Supplementary material for a framework to holistically investigate processes controlling the aerosol lifecycle using explainable AI techniques

Table of Contents

S1 Supplementary methods

S1.1 Aerosol size distribution filtering

Aerosol size distributions after initial processing, were further filtered using the filters described below considering only diameter ranges consistent over the size distribution whole timeseries used for analysis:

1. In order to remove PNSD with sections of the distribution missing, we consider the dlogdp of consecutive flagged groups in an individual size distribution. If any of these groups exceed the threshold (0.1) the whole distribution is removed.
2. The second filter applied is the percentage of NaN in a distribution, if a distribution consists of more than 10% NaN, then we consider the whole distribution to be contaminated and remove the distribution from analysis.
3. A total concentration percentage filter (99.5%) was implemented to ensure that any outlier distributions were removed from analysis. The total concentration threshold is calculated over the whole timeseries of PNSD and only considers bins within the consistent minimum and maximum bin limits to ensure a consistent comparison.

Total number concentration (N) is calculated as the sum of ΔN over all of the bins (Eq. (S1)).

$$N = \sum_{i=1}^{n} \Delta N_i \, \Delta log Dp_i \quad \text{(S1)}$$

4. DMPS measurements can experience abnormal behaviour in the largest size bins. As the concentrations in these bins are often lower, extreme values associated with the largest bins may not be addressed. Therefore, a separate filter is applied for the final two bins: implement a filter based on a threshold (99.5%) on the total concentration in the last two bins over the timeseries considered in each data file (with consistent bin midpoints).

Filter removal rates for each yearly aerosol file are shown in Table S1.

Year	Datapoints at start	Percentage points removed for 1	Percentage points removed for 2	Percentage points removed for 3	Percentage points removed for 4
2014	8760	20.16	0 (all removed during 1)	0.32	0.5
2015	8760	9.90	0 (all removed during 1)	0.22	0.5
2016	8784	7.66	0 (all removed during 1)	0.44	0.5
2017	8760	0.45	0 (all removed during 1)	1.38	0.5
2018	8760	2.95	0 (all removed during 1)	0.15	0.5

Table S1: The percentage of datapoints removed from each yearly aerosol file by each filter for Trollhaugen.

S1.2 Aerosol size distribution interpolation

Two issues must be considered when conducting the interpolation:

1. The range 1-1000nm often exceeds the original range of bin midpoints and the PCHIP algorithm will extrapolate past the bounds of the original distributions.
2. Any missing (NaN: Not a Number) values within the distribution will propagate through the whole distribution, resulting in an empty distribution.

To resolve these two issues data extrapolated during the PCHIP algorithm outside of the original grid limits are set to NaN. With the filters applied discussed in Sect. S1.2, any distributions with more than 10% NaN and with any groups of NaN with a $dlogD_p$ larger than 0.1 nm have been removed from analysis. Therefore, the distributions for interpolation have sufficiently small areas of missing data within the size distribution to interpolate over. For the PCHIP algorithm, all bins with missing values are removed from the original distribution before interpolation to the new grid.

The results after filtering and interpolating are shown in Fig. S1.

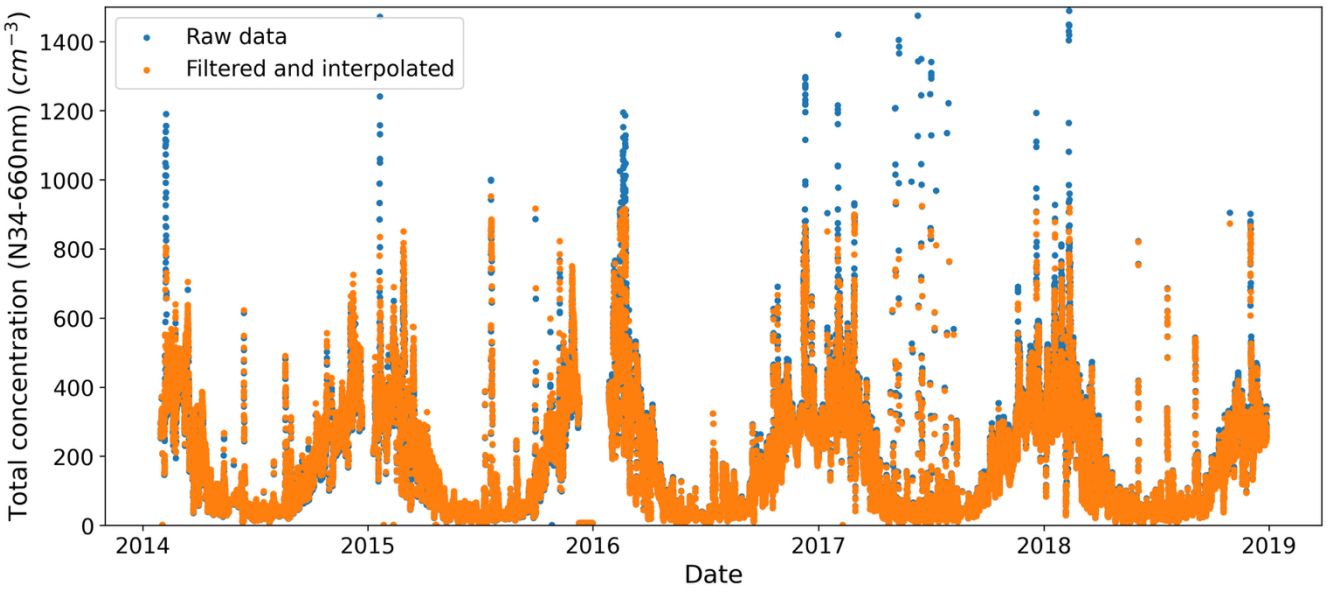


Figure S1: The total concentration 34-660nm. Raw data (blue) and data after filtering and interpolation to the new size grid (orange) for Trollhaugen.

