JGR Atmospheres

RESEARCH ARTICLE
10.1029/2024JD043116

Employing Machine Learning for New Particle Formation Identification and Mechanistic Analysis: Insights From a Six-Year Observational Study in the Southern Great Plains

Weixing Hao[1], Manisha Mehra[1], Gaurav Budhwani[1], T. C. Chakraborty[2], Fan Mei[2], and Yang Wang[1]

[1]Department of Chemical, Environmental and Materials Engineering, University of Miami, Coral Gables, FL, USA, [2]Pacific Northwest National Laboratory, Richland, WA, USA

Key Points:

- Machine learning identifies new particle formation (NPF) events with 90%–95% accuracy
- Key environmental factors associated with NPF: solar radiation, relative humidity, and temperature
- NPF frequency peaks in winter and spring, lowest in summer

Supporting Information:

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

Correspondence to:

Y. Wang and F. Mei,
yangwang@miami.edu;
fan.mei@pnnl.gov

Citation:

Hao, W., Mehra, M., Budhwani, G., Chakraborty, T. C., Mei, F., & Wang, Y. (2026). Employing machine learning for new particle formation identification and mechanistic analysis: Insights from a six-year observational study in the Southern Great Plains. *Journal of Geophysical Research: Atmospheres*, *131*, e2024JD043116. https://doi.org/10.1029/2024JD043116

Received 11 DEC 2024
Accepted 24 DEC 2025

Abstract We present a supervised machine learning (ML) framework to automatically identify new particle formation (NPF) events and analyze key atmospheric factors associated with their occurrence and growth. We applied ML to detect NPF events using start time and particle concentrations across size ranges, while identifying atmospheric variables including ambient temperature, relative humidity, solar radiation intensity (SRI), wind speed, wind direction, boundary layer height, total organics, sulfate, nitrate, total surface area concentration, sulfur dioxide, and turbulent kinetic energy (TKE). We analyzed a 6-year data set from the Atmospheric Radiation Measurement at the Southern Great Plains (SGP) site in Oklahoma, USA. Using long-term ground-based measurements, we identified NPF events and applied Random Forest Classifiers, which achieved 90%–95% prediction accuracy. Feature importance analysis highlighted SRI, relative humidity, and ambient temperature as the most influential variables, contributing normalized importances of 28%, 17%, and 10%. Partial Dependence Plots (PDPs) indicated that higher SRI and lower relative humidity were critical in promoting NPF formation at SGP. Seasonally, NPF events were more frequent in winter (42.1%) and spring (35.5%), and least in summer (4.0%). Particle growth rates also exhibited a seasonal variation, with the lowest in winter (below 2 nm hr^{-1}) and highest in late spring and early summer (exceeding 5 nm hr^{-1}). Temperature, turbulent kinetic energy, and aerosol properties were the primary factors of growth rate variability. This study advances predictive modeling of NPF, offers insights for future campaign deployments, and demonstrates the effectiveness of ML in understanding the formation and growth of atmospheric aerosols.

Plain Language Summary We used machine learning to automatically detect when and where NPF occurs. We also identified the environmental factors like temperature, humidity, and solar radiation that are associated with these processes. Our approach not only improves the ability to detect NPF events but also reveals how seasonal and weather-related factors influence particle growth. By revealing the conditions that influence particle formation, our study helps improve climate models and can guide future environmental research.

1. Introduction

Atmospheric new particle formation (NPF) is a critical process in generating aerosol particles that influence the Earth's climate (Lee et al., 2019; Leng et al., 2014; Oliveira et al., 2025; Xiao et al., 2023; Zhu et al., 2019). This process initiates with the formation of molecular clusters, which grow to several nanometers in diameter before further growth through condensation and coagulation into larger sizes. The particles generated from NPF events, particularly those within the sub-100 nm size range, substantially increase atmospheric particle concentrations (Kerminen et al., 2018; Kulmala et al., 2014; Ma & Birmili, 2015). Moreover, after growing in sizes, these NPF-generated particles (typically, newly formed are those $D_p < 10$ nm) can contribute to cloud condensation nuclei concentration, thereby affecting cloud microphysics and potentially inducing a net cooling effect on the climate (Fanourgakis et al., 2019; Joutsensaari et al., 2018; Kerminen et al., 2018).

Understanding NPF requires accurate identification of its occurrence, which is traditionally achieved by directly observing and analyzing aerosol particle size distributions. While expert-based manual visual analysis has proven to be a reliable method for classifying days as NPF events or non-events, this approach is both labor-intensive and subjective, particularly when applied to long-term data series (Dal Maso et al., 2005; Joutsensaari et al., 2018). These limitations underscore the need for automated, consistent classification systems. Recent advancements

have seen the adoption of image-based methods utilizing computer vision techniques, neural networks, and deep learning models to classify NPF events (Joutsensaari et al., 2018; Su et al., 2022a, 2022b; Zaidan et al., 2017). However, the success of these models hinges on extensive data sets and significant computational resources, with a high risk of overfitting when the number of model parameters is excessive, thereby limiting their generalizability.

Despite advancements in NPF detection and classification, the underlying mechanisms governing NPF events remain elusive due to their complex interplay and high nonlinearity with dependences on multiple environmental variables. These include meteorological conditions (Bousiotis, Brean, et al., 2021; Sorribas et al., 2015), atmospheric chemical composition (Bzdek et al., 2011; Du et al., 2022), and pre-existing aerosol loading (Cai et al., 2017; McMurry & Friedlander, 1967; Kuang et al., 2010). For example, while some studies suggest that low ambient relative humidity (RH) favors NPF (Dada et al., 2017; Li et al., 2019), NPF events have also been observed in high RH environments, likely because of the participation of water vapor in the early stages of NPF (Bousiotis, Pope, et al., 2021; O'dowd et al., 1998). Similarly, high ambient temperatures, often associated with intense solar radiation intensity (SRI), can enhance photochemical reactions and nucleation processes (Boy & Kulmala, 2002; Kürten et al., 2016), yet may also destabilize molecular clusters and inhibit NPF (Kürten et al., 2018). The role of mixed atmospheric chemical species, such as SO_2, NH_3, and oxidized volatile organic compounds (VOCs), further complicates the NPF process, with their influence varying based on nucleation mechanisms and nucleating precursor concentrations (Kürten et al., 2016; Laaksonen et al., 2008; Zhang et al., 2021).

These complexities indicate that our current understanding of NPF does not fully account for the diverse chemical and environmental conditions under which NPF events occur. Thus, there is a need to better understand the relationship between different environmental variables and NPF occurrences using advanced data processing methods. Furthermore, NPF events are influenced by distinct temporal patterns driven by daily and seasonal cycles, which are closely linked to fluctuations in environmental conditions, anthropogenic activities, and atmospheric chemistry (Bousiotis, Pope, et al., 2021; Mikkonen et al., 2020). Understanding how these temporal variations affect NPF is crucial, but remains unclear. Given the complexity and number of influencing variables, an automated, intelligent learning method is necessary to identify and quantify the significance of these factors in NPF processes.

Machine Learning (ML) has emerged as a powerful tool for analyzing the nonlinear features and complex relationships inherent in atmospheric processes (Chakraborty & Lee, 2021; Joutsensaari et al., 2018; Nair & Yu, 2020; Su, Joutsensaari, et al., 2022; Zhong et al., 2021). Among ML techniques, decision tree-based methods, particularly Random Forest (RF), have shown exceptional capability in handling large data sets with numerous input variables (Breiman, 2001). RF's ensemble approach, which aggregates predictions from multiple decision trees, makes it particularly well-suited for studying atmospheric phenomena like NPF, where the interactions between multiple environmental factors are intricate and non-linear. Furthermore, Partial Dependence Plots (PDPs) have proven useful as diagnostic tools for exploring the relationships between input variables and model outputs (Goldstein et al., 2015; Venter et al., 2021). Although RF and PDP methods have been successfully applied in various domains in the earth and environmental sciences, including the study of environmental factors affecting atmospheric pollutant concentrations (e.g., PM_{10} and SO_2), and the concentrations of primary and oxygenated organic aerosols at urban and rural sites in Hong Kong (Grange & Carslaw, 2019; Grange et al., 2018; Qin et al., 2022), their application in evaluating NPF classification and the influence of environmental variables remains unexplored.

In this study, we employ supervised ML techniques to predict the occurrence of NPF events and evaluate the impact of various environmental variables influencing the occurrence using large data sets from the Southern Great Plains (SGP) site. By integrating RF and PDP algorithms, we analyze raw particle size distribution data along with interpretable environmental inputs, eliminating the need for manual detection or complex neural networks. Importantly, we demonstrate that ML-based predictions can efficiently classify NPF events from multi-year aerosol data sets, inform when and under what conditions NPF is most likely to occur, offering a powerful tool for designing targeted field campaigns. While ML does not provide direct mechanistic insights, it offers a scalable and data-driven framework for anticipating NPF episodes, guiding more effective sampling strategies and resource allocation in atmospheric research.

2. Measurement and Methods

2.1. Measurement Site and Data Collection

This study utilized an extensive data set collected over 6 yrs, from 1 January 2018, to 31 December 2023, at the U. S. Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) SGP atmospheric observatory, located in Lamont, north-central Oklahoma, USA. Strategically situated at 36°61′N, 97°49′W, and 314 m above sea level, the SGP site is one of the world's largest and most comprehensive climate research facilities (Ackerman & Stokes, 2003; Mather & Voyles, 2013). Its relative remoteness from significant anthropogenic pollution sources makes it an ideal location for collecting diverse atmospheric data, essential for studying natural atmospheric processes such as NPF over mid-latitude continents. The site's exposure to various aerosol types, sources, and transport pathways makes it optimal for studying aerosol dynamics and characterizing aerosol properties at a representative rural, continental location in North America (Marinescu et al., 2019).

Aerosol particle size distribution data were measured by two scanning mobility particle sizers (SMPS, TSI Inc., Model 3,936), one equipped with a "long" differential mobility analyzer (DMA, TSI Inc., Model 3081A), and another equipped with a "nano" DMA (TSI Inc., Model 3085A). For ease of reference, we will designate these two systems as the "long-SMPS" and "nano-SMPS" systems, respectively. Measurements were recorded every 5 min, yielding 288 data points daily, except during occasional system malfunctions. Note that the long-SMPS and nano-SMPS overlap in certain size ranges. We used the size distribution measured by the nano-SMPS in the overlapping size range whenever nano-SMPS data were available, as it provides higher resolution and improved accuracy for ultrafine particles (<20 nm), which are essential for capturing nucleation-mode dynamics. This approach is consistent with standard aerosol measurement practice, given that the nano-SMPS is specifically optimized for newly formed particle detection in this range (Kulkarni et al., 2011). Meanwhile, to ensure broader applicability, we incorporated the observational long-SMPS data into our analysis.

In addition to aerosol particle size distribution data, the data set encompassed a broad range of atmospheric variables, including meteorological parameters and chemical variables such as ambient temperature (T), relative humidity (RH), SRI, wind speed (Wsp), wind direction (Wdir), boundary layer height (BLH), total organics (Organics), sulfate (Sulfate), nitrate (Nitrate), total surface area concentration (S_{tot}), sulfur dioxide (SO_2), and turbulent kinetic energy (TKE).

The data acquisition process began by retrieving the necessary files from the ARM data repository in netCDF format. These files were then converted to MATLAB (.mat) format using custom scripts to facilitate data processing and analysis. Subsequently, all relevant matrices were exported to Comma-Separated Values (CSV) files for further use in the ML process. Data interpolation was performed to align the various measured parameters with the aerosol number concentration at 288 timestamps per day, ensuring a temporally consistent data set.

2.2. NPF Events Identification

2.2.1. Manual Identification

The most commonly used method for classifying NPF is manual visual analysis labeling, where researchers visually analyze aerosol size-distribution plots to assess time series of size distributions and nucleation-mode particle evolution, identifying periods of NPF and growth (Kulmala et al., 2012). Using this method, we applied it to the SGP data by visualizing particle size distributions over time (0–24 hr UTC) on the x-axis, particle diameter on the y-axis, and particle concentration represented by color gradients, as example shown in Section 3.4.

