, the slope estimates derived from BLS consistently exceeded 1, contrary to what is expected in reality.

To address these limitations, we leveraged advanced machine learning techniques to identify key predictors influencing CDNC variability. Our analysis highlighted the critical role of updrafts in shaping the CCN-CDNC relationship. Specifically, stronger updrafts were associated with higher CDNC levels for a given CCN. However, when clouds with varying updraft strengths are included in the analysis, statistical methods like BLS can lead to an artificially steeper slope. This is because the differences in updraft strengths result in distinct cloud dynamics, and combining these variations in the analysis can exaggerate the slope between CCN and CDNC. Furthermore, this pattern in updraft and cloud types significantly influenced the variability in LWP. In order to account for these complexities, we recommend the use of a machine-learning based regularization method such as ENR, to accurately estimate the susceptibility of cloud droplets to aerosol perturbations. The use of ENR to analyze such data results in more accurate estimates of the susceptibility of CDNC to changes in CCN concentrations. By incorporating key predictors affecting CDNC variability, ENR provides a robust framework for analyzing complex data sets with diverse cloud types and atmospheric conditions. The use of ENR results in more accurate estimates of the susceptibility of CDNC to changes in CCN concentrations, addressing the limitations of traditional regression methods. Therefore, we suggest employing ENR as a preferred method for estimating the susceptibility of CDNC to changes in CCN concentration in studies involving non-homogeneous cloud populations. To further validate the physical realism of ENR-derived slope estimates, we compared them with the results from a cloud parcel model simulations and tested their robustness using a synthetic data set with a prescribed CCN–CDNC relationship. The parcel model provides a first-order theoretical benchmark based on well-understood activation physics. We found that the ENR slopes aligned closely with parcel model results, reinforcing that ENR is capturing the dominant microphysical processes in a physically meaningful way. Additionally, in synthetic experiments where the true slope was known, ENR produced estimates closer to the prescribed value despite the presence of noise and multivariate influences, whereas OLS and BLS showed systematic biases. These comprehensive evaluations of theoretical models, synthetic simulations, and observation strengthen our confidence that ENR provides not only statistically robust but also physically grounded estimates of CDNC susceptibility to aerosol perturbations.

It is also important to note that the stronger CCN effect on CDNC in the model compared to satellite observations is largely due to differences in updraft strength, which plays a crucial role in CDNC. As shown in Figure 4, the climate model systematically overestimates updraft velocity compared to satellite-derived values. Since updrafts directly influence supersaturation and subsequent CCN activation, this bias can lead to an enhanced CDNC in the model. In the model, where updrafts are stronger, it is more likely that activation is in the aerosol limited regime, which results in a stronger sensitivity of CDNC to CCN than in the case of lower updrafts.

It should also be noted that although the sensitivity estimates from OLS and ENR for MODIS data and climate model simulations appear to be numerically closer (e.g., 0.34 vs. 0.40 for MODIS and 0.58 vs. 0.61 for simulation data), the true advantage of ENR lies in its ability to handle the complexities of real-world data. ENR, unlike OLS, is effective at reducing the undue effects of noise and multicollinearity through regularization, thereby limiting the influence of less relevant predictors. This makes ENR particularly robust when extending analyses to observational data sets where retrieval uncertainties, spatial averaging, and other sources of variability are

present. Moreover, the limited variability in the x-direction, which is CCN in our case, partly accounts for why the slope estimates from ENR and OLS do not differ dramatically. With a fairly narrow range of CCN values, both methods may yield similar estimates, which makes the advantage of ENR's regularization less apparent when looking at the slope. These qualities further support the use of ENR as a more reliable and robust approach for estimating the susceptibility of CDNC to changes in CCN concentration, particularly in studies involving non-homogeneous cloud populations.

By integrating machine learning techniques with traditional statistical methods, our study advances the understanding of aerosol-cloud interactions and their implications for cloud microphysics. This approach provides valuable insights into how changes in aerosol concentrations and atmospheric dynamics modulate cloud properties, ultimately contributing to improved modeling of cloud behavior and climate impacts.

