npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-025-01181-y

Interpretable ensemble learning unveils main aerosol optical properties in predicting cloud condensation nuclei number concentration

Nan Wang[1], Yuying Wang[1] ✉, Chunsong Lu[1], Bin Zhu[1], Xing Yan[2], Yele Sun[3], Jialu Xu[1], Junhui Zhang[1] & Zhuoxuan Shen[1]

Variations in cloud condensation nuclei number concentration (N_{CCN}) significantly influence cloud microphysics, yet direct N_{CCN} measurements remain challenging. Here, we present an N_{CCN} ensemble learning (NEL) model utilizing ensemble learning and interpretability analysis on aerosol optical parameters. Validated at two land sites, two ocean sites and one polar site within the Atmospheric Radiation Measurement program, the mean absolute percentage error range of the NEL model across different environments is from 12% to 36%, demonstrating high accuracy. Key findings reveal that aerosol optical parameters can serve as predictors for N_{CCN}. Aerosol scattering and backscattering coefficients, absorption coefficient, backscatter fraction (BSF), and Ångström exponent (AE) are positively correlated with N_{CCN}, while single scattering albedo shows negative correlations. N_{CCN} prediction at land sites is highly sensitive to BSF, largely driven by the backscattering coefficient, as fine particles dominate in these sites. At ocean sites, N_{CCN} prediction is more sensitive to AE, primarily influenced by the scattering coefficient, due to the higher proportion of larger particles. At the polar site, N_{CCN} prediction shows sensitivity to both BSF and AE, mainly driven by the scattering coefficient, as polar sites are cleaner and contain larger particles. These differences reflect the variation in particle size and number concentration across different environments.

Defined as a mixture of solid and liquid particles suspended in the air, atmospheric aerosol is a major factor influencing the Earth's radiation balance. It can also affect the water cycle through influencing cloud and precipitation processes[1,2]. Cloud condensation nuclei (CCN) refer to aerosol particles that can activate to form cloud and fog droplets under supersaturated water vapor conditions. Changes in CCN number concentration (N_{CCN}) can lead to variations in cloud physics, further changing precipitation and cloud radiation balance[3,4]. Among the uncertainties in aerosol-related global climate effective radiative forcing, the aerosol-cloud interaction (ACI) contributes the most[5,6]. Therefore, accurately describing N_{CCN} in the atmosphere is crucial, as it will help reduce uncertainties in ACI modeling.

Köhler theory[7] provides a fundamental theory linking CCN activity to aerosol physicochemical properties. Numerous studies have shown that aerosol particle size, chemical composition, hygroscopicity and mixing state are the primary factors affecting CCN activity[8–12]. However, these factors exhibit significant spatial and temporal variability across the world[13,14], and due to the complexity of measurements and models, such data is not easily obtained with high accuracy, which adds to the uncertainty of N_{CCN} prediction.

Aerosol optical parameters are relatively easier to obtain through observational methods such as lidar and satellites. While utilizing these parameters to predict N_{CCN} holds significant appeal, it remains a challenging task across diverse environmental conditions[15,16]. For instance, Jefferson (2010) demonstrated that the uncertainty in N_{CCN} predictions tends to increase at lower particle concentrations[17]. Similarly, in the study by Shen et al. [18], despite the development of several complex models across multiple

[1]State Key Laboratory of Climate System Prediction and Risk Management/Key Laboratory for Aerosol—Cloud Precipitation of China Meteorological Administration/Special Test Field of National Integrated Meteorological Observation, Nanjing University of Information Science & Technology, Nanjing, China. [2]Faculty of Geographical Science, Beijing Normal University, Beijing, China. [3]State Key Laboratory of Atmospheric Environment and Extreme Meteorology, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China. ✉e-mail: yuyingwang@nuist.edu.cn

regions, significant errors persisted in some cases, particularly when supersaturation levels increased[18]. In contrast, the approach proposed by Shinozuka et al. [19] revealed even larger errors at lower supersaturation levels, with errors reaching up to three times the value of the best estimate. These findings underscore the complexity and variability in the relationship between CCN and aerosol optical properties[19]. Additionally, aerosol optical parameters can partly reflect characteristics such as particle size, shape, and changes in aerosol hygroscopicity[17,19–23]. Previous studies have established relationships between N_{CCN} and single aerosol optical parameters, such as aerosol optical depth (AOD)[20], or employed multiple aerosol optical parameters, such as backscatter fraction (BSF) and single scattering albedo (SSA) to predict N_{CCN}[17,19,21], achieving promising results. However, most of these studies are based on single-site measurements with limited variables, making the N_{CCN} prediction methods less universally applicable. Therefore, developing models based on multi-site observations across different environments is essential to provide more universally applicable N_{CCN} predictions.

Over the past few decades, the development and use of machine learning (ML) have been booming, and it has been applied to atmospheric sciences recently[24–27]. Previous scholars have studied the application of ML in predicting N_{CCN}[24,28,29], specifically using aerosol optical parameters for the prediction[28,29]. Their work demonstrated that ML achieved overall success in deriving N_{CCN} under different aerosol physical and chemical conditions. Notably, ML can extract information such as aerosol size from aerosol composition and aerosol optical parameters, indicating that the statistical learning of ML algorithms is rooted in fundamental physical and chemical principles[30]. However, the "black box" nature of ML makes it difficult to interpret how input features influence the output results[31]. The SHapley Additive exPlanations (SHAP) algorithm offers a promising solution to this challenge and has already shown significant progress in studies related to ozone formation and boundary layer height inversion[31–33], which has not been used to predict N_{CCN}.

This study aims to develop an N_{CCN} ensemble learning (NEL) model for predicting N_{CCN} and to enhance its interpretability using the SHAP algorithm. The approach begins by evaluating various models, selecting the top three for ensemble learning, and then training the ensemble model. The NEL model is subsequently applied to predict N_{CCN}, with SHAP used for interpretative analysis to quantify the contributions of different aerosol optical parameters in the prediction process. Finally, the study examines the importance and interactions of these aerosol optical parameters in predicting N_{CCN} over land, ocean and polar regions.

Results

Model preparation

In developing the NEL model, eXtreme Gradient Boosting (XGBoost)[34], Categorical Boosting (CatBoost)[35], and Random Forest (RF)[36] were selected due to their complementary strengths in addressing the complexities of environmental datasets. XGBoost and CatBoost utilize gradient boosting to refine predictions sequentially, excelling at capturing nonlinear relationships and complex feature interactions. In contrast, RF applies bagging and random feature selection to provide robust generalization and model diversity by emphasizing different aspects of the input space. Averaging predictions from these three models allows the NEL ensemble to mitigate individual model biases and errors, enhancing robustness and predictive accuracy. This design is particularly suitable for datasets with multifaceted characteristics, such as aerosol optical properties relevant to N_{CCN} estimation[37,38]. Details regarding model construction are provided in the "Methods" section.

To validate model performance, the models are trained under identical computational conditions using data from the atmospheric radiation measurement (ARM) SGP site, employing widely used ML techniques. The models tested include Decision Tree (DT), Support Vector Machine (SVM), RF, Bagging-SVM, Adaptive Boosting—Logistic Regression (AdaBoost-LR), CatBoost, Light Gradient Boosting Machine (LightGBM), XGBoost, and the NEL model. Further details are provided in Supplementary Text 1, and simulation results for each model are illustrated in Supplementary Fig. 1. The prediction accuracy is evaluated using five metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Relative Euclidean Distance (RED)[39,40], and the coefficient of determination (R^2) (details in the "Methods" section).

Performance evaluation (Fig. 1) shows that XGBoost, CatBoost, and RF achieve R^2 values of 0.57, 0.58, and 0.55, respectively, while the NEL model reaches an R^2 of 0.63 and the lowest RED (0.32), demonstrating superior predictive performance. Although the NEL model increases computational demand compared to individual algorithms, its improved accuracy and enhanced robustness justify the cost.

The aerosol optical parameters used in the models include scattering coefficient (σ_{sp}), backscattering coefficient (σ_{bsp}), absorption coefficient (σ_{ap}), backscatter fraction (BSF), Ångström exponent (AE) and single scattering albedo (SSA). The letters B, G, and R following these parameters represent measurements at three specific wavelengths: Blue (464 nm), Green (529 nm), and Red (648 nm). For instance, σ_{sp}_B represents the σ_{sp} at the blue wavelength. The AE parameter with letters indicates that it is computed from the scattering coefficients at two wavelengths. For example, AE_BR denotes the AE calculated using the blue and red scattering coefficients. Considering that aerosol optical parameters vary with relative humidity (RH)[41,42], all aerosol optical parameters are measured under dried conditions to ensure a more comprehensive and accurate analysis. In a separate study[43], the influence of RH on CCN estimation based on aerosol optical properties was explored, and a corresponding parameterization method was proposed. Table 1 outlines the specific instruments used to measure each parameter and explains how these parameters contribute to the prediction of N_{CCN}, ensuring a clear scientific basis for their inclusion in the model.

