nature communications

Article

https://doi.org/10.1038/s41467-024-52747-y

A machine learning paradigm for necessary observations to reduce uncertainties in aerosol climate forcing

Received: 23 February 2024

Accepted: 20 September 2024

Published online: 27 September 2024

Check for updates

Jens Redemann & Lan Gao

Uncertainties in estimates of climate cooling by anthropogenic aerosols have not decreased significantly in the last two decades, partly because observational constraints on crucial aerosol properties simulated in Earth System Models are insufficient. To help address this insufficiency in aerosol observations, we describe a paradigm for deriving higher-level aerosol properties with machine learning algorithms that use only lidar observations and reanalysis data as predictors. Our paradigm employs high-accuracy suborbital lidar and collocated in situ measurements to train and test two fully-connected neural network algorithms. We use two lidar data sets as input to our machine learning algorithms. The first data set consists of suborbital lidar observations not previously used in the training of the machine learning algorithms. The second data set consists of simulated UV-only observations to preview the algorithms' predictive capabilities in anticipation of data from the ATmospheric LIDar system on the EarthCARE satellite, which was launched in May 2024. Here we show that our algorithms predict two crucial aerosol properties, aerosol light absorption and cloud condensation nuclei concentrations with unprecedented accuracy, yielding mean relative errors of 21% and 13%, respectively, when suborbital lidar data are used as predictors. These errors represent significant improvements over conventional aerosol retrievals. Applied to future satellite missions, the paradigm presented here has great potential for constraining Earth System Models and reducing uncertainties in their estimates of aerosol climate forcing and future global warming.

Anthropogenically produced atmospheric aerosols are estimated to exert an average cooling of Earth's climate that offsets between one-fifth and one-half of the warming induced by greenhouse gasses[1]. While our understanding of the role of atmospheric aerosols in the Earth system has evolved steadily over the past two decades, uncertainties in aerosol forcing of climate have not decreased commensurately between recent IPCC (Intergovernmental Panel on Climate Change) reports[2,3]. This is partly due to continuous, new discoveries of previously unknown, complex interactions of aerosols with clouds that drive the aerosol indirect radiative forcing. Arguably the most significant hindrance to the reduction of uncertainties in aerosol forcing of climate though is the limitation of observational constraints on crucial aerosol properties simulated in Earth System Models (ESMs), e.g., refs. 1, 4. We note that past satellite observations have yielded essential global-scale insights into the distribution of atmospheric aerosols[5–8]. Likewise, suborbital measurements have provided insights into aerosol processes by providing high-accuracy measurements not obtainable from space. However, neither type of measurement has yielded the data to constrain aerosol properties and processes in global-scale ESMs to result in notable reductions in uncertainties of aerosol climate forcing, primarily because observations are often limited in accuracy and spatiotemporal resolution and coverage.

School of Meteorology, University of Oklahoma, Norman, OK, USA. e-mail: jredemann@ou.edu

Specifically, the vertical distribution of key aerosol properties, particularly in the remote atmosphere and near those types of clouds that are highly susceptible to aerosol-induced modifications, has proven an elusive measurement challenge for the observational community[9]. In this paper, we describe an ML-based paradigm for observations of crucial aerosol properties, specifically aerosol light absorption (ABS) and cloud condensation nuclei (CCN) concentrations, that can be applied to satellite lidar measurements in the next decade, but also to past suborbital lidar observations. These lidar-based ML predictions yield aerosol data in closer proximity to clouds because the active lidar measurements are not subject to some of the artifacts that afflict passive remote sensing of aerosols in the immediate vicinity of clouds[10]. ABS and CCN are aerosol properties that have not been traditionally derived from lidar measurements – we refer to these aerosol properties as higher-level aerosol properties hereafter. We note that our paper intends to present the paradigm of estimating higher-level aerosol properties with ML algorithms that use only lidar observations and reanalysis data as predictors. These ML algorithms are trained with collocated lidar measurements and in situ observations, the latter representing the truth regarding the higher-level aerosol properties. The paradigm is therefore only possible because of the recent collection of carefully coordinated lidar and in situ aerosol observations. We fully expect that future ML models that use lidar observations as input can be further optimized with evolving ML techniques and additional training data.

New satellite mission concepts are being developed to address some of the observational gaps in aerosol properties that determine aerosol climate forcing. They typically include multi-angle polarization-sensitive passive remote sensing of column-integrated aerosol quantities and active remote sensing of the vertical distributions of aerosol backscattering, extinction, and depolarization with lidars, e.g., NASA's Atmosphere Observing System (AOS)[11], and ESA/JAXA's EarthCARE (Earth Cloud, Aerosol and Radiation Explorer)[12]. These aerosol retrievals are typically complemented with observations of crucial cloud properties (e.g., cloud droplet number concentration and effective radius, liquid and ice water path) and meteorological conditions (e.g., updraft velocity) to elucidate aerosol-induced changes in cloud properties and lifetimes (e.g., ref. 13). We note that only the lidar observations provide the vertically resolved aerosol information that is crucial for addressing the most uncertain aerosol climate interactions. Hence, the paradigm developed here is primarily aimed at retrieving essential aerosol properties from spaceborne lidar observations expected to be available in the future, either from the recently launched (EarthCARE) or from the AOS project that is currently in development. Because EarthCARE and its lidar system ATLID[12] (Atmospheric LIDar) were launched recently (May 2024) and their Level 2 data is expected to be released publicly in the spring of 2025, we include ML predictions that use only UV observables in the training of the ML models and then test the models by using noise-added UV lidar observations as input. The actual error characteristics of the ATLID system will ultimately determine the errors in ML predictions that use ATLID observations as input. Nonetheless, we believe that our analysis provides a meaningful bounding exercise for the expected ML predictions based on ATLID observations.

Most remote sensing of aerosol properties relies on physics-based retrievals in an optimal estimation framework[14]. Physics-based refers to the fact that these retrievals use forward models that describe the complex interactions of electromagnetic radiation with the aerosol itself, but also the Earth's surface and the atmosphere[15]. More recently, such retrieval approaches have been applied to joint observations from lidars and polarimeters, with great promise for the increased information content from the combined observations[16]. The grand challenges for physics-based aerosol remote sensing include: (a) some details of the underlying electromagnetic scattering processes in forward models (e.g., scattering of polarized radiation with non-spherical aerosol particles and polarized surface reflectance) are unknown; (b) uncertainties in a priori retrieval assumptions propagate more or less directly into the retrieved quantities; and (c) the optimal estimation techniques are computationally expensive.

The paradigm shift presented here involves the training of ML models with lidar data and in situ measurements and then applying the models to an arbitrary lidar dataset for the retrieval of vertically resolved higher-level aerosol properties. The ML models are trained using high-accuracy lidar measurements and collocated in situ observations of key aerosol properties that are not traditionally derived from lidar observations alone. Atmospheric reanalysis data are used to constrain the ML predictions further. Because lidar systems anticipated for space deployment in the future will include only a subset of observations of the airborne lidar system we use in the training of our ML models presented here, our models can be applied directly to future satellite missions after retraining them with the available subset of observables for these missions. We illustrate this paradigm by using airborne lidar observations and collocated in situ measurements from four recent airborne field campaigns to train two sets of ML algorithms to predict aerosol light absorption (ABS) and Cloud Condensation Nuclei (CCN), respectively. These higher-level aerosol quantities are arguably the most important aerosol properties for determining aerosol effects on clouds and climate[1,17]. The majority of CCN data (>77%) available to this study was measured at supersaturations in the 0.35-0.4% SS range, hence our choice to focus the ML retrievals on $\text{CCN}_{\sim 0.4\%\text{SS}}$, although we omit the subscript in the remainder of the manuscript for readability. ABS determines the direct interactions of aerosols with solar radiation and it affects clouds by changing local atmospheric heating rates and atmospheric stability[18]. CCN is the fraction of aerosols capable of forming cloud droplets. Although a fraction of them are too small to be optically active (e.g., ref. 19), they are the key ingredient for understanding how atmospheric aerosols affect cloud droplet number concentrations, and hence the radiative properties, precipitation, and lifetimes of clouds[20,21].

Our work builds on research in which we have shown that linear regression models between lidar-derived aerosol extinction or backscattering and CCN are aerosol-type specific and only work for a limited range of relative humidities (RH)[22]. This is caused by the non-linear response of aerosol optical properties to RH increases, while CCN concentrations remain constant. The ML models we developed can predict this exact behavior from the in situ observations. Both ABS and CCN are difficult to constrain with current satellite measurements and physics-based retrieval approaches. However, global observations of their vertical distribution are urgently needed to constrain physical processes and aerosol climate-forcing estimates in ESMs. In the remainder of this paper, we use the term retrieval as a short-hand term for the predictions of the ML models, despite our awareness of the fundamental differences to the physics-based retrievals discussed above.