To assist this visual method and determine NPF start and end times, we applied a semi-quantitative criterion based on particle size distribution data. Specifically, an NPF event was defined to start when the arithmetic mean particle size fell below 9 nm and the smoothed percentage of particles below 15 nm showed a sustained positive growth trend. The event ended when the arithmetic mean particle size began to decrease, verified by a negative rate of change over the subsequent 10 timepoints. To ensure robustness, only events lasting at least 24 consecutive timepoints (2 hr at 5-min resolution) were considered. This approach combines objective size threshold criteria with visual inspection of nucleation-mode growth patterns, consistent with the characteristic "banana shape" used in manual NPF identification. Each day was then classified as either an NPF event (binary label of 1) or a non-event (binary label of 0). To enhance classification accuracy, days labeled as "undefined"—those that did not

Table 1
Summary of New Particle Formation (NPF) Event Occurrences From 2018 to 2023, Including the Number of Days With NPF Events, Total Measurement Days, and the Percentage of NPF Occurrence Per Year

Year	Days of NPF	Days of total measurement	Percentage (%)
2018	51	340	15
2019	39	333	12
2020	56	333	17
2021	51	320	16
2022	46	165	28
2023	42	308	14
Total	285	1799	16

clearly meet the criteria for either category (less than 10% of the data set)—as well as days with "bad data" due to measurement errors or data loss, were excluded from the analysis. Limiting the categorization to two classes (event and non-event) rather than three was also consistent with recommendations from previous studies, ensuring better classification and reducing ambiguity in identifying NPF events (Joutsensaari et al., 2018). These labels, along with the start time, end time, and duration of each event, were recorded in the corresponding CSV files for further analysis.

Table 1 summarizes the number of days with NPF events identified during the study period from 1 January 2018, to 31 December 2023, using the manual labeling method based on semi-quantitative criteria, along with the total number of measurement days and the percentage of days with NPF occurrences. Over the 6-year period, a total of 285 well-identified NPF event days were observed out of 1,799 measurement days, corresponding to an overall occurrence rate of 16%. The annual distribution of NPF events varied, with 2020 having the highest number of NPF days (56 days, 17%) and 2019 recording the fewest (39 days, 12%). In 2022, the proportion of NPF days peaked at 28%, primarily due to data gaps between 22 August 2022, and 17 January 2023, caused by instrumentation issues at the ARM SGP site. These gaps reduced the total measurement days for 2022, inflating the proportion of NPF events compared to other years, where the percentage of NPF event days more consistently ranged between 12% and 17%.

2.2.2. Random Forest

As discussed in the above section, each day during the study period was classified into the presence or absence of NPF events. This binary categorization facilitated a focused analysis of NPF occurrences. Given its suitability for binary classification tasks, we employed the RF algorithm. The methodological workflow for developing the RF model used for NPF identification is outlined in Figure 1a.

Since its introduction by Leo Breiman in 2001 (Breiman, 2001), RF has become a widely recognized ensemble learning method. Its strength lies in its structure, which consists of multiple decision trees that collectively contribute to the decision-making process. Each tree is trained independently on random subsets of data, reducing correlation among trees and enhancing the model's robustness and generalizability. This diverse learning approach improves the model's ability to generalize across different data sets and perform effectively in various settings. RF models offer several advantages compared to more advanced ML techniques, such as ease of interpretability and straightforward implementation. Their decision tree-based architecture effectively handles complex, nonlinear relationships between predictor variables, even when those variables exhibit interdependencies or correlations, making RF well-suited for this application. Furthermore, compared to black-box models like neural networks or deep learning methods, RF models are easier to visualize and interpret. In terms of predictive accuracy, RF ranks among the most effective ML tools, making it an ideal choice for our analysis.

For NPF event identification, we implemented the RF model using the Python *scikit-learn* package. Feature selection followed three principles: (a) prioritizing features known to influence NPF, (b) excluding features with minimal predictive impact, and (c) retaining minor contributors for comprehensive analysis. To adhere to these principles, we excluded the concentration of particle sizes above 22.5 nm, as they were found to have minimal impact during the initial feature selection process. A total of 45 features were used for training, including NPF start time and concentrations at 44 particle sizes (4.8–22.5 nm). We utilized the raw particle size distribution data obtained from the Nano-SMPS to assess the importance of various features in predicting NPF events. Given the nature of NPF events, particle concentration at different sizes plays a crucial role, and the NPF start time is also expected to be significant due to the influence of time of day on NPF occurrence. Note that only start times from NPF event days (binary label of (a) were included in the training data set. Since non-NPF days do not have defined start times, none were included. By combining these features, we sought to identify the variables with the greatest impact on predicting NPF events.

The data set was randomly split into a training set and a test set in a 70:30 ratio based on the hold-out method, a common practice to ensure effective model training and to assess the model's generalizability (Chen et al., 2022). To mitigate overfitting and robustly evaluate the model's predictive capabilities and its dependence on the training

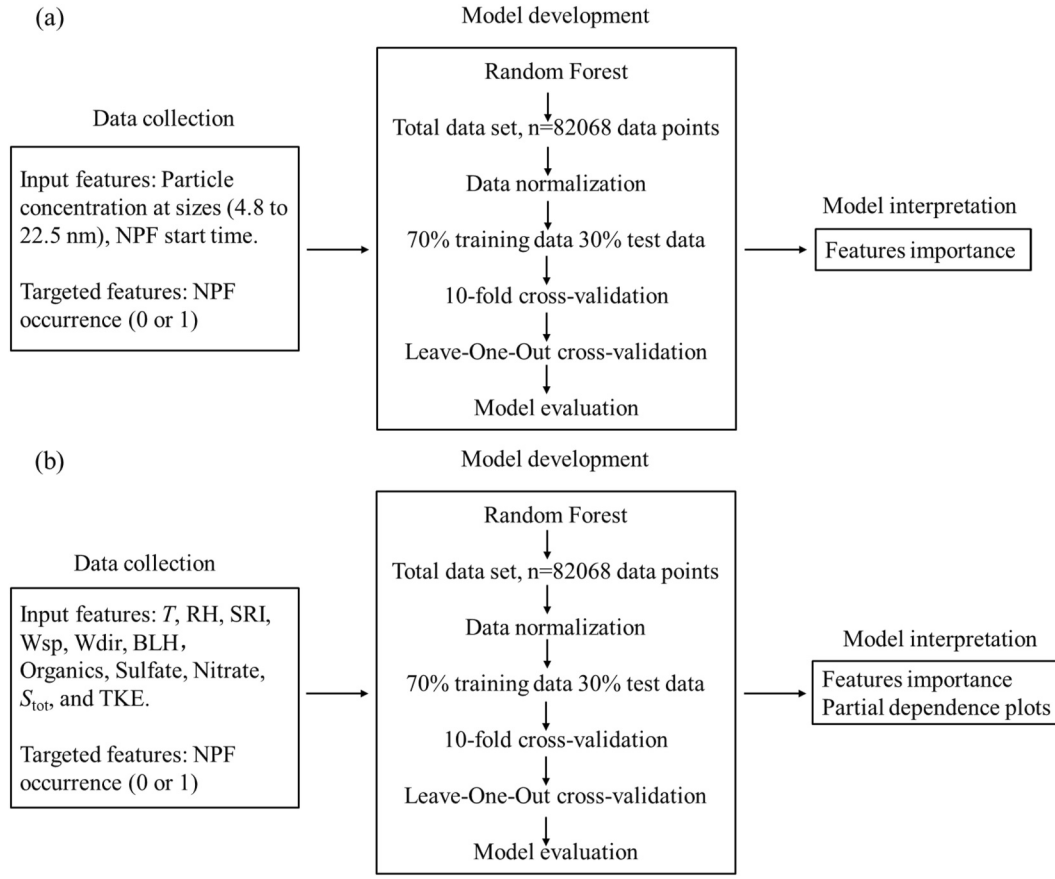

Figure 1. The methodology is outlined in two flowcharts, which cover data collection, model development, and data interpretation for RF model (a) automated identification of new particle formation (NPF) events, and RF model (b) mechanistic analysis of the factors influencing NPF occurrences.

set (Hastie et al., 2009), we employed k-fold cross-validation, dividing the training set into 10 subsamples. In each iteration of 10-fold cross-validation, one subsample was the validation set, while the remaining nine were used for training. Additionally, Leave-One-Out Cross-Validation (LOOCV) was applied, where each data point, representing a specific year, was iteratively excluded from the training set and used for testing. This method ensures the model's generalizability across different years, minimizing the impact of year-to-year variability. These validation methods are summarized in Table 2.

The model performance evaluation process relied on well-established metrics derived from the confusion matrix (Figure 2), which categorizes predictions into four key areas: True Positives (TP), where NPF events are accurately classified as events; False Negatives (FN), where NPF events are mistakenly classified as non-events; False Positives (FP), where non-events are incorrectly identified as events; and True Negatives (TN), where non-events are correctly classified (Chen et al., 2020, 2024). Using these classifications, we calculated performance metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive assessment of the model's efficacy, as shown in Supporting Information S1 and Table 3. Additionally, model interpretation was investigated through feature importance ranking using the Mean Decrease in Impurity (MDI) method, which measures how much each feature contributes to reducing Gini impurity—a measure of classification uncertainty in decision trees. MDI is particularly well-suited for capturing non-linear relationships between features and outcomes and was selected for its computational efficiency, interpretability, and compatibility with tree-based models. By calculating the total reduction in impurity attributed to each feature, this method helped identify the features that most significantly contributed to the occurrence of NPF events. Following the similar method, we conducted the feature importance analysis of the aerosol growth rate, which is derived from each identified NPF event.

Table 2
Accuracy Results for the Random Forest Model for Automated Identification of New Particle Formation Events, Using Different Model Evaluation Methods: Hold-Out Experiment With 70–30 Train-Test Split, 10-Fold Cross-Validation, and Leave-One-Out Cross-Validation, Applied Across Various Test Years

Test methods	Test year	Accuracy
Hold-Out Experiment (70-30 Split)	N/A	0.93
10-Fold Cross Validation	N/A	0.93
Leave-One-Out Cross Validation	2018	0.93
	2019	0.93
	2020	0.92
	2021	0.92
	2022	0.92
	2023	0.92

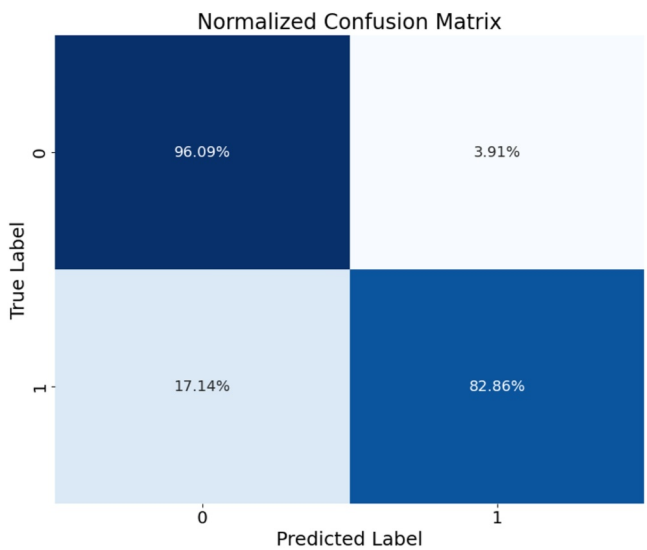

Figure 2. Normalized confusion matrix of the Random Forest model for automated identification of new particle formation (NPF) events. The matrix shows the percentage of correct and incorrect predictions for each class. Class 0 represents non-NPF events, and Class 1 represents NPF events.

2.3. NPF Events Mechanistic Analysis

2.3.1. Measured Variables

The mechanisms driving NPF events are not yet fully understood due to their complex interactions with various environmental factors. To address this, a total of 12 meteorological and chemical variables were selected for model development: T, RH, SRI, Wsp, Wdir, BLH, Organics, Sulfate, Nitrate, S_{tot}, SO_2, and TKE, following the feature selection principles outlined in Section 2.2.2.

