Supplementary Information

S1. Overview of measurements

Table S1. Overview of the measurements used in this study.

Parameter	Instrument	Campaigns
Particle number size distribution	Scanning Electrical Mobility Spectrometer 10-1000 nm (SEMS, Brechtel, USA)	March 2021, December 2021, Feb-Mar 2022, March 2023
	Scanning Mobility Particle Sizer 14-685 nm (SMPS, TSI, USA)	October 2016, October 2019
	Aerodynamic Particle Sizer 700-5000 nm (APS, TSI, USA)	October 2016, October 2019, December 2021, Feb-Mar 2022, March 2023
Particle number concentration	Condensation Particle Counter (CPC, 3782 TSI, USA)	March 2021
	Condensation Particle Counter (CPC, 3010 TSI, USA)	October 2016, October 2019
	Condensation Particle Counter (CPC, A20 Airmodus, Finland)	March 2023
	Mixing Condensation Particle Counter (mCPC, Brechtel, USA)	December 2021, Feb-Mar 2022
Cloud condensation nuclei concentration	Cloud Condensation Nuclei counter (CCN-100/200, Droplet Measurement Technologies, USA)	October 2016, October 2019, March 2021, December 2021, Feb-Mar 2022, March 2023
Black carbon concentration	Aethalometer (AE33, Aerosol Magee Scientific, Slovenia)	October 2016, October 2019, March 2023
	Tricolour Absorption Photometer (TAP, Brechtel, USA)	March 2021, December 2021, Feb-Mar 2022
CO_2 concentration	CO_2 analyser (CA-10, Sable, USA)	October 2016, March 2023
NO_x concentration	NO_x analyser (9841A, Ecotech, Australia)	March 2023
CO concentration	CO analyser (9830B, Ecotech, Australia)	October 2016, March 2023
O_3 concentration	O_3 analyser (9810B, Ecotech, Australia)	October 2016, October 2019, March 2023
Meteorological parameters	Weather station (Lufft WS800-UMB, Germany)	March 2021
	Weather station (Gill MaxiMet GMX501, Gill Instruments Ltd., UK)	December 2021, Feb-Mar 2022, March 2023
	RV Investigator's built-in suite of weather sensors	October 2016, October 2019

