Supporting information of

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], TC Chakraborty[2], Fan Mei[2*], Yang Wang[1*]

[1]Department of Chemical, Environmental and Materials Engineering, University of Miami, Coral Gables, FL 33146, USA

[2]Pacific Northwest National Laboratory, Richland, WA, 99352, USA

Corresponding author: Yang Wang (yangwang@miami.edu), Fan Mei (fan.mei@pnnl.gov)

Key Points:

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

The confusion matrix of classification is shown in **Fig. 2 and S1**. The confusion matrix is also known as an error matrix. It realizes the visualization of the classification performance. Each column of the matrix represents the instances in a predicted class while each row represents the instances in a ground truth class. Some commonly used metrics are adopted to evaluate the classification performance (Chen et al., 2020; Chen et al., 2024):

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = \ 2 \cdot \frac{precison \cdot \ recall}{precison + recall} \tag{4}$$

where in Eq. (1), (2), and (3), the True Positive (TP) refers to a sample x belonging to a class C that is correctly classified as C. True Negative (TN) indicates that a sample x from a 'not C' class is correctly classified as a member of the 'not C' class. The False Positive (FP) is when a sample x from a 'not C' class is incorrectly classified as class C. The False Negative (FN) describes a situation, in which a sample x from class C is misclassified as belonging to 'not C' classes. They are the four basic combinations of actual data categories and assigned categories in the classification. The values of the metrics of the classification results on the testing data set are shown in **Table 3 and S2**.

The Precision describes the exactness or quality of the method, whereas Recall can be seen as a measure of completeness or quantity. In Eq. (4), the F1-score can provide a more realistic measure of a test's performance by using both Precision and Recall. It represents the harmonic mean of the precision and recall, which ranges in the interval [0, 1].

Table S1. Accuracy results for the Random Forest model evaluating the factors influencing NPF occurrences, 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	0.97
10-Fold Cross Validation	N/A	0.97
Leave-One-Out Cross Validation	2018	0.87
	2019	0.87
	2020	0.87
	2021	0.85
	2022	0.86
	2023	0.84

Table S2. Performance metrics of the Random Forest model for mechanistic analysis of the factors influencing NPF occurrences. Class 0 represents the non-NPF event category, while Class 1 corresponds to NPF events. Metrics include precision, recall, F1-score, and overall accuracy of the model.

Class	Precision	Recall	F1-score	Accuracy
0	0.98	0.99	0.98	0.97
1	0.95	0.92	0.94	0.97