S1.3 Filtering for potential contamination from generators

As described in Sect. 2.1 additional filters were needed for Trollhaugen to mitigate potential contamination from nearby station generators, the results for these filters are shown in Fig. S2.

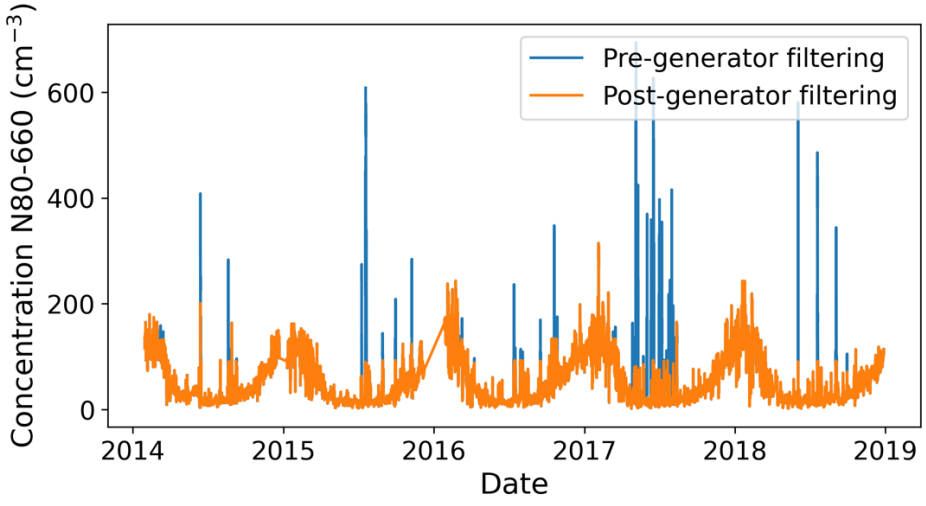


Figure S2: The concentration (N80-660nm) pre (blue) and post (orange) filtering for potential generator contamination at Trollhaugen.

S1.4 Land classification types

Explanatory variable	Class names
Deciduous Forest	"tree_broadleaved_deciduous_closed_to_open", "tree_broadleaved_deciduous_closed", "tree_broadleaved_deciduous_open", "tree_needleleaved_deciduous_closed_to_open", "tree_needleleaved_deciduous_closed", "tree_needleleaved_deciduous_open"
Evergreen Forest	"tree_broadleaved_evergreen_closed_to_open", "tree_needleleaved_evergreen_closed_to_open", "tree_needleleaved_evergreen_closed", "tree_needleleaved_evergreen_open"
Shrub and grass land	"mosaic_tree_and_shrub", "mosaic_herbaceous", "shrubland", "shrubland_evergreen", "shrubland_deciduous", "grassland", "lichens_and_mosses", "sparse_vegetation","sparse_shrub", "sparse_herbaceous"
Cropland	"cropland_rainfed", "cropland_rainfed_herbaceous_cover", "cropland_rainfed_tree_or_shrub_cover", "cropland_irrigated", "mosaic_cropland"
Urban	"urban"

Table S2: The names of Land Cover Classification System (LCCS) classes used to define each land type in this study.

S1.5 UKESM aerosol modes

Aerosol mode	Species	Diameter (nm)	Geometric standard deviation (σ)
Nucleation soluble	SO_4, OM	1–10	1.59
Aitken soluble	SO_4, BC, OM	10–100	1.59
Accumulation soluble	SO_4, BC, OM, SS	100–500	1.40
Coarse soluble	SO_4, BC, OM, SS	500–10000	2.00
Aitken insoluble	BC, OM	10–100	1.59

Table S3: UKESM1.0 aerosol mode descriptions. Species comprise sulphate (SO_4), organic matter (OM), black carbon (BC) and sea salt (SS).

S1.6 Weighting

A weighting was applied to summary statistics taken along the trajectories as described in main text Sect. 3.2, Eq. (1). Fig. S2 shows the weighting in a graphical form.

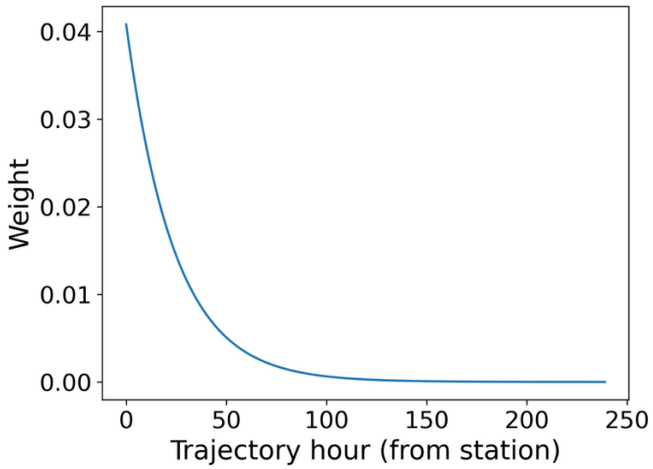


Figure S3: Weighting applied along trajectories.