NEL model performance and analysis of correlation between N_{CCN} and aerosol optical parameters

As shown in Fig. 1, the established NEL model outperforms other models in predicting N_{CCN}. Figure 2 presents density scatter plots comparing predicted and measured N_{CCN} values for the test sets across five sites (land sites: SGP, GUC; ocean sites: ENA, ASI; polar site: MOS). Additionally, line plots of 500 randomly selected test samples from each site are generated (Supplementary Fig. 2). These results demonstrate a high degree of consistency between the N_{CCN} predictions from the NEL model and the actual values, with R^2 values for the five sites being 0.63, 0.92, 0.70, 0.65 and 0.83. The model achieves low MAE and RMSE values, especially at ASI (Fig. 2b) with larger datasets. However, at SGP (Fig. 2a), the MAE and RMSE are highest, likely due to the strongest variation in N_{CCN} values at this site. Despite this, the MAPE and RED remain consistently low across all five sites. Even at GUC and MOS (Fig. 2c, e), where the sample size is smaller, the NEL model demonstrates strong performance, highlighting its robustness across both large and small datasets.

The SHAP method is employed to interpret the outputs of the NEL model, as illustrated in Fig. 3. It is found that the aerosol optical parameters with the highest contributions are all related to aerosol scattering parameters. During the prediction process at the five sites, it is observed that higher values of σ_{bsp}, σ_{sp}, and σ_{ap} correspond to larger SHAP values, indicating a positive correlation between these parameters and N_{CCN}. This correlation likely arises from the fact that these parameters are positively correlated with aerosol number concentration; higher values typically imply a greater number of particles that can be activated as CCN, leading to higher N_{CCN}. This finding aligns with previous studies[21,44,45]. BSF and AE also positively correlate with N_{CCN}, which is closely linked to particle size, but the relationship between BSF and N_{CCN} is more pronounced at land sites, which is also consistent with an earlier study[44]. A detailed comparison of the aerosol physicochemical properties across the five sites is provided in the Supplement (Supplementary Figs. 3 and 4).

Additionally, SSA shows a minimal contribution (Fig. 3), which may be due to SSA reflecting the influence of differences in aerosol chemical composition on N_{CCN}. The minor contribution of SSA suggests that aerosol

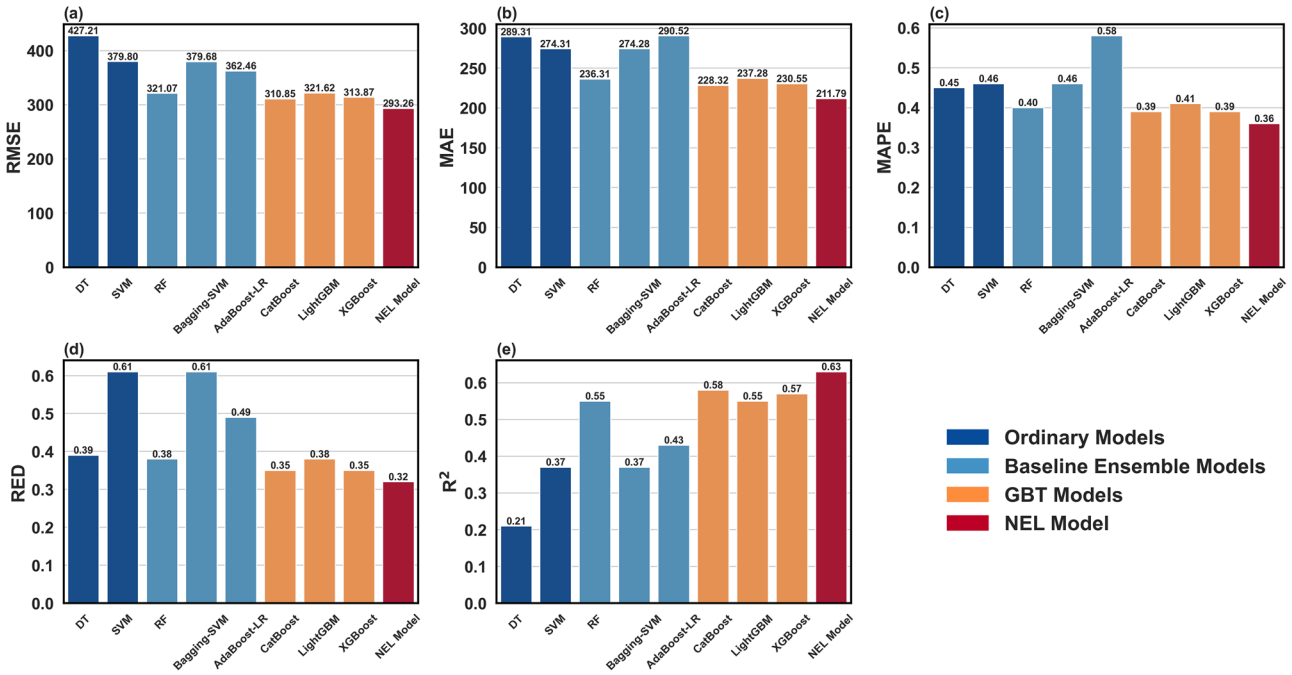

Fig. 1 | Performance comparison of different models. The performance comparison of different models using statistical parameters of **a** Root mean square error (RMSE), **b** mean absolute error (MAE), **c** mean absolute percentage error (MAPE), **d** relative euclidean distance (RED) and **e** determination coefficient (R^2), where smaller RMSE, MAE, MAPE, and RED, and larger R^2 indicate a better model performance.

Table 1 | This table outlines all the aerosol optical parameters and meteorological variables used in predicting N_{CCN}, including absorption coefficient (σ_{ap}), scattering coefficient (σ_{sp}), backscattering coefficient (σ_{bsp}), single scattering albedo (SSA), backscatter fraction (BSF) and Angstrom exponent (AE)

Variable	Instruments/Method	Role in N_{CCN} prediction
σ_{ap}	Particle soot absorption photometer (PSAP)	Reflects the aerosol number concentration, particularly the concentration of absorbing aerosols.
σ_{sp}	Nephelometer	Reflects the aerosol number concentration, especially sensitive to large particles.
σ_{bsp}	Nephelometer	Reflects the aerosol number concentration, especially sensitive to fine particles.
SSA	$\frac{\sigma_{sp}}{\sigma_{sp}+\sigma_{ap}}$	Reflects the aerosol chemical composition and hygroscopicity.
BSF	$\frac{\sigma_{bsp}}{\sigma_{sp}}$	Reflects the shape and size of particles, more sensitive to fine particles.
AE	$-\frac{\log(\sigma_{sp}(\lambda_1)/\sigma_{sp}(\lambda_2))}{\log(\lambda_1/\lambda_2)}$	Reflects the particle size, more sensitive to large particles.

It specifies the measurement instruments for each parameter and their respective roles in the prediction model.

chemical composition has a limited impact on N_{CCN}, indirectly indicating that aerosol number concentration and particle size are more significant factors for predicting N_{CCN}, consistent with previous studies[8,11,43].

At most sites, aerosols are predominantly composed of particles in the Aitken and accumulation modes. These smaller particles exhibit higher scattering efficiency at shorter wavelengths, making aerosol scattering parameters at the blue wavelength particularly effective predictors of N_{CCN}. Overall, SHAP values effectively clarify the correlation between each variable and N_{CCN} during the prediction process.

Importance of aerosol optical parameters to N_{CCN} prediction

An aerosol optical parameter is identified as a major driving predictor for a specific site type (land, ocean or polar) if its average relative contribution across sites of the same type is 15% or higher. At land sites, the primary driving predictor is σ_{bsp}_B, contributing 20.75% overall, with SGP at 23.32% and GUC at 18.17% (Supplementary Table 11). This pronounced influence is likely attributable to the greater complexity and heterogeneity of aerosol types and morphologies in continental environments, where fine particles are typically more abundant, thereby enhancing the sensitivity of σ_{bsp} environmental variability.

At the ocean sites, the major driving predictor is σ_{sp}_B, contributing 21.22%, with ENA at 20.20% and ASI at 22.24%. N_{CCN} is closely related to σ_{sp}_B, likely because N_{CCN} is controlled by Aitken mode and accumulation mode particles in ocean regions[46], where sulfate aerosols and organic aerosols are abundant[47–51]. The aerosol size spectrum at ocean sites is broader due to the presence of sea salt, which leads to greater variability in particle size distributions. The substantial presence of coarse-mode particles promotes the relationship between σ_{sp}_B and N_{CCN}.