The lidar data used in this study were collected by the NASA Langley HSRL-2 (High Spectral Resolution Lidar-2) system[23,24]. This system is the most advanced airborne HSRL system currently in operation, providing independent measurements of aerosol backscattering and depolarization at three wavelengths (355 nm, 532 nm, 1064 nm), and aerosol extinction at 355 and 532 nm. To train the ML algorithms to predict ABS and CCN, we used HSRL-2 data from four airborne measurement campaigns, collocated with accurate in situ observations of the higher-level aerosol observables to be predicted. The airborne measurement campaigns were conducted in the continental US[25], over the Southeast Atlantic Ocean[26], near the Philippines[27], and off the US Atlantic coast[28], capturing a wide range of aerosol types and environmental conditions (Fig. 1). As the inset in Fig. 1 indicates, the probability density functions in the training dataset of CCN and ABS cover a broad range of conditions from pristine to heavily polluted, including CCN concentrations of more than 2000 per cubic centimeter, and a broad range of aerosol types, including dust,

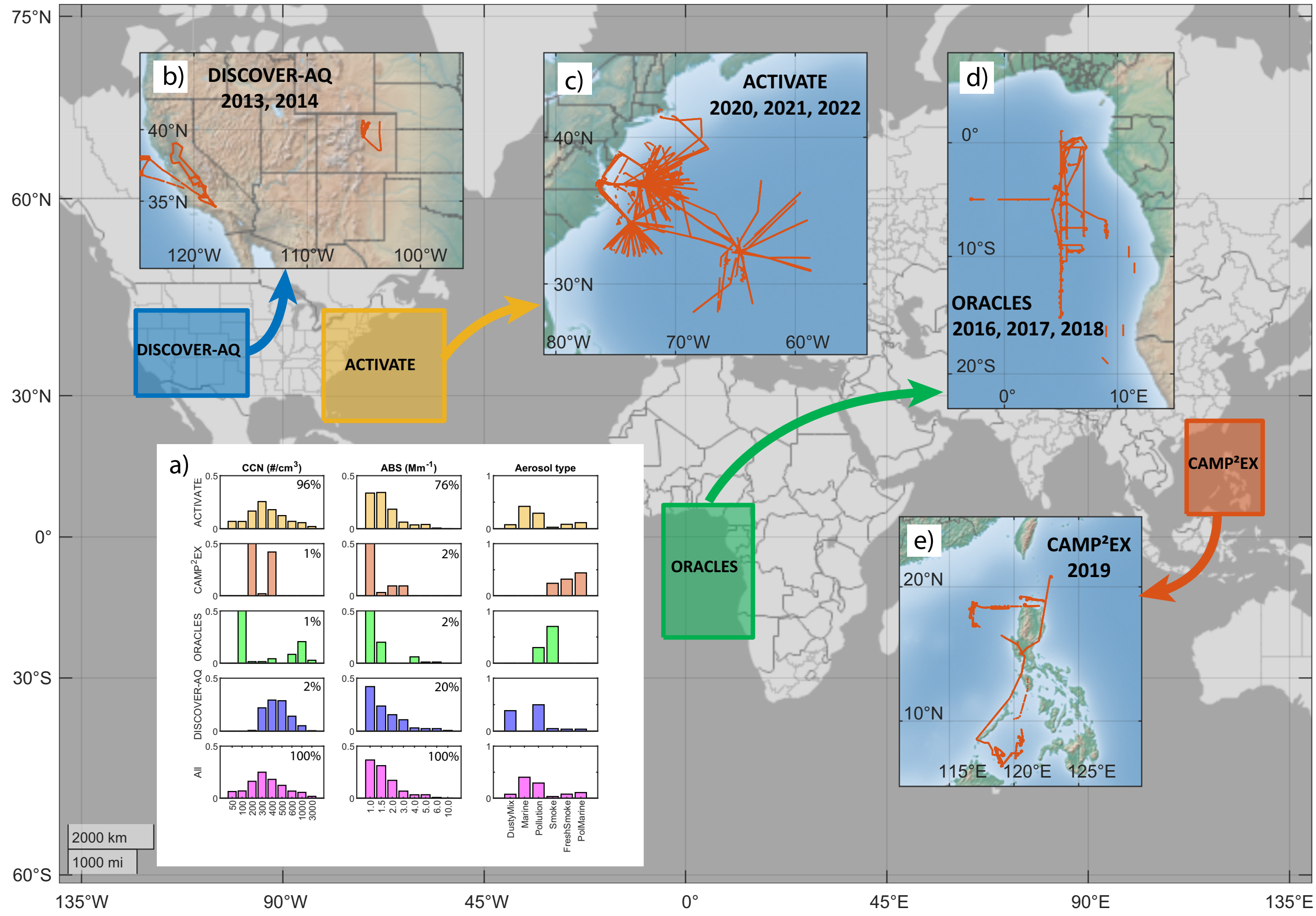

Fig. 1 | Sources of lidar and in situ data. Location of suborbital data collected in four airborne field campaigns. Inset (**a**) shows frequency distributions of CCN (cloud condensation nuclei) number concentrations (first column), aerosol absorption (ABS, second column), and lidar-derived aerosol type (third column) in each campaign. The acronyms for the field campaigns in the inset panels are defined as follows: **b** DISCOVER-AQ - Deriving Information on Surface Conditions from Column and Vertically Resolved Observations Relevant to Air Quality; **c** ACTIVATE - Aerosol Cloud meTeorology Interactions oVer the western ATlantic Experiment; **d** ORACLES - ObseRvations of Aerosols above CLouds and their intEractionS; **e** CAMP2Ex - Cloud, Aerosol and Monsoon Processes Philippines Experiment. Made with Natural Earth - free vector and raster map data @ naturalearthdata.com.

smoke, marine and pollution aerosols, and various mixtures thereof. The aerosol types in our study stem from the HSRL aerosol-type classification described in the literature[29]. This methodology uses four intensive aerosol properties derived from HSRL-2 observables, which vary independently for different aerosol types and can therefore be used to qualitatively classify observations into eight distinct aerosol types, including smoke, fresh smoke, urban, polluted marine, marine, pure dust, dusty mix, and ice.

While the majority of our collocated data (96% of CCN and 76% of ABS training data) was collected in one campaign off the US coast, we note that this dataset covers a large spatial domain and by itself a broad range of aerosol types and pollution levels. In additional analyses shown in Fig. S6, we compared the frequency of occurrence of CCN, ABS, RH, and T in our dataset to data collected at the Atmospheric Radiation Measurement (ARM) SGP (Southern Great Plains) site in the Central US. This site features one of the longest and most complete aerosol in situ datasets in the world and it is characterized by an even mix of background conditions with pollution transport events. We find strong similarities in the frequency of occurrence and the range of CCN and ABS measurements between our dataset and the SGP site data. Our dataset features a slightly larger relative occurrence of data in clean conditions which are important for CCN assessments in the context of aerosol-cloud interactions. As we demonstrate in Figs. S3 and S4, the ML models we developed have good skill in interpolating between incomplete training datasets. Therefore, we contend that the most important feature in our ML training data is the coverage of the complete range of commonly observed ABS and CCN data. Moreover, we contend that the inclusion of data from the other three airborne campaigns provides training data for specific aerosol types that may not be comprehensively represented in the ACTIVATE data alone, e.g., smoke aerosol over the Southeast Atlantic Ocean and pollution-type aerosol in the continental US and near the Philippines. While we have no global climatology of ABS and CCN to compare to, we believe that our data is generally representative of global CCN and ABS distributions in terms of data range and frequency of distributions. Nonetheless, an analysis of CCN and ABS frequency distributions stratified by aerosol type revealed some remaining training data insufficiencies, which we define as data regions with less than 2% frequency of occurrence, as illustrated in Fig. S7. The text in the supplements describes the limitations in the aerosol-type-specific representation of CCN and ABS in the training data. The impacts of sparse data on ML CCN predictions in low pollution conditions are discussed below.