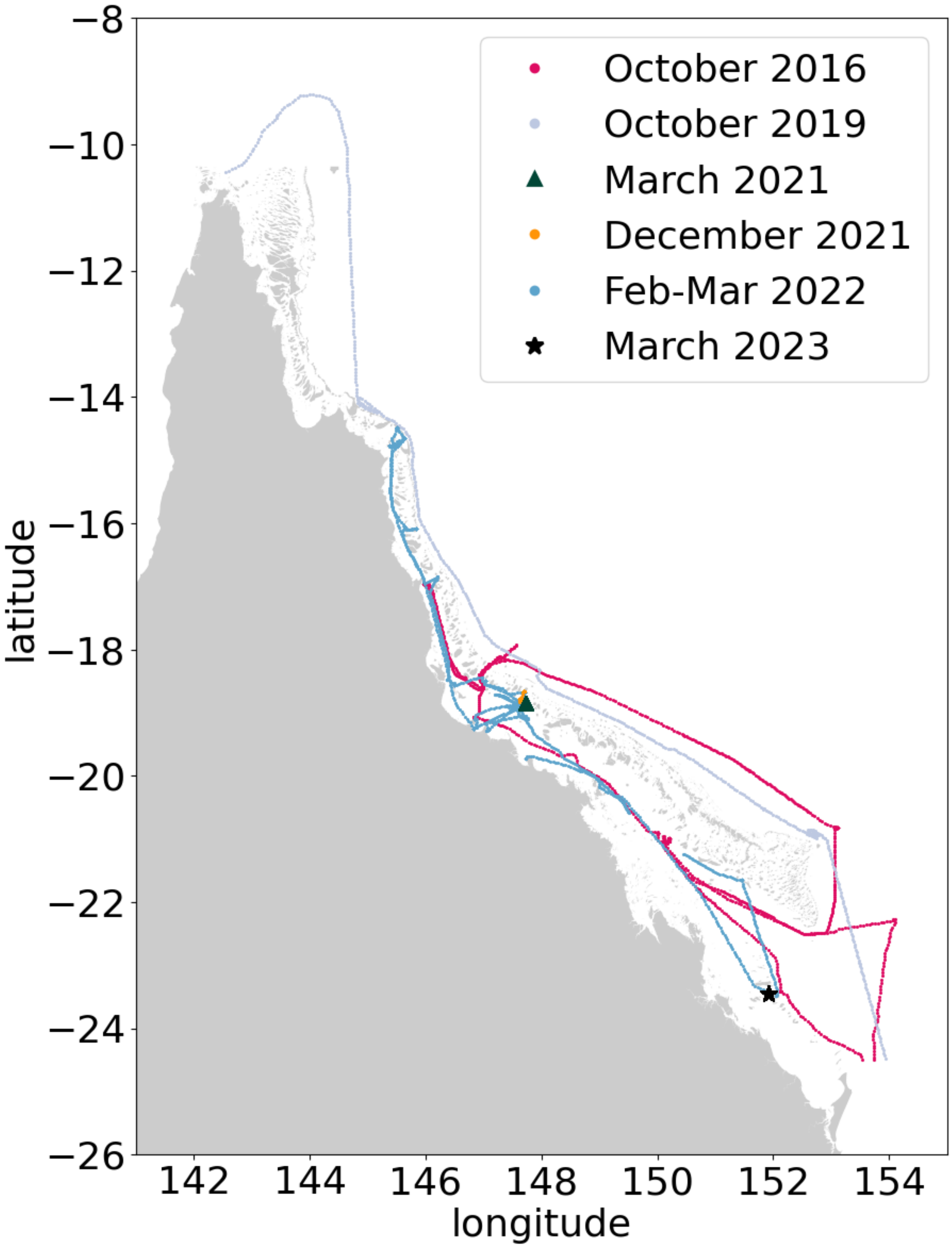


Figure S1: Spatial distribution of measurements used in this study, grouped by their respective campaigns.

