



*Supplement of*

## **A 23-year nationwide study revealing aerosol-driven light rain shifts in China's emission control era**

**Rou Zhang et al.**

*Correspondence to:* Fang Zhang (zhangfang2021@hit.edu.cn)

The copyright of individual parts of the supplement might differ from the article licence.

1 **This PDF file includes:**

2 Method descriptions

3 Figs. S1 to S17

4 References

5 **Sect. S1**

6 **Machine learning methods**

7 The XGBoost (eXtreme Gradient Boosting) model is an advanced machine learning algorithm that  
8 has gained significant popularity and achieved state-of-the-art results in various predictive modeling tasks  
9 (Chen and Guestrin, 2016). It belongs to the family of gradient boosting algorithms and is known for its  
10 efficiency, flexibility, and high performance. XGBoost is designed to handle both classification and  
11 regression problems. It works by sequentially adding weak prediction models, typically decision trees, to  
12 an ensemble in a process known as boosting. Each subsequent model is built to correct the mistakes made  
13 by the previous models, gradually improving the overall predictive accuracy. What sets XGBoost apart  
14 is its focus on optimization and regularization techniques. It incorporates a regularized objective function  
15 that combines a loss function and a penalty term to control model complexity and prevent overfitting (Gui  
16 et al., 2020; Si and Du, 2020; Wong et al., 2021).

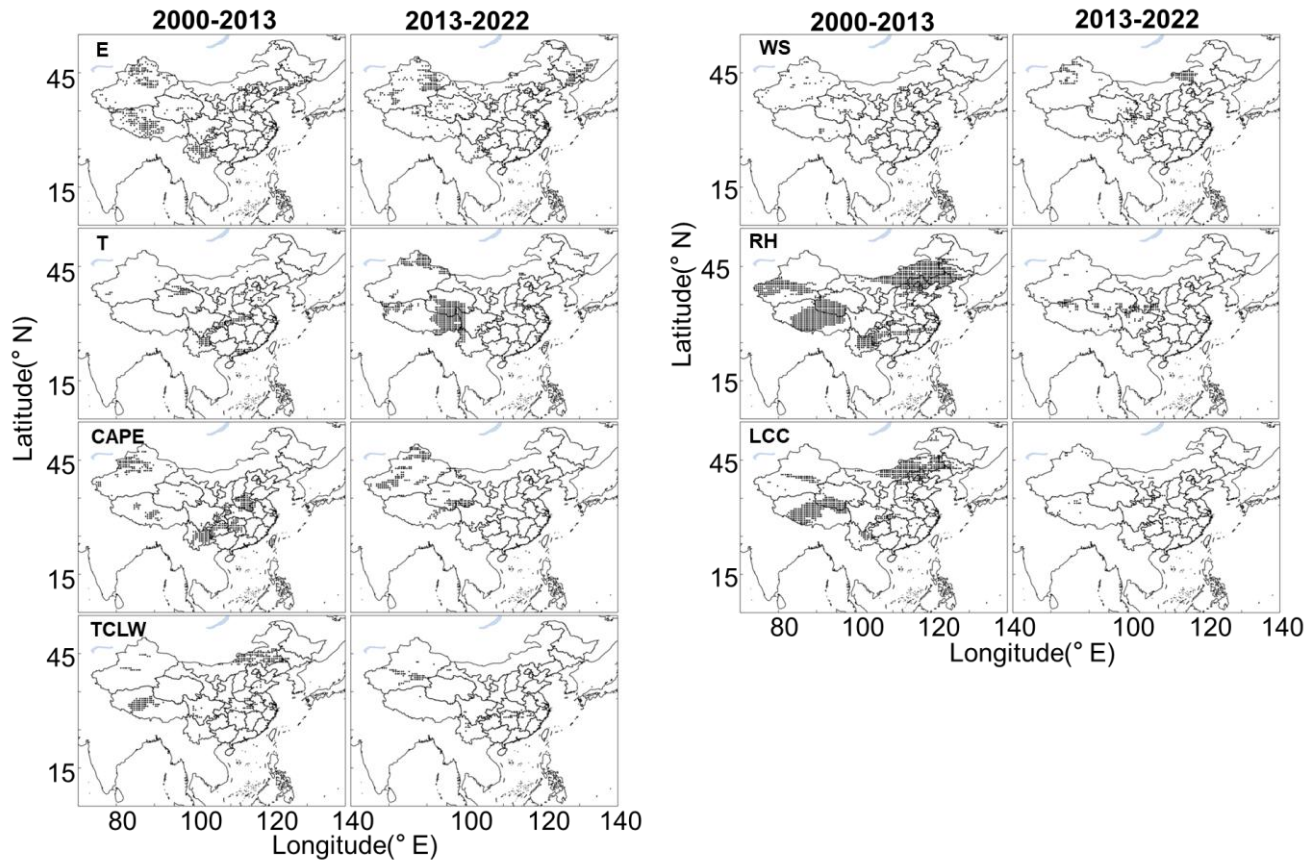
17 Building upon this robust theoretical foundation, we implemented the XGBoost model with a  
18 rigorous hyperparameter tuning strategy to ensure the reliability of our results. The hyperparameters of  
19 the XGBoost model were meticulously optimized to ensure maximum predictive performance and the  
20 reliability of the subsequent feature importance analysis. For each of the six study regions and both time

21 periods, we conducted an independent hyperparameter tuning process due to variations in local sample  
22 characteristics.

23 The optimization was performed using RandomizedSearchCV with 5-fold cross-validation,  
24 employing R2 as the scoring metric to identify the parameter set that yielded the highest and most  
25 generalizable model performance. We defined a comprehensive search space for the key hyperparameters  
26 to balance model complexity and prevent overfitting:

- 27 1. max\_depth: [6, 8, 10] (Controls the depth of trees, balancing complexity)
- 28 2. learning\_rate: [0.01, 0.05, 0.1, 0.2] (Shrinks the contribution of each tree for smoother  
29 convergence)
- 30 3. n\_estimators: [100, 150, 200, 250] (Number of boosting rounds)
- 31 4. subsample: [0.6, 0.8, 1.0] (Fraction of samples used for fitting each tree)
- 32 5. colsample\_bytree: [0.6, 0.8, 1.0] (Fraction of features available for each tree)

33 This rigorous approach ensured that the final models used for interpreting factor contributions were  
34 neither underfitted nor overfitted, but were optimally calibrated for each unique regional dataset. The  
35 consistently high model performance (as shown in Fig. S3) validates the effectiveness of this tuning  
36 strategy.



