https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

A framework to holistically investigate processes controlling the aerosol lifecycle using explainable AI techniques

Eliza K. Duncan[1], Jonathan E. Fieldsend[2], Alistair Sellar[3], Emanuele Tovazzi[1], Paul Kim[1], James M. Haywood[1] and Daniel G. Partridge[1].

[1]Department of Mathematics and Statistics, University of Exeter, Exeter, EX4 4QF, United Kingdom
[2]Department of Computer Science, University of Exeter, Exeter, EX4 4RN, United Kingdom
[3]Met Office, Exeter, EX1 3PB, United Kingdom

Correspondence to: Eliza K. Duncan (e.k.duncan@exeter.ac.uk) and Daniel G. Partridge (d.g.partridge@exeter.ac.uk)

Abstract. General circulation models (GCMs) face significant uncertainties in estimating Earth's radiative budget due to aerosol-cloud interactions (ACI). To improve the representation of ACI in GCMs it is crucial to constrain processes controlling the aerosol lifecycle and the resulting size distribution. This is challenging due to the complexity and number of competing atmospheric processes that interact over large spatial and temporal scales which require untangling to elucidate dominant processes controlling aerosol properties. This study aims to (a) develop a generic explainable AI framework from air-mass history to build an accurate representation of processes controlling aerosol properties, from this, (b) identify key relationships between aerosol processes and their impacts on observed aerosol number concentrations, and (c) provide robust process-based observational constraints to aid in the isolation of GCM structural uncertainties. This is achieved by developing XGBoost regression models to simulate Aitken and accumulation mode number concentrations for receptor surface stations and application of TreeSHAP to identify key processes from explanatory variables describing meteorological and aerosol processes collocated to Lagrangian air-mass trajectories. The fidelity of this framework is demonstrated for the Antarctic station Trollhaugen, situated in a pristine region in which GCMs exhibit significant biases. Aerosol number concentrations at Trollhaugen were shown to be dominated by marine sources as well as transport from the free troposphere. The contribution from aloft dominates aerosol burden of the Aitken mode in the transitions between summer and winter, in contrast to a larger contribution in the summer from local marine sources from transport in the boundary layer.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

1 Introduction

Aerosols remain one of the largest sources of uncertainty in climate modelling, with the difficulty in constraining the natural baseline being a significant compounding factor for accurately predicting future climate scenarios (Bellouin et al., 2020; Carslaw et al., 2013; Sherwood et al., 2020). Aerosol-cloud interactions (ACI) contribute the largest uncertainty to the radiative forcing (RF) from aerosols (Bellouin et al., 2020), with the strong sensitivity in pristine conditions contributing significantly to the total uncertainty in ACI (Gryspeerdt et al., 2023; McCoy et al., 2020; Regayre et al., 2020). The non-linearity of the sensitivity of the anthropogenic forcing to the natural baseline significantly impedes our ability to accurately predict future climate scenarios (Carslaw et al., 2013).

The lack of observations in the pre-industrial period is one of the factors that limits our ability to constrain the natural aerosol baseline in general circulation models (GCMs) and to enable estimations of ACI forcing. To improve understanding of natural aerosol processes, case studies of pristine regions are frequently analysed to minimise the anthropogenic influence (Hamilton et al., 2014; McCoy et al., 2020; Schmale et al., 2019). However, GCMs have been shown to demonstrate significant bias in these regions with significant underpredictions of aerosol concentrations being reported over the Antarctic and Southern Ocean (e.g. McCoy et al., 2020, 2021; Mulcahy et al., 2020; Regayre et al., 2020). Improving the poor representation of aerosols and thus cloud properties is of paramount importance to reduce known biases in the planetary albedo over the Southern Ocean (Fiddes et al. 2024). Improvement in representation of natural aerosol in pristine regions promises to improve natural aerosol processes globally and thus provide an improved constraint on RF.