S2. Density distributions and correlation analysis

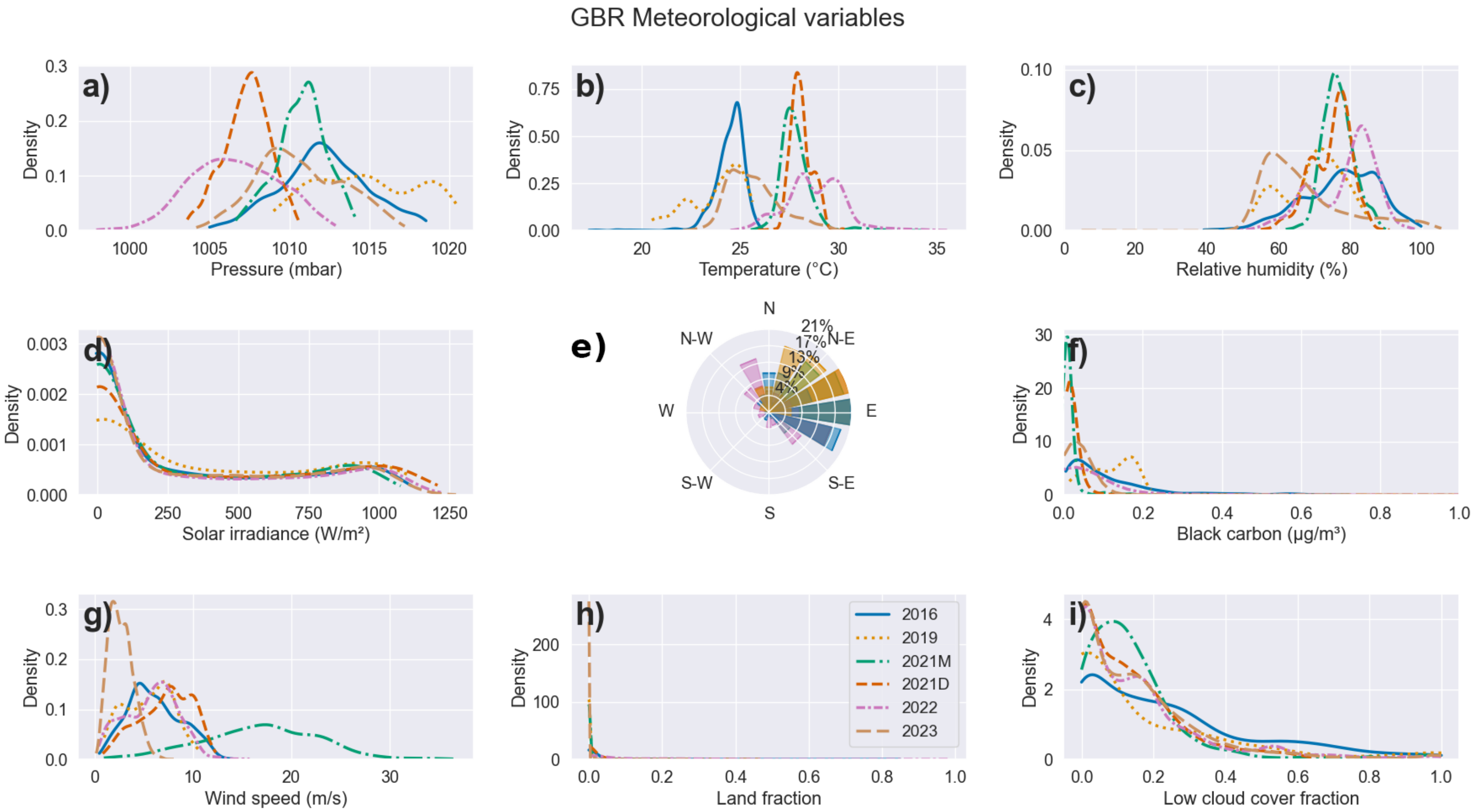


Figure S2: Kernel density estimates for in-situ meteorological variables collected at the GBR during spring (blue) and summer (orange). The variables shown are, pressure (a), air temperature (b), relative humidity (c), solar irradiance (d), wind direction (e), black carbon mass concentration (f), wind speed (g), land influence as fraction of land influenced back trajectory (h) and low cloud cover fraction (i). Land fraction is calculated from 72-hour back trajectories using the HYPSLIT model. Low cloud cover fraction is from ERA5 reanalysis data product.

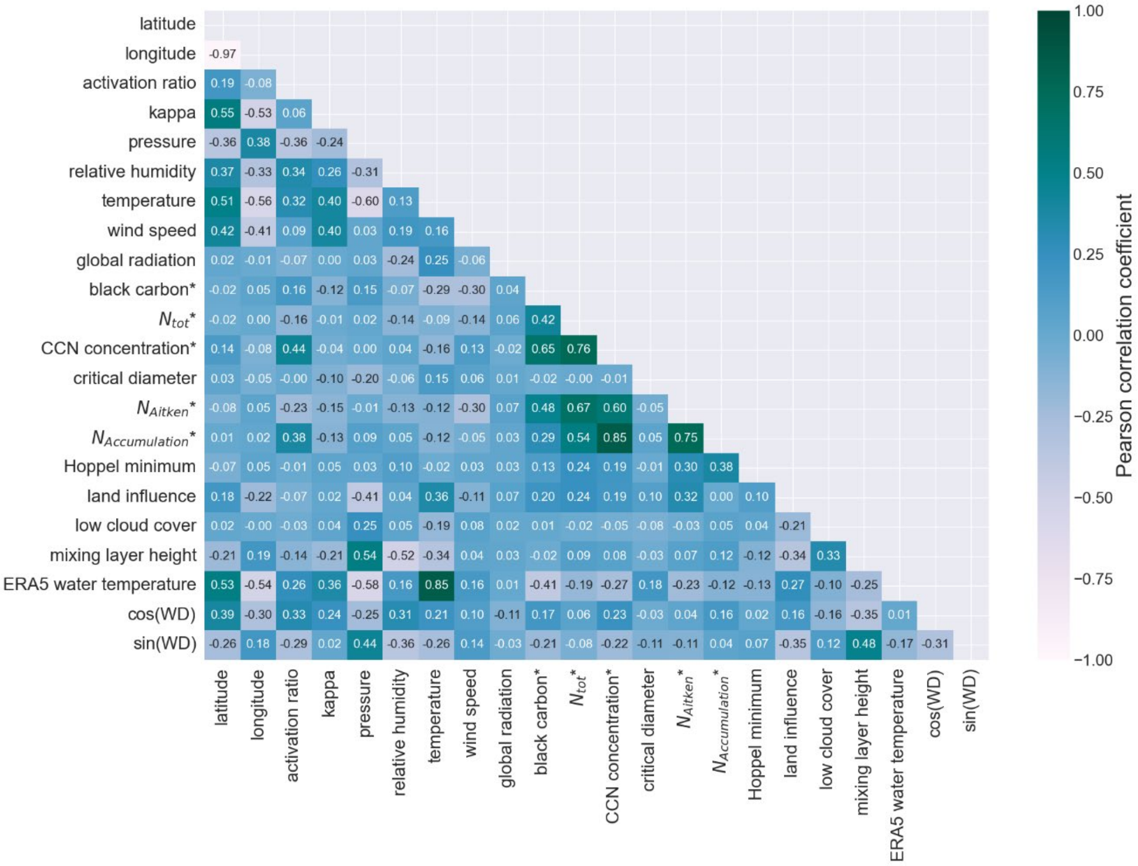


Figure S3: Correlation heatmap for main variables analysed in this study. Logarithmic values were used to calculate Pearson correlation coefficients for variables marked with *. Wind direction is presented by cosine and sine values of the wind direction as linear variables instead of its nominal polar coordinate representation.