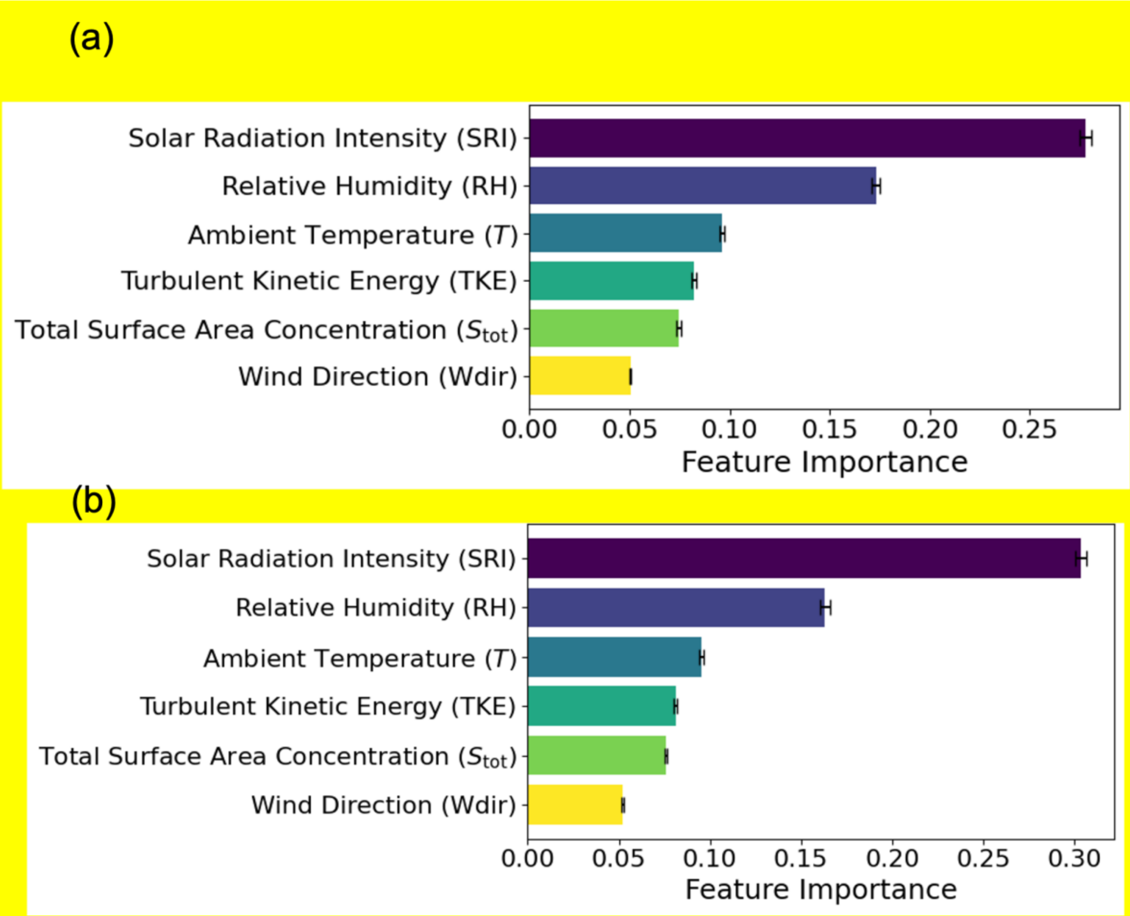

Figure S1. Top six most influential atmospheric variables ranked by feature importance using the Random Forest model. The horizontal bars represent the relative importance of each feature. (a) with SO_2 as an input variable. (b) without SO_2 as an input variable.