Additionally, in future climates it is thought that aerosol-climate feedbacks will have significant impacts, acting to enhance or dampen RF but these are currently poorly constrained in GCMs (e.g. Blichner et al., 2024). A study by Paasonen et al. (2013) across different regions found that concentrations of cloud condensation nuclei (CCN) increase with increasing temperatures, this feedback was found to be particularly strong for clean environments. Polar regions are some of the most pristine regions on Earth and are also experiencing rapid changes owing to polar amplification effects. These changes such as retreating sea ice (Dall'Osto et al., 2017; Struthers et al., 2011) and changes in atmospheric transport patterns (Pernov et al., 2022) may strongly impact aerosol sources (Schmale et al., 2022). Both climate induced, and policy induced changes could result in rapid alterations to dominant aerosol processes globally. To understand the full impacts of climate change, understanding of aerosol processes and therefore potential amplification or dampening of these processes under future scenarios will be key.

The importance of airmass history for understanding aerosol processes has been demonstrated in many studies (e.g. Sogacheva et al., 2005; Tunved et al., 2006, 2013). As aerosol populations undergo significant transformations during transport, Lagrangian frameworks are employed to investigate the potential sources and sinks experienced during transport, and thus understand the resultant measured or modelled population of aerosol. Source-receptor models have been used in aerosol studies to identify potential source regions by linking back trajectories to concentrations measured at the receptor site, for example

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

using a concentration-weighted trajectory framework (CWT). These are powerful tools to provide an overview of potential source regions, particularly when applied to perform climate model evaluation using air-mass trajectories calculated from climate model meteorological and aerosol outputs for climate model evaluation (Kim et al., 2020; Talvinen et al., 2025). However, receptor modelling alone poses challenges for informing process understanding and pinpointing which processes contribute to biases in GCM aerosol representation as they only provide the combined effect of sources and sinks during transport.

Further understanding of the role of specific processes can be gained from deriving source-receptor relationships. One at a time (OAT) based studies have long been used to understand aerosol processes (e.g. Tunved et al., 2006, 2013), focusing on the relationship between a variable, particularly during transport, and the measured aerosol properties at a receptor site. For example, the role of removal via wet scavenging during transport (Khadir et al., 2023; Tunved et al., 2013) and the importance of emissions from the boreal forest as a source of secondary organic aerosols during transport (Liao et al., 2014; Tunved et al., 2006). This has recently been extended to evaluation of model representation of wet scavenging source-receptor relationships using GCM airmass transport (Talvinen et al., 2025) over the boreal forest. Whilst these techniques useful to improve understanding of model bias, these studies do not account for interactions or the potential for compensating errors in GCMs. Additionally, the high dimensionality of aerosol modelling means that applying OAT approaches to constrain all processes would be prohibitively time consuming. Statistical models have previously been employed to investigate relationships of multiple variables at a time for specific aerosol environments. For example, a study by Isokääntä et al. (2022) developed a mixed effects model to examine the drivers of aerosol mass in the boreal forest. The benefit of statistical models is interpretability; however, they commonly rely on an assumption of linearity.

The use of machine learning techniques to model complex, difficult to constrain aerosol properties and processes, has been increasing rapidly, with examples of regressing aerosol particle number concentrations (e.g. Kulkarni et al., 2022), aerosol optical properties (e.g. Geiss et al., 2023; Kumar et al., 2024), CCN concentrations (e.g. Arjunan Nair & Yu, 2020) and aerosol-cloud interactions (Chen et al., 2022, 2024; Watson-Parris et al., 2019). Machine learning regression models using meteorological inputs and source proxies have been used to predict aerosol concentrations, and several studies have considered direction of transport (Chen et al., 2020; Gao & Li, 2021; Karimian et al., 2019; Qin et al., 2019; Qiu et al., 2023; Zhao et al., 2019) but have not considered a full airmass history. Whilst these deep learning models, such as neural networks, have been found to provide accurate predictions of aerosol properties, these architectures are often referred to as ‘black boxes’ due to the difficulty in probing the hidden architecture to unpick drivers of predictions. To improve both real-world understanding of natural aerosol processes and their representation by numerical models, understanding the drivers of predictive models is key. Numerous methods for interpreting machine learning models have been developed, for example permutation feature importance, SHAP (Lundberg & Lee, 2017), Sobols (Jaxa-Rozen & Kwakkel, 2018) and LIME (Ribeiro et al., 2016), however,

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

these rely on the assumption of non-correlated explanatory variables for implementation which cannot be assumed for processes controlling the aerosol lifecycle.