S3. Spatiotemporal distibutions

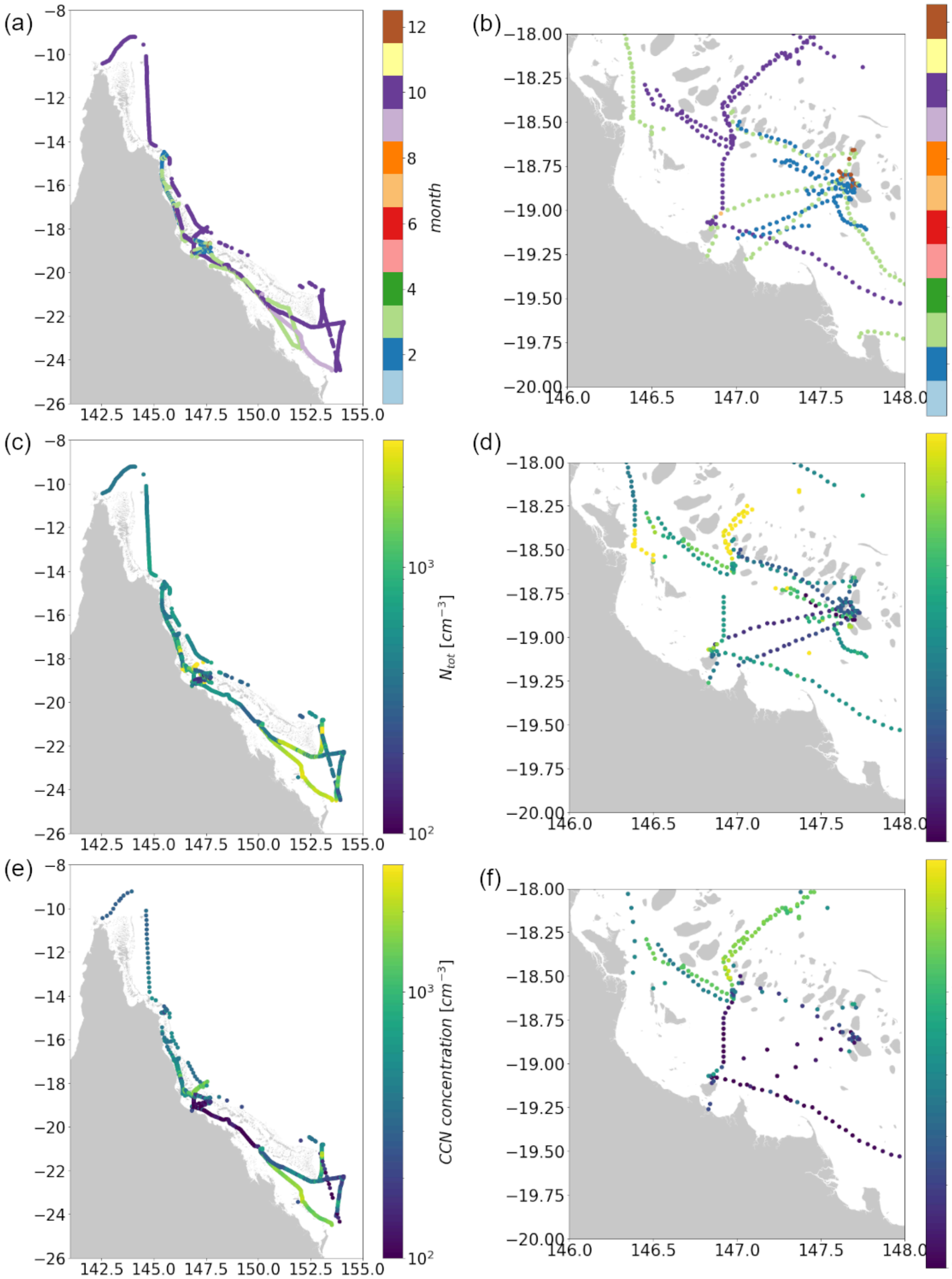


Figure S4: Spatial distribution of measurement month (a, b), N_{tot} (c, d) and CCN concentration (e, f) over GBR. Subplots (a,c,e) shows spatial distribution for entire GBR, while subplots (b,d,f) focus on central part of GBR where most of campaigns overlap. The median of data with resolution of 0.01 latitude and longitude was calculated for clearer visualization of stationary and overlapping data.