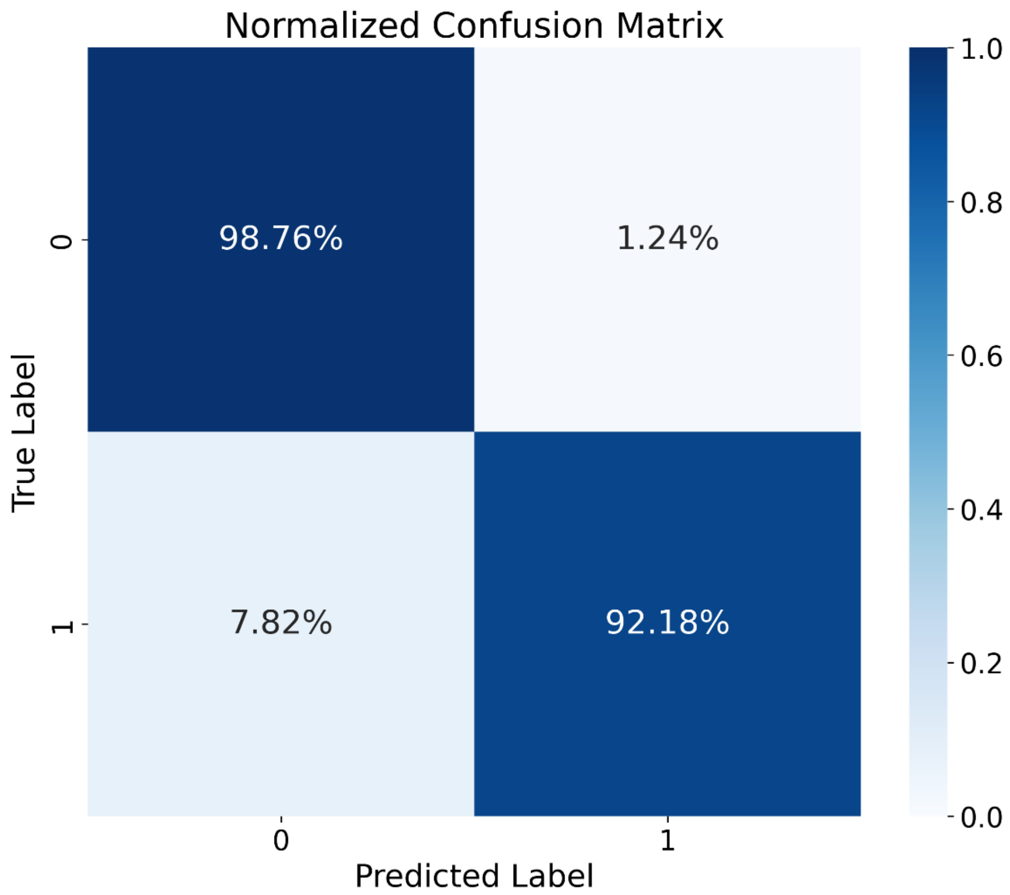

Figure S2. Normalized confusion matrix of the Random Forest model for mechanistic analysis of the factors influencing NPF occurrences. The matrix shows the percentage of correct and incorrect predictions for each class. Class 0 represents non-NPF events, and Class 1 represents NPF events.

