,

Supplement for "On the Prediction of Aerosol-Cloud Interactions within a Data-driven Framework"

Xiang-Yu Li,[1, *] Hailong Wang,[1, †] TC Chakraborty,[1] Armin Sorooshian,[2, 3] Luke
D. Ziemba,[4] Christiane Voigt,[5] Kenneth Lee Thornhill,[4] and Emma Yuan[1, 6]

[1]*Pacific Northwest National Laboratory, Richland, WA, USA*

[2]*Department of Chemical and Environmental Engineering and Department of Hydrology and Atmospheric Sciences,*
University of Arizona, Tucson, AZ, United States

[3]*University of Arizona, Department of Hydrology and Atmospheric Sciences, Tucson, AZ, United States*

[4]*NASA Langley Research Center, Hampton, VA, United States*

[5]*Institut für Physik der Atmosphäre, Deutsches Zentrum für Luft- und Raumfahrt (DLR), Oberpfaffenhofen,*
Germany, and Institute for Physics of the Atmosphere, Johannes Gutenberg-University Mainz, Germany

[6]*Hanford High School, Richland, WA, United States*

Contents
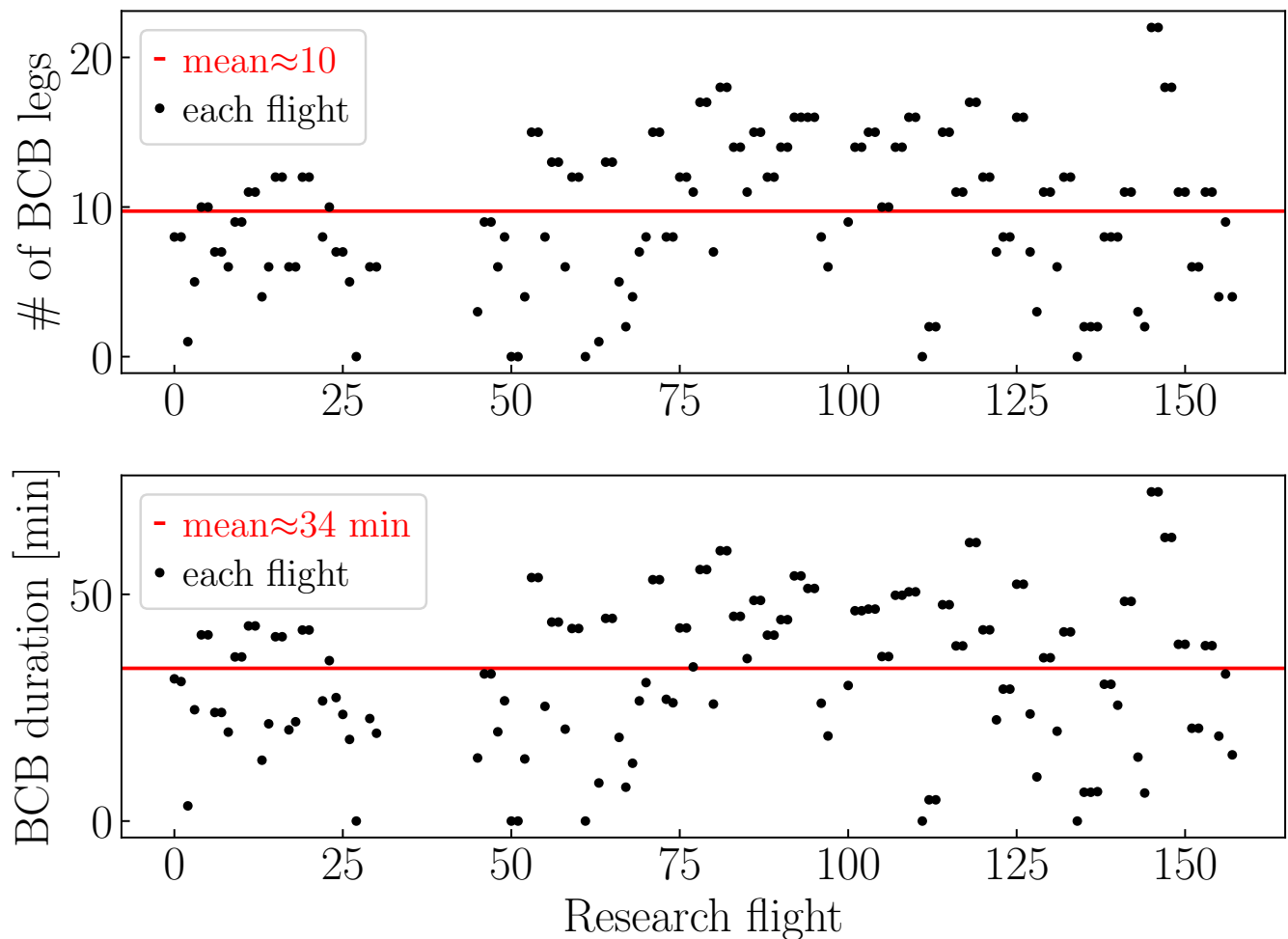
---

[*] xiangyu.li@pnnl.gov

[†] hailong.wang@pnnl.gov

FIG. S1: Number and the total duration of BCB flight legs for each flight. The mean duration for each BCB leg of a flight is $34/10 \approx 3$ min.

I. Limitations to the ACTIVATE dataset

There are limitations to the ACTIVATE dataset, which are generally innate to in-situ aircraft datasets, including how the sub-cloud measurements are not simultaneously collected directly underneath in-cloud measurements but are slightly offset in time and space. Moreover, there are instrument uncertainties (see Table S1) that potentially contribute to biases albeit state-of-the-art instruments were deployed in the ACTIVATE field campaign. Although ACTIVATE collected the largest in-situ dataset with spatially coordinated aircraft to our knowledge, statistics and spatiotemporal coverage (3 years over 25°50°N, 60°85°W) can still improve.

II. Large-eddy simulation

A. Numerical experiments setup

We use the Weather Research and Forecasting (WRF) model [1] in its LES configuration (WRF-LES) with doubly periodic boundary conditions [2]. The LES domain has a lateral size of $L_x = L_y = 20$ km (60 km) with a grid spacing of $dx = dy = 100$ m (300 m) for the cumulus (cold-air outbreak) cases and a vertical extent of $z_{\text{top}} = 7$ km with 153 vertical layers. Initial profiles of temperature, humidity, and horizontal wind components and the time-varying

TABLE S1: List of data and the corresponding instruments used in this study. n/a – not applicable.

Instruments	quantities	resolution (uncertainty)	frequency (Hz)
FCDP	cloud droplet diameter	$3 - 50\,\mu m$ ($\leq 20\%$)	1
2DS	cloud droplet diameter	$11.4 - 1464.9\,\mu m$ ($\leq 20\%$)	1
Five-port pressure system (TAMMS)	u, v, w	n/a($w: 10\mathrm{cm\,s^{-1}}; u, v: 50\mathrm{cm\,s^{-1}}$)	20
Rosemount 102 sensor	T	n/a($0.5°$)	1
Diode laser hygrometer (DLH)	q_v	n/a (5% or 0.1 ppmv)	< 0.05
AMS	non-refractory aerosol mass concentration	60-600 nm (50%)	1
SMPS	aerosol particles with $d \leq 100$ nm	$\leq \pm 10\% - 20\%$	1/60
LAS	aerosol particles with $d \geq 100$ nm	$\leq \pm 10\% - 20\%$	1

TABLE S2: Scores for different average window size of data. $N_c = \mathcal{G}(m_\mathcal{X}, N_a, w', u', v', T, q_v, \boldsymbol{x}, \theta_z)$. Comparison between the prediction and observation is shown in Figure S4.