Data Availability Statement

All MODIS data used in this study are open data and were obtained from the NASA Level-2 and Atmosphere Archive & Distribution System Distributed Active Archive Center (LAADS DAAC) https://ladsweb.modaps.eosdis.nasa.gov/. Aerosol-climate model data and codes used to produce figures are openly available (Irfan et al., 2024). Libradtran radiative transfer model is freely available for public usage at https://www.libradtran.org/doku.php?id=start (Mayer et al., 2017). ERA5 data were obtained from Copernicus Climate Change Service (C3S) Climate Data Store accessible at https://cds.climate.copernicus.eu/ (Hersbach et al., 2020). Cloud parcel model is openly available at https://zenodo.org/records/13819445 (Kokkola, 2024). All input files for ECHAM-SALSA are standard and are available from the HAMMOZ repository (HAMMOZ consortium, 2020). The ECHAM6-HAMMOZ model is made available to the scientific community under the HAMMOZ Software Licence Agreement, which defines the conditions under which the model can be used. The licence can be accessed from https://redmine.hammoz.ethz.ch/projects/hammoz/wiki/1_Licencing_conditions (HAMMOZ consortium, 2012).

Acknowledgments

This work was supported by the Horizon Europe programme under Grant Agreement No 101137680 via project CERTAINTY (Cloud-aERosol inTeractions and their impActs IN The earth sYstem); the University of Eastern Finland Doctoral Program in Environmental Physics, Health and Biology; the Research Council of Finland (project numbers 339885, 337550, 357904); the European Research Council (ERC) project "PyroTRACH" (Grant agreement No. 726165), and the European Union's Horizon Europe project "CleanCloud" (Grant agreement No. 101137639). We acknowledge CSC–IT Center for Science, Finland for computational resources. The ECHAM-HAMMOZ model is developed by a consortium composed of ETH Zurich, Max Planck Institut für Meteorologie, Forschungszentrum Jülich, University of Oxford, the Finnish Meteorological Institute and the Leibniz Institute for Tropospheric Research, and managed by the Leibniz Institute for Tropospheric Research (TROPOS). Open access publishing facilitated by Ita-Suomen yliopisto, as part of the Wiley - FinELib agreement.

References

Abdul-Razzak, H., & Ghan, S. J. (2002). A parameterization of aerosol activation 3. Sectional representation. *Journal of Geophysical Research*, *107*(D3), AAC-1. https://doi.org/10.1029/2001jd000483

Ahola, J., Raatikainen, T., Alper, M. E., Keskinen, J.-P., Kokkola, H., Kukkurainen, A., et al. (2022). Technical note: Parameterising cloud base updraft velocity of marine stratocumuli. *Atmospheric Chemistry and Physics*, *22*(7), 4523–4537. https://doi.org/10.5194/acp-22-4523-2022

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Arola, A., Lipponen, A., Kolmonen, P., Virtanen, T. H., Bellouin, N., Grosvenor, D. P., et al. (2022). Aerosol effects on clouds are concealed by natural cloud heterogeneity and satellite retrieval errors. *Nature Communications*, *13*(1), 7357. https://doi.org/10.1038/s41467-022-34948-5

Bai, H., Gong, C., Wang, M., Zhang, Z., & L'Ecuyer, T. (2018). Estimating precipitation susceptibility in warm marine clouds using multi-sensor aerosol and cloud products from a-train satellites. *Atmospheric Chemistry and Physics*, *18*(3), 1763–1783. https://doi.org/10.5194/acp-18-1763-2018

Bellomo, K., Clement, A. C., Norris, J. R., & Soden, B. J. (2014). Observational and model estimates of cloud amount feedback over the Indian and Pacific oceans. *Journal of Climate*, *27*(2), 925–940. https://doi.org/10.1175/jcli-d-13-00165.1

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., et al. (2019). Bounding global aerosol radiative forcing of climate change. *Reviews of Geophysics*, *58*(1), 1234–1265. https://doi.org/10.1029/2019RG000660

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

Brenguier, J.-L., Pawlowska, H., Schüller, L., Preusker, R., Fischer, J., & Fouquart, Y. (2000). Radiative properties of boundary layer clouds: Droplet effective radius versus number concentration. *Journal of the Atmospheric Sciences*, *57*(6), 803–821. https://doi.org/10.1175/1520-0469(2000)057⟨0803:RPOBLC⟩2.0.CO;2

Calderón, S. M., Tonttila, J., Buchholz, A., Joutsensaari, J., Komppula, M., Leskinen, A., et al. (2022). Aerosol–stratocumulus interactions: Towards a better process understanding using closures between observations and large eddy simulations. *Atmospheric Chemistry and Physics*, *22*(18), 12417–12441. https://doi.org/10.5194/acp-22-12417-2022

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, Y., Haywood, 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

Dismuke, C., & Lindrooth, R. (2006). Ordinary least squares. *Methods and Designs for Outcomes Research*, *93*(1), 93–104.

Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, *10*(5), 1048–1054. https://doi.org/10.1109/72.788645

Emde, C., Buras-Schnell, R., Kylling, A., Mayer, B., Gasteiger, J., Hamann, U., et al. (2016). The libradtran software package for radiative transfer calculations (version 2.0.1). *Geoscientific Model Development*, *9*(5), 1647–1672. https://doi.org/10.5194/gmd-9-1647-2016

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Frost, C., & Thompson, S. G. (2000). Correcting for regression dilution bias: Comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society - Series A: Statistics in Society*, *163*(2), 173–189. https://doi.org/10.1111/1467-985x.00164

Fu, Q., & Liou, K. N. (1992). On the correlated k-distribution method for radiative transfer in nonhomogeneous atmospheres. *Journal of the Atmospheric Sciences*, *49*(22), 2139–2156. https://doi.org/10.1175/1520-0469(1992)049⟨2139:OTCDMF⟩2.0.CO;2

Gao, S., Lu, C., Liu, Y., Yum, S. S., Zhu, J., Zhu, L., et al. (2021). Comprehensive quantification of height dependence of entrainment mixing between stratiform cloud top and environment. *Atmospheric Chemistry and Physics*, *21*(14), 11225–11241. https://doi.org/10.5194/acp-21-11225-2021

Gao, S., Lu, C., Zhu, J., Li, Y., Liu, Y., Zhao, B., et al. (2024). Using machine learning to predict cloud turbulent entrainment-mixing processes. *Journal of Advances in Modeling Earth Systems*, *16*(8), e2024MS004225. https://doi.org/10.1029/2024MS004225

Gassó, S., & Hegg, D. A. (2003). On the retrieval of columnar aerosol mass and CCN concentration by MODIS. *Journal of Geophysical Research*, *108*(D1), AAC-6. https://doi.org/10.1029/2002jd002382

Grosvenor, D. P., Sourdeval, O., Zuidema, P., Ackerman, A., Alexandrov, M. D., Bennartz, R., et al. (2018). Remote sensing of droplet number concentration in warm clouds: A review of the current state of knowledge and perspectives. *Reviews of Geophysics*, *56*(2), 409–453. https://doi.org/10.1029/2017RG000593

Gryspeerdt, E., Goren, T., Sourdeval, O., Quaas, J., Mülmenstädt, J., Dipu, S., et al. (2019). Constraining the aerosol influence on cloud liquid water path. *Atmospheric Chemistry and Physics*, *19*(8), 5331–5347. https://doi.org/10.5194/acp-19-5331-2019

Gryspeerdt, E., McCoy, D. T., Crosbie, E., Moore, R. H., Nott, G. J., Painemal, D., et al. (2022). The impact of sampling strategy on the cloud droplet number concentration estimated from satellite data. *Atmospheric Measurement Techniques*, *15*(12), 3875–3892. https://doi.org/10.5194/amt-15-3875-2022

Gryspeerdt, E., Povey, A. C., Grainger, R. G., Hasekamp, O., Hsu, N. C., Mulcahy, J. P., et al. (2023). Uncertainty in aerosol–cloud radiative forcing is driven by clean conditions. *Atmospheric Chemistry and Physics*, *23*(7), 4115–4122. https://doi.org/10.5194/acp-23-4115-2023

Gryspeerdt, E., Quaas, J., Ferrachat, S., Gettelman, A., Ghan, S., Lohmann, U., et al. (2017). Constraining the instantaneous aerosol influence on cloud albedo. *Proceedings of the National Academy of Sciences*, *114*(19), 4899–4904. https://doi.org/10.1073/pnas.1617765114

HAMMOZ consortium. (2012). HAMMOZ software licence agreement. Retrieved from https://redmine.hammoz.ethz.ch/attachments/291/License_ECHAM-HAMMOZ_June2012.pdf

HAMMOZ consortium. (2020). Input files for ECHAM-SALSA. (Standard input files available from the HAMMOZ repository) Retrieved from https://redmine.hammoz.ethz.ch/projects/hammoz/repository/1/show/echam6-hammoz/branches/fmi/fmi_trunk