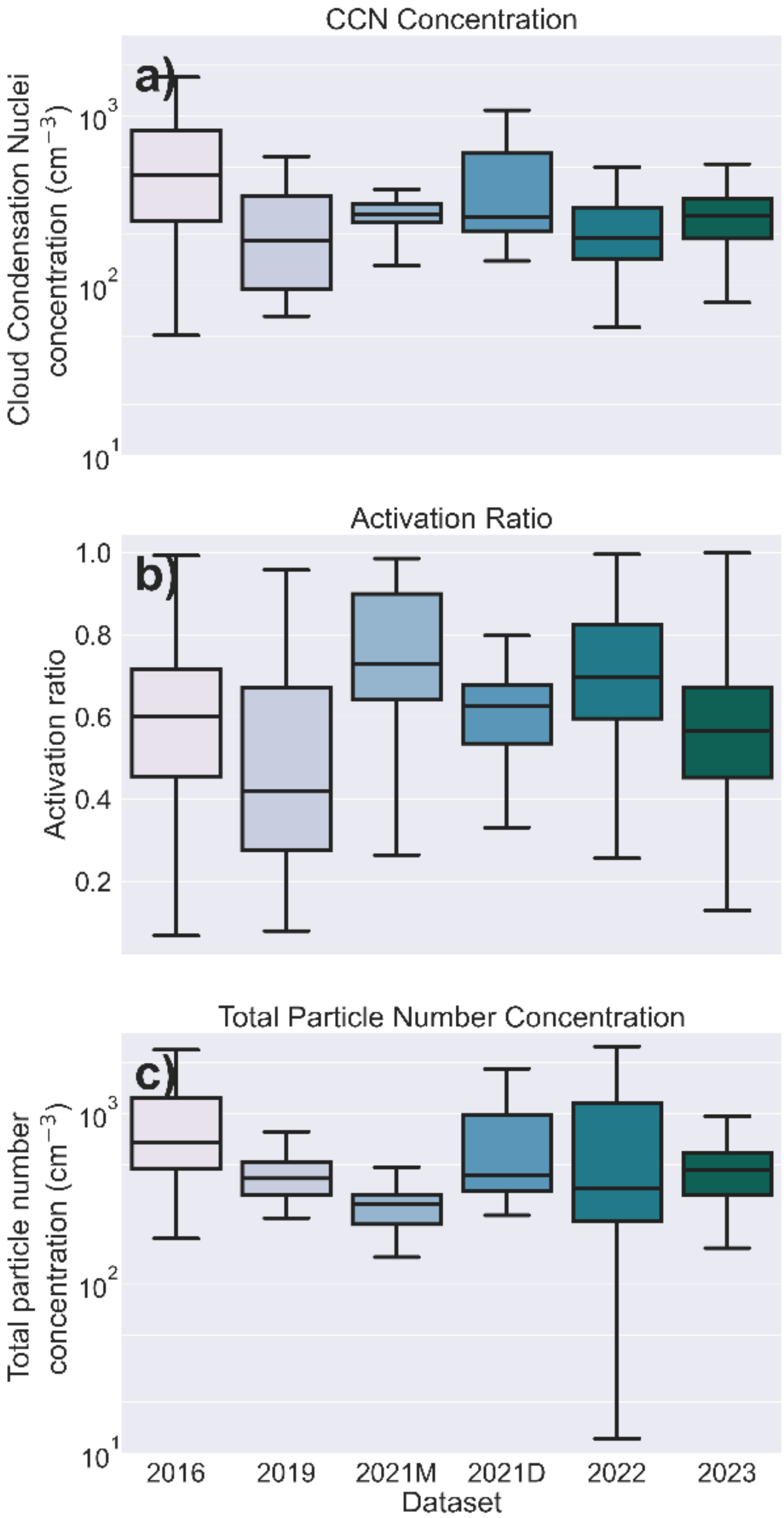


Figure S5: Temporal variability of cloud condensation nuclei (CCN) concentration (a), activation ratio (b), and total particle number concentration (c) excluding data from northern part of the GBR. The middle line of a boxplot represents the median value, the upper and lower edges of a box show the interquartile range, and the whiskers show the entire range.

S4. CCN concentration drivers analysis

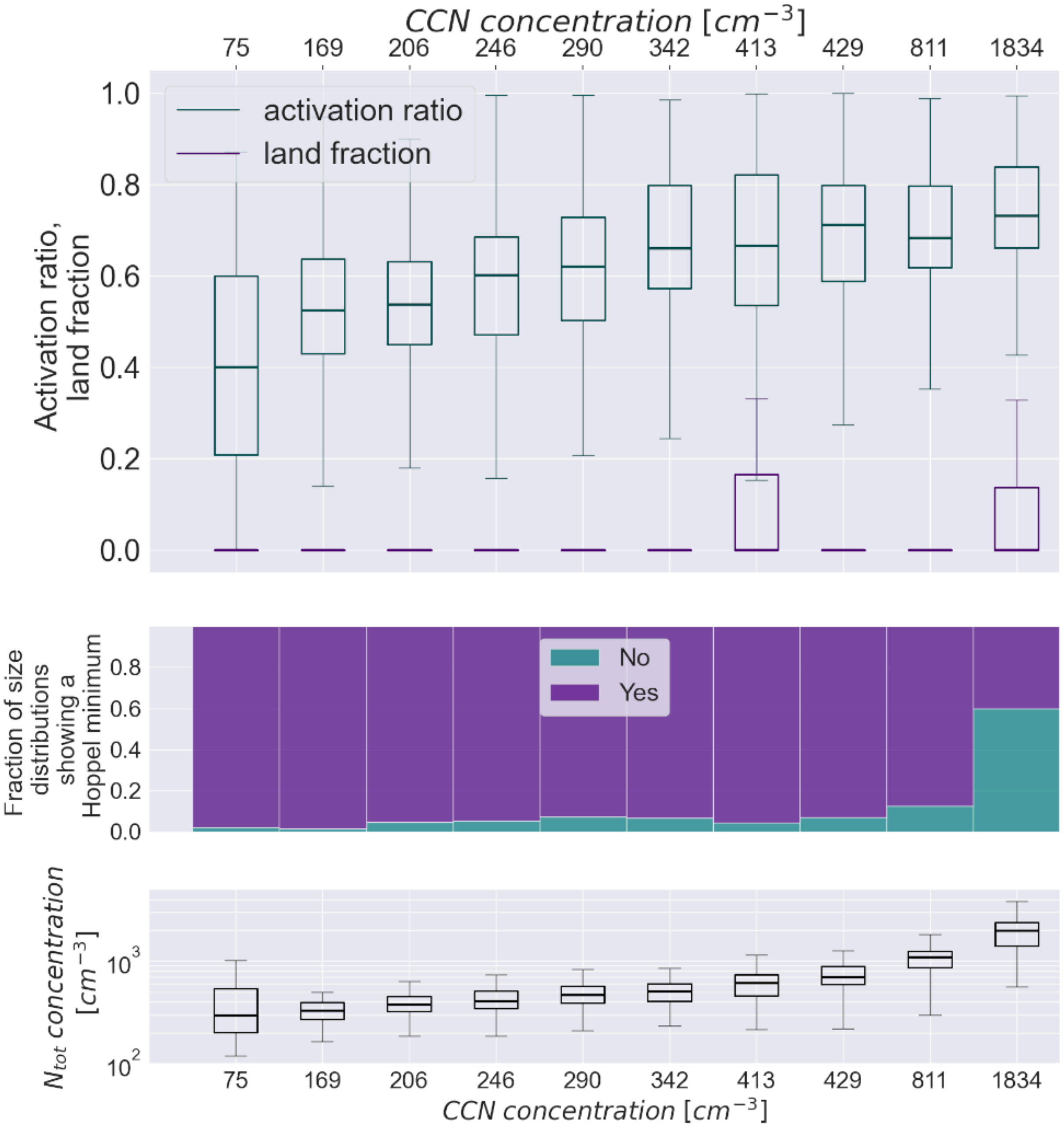


Figure S6: Activation ratio, land fraction, and fraction of data with no Hoppel minimum present as well as total number concentration of particles for different CCN concentration bins. CCN concentration bins were created based on CCN concentration quantiles, ensuring equal distribution of data points between analysed bins. Values of activation ratio, land fraction, and total number concentration of particles are visualized by boxplots in which the middle line represents the median value, the upper and lower edges of a box show 75 % and 25% percentiles, and the whiskers show the entire range.

Table S2. Quantiles of CCN concentration and tertile ranges of activation ratio and the number of data points used for cloud processing analysis.