In the following section ("Results" section), we provide two separate analyses. The first analysis describes the performance of ML models that use the full set of HSRL-2 observations for training and as input to the ML predictions ("ML predictions using HSRL-2 data as input" section). The results in the "ML predictions using simulated

EarthCARE/ATLID observations as input" section describe the potential performance of ML algorithms trained with the lidar observables anticipated to be available for the EarthCARE lidar system ATLID (i.e., aerosol backscatter, depolarization, and extinction at only one wavelength, 355 nm), which was launched in May 2024. For this configuration, we provide the predictions of CCN and ABS using the airborne HSRL-2 UV observations as predictors after applying random uncertainties to the extinction and backscatter data as described in the "Results" section. This exercise is meant to estimate the probable decrease in predictive capabilities of our ML models when they are applied to simulated satellite observations which necessarily possess larger uncertainties.

Both analyses provide insights into the usefulness of the ML retrieved higher-level aerosol properties. The airborne HSRL-2 data can be used with their inherent error characteristics, for example, to study the cloud nucleating properties of aerosols along the lidar curtain. The noise-added UV data in the second analysis are a proxy for future spaceborne UV-only lidar observations, which are certain to possess larger uncertainties due to lower signal-to-noise ratios for spaceborne lidars in general, because of the range-squared dependence of the lidar signal from the laser transmitter system. Hence, the results shown in the "ML predictions using HSRL-2 data as input" section are best-case-scenario ML prediction assessments and are mostly representative of lidars with the HSRL-2 system's information content and error characteristics. Adaptation to lidar observables available for a single-channel HSRL (HSRL-1) or Raman lidar system, i.e., aerosol backscatter and depolarization at two wavelengths, 532 nm and 1064 nm, and aerosol extinction at one wavelength, 532 nm, is straight-forward and will be presented in a separate paper. Such a system may better represent the capabilities of the Raman lidar system that may be contributed to the NASA AOS project by the Italian Space Agency. Conceptually, the approach we outline here can be applied to any set of lidar observations. However, the predictive capabilities of the ML model will depend on the accuracy and precision of the lidar observations, the independence of aerosol backscatter and extinction retrievals, and the availability and accuracy of collocated in situ observations to train the ML model.

Results

In this section, we discuss the performance of the ML algorithms for CCN and ABS predictions jointly, because of the general similarities in the algorithms' training and validation. We present results that use the full suite of HSRL-2 observables to train our ML models and as input to the ML predictions in the "ML predictions using HSRL-2 data as input" section. In the "ML predictions using simulated EarthCARE/ATLID observations as input" section, we discuss the performance of ML models trained with the lidar observables available for the EarthCARE lidar system ATLID (i.e., aerosol backscatter, depolarization, and extinction at only one wavelength, 355 nm). For this configuration, we provide predictions of CCN and ABS using the HSRL-2 UV observations as input to the model after applying random noise to each observation. The noise generation is described in the "Data availability" section.

ML Predictions using HSRL-2 data as input

Using the HSRL-2 data as input to the ML predictions, Fig. 2 indicates that the ML predictions yield correlations of 0.93 (0.80) for CCN (ABS) when lidar observables are used to train the ML models; these correlations increase to 0.97 (0.90) for CCN (ABS) when the T and RH reanalysis data is used as additional constraints. The density plot function indicating data density in Fig. 2 is adapted from the original dscatter function provided by MATLAB. First, the natural logarithms of the original X and Y variables are calculated. These values are then divided into 200 equal logarithmic bins to create a 2D histogram for density calculation. Finally, the individual data point densities are scaled by the maximum density to a range of 0–1.

For CCN (ABS), the ML predictions based on lidar observations alone [panels a (c), also Fig. 3, and Table 1] have mean relative errors of 22% (25%). These errors are reduced to 13% (21%) when reanalysis products are used as additional predictors [panels b (d)]. Other performance metrics, including the predictions' mean absolute error (MAE), the correlation coefficient R, and the number and fraction of predictions within either 30% or 50% error are summarized in Table 1.

ML predictions using simulated EarthCARE/ATLID observations as input

In analogy to Figs. 2, 4 shows the ML predictions for ML models trained with EarthCARE/ATLID observables only and using noise-added HSRL −2 UV observations as input to simulate an application to satellite measurements anticipated to be released publicly in the spring of 2025, as described above. Figure 4 indicates that ML predictions thus produced yield correlations with the test data of 0.57 for CCN and 0.55 for ABS, respectively, when only the lidar observables are used to train the ML models (panels a and c); these correlations increase to 0.88 and 0.74, respectively, when the reanalysis data is incorporated as additional constraints.

For CCN (ABS), the ML predictions based on lidar observations alone and applied to simulated ATLID measurements [panels a (c), also Fig. 5, and Table 1] have mean relative errors of 51% (40%). These errors are reduced to 23% and 28%, respectively, when reanalysis products are used as additional predictors [panels b (d)]. Other performance metrics, including the predictions' mean absolute error (MAE), and the number and fraction of predictions within either 30% or 50% error are summarized in Table 1.

Discussion

The ML model trained with, and applied to, the full set of airborne HSRL-2 observations is able to provide vertically resolved CCN to within 30% uncertainty in 85% of all retrievals, when reanalysis data of T and RH are used as additional predictors. This constitutes a tremendously useful tool for studying CCN concentrations in the vicinity of clouds from past and future airborne HSRL observations. The model that is trained with UV observations and applied to noise-added observations in order to simulate EarthCARE ATLID observations still yields vertically resolved CCN to within 50% uncertainty in 83% of all retrievals. This capability is far beyond any published physics-based lidar retrievals, including recently published efforts using CALIOP lidar measurements by Choudhury and Tesche[30], hereafter C&T, which provide a general CCN uncertainty of a factor of 2. We note that the methodology developed in C&T was applied to actual satellite observations, while our study applies the ML models to simplistically simulated satellite observations. We note that the lidar measurements we used for training and input to our ML models provide independent measurements of aerosol extinction and backscatter, as a benefit of the HSRL technique. We note further that the uncertainty comparison above is inadequate in a number of ways. The physics-based retrievals are subject to the full range of parametric, structural, forward model, and measurement uncertainties, while our ML prediction errors primarily capture uncertainty in the suborbital data and the ML algorithms themselves, with the latter affected by the quality and comprehensiveness of the training dataset. We include the comparison to the case-by-case analyses in the literature merely as a reference to the current state of global CCN and ABS retrievals. Future applications of ML models to actual satellite observations and their comparison to suborbital observations will have to provide a more complete uncertainty assessment.

For ABS retrievals, there are very few physics-based lidar retrieval studies available in the peer-reviewed literature. Recent satellite projects, including the NASA AOS mission, aimed for uncertainties in the retrieved aerosol single scattering albedo (SSA, ratio of aerosol scattering to aerosol extinction, where extinction is the sum of scattering