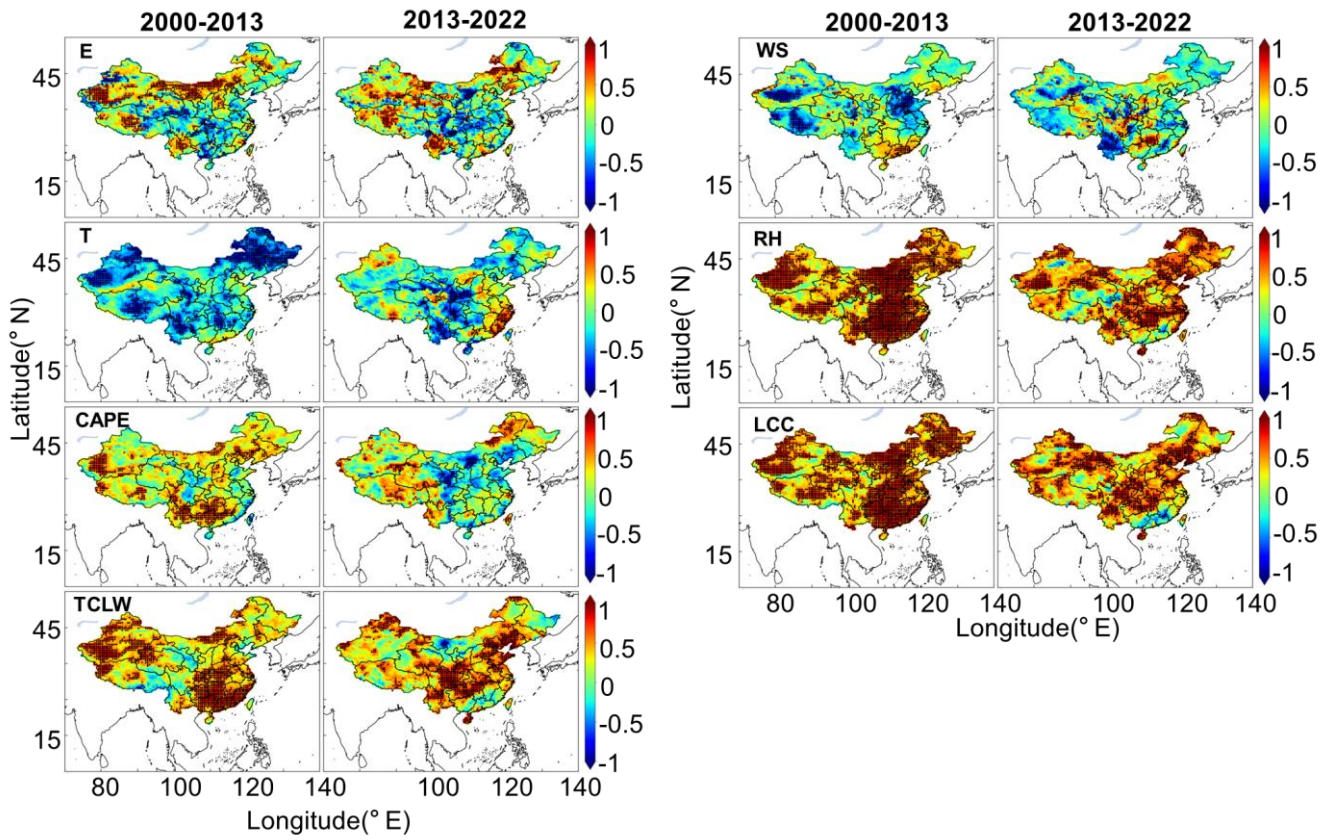
38

39

**Fig. S1.** The significance of the fitted variation trends of other influencing factors in China during the two

40

periods of 2000 - 2013 and 2013 - 2022 (black dots indicate passing the 95% significance test).



41

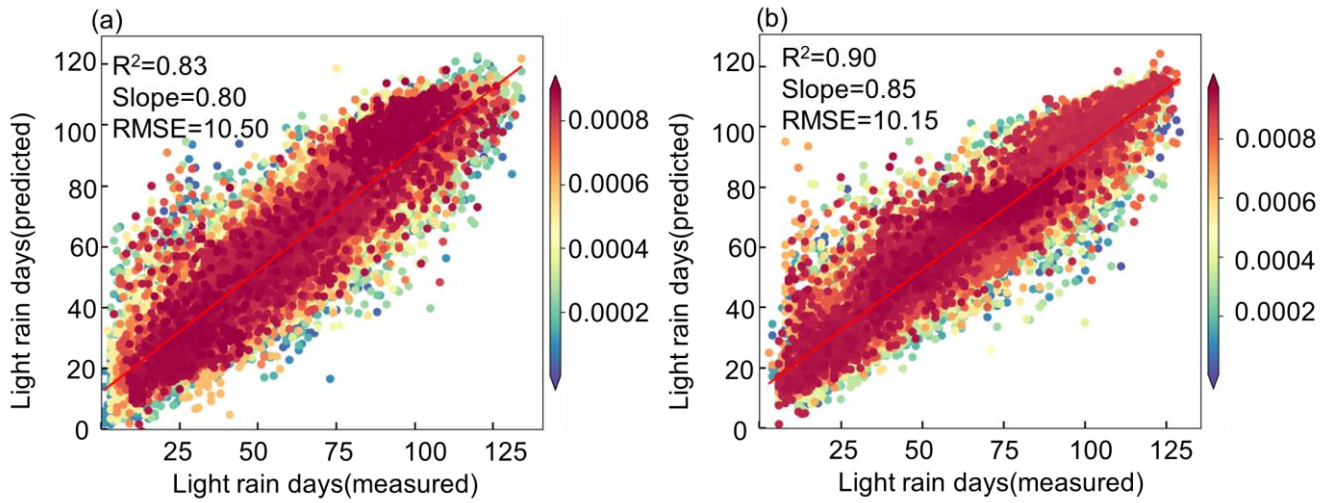
42

**Fig. S2.** correlation of between Meteorological factors and light rain frequency in 2000 - 2013 and 2013

43

- 2022.

44

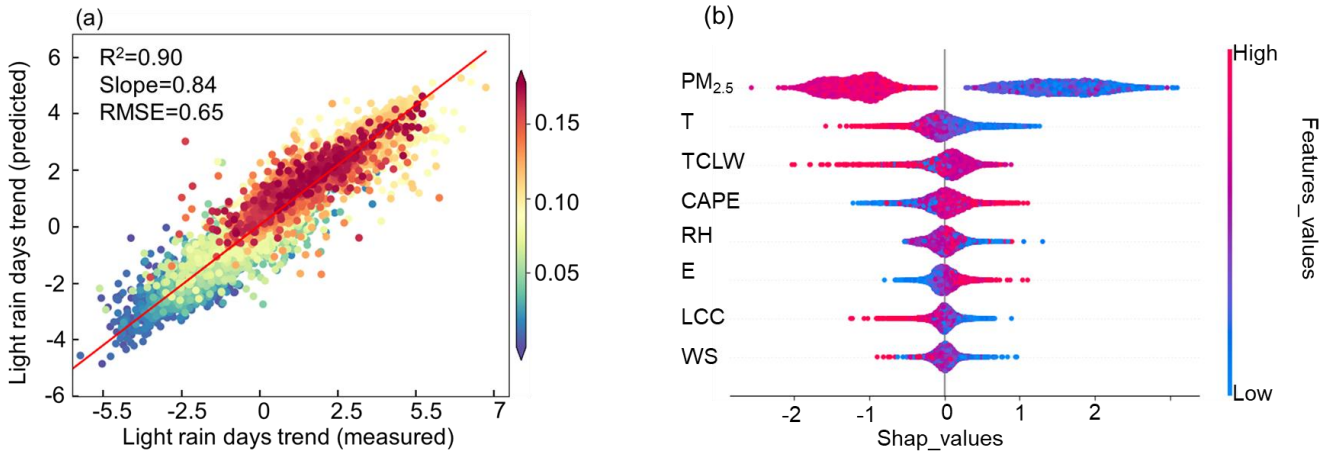


45

46 **Fig. S3.** The predicted values of light rain days during (a) 2000 - 2013 and (b) 2013 - 2022 by the  
 47 XGBoost method.

48

49

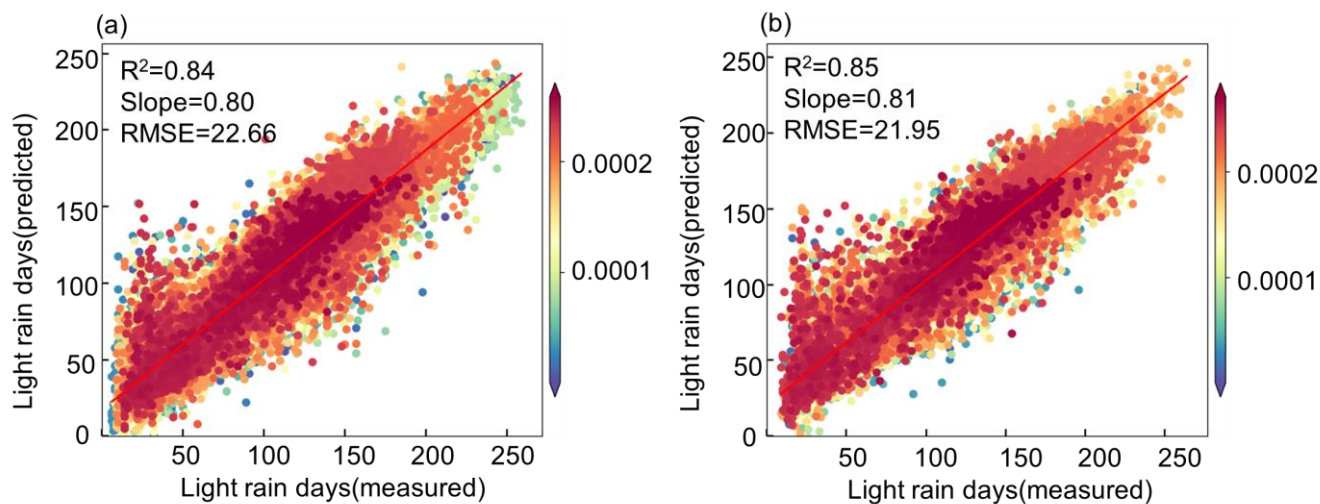


50

51 **Fig. S4.** The predicted values of light rain days trend by the XGBoost method (a) and SHAP summary  
 52 plot (b).

53

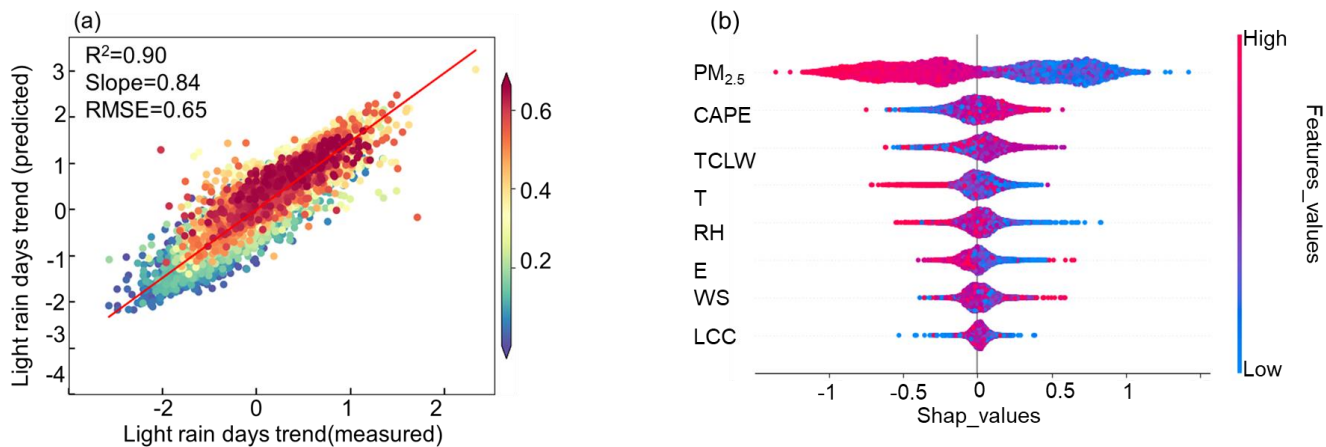
54



55

56 **Fig. S5.** The predicted values of light rain days in the warm season (Jun.-Oct.) during (a) 2000 - 2013  
57 and (b) 2013 - 2022 by the XGBoost method.

58

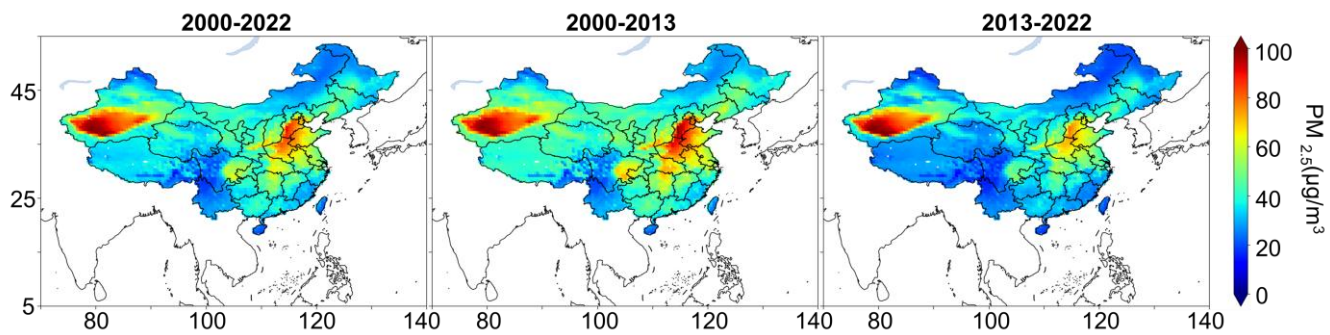


59

60 **Fig. S6.** The predicted values of light rain days trend in the warm season (Jun.-Oct.) by the XGBoost  
61 method (a) and SHAP summary plot (b).

62

63

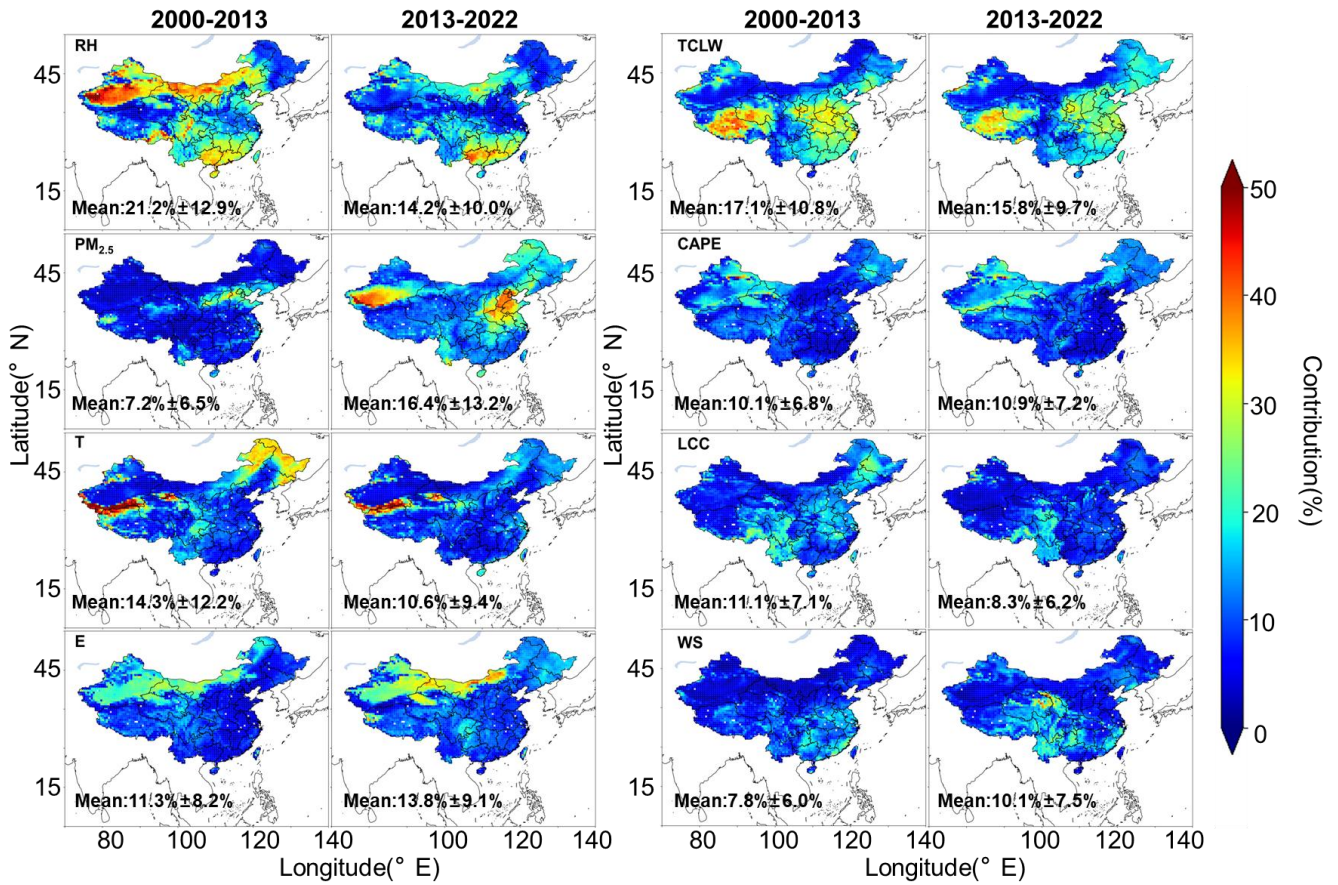


64

65 **Fig. S7.** Spatial distribution of average PM<sub>2.5</sub> in 2000 - 2022,2000 - 2013 and 2013 - 2022 respectively.

66

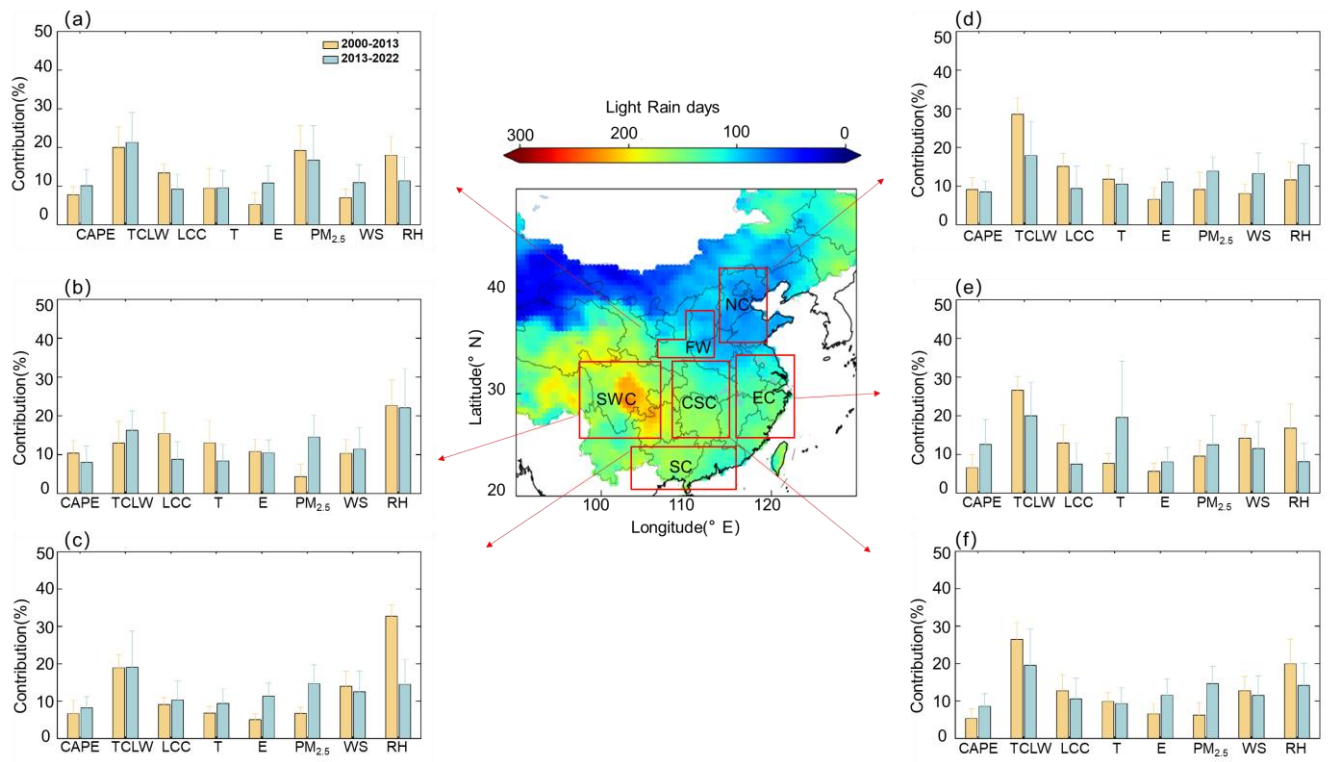
67



68

69 **Fig. S8.** The relative contribution of relative humidity (RH), PM<sub>2.5</sub> mass concentration, temperature (T),  
 70 evaporation (E), total column liquid water (TCLW), CAPE, low cloud cover (LCC) and wind speed (WS)  
 71 to light rain days in the warm season during 2000 - 2013 and 2013 - 2022.

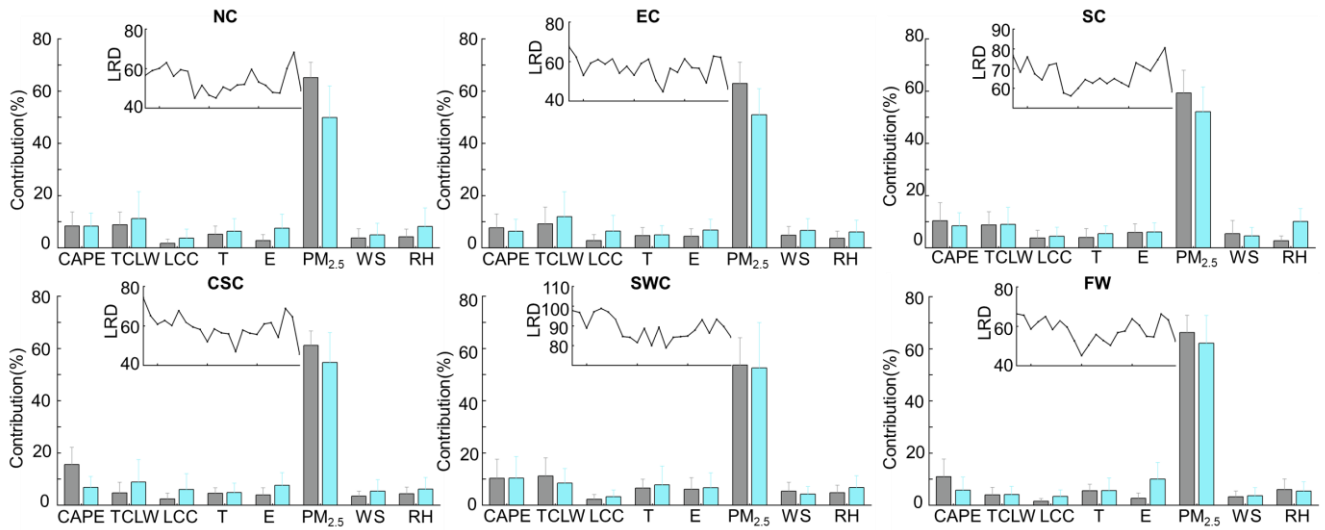
72



73

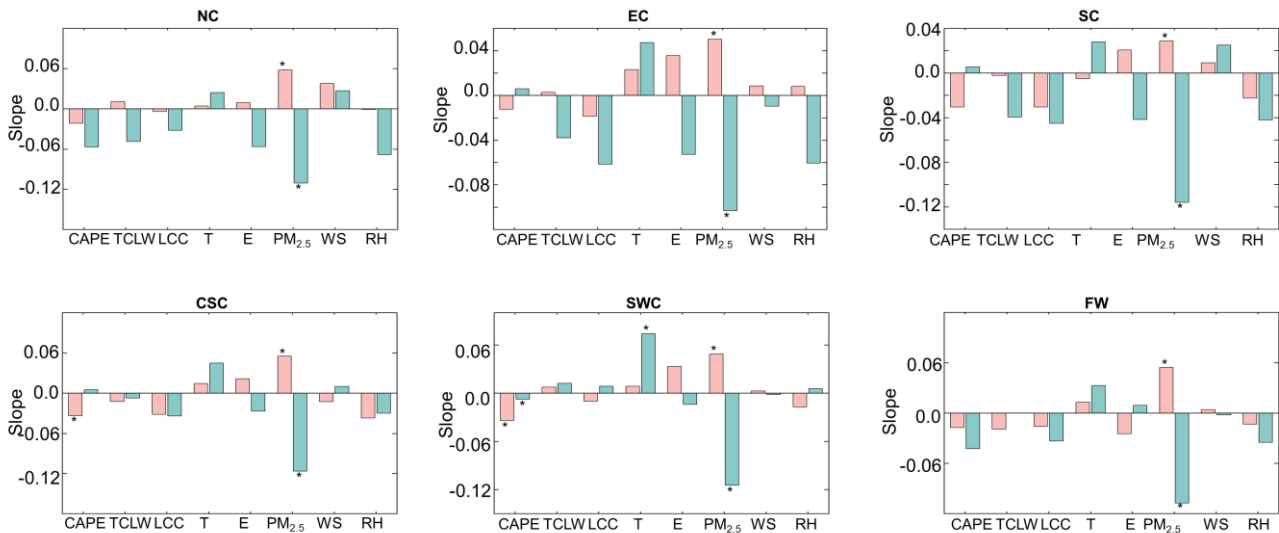
74 **Fig. S9.** Comparison of the contribution of individual factors to light rain days in the warm season over  
 75 2000 - 2013 and 2013 - 2022 in the selected six regions of China.

76



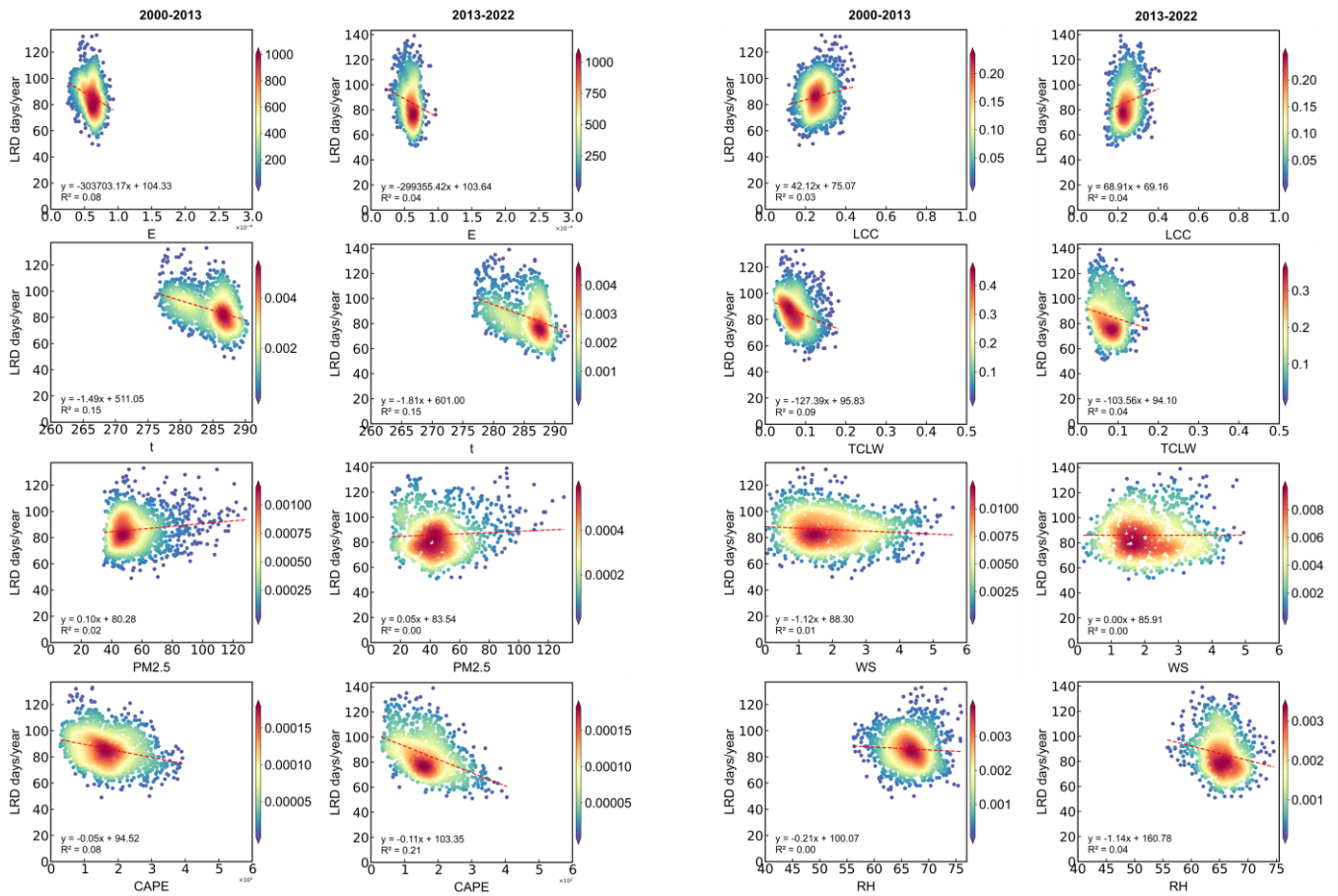
77

78 **Fig. S10.** Contribution of variations of each individual factor to the long-term trends of light rain days of  
 79 warm season in the selected six regions of China. The small graphs embedded in the middle represent the  
 80 average trends of light rain days of warm season over 2000 - 2022 in the six regions.

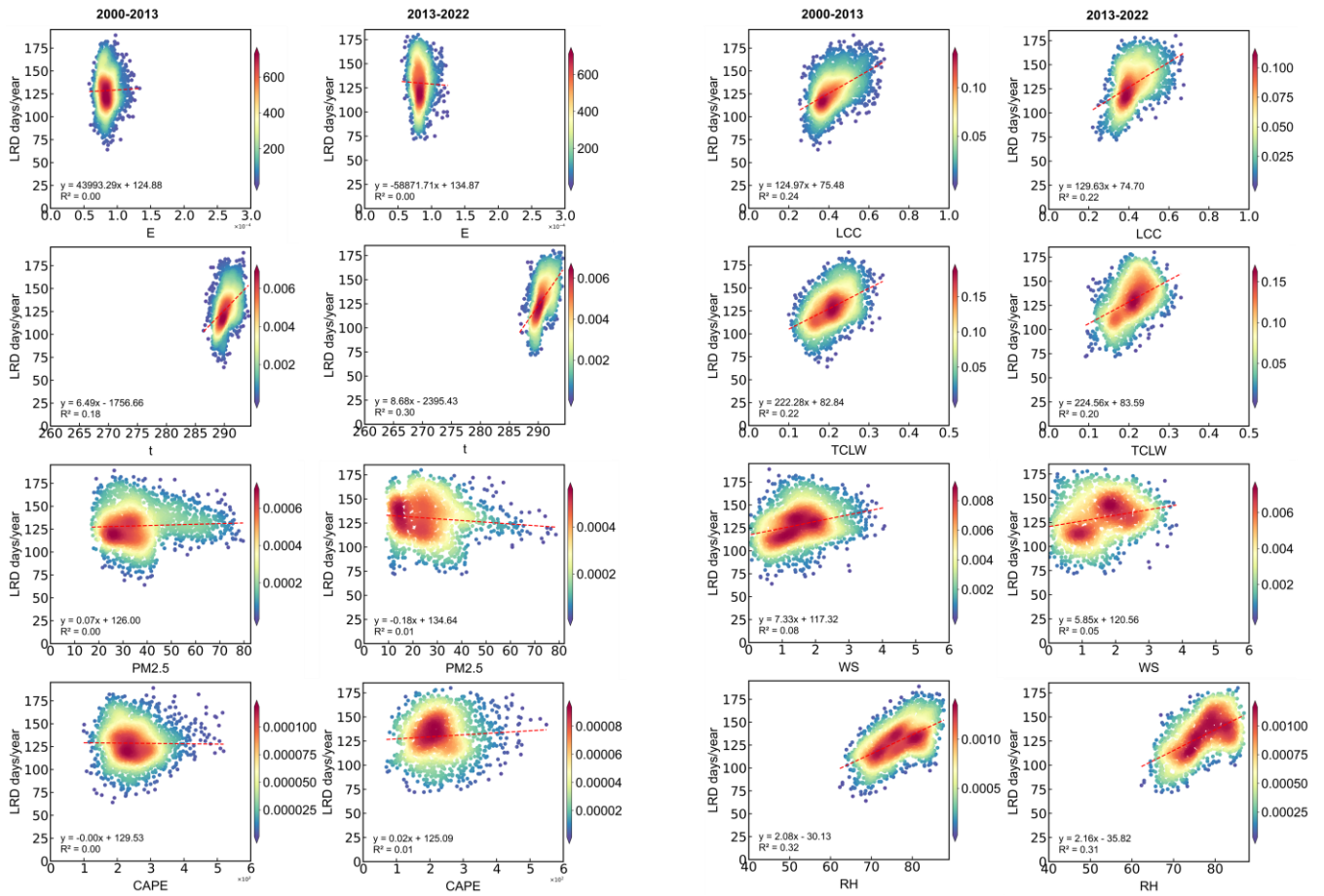


81

82 **Fig. S11.** The long-term trends of each individual factor of warm season in the selected six regions of  
 83 China. The black asterisk (\*) represent the trend pass the 95% significance test.



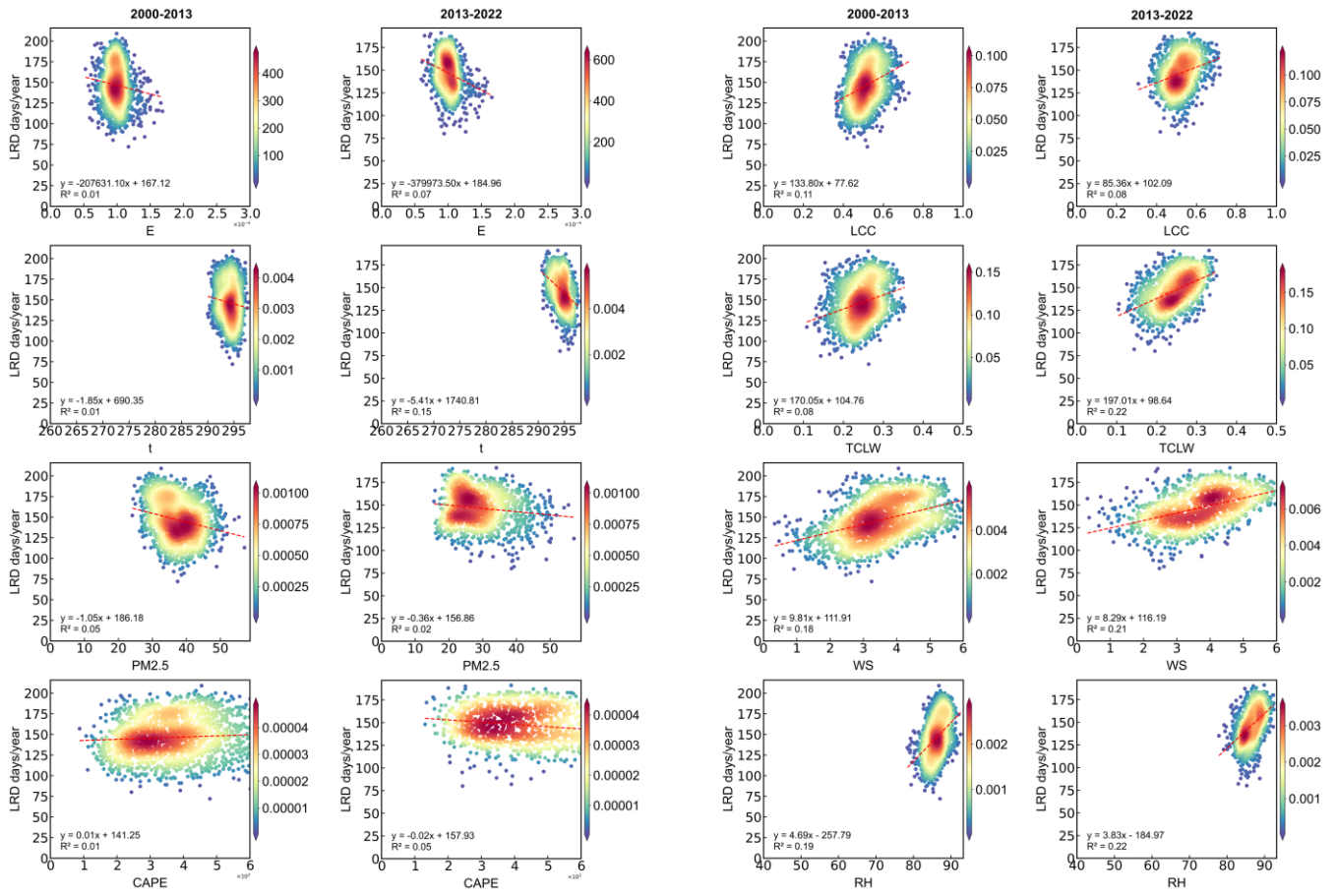
86 **Fig. S12.** Scatter plots showing the relationships between various factors and the number of light rain  
 87 days in NC.



88

89 **Fig. S13.** Scatter plots showing the relationships between various factors and the number of light rain  
 90 days in EC.

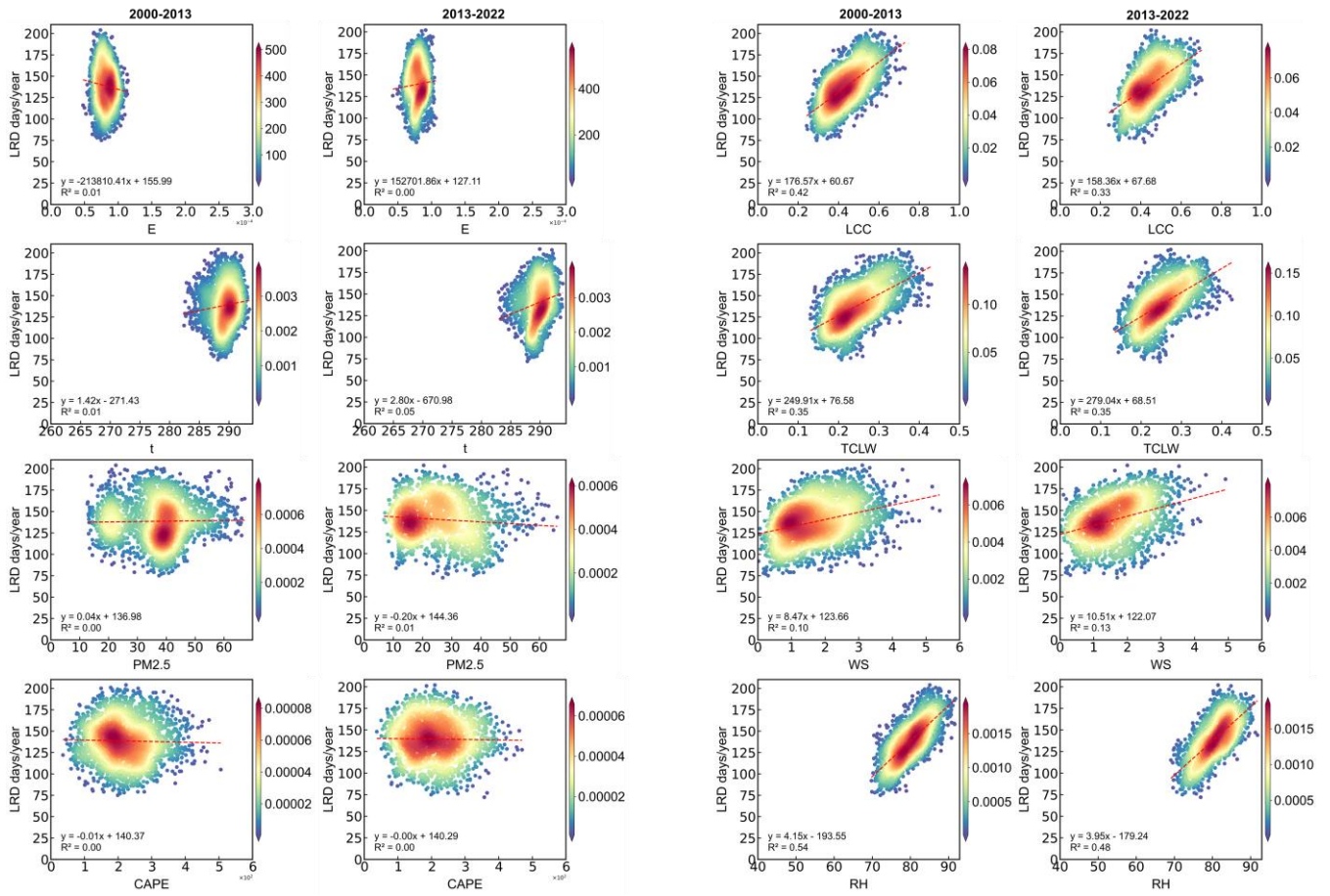
91



92

93 **Fig. S14.** Scatter plots showing the relationships between various factors and the number of light rain  
 94 days in SC.

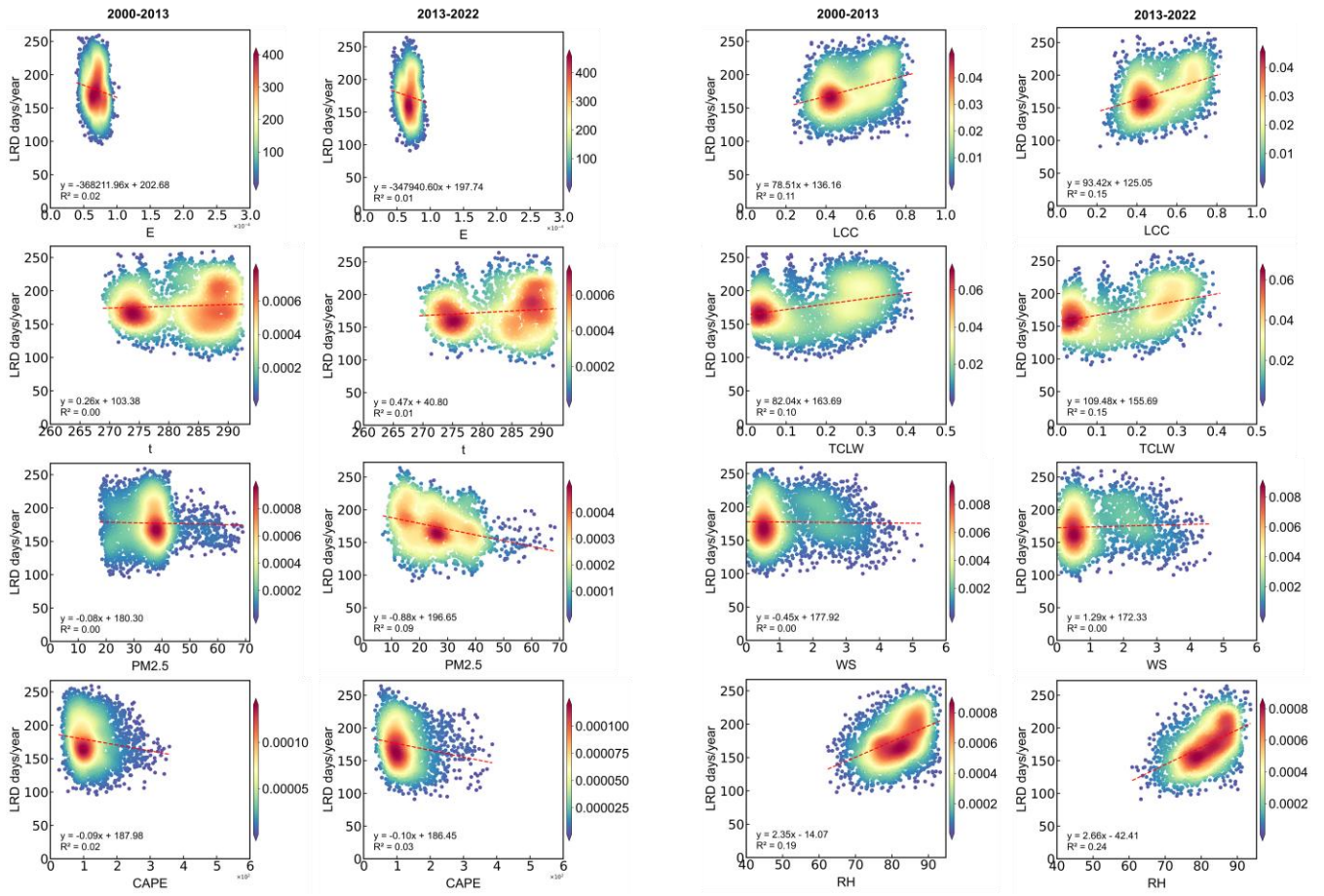
95



96

97 **Fig. S15.** Scatter plots showing the relationships between various factors and the number of light rain  
 98 days in CSC.

99



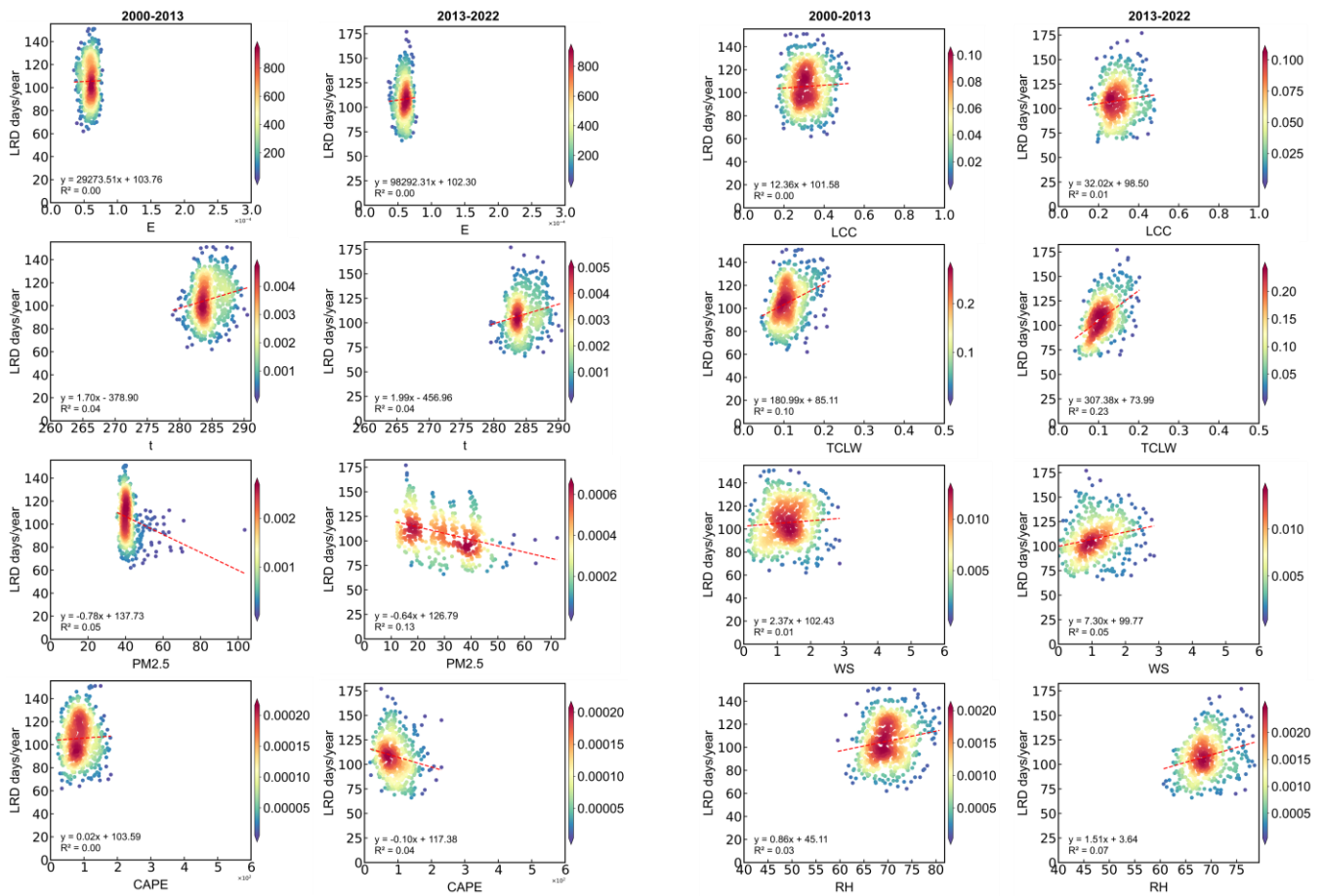
100

101

**Fig. S16.** Scatter plots showing the relationships between various factors and the number of light rain days in SWC.

102

103



104

105

**Fig. S17.** Scatter plots showing the relationships between various factors and the number of light rain days in FW.

106

107

108

## References

109

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>

110

111

Gui, K., Che, H., Zeng, Z., Wang, Y., Zhai, S., Wang, Z., Luo, M., Zhang, L., Liao, T., Zhao, H., Li, L., Zheng, Y., Zhang, X., 2020. Construction of a virtual PM2.5 observation network in China based on high-density surface meteorological observations using the Extreme Gradient Boosting model. *Environment International* 141, 105801. <https://doi.org/10.1016/j.envint.2020.105801>

112

113

114

115

Si, M., Du, K., 2020. Development of a predictive emissions model using a gradient boosting machine learning method. *Environmental Technology & Innovation* 20, 101028. <https://doi.org/10.1016/j.eti.2020.101028>

116

117

118 Wong, P., Lee, H., Chen, Y., Zeng, Y., Chern, Y., Chen, N., Candice Lung, S.C., Su, H., Wu, C., 2021. Using a land use  
119 regression model with machine learning to estimate ground level PM2.5. *Environmental Pollution* 277, 116846.  
120 <https://doi.org/10.1016/j.envpol.2021.116846>  
121