Hasekamp, O. P., Gryspeerdt, E., & Quaas, J. (2019). Analysis of polarimetric satellite measurements suggests stronger cooling due to aerosol-cloud interactions. *Nature Communications*, *10*(1), 5405. https://doi.org/10.1038/s41467-019-13372-2

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1994–2049. https://doi.org/10.1002/qj.3803

Hsieh, W.-C., Nenes, A., Flagan, R. C., Seinfeld, J. H., Buzorius, G., & Jonsson, H. (2009). Parameterization of cloud droplet size distributions: Comparison with parcel models and observations. *Journal of Geophysical Research*, *114*(D11), D11205. https://doi.org/10.1029/2008jd011387

Irfan, M., Kokkola, H., & Holopainen, E. (2024). Dataset for the study "Using machine learning approach for isolating aerosol effects on cloud droplet number concentration in marine stratocumulus clouds" by Irfan et al [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.13787708

Jia, H., & Quaas, J. (2023). Nonlinearity of the cloud response postpones climate penalty of mitigating air pollution in polluted regions. *Nature Climate Change*, *13*(9), 943–950. https://doi.org/10.1038/s41558-023-01775-5

Jia, H., Quaas, J., Gryspeerdt, E., Böhm, C., & Sourdeval, O. (2022). Addressing the difficulties in quantifying droplet number response to aerosol from satellite observations. *Atmospheric Chemistry and Physics*, *22*(11), 7353–7372. https://doi.org/10.5194/acp-22-7353-2022

Kalbande, R., Kumar, B., Maji, S., Yadav, R., Atey, K., Rathore, D. S., & Beig, G. (2023). Machine learning based quantification of VOC contribution in surface ozone prediction. *Chemosphere*, *326*, 138474. https://doi.org/10.1016/j.chemosphere.2023.138474

Kato, S., Ackerman, T. P., Mather, J. H., & Clothiaux, E. E. (1999). The k-distribution method and correlated-k approximation for a shortwave radiative transfer model. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *62*(1), 109–121. https://doi.org/10.1016/S0022-4073(98)00075-2

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.

Khatri, P., Yoshida, K., & Hayasaka, T. (2023). Aerosol effects on water cloud properties in different atmospheric regimes. *Journal of Geophysical Research: Atmospheres*, *128*(24), e2023JD039729. https://doi.org/10.1029/2023jd039729

Kokkola, H. (2024). *TUT-ISI/CPM: CPM 1.1*. Zenodo. https://doi.org/10.5281/zenodo.13819445

Kokkola, H., Kühn, T., Laakso, A., Bergman, T., Lehtinen, K. E. J., Mielonen, T., et al. (2018). Salsa2.0: The sectional aerosol module of the aerosol–chemistry–climate model echam6.3.0-ham2.3-moz1.0. *Geoscientific Model Development*, *11*(9), 3833–3863. https://doi.org/10.5194/gmd-11-3833-2018

Kokkola, H., Tonttila, J., Calderón, S. M., Romakkaniemi, S., Lipponen, A., Peräkorpi, A., et al. (2025). Model analysis of biases in the satellite-diagnosed aerosol effect on the cloud liquid water path. *Atmospheric Chemistry and Physics*, *25*(3), 1533–1543. https://doi.org/10.5194/acp-25-1533-2025

Korhola, T., Kokkola, H., Korhonen, H., Partanen, A.-I., Laaksonen, A., Lehtinen, K. E. J., & Romakkaniemi, S. (2014). Reallocation in modal aerosol models: Impacts on predicting aerosol radiative effects. *Geoscientific Model Development*, *7*(1), 161–174. https://doi.org/10.5194/gmd-7-161-2014

Kumar, B., Abhishek, N., Chattopadhyay, R., George, S., Singh, B. B., & Samanta, A. (2022). Deep learning based short-range forecasting of Indian summer monsoon rainfall using Earth observation and ground station datasets. *Geocarto International*, *37*(27), 17994–18021. (Published online 28 Oct 2022). https://doi.org/10.1080/10106049.2022.2138732

Levy, R., Mattoo, S., Munchak, L., Remer, L., Sayer, A., Patadia, F., & Hsu, N. (2013). The collection 6 MODIS aerosol products over land and ocean. *Atmospheric Measurement Techniques*, *6*(11), 2989–3034. https://doi.org/10.5194/amt-6-2989-2013

Li, X.-Y., Wang, H., Chakraborty, T., 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*(24), e2024GL110757. https://doi.org/10.1029/2024gl110757