Window size	r^2 training score	r^2 validation score	OOB score
1	0.95	0.69	0.68
2	0.97	0.81	0.80
5	0.99	0.93	0.93
10	1.00	0.98	0.98
20	1.00	0.99	0.99
50	1.00	1.00	1.00

large-scale forcings and surface turbulent heat fluxes are obtained from the fifth generation of European Centre for Medium-Range Weather Forecasts's Integrated Forecast System (ERA5) reanalysis. We use the two-moment Morrison cloud microphysics scheme [3] with prescribed aerosol size modes and hygroscopicity derived from the ACTIVATE campaign measurements. All simulations start at 06:00 UTC and end at 21:00 UTC with a fixed time step of 1 s. The two cold-air outbreak cases on 28 February 2020 and 1 March 2020 are characterized by a well-defined boundary layer while the two precipitating summertime cumulus cases on 02 and 07 June 2021 are characterized by strong spatiotemporal variations of cloud top height. We refer readers to Li *et al.* [4] and Li *et al.* [5] for details of the simulations used in the present study.

B.   LES validation data for the observation-RFM emulator

To use the observation-RFM as an emulator for the LES N_c , we prepared the input data $(N_a, T, q_v, \text{and } w')$ from LES. These predictors are chosen for the following reasons: 1. they represent our current understanding of physical processes (e.g., aerosol activation and condensation and collision-coalescence of cloud droplets) governing the ACI; 2. the RFM can successfully predict N_c use these independent predictors as will be discussed in section 3.1; 3. to apply the observation-RFM emulator to LES, the same variables from LES are required as input. The WRF-LES used in this study does not have prognostic aerosols, i.e., $m_\mathcal{X}$. The predicted N_c from the observation-RFM emulator will be compared to the observation. This requires a point-to-point data comparison between LES and FCDP. The same threshold values of LWC $= 0.02\,\mathrm{g\,m^{-3}}$, $3.5\,\mu\mathrm{m} \leq d_{\mathrm{eff}} \leq 50\,\mu\mathrm{m}$, and $N_c = 20\,\mathrm{cm^{-3}}$ as in section 2.1 are applied to LES data to define clouds. The measurement was taken during 16:00-17:00 UTC on 28 February 2020 and 15:00-16:00 UTC on 1 March 2020 and 19:00-20:00 UTC on both 02 and 07 June 2021. The LES data are based on three snapshots 30 minutes apart. We only use ACB and BCT flight legs. The data are binned at those heights of flight legs with a residual range of $\pm 40\,\mathrm{m}$ such that at least one model layer is counted at the height of each flight leg. This also avoids double counting. At each height, we randomly draw the same number of data points from the LES domain as in the FCDP field measurements. A single random sampling is representative as shown in Figure S7.

III.   Results from the XGBoost model
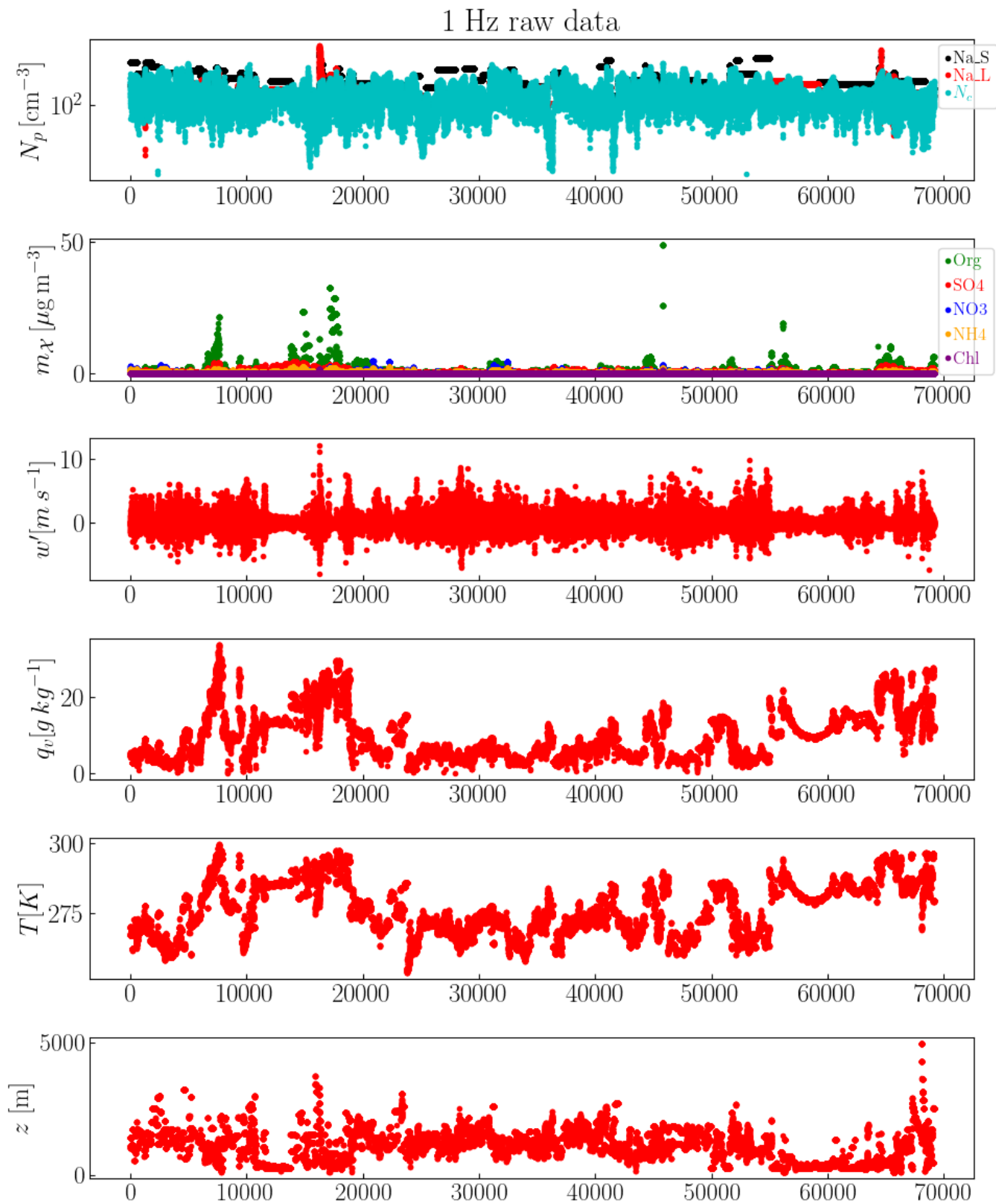
IV.   Explainable artificial intelligence

FIG. S2: 1 Hz raw data of assumed physically important quantities. The sample size is 69159 with a sampling rate of 1 Hz, which is from all the ACTIVATE flights.
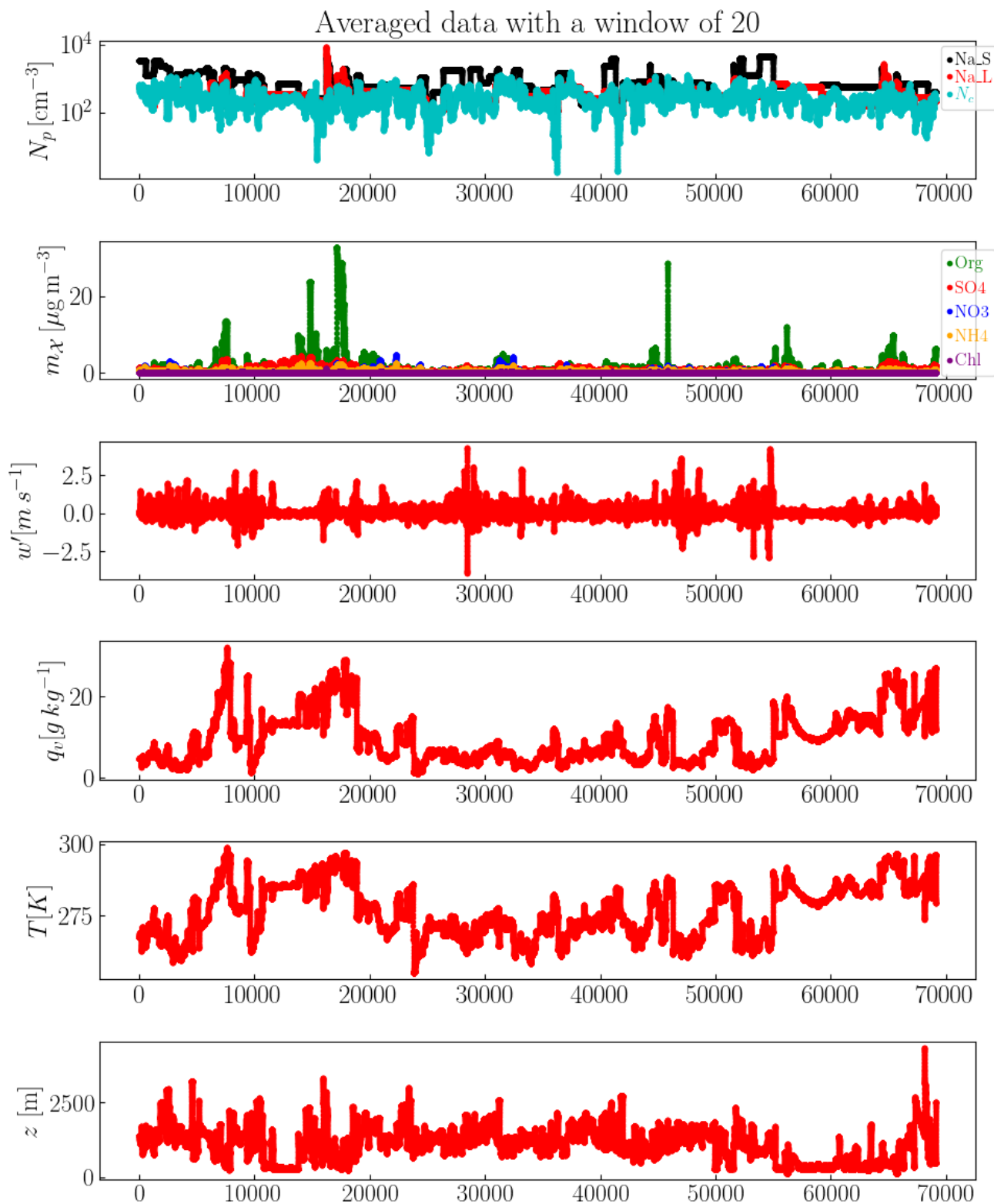
FIG. S3: Averaged data with a window size of 20 (data points) from the raw data shown in Figure S2 as the input dataset for the RFM algorithm.
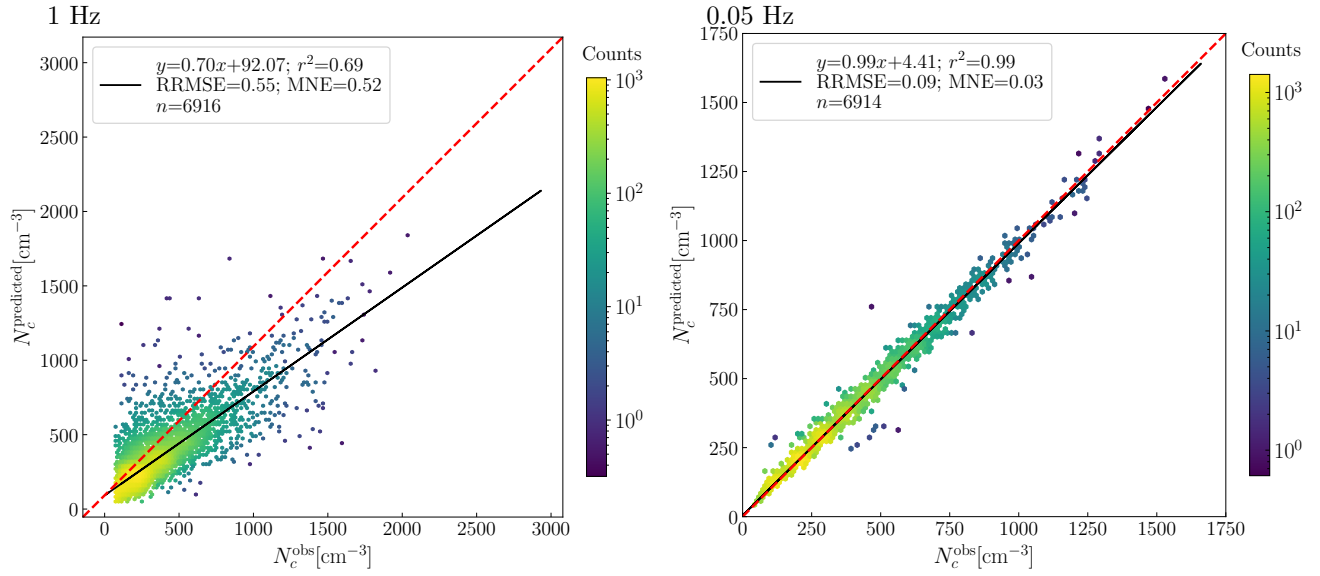
FIG. S4: Comparison between N_c^{obs} and $N_c^{\mathrm{predicted}}$ for different average window size as shown in Table S2 for the full dataset ($N_c = \mathcal{G}(m_{\mathcal{X}}, N_a, w', u', v', T, q_v, \boldsymbol{x}, \theta_z)$).

FIG. S5: r^2 of 1000 random realizations for $0.0005N$ in Figure 4(a). Note that the samples are drawn from the averaged (window size 20) dataset.

FIG. S6: Same as Figure 2 but with the same predictors as in Figure 1(a).

FIG. S7: r^2 between the FCDP measurements and LES. As the LES data were sampled randomly at each height, we test how the random selection affects the r^2 by performing 1000 realizations for each case.
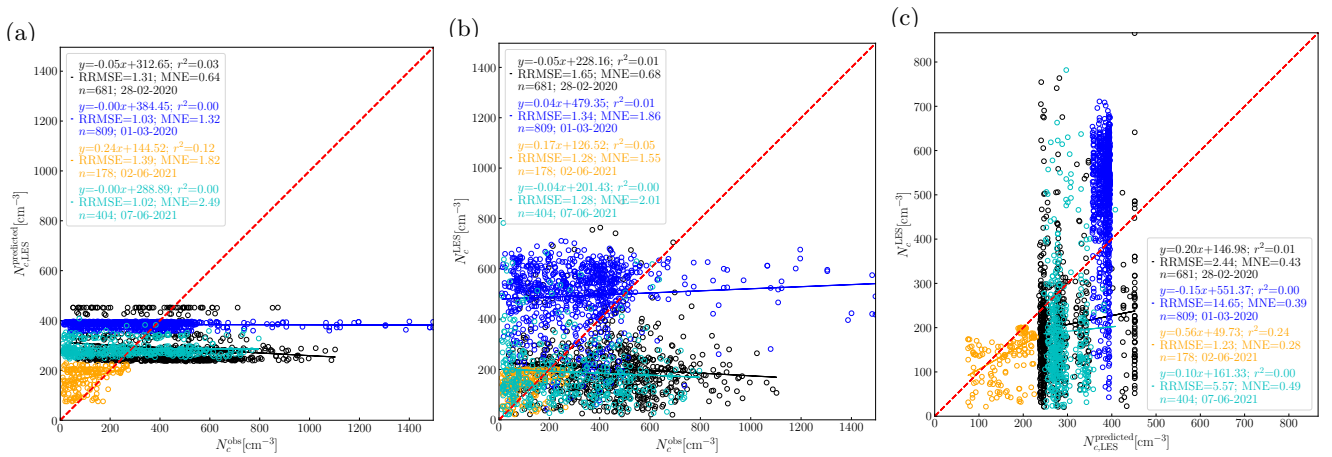

FIG. S8: N_c predicted ($N_{c,\text{LES}}^{\text{predicted}}$) from the WRF-LES as input (LES N_a, w', T, q_v) using the full-data observation-RF compared to the LES-N_c (N_c^{LES}) or FCDP-N_c (N_c^{obs}).
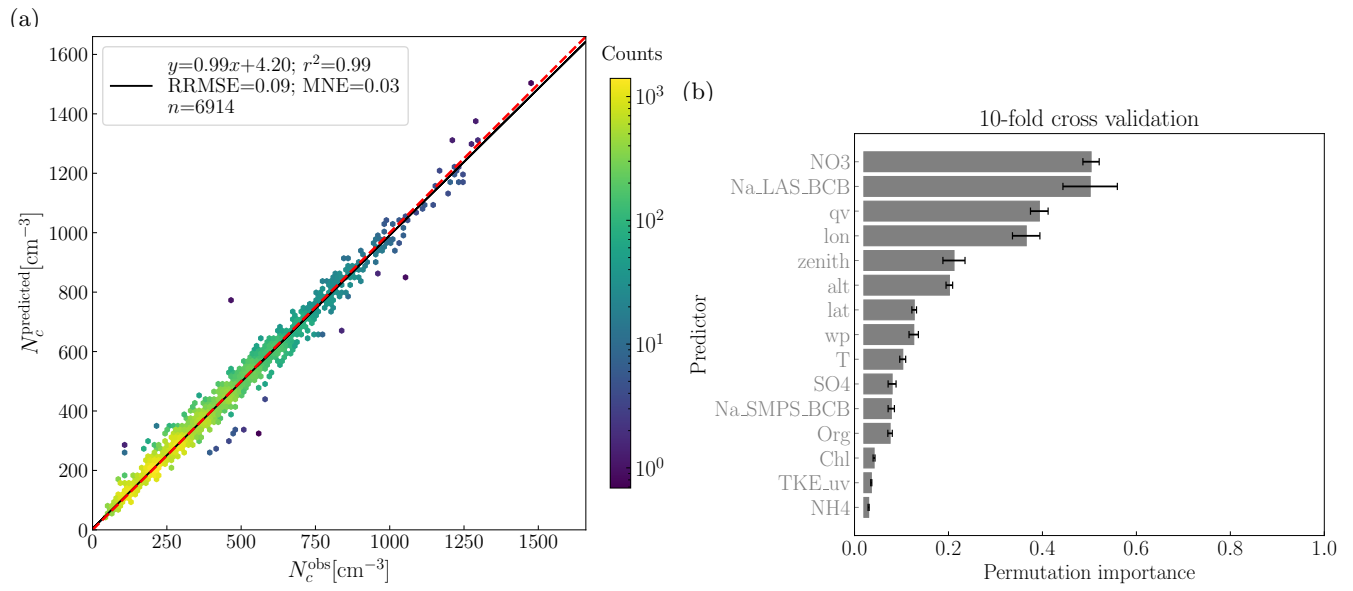
(a)

(b)

FIG. S9: Same as Figure 1(b) but with $N_c = \mathcal{G}(m_{\mathcal{X}}, N_a, w', \mathrm{TKE\_uv}, T, q_v, \boldsymbol{x}, \theta_z)$.

FIG. S10: Same as Figure 1 but with XGBoost.
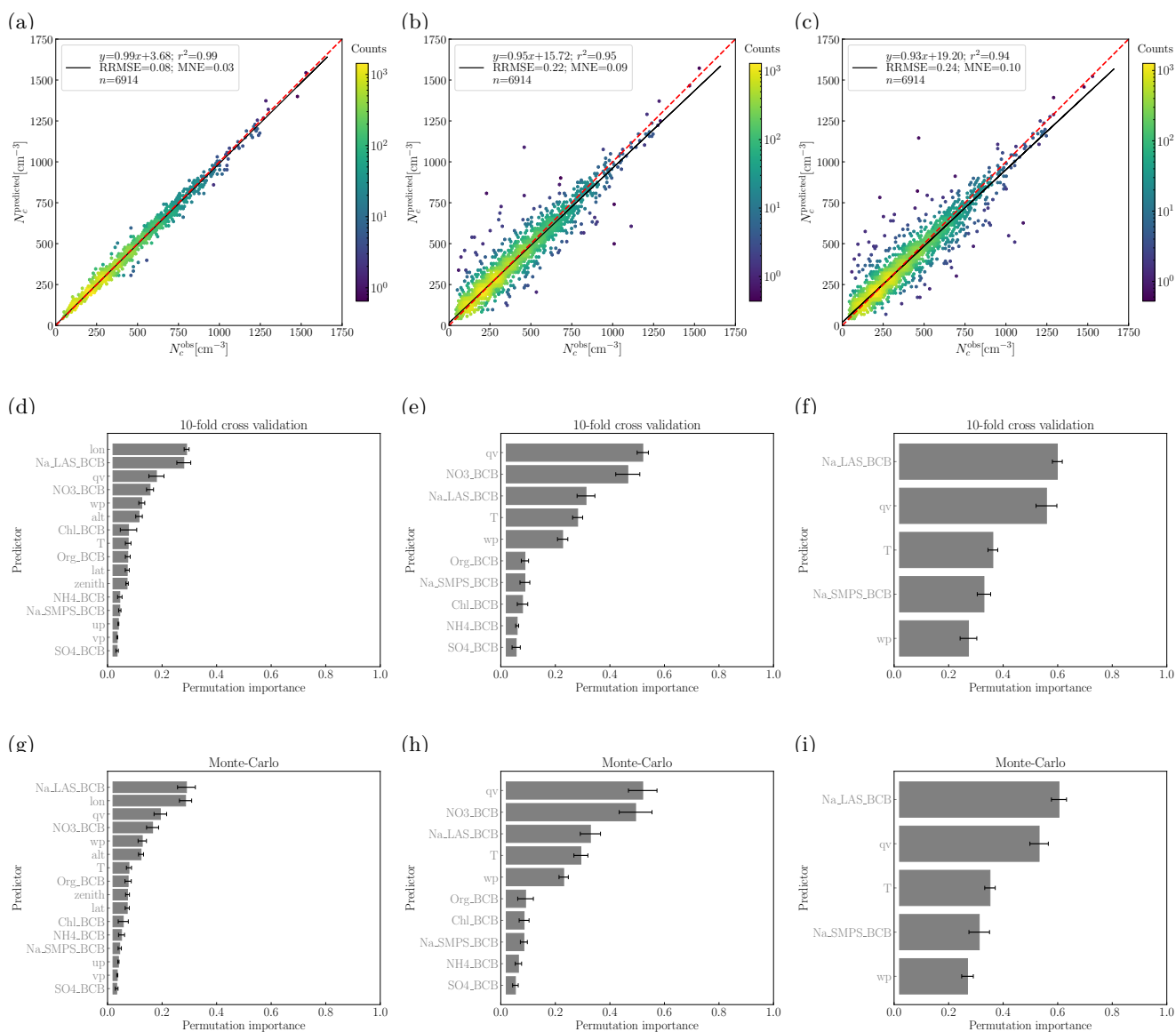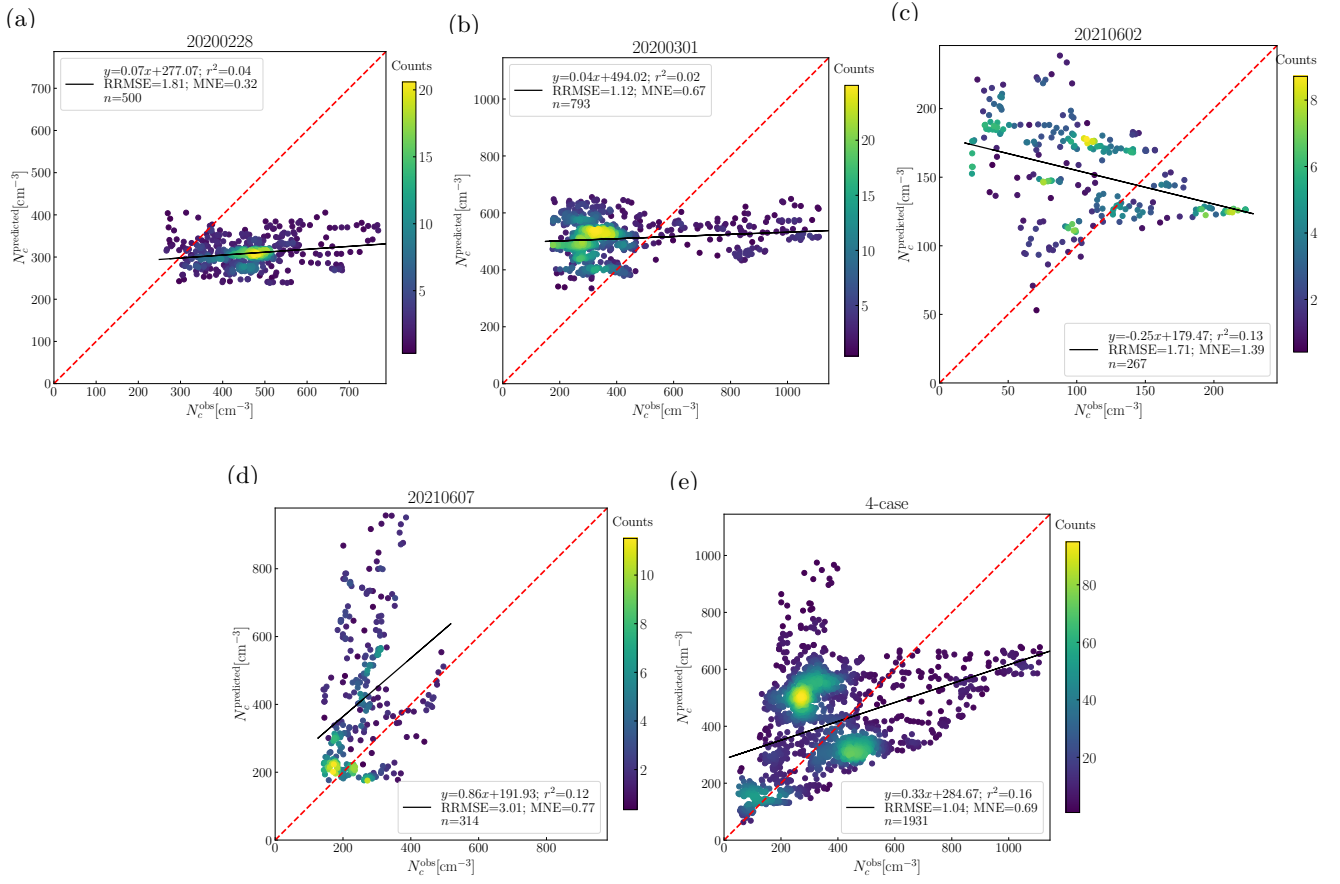
(a)

(b)

(c)

(d)

(e)

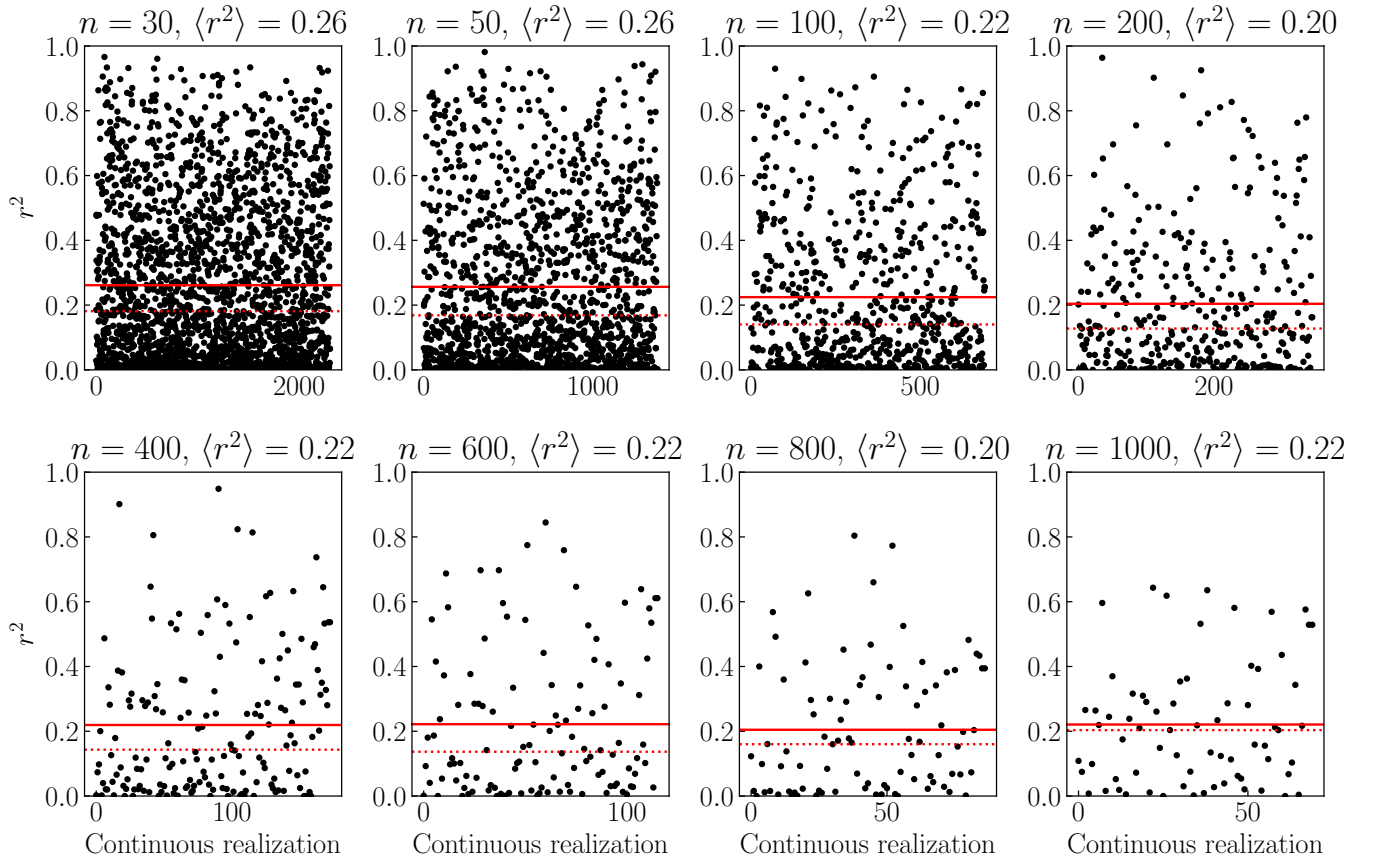FIG. S11: Same as Figure 2 but with XGBoost.
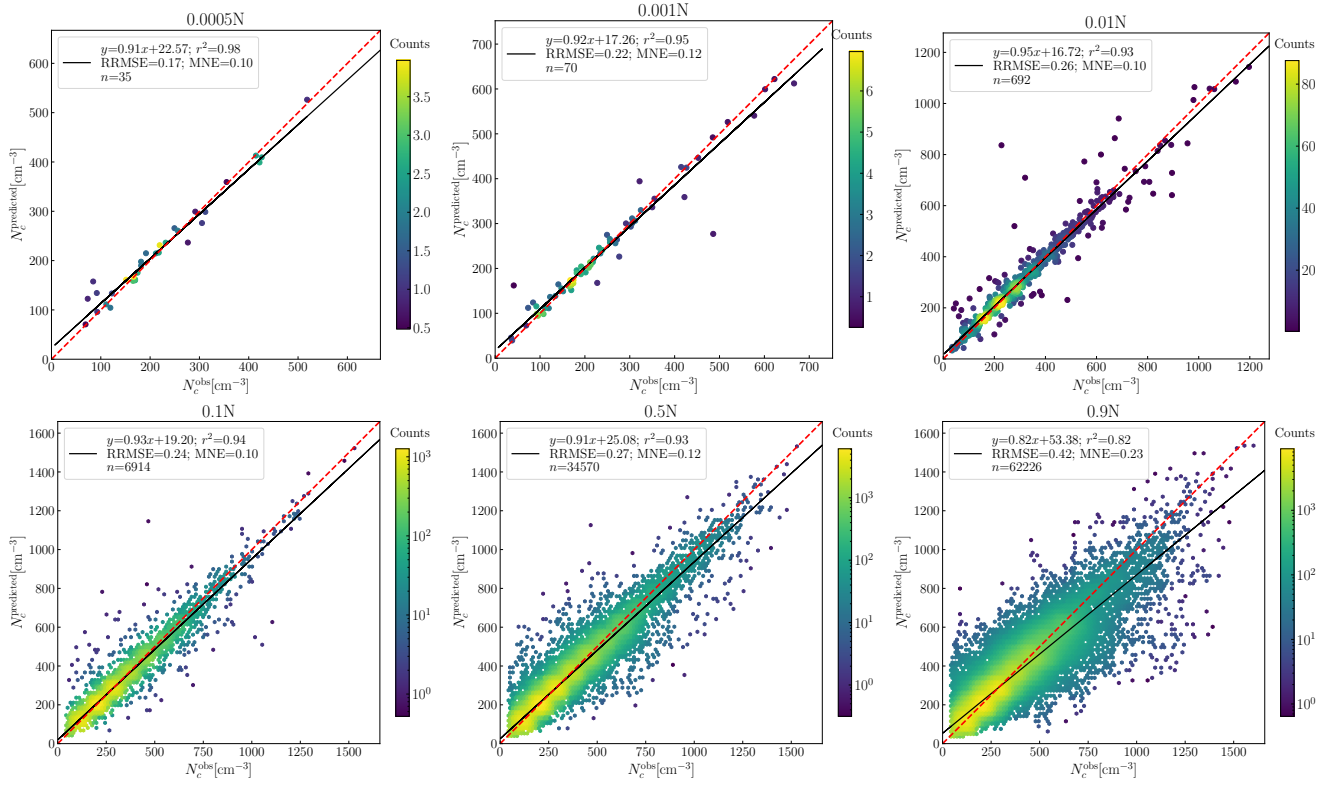
FIG. S12: Same as Figure 3 but with XGBoost.

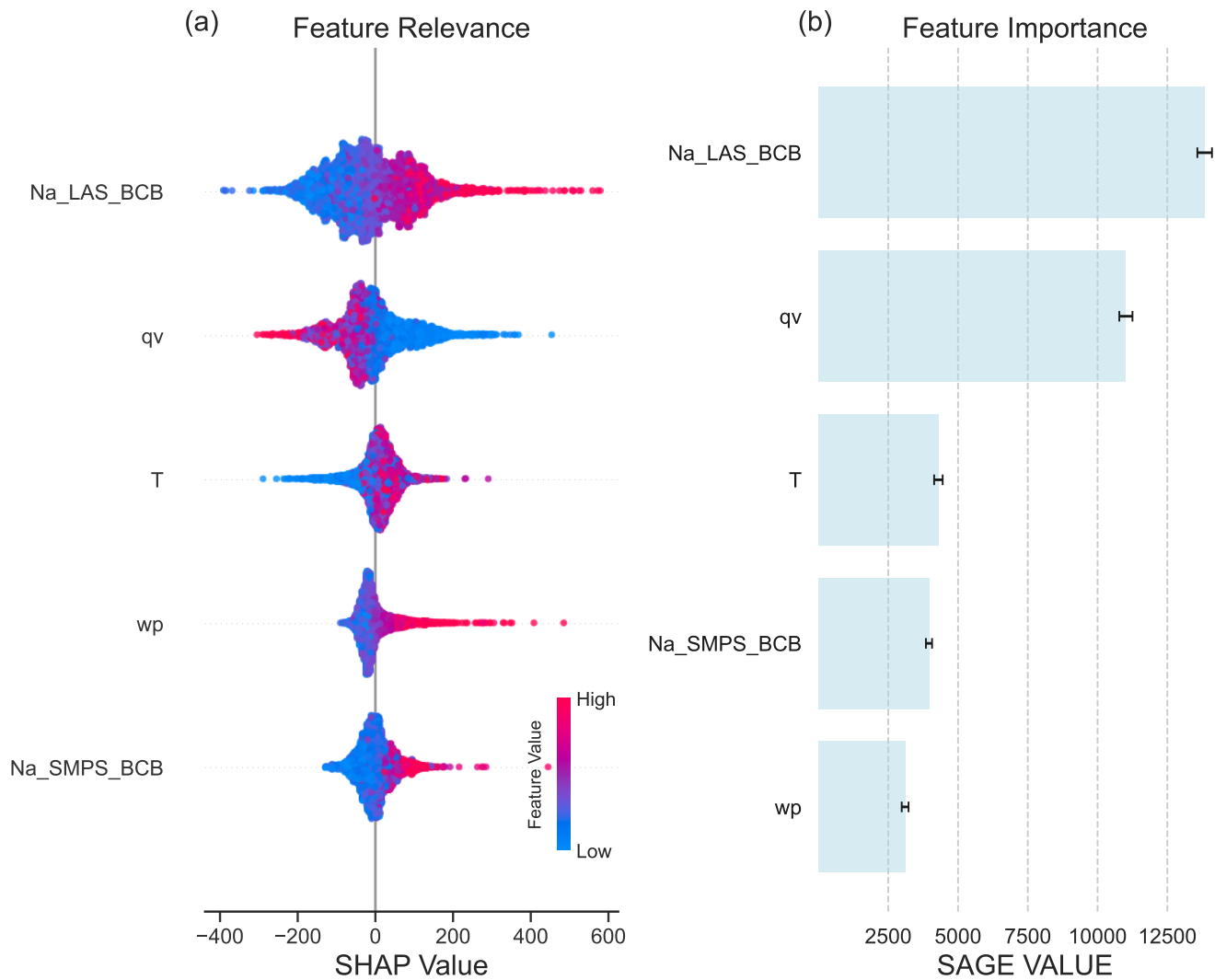FIG. S13: Same as Figure 4 but with XGBoost.

FIG. S14: (a) Shapley additive explanations (SHAP) and (b) Shapley additive global explanation (SAGE) feature rankings for the RFM. For (a), scatter points are SHAP values, while the color coding indicates the normalized feature value (0 - 1). To approximate a violin-style plot, vertical spread is applied to dense regions as in Flora *et al.* [6]. Features are ranked by their mean absolute SHAP value. SHAP is computed for a single example (local explainability), while SAGE is computed over a dataset (global explainability).

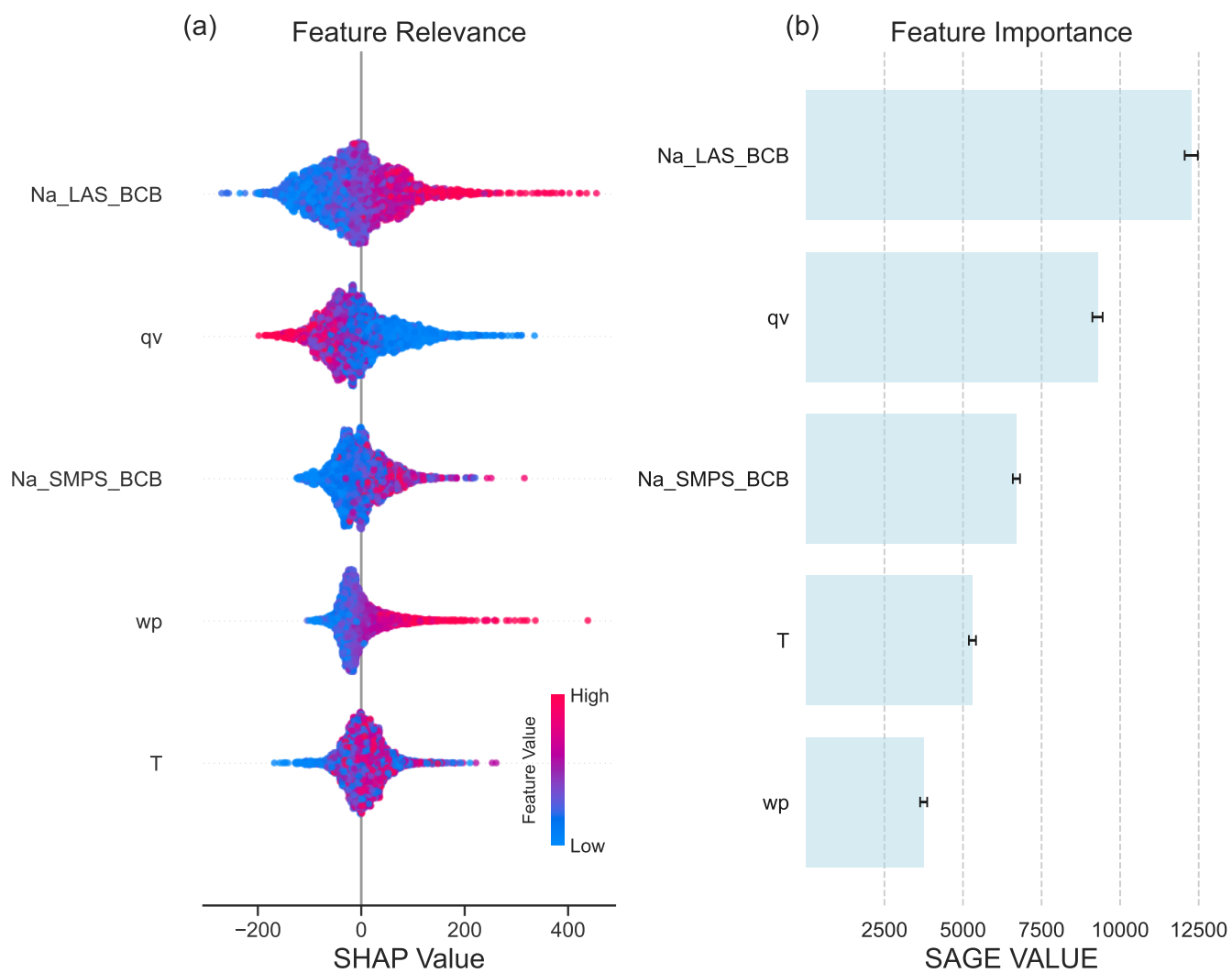(a) Feature Relevance

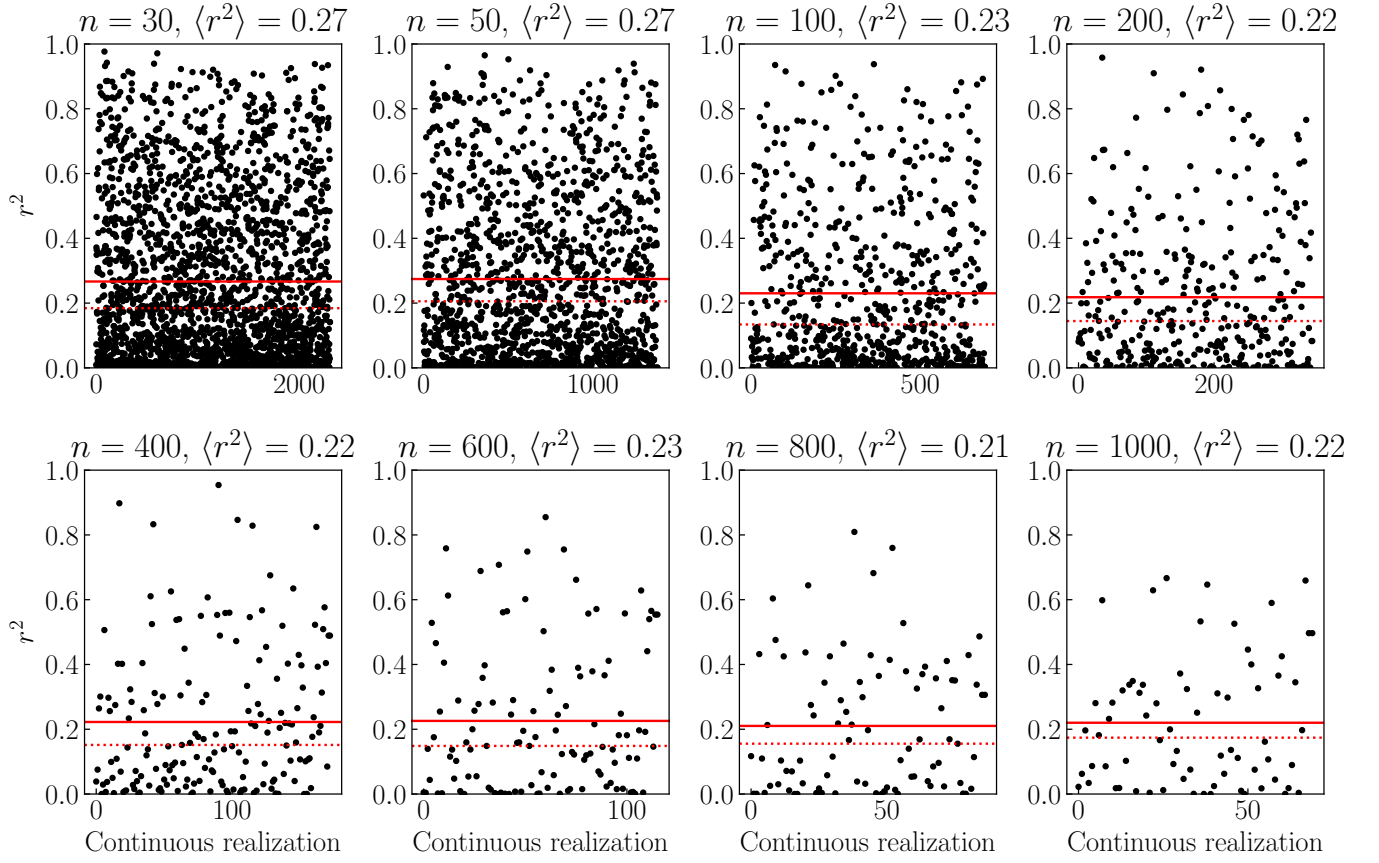(b) Feature Importance

FIG. S15: The same as Figure S14 but for the XGBoost.

FIG. S16: The same as Figure S12 but with a maximum tree depth of 10 for the XGBoost model. The corresponding binned scatter-plot of N_c^{obs} and $N_c^{\mathrm{predicted}}$ is shown in Figure S17, which yields a worse prediction compared to Figure S10c. The test of the maximum tree depth is shown in Figure S18.
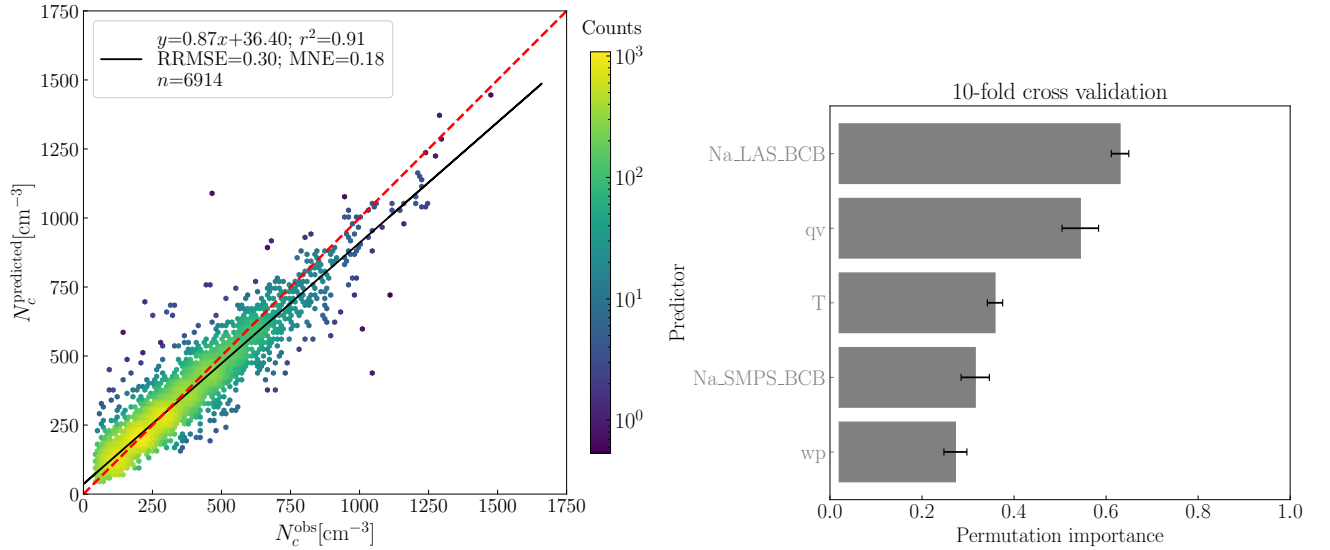

FIG. S17: The same as Figure S10c but with a maximum tree depth of 10 for the XGBoost model. The corresponding r^2 of $N_c = \mathcal{G}(N_a, w', T, q_v)$ predictions from continuous sampling with different sampling size n using the full-data observation-XGBoost is shown in Figure S16.
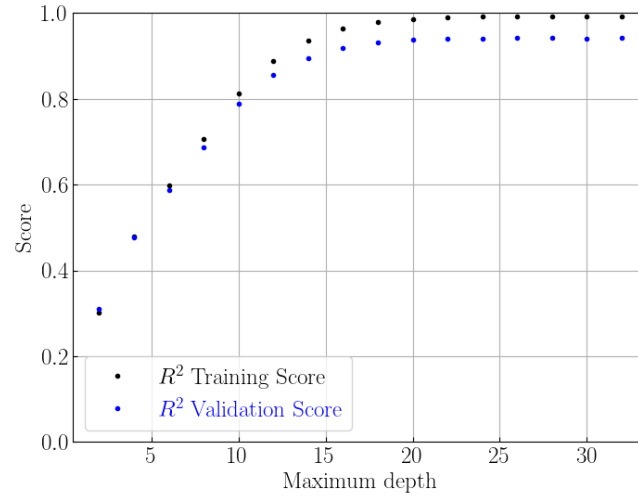
FIG. S18: Test of the maximum tree depth for the XGBoost with fixed number of trees (98) and test size (0.1).

[1] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, Z. Liu, J. Berner, W. Wang, J. G. Powers, M. G. Duda, D. M. Barker, and others, A description of the advanced research WRF model version 4, National Center for Atmospheric Research: Boulder, CO, USA **145**, 550 (2019).

[2] H. Wang, W. C. Skamarock, and G. Feingold, EnglishEvaluation of Scalar Advection Schemes in the Advanced Research WRF Model Using Large-Eddy Simulations of Aerosol–Cloud Interactions, Monthly Weather Review **137**, 2547 (2009).

[3] H. Morrison, G. Thompson, and V. Tatarskii, Impact of Cloud Microphysics on the Development of Trailing Stratiform Precipitation in a Simulated Squall Line: Comparison of One- and Two-Moment Schemes, Monthly Weather Review **137**, 991 (2009).

[4] X.-Y. Li, H. Wang, J. Chen, S. Endo, S. Kirschler, C. Voigt, E. Crosbie, L. D. Ziemba, D. Painemal, B. Cairns, J. W. Hair, A. F. Corral, C. Robinson, H. Dadashazar, A. Sorooshian, G. Chen, R. A. Ferrare, M. M. Kleb, H. Liu, R. Moore, A. J. Scarino, M. A. Shook, T. J. Shingler, K. L. Thornhill, F. Tornow, H. Xiao, and X. Zeng, Large-Eddy Simulations of Marine Boundary-Layer Clouds Associated with Cold-Air Outbreaks During the ACTIVATE Campaign. Part II: Aerosol–Meteorology–Cloud Interaction, Journal of the Atmospheric Sciences **80**, 1025 (2023).

[5] X.-Y. Li, H. Wang, M. W. Christensen, J. Chen, S. Tang, S. Kirschler, E. Crosbie, L. D. Ziemba, D. Painemal, A. F. Corral, K. A. McCauley, S. Dmitrovic, A. Sorooshian, M. Fenn, J. S. Schlosser, S. Stamnes, J. W. Hair, B. Cairns, R. Moore, R. A. Ferrare, M. A. Shook, Y. Choi, G. S. Diskin, J. DiGangi, J. B. Nowak, C. Robinson, T. J. Shingler, K. Lee Thornhill, and C. Voigt, Process Modeling of Aerosol-Cloud Interaction in Summertime Precipitating Shallow Cumulus Over the Western North Atlantic, Journal of Geophysical Research: Atmospheres **129**, e2023JD039489 (2024).

[6] M. L. Flora, C. K. Potvin, A. McGovern, and S. Handler, A Machine Learning Explainability Tutorial for Atmospheric Sciences, Artificial Intelligence for the Earth Systems **3**, 10.1175/aies-d-23-0018.1 (2023).