Tree-based models, whilst less complex than neural networks, offer interpretability and still maintain the ability to represent non-linear relationships. A study by Song et al. (2022) leveraged information from airmass history with random forests to predict species of aerosol and investigate the driving processes of concentrations of species of aerosol. This covered an 8-month period in 2015, during the operating period at Gruvebadet Observatory on Svalbard, so could not explore the full seasonal cycle. To predict species of aerosol some of the airmass history was included as explanatory variables (fraction of trajectory over each land type and cluster of direction of transport) as well as meteorological parameters at the site. However, meteorological parameters during transport, such a precipitation, have been found to play a significant role in aerosol prediction (Isokääntä et al., 2022; 2024, Tunved et al., 2013), so key relationships could be missed by only considering meteorology at the measurement site. A recent study utilised combined meteorological weighted trajectory (MWT) maps to as the explanatory variables to predict CWT maps of methanesulfonic acid aerosol (MSA) across the Arctic (Pernov et al. 2024). This allowed for the exploration of the impact that meteorological variables had during transport, but on a monthly scale, using averaged MWT maps. Aerosol processes act much shorter timescales, therefore, in order to perform a process-based analysis, a higher temporal resolution is required.

Whilst significant progress has been made in leveraging Lagrangian-based frameworks to constrain aerosol processes, a number of gaps have been identified that we will address by developing a new framework that considers:

1. High dimensionality
2. Non – linearity
3. Explanatory (allowing for correlations)
4. Generic applicability
5. Process driven explanations

Development of a Lagrangian machine learning (ML) framework that accounts for (1-5) is paramount to untangle the complex, non-linear processes that govern natural aerosol properties and provide robust observationally derived constraints on process level timescales for GCM evaluation. This is important as it will allow us to pinpoint key processes leading to targeted GCM improvements which will be the focus of future work by leveraging a recently developed modelling framework to obtain air-mass history from GCM meteorological data (Kim et al., 2020).

To achieve this, the focus of this study is the development of a holistic, generic ML framework for aerosol lifecycle process understanding that can be subsequently applied to provide unparalleled constraints for aerosols in GCMs to reduce the longstanding uncertainty in ACI. To achieve a process driven ML framework (point 5 above) it is important to note that this

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

framework does not include date information or wind direction as proxies for processes, commonly utilised in PM2.5 modelling studies (e.g. Xiao et al., 2020; Yazdi et al., 2020). Instead, we link directly to data describing underlying processes, which will inherently encapsulate the information represented by these proxies, such as the seasonal cycle, as the aim of the study is to investigate natural aerosol processes and in future work elucidate the potential sources of bias in GCMs. The criterion for the development of a framework is that it is applicable to any in-situ aerosol measurement site (point 4 above), with a sufficient multi-annual timeseries of high temporal resolution aerosol size distribution data, to build a greater understanding of the driving processes in any environment.

The framework developed uses tree-based regression models, selected as the most suitable architecture to achieve points 1-3, to predict aerosol concentrations at in-situ measurement sites from the airmass history. We build a comprehensive airmass history using satellite and reanalysis data during transport to the measurement site. This breadth of data allows us to conduct a holistic study into the drivers of natural aerosols and facilitates application of the framework to different environments. We utilise model interrogation techniques to build on understanding from observational studies and further unpick the complex interactions of the aerosol-climate system. We demonstrate the capability of this framework at Trollhaugen in the Antarctic; however, this framework has been developed to be globally applicable.

Our objective is to improve understanding of the processes governing aerosol concentrations in Antarctic which will be achieved by the following contributions:

1. Identifying the seasonal cycle of aerosol number concentration and the dominant source regions for the measurement site.
2. Building and evaluating a model for accurate predictions of aerosol concentrations from airmass history.
3. Elucidating the relationships between meteorological parameters, source proxies and aerosol concentrations.
4. Identifying dominant source and sink processes controlling the number concentration for Aitken and accumulation modes particle number concentration for the case study environment.

These results could have significant impact for representation of natural aerosol sources in GCM parameterisations, by improving understanding of aerosol processes, and therefore highlighting underrepresented or missing sources and sinks in GCMs.

In Sect. 2 we present the datasets utilised in the study, in Sect. 3 the framework is described, results are discussed in Sect. 4 and finally, conclusions are presented in Sect. 5.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

2 Datasets

We now set out the properties and sources of the datasets used in this study as the explanatory and response variables in the ML framework.

2.1 Particle number size distribution measurements

Particle number size distribution (PNSD) measurements used for this study were obtained from a ground-based aerosol measurement site on the Antarctic continent (Fiebig et al., 2014) through the ACTRIS network (Laj et al., 2024). Trollhaugen (TRH) (72°00′42″S 02°32′06″E, 1553 m a.s.l.) is located at Trollhaugen Mountain, between the Antarctic plateau and the coast, and experiences both air masses from the ocean and continent (Hansen et al., 2009). The location is shown in Fig. 1a. The observatory was established taking permanent measurements at this location in February 2014, so is the longest running timeseries of Differential Mobility Particle Sizer (DMPS) data in the Antarctic region.

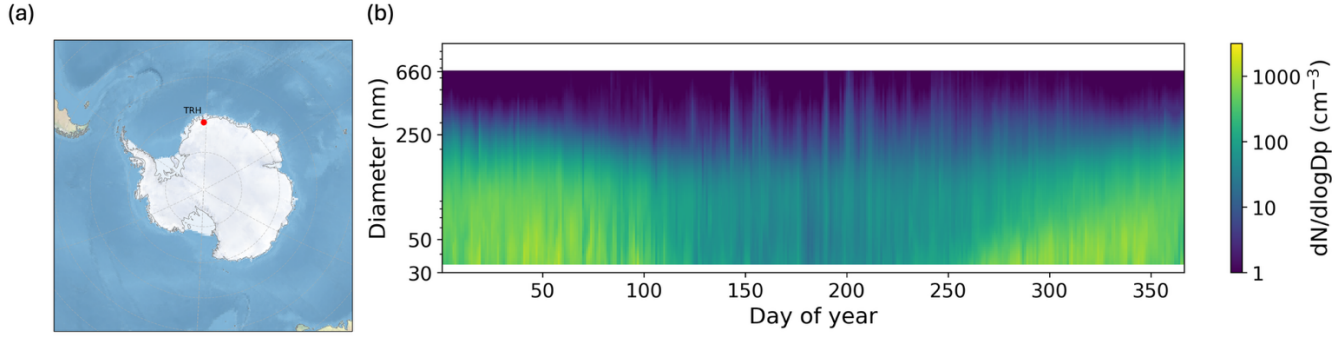


Figure 1: (a) Map of aerosol measurement site location used in this study. (b) The median size distribution per day of year (2014-2018).

Dry PNSD measurements obtained at the site using DMPS instruments at ambient humidity and temperature are then filtered for this study to remove artifacts and contamination. Filters provided with the size distribution datasets were applied to remove those data entries flagged as likely containing instrumental artifacts. Four additional filters were applied; these are described along with the selected flags in Sect. S1.1 and the results are described in Table S1. Once filtered, the size distributions were interpolated to a consistent grid from 1-1000nm with a dlogDp of 0.015 for consistency between instruments at sites and to facilitate consistent comparison to data from multiple GCMs in future studies, the results of which are shown in Fig. S1. The interpolation was conducted using the Piecewise Cubic Hermite Interpolation Polynomial (PCHIP) function (Fritsch & Butland, 1984) from the SciPy package (Virtanen et al. 2020). However, we note that the consistent instrument measurement limits over the period considered spans 34-660 nm, and we do not extrapolate beyond these limits, additional considerations for interpolation are discussed in Sect. S1.2. We note that the interpolation can result in negligible differences of the total

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

concentration calculation where the bin midpoints of the new and original grid do not match (Fig. S1). The resultant PNSD are summarised in Fig. 1b, showing the median size distribution for each day of the year.