Liu, Y., Zhang, J., Zhou, P., Lin, T., Hong, J., Shi, L., et al. (2018). Satellite-based estimate of the variability of warm cloud properties associated with aerosol and meteorological conditions. *Atmospheric Chemistry and Physics*, *18*(24), 18187–18202. https://doi.org/10.5194/acp-18-18187-2018

Lu, C., Liu, Y., Niu, S., & Vogelmann, A. M. (2012). Observed impacts of vertical velocity on cloud microphysics and implications for aerosol indirect effects. *Geophysical Research Letters*, *39*(21), L21808. https://doi.org/10.1029/2012gl053599

Lu, C., Liu, Y., Yum, S. S., Chen, J., Zhu, L., Gao, S., et al. (2020). Reconciling contrasting relationships between relative dispersion and volume-mean radius of cloud droplet size distributions. *Journal of Geophysical Research: Atmospheres*, *125*(9), e2019JD031868. https://doi.org/10.1029/2019jd031868

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). Curran Associates Inc.

Ma, P.-L., Rasch, P. J., Chepfer, H., Winker, D. M., & Ghan, S. J. (2018). Observational constraint on cloud susceptibility weakened by aerosol retrieval limitations. *Nature Communications*, *9*(1), 2640. https://doi.org/10.1038/s41467-018-05028-4

Malavelle, F. F., Haywood, J. M., Field, P. R., Hill, A. A., Abel, S. J., Lock, A. P., et al. (2014). A method to represent subgrid-scale updraft velocity in kilometer-scale models: Implication for aerosol activation. *Journal of Geophysical Research: Atmospheres*, *119*(7), 4149–4173. https://doi.org/10.1002/2013jd021218

Mayer, B., Emde, C., Gasteiger, J., & Kylling, A. (2017). Libradtran software package. Retrieved from http://www.libradtran.org

McComiskey, A., & Feingold, G. (2008). Quantifying error in the radiative forcing of the first aerosol indirect effect. *Geophysical Research Letters*, *35*(2). https://doi.org/10.1029/2007gl032667

McCoy, D., Bender, F.-M., Mohrmann, J., Hartmann, D., Wood, R., & Grosvenor, D. (2017). The global aerosol-cloud first indirect effect estimated using MODIS, MERRA, and AEROCOM. *Journal of Geophysical Research: Atmospheres*, *122*(3), 1779–1796. https://doi.org/10.1002/2016jd026141

Merk, D., Deneke, H., Pospichal, B., & Seifert, P. (2016). Investigation of the adiabatic assumption for estimating cloud micro- and macrophysical properties from satellite and ground observations. *Atmospheric Chemistry and Physics*, *16*(2), 933–952. https://doi.org/10.5194/acp-16-933-2016

Miinalainen, T., Kokkola, H., Lipponen, A., Hyvärinen, A.-P., Soni, V. K., Lehtinen, K. E., & Kühn, T. (2023). Assessing the climate and air quality effects of future aerosol mitigation in India using a global climate model combined with statistical downscaling. *Atmospheric Chemistry and Physics*, *23*(6), 3471–3491. https://doi.org/10.5194/acp-23-3471-2023

Mikkonen, S., Pitkänen, M. R., Nieminen, T., Lipponen, A., Isokääntä, S., Arola, A., & Lehtinen, K. E. (2019). Effects of uncertainties and number of data points on line fitting–A case study on new particle formation. *Atmospheric Chemistry and Physics*, *19*(19), 12531–12543.

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-20-12853-2020

Painemal, D., & Zuidema, P. (2010). Microphysical variability in southeast Pacific stratocumulus clouds: Synoptic conditions and radiative response. *Atmospheric Chemistry and Physics*, *10*(13), 6255–6269. https://doi.org/10.5194/acp-10-6255-2010

Park, J., Dall'Osto, M., Park, K., Gim, Y., Kang, H. J., Jang, E., et al. (2020). Shipborne observations reveal contrasting arctic marine, arctic terrestrial and Pacific marine aerosol properties. *Atmospheric Chemistry and Physics*, *20*(9), 5573–5590. https://doi.org/10.5194/acp-20-5573-2020

Park, S.-Y., & Kim, C.-H. (2021). Interpretation of aerosol effects on precipitation susceptibility in warm clouds inferred from satellite measurements and model evaluation over northeast Asia. *Journal of the Atmospheric Sciences*, *78*(6), 1947–1963. https://doi.org/10.1175/jas-d-20-0293.1