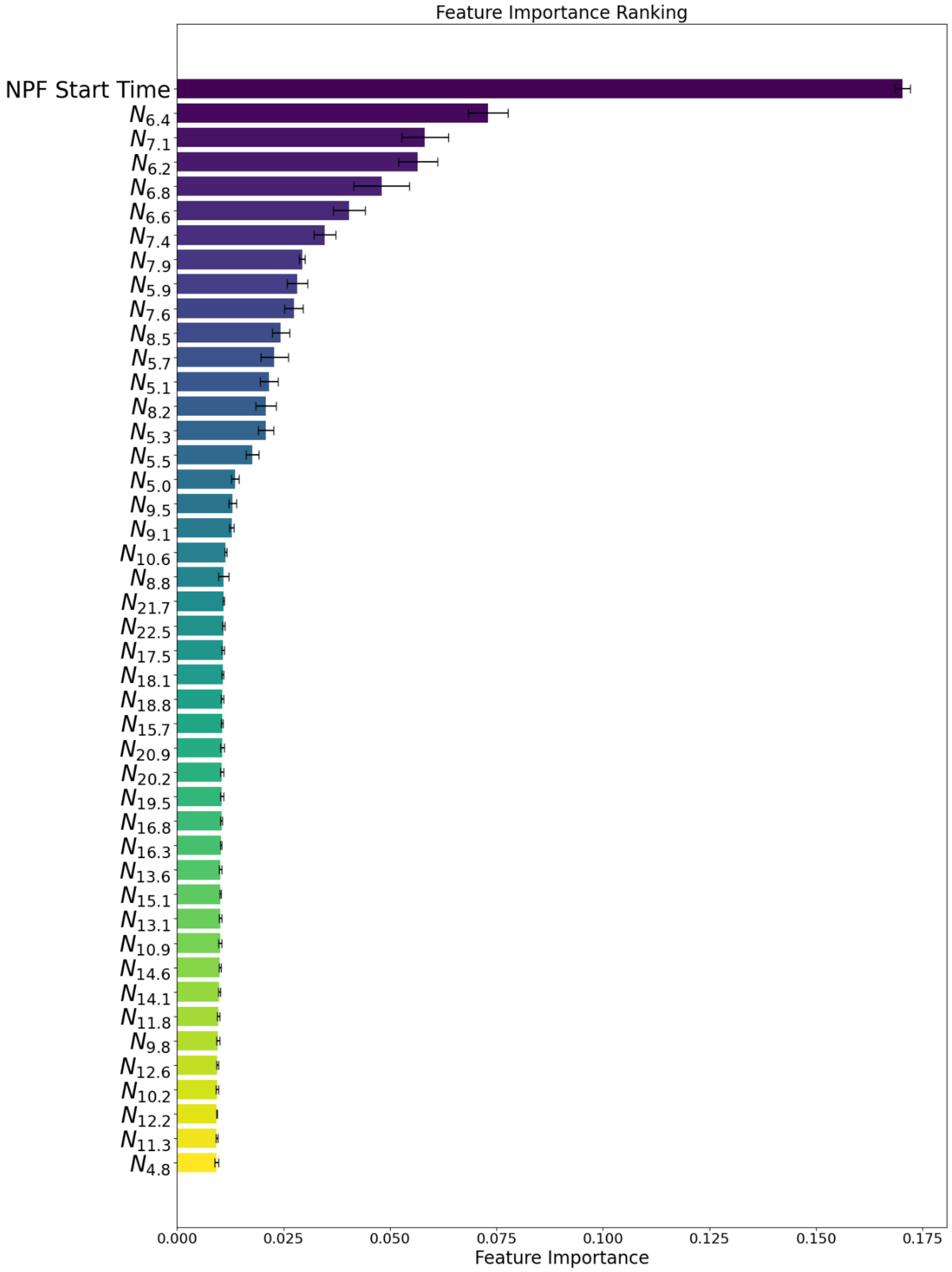

Figure S3. Feature importance ranking of the total of 45 features influencing NPF occurrence based on the Random Forest model.

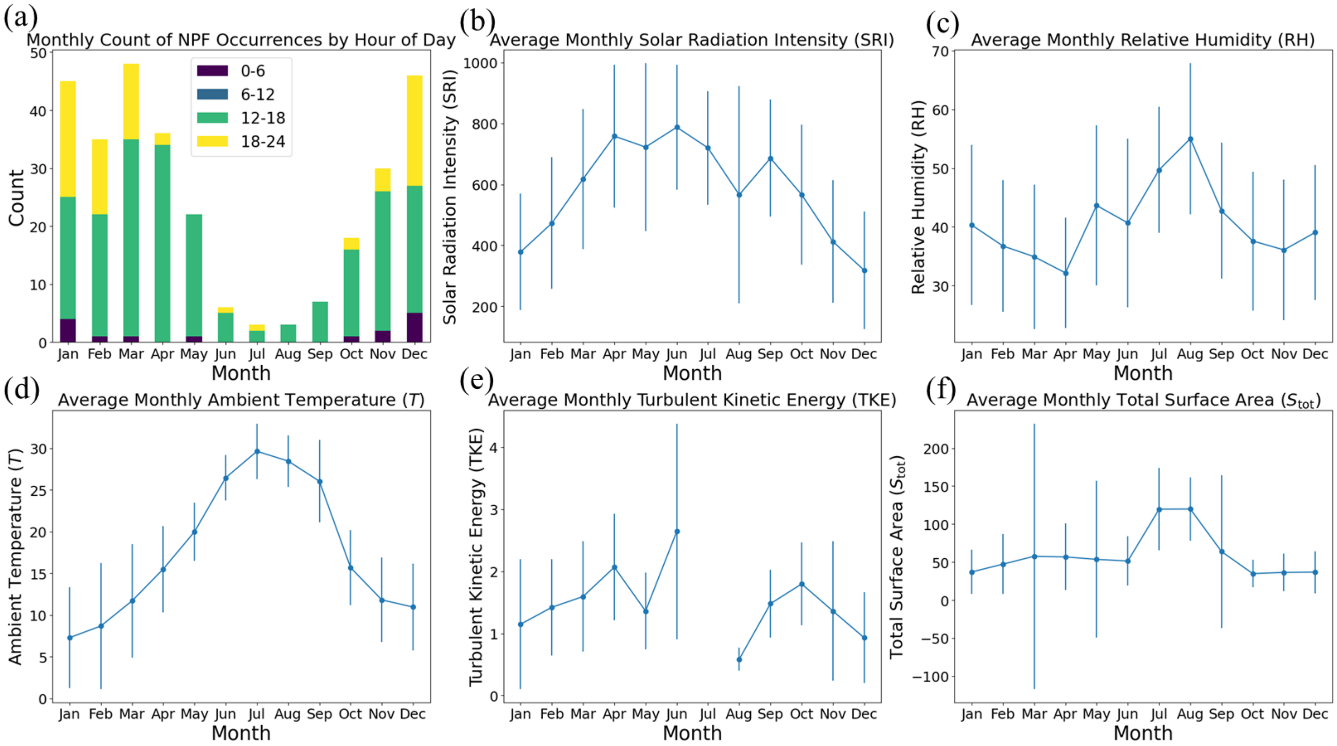

Figure S4. **Monthly analysis of NPF occurrences and atmospheric conditions. Panel (a) shows the monthly count of NPF occurrences categorized by hour of day (UTC). Panels (b) to (f) depict the average monthly trends of key atmospheric variables: solar radiation intensity (SRI), relative humidity (RH), ambient temperature (*T*), turbulent kinetic energy (TKE), and total surface area concentration (S_{tot}). Error bars represent variability in each parameter across the months.**