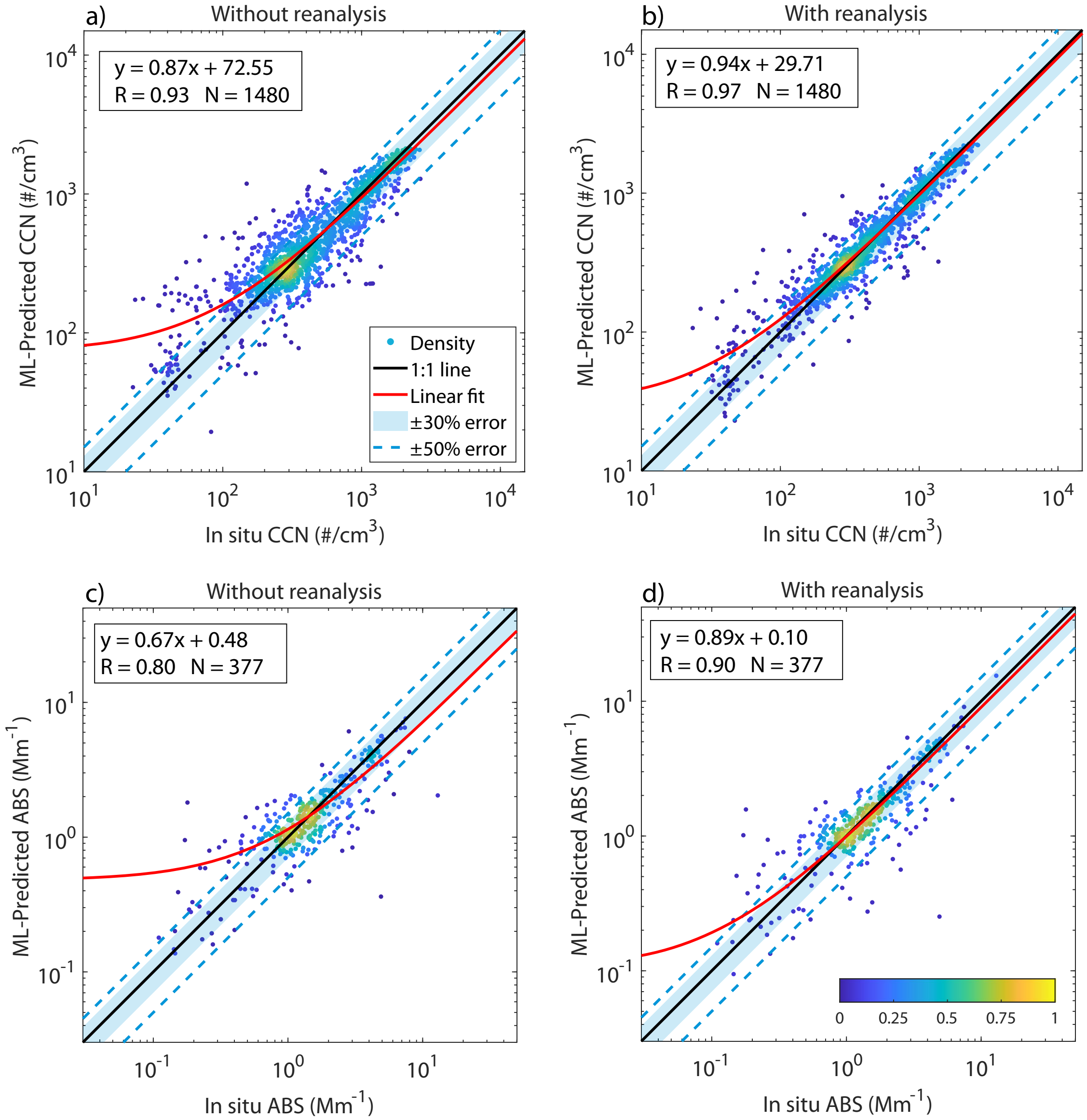

Fig. 2 | Evaluation of machine learning (ML)-based aerosol property predictions. ML-predicted CCN (cloud condensation nuclei) concentrations (**a**, **b**) and aerosol absorption (ABS, **c**, **d**) versus in situ observed quantities. **a** and **c** depict the ML algorithms' performance when only lidar observables are used for training; **b** and **d** depict the ML algorithms' performance when reanalysis data of temperature (T) and relative humidity (RH) are used as additional predictors. Data points are color-coded based on data density, with yellow indicating higher density and blue indicating lower density.

and absorption) of ±0.03 in the mid-visible and recent theoretical work[30] indicated SSA uncertainties in physics-based HSRL retrievals of approximately 0.025. Propagating these uncertainties into an absorption uncertainty for the simulated ATLID dataset yields an ABS uncertainty range of $1.7–2.0 \times 10^{-6}$ m^{-1}. Therefore, our ML-based mean ABS retrieval error for the simulated ATLID observations of 0.52×10^{-6} m^{-1} (see Table 1) represents a lower uncertainty by more than a factor of three.

As a visualization of the great potential of the ML algorithms introduced here, Fig. 6 shows a curtain plot of CCN estimated using HSRL observations from a transatlantic flight on August 26, 2016. The results in Fig. 6 were produced with an algorithm that was trained without the use of lidar depolarization because such measurements were not available for some of the most pristine altitudes of the particular flight shown. The algorithm trained without depolarization is only marginally less capable and it reveals a level of detail in the vertical distribution of CCN that is unparalleled in the published literature. We note that no example of a CCN curtain retrieval of similar resolution or accuracy exists in the peer-reviewed literature to date.

The ML-based predictions of higher-level aerosol properties (i.e., ABS and CCN) from lidar observations and reanalysis data presented in this study provide a paradigm shift and hence a unique aerosol remote sensing retrieval capability with significant advantages over currently available physics-based retrieval methods. The paradigm of estimating higher-level aerosol properties with ML algorithms that use only lidar observations and reanalysis data as predictors represents a quantum leap forward in accuracy and coverage of the retrieved aerosol properties by comparison to currently available aerosol retrievals. Specific advantages of this paradigm include the provision of data in close proximity to clouds due to the availability of lidar measurements near clouds, and the provision of higher-level aerosol properties not traditionally derived from lidar observations. The core ingredient for the development of the underlying ML models is a high-accuracy lidar dataset collocated with in situ observations of targeted aerosol

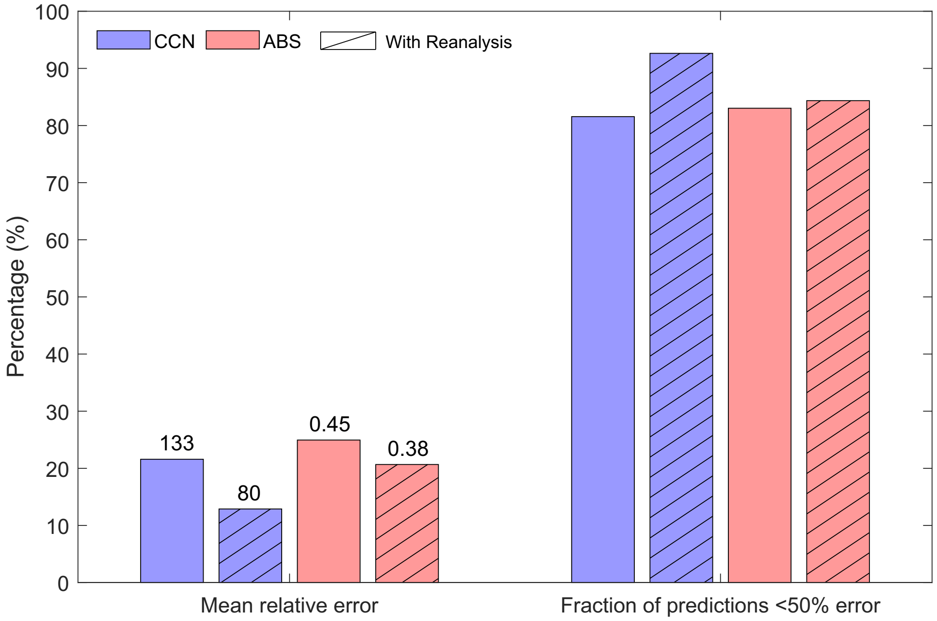

Fig. 3 | Statistical machine learning (ML) prediction performance for ML models applied to complete airborne High Spectral Resolution Lidar (HSRL)-2 data. Mean relative error (MRE, left set of bars) and fraction of predictions within 50% uncertainty (right set of bars) for CCN (cloud condensation nuclei, blue) and aerosol absorption (ABS, red), with reanalysis data used as predictors (hashed) and without (solid). Numbers above MRE bars indicate mean absolute errors (MAE) in cm^{-3} for CCN and Mm^{-1} (1/10^{6} m) for ABS.

Table. 1 | Machine learning model prediction performance for two lidar configurations for aerosol absorption (ABS) and cloud condensation nuclei (CCN), without and with reanalysis data used as predictors

Predictor dataset →		**Lidar-only**				**Lidar + Reanalysis data**			
Lidar configuration ↓	Performance Indicator →	Mean Absolute Error (Relative)	Corr. Coef.	*N* within ±30%	*N* within ±50%	Mean Absolute Error (Relative)	Corr. Coef.	*N* within ±30%	*N* within ±50%
HSRL^{-2}	CCN [1/cm^{3}]	133 (22%)	0.93	975 (66%)	1207 (82%)	80 (13%)	0.97	1261 (85%)	1371 (93%)
	ABS [10^{-6} m^{-1}]	0.45 (25%)	0.80	245 (65%)	313 (83%)	0.38 (21%)	0.90	276 (73%)	318 (84%)
ATLID observables (UV HSRL^{-2} + noise)	CCN [1/cm^{3}]	310 (51%)	0.57	406 (32%)	669 (52%)	141 (23%)	0.88	900 (70%)	1059 (83%)
	ABS [10^{-6} m^{-1}]	0.76 (40%)	0.55	139 (46%)	203 (67%)	0.52 (28%)	0.74	202 (66%)	250 (82%)

properties, with reanalysis offering a boost in performance to the ML predictions.