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*, 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

Pinkus, A. (1999). Approximation theory of the mlp model in neural networks. *Acta Numerica*, *8*, 143–195. https://doi.org/10.1017/s0962492900002919

Pitkänen, M. R., Mikkonen, S., Lehtinen, K. E., Lipponen, A., & Arola, A. (2016). Artificial bias typically neglected in comparisons of uncertain atmospheric data. *Geophysical Research Letters*, *43*(18), 10003–10011. https://doi.org/10.1002/2016gl070852

Platnick, S., Ackerman, S., King, M., Wind, G., Meyer, K., Menzel, P., et al. (2017). Modis atmosphere l2 cloud product (06_l2). NASA MODIS adaptive processing system, Goddard Space Flight Center. https://doi.org/10.5067/MODIS/MOD06_L2.061

Pruppacher, H. R., Klett, J. D., & Wang, P. K. (1998). *Microphysics of clouds and precipitation*. Taylor & Francis.

Quaas, J., Boucher, O., & Lohmann, U. (2006). Constraining the total aerosol indirect effect in the LMDZ and ECHAM4 GCMS using MODIS satellite data. *Atmospheric Chemistry and Physics*, *6*(4), 947–955. https://doi.org/10.5194/acp-6-947-2006

Quaas, J., Ming, Y., Menon, S., Takemura, T., Wang, M., Penner, J. E., et al. (2009). Aerosol indirect effects – General circulation model intercomparison and evaluation with satellite data. *Atmospheric Chemistry and Physics*, *9*(22), 8697–8717. https://doi.org/10.5194/acp-9-8697-2009

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

Redemann, J., & Gao, L. (2024). A machine learning paradigm for necessary observations to reduce uncertainties in aerosol climate forcing. *Nature Communications*, *15*(1), 8343. https://doi.org/10.1038/s41467-024-52747-y

Reutter, P., Su, H., Trentmann, J., Simmel, M., Rose, D., Gunthe, S. S., et al. (2009). Aerosol- and updraft-limited regimes of cloud droplet formation: Influence of particle number, size and hygroscopicity on the activation of cloud condensation nuclei (CCN). *Atmospheric Chemistry and Physics*, *9*(18), 7067–7080. https://doi.org/10.5194/acp-9-7067-2009

Romakkaniemi, S., McFiggans, G., Bower, K., Brown, P., Coe, H., & Choularton, T. (2009). A comparison between trajectory ensemble and adiabatic parcel modeled cloud properties and evaluation against airborne measurements. *Journal of Geophysical Research*, *114*(D6), D06214. https://doi.org/10.1029/2008jd011286

Rosa, G. J. (2010). *The elements of statistical learning: Data mining, inference, and prediction by Hastie, T., Tibshirani, R., and Friedman, J.* Oxford University Press.

Sanchez, K. J., Russell, L. M., Modini, R. L., Frossard, A. A., Ahlm, L., Corrigan, C. E., et al. (2016). Meteorological and aerosol effects on marine cloud microphysical properties. *Journal of Geophysical Research: Atmospheres*, *121*(6), 3178–3193. https://doi.org/10.1002/2015JD024595

Schultz, M. G., Stadtler, S., Schröder, S., Taraborrelli, D., Franco, B., Krefting, J., et al. (2018). The chemistry–climate model echam6.3-ham2.3-moz1.0. *Geoscientific Model Development*, *11*(5), 1695–1723. https://doi.org/10.5194/gmd-11-1695-2018

Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric chemistry and physics: From air pollution to climate change*. John Wiley & Sons.

Shettle, E. P. (1990). Models of aerosols, clouds, and precipitation for atmospheric propagation studies. In (Vol. *1*).

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., et al. (2013). Atmospheric component of the MPI-M Earth system model: Echam6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172. https://doi.org/10.1002/jame.20015

Stier, P., van den Heever, S. C., Christensen, M. W., Gryspeerdt, E., Dagan, G., Saleeby, S. M., et al. (2024). Multifaceted aerosol effects on precipitation. *Nature Geoscience*, *17*(8), 719–732. https://doi.org/10.1038/s41561-024-01482-6

Stull, R. B. (2012). *An introduction to boundary layer meteorology* (Vol. 13). Springer Science & Business Media.