At the polar site (MOS), the primary driving predictor is σ_{sp}_B, which contributes 18.85% to the model performance. Although the Arctic atmosphere is generally characterized by lower aerosol concentrations[52], it is predominantly influenced by fine particles, with occasional contributions from sea salt. These conditions enhance the relevance of σ_{sp}, which effectively captures the substantial variability in aerosol concentration despite the overall lower loading. This emphasizes that in most environments, the aerosol number concentration is a key factor in predicting N_{CCN}.

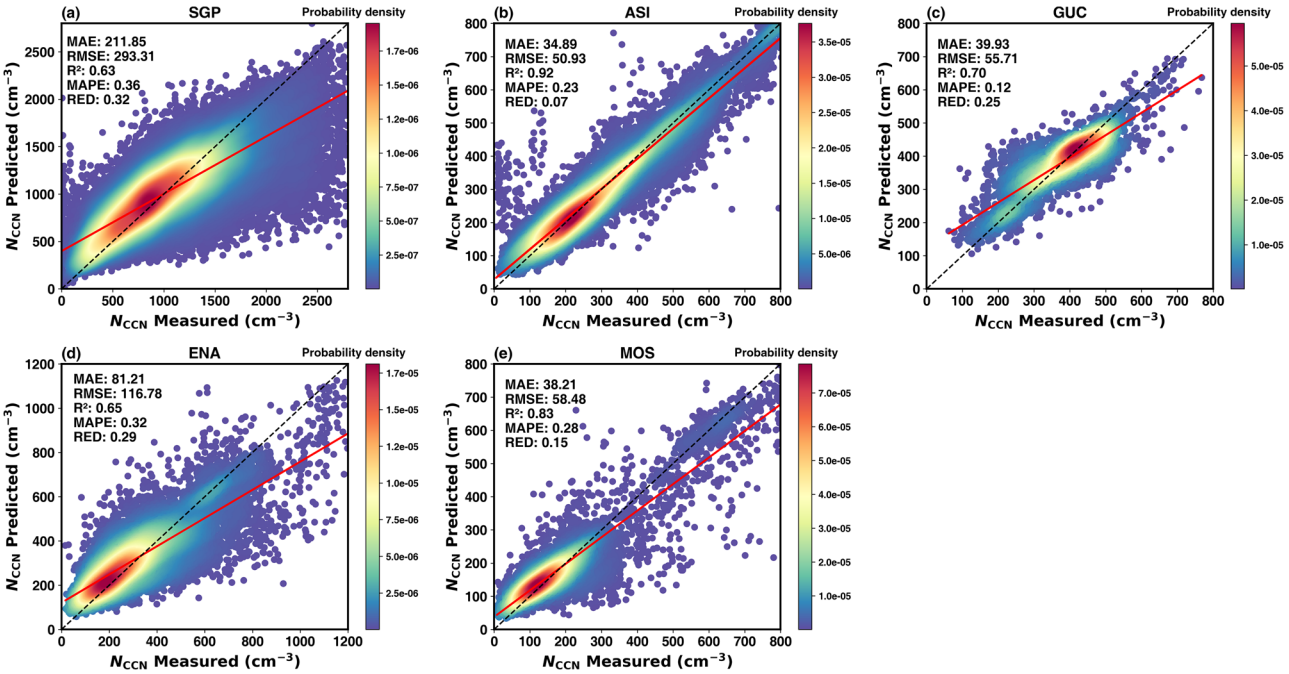

Fig. 2 | Comparison of N_{CCN} predictions and observations at five sites. Density scatter plots of N_{CCN} predicted by the NEL model are shown in (**a**–**e**), representing the results for SGP, ASI, GUC, ENA, and MOS, respectively. The horizontal axis shows the observed N_{CCN} at 0.4% supersaturation, while the vertical axis shows the model-predicted N_{CCN} at 0.4% supersaturation. The black dashed line denotes the fitted line for the ideal case, and the red solid line is the actual fitted line. The point colors indicate density, with representing higher point density.

The mean relative contributions of each aerosol optical parameter at three wavelengths across five sites are determined (Fig. 4 and Supplementary Fig. 5). The results indicate that the relative contributions of BSF and σ_{bsp} are higher at land sites compared to other sites. Specifically, BSF at the SGP site contributes 21.33%, which is more than twice the contributions at other sites. Over land, aerosols originate from diverse sources such as biomass burning and urban pollution, resulting in fine particles with irregular shapes[53]. The σ_{bsp} effectively captures the scattering behavior of these irregular fine particles, indicating that their number concentration plays a critical role in CCN activation in continental environments. Previous studies indicate BSF is more sensitive to smaller particles, while AE responds more to larger ones[54,55].

In contrast, the relative contributions of AE and σ_{sp} are greater at ocean and polar sites due to larger-sized particles prevalent in these regions. The elevated contribution of AE underscores the pivotal role of particle size distribution in governing N_{CCN} levels over ocean and polar environments. Unlike land sites dominated by complex organic aerosols, the ocean and polar atmosphere contain a higher proportion of regularly shaped particles[56], thereby diminishing the enhancement of BSF typically caused by particle shape irregularity. Additionally, at ASI and MOS, σ_{bsp} also proves to be particularly influential, likely because aerosols are primarily composed of long-range transported fine particles, with occasional contributions from sea salt aerosols[52,57,58].

Notably, the relative contribution of σ_{ap} is higher at the GUC site, likely due to the prolonged wildfires nearby during the observation period, which generated substantial amounts of black carbon and brown carbon aerosols[59]. These aerosols can become CCN after undergoing aging and growth processes[60]. High relative contributions of σ_{ap} are also observed at SGP and ENA, likely due to the presence of carbonaceous aerosols from biomass burning at SGP[53]. ENA, with its large population of permanent residents, experiences significant contributions of fresh black carbon from both traffic and daily activities[47]. Additionally, the relative contribution of SSA is minimal, with only slight variations between land and ocean sites. However, a deeper analysis of the impact of aerosol optical parameters on aerosol activation rate (AR), defined as the ratio of N_{CCN} to the total aerosol number concentration, reveals a significant contribution of SSA to AR (Supplementary Figs. 6, 7, and 8), with an average contribution of 21.46%. This indicates that SSA indirectly influences N_{CCN} by affecting AR, although its contribution is limited.

Furthermore, our results indicate that the contribution of aerosol optical parameters, such as SSA and BSF, to the model's performance is not the most significant. This suggests that minor errors in these parameters from remote sensing data are unlikely to substantially affect the overall model performance. However, uncertainty in σ_{bsp} may introduce errors in predictions for land sites. For example, the uncertainty in σ_{bsp} from satellites is approximately 30%[61], which could lead to 7% error in the NEL model's prediction of N_{CCN}. Similarly, uncertainties in AE could also lead to errors in the N_{CCN} prediction for ocean sites. To achieve more accurate predictions, it may be necessary to apply more precise estimation algorithms to satellite data or rely on accurate ground-based observational data.

In summary, the differences in aerosol physicochemical properties among different sites lead to varying contributions of aerosol optical parameters to N_{CCN} prediction. The NEL model, combined with SHAP analysis, effectively captures these differences, enabling a detailed assessment of their relative contributions across different sites.

Interaction effects between aerosol optical parameters

To further investigate the interaction effects between aerosol optical parameters on N_{CCN} prediction, this study utilizes SHAP dependency plots to analyze the main effects of individual variables and their interactions (Figs. 5 and 6). Detailed interaction processes are illustrated in Supplementary Figs. 9–13. Specifically, given the significant differences in AE and BSF between different sites, the interactions between the aerosol optical parameters with the largest contributions, σ_{bsp}_B for land sites, σ_{sp}_B for ocean sites and σ_{sp}_B for polar sites, are analyzed alongside AE_BR and BSF_G, which have the highest contributions at most sites.