When the ML models are applied to airborne HSRL-2 data, the predictions of CCN in particular represent a capability that has not been realized from remote sensing observations before. Applied to complete airborne lidar datasets, this capability is bound to provide vastly improved aerosol observations near clouds, which will elucidate aerosol-cloud interactions with accuracy and coverage well beyond previously available datasets. The preview of ML-based observational capabilities for higher-level aerosol properties using simulated ATLID observation also yields auspicious results. While our ATLID data simulation is simplistic, we made various conservative choices in quantifying the potential error characteristics of such a dataset, rendering our ML predictions of aerosol properties on the basis of ATLID observations entirely plausible.

We note that the ML methodology presented here can be readily applied to any spaceborne or suborbital lidar observations in the future. An application to a particular lidar system, similar to the application of specifically trained ML models to simulated ATLID UV observations, will only require the retraining of the ML models with the exact set of observables available in such a system. This process is feasible because currently considered spaceborne lidar systems will observe only a subset of the airborne HSRL-2 observables.

Nonetheless, we emphasize that our ML-based solution to estimating key aerosol properties from lidar observations is not a replacement for physics-based retrievals, because ML models in general provide no insight into the underlying physics of the retrieval problem. Instead, in our vision, the ML models will augment physics-based retrievals by providing computationally inexpensive ways to retrieve aerosol properties with high accuracy where physics-based retrievals have not yet been perfected or may not be possible at an acceptable computational expense, or in physics-based retrievals where the uncertainty in a priori assumptions causes unacceptable uncertainties in the derived products. It is likely that physics-based retrievals in combination with evolved ML methods will provide the best approach to aerosol retrievals from satellite observations in the future, as recently opined[31].

At the current state of our ML model development, we have evidence that independent lidar observations of aerosol extinction and backscatter, such as those afforded by HSRL and Raman lidar systems, are crucial for the training of the models and their subsequent use to derive aerosol properties. It is reasonable to assume that the increased information content inherent in HSRL observations is one of the fundamental features that enable the ML algorithms to discover the nonlinear, multivariate correlations between the aerosol optical properties observable by the HSRL and the microphysical and chemical

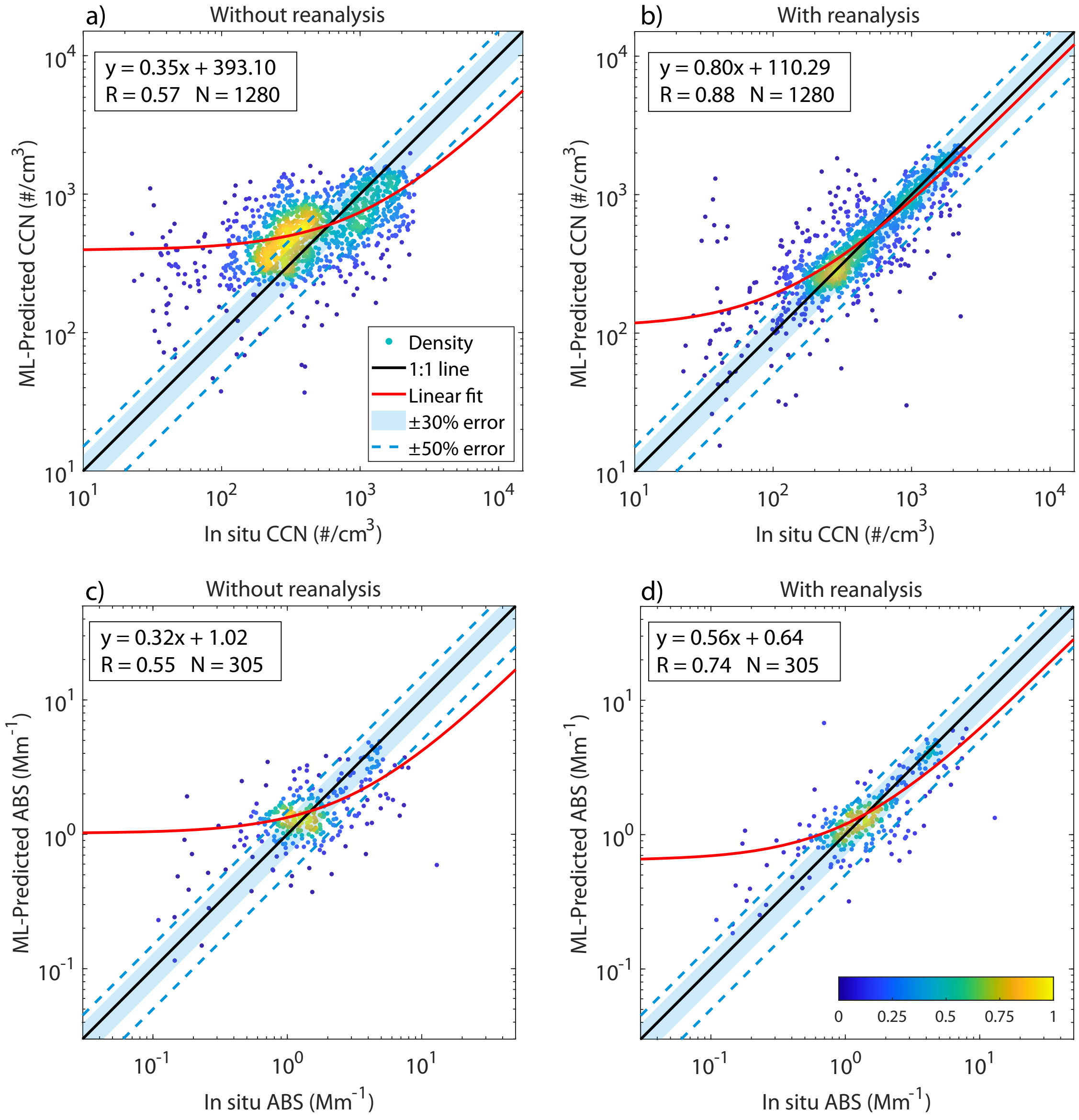

Fig. 4 | Evaluation of ML-based aerosol property predictions for machine learning (ML) models applied to simulated ATLID (Atmospheric LIDar) observations. ML-predicted CCN (cloud condensation nuclei) concentrations (**a**, **b**) and aerosol absorption (ABS, **c**, **d**) versus in situ observed quantities. **a**, **c** depict the ML algorithms' performance when only lidar observables are used for training; **b**, **d** depict the ML algorithms' performance when reanalysis data of temperature (T) and relative humidity (RH) are used as additional predictors. Data points are color-coded based on data density, with yellow indicating higher density and blue indicating lower density.

properties of the aerosol population, including properties in the CCN size range. As such, the application of similarly trained ML algorithms to future spaceborne lidar observations, as previewed by our application of the models to simulated EarthCARE ATLID observations, may unlock a view of the global distribution of aerosol properties that was previously shrouded in great uncertainty or completely concealed. Using such aerosol properties to confront ESMs will invariably lead to improved simulations of the global distribution of aerosols and assessments of their role in climate forcing and future global warming.

Methods

Details of the suborbital data used for the training and testing of ML models

For the training and testing of our ML algorithms, we used lidar data that was spatially collocated with in situ observations within horizontal distances of 1100 m and vertical distances of 45 m, and coincident within 30 min. These collocation criteria yielded 9873 lidar measurements paired with in situ CCN measurements in the supersaturation range of 0.35%–0.4% at which the majority of in situ CCN data were measured, and 2516 lidar measurements coincident with ABS measurements to train our ML models. The in situ observations of ABS and CCN were taken from a second aircraft that flew in close spatio-temporal coordination below the HSRL platform. For a small fraction of data (<2%), the in situ observations were collected from the same aircraft as the HSRL after maneuvering the aircraft into the column previously observed below the aircraft by the HSRL.