Sullivan, S. C., Lee, D., Oreopoulos, L., & Nenes, A. (2016). Role of updraft velocity in temporal variability of global cloud hydrometeor number. *Proceedings of the National Academy of Sciences*, *113*(21), 5791–5796. https://doi.org/10.1073/pnas.1514039113

Tegen, I., Neubauer, D., Ferrachat, S., Drian, C. S.-L., Bey, I., Schutgens, N., et al. (2019). The global aerosol–climate model ECHAM6.3–HAM2.3 – Part 1: Aerosol evaluation. *Geoscientific Model Development*, *12*(4), 1643–1677. https://doi.org/10.5194/gmd-12-1643-2019

Toll, V., Christensen, M., Quaas, J., & Bellouin, N. (2019). Weak average liquid-cloud-water response to anthropogenic aerosols. *Nature*, *572*(7767), 51–55. https://doi.org/10.1038/s41586-019-1423-9

Virtanen, A., Joutsensaari, J., Kokkola, H., Partridge, D. G., Blichner, S., Seland, y., et al. (2025). High sensitivity of cloud formation to aerosol changes. *Nature Geoscience*, *18*(4), 289–295. https://doi.org/10.1038/s41561-025-01662-y

Wang, Y., Jia, H., Zhang, P., Fang, F., Li, J., Zhu, L., et al. (2024). Sensitivity of cloud microphysics to aerosol is highly associated with cloud water content: Implications for indirect radiative forcing. *Atmospheric Research*, *309*, 107552. https://doi.org/10.1016/j.atmosres.2024.107552

Wang, Y., Li, J., Fang, F., Zhang, P., He, J., Pöhlker, M. L., et al. (2024). In-situ observations reveal weak hygroscopicity in the southern Tibetan Plateau: Implications for aerosol activation and indirect effects. *npj Climate and Atmospheric Science*, *7*(1), 77. https://doi.org/10.1038/s41612-024-00629-x

Wang, Y., Lu, C., Niu, S., Lv, J., Jia, X., Xu, X., et al. (2023). Diverse dispersion effects and parameterization of relative dispersion in urban fog in eastern China. *Journal of Geophysical Research: Atmospheres*, *128*(6), e2022JD037514. https://doi.org/10.1029/2022jd037514

Warren, S. G., Hahn, C. J., London, J., Chervin, R. M., & Jenne, R. L. (1986). *Global distribution of total cloud cover and cloud type amounts over land (Tech. Rep.)*. Washington Univ., Seattle (USA). Dept. of Atmospheric Sciences; Colorado.

West, R., Stier, P., Jones, A., Johnson, C., Mann, G., Bellouin, N., et al. (2014). The importance of vertical velocity variability for estimates of the indirect aerosol effects. *Atmospheric Chemistry and Physics*, *14*(12), 6369–6393. https://doi.org/10.5194/acp-14-6369-2014

Wilcox, E. M. (2010). Stratocumulus cloud thickening beneath layers of absorbing smoke aerosol. *Atmospheric Chemistry and Physics*, *10*(23), 11769–11777. https://doi.org/10.5194/acp-10-11769-2010

Wood, R., Kubar, T. L., & Hartmann, D. L. (2009). Understanding the importance of microphysics and macrophysics for warm rain in marine low clouds. Part II: Heuristic models of rain formation. *Journal of the Atmospheric Sciences*, *66*(10), 2973–2990. https://doi.org/10.1175/2009jas3072.1

York, D., Evensen, N. M., Martinez, M. L., & De Basabe Delgado, J. (2004). Unified equations for the slope, intercept, and standard errors of the best straight line. *American Journal of Physics*, *72*(3), 367–375. https://doi.org/10.1119/1.1632486

Zheng, Y., Rosenfeld, D., & Li, Z. (2016). Quantifying cloud base updraft speeds of marine stratocumulus from cloud top radiative cooling. *Geophysical Research Letters*, *43*(21), 11–407. https://doi.org/10.1002/2016gl071185

Zheng, Y., Rosenfeld, D., Zhu, Y., & Li, Z. (2019). Satellite-based estimation of cloud top radiative cooling rate for marine stratocumulus. *Geophysical Research Letters*, *46*(8), 4485–4494. https://doi.org/10.1029/2019gl082094

Zheng, Y., Zhu, Y., Rosenfeld, D., & Li, Z. (2021). Climatology of cloud-top radiative cooling in marine shallow clouds. *Geophysical Research Letters*, *48*(19), e2021GL094676. https://doi.org/10.1029/2021gl094676

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x