S1.7 Dataset references

Citations for each ACTRIS dataset

Station	Years	Citation
Trollhaugen	2014-2016	Lunder, C., Fiebig, M., GAW-WDCA, NILU, ACTRIS, 2014-2016, Particle_number_size_distribution at Trollhaugen, data hosted by EBAS at NILU, DOI: https://doi.org/10.48597/P8UE-H3XQ
	2016-2019	Fiebig, M., Lunder, C. GAW-WDCA, NILU, ACTRIS, 2016-2020, Particle_number_size_distribution at Trollhaugen, data hosted by EBAS at NILU, DOI: https://doi.org/10.48597/ZYY4-2JFW
Varrio	2009-2016	Kulmala, M., Aalto, P. GAW-WDCA, 2000-2017, Particle_number_size_distribution at Värriö, data hosted by EBAS at NILU, DOI: https://doi.org/10.48597/G847-CW5N
	2016-2018	Kulmala, M., Petäjä, T. GAW-WDCA, ACTRIS, 2015-2020, Particle_number_size_distribution at Värriö, data hosted by EBAS at NILU, DOI: https://doi.org/10.48597/9Z23-MPE6

Mace Head	2009-2010	Jennings, P., GAW-WDCA, CREATE, EUSAAR, 2009-2010, Particle_number_size_distribution at Mace Head, data hosted by EBAS at NILU, DOI: https://doi.org/10.48597/7Y43-TTEN
	2010-2011	Monahan, C., GAW-WDCA, EUSAAR, EMEP, 2010-2011, Particle_number_size_distribution at Mace Head, data hosted by EBAS at NILU, DOI: https://doi.org/10.48597/SG52-MN3P
	2011-2012	O'Dowd, C., GAW-WDCA, ACTRIS, EMEP, 2011-2012, Particle_number_size_distribution at Mace Head, data hosted by EBAS at NILU, DOI: https://doi.org/10.48597/ZQ5A-MMDK
	2012-2013	O'Dowd, C., GAW-WDCA, ACTRIS, EMEP, 2012-2013, Particle_number_size_distribution at Mace Head, data hosted by EBAS at NILU, DOI: https://doi.org/10.48597/TX72-EJEZ
	2016-2018	Personal correspondence with Dr Ceburnis, University of Galway

Table S4: References for each ACTRIS dataset used in the study.

Dataset/ Network	**Database**
Actris	http://ebas.nilu.no/
ERA5	https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels
GLOB-Colour	https://doi.org/10.48670/moi-00281
SPOT VGT	10.24381/cds.7e59b01a
Ostia	ttps://doi.org/10.48670/moi-00168?
ESACCI	https://maps.elie.ucl.ac.be/CCI/viewer/download.php
GEBCO	https://www.gebco.net/data-products/gridded-bathymetry-data/gebco-2019
CEDS	https://data.pnnl.gov/group/nodes/dataset/13488
GFEDS	https://www.geo.vu.nl/~gwerf/GFED/GFED4/

Table S5: Databases for the datasets used in the study.

S2 Supplementary results

S2.1 Feature selection results

	TRH 30-80nm	**TRH 80-660nm**
Height weighted mean	**X**	**X**
BL height weighted mean	**X**	**X**
RH weighted mean	**X**	**X**
Temperature weighted mean	**X**	**X**
Time in boundary layer	**X**	**X**
Time in cloud	**X**	**X**
Surface solar radiation weighted sum	**X**	**X**
Trajectory speed mean	**X**	**X**
Precipitation weighted sum	**X**	**X**
Precipitation sum	**X**	**X**
Precipitation reverse weighted sum	**X**	**X**
Snowfall weighted sum		**X**
Chlorophyll weighted sum	**X**	**X**
LAI weighted sum		
10m windspeed weighted mean	**X**	**X**
10m windspeed std	**X**	**X**
10m windspeed mean	**X**	**X**
10m windspeed max	**X**	**X**
2m temperature weighted mean	**X**	**X**
Time over sea ice	**X**	**X**
Sea ice weighted mean	**X**	**X**
Time over sea	**X**	**X**
Time over deciduous forest	**X**	**X**
Time over evergreen forest		
Time over shrub	**X**	**X**
Time over cropland		**X**
Time over urban		

BC weighted sum	**X**	**X**
SO2 weighted sum	**X**	**X**
NH3 weighted sum	**X**	**X**
NOx weighted sum	**X**	**X**
OC weighted sum	**X**	**X**
BC biomass low sum		
BC biomass high sum		
OC biomass low sum		
OC biomass high sum		
Arrival hour	**X**	**X**

Table S6: Results from the recursive feature selection described in Sect. 3.4, where an X indicates that the feature was included in the final model.

S2.2 Hyperparameter tuning results

	TRH 30-80nm	**TRH 80-660nm**
Number of estimators	200	450
Max depth	9	9
Learning rate	0.07	0.06
Min child weight	3	5
Min split loss	2	0
Subsample	0.94	0.62
Colsample by tree	0.97	0.81
Lambda	33.19	34.48
Alpha	9.73	0.50
Test score	0.65	0.71

Table S7: Results of the hyperparameter tuning performed using Tree-based Parzen Estimators, as described in Sect. 3.4 of the main text.