The archived horizontal resolution of HSRL-2 aerosol backscatter, depolarization, and extinction measurements depends on aircraft speed and is on average 1500–2000 m corresponding to a 10-s average of the native lidar observations; the vertical resolution is 15 m. Our spatial averaging and collocation scheme is intended to find all in situ observations within an 1100 m horizontal distance of the central lidar

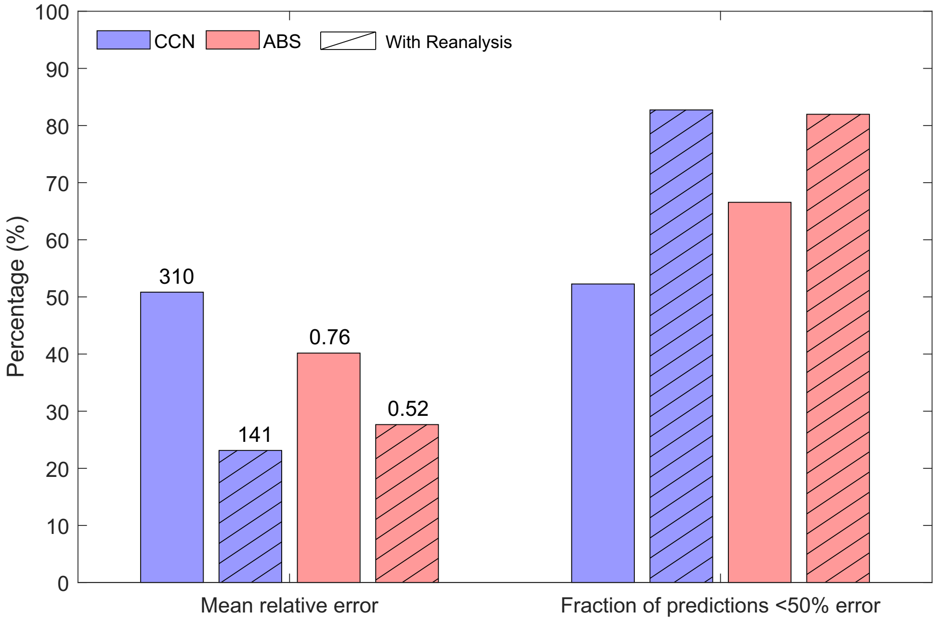

Fig. 5 | Statistical machine learning (ML) model prediction performance for ML models applied to simulated Atmospheric LIDar (ATLID) observations. Mean relative error (MRE, left set of bars) and fraction of predictions within 50% uncertainty (right set of bars) for cloud condensation nuclei (CCN, blue) and aerosol absorption (ABS, red), with reanalysis data used as predictors (hashed) and without (solid). Numbers above MRE bars indicate mean absolute errors (MAE) in cm^{-3} for CCN and Mm^{-1} ($1/10^6$ m) for ABS.

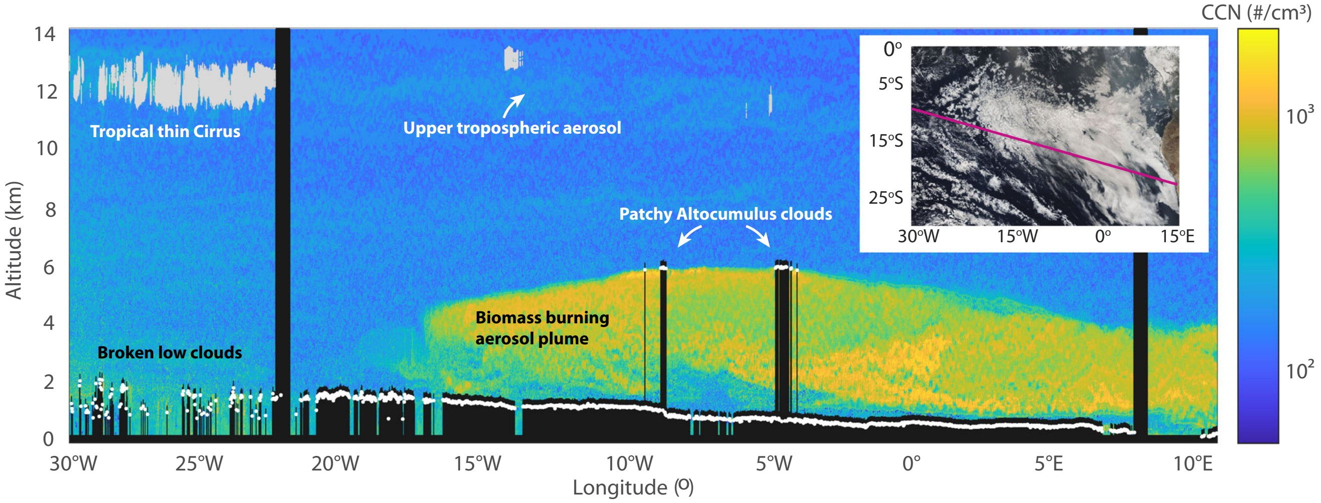

Fig. 6 | Curtain plot of cloud condensation nuclei (CCN). Data was derived using the machine learning (ML) model based on High Spectral Resolution Lidar (HSRL)-2 observations during a NASA ER−2 transatlantic flight on August 26, 2016, from Recife, Brazil, to Walvis Bay, Namibia. The white dots indicate locations identified by the lidar returns as the tops of clouds where no lidar returns are present below the indicated altitude.

data location, and it allows for vertical variations of the in situ aircraft inside a horizontal HSRL bin. The vertical averaging is set to 45 m, so that a maximum of 3 vertical lidar measurement bins and 20 in situ measurement bins may be located inside each collocation bin, using the in situ data sampling of 1 Hz and an assumed average aircraft speed of 100 m/s for the in situ aircraft. These collocation criteria are adapted from our recent work on simple correlation studies[22].

We used CCN data from continuous flow CCN counters and absorption data from Particle Soot Absorption Photometers (PSAP). CCN counters measure CCN concentrations at different levels of supersaturation (SS) in standard mode and scanning-flow mode[32]. The CCN counters in each of the four experiments were operated in similar modes, typically alternating between periods of constant supersaturation (SS) or scanning through SS ranges to create CCN spectra. The majority of CCN data (>77%) available to this study was measured in the 0.35%–0.4% SS range, hence our choice to focus the ML retrievals on $CCN_{0.4\%SS}$, although we omit the subscript in the remainder of the manuscript for readability. The uncertainty associated with the CCN number concentration is typically 10% at high signal-to-noise ratios. The PSAP is a filter-based instrument that collects aerosol particles on a substrate and measures the change in light transmission at three wavelengths (467, 530, and 660 nm) relative to a reference filter. Bulk aerosol absorption is derived after correcting for scattering effects. PSAP uncertainty is an ongoing research topic and depends on filter loading, aerosol type, and corrections applied to data. Under relatively clean conditions, a recent study[33] estimated PSAP ABS uncertainties of 40%. In aircraft operations, PSAP uncertainty is highly sensitive to pressure changes. Thus, we carefully filtered the PSAP data and discarded measurements when consecutive aircraft altitude variations exceeded 5 meters vertically. Both instrument types were operated in all four campaigns, providing data at a temporal resolution of 1 s. Before collocating with HSRL lidar observations, CCN concentrations below 10 cm^{-3} and ABS below 0.1 Mm^{-1} were discarded to account for the low sensitivity and relatively larger uncertainty of both measurements in clean conditions.

We note that in our methodology the lidar data used as input to the ML algorithms are independent of the lidar data used to train the algorithms in the first place unless the dataset shows strong spatial autocorrelation at the horizontal sampling scales (2.2 km) we applied. We tested the potential spatial autocorrelation in our dataset by removing any data points that were located within a horizontal distance of ±5 km from any other data points in the dataset. The results of this exercise show a slight decrease in the predictive capability of our algorithms (see Fig. S2), for example from 12.9% (see Table 1) mean relative error in CCN to 16.9% (see Fig. S2a). However, even this slight decrease in predictive capability cannot be attributed to spatial autocorrelation of the original dataset, because the simple reduction of the number of data points in the reduced dataset may have a similarly negative impact on the algorithms' predictive performance. To further test the fidelity of our ML retrieval results, we investigated the impact of the completeness of the training dataset in terms of the range and continuity of the suborbital in situ measurements of ABS and CCN. These tests indicate sufficient sampling completeness for the ABS and CCN datasets used in the training of our models and are described in the supplementary materials, specifically Figs. S3 and S4 and the accompanying text.

Reanalysis data used as additional predictors in the ML models

In addition to using the lidar observables as predictors in our ML algorithms, we explored the use of reanalysis data, in particular temperature (T) and relative humidity (RH), as additional predictors. They were added in the input layer at the same level as the lidar observables and their addition did not change the architecture of our ML algorithms. The reanalysis data, sourced from the ERA5 (ECMWF Reanalysis v5) product[34], were spatiotemporally interpolated to match the lidar data locations. ERA5 reanalysis data has a spatial resolution of 0.25° × 0.25° in the horizontal, 37 pressure levels in the vertical, and a temporal resolution of 1 h. The product is spatially and temporally interpolated to the exact time and location of the airborne data using linear interpolation. Figure S6 shows the frequency distribution of RH and T in the reanalysis dataset collocated with our lidar and in situ observations in each campaign, similar to the frequency distributions of CCN and ABS in the inset of Fig. 1. We chose T and RH as additional predictors because they describe the thermodynamic state of the atmosphere in which the lidar observations were collected. RH, in particular, determines the size and refractive index of atmospheric aerosols under ambient conditions and thereby the lidar-observed aerosol optical properties, i.e., extinction and backscattering[35]. We note that the frequency distributions of RH and T between the four airborne campaigns in Fig. S6 showed many similarities and significant overlap in RH frequency distributions in particular. Therefore, it seems unlikely that the Fully Connected Neural Networks (FCNN) that we employ inadvertently distinguish and preferentially select training data from any one of the four flight campaigns. The addition of reanalysis data improves the ML retrieval capabilities significantly, as we show above.