At land sites (SGP and GUC), the dispersion of the blue sample dots above the *y*-axis zero line in Fig. 5a, b shows that when σ_{bsp}_B is less than 2 Mm^{-1}, a low BSF_G amplifies the positive impact of σ_{bsp}_B, resulting in an increase in N_{CCN} (high SHAP value). In the range of 2 to 4 Mm^{-1}, a low BSF_G causes σ_{bsp}_B to have a negative contribution. When

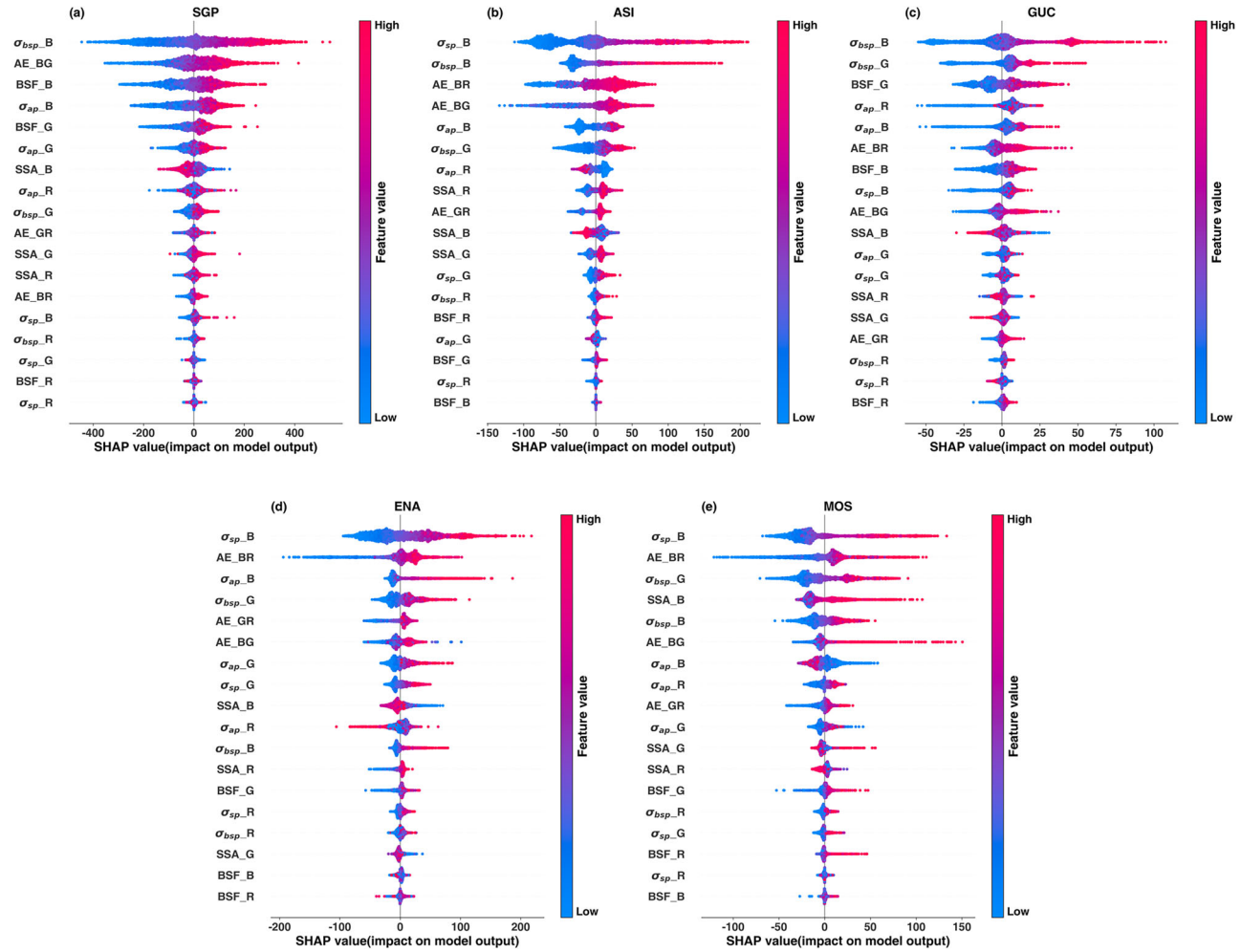

Fig. 3 | The positive and negative correlations between aerosol optical parameters and N_{CCN}, as well as the importance of aerosol optical parameters. In predicting N_{CCN}, the SHAP values for each feature are denoted as follows: **a**–**e** represent SGP, GUC, ASI, ENA, and MOS. The letters B, G, and R following these parameters represent measurements at three specific wavelengths: Blue (464 nm), Green (529 nm), and Red (648 nm). For instance, σ_{sp}_B represents the σ_{sp} at the blue wavelength. The AE parameter with letters indicates that it is computed from the total scattering at two wavelengths. For example, AE_BR denotes the AE calculated using the blue and red total scattering values. The mean of the absolute values of the SHAP values indicates the importance of each variable to N_{CCN} prediction. In each plot, feature importance is arranged from top to bottom, with the width of the bars indicating the sample size. The color of the points represents the value of the corresponding variable, with warmer colors indicating higher values and cooler colors indicating lower values. Full variable abbreviations can be found in Supplementary Tables 1–5.

σ_{bsp}_B < 2 Mm^{-1}, it typically reflects low aerosol loading conditions. In this regime, a low BSF_G indicates a dominance of larger particles. The contribution of σ_{bsp}_B is associated with enhanced activation of these larger particles, thereby contributing positively to N_{CCN} (as shown by positive SHAP values). In contrast, a high BSF_G (with smaller particles) points to the prevalence of smaller particles, implying that the contribution of σ_{bsp}_B may result from particles too small to be efficiently activated as CCN, which leads to a weaker or even negative effect on N_{CCN} (as shown by negative SHAP values).

When σ_{bsp}_B > 2 Mm^{-1}, the overall aerosol number concentration is likely higher, and its influence on N_{CCN} becomes more pronounced. In this context, for a given σ_{bsp}_B, a lower BSF (indicative of a greater fraction of larger particles) generally corresponds to a lower particle number concentration, thereby reducing N_{CCN}. Conversely, a higher BSF suggests a greater abundance of smaller particles, which increases the number of potential CCN and thus enhances N_{CCN}. Notably, when σ_{bsp}_B exceeds 4 Mm^{-1} and BSF_G is below 0.14, indicating a strong dominance of larger particles, the interaction effect between these variables tends to plateau, suggesting a diminishing marginal impact on N_{CCN}.

Overall, at land sites, the contribution of σ_{bsp}_B, whether positive or negative, fluctuates with changes in BSF_G. In contrast, the contribution of σ_{bsp}_B is minimal when influenced by AE_BR (Fig. 5c, d). These findings suggest that the NEL model effectively captures the roles of aerosol number concentration and particle size, as reflected by σ_{bsp}_B and BSF_G, in influencing N_{CCN}.

In contrast, at the ocean sites (ASI and ENA), when σ_{sp}_B is below ~ 10 Mm^{-1}, higher AE_BR is associated with a decrease in N_{CCN} (Fig. 5g, h). This is likely because smaller particles (indicated by higher AE_BR) are less likely to activate as CCN, while a greater presence of larger particles (lower AE_BR) enhances CCN activation. When σ_{sp}_B exceeds 10 Mm^{-1}, the effect reverses, possibly due to severe pollution leading to a higher number concentration of larger particles, which typically exhibit greater scattering ability and higher activation potential. This effect is similar to that of land sites. Overall, N_{CCN} increases are strongly influenced by particle size when σ_{sp}_B is below 10 Mm^{-1}, whereas number concentration becomes more significant when σ_{sp}_B exceeds 10 Mm^{-1}. Additionally, the contribution of σ_{sp}_B is minimal when influenced by BSF_G (Fig. 5e, f).

At the MOS site (Fig. 6), the variation in σ_{sp} is significantly influenced by AE and BSF, especially under more polluted conditions (σ_{sp} > 18 Mm^{-1}). Under cleaner conditions (σ_{sp} < 18 Mm^{-1}), although the positive and

negative contributions are clearly distinguished, the variation is relatively gentle. These further highlight that aerosol number concentration is the dominant factor influencing N_{CCN} in polar regions.

These regional differences provide valuable insights into cloud microphysical processes and associated climate feedback. Over land, the strong dependence on BSF and σ_{bsp} indicates that aerosol shape, size and number concentration play a dominant role in regulating N_{CCN}, thereby influencing cloud albedo and lifetime. In contrast, in ocean regions, the greater importance of AE and σ_{sp} suggests that variability in particle size distribution and aerosol number concentration are the primary drivers of N_{CCN}, potentially altering cloud droplet formation and subsequent radiative properties. In polar regions, the aerosol number concentration has a greater influence on the prediction of N_{CCN}.

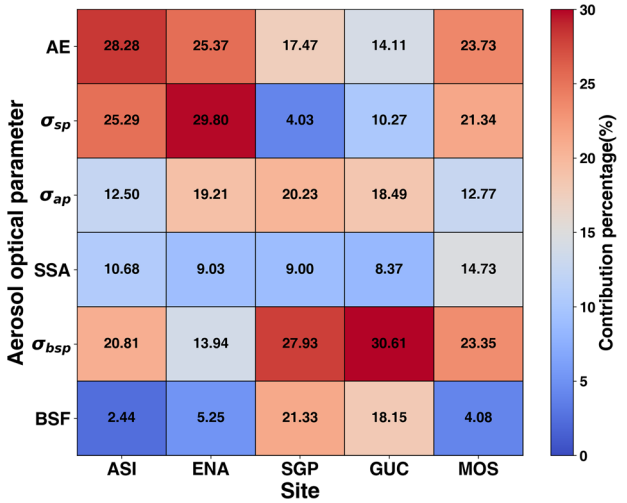

Fig. 4 | The mean relative contributions of each aerosol optical parameter at three wavelengths across five sites, with the color indicates contribution percentage. This figure illustrates the contribution of different aerosol optical parameters in various sites to the N_{CCN} prediction. Warm colors indicate a greater contribution, while cool colors indicate a smaller contribution.