S2.3 Model evaluation

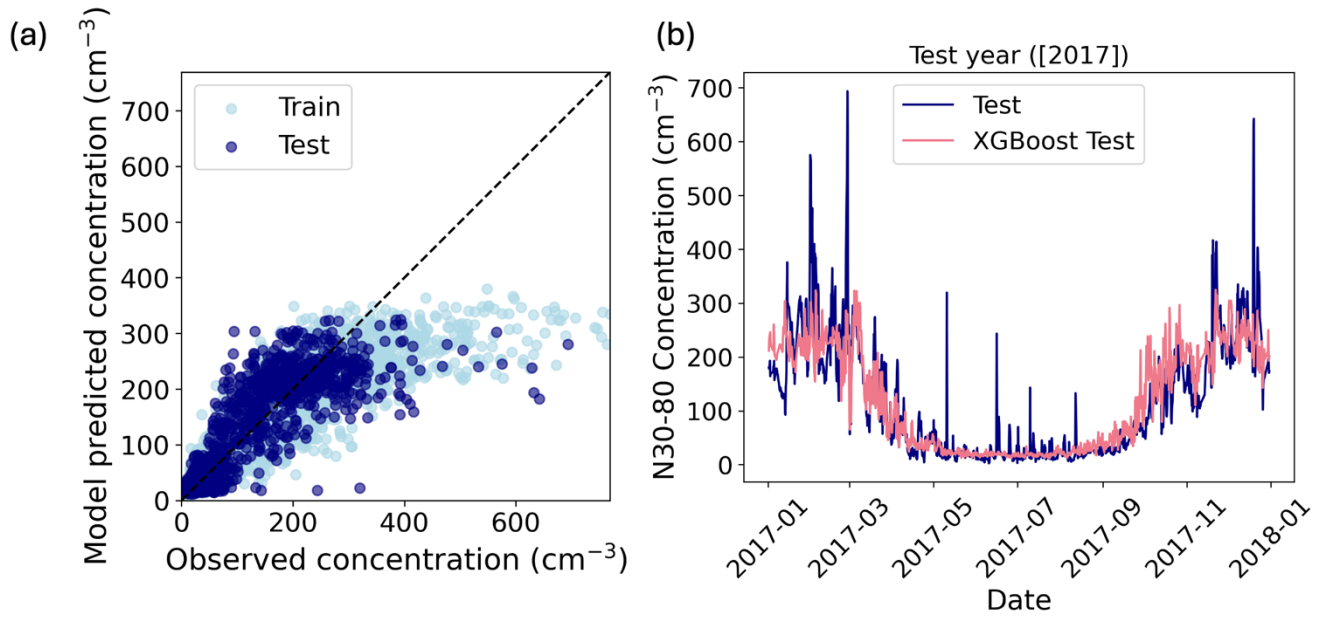


Figure S4: (a) Correlation plot for observed concentrations (30-80nm) against ML model prediction concentration (30-80nm) for the training data set (light blue) and test dataset (navy). (b) Timeseries for observed concentrations (navy) and ML model predictions (pink). For Trollhaugen.

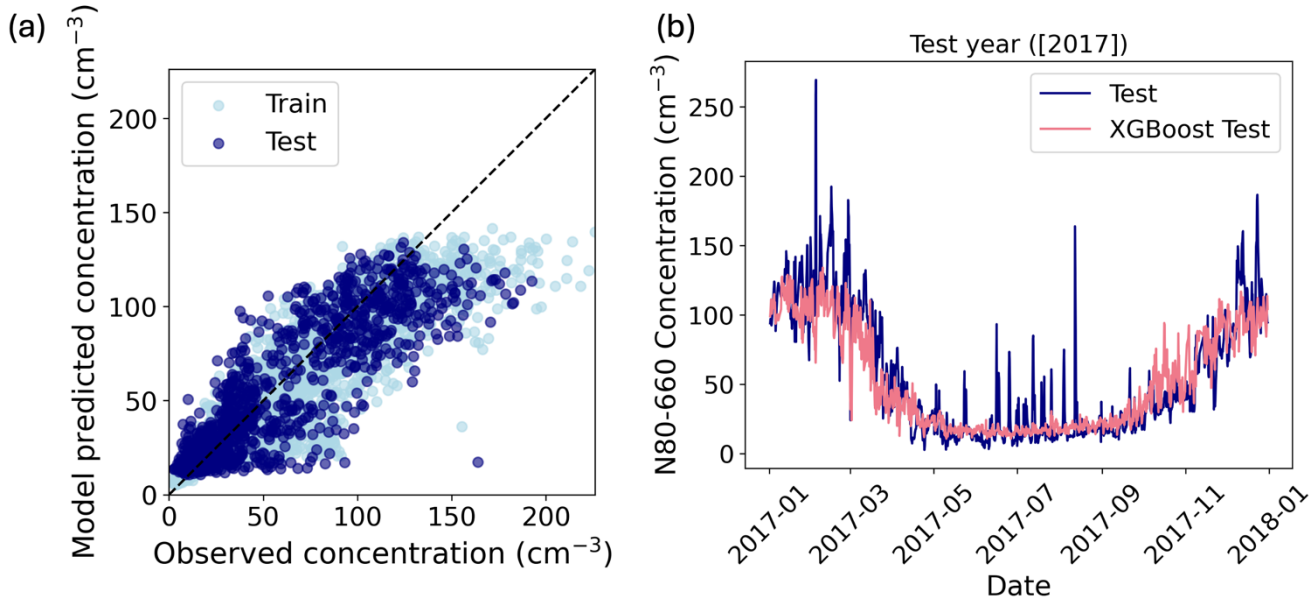
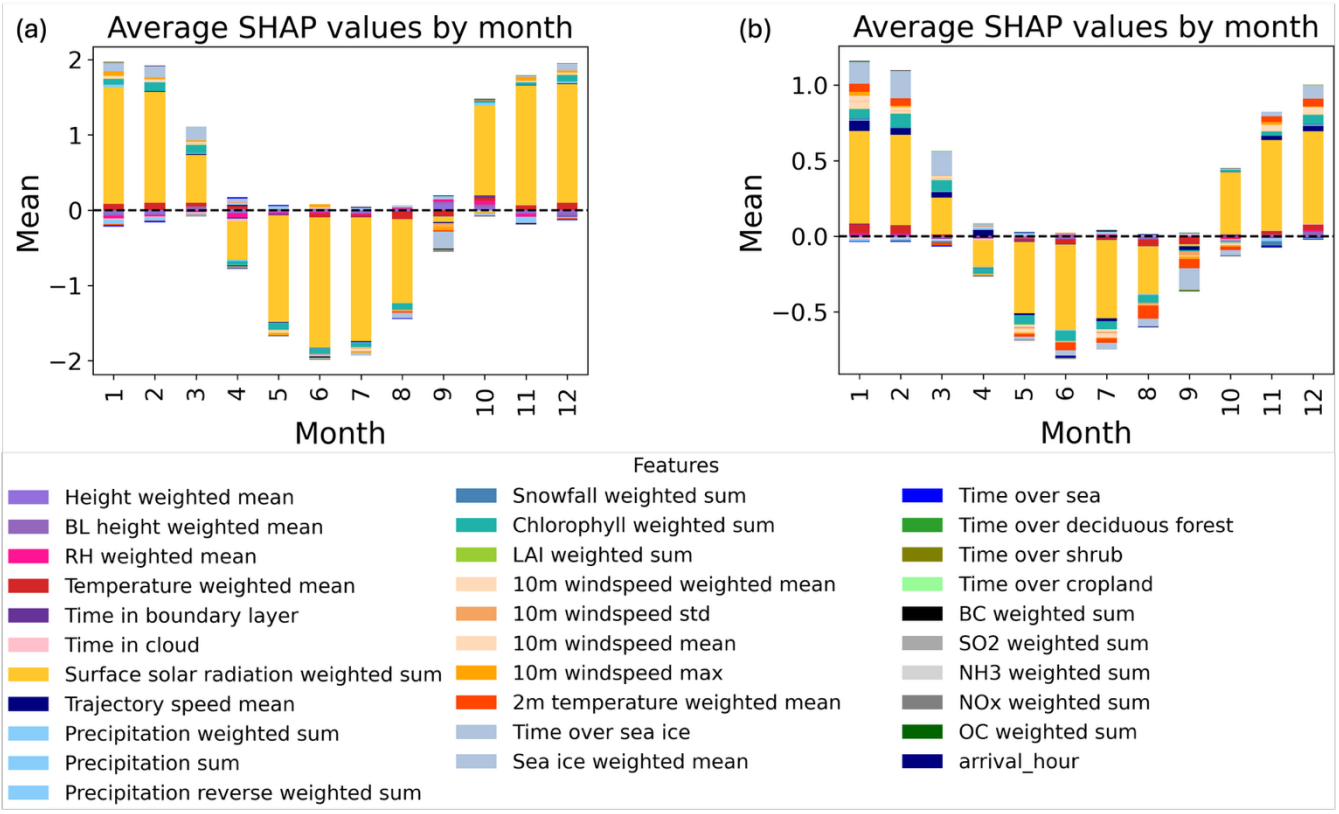