		Activation Ratio		
	Number of data points	Low (0.03 – 0.53)	Medium (0.53 – 0.69)	High (0.69 – 0.99)
CCN concentration	Q1 (51 - 240 #/cm3)	530	268	112
	Q2 (240 - 450 #/cm3)	317	344	219
	Q3 (450 - 810 #/cm3)	166	274	378
	Q4 (810-2650 #/cm3)	117	243	455

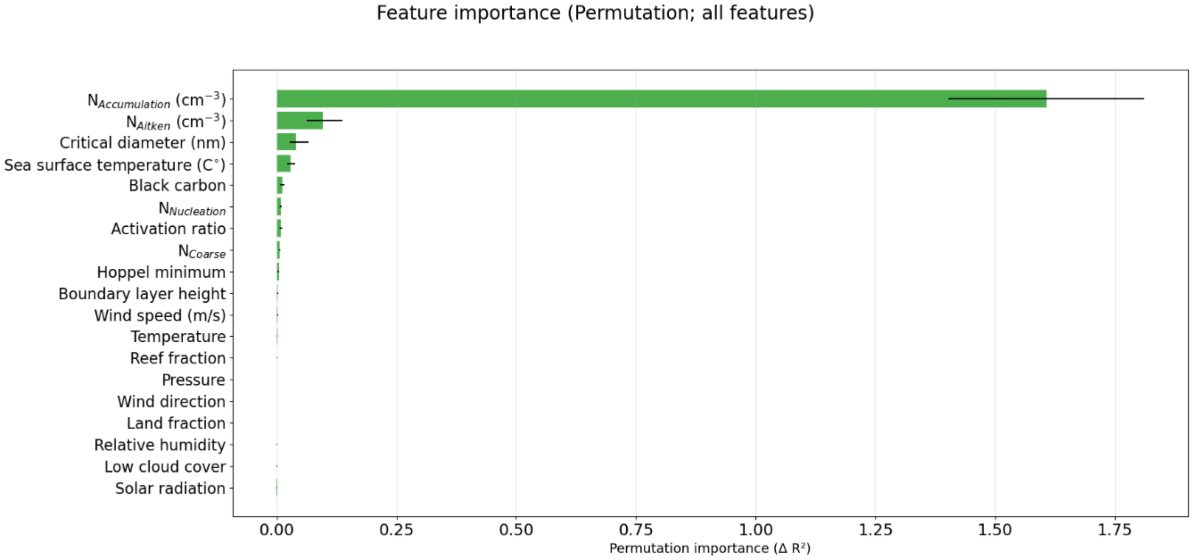


Fig S7: Permutation feature importance of climate variables in explaining CCN concentration using every variable in our dataset for spring (left) and summer (right). The differences between original and permuted mean square errors on the x-axis indicate how important each feature is in predicting the CCN concentration, with larger difference between original and permuted values indicating a more important feature. The black whiskers indicate 95% confidence intervals. The number of permutations used was 10000.

S5. SHAP estimates and partial dependency plots

The directionality of the significant (95% confidence intervals > 0) features in the gradient boosting regression algorithm model explaining CCN concentration over GBR (Fig. 4) is visualized by SHAP estimates (Fig. S8a) and partial dependency plots (Fig. S8 b-f). The SHAP estimate plot shows the impact on model output (SHAP value) by each of features and the datapoints are coloured by relative value for each feature. Color scaling is applied independently to each feature, giving a relative scale. For example, looking at the accumulation mode particles (Fig. S8a), we can see the highest (red) values are over 1000 in SHAP value, which means that high values of accumulation mode particles contribute strongly and positively to estimated CCN concentration. At the same time, critical diameter has the highest values (red) for negative SHAP values, which means that high values of critical diameter decrease predicted CCN concentration values. Furthermore, the absolute value of SHAP values for critical diameter is on average lower than for accumulation mode particle concentration which means that the impact of critical diameter on model output is smaller than accumulation mode particle concentration. Relations between feature values and model output can be further investigated by partial dependency plots that show the relation between measured values of features and their impact on model outcome. Partial dependency plots show values of both the feature and its corresponding effected on the predicted value. Hence, partial dependency plots (Fig. S8 b,d) confirm previous observations about accumulation mode particle concentration and critical diameter effects on modelled CCN concentration.

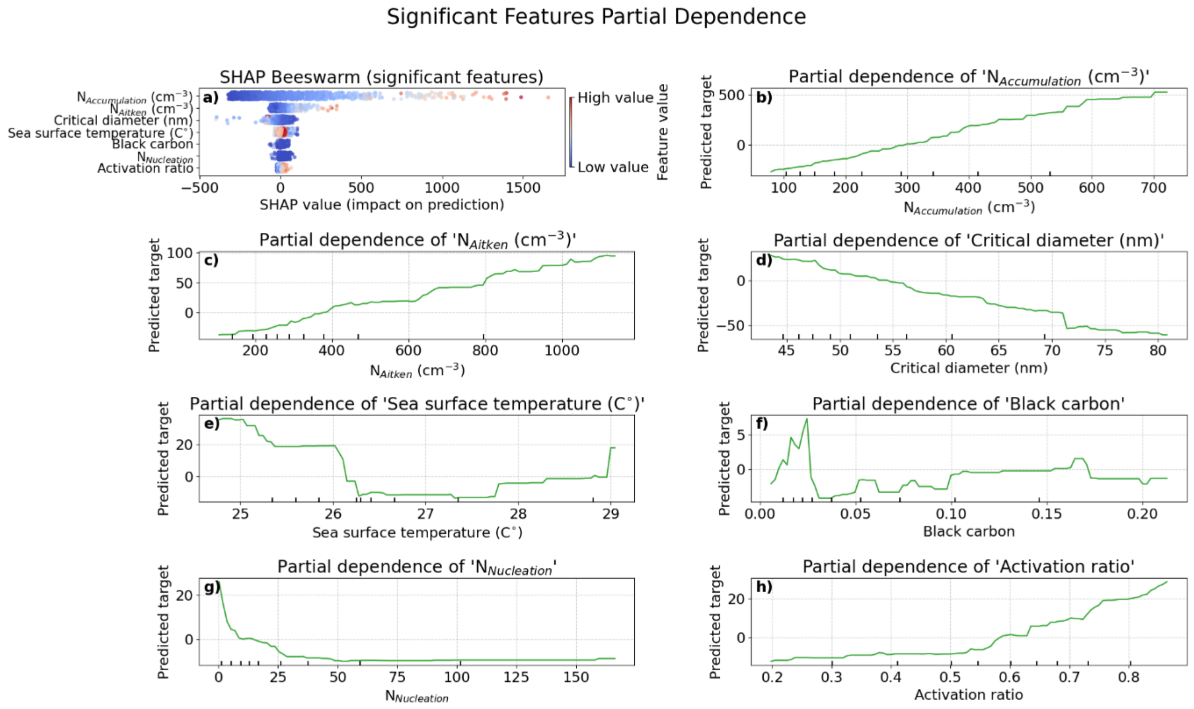
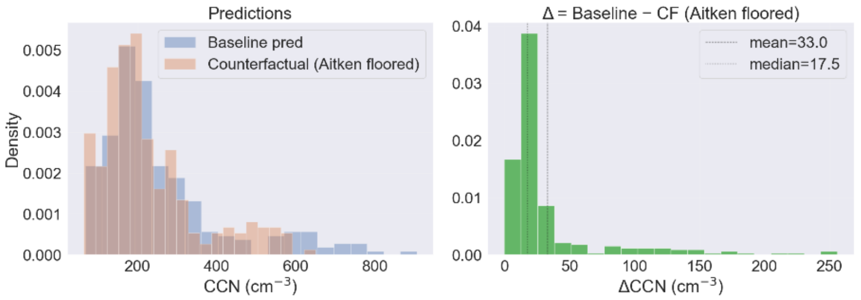


Figure S8: SHAP estimates (a) and partial dependency plots (b – h) for the seven statistically significant model inputs and how they drive CCN concentrations. a) Positive values represent a positive contribution to model prediction value and negative values a negative contribution. SHAP features are coloured by the relative values of each feature. b – h) Predicted contribution to model prediction as a function of feature values.

S6. Counterfactual modelling



Counterfactual: minimize Aitken
Feature: N_{Aitken} (cm^{-3})
Floor: 1.0th pct = 68.47

Mean ΔCCN: 33.0 (95% CI 13.5, 24.3)
Median ΔCCN: 17.5 (95% CI 14.3, 18.2)
n = 252
Batches kept: 16

Fig S9: Counterfactual modelling results for CCN predicting gradient boosting regression model. On the left, histograms of the baseline model (blue) and the counterfactual model (orange). In the middle, histograms of the difference between the baseline and the counterfactual model CCN predictions. On the right, numerical statistics for the counterfactual model, mean and median differences between the baseline and counterfactual models including confidence intervals, number of data points in the test set for each season and the amount of continuous batches used for block bootstrapping. The number of permutations used was 10,000.