Discussion

This study employed ARM observational data to apply the NEL model, developed using a combination of three machine learning methods and SHAP analysis, to predict N_{CCN} based on aerosol optical parameters. The model was tested at two land sites (SGP and GUC), two ocean sites (ENA and ASI) and one polar site (MOS), providing a comprehensive comparison of aerosol characteristics across diverse environments. The results demonstrate that the NEL model accurately predicts N_{CCN} throughout the sampling period, with R^2 values of 0.63, 0.92, 0.70, 0.65 and 0.83 for SGP, GUC, ASI, ENA, and MOS, respectively. These strong correlations highlight the model's capability to predict N_{CCN} under varying environmental conditions. Overall, σ_{sp}, σ_{bsp}, σ_{ap}, BSF, and AE show positive correlations with N_{CCN}. Although SSA has weaker associations with N_{CCN}, SSA indirectly influences N_{CCN} by affecting aerosol activation ability.

SHAP analysis identified the key aerosol optical parameters influencing N_{CCN}, revealing distinct differences between different environments. At land sites, σ_{bsp}_B emerged as the primary driver of N_{CCN} (20.75%), particularly at SGP (23.32%) and GUC (18.17%), where local sources such as biomass burning may elevate the significance of smaller backscattering particles. In contrast, at ocean sites, σ_{sp}_B (21.22%) was the dominant predictor in N_{CCN} prediction, reflecting the larger particle sizes commonly found over oceans. At the polar site (MOS), σ_{sp}_B was the primary driver of N_{CCN}, contributing 18.85%. All these underscore the importance of aerosol number concentration as a crucial factor for CCN formation across most environments.

The study also highlights key differences between different environments in the contributions of σ_{sp} and σ_{bsp}, modulated by AE and BSF. In both land and ocean regions, when the environment is relatively clean, the contribution to N_{CCN} is primarily driven by particle size. However, as pollution levels increase, the contribution of aerosol number concentration to N_{CCN} gradually becomes more significant. Notably, BSF is more sensitive at land sites, while AE has a greater impact at ocean sites. In polar regions,

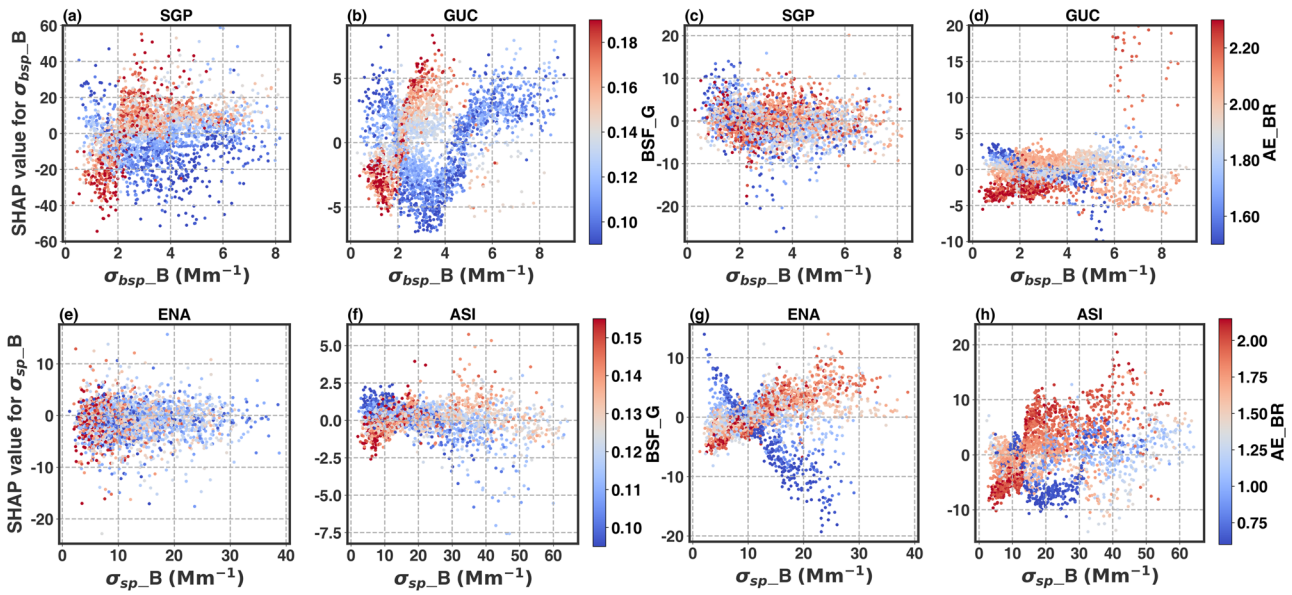

Fig. 5 | SHAP dependence plots for key aerosol optical parameters, with SGP and GUC corresponding to σ_{bsp}_B, and ASI and ENA corresponding to σ_{sp}_B. The *x*-axis represents the primary feature, while the *y*-axis represents the SHAP value of the primary feature. The color indicates the interaction feature. The whole represents the contribution of the main feature under the influence of the interaction feature. **a** SGP and **b** GUC show the contribution of σ_{bsp}_B under the influence of BSF_G. **c** SGP and **d** GUC display the contribution of σ_{bsp}_B under the influence of AE_BR. **e** ENA and **f** ASI illustrate the contribution of σ_{sp}_B under the influence of BSF_G. **g** ENA and **h** ASI depict the contribution of σ_{sp}_B under the influence of AE_BR. Plots **a** and **b**, **c** and **d**, **e** and **f**, and **g** and **h** share a common *y*-axis label and color bar, respectively.

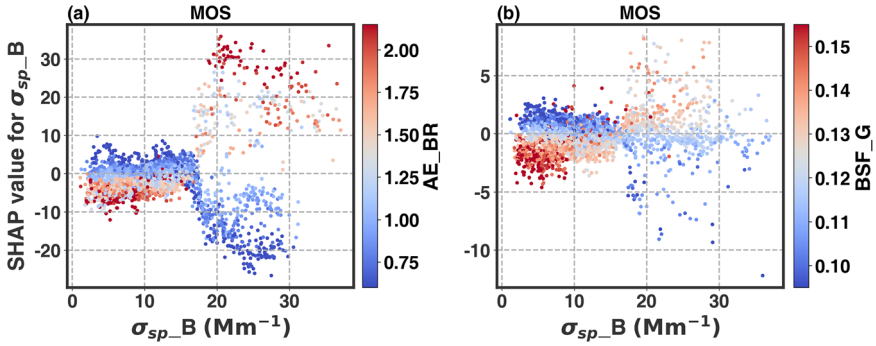

Fig. 6 | SHAP dependence plots for key aerosol optical parameters, with MOS corresponding to σ_{sp}_B. The *x*-axis represents the primary feature, while the *y*-axis represents the SHAP value of the primary feature. The color indicates the interaction feature. The whole represents the contribution of the main feature under the influence of the interaction feature. **a** show the contribution of σ_{sp}_B under the influence of AE_BR. **b** show the contribution of σ_{sp}_B under the influence of BSF_G.

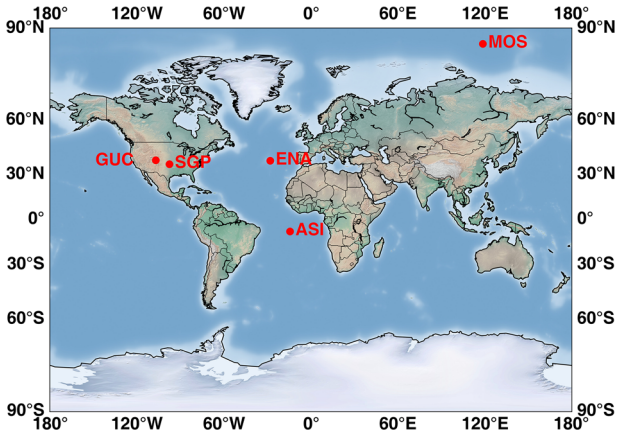

Fig. 7 | The geographical distribution of five sites. ENA (Eastern North Atlantic, 39°5'N, 28°1'W), ASI (Ascension Island, 7°58'S, 14°20'W), SGP (Southern Great Plain, 36°36'N, 97°29'W), GUC (Gunnison, CO, USA, 38°53'N, 106°56'W), MOS (Arctic Ocean; Mobile Facility, 86°37'8"N, 118°6'46"E). Land sites: SGP, GUC; ocean sites: ENA, ASI; polar site: MOS.

under polluted conditions, the contribution of σ_{sp} to N_{CCN} shows significant changes due to the influence of BSF and AE, further indicating that N_{CCN} in this region is mainly controlled by aerosol number concentration. These findings indicate that the NEL model has identified differences in CCN activation across different regions at varying pollution levels.