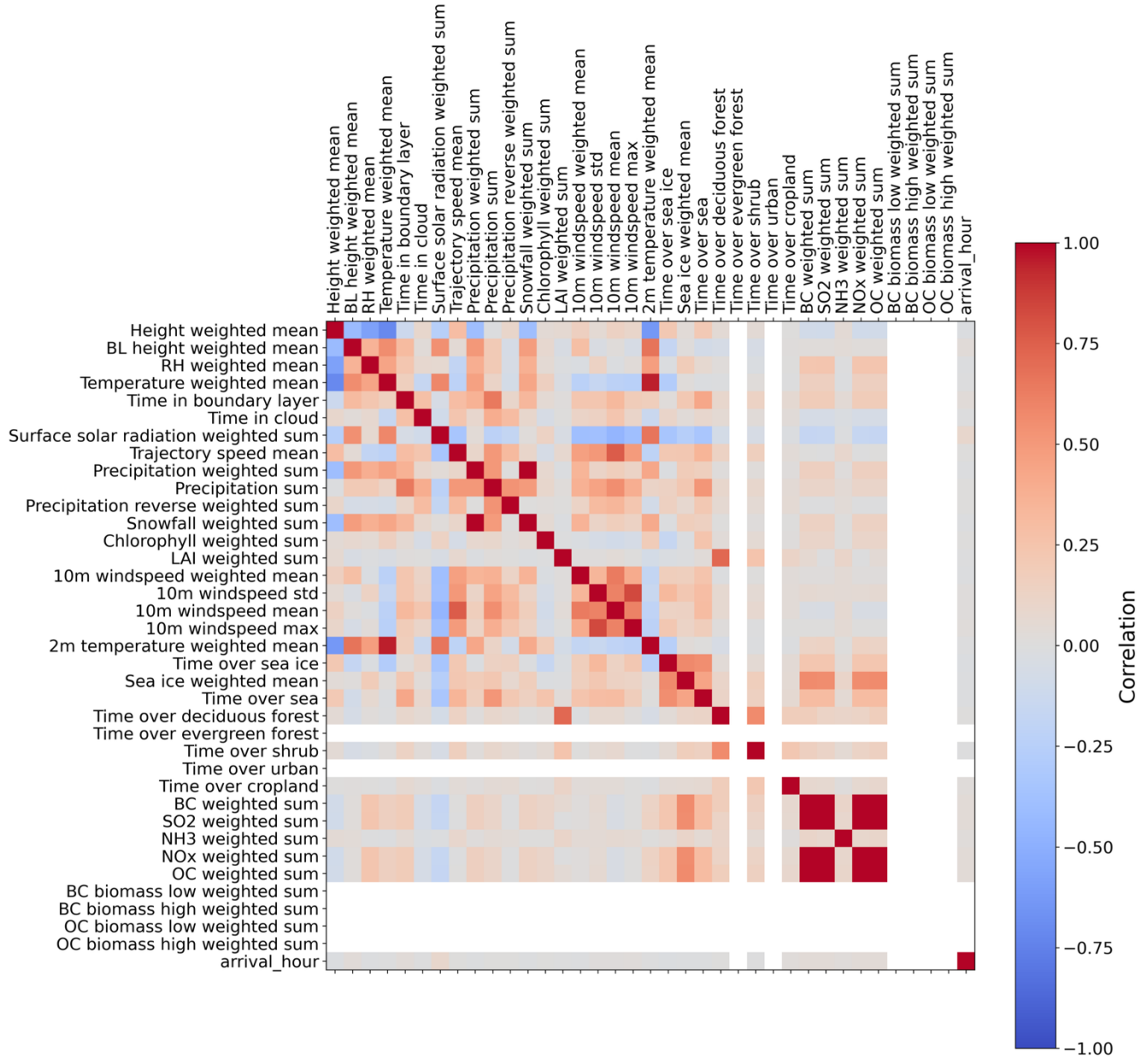
Figure S5: (a) Correlation plot for observed concentrations (80-660nm) against ML model prediction concentration (80-660nm) for the training data set (light blue) and test dataset (navy). (b) Timeseries for observed concentrations (navy) and ML model predictions (pink).

S2.4 Seasonal cycle of SHAP



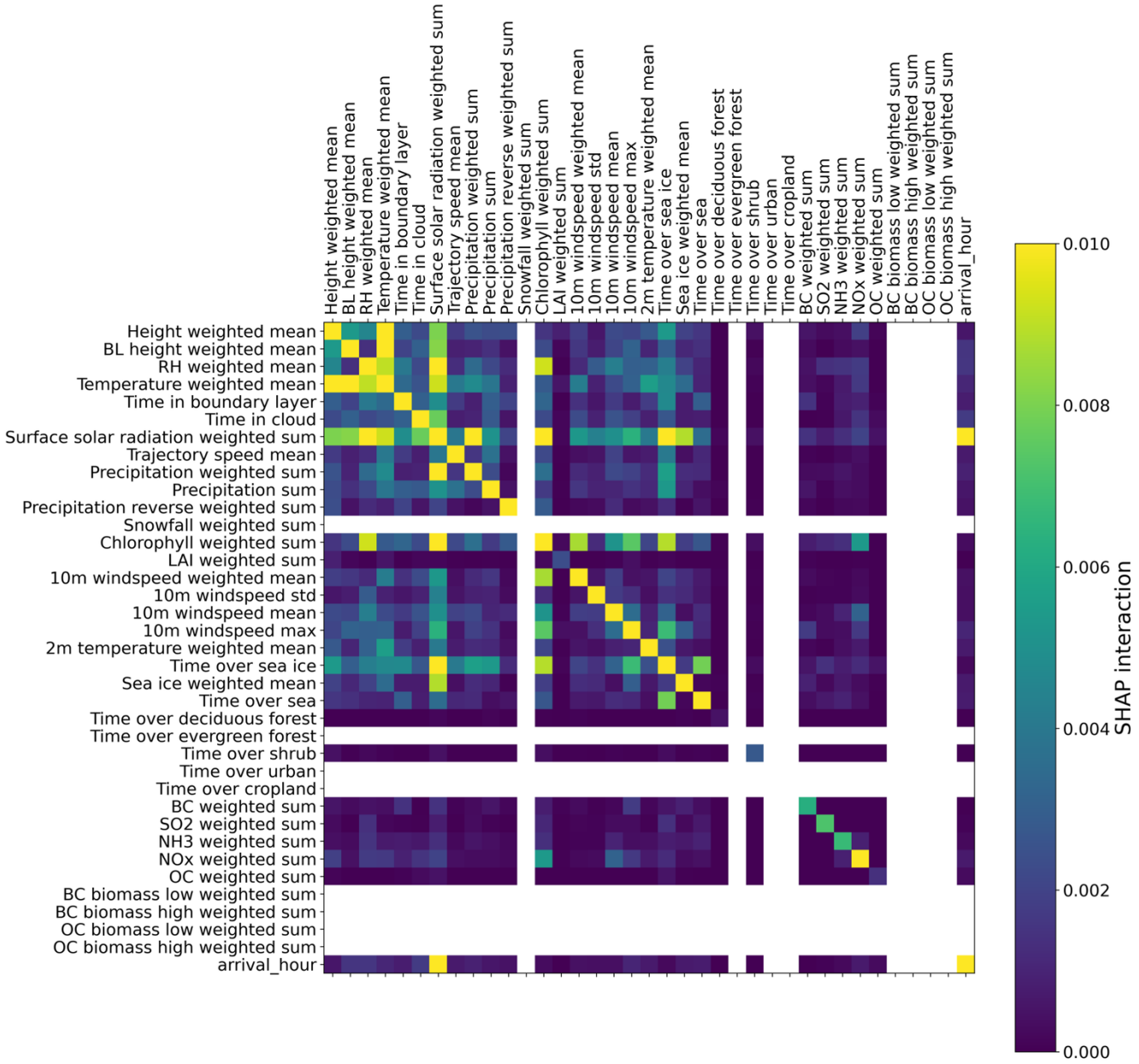
125 **Figure S6: The average monthly SHAP value for each feature. The size of the bar corresponds to the magnitude of the mean SHAP value of the feature for each month. Note that the placement above or below the x-axis demonstrates the sign of the mean SHAP value of the feature, however that the order of features within the bar does not have significance.**

S2.5 Correlation matrix



130 **Figure S7: Correlation coefficients for all explanatory variables in the test dataset (2017). White indicates that the feature was not included in the final models.**

S2.6 SHAP interaction matrix



135 **Figure S8: Absolute mean SHAP results for the 30-80nm model. Where the diagonal elements are the total SHAP value for the feature and the other elements represent the other features contribution to the total SHAP of a feature. White indicates that the feature was not included in the final model.**

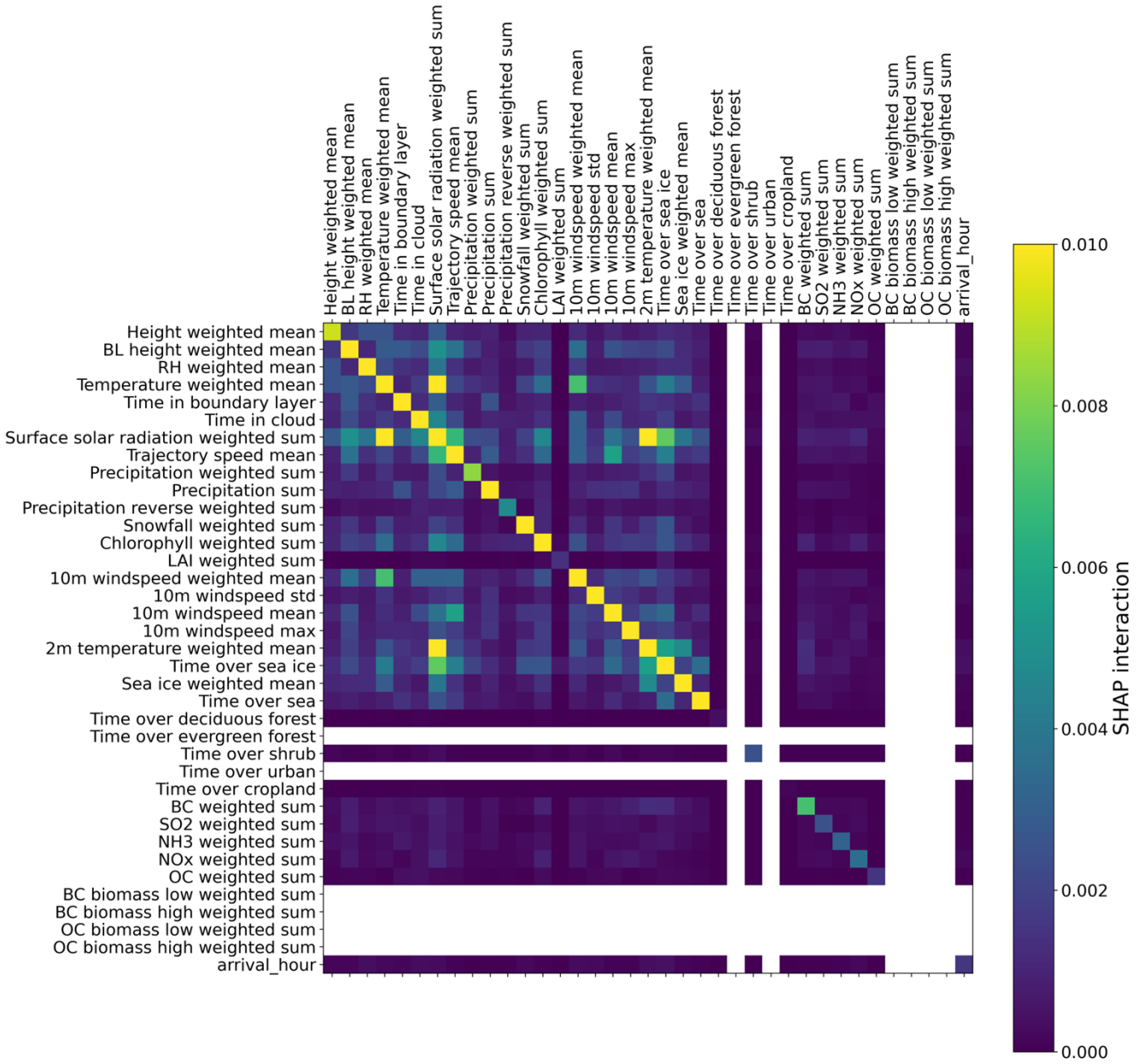


Figure S9: Absolute mean SHAP results for the 80-660nm model. Where the diagonal elements are the total SHAP value for the feature and the other elements represent the other features contribution to the total SHAP of a feature. White indicates that the feature was not included in the final model.

S2.7 Interventional SHAP

We also conduct the TreeSHAP analysis with the 'interventional' approach to ensure the robustness of the SHAP results.

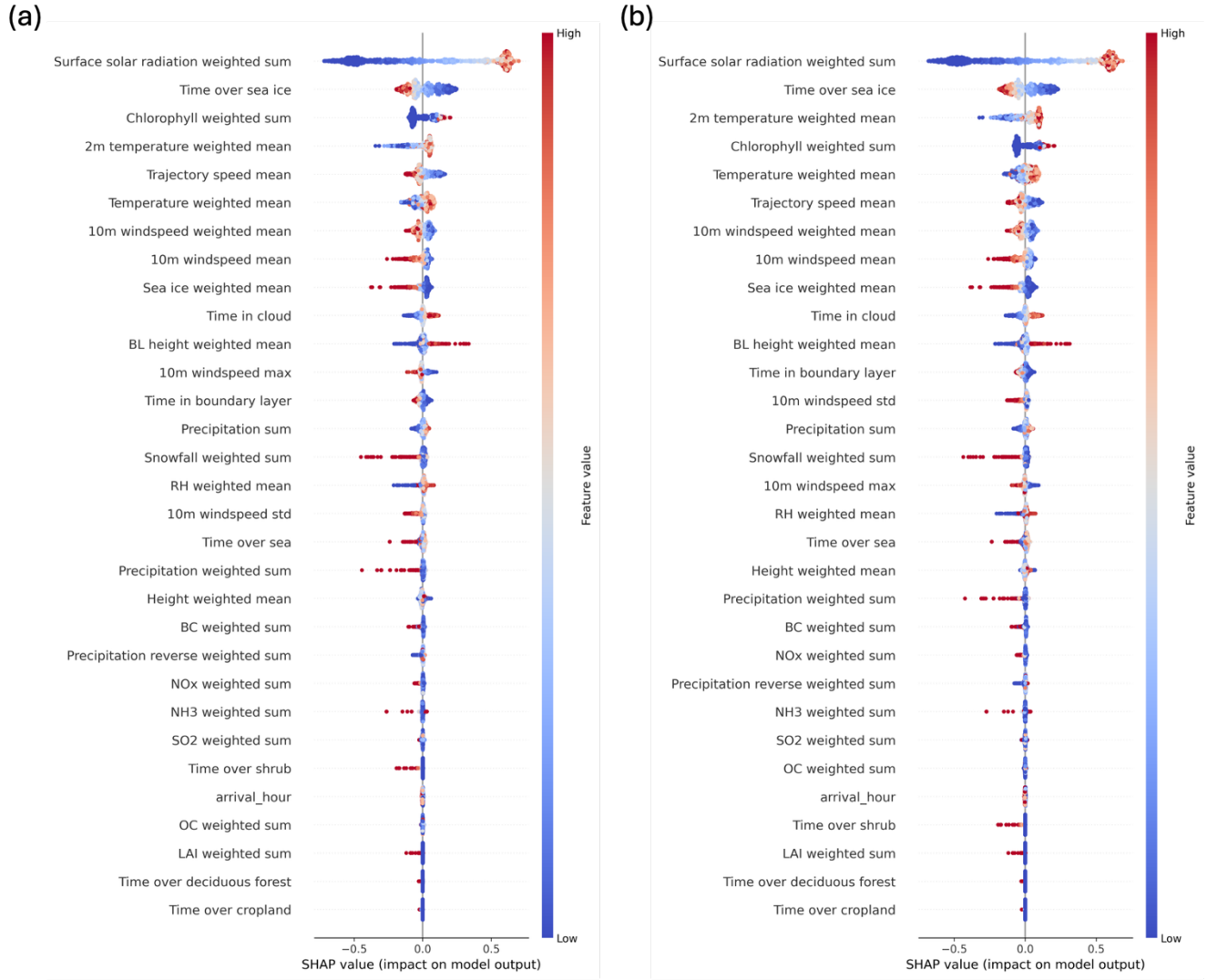


Figure S10: SHAP beeswarm for Trollhaugen N80-660nm model (a) path-dependent TreeSHAP and (b) interventional. Plots are ordered by the rank of each feature as determined by the TreeSHAP analysis. The points are coloured by the feature value corresponding to the data point for the SHAP value, and the 'violin' shape of the distribution represents the density of points.

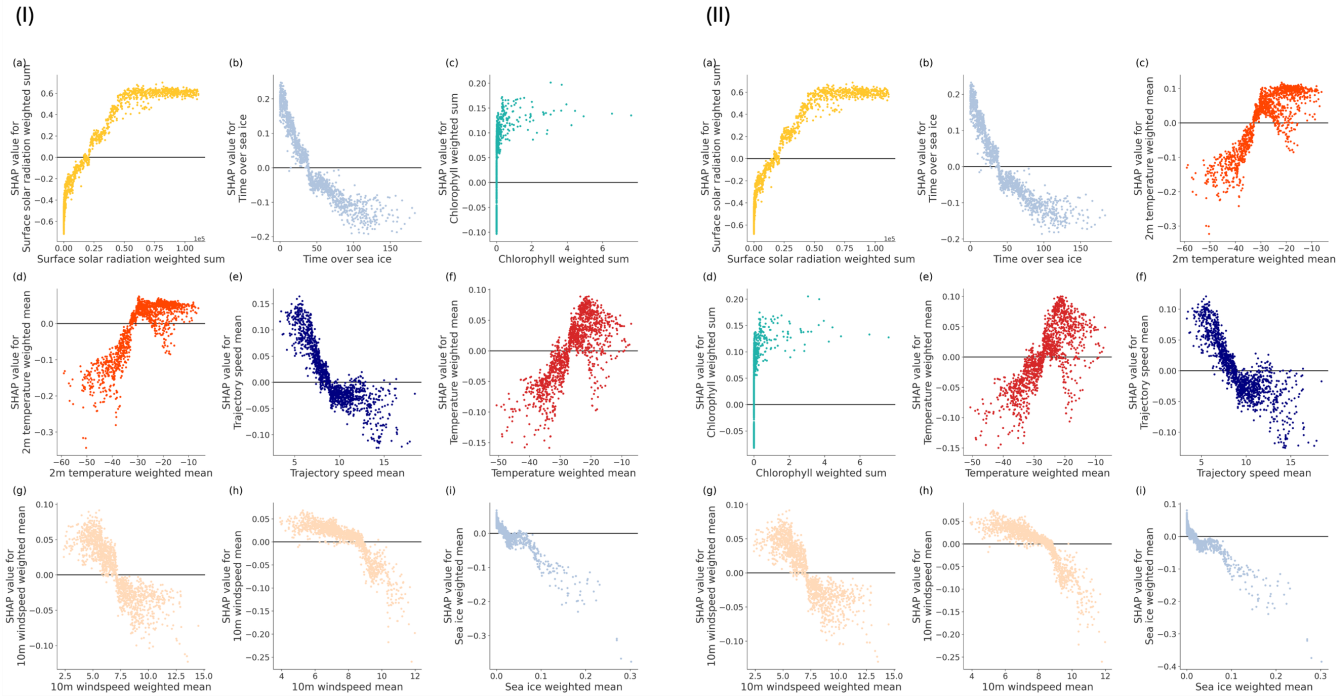


Figure S11: SHAP dependence plots for the Trollhaugen N80-660nm model. Plots are ordered by the rank of each feature as determined by the TreeSHAP analysis and the top 9 ranked features are shown. Colours are associated with each variable. (I) Path-dependent TreeSHAP, (II) interventional TreeSHAP.