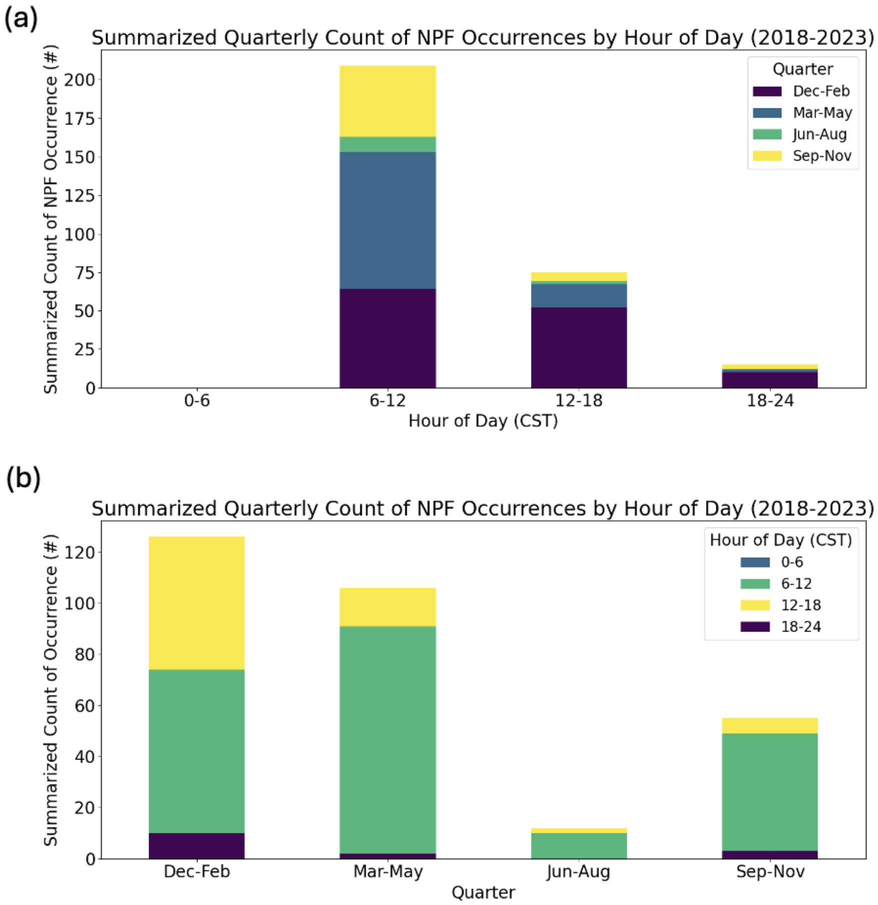

Figure S5. **Temporal distribution of NPF occurrences: (a) summarized quarterly count of NPF occurrences by hour of the day across four quarters: Dec-Feb, Mar-May, Jun-Aug, and Sep-Nov, (b) summarized quarterly count of NPF occurrences by hour of the day, grouped by quarter, illustrating the distribution of NPF events within each time interval.**

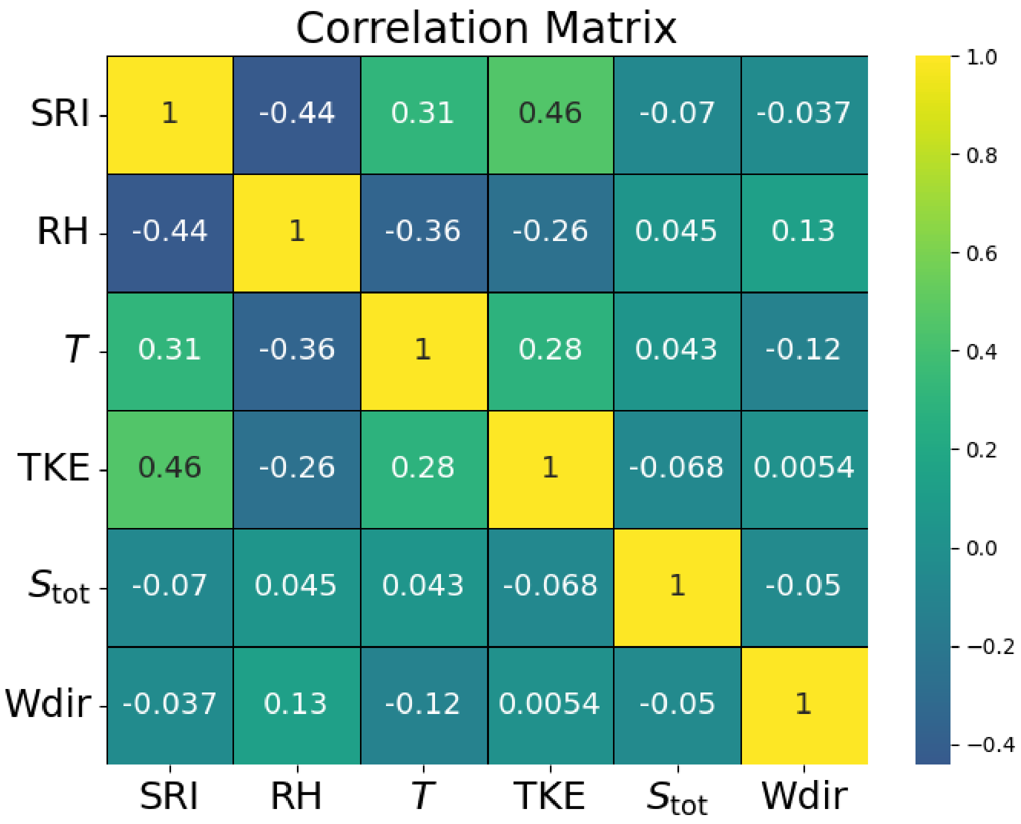

Figure S6. Correlation matrix displaying the relationships between key atmospheric variables: solar radiation intensity (SRI), relative humidity (RH), ambient temperature (*T*), turbulent kinetic energy (TKE), total surface area concentration (S_{tot}), and wind direction (Wdir).

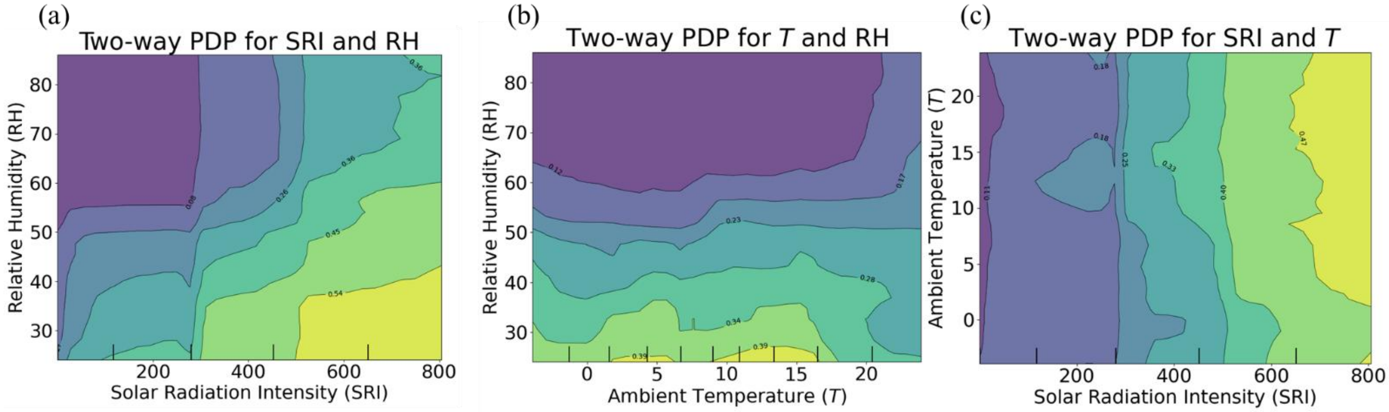

Figure S7. Two-way Partial Dependence Plots (PDPs) show the interaction effects of solar radiation intensity (SRI), relative humidity (RH), and ambient temperature (*T*) on the probability of NPF events. The first plot illustrates the interaction between SRI and RH, the second shows *T* and RH, and the third depicts the relationship between SRI and *T*.