Features of the ML architecture

Our ML algorithms are based on Deep Learning (DL) with FCNN. We trained a feedforward neural network regression model using the Levenberg-Marquardt algorithm[36,37]. While such ML algorithms have been used extensively in the atmospheric sciences and specifically in Meteorology[38], their use in retrievals of geophysical variables from remote sensing observations is relatively new (e.g., ref. 39). We explored the use of FCNN and Random Forest algorithms for our application. Both methods demonstrated comparable performance in predicting CCN and ABS. However, the training time for the Random Forest algorithm was nearly twice as long as that for the FCNN. Consequently, FCNNs were chosen for the final model training in this study. For future applications of this proposed paradigm, it is possible to implement other training algorithms within this framework.

In our DL architecture, the lidar observables and reanalysis data constitute the predictors, while the in situ measured CCN and ABS constitute the algorithm responses. For each algorithm, we use 70% of the available collocated lidar and in situ observations as the training dataset, 15% as the validation dataset, and 15% as the test dataset. The models' hyperparameters are tuned iteratively using Bayesian optimization. We use 10-fold cross-validation, which means that we train the ML models on subsets of the training data and validate them with the complementary subsets ten times. The training of neural networks largely depends on the selection of optimal hyperparameters (including the number of layers, number of neurons per hidden layer, activation functions, etc.). Therefore, we have included the optimal sets of these important hyperparameters specific to the models we trained for this study in Table S2. Further details of our data selection, along with a description of the FCNN and its setup can be found in Fig. S1, and Tables S1 and S2 in the supplementary information.

During the review of our manuscript, several reviewers commented on the relatively larger uncertainties in our ML predictions of CCN and ABS in clean conditions (i.e., CCN < 100/cm^3; ABS < 1 Mm^{-1}). In response, we pursued a technique that is commonly used in ML algorithms and entails an extra weighting of data points in data-sparse regions, in this case, the clean conditions. Specifically, we explored the retraining of our ML models for CCN/ABS with a training dataset that assigned extra weighting to the data in clean conditions (CCN < 100/cm^3; ABS < 1 Mm^{-1}). A risk to such weighting of data in specific data ranges is that it can result in less good fitting of the data outside of the extra-weighted data range. We investigated additional weighting of data in clean conditions by up to a factor of 10 and found that the mean relative uncertainty (MRE) in CCN could be reduced significantly with an increased weighting of data in clean conditions by a factor of 3, with relatively little impact on the MRE in overall conditions (i.e., all predictions). The improvement in ML predictions at the low CCN range between equal weighting of all data versus triple weighting of the data in clean conditions is illustrated in Fig. S5. As a result of this analysis, all ML model predictions of CCN in this study use a triple weighting of training data in the clean CCN conditions.

An analogous analysis for extra weighting of ABS training data in clean conditions indicated that extra weighting of ABS training data in clean conditions resulted in significantly increased MRE in ABS predictions for all conditions. The reason for the greater impact of the extra weighting of the ABS training data in clean conditions on the overall ABS prediction performance (by comparison to CCN retrievals) is that a relatively larger fraction of ABS data resides in clean conditions (more than 60% of overall ABS data). Unlike the CCN predictions, the ABS predictions in clean conditions are climatically less important than the predictions at higher absolute ABS. In combination with the increased penalties, i.e., the higher MRE for overall ABS conditions, we decided to use equal weighting of all ABS training data.

Simulating EarthCARE/ATLID error characteristics

In the "ML predictions using simulated EarthCARE/ATLID observations as input" section, we discussed the performance of ML algorithms trained with the set of lidar observables that is anticipated to be available for the EarthCARE lidar system ATLID (i.e., aerosol backscatter, depolarization, and extinction at only one wavelength, 355 nm). The uncertainties we applied are derived from actual comparisons between airborne HSRL-2 observations and collocated spaceborne CALIOP (Cloud-Aerosol LIdar with Orthogonal Polarization) extinction and backscatter retrievals. For the HSRL backscatter data used as input to the ML models, we assigned random noise by assuming a Gaussian distribution, such that the mean of the noise added equals the mean relative difference between the collocated CALIOP and HSRL data. We then propagated the noise in the backscatter data to the extinction observations using uncertainties in lidar ratios published in the literature[40]. In this manner, we created a UV-only dataset of aerosol backscatter, extinction, and depolarization from airborne HSRL-2 measurements that is noisier than the HSRL-2 airborne observations themselves, in an attempt to simulate the error characteristics of the EarthCARE/ATLID system. Such a noise envelope may well be an overestimate of the uncertainties that we can expect from future spaceborne HSRL and night-time Raman observations because the CALIOP extinction retrievals are subject to systematic errors in the assumptions of lidar ratios, which are necessary in the inversion of elastic backscatter lidar measurements as is the case for CALIOP.

Data availability

All processed data, including the CCN and ABS ML predictions generated in this study, have been deposited in a designated CodeOcean capsule[41] (https://doi.org/10.24433/CO.3891939.V1). The lidar and in situ datasets used in this study as input to the ML model training and evaluation are available for download from the NASA ESPO (Earth Science Project Office) data archive websites as indicated below. ORACLES datasets are available at the following links: https://espo.nasa.gov/oracles/archive/browse/oracles/id22/P3. https://espo.nasa.gov/oracles/archive/browse/oracles/id14/P3. https://espo.nasa.gov/oracles/archive/browse/oracles/id8/ER2. https://espo.nasa.gov/oracles/archive/browse/oracles/id8/P3. ACTIVATE datasets are available at the following links. https://www-air.larc.nasa.gov/cgi-bin/ArcView/activate.2022?KINGAIR=1. https://www-air.larc.nasa.gov/cgi-bin/ArcView/activate.2022?MERGE=1. https://www-air.larc.nasa.gov/cgi-bin/ArcView/activate.2021?UC12=1. https://www-air.larc.nasa.gov/cgi-bin/ArcView/activate.2021?MERGE=1. https://www-air.larc.nasa.gov/cgi-bin/ArcView/activate.2019?UC12=1. https://www-air.larc.nasa.gov/cgi-bin/ArcView/activate.2019?MERGE=1. CAMP2EX datasets are available at the following links: https://www-air.larc.nasa.gov/cgi-bin/ArcView/camp2ex?P3B=1. https://www-air.larc.nasa.gov/cgi-bin/ArcView/camp2ex?MERGE=1. DISCOVER-AQ datasets are available at the following links: https://www-air.larc.nasa.gov/cgi-bin/ArcView/discover-aq.co-2014?B200=1. https://www-air.larc.nasa.gov/cgi-bin/ArcView/discover-aq.co-2014?MERGE=1. https://www-air.larc.nasa.gov/cgi-bin/ArcView/discover-aq.ca-2013?B200=1. https://www-air.larc.nasa.gov/cgi-bin/ArcView/discover-aq.ca-2013?MERGE=1. ECMWF ERA5 datasets are available at the following link: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=form.

Code availability

Data, custom codes, and a Readme file to reproduce all results and figures in this paper are available in a designated CodeOcean capsule[41] (https://doi.org/10.24433/CO.3891939.V1).