The findings of this study have several implications for both scientific understanding and practical applications. First, the ability of the NEL model, combined with SHAP analysis, to accurately predict N_{CCN} across diverse environments offers a significant advancement in aerosol-cloud interaction research. Direct N_{CCN} measurements are costly and logistically challenging, particularly over oceans and remote areas, making the development of a reliable prediction framework crucial. By using commonly measured aerosol optical properties, the NEL model provides a cost-effective and scalable alternative to direct measurements, facilitating broader research on cloud microphysics and climate modeling.

The study's identification of key aerosol optical parameters and their interactions in influencing N_{CCN} has significant implications for improving climate models. Aerosols are crucial in modulating cloud properties, which affect radiation balance and precipitation patterns. Accurate N_{CCN} prediction is essential for understanding aerosol-cloud-climate feedback mechanisms. The varying sensitivities of aerosol parameters between land and ocean environments, as revealed in this study, emphasize the importance of considering regional and environmental contexts in cloud and climate modeling. Incorporating these insights into global climate models could enhance the accuracy of cloud formation predictions and their effects on climate systems. Furthermore, the study highlights the impact of specific aerosol sources, such as biomass burning and wildfire events, on local CCN concentrations, which has implications for air quality and regional climate forecasting, particularly in wildfire-prone or heavily polluted areas. Understanding how these events influence aerosol properties and cloud formation can aid in developing mitigation strategies and improving early warning systems for climate-related impacts.

While this study provides valuable insights, it is primarily based on ground-based observations, which, despite their comprehensiveness, may not fully capture the vertical and spatial variability of aerosols. Future research should focus on integrating satellite data, aircraft observations, and multi-dimensional simulations to improve the accuracy of N_{CCN} retrievals across different spatial and temporal scales. Additionally, expanding the NEL model to encompass more diverse environments and varying supersaturation levels would extend its applicability. Incorporating various climatic and aerosol regimes would allow for further validation and refinement, advancing the model toward becoming a universal tool for global N_{CCN} prediction. Aerosol activation schemes, which predict the number and mass of activated particles crucial for cloud formation and climate studies, should also focus on mass activation efficiency in future research to improve estimates of cloud droplet formation. Moreover, this research offers practical recommendations for enhancing climate models. Current models often rely on simplified aerosol activation schemes that overlook environmental variability. We suggest incorporating BSF and σ_{bsp} into parameterizations for land regions, AE and σ_{sp} for ocean regions and σ_{sp} for polar regions to improve N_{CCN} predictions. For example, models like the Community Earth System Model (CESM) could integrate environment-specific weightings of these properties or adopt the NEL model to enhance simulations of cloud formation and aerosol indirect effects, reducing uncertainties in climate predictions.

Methods

Data sources and preprocessing

The U.S. Department of Energy (DOE) is responsible for deploying the Atmospheric Radiation Measurement (ARM) Climate Research Facility (at both fixed and mobile sites). In recent years, ARM has measured cloud condensation nuclei (CCN) and numerous related variables. This study utilizes observational data from five ARM sites (Fig. 7), each characterized by distinct aerosol types: Eastern North Atlantic (ENA, a long-term fixed site with marine aerosols, 39°5'N, 28°1'W), Ascension Island, South Atlantic Ocean (ASI, a mobile site with marine aerosols and long-range transported biomass-burning aerosols from southern Africa, 7°58'S, 14°20'W), Southern Great Plains (SGP, a permanent site with typical rural continental aerosols over farmland, 36°36'N, 97°29'W), Gunnison, CO, USA (GUC, a mobile site with mountain forest aerosols, 38°53'N, 106°56'W), and Arctic Ocean; Mobile Facility (MOSAiC) (MOS, a mobile site with polar aerosols, 86°37'8"N, 118°6'46"E)[47,52,53,57,59,62].

This study collected N_{CCN} data observed by the Cloud Condensation Nuclei Counter (CCNc) and aerosol optical data measured by various instruments[63]. The data spans different periods for each site: ENA data from June 2021 to June 2023, SGP data from April 2017 to January 2021, ASI data from May 2016 to October 2017, GUC data from September 2021 to

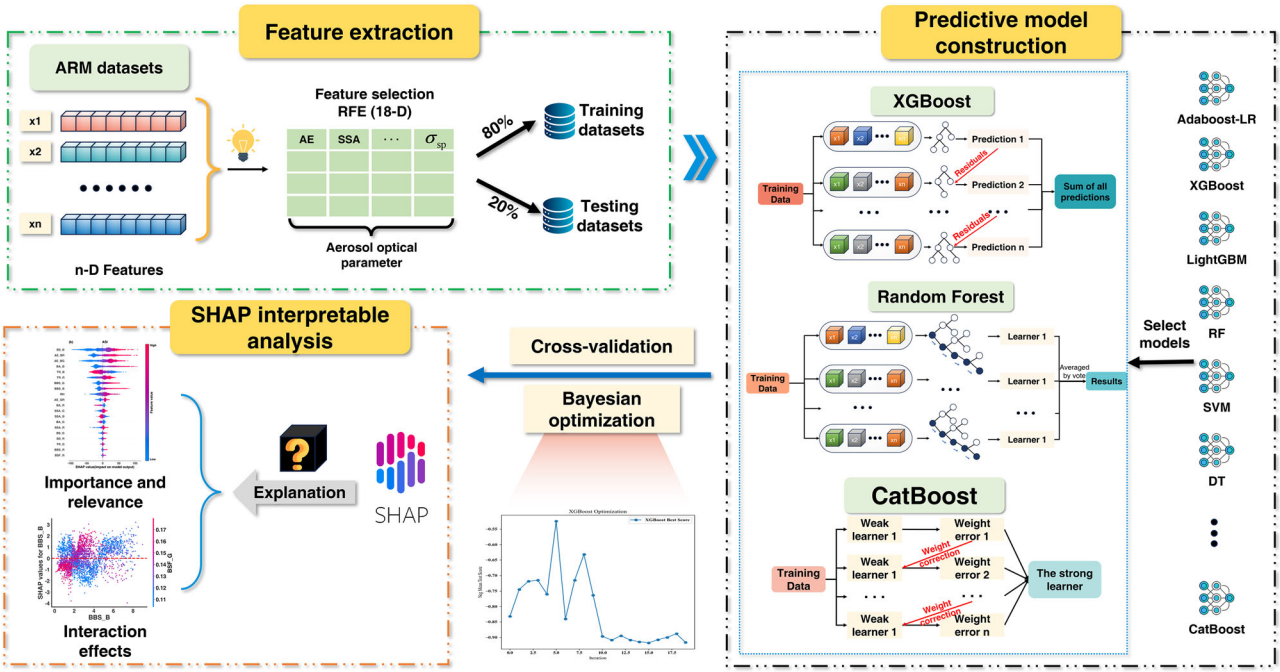

Fig. 8 | The framework of N_{CCN} Ensemble Learning (NEL) model. The figure shows the workflow of the NEL model, including data collection, preprocessing, modeling, and SHAP analysis.

October 2021 and MOS data from October 2019 to October 2020. The total number of data points for each site is 519,375 for SGP, 26,741 for GUC, 98,299 for ENA, 125,806 for ASI, and 55,163 for MOS. The instrumentation used across all sites was consistent. Detailed information about the data and instruments can be found at https://adc.arm.gov/discovery.

After aligning data from multiple instruments based on observation times, the mean and standard deviation are computed for each site. To ensure the accuracy of the results, quality control procedures are implemented to screen all data. Any data point exceeding three standard deviations from the mean is considered an outlier and removed, along with any missing values. Since the majority of N_{CCN} measurements across the sites are obtained at a supersaturation (SS) level of 0.4%, and SS = 0.4% is more representative of convective clouds[21], only N_{CCN} data at SS = 0.4% are retained for subsequent analysis, with data at other supersaturation levels excluded. Variable names, abbreviations, data ranges, and means for all sites (SGP, ENA, ASI, GUC, and MOS) are provided in Supplementary Tables 1–5.

Model framework

The framework of the N_{CCN} ensemble learning (NEL) model used for predicting N_{CCN} is illustrated in Fig. 8. To prevent overfitting from the inclusion of excessive variables, this study employs Recursive Feature Elimination (RFE) combined with manual selection for dimensionality reduction (Supplementary Text 2). The feature selection process primarily focuses on the impact of each feature on model accuracy, selecting the most relevant features from the initial dataset[64]. Ultimately, six aerosol optical parameters across three wavelengths are chosen as feature variables, with N_{CCN} as the target for prediction.

The temporal resolution of the data used for training the NEL model is 1 minute. This high temporal resolution ensures that the data captures detailed variations in aerosol optical properties and CCN concentrations over short time intervals, which is important for accurately predicting N_{CCN}. Such fine-grained data ensures that rapid changes in atmospheric conditions are reflected in the model, contributing to more precise predictions. Given the large dataset, the data is split into training and testing sets in an 8:2 ratio, with both sets shuffled to prevent overfitting and mitigate the effects of time series data. Five-fold cross-validation and Bayesian optimization are employed to adaptively adjust model hyperparameters and initial values, maximizing the coefficient of determination (R^2) to enhance model performance. The optimization process for the five sites is illustrated in Supplementary Figs. 14–18, with specific model parameters listed in Supplementary Tables 6–10. A consistent random seed of 2024 is used throughout the process. Final predictions are obtained by averaging the outputs from three models (XGBoost, CatBoost and RF), forming the NEL model.

The NEL model trains individual models for each site. By training separate models for different environments, each model is optimized to account for the unique atmospheric conditions of that site. This approach ensures that the models can be directly applied to similar environments, enhancing their applicability and accuracy in predicting N_{CCN} across a wide range of atmospheric backgrounds. The model's performance is evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), the coefficient of determination (R^2), Mean Absolute Percentage Error (MAPE), and Relative Euclidean Distance (RED).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(N_{CCN_{True_i}} - N_{CCN_{Predict_i}}\right)^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|N_{CCN_{True_i}} - N_{CCN_{Predict_i}}\right| \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(N_{CCN_{True_i}} - N_{CCN_{Predict_i}}\right)^2}{\sum_{i=1}^{n}\left(N_{CCN_{True_i}} - \overline{N_{CCN_{Predict}}}\right)^2} \quad (3)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{N_{CCN_{True_i}} - N_{CCN_{Predict_i}}}{N_{CCN_{True_i}}}\right| \quad (4)$$

$$RED = \sqrt{\left(\frac{\overline{N_{CCN_{Predict}}} - \overline{N_{CCN_{True}}}}{\overline{N_{CCN_{True}}}}\right)^2 + \left(\frac{\sigma_{Predict} - \sigma_{True}}{\sigma_{True}}\right) + (1-R)^2} \quad (5)$$

Here, n represents the number of input samples. $N_{CCN_{True_i}}$ denotes the measured N_{CCN} value for the ith sample. while $N_{CCN_{Predict_i}}$ represents the predicted N_{CCN} value for the ith sample. $\overline{N_{CCN\,Predict}}$ refers to the mean N_{CCN} value predicted by the model. $\overline{N_{CCN\,True}}$ refers to the mean N_{CCN} value measured. $\sigma_{Predict}$ and σ_{True} indicate the standard deviations of the predicted and measured N_{CCN} values, respectively. R represents the correlation coefficient.

After completing the basic training, the SHAP algorithm is employed to conduct an interpretability analysis on the predicted N_{CCN} values. SHAP operates by utilizing Shapley values to quantitatively evaluate the contribution of each feature within a machine learning model[65]. SHAP evaluates the contribution of each feature by measuring how it changes the model's prediction across all possible combinations of features. In the absence of any features (e.g., aerosol optical parameters), the NEL model outputs a baseline prediction, typically the average N_{CCN} value across the dataset. When a single feature, such as σ_{sp}, is added, the model's prediction may shift. This shift represents the marginal contribution of σ_{sp}. SHAP quantifies this contribution by computing the prediction difference introduced by σ_{sp} across all possible feature subsets in which it is included. By averaging these marginal contributions, SHAP assigns an importance value to each feature, providing a consistent and interpretable measure of how each aerosol parameter influences N_{CCN} predictions both individually and in combination with others. Further details on the SHAP algorithm are provided in Supplementary Text 3.

SHAP can be influenced by multicollinearity, where strongly correlated features may distort the attribution of importance. To address this, feature selection strategies, such as RFE and artificial selection of aerosol optical parameters, were employed to reduce redundancy and ensure that the selected features contribute independently and meaningfully.

Data availability

The U.S. Department of Energy (DOE) initiated the Atmospheric Radiation Measurement (ARM) program at the end of the 20th century. Over the past two decades, the program has conducted continuous observational experiments through a network of fixed and mobile sites worldwide. The ARM program performs long-term comprehensive observations of meteorological conditions, radiation, ground-based aerosol optical properties, and cloud condensation nuclei. The data collected are freely available online to researchers globally, providing a solid foundation for studying the spatiotemporal distribution and long-term changes in aerosol properties[66]. ARM data can be downloaded from the ARM website (https://adc.arm.gov/discovery).

Code availability

The NEL model and Python codes used for performing analyses can be accessed here: https://github.com/dtnan/NEL.

Received: 5 November 2024; Accepted: 27 July 2025;
Published online: 14 August 2025

References

1. Charlson, R. J. et al. Climate forcing by anthropogenic aerosols. *Science* **255**, 423–430 (1992).
2. Li, Z. et al. Long-term impacts of aerosols on the vertical development of clouds and precipitation. *Nat. Geosci.* **4**, 888–894 (2011).
3. Tao, W. K., Chen, J. P., Li, Z., Wang, C. & Zhang, C. Impact of aerosols on convective clouds and precipitation. *Rev. Geophys.* **50**, RG2001 (2012).
4. Rosenfeld, D. et al. Flood or drought: how do aerosols affect precipitation?. *Science* **321**, 1309–1313 (2008).
5. Malavelle, F. F. et al. Strong constraints on aerosol–cloud interactions from volcanic eruptions. *Nature* **546**, 485–491 (2017).
6. IPCC. in *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* (eds. Masson-Delmotte, V. et al.) 2061–2086 (Cambridge University Press, 2021).
7. Köhler, H. The nucleus in and the growth of hygroscopic droplets. *Trans. Faraday Soc.* **32**, 1152–1161 (1936).
8. Dusek, U. et al. Size matters more than chemistry for cloud-nucleating ability of aerosol particles. *Science* **312**, 1375–1378 (2006).
9. Farmer, D. K., Cappa, C. D. & Kreidenweis, S. M. Atmospheric processes and their controlling influence on cloud condensation nuclei activity. *Chem. Rev.* **115**, 4199–4217 (2015).
10. Ren, J. et al. Using different assumptions of aerosol mixing state and chemical composition to predict CCN concentrations based on field measurements in urban Beijing. *Atmos. Chem. Phys.* **18**, 6907–6921 (2018).
11. Wang, Y. et al. Characterization of aerosol hygroscopicity, mixing state, and CCN activity at a suburban site in the central North China Plain. *Atmos. Chem. Phys.* **18**, 11739–11752 (2018).
12. Zhang, F. et al. Uncertainty in predicting CCN activity of aged and primary aerosols. *J. Geophys. Res. Atmos.* **122**, 11723–11736 (2017).
13. Jurányi, Z. et al. A 17 month climatology of the cloud condensation nuclei number concentration at the high alpine site Jungfraujoch. *J. Geophys. Res.* **116**, D10204 (2011).
14. Paramonov, M. et al. A synthesis of cloud condensation nuclei counter (CCNC) measurements within the EUCAARI network. *Atmos. Chem. Phys.* **15**, 12211–12229 (2015).
15. Ghan, S. J. et al. Use of in situ cloud condensation nuclei, extinction, and aerosol size distribution measurements to test a method for retrieving cloud condensation nuclei profiles from surface measurements. *J. Geophys. Res. Atmos.* **111**, D05S10 (2006).
16. Kapustin, V. N. et al. On the determination of a cloud condensation nuclei from satellite: Challenges and possibilities. *J. Geophys. Res. Atmos.* **111**, D04202 (2006).
17. Jefferson, A. Empirical estimates of CCN from aerosol optical properties at four remote sites. *Atmos. Chem. Phys.* **10**, 6855–6861 (2010).
18. Shen, Y. et al. Estimating cloud condensation nuclei number concentrations using aerosol optical properties: role of particle number size distribution and parameterization. *Atmos. Chem. Phys.* **19**, 15483–15502 (2019).
19. Shinozuka, Y. et al. The relationship between cloud condensation nuclei (CCN) concentration and light extinction of dried particles: indications of underlying aerosol processes and implications for satellite-based CCN estimates. *Atmos. Chem. Phys.* **15**, 7585–7604 (2015).
20. Andreae, M. O. Correlation between cloud condensation nuclei concentration and aerosol optical thickness in remote and polluted regions. *Atmos. Chem. Phys.* **9**, 543–556 (2009).
21. Liu, J. & Li, Z. Estimation of cloud condensation nuclei concentration from aerosol optical quantities: influential factors and uncertainties. *Atmos. Chem. Phys.* **14**, 471–483 (2014).
22. Shinozuka, Y. et al. Aerosol optical properties relevant to regional remote sensing of CCN activity and links to their organic mass fraction: airborne observations over Central Mexico and the US West Coast during MILAGRO/INTEX-B. *Atmos. Chem. Phys.* **9**, 6727–6742 (2009).
23. Tao, J. et al. A new method for calculating number concentrations of cloud condensation nuclei based on measurements of a three-wavelength humidified nephelometer system. *Atmos. Meas. Tech.* **11**, 895–906 (2018).

24. Nair, A. A. & Yu, F. Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements. *Atmos. Chem. Phys.* **20**, 12853–12869 (2020).
25. Yang, Y. et al. Revolutionizing clear-sky humidity profile retrieval with multi-angle aware networks for ground-based microwave radiometers. *J. Remote Sens*. **5**, 0736 (2025).
26. Xin, J. et al. AI model to improve the mountain boundary layer height of ERA5. *Atmos. Res.* **304**, 107352 (2024).
27. Yu, S. & Ma, J. Deep learning for geophysics: Current and future trends. *Rev. Geophys.* **59**, e2021RG000742 (2021).
28. Redemann, J. & Gao, L. A machine learning paradigm for necessary observations to reduce uncertainties in aerosol climate forcing. *Nat. Commun.* **15**, 8343 (2024).
29. Liang, M. et al. Prediction of CCN spectra parameters in the North China Plain using a random forest model. *Atmos. Environ.* **289**, 119323 (2022).
30. Nair, A. A. et al. Machine learning uncovers aerosol size information from chemistry and meteorology to quantify potential cloud-forming particles. *Geophys. Res. Lett.* **48**, e2021GL094133 (2021).
31. Zhang, L. et al. Explainable ensemble machine learning revealing the effect of meteorology and sources on ozone formation in megacity Hangzhou, China. *Sci. Total Environ.* **922**, 171295 (2024).
32. Tao, C. et al. Diagnosing ozone–NOx–VOC–aerosol sensitivity and uncovering causes of urban–nonurban discrepancies in Shandong, China, using transformer-based estimations. *Atmos. Chem. Phys.* **24**, 4177–4192 (2024).
33. Peng, K. et al. Machine learning model to accurately estimate the planetary boundary layer height of Beijing urban area with ERA5 data. *Atmos. Res.* **293**, 106925 (2023).
34. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794 (2016).
35. Hancock, J. T. & Khoshgoftaar, T. M. CatBoost for big data: an interdisciplinary review. *J. Big Data* **7**, 94 (2020).
36. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
37. Wang, L. et al. Predicting ozone formation in petrochemical industrialized Lanzhou city by interpretable ensemble machine learning. *Environ. Pollut.* **318**, 120798 (2023).
38. Requia, W. J. et al. An ensemble learning approach for estimating high spatiotemporal resolution of ground-level Ozone in the contiguous United States. *Environ. Sci. Technol.* **54**, 11037–11047 (2020).
39. Shan, Y., Liu, Y. & Zhou, X. Comparative evaluation of the ability of the MYNN-EDMF PBL scheme in WRF model to reproduce near surface wind speed over different topographical types. *J. Geophys. Res. Atmos.* **130**, e2023JD040620 (2025).
40. Elmore, K. L. & Richman, M. B. Euclidean distance as a similarity metric for principal component analysis. *Mon. Weather Rev.* **129**, 540–549 (2001).
41. Fierz-Schmidhauser, R. et al. Light scattering enhancement factors in the marine boundary layer (Mace Head, Ireland). *J. Geophys. Res. Atmos.* **115**, D20204 (2010).
42. Song, X. et al. The impacts of dust storms with different transport pathways on aerosol chemical compositions and optical hygroscopicity of fine particles in the Yangtze River Delta. *J. Geophys. Res. Atmos.* **128**, e2023JD039679 (2023).
43. Wang, Y. et al. The role of relative humidity in estimating cloud condensation nuclei number concentration through aerosol optical data: mechanisms and parameterization strategies. *Geophys. Res. Lett.* **52**, e2024GL112734 (2025).
44. Shinozuka, Y. Relations between cloud condensation nuclei and aerosol optical properties relevant to remote sensing. *Atmos. Environ* **267**, 118748 (2008).
45. Zhang, R. et al. Vertical profiles of cloud condensation nuclei number concentration and its empirical estimate from aerosol optical properties over the North China Plain. *Atmos. Chem. Phys.* **22**, 14879–14891 (2022).
46. Zheng, G. et al. Marine boundary layer aerosol in the eastern North Atlantic: seasonal variations and key controlling processes. *Atmos. Chem. Phys.* **18**, 17615–17635 (2018).
47. Ghate, V. P. et al. Drivers of cloud condensation nuclei in the Eastern North Atlantic as observed at the ARM site. *J. Geophys. Res. Atmos.* **128**, e2023JD038636 (2023).
48. Charlson, R. J., Lovelock, J. E., Andreae, M. O. & Warren, S. G. Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* **326**, 655–661 (1987).
49. Frossard, A. A. et al. Sources and composition of submicron organic mass in marine aerosol particles. *J. Geophys. Res.: Atmos.* **119**, 12977–13003 (2014).
50. Novakov, T. & Penner, J. E. Large contribution of organic aerosols to cloud-condensation-nuclei concentrations. *Nature* **365**, 823–826 (1993).
51. Zheng, G. et al. New particle formation in the remote marine boundary layer. *Nat. Commun.* **12**, 527 (2021).
52. Heutte, B. et al. Measurements of aerosol microphysical and chemical properties in the central Arctic atmosphere during MOSAiC. *Sci. Data* **10**, 690 (2023).
53. Marinescu, P. J., Levin, E. J. T., Collins, D., Kreidenweis, S. M. & van den Heever, S. C. Quantifying aerosol size distributions and their temporal variability in the Southern Great Plains, USA. *Atmos. Chem. Phys.* **19**, 11985–12006 (2019).
54. Rejano, F. et al. Activation properties of aerosol particles as cloud condensation nuclei at urban and high-altitude remote sites in southern Europe. *Sci. Total Environ.* **762**, 143100 (2021).
55. Collaud Coen, M. et al. Long-term trend analysis of aerosol variables at the high-alpine site Jungfraujoch. *J. Geophys. Res. Atmos.* **112**, D13213 (2007).
56. Willis, M. D., Leaitch, W. R. & Abbatt, J. P. D. Processes controlling the composition and abundance of arctic aerosol. *Rev. Geophys.* **56**, 621–671 (2018).
57. de Graaf, M. et al. Aerosol first indirect effect of African smoke at the cloud base of marine cumulus clouds over Ascension Island, southern Atlantic Ocean. *Atmos. Chem. Phys.* **23**, 5373–5391 (2023).
58. Zuidema, P. et al. The Ascension Island boundary layer in the remote southeast Atlantic is often smoky. *Geophys. Res. Lett.* **45**, 4456–4465 (2018).
59. Feldman, D. R. et al. The surface atmosphere integrated field laboratory (SAIL) campaign. *Bull. Am. Meteorol. Soc.* **104**, E2192–E2222 (2023).
60. Zheng, G. et al. Long-range transported North American wildfire aerosols observed in marine boundary layer of eastern North Atlantic. *Environ. Int.* **139**, 105680 (2020).
61. Chipade, R. A. & Pandya, M. R. Theoretical derivation of aerosol lidar ratio using Mie theory for CALIOP-CALIPSO and OPAC aerosol models. *Atmos. Meas. Tech.* **16**, 5443–5459 (2023).
62. Logan, T. et al. Assessing radiative impacts of African smoke aerosols over the southeastern Atlantic Ocean. *Earth Space Sci.* **11**, e2023EA003138 (2024).
63. McComiskey, A. & Ferrare, R. A. Aerosol physical and optical properties and processes in the ARM program. *Meteorol. Monogr.* **57**, 21.21–21.17 (2016).
64. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
65. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
66. Mather, J. H. & Voyles, J. W. The arm climate research facility: a review of structure and capabilities. *Bull. Am. Meteorol. Soc.* **94**, 377–392 (2013).

Acknowledgements
This research has been supported by the Key Program of the National Natural Science Foundation of China (Grant No. 42030606), National Key Laboratory of Science and Technology on Near-surface Detection (Grant No. 6142414221302), and the Natural Science Foundation of Jiangsu Province (Grant No. BK20220226).

Author contributions
N.W. and Y.W. conceived the research, performed the analysis, and wrote the manuscript. C.L., B.Z., X.Y., Y.S., J.X., J.Z., and Z.S. assisted in the interpretation of the results and revision of 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/s41612-025-01181-y.

Correspondence and requests for materials should be addressed to Yuying Wang.