The impact of these variables on atmospheric processes governing NPF has been demonstrated in previous research (Bousiotis, Brean, et al., 2021; Hamed et al., 2011; Kerminen et al., 2018; Jimenez et al., 2009). T and RH influence nucleation and condensation rates indirectly by modulating key atmospheric processes. T affects vapor pressure, BLH, and biogenic VOC emissions, while RH influences the condensation sink through hygroscopic growth and cloud-related mixing. Together, they serve as proxies for broader meteorological and chemical dynamics. SRI drives photochemical reactions essential for producing precursor gases involved in NPF. Wsp and Wdir influence the transport and dispersion of aerosols and precursor gases, while BLH controls the vertical mixing of these substances in the atmosphere. TKE also plays a key role in vertical mixing, enhancing the distribution of aerosols and precursor gases within the boundary layer. Additionally, chemical components such as total organics, sulfate, and nitrate serve as indicators of their corresponding gas-phase precursors (VOCs, SO_2, and NO_x), which can contribute to the pool of condensable vapors involved in nucleation and growth processes under suitable T and RH conditions. On the other hand, S_{tot} affects the scavenging of the nucleating vapors and newly formed particles, thereby influencing the occurrence of NPF and the subsequent particle growth. By analyzing these variables, the study aims to provide a comprehensive understanding of the environmental conditions that either promote or inhibit NPF. SO_2 was retained as an exploratory predictor to allow the Random Forest model to evaluate its potential relevance; however, because ARM SO_2 measurements exhibit known sensitivity limitations, we do not interpret SO_2 mechanistically in the context of NPF drivers. To further assess its influence, we repeated the feature importance analysis by omitting SO_2 from the predictor set, and the resulting rankings were nearly identical to those obtained using the full set of variables (Figure S1 in Supporting Information S1). This consistency indicates that the SO_2 measurements available at SGP do not substantially contribute to explaining NPF occurrence within this data set.

2.3.2. Random Forest and Partial Dependence Plots

The flowchart illustrating the methodology used for the mechanistic analysis of NPF events is presented in Figure 1b. Similar to the RF methodology applied for NPF identification, the flowchart outlines the steps of data collection, model development, and interpretation. Accuracy results for the RF model were also obtained using different test methods, including a hold-out experiment with a 70–30 train-test split, 10-fold cross-validation, and LOOCV, as detailed in Table S1 of Supporting Information S1. The model performance was evaluated using a confusion matrix, as shown in Figure S2 and Table S2 of Supporting Information S1. For model interpretation, a feature importance analysis was performed to rank variables based on their significance in predicting NPF events. While this analysis highlighted the importance of variables such as RH and SRI, further information on how these factors influence NPF is still missing.

To provide additional insights, PDPs were generated to visualize the marginal effects of individual variables on NPF occurrence. These plots allowed for a more nuanced understanding of how atmospheric variables influence NPF events across their respective ranges. PDPs indicated whether the effects of the variables increased, decreased, or remained stable under varying

Table 3
Performance Metrics of the Random Forest Model in Predicting New Particle Formation Events

Class	Precision	Recall	F1-score	Accuracy
0	0.95	0.96	0.96	0.93
1	0.85	0.83	0.84	0.93

Note. Class 0 represents the non-NPF event category, while Class 1 corresponds to NPF events. Metrics include precision, recall, F1-score, and accuracy of the model.

conditions. For example, PDPs revealed how RH and SRI contribute to NPF formation at different values, highlighting potential thresholds and nonlinear behaviors in their impact on the process.

3. Results and Discussion

3.1. NPF Events Identification Using RF

To ensure the robustness and reliability of the RF model in identifying NPF events, several evaluation and validation steps were conducted after its development. Table 2 presents three test methods employed: hold-out experiment with a 70–30 train-test split, 10-fold cross-validation with the entire data set, and LOOCV applied year-by-year from 2018 to 2023. The RF classifier achieved an identification accuracy of approximately 92%–93%, demonstrating consistent performance across these validation techniques and confirming the model's robustness in predicting NPF events.

To assess the model's performance across different classes, standard metrics derived from the confusion matrix were utilized. As shown in Figure 2, the normalized confusion matrix indicates that the model correctly identified 96.09% of non-NPF events (Class 0) and 82.86% of NPF events (Class 1). A summary of accuracy, precision, recall, and F1-score is provided in Table 3. Overall, the results indicate strong performance in detecting non-NPF events, but comparatively lower performance in identifying NPF events, highlighting some limitations in accurately identifying all occurrences. Nevertheless, the accuracy of 93% (Table 2) suggests the model is effective at capturing key factors identifying NPF events.

Feature importance analysis, depicted in Figure 3, evaluates several key feature groups, including particle number concentrations across different size ranges ($N_{4.8\text{-}6.8}$, $N_{7.1\text{-}10.2}$, $N_{10.6\text{-}15.7}$, and $N_{16.3\text{-}22.5}$) as well as the NPF start time. The results indicate that smaller particle concentrations, particularly those in the 4.8–6.8 nm range, are the most critical factors influencing the model's predictions, contributing 34% to the model, followed by particle concentration in the 7.1–10.2 nm, 10.6–15.7 nm, and 16.3–22.5 nm size ranges, underscoring the critical role of specific size distributions for capturing the onset of NPF events. The detailed feature importance for each particle size is presented in Figure S3 of Supporting Information S1. Notably, the particle concentration at 4.8 nm shows low feature importance, likely due to limited data availability and the tendency for NPF to occur in the upper boundary layer, where the lower condensation and coagulation sinks, lower ambient temperature, higher actinic flux, and availability of nucleating precursors can promote NPF (Chen et al., 2018; Nilsson et al., 2001; Platis et al., 2016; O'Donnell et al., 2023 Wang et al., 2023; Zheng et al., 2021). As a result, newly formed particles are often detected at the surface only after they have grown to larger sizes.

The 4.8–10.2 nm size range is characteristic of freshly nucleated particles, which typically form from the gas phase and either experience rapid growth via condensation and coagulation or are scavenged by larger pre-existing particles through coagulation. In contrast, while larger particles (10.6–22.5 nm) remain relevant, they may not provide as clear a signal of initial nucleation events. Larger particles are typically associated with significant growth and may be influenced by coagulation or vapor condensation onto pre-existing particles. Thus, the model's sensitivity to smaller particles aligns with the need to capture the early stages of particle formation, highlighting the importance of detecting particles within the 4.8–10.2 nm range for early NPF identification.

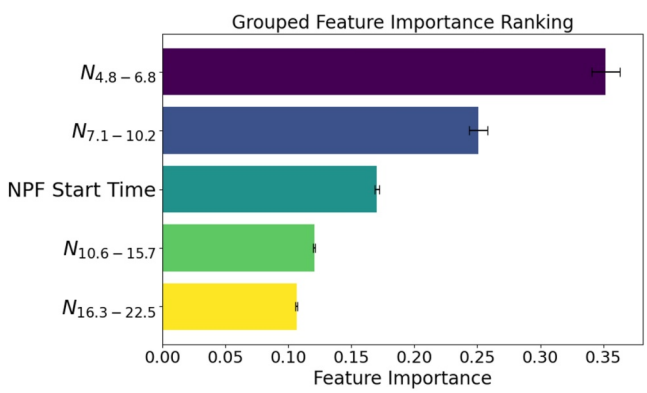

Figure 3. Feature importance ranking for several key feature groups influencing new particle formation (NPF) events identification, including particle number concentrations across different size ranges ($N_{4.8\text{-}6.8}$, $N_{7.1\text{-}10.2}$, $N_{10.6\text{-}15.7}$, and $N_{16.3\text{-}22.5}$) as well as the NPF start time.

Additionally, the start time of NPF events also impacts the model's predictions. NPF events are strongly influenced by diurnal cycles, with nucleation and particle growth typically occurring during specific times of the day when atmospheric conditions are optimal, often coinciding with the increased availability of surrounding vapors, such as sulfuric acid, organic compounds, and ammonia, which facilitate rapid particle growth. By capturing these temporal patterns, the model enhances its ability to predict the timing of NPF events. A more detailed discussion of these temporal influences is provided in Section 3.2.

By prioritizing both time and particle size, the model is fine-tuned to detect the most defining characteristics of NPF events. Smaller particles reflect early nucleation, while larger particles represent later stages of growth influenced by additional factors. The model's attention to diurnal cycles further strengthens its predictive power, as NPF events often follow distinct daily patterns. This combination of temporal and particle size distributions makes the model effective in identifying NPF events, consistent with prior research emphasizing the role of time-dependent atmospheric conditions and particle size distributions in NPF prediction (Mikkonen et al., 2020). Beyond accuracy, this ML-based approach offers practical advantages over traditional visual inspection of banana plots, which can be subjective and labor-intensive for long-term data sets. Our RF model operates directly on raw nano-SMPS size distribution data and interpretable environmental variables, eliminating the need for extensive manual mode detection or the additional data-processing steps by image-based or deep-learning methods. This approach enhances scalability, reproducibility, and potential integration into broader aerosol-climate modeling frameworks, while providing a transparent and computationally efficient framework suited for mechanistic interpretation.

3.2. Temporal Characteristics of NPF Frequency and Growth Rates

When estimating the importance of atmospheric NPF in different environments, it is essential to determine the frequency of its occurrence (Kerminen et al., 2018), defined as the fraction of days with NPF events over a given period. Therefore, in addition to using ML to predict NPF event occurrences and assess the impact of various environmental variables, we analyzed the temporal distribution of NPF events at the SGP site over the 6-year period (2018–2023). The analysis examined diurnal, monthly, and seasonal variations, with the data divided into four meteorological quarters—December to February, March to May, June to August, and September to November—and four time intervals: 00:00–06:00, 06:00–12:00, 12:00–18:00, and 18:00–24:00 Central Standard Time (CST). Although the SGP site observes daylight saving time (CDT) from March to November, we adopted CST (UTC−6) uniformly to ensure temporal consistency across seasons and years and to avoid confusion related to periodic time zone changes. It is important to note that the reported time intervals correspond to the typical start time of NPF events, rather than the entire event duration.

Figure 4a illustrates the monthly distribution of NPF occurrences and their respective start times. The result shows a clear monthly variation in NPF frequency, with events being more frequent between December and April, and with March recording the highest number of occurrences. In contrast, the lowest frequency of NPF events is observed during the summer months (June to August), likely due to atmospheric changes such as higher RH, T fluctuations (see detailed mechanistic discussion in Section 3.3), and/or increased concentrations of pre-existing particles (Liu et al., 2021; Wang et., 2023), as supported by Figure S4 in Supporting Information S1, which provides additional details on the monthly distribution of relevant atmospheric factors. It should be noted that the NPF events predominantly occur between 6:00–12:00 CST. This time frame highlights the strong dependence of NPF on solar radiation, as increased sunlight triggers the photochemical reactions necessary for particle nucleation. The details about the relationship between NPF probability and SRI are discussed in Section 3.3.

Evening events remain rare across all seasons and likely reflect boundary-layer transitions rather than classical photochemically driven NPF. Although daytime photochemistry typically dominates NPF formation, previous field studies have demonstrated that nighttime NPF can occur under specific atmospheric conditions. Observations from urban and forested environments (Kammer et al., 2018; Salimi et al., 2017) as well as from the free troposphere (Lee et al., 2008) show that nocturnal nucleation is possible when low condensation sinks, stable nocturnal boundary layers, and alternative oxidation pathways, particularly NO_3-radical oxidation of biogenic or anthropogenic VOCs, enhance the formation and survival of molecular clusters. While nighttime events at SGP are infrequent and generally weaker than their daytime events, their occurrence patterns and slow growth rates are consistent with the characteristics described in these prior studies.

To further examine seasonal and diurnal trends, Figure S5 in Supporting Information S1 presents complementary visualizations of quarterly distributions and the interaction between seasonal and hourly NPF frequencies. These additional panels provide insight into subtle patterns in NPF occurrence in different seasons, although these trends are also observable in monthly plots (Figure 4). Figure 4 and Figure S5 in Supporting Information S1 show that NPF is predominantly a daytime phenomenon, with most events occurring between sunrise and sunset. The first quarter (December to February, winter) records the highest number of NPF events, while the third quarter (June to August, summer) shows the fewest. Similar seasonal dependencies have been documented at other mid-latitude sites. Multi-site analyses across Europe, for example, show that NPF occurrence reduces during summer periods

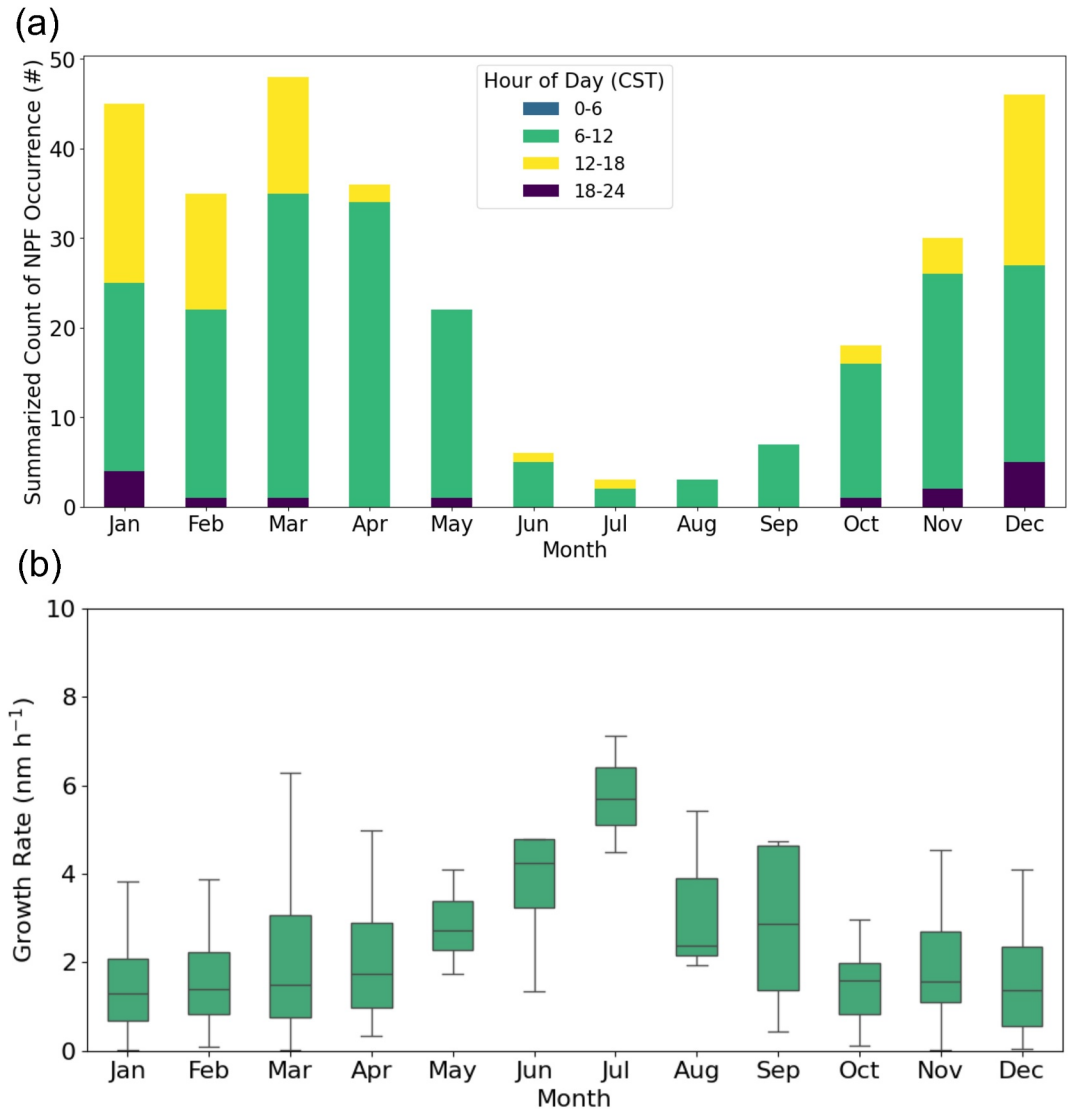

Figure 4. Temporal characterization of new particle formation (NPF) at the Southern Great Plains site from 2018 to 2023: (a) summarized monthly count of NPF occurrences by hour of the day, with the color-coded bars representing time intervals (0–6, 6–12, 12–18, and 18–24 Central Standard Time); (b) summarized monthly variation of NPF growth rates (nm h^{-1}).

characterized by higher condensation sinks, while for the year, enhances during times with elevated solar radiation and reduced humidity (Bousiotis, Brean, et al., 2021). This enhanced sink during the summer was also observed by Marinescu et al. (2019), which examined the 5-year temporal variability of aerosols at SGP. The study also linked the enhanced sink to the suppressed nucleation mode aerosols with sizes between 7 and 30 nm.

Therefore, the seasonal pattern of NPF occurrence is likely a result of precursor availability, sinks, and local meteorology, and this pattern is dependent on the geographical locations. For example, synthesis studies of high-altitude mountain background stations in Europe and other regions (Sellegri et al., 2019) demonstrate a predominance of daytime and warm-season NPF, largely attributed to stronger solar radiation and oxidant production during summer, deeper and better-ventilated boundary layers, and reduced cloudiness and humidity that collectively enhance nucleation conditions. Furthermore, multi-decadal observations at the boreal forest site in Hyytiälä (Dada et al., 2017) highlight the persistence of strong springtime maxima and muted wintertime activity, emphasizing that seasonal modulation of NPF is a robust characteristic of mid-latitude atmospheric environments. Long-term observations in the Yangtze River Delta region of China (Shen et al., 2022) likewise reveal that NPF

events occur most frequently in spring, adding further evidence that the seasonal influence on NPF is common across mid-latitude sites, even though the dominant season and its underlying drivers vary regionally.

A notable feature in Figure 4 and Figure S5 in Supporting Information S1 is that a subset of NPF events in winter months (December-February) occur during the early afternoon (12:00-18:00 CST), a pattern less common in other seasons. This shift may be caused by the reduced solar elevation (where solar radiation is weak in the morning) and lower temperatures, which suppress photochemical precursor production while stabilizing low-volatility vapors and modifying boundary-layer dynamics. Similar wintertime influences on aerosol evolution at the SGP site have been reported by Marinescu et al. (2019), and are consistent with analyses showing that weak wintertime solar radiation and stable boundary-layer structure facilitate precursor buildup and particle formation in mid-latitude continental environments (Boy & Kulmala, 2002; Kulmala et al., 2004). These comparisons indicate that the temporal behavior observed at SGP aligns with patterns documented at other mid-latitude sites.

It is worth noting that the pronounced afternoon occurrence of NPF events during winter may also be influenced by slower particle growth rates (GRs) under colder and more stable atmospheric conditions. To investigate this, we quantitatively evaluated particle GRs of NPF at the SGP site from 2018 to 2023 and identified a clear seasonal pattern, with lower GRs in winter (below 2 nm hr^{-1}) and enhanced growth during late spring and early summer (exceed 5 nm hr^{-1}), as shown in Figure 4b. These findings reinforce the role of photochemical activity and precursor availability in driving particle growth. The reduced GRs observed from August through October likely reflect elevated background aerosol concentrations, which increase the condensation sink and suppress further growth.

In summary, our work extends previous efforts at the SGP site by incorporating 6 yrs of high-resolution observations that span all seasons. Whereas earlier studies focused on short-term spring or summer campaigns (Chen et al., 2018; Hodshire et al., 2016; Liu et al., 2021; O'Donnell et al., 2023) or addressed general aerosol properties (Marinescu et al., 2019), our integrated manual and ML-based classification enabled detailed quantification of seasonal NPF occurrence frequency and particle growth rate analysis. By additionally resolving daytime versus nighttime events, we offer a more nuanced view of diurnal and seasonal variability, enriching the current understanding of temporal NPF behavior in a rural continental setting.

3.3. Analyzing Environmental Variable Relationships Using RF and PDP

The RF algorithm was applied to predict NPF events using 12 atmospheric variables (discussed in Section 2.3.1), achieving a prediction accuracy of 95%. Initially, the model included a broader range of atmospheric variables, but it was refined to focus on these 12 based on the availability of consistent data and their physical relevance to NPF processes. The feature importance rankings were consistent across all folds of the 10-fold cross-validation, reinforcing the robustness of the model's conclusions. Figure 5a highlights the top six atmospheric variables influencing NPF events. SRI emerged as the most important predictor (28%), consistent with previous studies (Boy et al., 2008; Kerminen et al., 2018). RH was the second most important, with a feature importance score of around 17%, followed by T, TKE, S_{tot}, and Wdir, with scores of 10%, 8%, 8%, and 5%, respectively. These variables collectively account for 76% of the model's predictive power, underscoring the complex interplay of multiple atmospheric factors driving NPF events. Notably, SO_2 ranked lowest. However, this does not necessarily indicate that nucleating precursors are not a limiting factor for NPF at the SGP site, as SO_2 alone may not fully represent all relevant precursor species, such as organics, ammonia, or other sulfur compounds, in the initial nucleation and continuous growth of the newly formed particles. Due to limited instrumentation for gas-phase precursor measurements, SO_2 was the only NPF-relevant gas species continuously monitored. The low ranking of SO_2 likely rules out the possibility that NPF was driven by SO_2 oxidation at the surface, a mechanism commonly observed in urban environments (Alam et al., 2003; Meng et al., 2015; Xiao et al., 2015).

To further examine the environmental factors of particle growth during NPF events, we performed RF analysis using key atmospheric variables. As shown in Figure 5b, T was identified as the most important factor, followed by TKE and S_{tot}. These variables reflect the role of thermodynamic conditions and atmospheric turbulence in regulating the condensation and diffusion of vapors onto nucleating particles. RH was also a key factor, consistent with its known impact on the condensation sink and chemical transformation processes. Additionally, nitrate and wind direction were also among the top six features, suggesting that aerosol chemical composition and regional air mass transport play roles in growth dynamics. Together, these results suggest both meteorological and

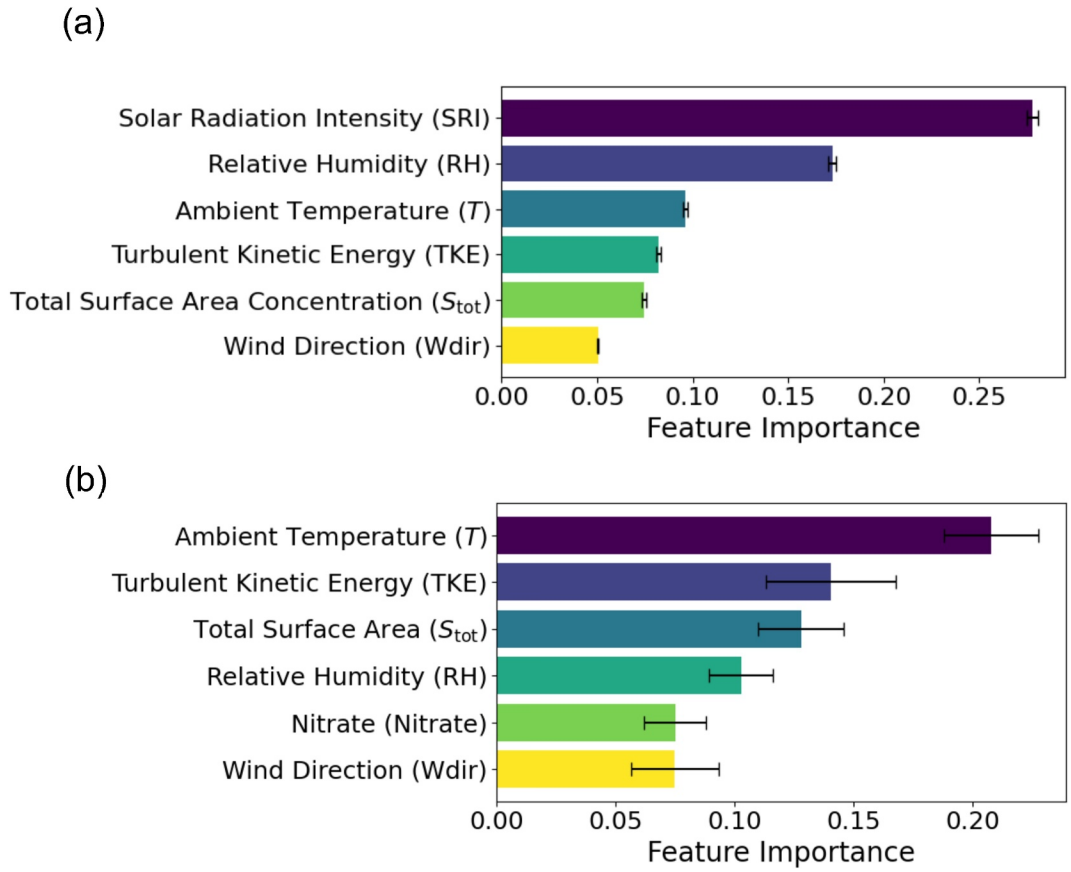

Figure 5. Top six most influential atmospheric variables ranked by feature importance using the Random Forest model. The horizontal bars represent the relative importance of each feature. (a) new particle formation event occurrence. (b) Particle growth rates.

compositional factors contribute to the observed seasonal variability in GRs, with lower GRs in winter and higher GRs during late spring and early summer, as reflected in Figure 4b.

To visualize the impact of individual variables, PDPs for the top six features are shown in Figure 6, revealing influences consistent with physical explanations and highlighting the alignment between the ML model and underlying physical phenomena. These plots focus primarily on daytime events, as 95% of the observed NPF events occur during the day. Each PDP isolates the effect of a single variable while holding others constant, offering insight into how each factor influences NPF probability. In Figure 6a, SRI shows a strong positive correlation with NPF likelihood, consistent with its role in driving photochemical reactions and diurnal NPF cycles (Marinescu et al., 2019). Solar radiation also drives atmospheric chemistry, producing low-volatility vapors, such as sulfuric acid and extremely low volatility organic compounds (ELVOCs), which are key to NPF processes (Boy et al., 2008; Ehn et al., 2014). The steep rise in NPF probability for SRI values above 200 W m^{-2} suggests that intense sunlight substantially enhances the likelihood of particle formation.

In contrast, RH (Figure 6b) negatively correlates with NPF probability. This effect is likely due to increased condensation sink under higher moisture conditions, where vapors condense onto the surface of pre-existing particles, reducing the concentration of condensable vapors, such as sulfuric acid and organics, which are needed for nucleation and growth. The hygroscopic growth of pre-existing particles at high RH increases their surface area, accelerating the depletion of these vapors. Similarly, due to the hygroscopic growth of the particles, the coagulation scavenging of sub-3 nm clusters is also enhanced. Moreover, high RH can be under cloud cover, and thus reduce SRI, weakening photochemical reactions critical for producing aerosol precursor gases. This reduction in solar radiation diminishes gas-phase oxidation chemistry and, consequently, reduces NPF activity (Hamed et al., 2011).

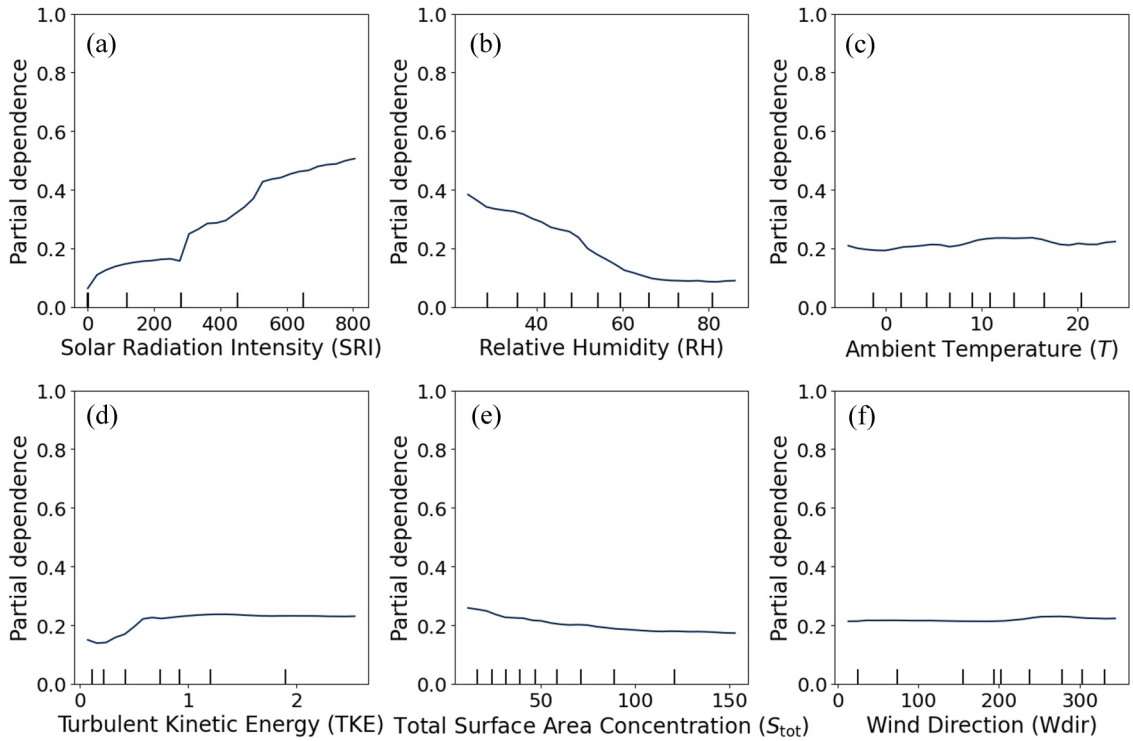

Figure 6. Partial Dependence Plots for the six most influential atmospheric variables affecting new particle formation (NPF) events. Panels (a) to (f) illustrate the marginal effects of solar radiation intensity, relative humidity (RH), ambient temperature (T), turbulent kinetic energy, total surface area concentration (S_{tot}), and wind direction (Wdir) on the probability of NPF events.

T, as shown in Figure 6c, has a minimal impact on NPF occurrences within the observed range (0°C–25°C). This is likely since T is associated with other atmospheric variables, such as SRI, RH, and TKE, indicating that the observed correlation with NPF might be influenced by several interacting factors rather than T alone, as shown in Figure S6 of Supporting Information S1. While some studies (Lee et al., 2019) indicate that T can influence gas-phase chemistry involved in particle nucleation, in this data set, the impact of T may be overshadowed by stronger influences such as SRI and RH.

Figure 6d shows a modest positive relationship between TKE and NPF probability, indicating that turbulence could enhance precursor mixing but plays a secondary role compared to factors like solar radiation. Figure 6e reveals that S_{tot} has a negative effect on NPF probability, due to pre-existing aerosols acting as sinks for condensable vapors and newly formed particles, thereby suppressing nucleation. Note that in our study, S_{tot} is less important than SRI and RH, differing from previous research (Cai et al., 2017). This suggests that NPF at SGP is not limited by S_{tot}, indicating the abundance of nucleating vapor molecules, making the condensation sink effect less critical in this environment. Figure 6f shows that Wdir has little impact on NPF probability within the 0–360° range, suggesting that wind direction is not a significant factor of NPF events.

To explore the combined effects of variables on NPF events, three two-way PDPs are presented in Figure S7 of Supporting Information S1. Figure S7a in Supporting Information S1 shows that NPF probability increases with higher SRI, particularly when RH is below 50%, likely due to enhanced photochemical activity and reduced moisture content. Figure S7b in Supporting Information S1 shows that lower T and RH below 50% favor NPF, likely due to the reduced volatility of nucleating vapors. Conversely, higher T and RH are associated with reduced NPF probability, which may reflect enhanced gas-phase dilution and an increased condensation sink that suppress nucleation (Kulmala et al., 2004; Zhang et al., 2012). Figure S7c in Supporting Information S1 demonstrates that NPF probability rises with higher SRI and moderate T (10–20°C), which may coincide with enhanced photochemical activity conducive to nucleation processes, although the interaction is likely influenced by SRI, T, and RH feedbacks and should be interpreted as a correlation rather than causation.

We also quantitatively analyzed correlations between six atmospheric variables using Pearson's correlation coefficient (Figure S6 in Supporting Information S1). SRI is negatively correlated with RH (−0.44), reflecting diurnal patterns, and positively correlated with T (0.31) and TKE (0.46), suggesting that solar radiation enhances atmospheric mixing and aerosol distribution. While these relationships align with previous studies on the role of solar radiation in shaping conditions favorable for NPF (Kulmala et al., 2004; Zhang et al., 2012), we emphasize that such correlations should be interpreted as potential indicators rather than definitive causal drivers. Additionally, S_{tot} shows negligible correlation with other variables, implying its dependence on aerosol sources or removal mechanisms rather than meteorological factors. Despite moderate correlations between some variables, the RF model effectively captures the non-linear and multi-dimensional interactions between these factors, as supported by previous studies (Yang et al., 2023).

It is worth noting that some variables exhibit high feature importance but relatively flat PDPs. This discrepancy is due to how feature importance in RF models is determined by factors like the frequency and depth of splits, which do not always correspond to the variable's direct influence on predictions. High feature importance indicates that the model frequently relies on the feature for splits, but this does not necessarily reflect the feature's direct influence on the output since the RF model does not have a fundamental physical understanding of the processes involved. Conversely, a PDP shows the average effect of a feature on the model's predictions across its range, marginalizing other features. For example, although T ranks third in RF feature importance, the PDP shows it does not significantly affect NPF probability alone, suggesting its influence likely arises through interactions with other variables, such as SRI.

In summary, the RF model accurately predicted NPF events at the SGP site, with SRI and RH emerging as key factors. These results emphasize the importance of these variables in understanding NPF mechanisms and suggest they should be prioritized in future studies. While the understanding that these variables are important for NPF is not new, our approach offers a more interpretable and quantitative framework. We apply RF directly to raw particle size distribution data, combined with PDP analysis, to reveal nonlinear behaviors and threshold effects. Notably, no prior study has applied this ML-based framework at SGP. With over two decades of continuous operation and growing interest in aerosol processes at DOE sites, our study at SGP offers timely, site-specific insights to guide broader applications at other long-term observatories.

Additionally, future analyses may benefit from incorporating composite proxy variables that reflect the balance between precursor production and loss processes. For instance, the ratio of solar radiation and SO_2 to the condensation sink ($SRI \times SO_2/S_{tot}$) could serve as an indicator of the net availability of condensable vapors, particularly sulfuric acid, helping to approximate the effective lifetime of nucleating species. Expanding this methodology to include such composite indicators and applying it across different geographic regions could enhance the generalizability and predictive power of ML models for NPF prediction. Indeed, our RF model further demonstrates that, under a log-linear regression framework, NPF probability is proportional to $SRI^{0.4} \times RH^{-5.2} \times T^{0.05} \times S_{tot}{}^{-0.44}$. While this derived probability does not directly reflect the fundamental mechanisms of NPF and may vary by geographic region, the equation highlights the complex non-linearity of the NPF process and the influence of key atmospheric variables.

3.4. Case Studies Analysis

In our analysis of NPF events, we considered measurements collected during both daylight and nighttime hours. We observed that 95% of NPF events occur during daylight and 5% at nighttime at the SGP site. To understand the similarities and differences between the daytime and nighttime NPF events, we compared two distinct cases: a daytime NPF event on 18 May 2020, and a nighttime NPF event on 20 December 2018, as shown in Figure 7. These cases were identified based on particle size distribution and characterized with various environmental parameters, including SRI, RH, T, TKE, S_{tot}, and Wdir. This comparison provides insights into the factors affecting particle formation under different atmospheric conditions.

Figure 7a presents a time series of key atmospheric parameters for the daytime NPF event on 18 May 2020. The particle size distribution shows a burst of particles below 10 nm, particularly between 16:00 and 24:00 UTC. This NPF event aligns with an increase in SRI and TKE, as reflected in rising T and decreasing RH. Solar-driven photochemistry played a critical role in initiating the event by facilitating the production of nucleating species such as sulfuric acid and ELVOCs. Diurnal patterns in Wdir, T, and RH further highlight the influence of SRI, with a sharp decrease in RH and an increase in T during the afternoon coinciding with the peak nucleation period.

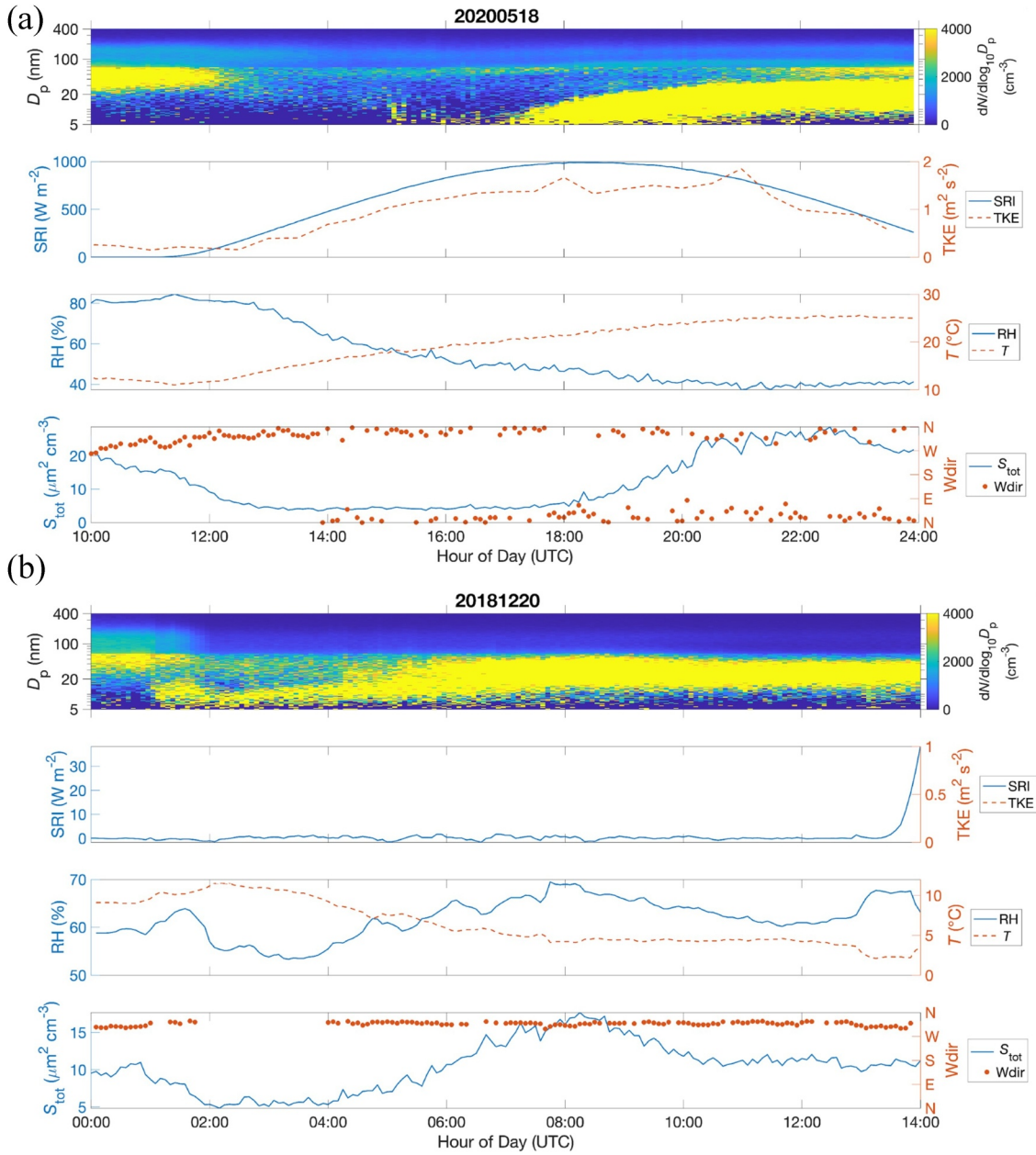

Figure 7. Time series of various atmospheric parameters recorded on (a) daytime new particle formation (NPF) event on 18 May 2020, and (b) nighttime NPF event on 20 December 2018. The panels from top to bottom represent particle size distribution, solar radiation intensity, turbulent kinetic energy (TKE), relative humidity (RH), ambient temperature (T), total surface area concentration (S_{tot}), and wind direction (Wdir). Note: TKE data was not available for the nighttime NPF event, and wind direction data has some gaps during this time period.

These conditions accelerated particle formation and growth, demonstrating the critical role of meteorological conditions in daytime NPF events.

Figure 7b presents a time series of key atmospheric parameters for the nighttime event on 20 December 2018. The particle size distribution indicates a distinct burst of particles below 10 nm, particularly between 00:00 and 06:00 UTC. The absence of SRI, combined with cooler T and higher RH, likely facilitated a different nucleation mechanism compared to daytime events. Without the influence of photochemical reactions, nocturnal oxidation mechanisms—such as the oxidation of organic compounds—may have driven particle formation (Liu et al., 2024). We observed RH during nighttime NPF events tends to be higher than during daytime events, despite fluctuations in RH during the NPF period. The higher RH suggests a stronger condensation sink, limiting the availability of

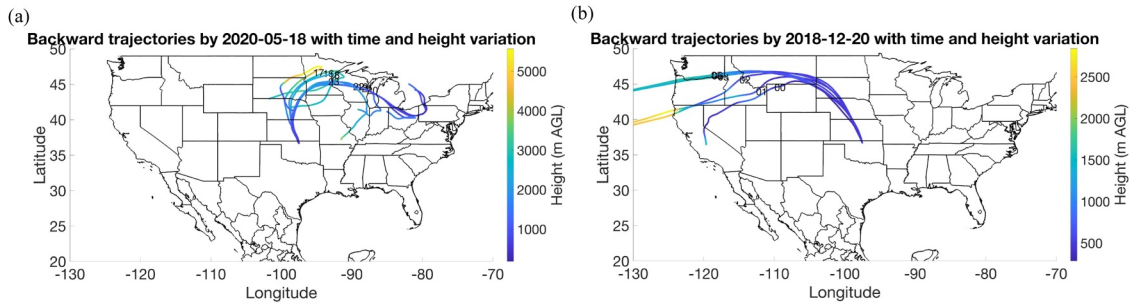

Figure 8. 72-hr HYSPLIT back trajectories for (a) a daytime new particle formation (NPF) event on 18 May 2020, and (b) a nighttime NPF event on 20 December 2018. The trajectories were calculated using the NOAA HYSPLIT model, with initialization at 500 m above ground level (a.g.l.). The color gradient along the trajectories represents different heights, while distinct lines correspond to different back trajectory start times during the respective NPF event periods.

precursor gases and slowing particle growth. These conditions imply that nighttime NPF events may involve slower nucleation pathways, resulting in less intense particle growth (1.75 nm hr^{-1}) compared to daytime events (3.36 nm hr^{-1}). Note that these growth rates are based on two selected case studies and illustrate typical differences between daytime and nighttime events, but they do not reflect the full variability across all observations and have limited generalizability. Assuming that the particles maintained a consistent growth rate following the onset of NPF (when they form the stable critical nucleus of 1 nm), we estimated the onset time of these two NPF events, which were approximately 10:30 CDT for the daytime event and around 17:00 CST for the nighttime event.

Additionally, we assessed the influence of air mass sources on these NPF events by conducting a 72-hr air mass back trajectory analysis for each case using the NOAA Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model. The HYSPLIT model analysis was performed to trace the potential source regions of air masses contributing to particle formation and to determine aerosol transport pathways to the SGP site. For each case, the model was initialized at ~500 m above ground level (a.g.l.) at the onset of the NPF event, allowing us to trace the trajectory of the contributing air masses.

For the daytime NPF event on 18 May 2020 (Figure 8a), the trajectory analysis reveals that the air masses originated from the northeastern part of the country, an area characterized by urban and industrial emissions, at altitudes ranging between 1,500 and 4,000 m a.g.l. These areas consistently report elevated SO_2 and VOC emissions, based on long-term inventories from the U.S. Environmental Protection Agency (EPA, 2017) and satellite-based retrievals from the Ozone Monitoring Instrument (OMI) (Krotkov et al., 2016). The long-range transport of air masses likely facilitated the accumulation of both natural and anthropogenic precursors, further supporting the NPF event at SGP. The passage of these air parcels over industrial areas may have increased the concentrations of sulfuric acid, VOCs, and other nucleating agents, which, in the presence of solar radiation, enhanced particle formation.

In contrast, the HYSPLIT back-trajectory paths for the nighttime NPF event on 20 December 2018 (Figure 8b), show that air masses moved eastward from the Pacific Northwest, traversing the Rocky Mountains before arriving at the central plains. The height of these air masses decreased significantly early in their journey, maintaining low altitudes for a prolonged period before arriving at the site. These air masses likely experienced significant changes in *T*, RH, and atmospheric pressure, all of which could impact aerosol processes like nucleation and growth (Liu et al., 2021). Coastal regions may have introduced marine aerosols or moisture into the air parcels, further influencing the conditions favorable for NPF at the SGP site (Peltola et al., 2023). Although solar radiation is absent, the retention of precursor gases and reduced condensation sink conditions may enable nighttime NPF (Salimi et al., 2017).

Overall, this case study demonstrates the complex interaction between atmospheric conditions and air mass origins in driving NPF events. The daytime event was driven by solar radiation, enhancing photochemical reactions and resulting in faster particle growth, while the nighttime event was influenced by non-photochemical processes, such as nocturnal oxidation and higher RH, which limited particle growth. The HYSPLIT trajectory analysis highlights how air masses transported precursors from distinct regions, each contributing uniquely to NPF dynamics. This comparison underscores the need for further research into the mechanisms behind particle formation under varying environmental conditions. We also acknowledge the limitation that source attribution

remains uncertain due to the absence of in-trajectory chemical processing. In future work, chemical transport modeling and satellite-derived data sets will be integrated into HYSPLIT to better quantify source contributions and further characterize the environmental drivers of NPF.

4. Conclusions

In this study, we introduced a Machine Learning (ML)-based approach for the automated identification and mechanistic analysis of NPF events. Utilizing a Random Forest Classifier, we analyzed the 6-year data set from the SGP site in Oklahoma, USA, effectively distinguishing between NPF events and non-events across various temporal segments. By utilizing raw aerosol data rather than derived variables, this method demonstrated significant potential for long-term NPF event detection, achieving high accuracy (90%–95%). SRI, relative humidity (RH), and ambient temperature (T) as the most influential factors in determining NPF occurrences, with normalized importance values of 28%, 17%, and 10%, respectively. Seasonal patterns revealed that NPF events were more frequent in winter (December to February, 42.1%) and spring (March to May, 35.5%), with a notably lower frequency in summer (June to August, 4.0%). Growth rates (GRs) also exhibited a clear seasonal trend, with the lowest values in winter (below 2 nm hr^{-1}) and the highest in late spring and early summer (exceeding 5 nm hr^{-1}). T, TKE, and aerosol properties were key drivers of GRs, supporting the observed seasonal patterns and demonstrating the broader utility of ML in characterizing both the occurrence and evolution of NPF events. These findings highlight the robustness of our approach for future airborne missions and atmospheric studies.

While the model performs well in classifying NPF events, several limitations should be noted. The model's effectiveness depends on the characteristics of the training data set, and it currently lacks an embedded understanding of the underlying physical processes governing NPF events. This study focuses primarily on strong and clear NPF events, leaving the detection and analysis of weaker events, such as "apple-case" and incomplete NPF, for future work (Manninen et al., 2010). Additionally, as most NPF studies are based on ground-level observations, questions remain regarding the vertical extent of these processes. Incorporating vertical profiles would provide a more comprehensive understanding of NPF dynamics and their spatial distributions. This study is also limited to the SGP site, which has limited atmospheric diversity, and thus, the generalizability of the findings to other regions with different meteorological and emission profiles remains to be validated.

Looking ahead, integrating physical insights into ML frameworks and expanding the model to include other geographic regions will improve its generalizability and robustness. Advanced methods such as recurrent neural networks (RNNs) for time-series prediction or unsupervised learning for discovering new NPF classes could further enhance model capability. Reinforcement learning also holds promise for addressing more complex or weakly labeled NPF scenarios. Collectively, this work provides a strong foundation for predictive NPF modeling and offers actionable insights for understanding aerosol-climate interactions and designing targeted field campaigns.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The observational data from the SGP site are accessible via the ARM's Data Discovery platform upon free registration (Atmospheric Radiation Measurement User Facility, 2024, https://arm.gov/capabilities/observatories/sgp). The code for data analysis, machine learning, and processed data is available at https://github.com/pmtlum/NPF_ML, with the associated DOI https://doi.org/10.5281/zenodo.14392940 (v1.0.0 - Initial Release).

Acknowledgments
This research was primarily supported by the U.S. Department of Energy's Atmospheric System Research (ASR), an Office of Science Biological and Environmental Research program, under DE-SC0023668 and DE-SC0024084. WH was also supported by the National Science Foundation (NSF) Non-Academic Research Internships for Graduate Students (INTERN) under Award 2324142. We would like to thank Battelle and the Pacific Northwest National Laboratory (PNNL) for their contributions to this research. Data were obtained from the ARM user facility, a U.S. Department of Energy (DOE) Office of Science user facility managed by the Biological and Environmental Research Program. PNNL is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830.

References

Ackerman, T. P., & Stokes, G. M. (2003). The atmospheric radiation measurement program. *Physics Today*, *56*(1), 38–44. https://doi.org/10.1063/1.1554135

Alam, A., Shi, J. P., & Harrison, R. M. (2003). Observations of new particle formation in urban air. *Journal of Geophysical Research: Atmospheres*, *108*(D3). https://doi.org/10.1029/2001JD001417

Atmospheric Radiation Measurement User Facility. (2024). *Southern Great Plains (SGP) observatory data*. U.S. Department of Energy. Retrieved from https://arm.gov/capabilities/observatories/sgp

Bousiotis, D., Brean, J., Pope, F. D., Dall'Osto, M., Querol, X., Alastuey, A., et al. (2021). The effect of meteorological conditions and atmospheric composition in the occurrence and development of new particle formation (NPF) events in Europe. *Atmospheric Chemistry and Physics*, *21*(5), 3345–3370. https://doi.org/10.5194/acp-21-3345-2021

Bousiotis, D., Pope, F. D., Beddows, D., Dall'Osto, M., Massling, A., Nøjgaard, J. K., et al. (2021). A phenomenology of new particle formation (NPF) at 13 European sites. *Atmospheric Chemistry and Physics*, *21*(15), 11905–11925. https://doi.org/10.5194/acp-21-11905-2021

Boy, M., Karl, T., Turnipseed, A., Mauldin, R. L., Kosciuch, E., Greenberg, J., et al. (2008). New particle formation in the Front Range of the Colorado Rocky Mountains. *Atmospheric Chemistry and Physics*, *8*(6), 1577–1590. https://doi.org/10.5194/acp-8-1577-2008

Boy, M., & Kulmala, M. (2002). Nucleation events in the continental boundary layer: Influence of physical and meteorological parameters. *Atmospheric Chemistry and Physics*, *2*, 1–16. https://doi.org/10.5194/acp-2-1-2002

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

Bzdek, B. R., Zordan, C. A., LutherIII, G. W., & Johnston, M. V. (2011). Nanoparticle chemical composition during new particle formation. *Aerosol Science & Technology*, *45*(8), 1041–1048. https://doi.org/10.1080/02786826.2011.580392

Cai, R., Yang, D., Fu, Y., Wang, X., Li, X., Ma, Y., et al. (2017). Aerosol surface area concentration: A governing factor in new particle formation in Beijing. *Atmospheric Chemistry and Physics*, *17*(20), 12327–12340. https://doi.org/10.5194/acp-17-12327-2017

Chakraborty, T., & Lee, X. (2021). Using supervised learning to develop BaRAD, a 40-year monthly bias-adjusted global gridded radiation dataset. *Scientific Data*, *8*(1), 238. https://doi.org/10.1038/s41597-021-01016-4

Chen, H., Hodshire, A. L., Ortega, J., Greenberg, J., McMurry, P. H., Carlton, A. G., et al. (2018). Vertically resolved concentration and liquid water content of atmospheric nanoparticles at the US DOE Southern Great Plains site. *Atmospheric Chemistry and Physics*, *18*(1), 311–326. https://doi.org/10.5194/acp-18-311-2018

Chen, H., Leu, M. C., & Yin, Z. (2022). Real-time multi-modal human–robot collaboration using gestures and speech. *Journal of Manufacturing Science and Engineering*, *144*(10), 101007. https://doi.org/10.1115/1.4054297

Chen, H., Tao, W., Leu, M. C., & Yin, Z. (2020). Dynamic gesture design and recognition for human–robot collaboration with convolutional neural networks. In *Proceedings of the international symposium on flexible automation* (Vol. 83617). American Society of Mechanical Engineers. V001T09A001. https://doi.org/10.1115/isfa2020-9609

Chen, H., Zendehdel, N., Leu, M. C., & Yin, Z. (2024). Fine-grained activity classification in assembly based on multi-visual modalities. *Journal of Intelligent Manufacturing*, *35*(5), 2215–2233. https://doi.org/10.1007/s10845-023-02152-x

Dada, L., Paasonen, P., Nieminen, T., Buenrostro Mazon, S., Kontkanen, J., Peräkylä, O., et al. (2017). Long-term analysis of clear-sky new particle formation events and nonevents in Hyytiälä. *Atmospheric Chemistry and Physics*, *17*(10), 6227–6241. https://doi.org/10.5194/acp-17-6227-2017

Dal Maso, M., Kulmala, M., Riipinen, I., Wagner, R., Hussein, T., Aalto, P. P., & Lehtinen, K. E. (2005). Formation and growth of fresh atmospheric aerosols: Eight years of aerosol size distribution data from SMEAR II, Hyytiälä. *Boreal Environment Research*, *10*, 323.

Du, W., Cai, J., Zheng, F., Yan, C., Zhou, Y., Guo, Y., et al. (2022). Influence of aerosol chemical composition on condensation sink efficiency and new particle formation in Beijing. *Environmental Science and Technology Letters*, *9*(5), 375–382. https://doi.org/10.1021/acs.estlett.2c00159

Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., et al. (2014). A large source of low-volatility secondary organic aerosol. *Nature*, *506*(7489), 476–479. https://doi.org/10.1038/nature13032

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

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational & Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Grange, S. K., & Carslaw, D. C. (2019). Using meteorological normalisation to detect interventions in air quality time series. *Science of the Total Environment*, *653*, 578–588. https://doi.org/10.1016/j.scitotenv.2018.10.344

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

Hamed, A., Korhonen, H., Sihto, S. L., Joutsensaari, J., Järvinen, H., Petäjä, T., et al. (2011). The role of relative humidity in continental new particle formation. *Journal of Geophysical Research*, *116*(D3), D03202. https://doi.org/10.1029/2010jd014186

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, Springer

Hodshire, A. L., Lawler, M. J., Zhao, J., Ortega, J., Jen, C., Yli-Juuti, T., et al. (2016). Multiple new-particle growth pathways observed at the US DOE Southern Great Plains field site. *Atmospheric Chemistry and Physics*, *16*(14), 9321–9348. https://doi.org/10.5194/acp-16-9321-2016

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., et al. (2009). Evolution of organic aerosols in the atmosphere. *Science*, *326*(5959), 1525–1529. https://doi.org/10.1126/science.1180353

Joutsensaari, J., Ozon, M., Nieminen, T., Mikkonen, S., Lähivaara, T., Decesari, S., et al. (2018). Identification of new particle formation events with deep learning. *Atmospheric Chemistry and Physics*, *18*(13), 9597–9615. https://doi.org/10.5194/acp-18-9597-2018

Kammer, J., Perraudin, E., Flaud, P. M., Lamaud, E., Bonnefond, J. M., & Villenave, E. (2018). Observation of nighttime new particle formation over the French Landes forest. *Science of the Total Environment*, *621*, 1084–1092. https://doi.org/10.1016/j.scitotenv.2017.10.118

Kerminen, V.-M., Chen, X., Vakkari, V., Petäjä, T., Kulmala, M., & Bianchi, F. (2018). Atmospheric new particle formation and growth: Review of field observations. *Environmental Research Letters*, *13*(10), 103003. https://doi.org/10.1088/1748-9326/aadf3c

Krotkov, N. A., McLinden, C. A., Li, C., Lamsal, L. N., Celarier, E. A., Marchenko, S. V., et al. (2016). Aura OMI observations of regional SO_2 and NO_2 pollution changes from 2005 to 2015. *Atmospheric Chemistry and Physics*, *16*(7), 4605–4629. https://doi.org/10.5194/acp-16-4605-2016

Kuang, C., Riipinen, I., Sihto, S.-L., Kulmala, M., McCormick, A., & McMurry, P. (2010). An improved criterion for new particle formation in diverse atmospheric environments. *Atmospheric Chemistry and Physics*, *10*(17), 8469–8480. https://doi.org/10.5194/acp-10-8469-2010

Kulkarni, P., Baron, P. A., & Willeke, K. (2011). *Aerosol measurement: Principles, techniques, and applications* (3rd ed.). John Wiley & Sons.

Kulmala, M., Petäjä, T., Ehn, M., Thornton, J., Sipilä, M., Worsnop, D., & Kerminen, V.-M. (2014). Chemistry of atmospheric nucleation: On the recent advances on precursor characterization and atmospheric cluster composition in connection with atmospheric new particle formation. *Annual Review of Physical Chemistry*, *65*(1), 21–37. https://doi.org/10.1146/annurev-physchem-040412-110014

Kulmala, M., Petäjä, T., Nieminen, T., Sipilä, M., Manninen, H. E., Lehtipalo, K., et al. (2012). Measurement of the nucleation of atmospheric aerosol particles. *Nature Protocols*, *7*(9), 1651–1667. https://doi.org/10.1038/nprot.2012.091

Kulmala, M., Vehkamäki, H., Petäjä, T., Dal Maso, M., Lauri, A., Kerminen, V.-M., et al. (2004). Formation and growth rates of ultrafine atmospheric particles: A review of observations. *Journal of Aerosol Science*, *35*(2), 143–176. https://doi.org/10.1016/j.jaerosci.2003.10.003

Kürten, A., Bergen, A., Heinritzi, M., Leiminger, M., Lorenz, V., Piel, F., et al. (2016). Observation of new particle formation and measurement of sulfuric acid, ammonia, amines and highly oxidized organic molecules at a rural site in central Germany. *Atmospheric Chemistry and Physics*, *16*(19), 12793–12813. https://doi.org/10.5194/acp-16-12793-2016

Kürten, A., Li, C., Bianchi, F., Curtius, J., Dias, A., Donahue, N. M., et al. (2018). New particle formation in the sulfuric acid–dimethylamine–water system: Reevaluation of CLOUD chamber measurements and comparison to an aerosol nucleation and growth model. *Atmospheric Chemistry and Physics*, *18*(2), 845–863. https://doi.org/10.5194/acp-18-845-2018

Laaksonen, A., Kulmala, M., O'Dowd, C., Joutsensaari, J., Vaattovaara, P., Mikkonen, S., et al. (2008). The role of VOC oxidation products in continental new particle formation. *Atmospheric Chemistry and Physics*, *8*(10), 2657–2665. https://doi.org/10.5194/acp-8-2657-2008

Lee, S. H., Gordon, H., Yu, H., Lehtipalo, K., Haley, R., Li, Y., & Zhang, R. (2019). New particle formation in the atmosphere: From molecular clusters to global climate. *Journal of Geophysical Research: Atmospheres*, *124*(13), 7098–7146. https://doi.org/10.1029/2018jd029356

Lee, S. H., Young, L. H., Benson, D. R., Suni, T., Kulmala, M., Junninen, H., et al. (2008). Observations of nighttime new particle formation in the troposphere. *Journal of Geophysical Research*, *113*(D10). https://doi.org/10.1029/2007jd009351

Leng, C., Zhang, Q., Tao, J., Zhang, H., Zhang, D., Xu, C., et al. (2014). Impacts of new particle formation on aerosol cloud condensation nuclei (CCN) activity in Shanghai: Case study. *Atmospheric Chemistry and Physics*, *14*(20), 11353–11365. https://doi.org/10.5194/acp-14-11353-2014

Li, X., Chee, S., Hao, J., Abbatt, J. P., Jiang, J., & Smith, J. N. (2019). Relative humidity effect on the formation of highly oxidized molecules and new particles during monoterpene oxidation. *Atmospheric Chemistry and Physics*, *19*(3), 1555–1570. https://doi.org/10.5194/acp-19-1555-2019

Liu, J., Alexander, L., Fast, J., Lindenmaier, R., & Shilling, J. (2021). Aerosol characteristics at the Southern Great Plains site during the HI-SCALE campaign. *Atmospheric Chemistry and Physics*, *21*(6), 5101–5116. https://doi.org/10.5194/acp-21-5101-2021

Liu, L., Hohaus, T., Franke, P., Lange, A. C., Tillmann, R., Fuchs, H., et al. (2024). Observational evidence reveals the significance of nocturnal chemistry in seasonal secondary organic aerosol formation. *npj Climate and Atmospheric Science*, *7*(1), 207. https://doi.org/10.1038/s41612-024-00747-6

Ma, N., & Birmili, W. (2015). Estimating the contribution of photochemical particle formation to ultrafine particle number averages in an urban atmosphere. *Science of the Total Environment*, *512*, 154–166. https://doi.org/10.1016/j.scitotenv.2015.01.009

Manninen, H., Nieminen, T., Asmi, E., Gagné, S., Häkkinen, S., Lehtipalo, K., et al. (2010). EUCAARI ion spectrometer measurements at 12 European sites–analysis of new particle formation events. *Atmospheric Chemistry and Physics*, *10*(16), 7907–7927. https://doi.org/10.5194/acp-10-7907-2010

Marinescu, P. J., Levin, E. J., Collins, D., Kreidenweis, S. M., & van den Heever, S. C. (2019). Quantifying aerosol size distributions and their temporal variability in the Southern Great Plains, USA. *Atmospheric Chemistry and Physics*, *19*(18), 11985–12006. https://doi.org/10.5194/acp-19-11985-2019

Mather, J. H., & Voyles, J. W. (2013). The ARM Climate Research Facility: A review of structure and capabilities. *Bulletin of the American Meteorological Society*, *94*(3), 377–392. https://doi.org/10.1175/bams-d-11-00218.1

McMurry, P., & Friedlander, S. (1967). New particle formation in the presence of an aerosol. *Atmospheric Environment*, *13*(12), 1635–1651. https://doi.org/10.1016/0004-6981(79)90322-6

Meng, H., Zhu, Y., Evans, G. J., Jeong, C. H., & Yao, X. (2015). Roles of SO_2 oxidation in new particle formation events. *Journal of Environmental Sciences*, *30*, 90–101. https://doi.org/10.1016/j.jes.2014.12.002

Mikkonen, S., Németh, Z., Varga, V., Weidinger, T., Leinonen, V., Yli-Juuti, T., & Salma, I. (2020). Decennial time trends and diurnal patterns of particle number concentrations in a central European city between 2008 and 2018. *Atmospheric Chemistry and Physics Discussions*, 1–27.

Nair, A. A., & Yu, F. (2020). Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements. *Atmospheric Chemistry and Physics*, *20*(21), 12853–12869. https://doi.org/10.5194/acp-20-12853-2020

Nilsson, E. D., Rannik, Ü., Kumala, M., Buzorius, G., & O'Dowd, C. D. (2001). Effects of continental boundary layer evolution, convection, turbulence and entrainment on aerosol formation. *Tellus B: Chemical and Physical Meteorology*, *53*(4), 441–461. https://doi.org/10.3402/tellusb.v53i4.16617

O'Donnell, S., Akherati, A., He, Y., Hodshire, A., Shilling, J., Kuang, C., et al. (2023). Look up: Probing the vertical profile of new particle Formation and growth in the planetary boundary layer with models and observations. *Journal of Geophysical Research: Atmospheres*, *128*(3), e2022JD037525. https://doi.org/10.1029/2022jd037525

O'Dowd, C. D., Geever, M., Hill, M. K., Smith, M. H., & Jennings, S. G. (1998). New particle formation: Nucleation rates and spatial scales in the clean marine coastal environment. *Geophysical Research Letters*, *25*(10), 1661–1664. https://doi.org/10.1029/98gl01005

Oliveira, M. V. B., Wang, Y., Mehra, M., Zhang, D., Abel, S. J., & Zuidema, P. (2025). New particle formation events over the Southeast Atlantic coincide with the African biomass burning season. *Geophysical Research Letters*, *52*(8), e2024GL113235. https://doi.org/10.1029/2024gl113235

Peltola, M., Rose, C., Trueblood, J. V., Gray, S., Harvey, M., & Sellegri, K. (2023). Chemical precursors of new particle formation in coastal New Zealand. *Atmospheric Chemistry and Physics*, *23*(7), 3955–3983. https://doi.org/10.5194/acp-23-3955-2023

Platis, A., Altstädter, B., Wehner, B., Wildmann, N., Lampert, A., Hermann, M., et al. (2016). An observational case study on the influence of atmospheric boundary-layer dynamics on new particle formation. *Boundary-Layer Meteorology*, *158*(1), 67–92. https://doi.org/10.1007/s10546-015-0084-y

Qin, Y., Ye, J., Ohno, P., Liu, P., Wang, J., Fu, P., et al. (2022). Assessing the nonlinear effect of atmospheric variables on primary and oxygenated organic aerosol concentration using machine learning. *ACS Earth and Space Chemistry*, *6*(4), 1059–1066. https://doi.org/10.1021/acsearthspacechem.1c00443

Salimi, F., Clifford, S., Choy, S. L., Hussein, T., Mengersen, K., & Morawska, L. (2017). Nocturnal new particle formation events in urban environments. *Atmospheric Chemistry and Physics*, *17*(1), 521–530. https://doi.org/10.5194/acp-17-521-2017

Sellegri, K., Rose, C., Marinoni, A., Lupi, A., Wiedensohler, A., Andrade, M., et al. (2019). New particle formation: A review of ground-based observations at mountain research stations. *Atmosphere*, *10*(9), 493. https://doi.org/10.3390/atmos10090493

Shen, X., Sun, J., Ma, Q., Zhang, Y., Zhong, J., Yue, Y., et al. (2022). Long-term trend of new particle formation events in the Yangtze River Delta, China, and its influencing factors: A 7-year dataset analysis. *Science of the Total Environment*, *807*, 150783. https://doi.org/10.1016/j.scitotenv.2021.150783

Sorribas, M., Adame, J., Olmo, F., Vilaplana, J., Gil-Ojeda, M., & Alados-Arboledas, L. (2015). A long-term study of new particle formation in a coastal environment: Meteorology, gas phase and solar radiation implications. *Science of the Total Environment*, *511*, 723–737. https://doi.org/10.1016/j.scitotenv.2014.12.011

Su, P., Joutsensaari, J., Dada, L., Zaidan, M. A., Nieminen, T., Li, X., et al. (2022). New particle formation event detection with Mask R-CNN. *Atmospheric Chemistry and Physics*, *22*(2), 1293–1309. https://doi.org/10.5194/acp-22-1293-2022

Su, P., Liu, Y., Tarkoma, S., Rebeiro-Hargrave, A., Petäjä, T., Kulmala, M., & Pellikka, P. (2022a). Retrieval of multiple atmospheric environmental parameters from images with deep learning. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/lgrs.2022.3149045

U.S. Environmental Protection Agency (EPA). (2017). *National Emissions Inventory (NEI) report*. U.S. EPA. Retrieved from https://www.epa.gov/air-emissions-inventories

Venter, Z. S., Chakraborty, T., & Lee, X. (2021). Crowdsourced air temperatures contrast satellite measures of the urban heat island and its mechanisms. *Science Advances*, *7*(22), eabb9569. https://doi.org/10.1126/sciadv.abb9569

Wang, Y., Bagya Ramesh, C., Giangrande, S. E., Fast, J., Gong, X., Zhang, J., et al. (2023). Examining the vertical heterogeneity of aerosols over the Southern Great Plains. *Atmospheric Chemistry and Physics*, *23*(24), 15671–15691. https://doi.org/10.5194/acp-23-15671-2023

Xiao, Q., Zhang, J., Wang, Y., Ziemba, L. D., Crosbie, E., Winstead, E. L., et al. (2023). New particle formation in the tropical free troposphere during CAMP2Ex: Statistics and impact of emission sources, convective activity, and synoptic condition. *Atmospheric Chemistry and Physics Discussions*, 1–34.

Xiao, S., Wang, M. Y., Yao, L., Kulmala, M., Zhou, B., Yang, X., et al. (2015). Strong atmospheric new particle formation in winter in urban Shanghai, China. *Atmospheric Chemistry and Physics*, *15*(4), 1769–1781. https://doi.org/10.5194/acp-15-1769-2015

Yang, C., Dong, H., Chen, Y., Xu, L., Chen, G., Fan, X., et al. (2023). New insights on the Formation of nucleation mode particles in a coastal City based on a machine learning approach. *Environmental Science and Technology*, *58*(2), 1187–1198. https://doi.org/10.1021/acs.est.3c07042

Zaidan, M., Haapsilta, V., Relan, R., Junninen, H., Aalto, P., Canova, F., et al. (2017). Neural network classifier on time series features for predicting atmospheric particle formation days. In *The 20th international conference on nucleation and atmospheric aerosols (Report Series in Aerosol Science 200)* (pp. 687–690).

Zhang, Q., Jia, S., Yang, L., Krishnan, P., Zhou, S., Shao, M., & Wang, X. (2021). New particle formation (NPF) events in China urban clusters given by sever composite pollution background. *Chemosphere*, *262*, 127842. https://doi.org/10.1016/j.chemosphere.2020.127842

Zhang, R., Khalizov, A., Wang, L., Hu, M., & Xu, W. (2012). Nucleation and growth of nanoparticles in the atmosphere. *Chemical Reviews*, *112*(3), 1957–2011. https://doi.org/10.1021/cr2001756

Zheng, G., Wang, Y., Wood, R., Jensen, M. P., Kuang, C., McCoy, I. L., et al. (2021). New particle formation in the remote marine boundary layer. *Nature Communications*, *12*(1), 527. https://doi.org/10.1038/s41467-020-20773-1

Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., et al. (2021). Machine learning: New ideas and tools in environmental science and engineering. *Environmental Science and Technology*, *55*(19), 12741–12754. https://doi.org/10.1021/acs.est.1c01339

Zhu, Y., Li, K., Shen, Y., Gao, Y., Liu, X., Yu, Y., et al. (2019). New particle formation in the marine atmosphere during seven cruise campaigns. *Atmospheric Chemistry and Physics*, *19*(1), 89–113. https://doi.org/10.5194/acp-19-89-2019