References

1. Bellouin, N. et al. Bounding global aerosol radiative forcing of climate change. *Rev. Geophys.* **58**, e2019RG000660 (2020).
2. Intergovernmental Panel on Climate Change. Clouds and aerosols. In *Climate Change 2013 – The Physical Science Basis* 571–658 (Cambridge Univ. Press, 2014). https://doi.org/10.1017/CBO9781107415324.016.
3. Intergovernmental Panel on Climate Change. *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2023). https://doi.org/10.1017/9781009157896.
4. Thornhill, G. D. et al. Effective radiative forcing from emissions of reactive gases and aerosols – a multi-model comparison. *Atmos. Chem. Phys.* **21**, 853–874 (2021).
5. Yu, H. et al. Global view of aerosol vertical distributions from CALIPSO lidar measurements and GOCART simulations: regional and seasonal variations. *J. Geophys. Res.* **115**, D00H30 (2010).
6. Remer, L. A. et al. The MODIS aerosol algorithm, products, and validation. *J. Atmos. Sci.* **62**, 947–973 (2005).
7. Kahn, R. A. et al. Satellite-derived aerosol optical depth over dark water from MISR and MODIS: comparisons with AERONET and implications for climatological studies. *J. Geophys. Res.* **112**, D18205 (2007).
8. Deuzé, J. L. et al. Estimate of the aerosol properties over the ocean with POLDER. *J. Geophys. Res.* **105**, 15329–15346 (2000).
9. Stier, P. Limitations of passive remote sensing to constrain global cloud condensation nuclei. *Atmos. Chem. Phys.* **16**, 6595–6607 (2016).
10. Marshak, A. et al. Aerosol properties in cloudy environments from remote sensing observations: a review of the current state of knowledge. *Bull. Am. Meteorol. Soc.* **102**, E2177–E2197 (2021).
11. NASA AOS - Home. https://aos.gsfc.nasa.gov/ (2023).
12. Wehr, T. et al. The EarthCARE mission – science and system overview. *Atmos. Meas. Tech.* **16**, 3581–3608 (2023).
13. Seinfeld, J. H. et al. Improving our fundamental understanding of the role of aerosol-cloud interactions in the climate system. *Proc. Natl Acad. Sci. USA* **113**, 5781–5790 (2016).
14. Rodgers, C. D. *Inverse Methods for Atmospheric Sounding: Theory and Practice*, Vol. 2 (World Scientific, 2000).
15. Stamnes, S. et al. Simultaneous polarimeter retrievals of microphysical aerosol and ocean color parameters from the "MAPP" algorithm with comparison to high-spectral-resolution lidar aerosol and ocean products. *Appl. Opt.* **57**, 2394 (2018).
16. Xu, F. et al. A combined lidar-polarimeter inversion approach for aerosol remote sensing over ocean. *Front. Remote Sens.* **2**, 620871 (2021).
17. Rosenfeld, D. et al. Global observations of aerosol-cloud-precipitation-climate interactions: aerosol-cloud-climate interactions. *Rev. Geophys.* **52**, 750–808 (2014).
18. Myhre, G. et al. Quantifying the importance of rapid adjustments for global precipitation changes. *Geophys. Res. Lett.* **45**, 11399–11405 (2018).
19. Riemer, N., Ault, A. P., West, M., Craig, R. L. & Curtis, J. H. Aerosol mixing state: measurements, modeling, and impacts. *Rev. Geophys.* **57**, 187–249 (2019).
20. Rosenfeld, D. et al. Satellite retrieval of cloud condensation nuclei concentrations by using clouds as CCN chambers. *Proc. Natl Acad. Sci. USA* **113**, 5828–5834 (2016).
21. Marinescu, P. J. et al. Impacts of varying concentrations of cloud condensation nuclei on deep convective cloud updrafts—a multimodel assessment. *J. Atmos. Sci.* **78**, 1147–1172 (2021).
22. Lenhardt, E. D. et al. Use of lidar aerosol extinction and backscatter coefficients to estimate cloud condensation nuclei (CCN) concentrations in the southeast Atlantic. *Atmos. Meas. Tech.* **16**, 2037–2054 (2023).
23. Hair, J. W. et al. Airborne high spectral resolution lidar for profiling aerosol optical properties. *Appl. Opt.* **47**, 6734 (2008).
24. Burton, S. P. et al. Information content and sensitivity of the 3β + 2α lidar measurement system for aerosol microphysical retrievals. *Atmos. Meas. Tech.* **9**, 5555–5574 (2016).

25. Crawford, J. H. & Pickering, K. E. Discover-AQ: advancing strategies for air quality observations in the next decade. *EM Air Waste Manag. Assoc.* **9**, 4–7 (2014).
26. Redemann, J. et al. An overview of the ORACLES (ObseRvations of Aerosols above CLouds and their intEractionS) project: aerosol–cloud–radiation interactions in the southeast Atlantic basin. *Atmos. Chem. Phys.* **21**, 1507–1563 (2021).
27. Reid, J. S. et al. The coupling between tropical meteorology, aerosol lifecycle, convection, and radiation during the Cloud, Aerosol and Monsoon Processes Philippines Experiment (CAMP2Ex). *Bull. Am. Meteorol. Soc.* **104**, E1179–E1205 (2023).
28. Sorooshian, A. et al. Spatially coordinated airborne data and complementary products for aerosol, gas, cloud, and meteorological studies: the NASA ACTIVATE dataset. *Earth Syst. Sci. Data* **15**, 3419–3472 (2023).
29. Burton, S. P. et al. Aerosol classification using airborne high spectral resolution lidar measurements – methodology and examples. *Atmos. Meas. Tech.* **5**, 73–98 (2012).
30. Choudhury, G. & Tesche, M. Estimating cloud condensation nuclei concentrations from CALIPSO lidar measurements. *Atmos. Meas. Tech.* **15**, 639–654 (2022).
31. Remer, L. A., Levy, R. C. & Martins, J. V. Opinion: Aerosol Remote Sensing Over The Next Twenty Years. *Atmos. Chem. Phys.* **24**, 2113–2127 (2024).
32. Roberts, G. C. & Nenes, A. A continuous-flow streamwise thermal-gradient CCN chamber for atmospheric measurements. *Aerosol Sci. Technol.* **39**, 206–221 (2005).
33. Asmi, E. et al. Absorption instruments inter-comparison campaign at the Arctic Pallas station. *Atmos. Meas. Tech.* **14**, 5397–5413 (2021).
34. Hersbach, H. et al. The ERA5 global reanalysis. *Quart. J. R. Meteor. Soc.* **146**, 1999–2049 (2020).
35. Düsing, S. et al. Measurement report: comparison of airborne, in situ measured, lidar-based, and modeled aerosol optical properties in the central European background – identifying sources of deviations. *Atmos. Chem. Phys.* **21**, 16745–16773 (2021).
36. Hagan, M. T. & Menhaj, M. B. Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Netw.* **5**, 989–993 (1994).
37. Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* **11**, 431–441 (1963).
38. McGovern, A. et al. Making the black box more transparent: understanding the physical implications of machine learning. *Bull. Am. Meteorol. Soc.* **100**, 2175–2199 (2019).
39. Chase, R. J., Nesbitt, S. W. & McFarquhar, G. M. A dual-frequency radar retrieval of two parameters of the snowfall particle size distribution using a neural network. *J. Appl. Meteorol. Climatol.* **60**, 341–359 (2021).
40. Kim, M.-H. et al. The CALIPSO version 4 automated aerosol classification and lidar ratio selection algorithm. *Atmos. Meas. Tech.* **11**, 6107–6135 (2018).
41. Redemann, J. & Gao, L. A machine learning paradigm for necessary observations to reduce uncertainties in aerosol climate forcing. *Code Ocean* https://doi.org/10.24433/CO.3891939.V1 (2024).

Acknowledgements

We gratefully acknowledge the extensive efforts of instrument teams that have collected the suborbital data used for the training of ML models in this study. As the main component of our analysis relies on the HSRL data as predictors, we are grateful for the NASA Langley HSRL team led by Richard Ferrare, Chris Hostetler, and John Hair for their tireless efforts to maintain a truly unique set of instruments, producing high-quality data with remarkable reliability. We are further indebted to the teams deploying the in situ instrument teams in the four field deployments mentioned above, as well as the teams that made these large-scale field deployments possible. Furthermore, we are grateful for the support of the OU (University of Oklahoma) OSCER (OU Supercomputing Center for Education & Research) team who maintain the supercomputing capabilities used for the ML simulations in this study. Funding for this work was provided by the National Aeronautics and Space Administration (NASA) through grants 80NSSC19K1086 and 80NSSC23M0083 (J.R., L.G.). We gratefully acknowledge the help with initial data processing by Emily Lenhardt and help with the layout of Fig. 1 by Benting Chen, both graduate students in the University of Oklahoma's School of Meteorology.

Author contributions

J.R. contributed to this work by acquiring research funding, providing the basic research concept, helping to devise the ML methodology and data visualization, and leading the development of this manuscript. L.G. made equal contributions to the concept, led the development of the Machine Learning models and the creation of all research results, and made substantial contributions to the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52747-y.

Correspondence and requests for materials should be addressed to Jens Redemann.

Peer review information *Nature Communications* thanks Yu Wang and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints