Supplement of Atmos. Chem. Phys., 24, 13603–13631, 2024
https://doi.org/10.5194/acp-24-13603-2024-supplement

Atmospheric Chemistry and Physics
Open Access
EGU

Supplement of

On the dynamics of ozone depletion events at Villum Research Station in the High Arctic

Jakob Boyd Pernov et al.

Correspondence to: Jakob Boyd Pernov (jakob.pernov@epfl.ch) and Henrik Skov (hsk@envs.au.dk)

Table S1. Percentage of missing data before imputation for 2007-2019.

Variable	% Missing before imputation
Above Mixed Layer	0
Pressure	12.76
Radiation	0
RH	18.269
Sea Ice	0
Temperature	13.07
Wind Direction	21.60
Wind Speed	21.49
Snow	0

Table S2. Overview of the hyperparameter optimization for the ML model.

Hyperparameter	Short name	Range	Optimal Value
Number of estimators	n_estimators	100-1000	550
Maximum tree depth	max_depth	3-7	5
L1 regularization	reg_alpha	0-10	8.1
L2 regularization	reg_lambda	0-10	1.8
Minimum child weight	min_child_weigh	1-10	8
Gamma	gamma	0-10	0.6
Learning rate	learning_rate	0.01-0.1	0.092
Subsample fraction	subsample	0.5-1.0	0.95
Column sample by tree fraction	colsample_bytree	0.5-1.0	0.9
Positive Label Scalar	scale_pos_weight	1-10	5

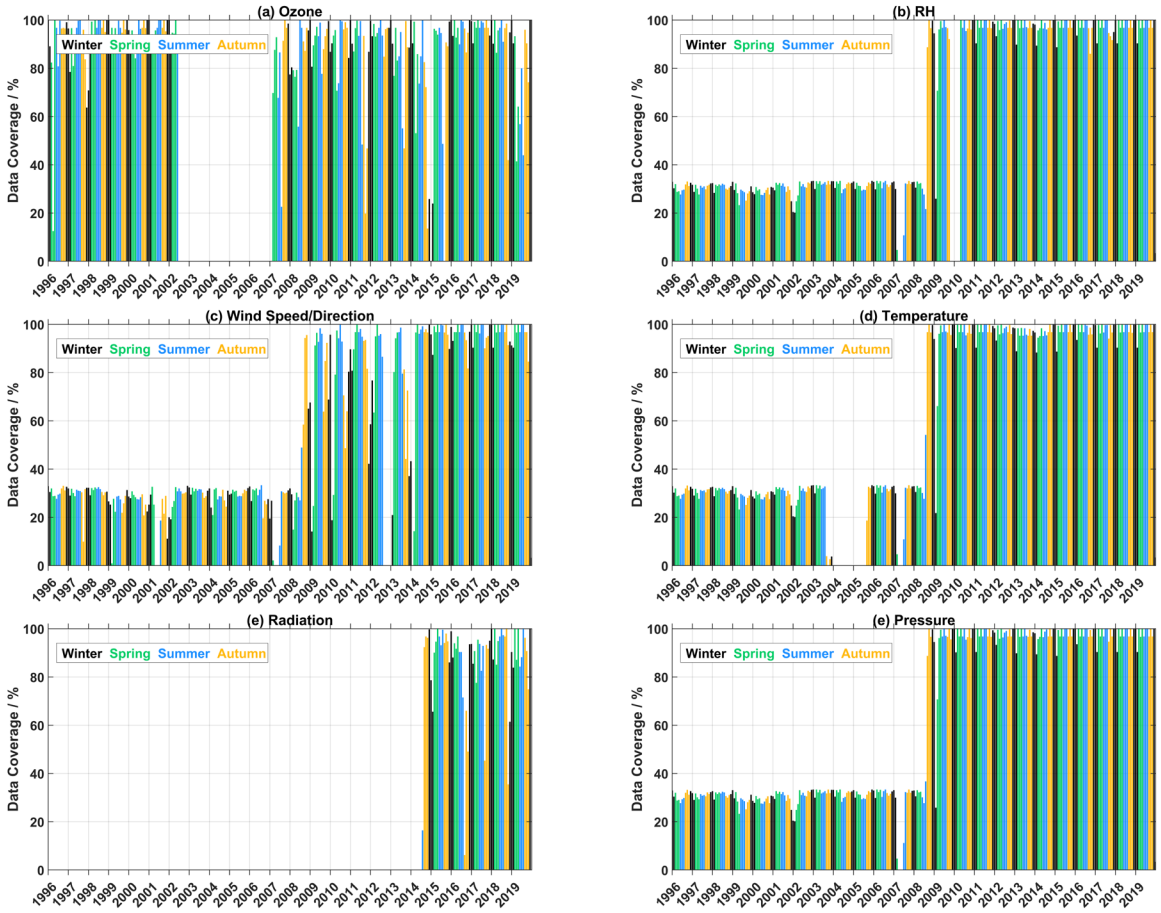

Figure S1. Data coverage over 1996-2019 for (**a**) ozone and (**b-e**) the meteorological variables as expressed as a percentage of available measurements relative to the total possible number of measurements for each month. Months are color-coded by season.

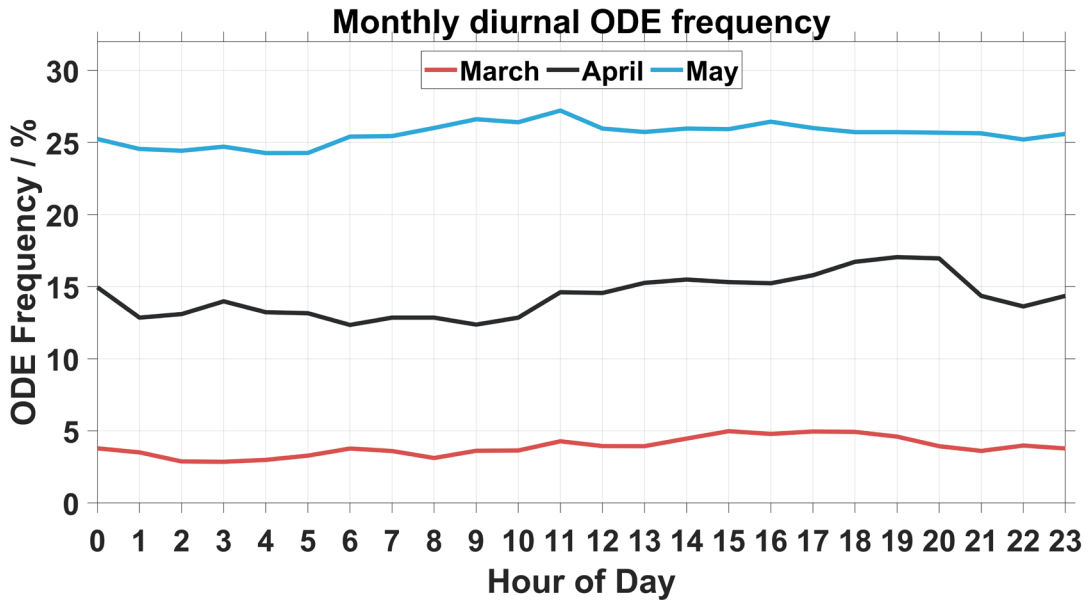

Figure S2. Diurnal ODE frequency by month.

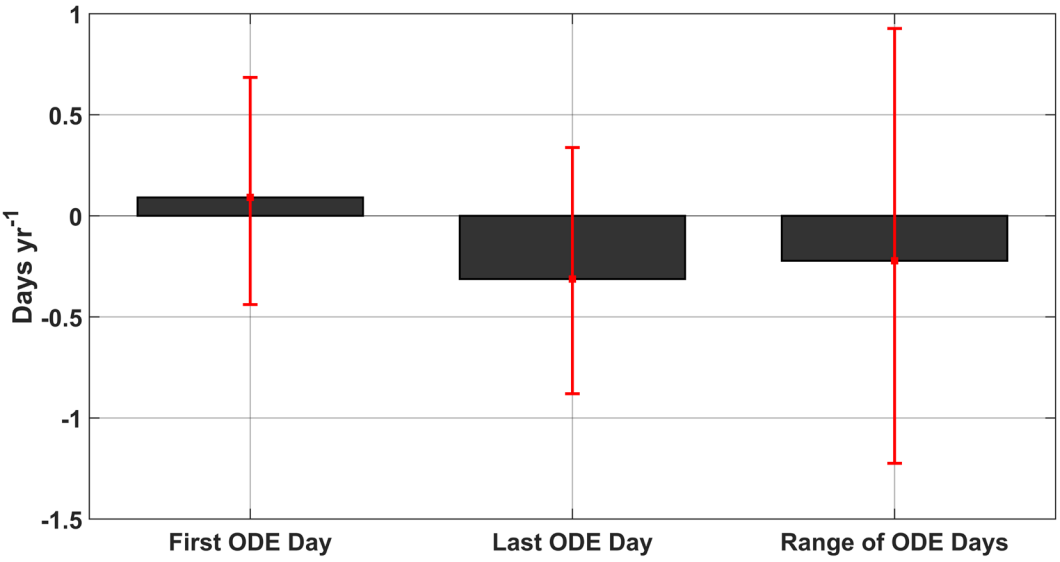

Figure S3. Trend analysis of the first ODE day (defined as the first day of the year with an ozone measurement < 10 ppbv), the last ODE day (defined as the last day of the year with an ozone measurement < 10 ppbv), and the range of the ODE season (last day of the year minus the first day of the year). The black bars represent trends that are not significant on the 95th % confidence level. The red error bars represent the 95th % confidence intervals of the slope. The p values for first, last, and range of ODE days are 0.78, 0.20, and 0.65, respectively.

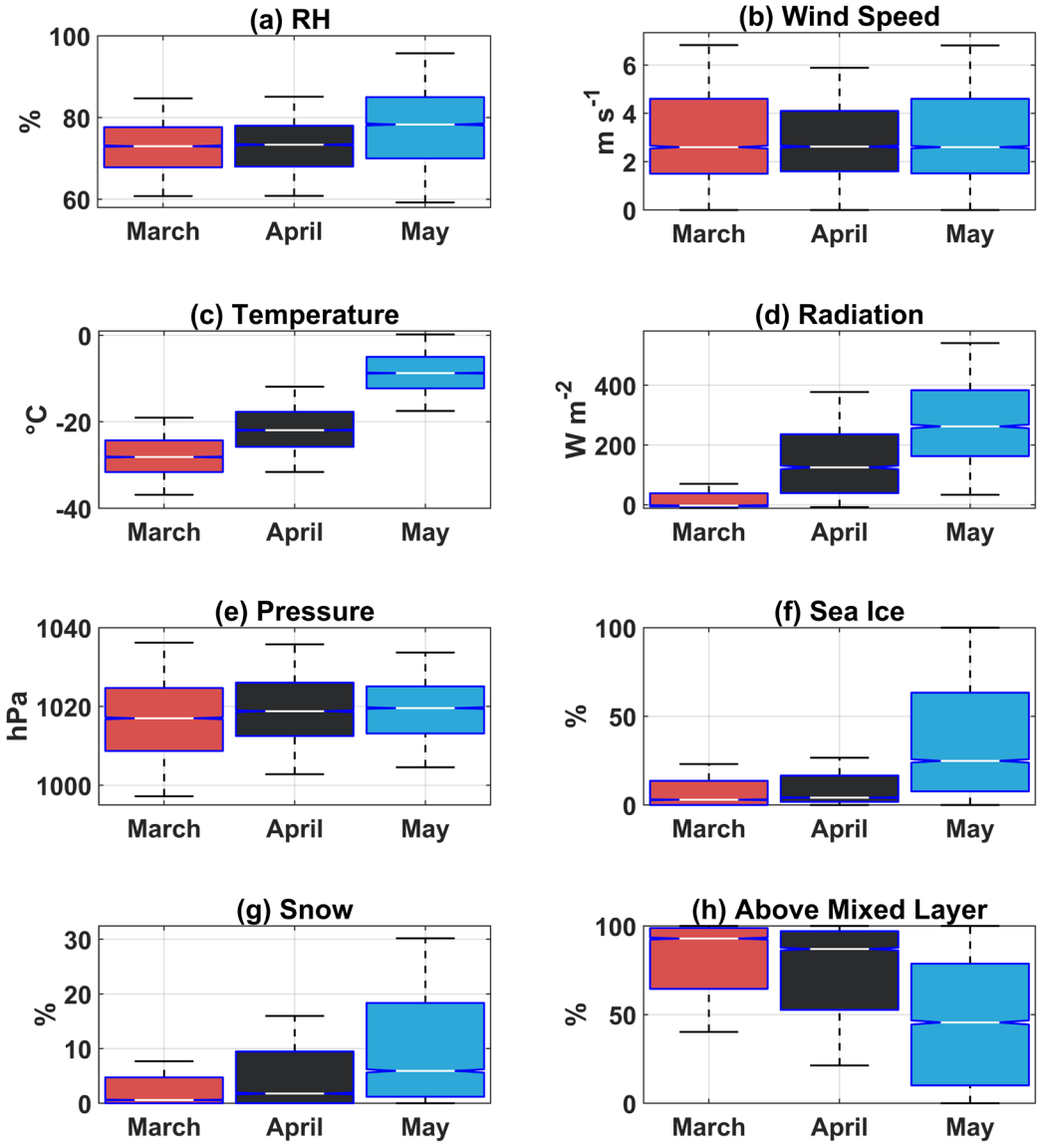

Figure S4. Distribution of meteorological and air mass history variables during each spring month including (**a**) RH, (**b**) wind speed, (**c**) temperature, (**d**) radiation, (**e**) pressure, (**f**) time over sea ice, (**g**) time over snow, and (**h**) time above the mixed layer. The line in the middle of the box represents the median, the boxes represent the interquartile range, the medians of boxes whose notches do not overlap differ with 95% confidence. For a description of how the time spent over different surface types is calculated see the methods section. All available data for each variable collocated with ozone measurements was used, resulting in different years used for each variable, with the minimum number of years included was 5 for radiation.

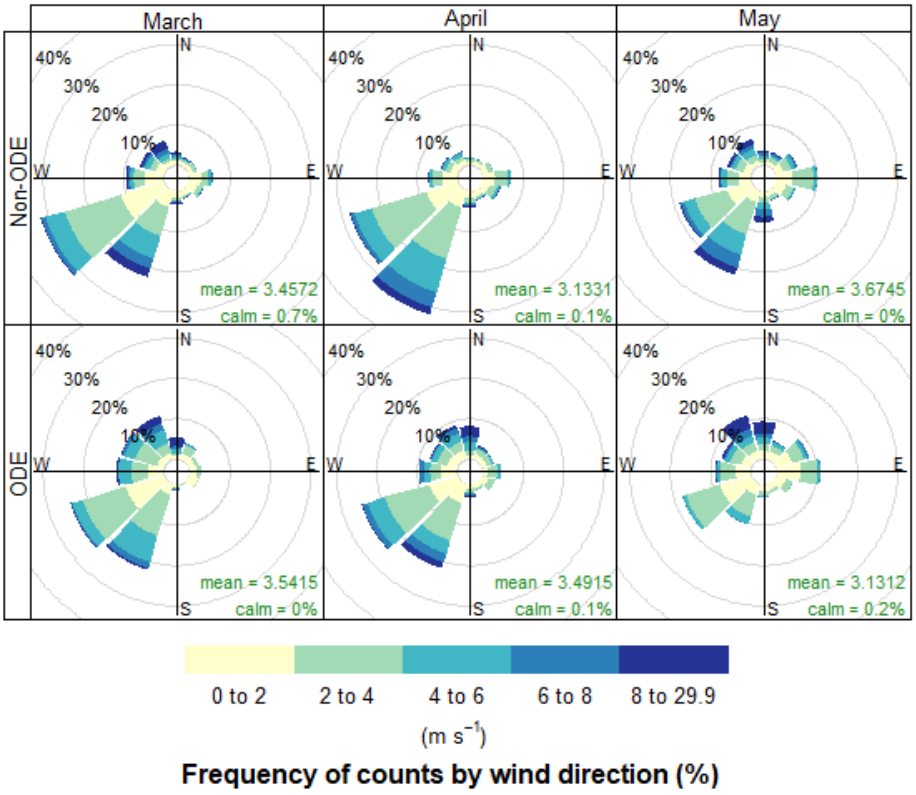

Figure S5. Wind roses for Non-ODEs (top row) and ODEs (bottom row) for the spring months. The mean wind speed and the percentage of time the wind speed is zero (or calm) is given in each panel.

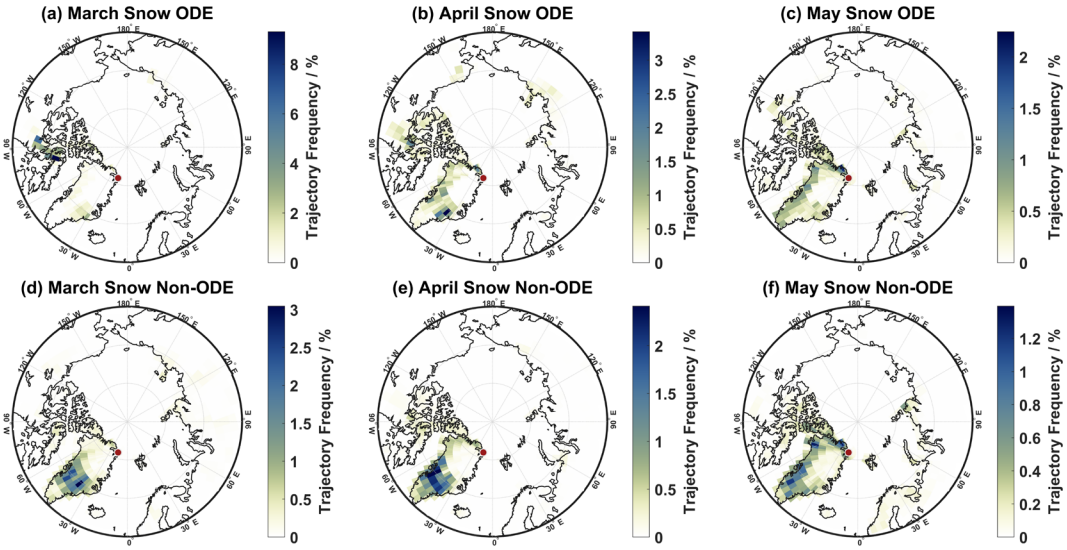

Figure S6. Trajectory frequency maps for trajectory steps below the mixed layer and over snow for (**a-c**) ODEs and (**d-f**) Non-ODEs during March, April, May at Villum (indicated by the red and white circle).

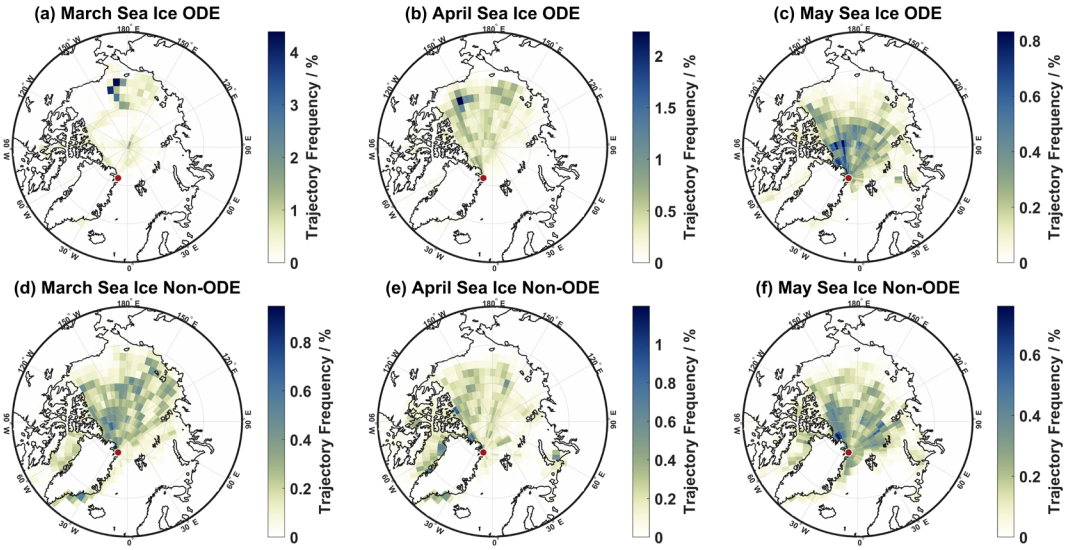

Figure S7. Trajectory frequency maps for trajectory steps below the mixed layer and over sea ice for (**a-c**) ODEs and (**d-f**) Non-ODEs during March, April, May at Villum (indicated by the red and white circle).

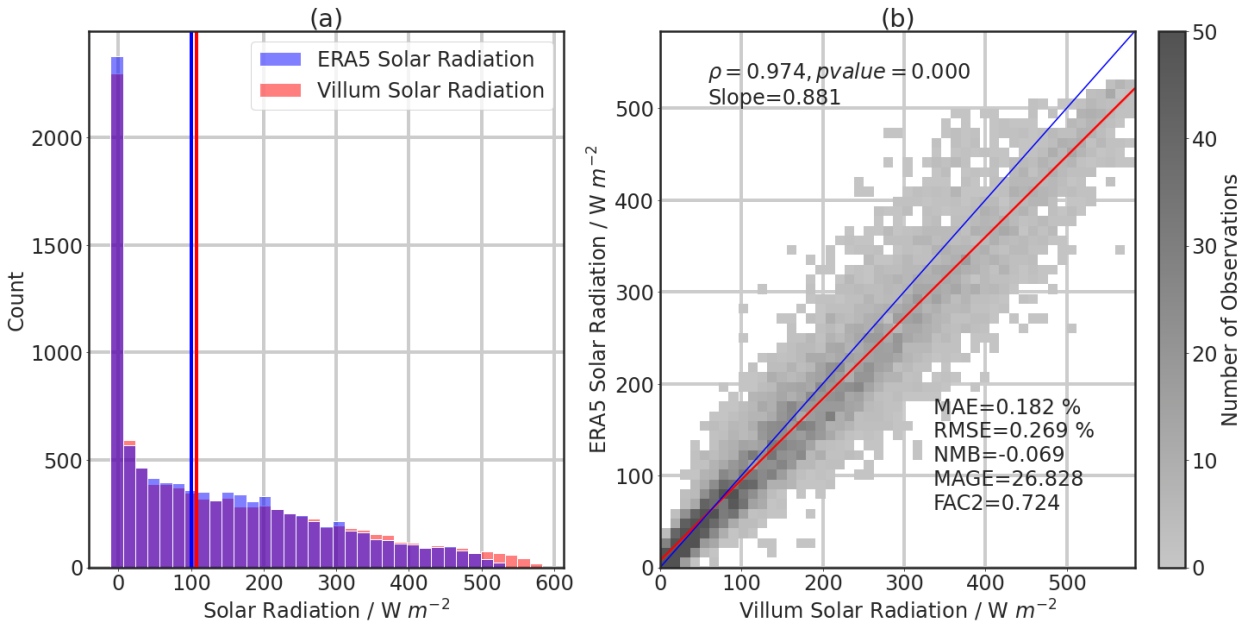

Figure S8. Comparison between observations of solar radiation and shortwave solar radiation downwelling from ERA5 Reanalysis, with (**a**) a histogram and (**b**) scatterplot. On the scatterplot, the spearman rank correlation coefficient along with its associated p-value are presented in the top left corner while the mean absolute error (MAE), root mean square error (RMSE), normalized mean bias (NMB), mean absolute gross error (MAGE), and fraction of modeled data with a factor of 2 (FAC2) of the observations are presented in the bottom right corner. The MAE and RMSE are given as percentages of the mean observational value.

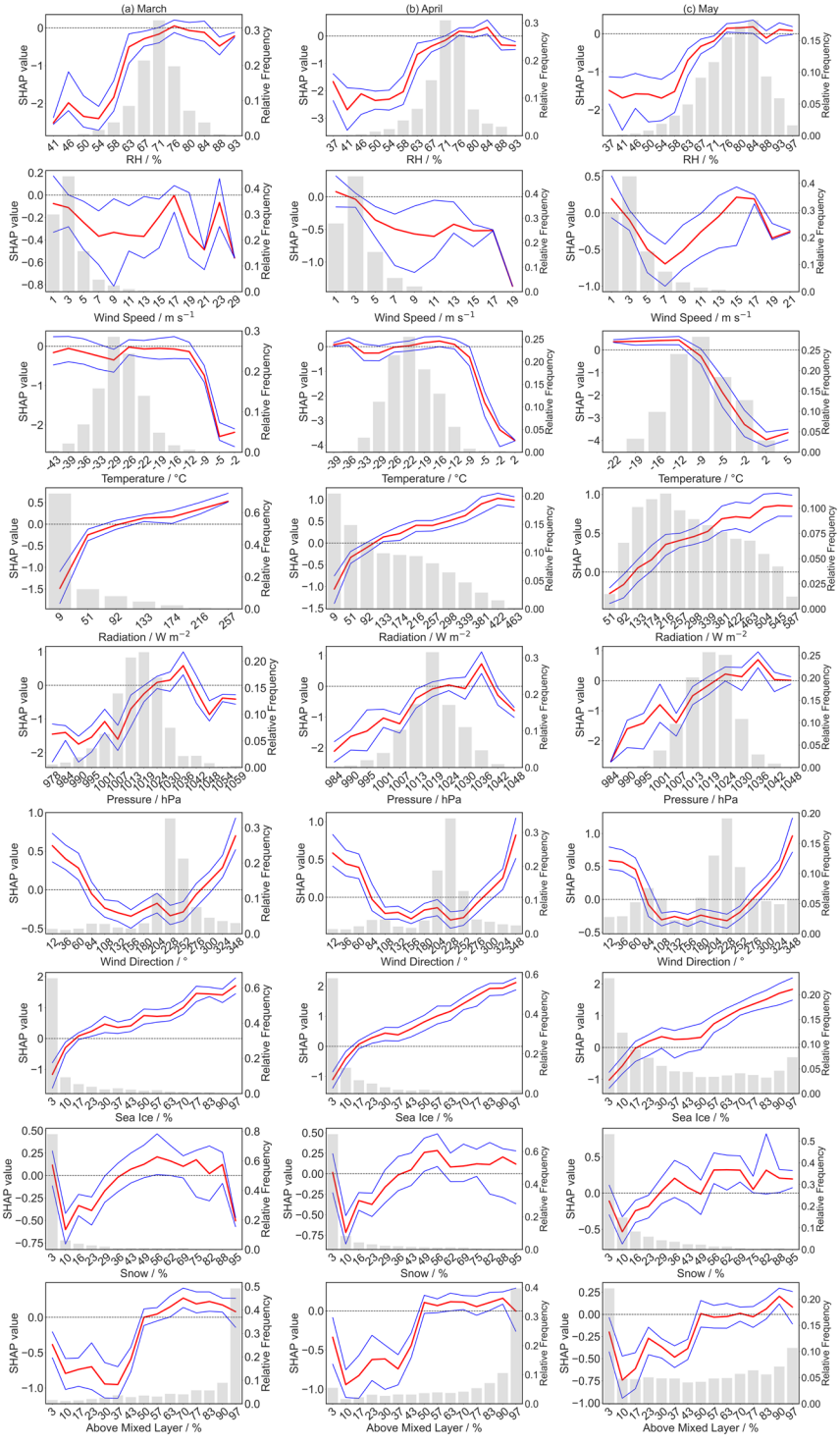

Figure S9. The relationships between SHAP and ambient values for (**a**) March, (**b**) April, and (**c**) May. 15 equally spaced bins were calculated for each feature, the median (red line) and IQR (blue lines) of the SHAP values were computed for each bin. The value listed on the x-axis is the midpoint of each bin. The relative frequency of each histogram bin for each variable is displayed on the right axis. The bins are the same used in Fig. 9.

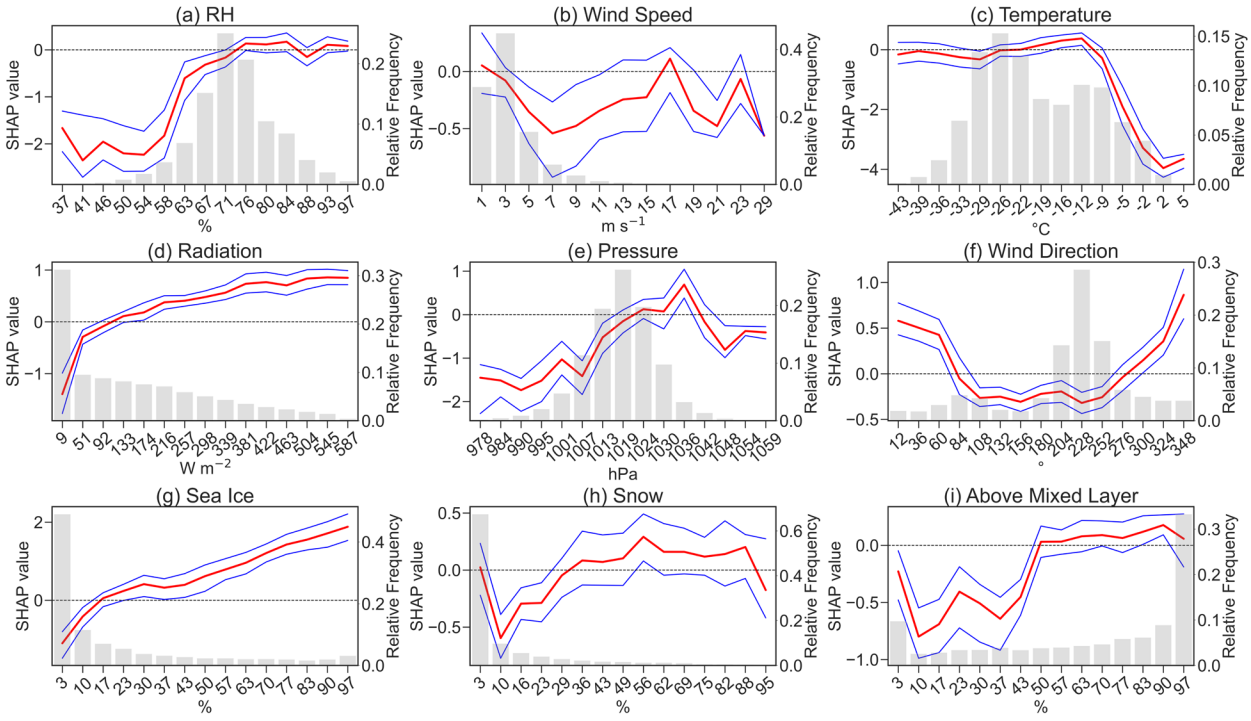

Figure S10. The relationships between SHAP and ambient values for March – May combined for (**a**) time above the mixed layer, (**b**) pressure, (**c**) radiation, (**d**) RH, (**e**) time air masses spent over sea ice, (**f**) time air masses spent over snow, (**g**) temperature, (**h**) wind direction, and (**i**) wind speed. Fifteen equally spaced bins were calculated for each feature, the median (red line) and IQR (blue lines) of the SHAP values were computed for each bin. The value listed on the x-axis is the midpoint of each bin. The relative frequency of each histogram bin for each variable is displayed on the right axis.

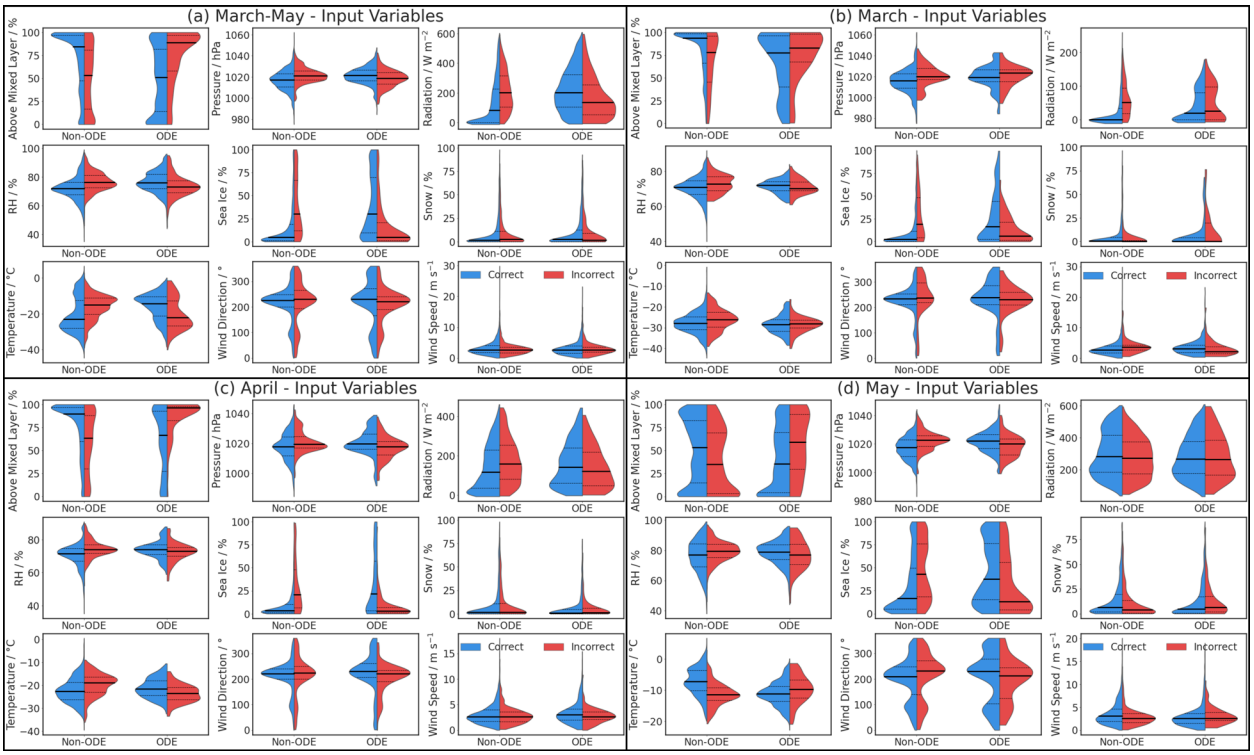

Figure S11. Distributions of input variables for Non-ODEs and ODEs for (**a**) March-May, (**b**) March, (**c**) April, (**d**) May, the thick black line represents the median and thin dashed lines represent the 25th and 75th percentiles. The distributions are color-coded by correct (blue) or incorrect (red) prediction by the ML model. All data over 2007-2019 was used to predict ODEs and was used in this analysis.

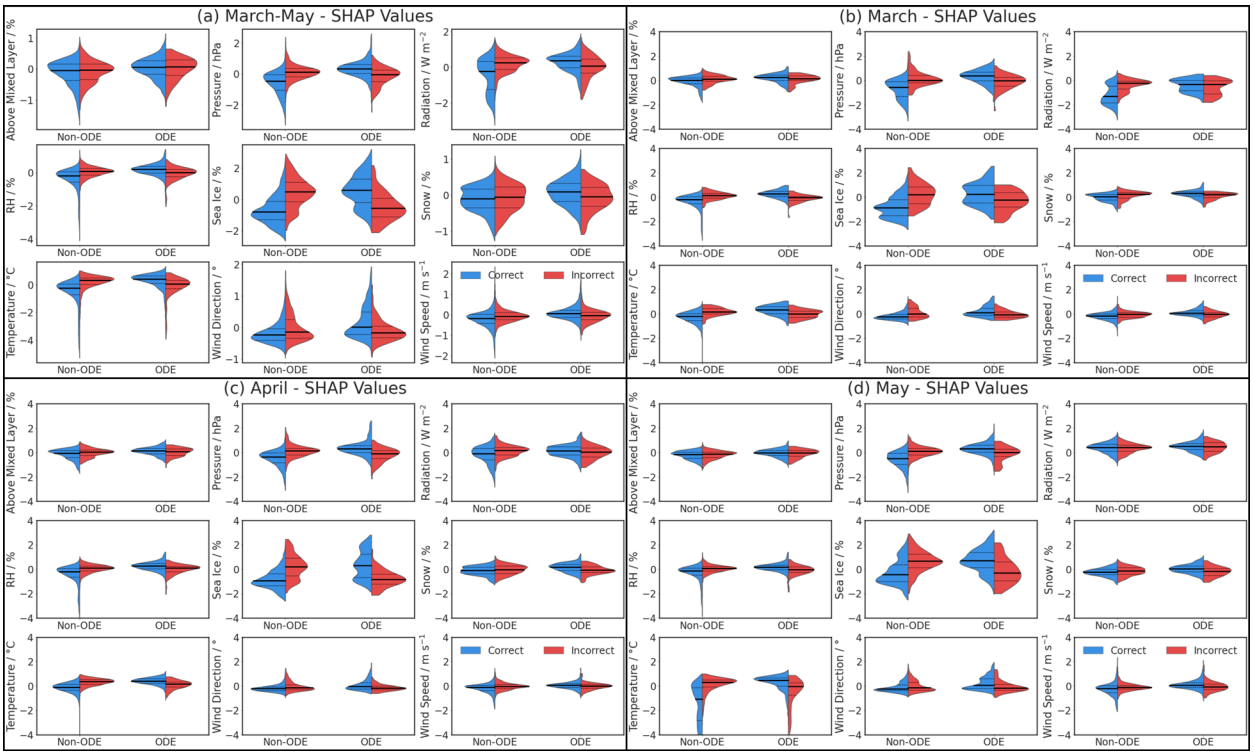

Figure S12. Distributions of SHAP values for Non-ODEs and ODEs for (**a**) March-May, (**b**) March, (**c**) April, (**d**) May, the thick black line represents the median and thin dashed lines represent the 25th and 75th percentiles. The distributions are color-coded by correct (blue) or incorrect (red) prediction by the ML model. All data over 2007-2019 was used to predict ODEs and was used in this analysis.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-Generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, event-place: Anchorage, AK, USA, 2623–2631, https://doi.org/10.1145/3292500.3330701, 2019.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B.: Algorithms for Hyper-Parameter Optimization, in: Advances in Neural Information Processing Systems, 2011.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, event-place: San Francisco, California, USA, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems, https://doi.org/10.48550/arXiv.1705.07874, 2017.

Lundberg, S. M., Erion, G. G., and Lee, S.-I.: Consistent Individualized Feature Attribution for Tree Ensembles, 2019.

Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022.

Shapley, L. S.: 17. A Value for n-Person Games, in: Contributions to the Theory of Games (AM-28), Volume II, edited by: Kuhn, H. W. and Tucker, A. W., Princeton University Press, Princeton, 307–318, https://doi.org/doi:10.1515/9781400881970-018, 1953.