Trollhaugen showed surprising pollution events during the austral winter (Fig. S2), some of which have previously been linked to local pollution from other Antarctic stations (Myhre et al., 2019). We utilise back trajectories to isolate events in the concentration timeseries considered at Trollhaugen corresponding to transport from nearby Antarctic stations that use diesel generators year-round: Novolazarevskaya (70°46′37″S 11°49′26″E), Maitri (70°46′00″S 11°43′55″E), Showa (69°00′15″S 39°34′55″E), SANAE IV (69°00′15″S 39°34′55″E) and Neumayer (70°40′28″S 8°16′27″W). These local anthropogenic sources will not be represented by the explanatory variables so need to be removed before fitting the regression model. We consider potential events to be the highest 2% of N80 concentrations for April-September transport, and consider contamination, and therefore a flagged pollution event, within 0.4 degrees of each site. Additionally, some extreme low concentrations ($N80 < 4 cm^{-3}$) consistent for several timesteps appeared anomalous in the timeseries and upon inspection of the raw data these were removed from analysis. The results for this contamination filtering step are shown in Sect. S1.3.

The response variable selected for the regression models was the aerosol concentration for two size ranges: ~30-80 nm to represent the Aitken mode (34 nm the actual lower limit due to PNSD coverage), and ~80-660nm to represent the accumulation mode. These size ranges were selected as the concentration of aerosol above 80nm (N80) is commonly used to represent a proxy for CCN (e.g. Asmi et al., 2011; Kerminen et al., 2012) and understanding the drivers of CCN in different regions is essential when considering the climatic impact. We also consider the Aitken mode, as in pristine regions these can contribute to CCN burden (e.g. Karlsson et al., 2022). It is also important to consider the two size ranges in tandem when exploring the dominating processes controlling the observed aerosol particle number size distribution to contribute to a greater understand of aerosols in the region. The aerosol concentrations were averaged to a 6-hourly resolution, using the mean, to reduce noise in the dataset.

2.2 Air-mass trajectories

Single-particle backwards trajectories from each aerosol measurement (receptor) site were calculated using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model developed by the National Oceanic and Atmospheric Administration (NOAA) Air Resources Laboratory (Stein et al., 2015). The HYSPLIT version used in this study (5.1.0) includes a minor bugfix to improve the velocity interpolation calculation of trajectories passing near the poles. Trajectory release heights are selected based on the receptor site orography and the representation of the orography in the driving meteorological data. To reduce the likelihood of trajectories having contact with the surface, a release height of 100m above ground level (AGL) is used, the decimal coordinates and starting altitude for trajectory calculations are reported in Table 1.

For the driving meteorological data, 6-hourly ERA-Interim reanalysis data (Dee et al., 2011) was re-gridded onto a 1.0° x 1.0° latitude-longitude grid. For this study 4D variables (latitude, longitude, pressure, time) from ERA-Interim are output on 30

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

fixed pressure levels. ERA-Interim was selected for this study following the setup for the Aerosol Comparisons between Observations and Models (AeroCom) phase III experiment. This facilitates comparison to GCM simulations nudged to ERA-interim horizontal winds, to ensure any biases identified are not due to differences in transport.

Our analysis, like all frameworks reliant on single-particle trajectories (e.g. Dal Maso et al., 2007; Tunved et al., 2013), is susceptible to the inherent uncertainty linked with individual trajectories (Engström & Magnusson, 2009; Stohl, 1998). Some uncertainty can be mitigated with the averaging of summary statistics from hourly trajectories to create a 6-hourly average airmass history (see Sect. 2), however the uncertainty in the input meteorology is often regarded as the dominant contribution (Bowman et al., 2013).

Station	**Description**	**Years included in study (inclusive)**	**Diameter range included (nm)**	**Trajectory release location**			**Trajectory length (hrs)**
				Latitude	**Longitude**	**Height m A.G.L.**	
Trollhaugen	Antarctic (pristine)	2014-2018	34-660	-72.012	2.535	100	240

Table 1: The receptor site with the site description, years included in the study, diameter range and the trajectory release locations.

2.3 Explanatory variables

The airmass history provided by HYSPLIT trajectories is further supplemented by considering sources and sinks of aerosols during transport (Table 2). We collocate 2D satellite and reanalysis variables identified as proxies for aerosol processes along trajectories in 3 dimensions (latitude, longitude, and time) using linear interpolation over rectilinear grids provided by the SciPy library (Virtanen et al., 2020), in a 3D adaption of the 4D approach described in Talvinen et al. (2025). For categorical variables such as land classes, we employ a nearest neighbour collocation approach. We consider a wide range of potential aerosol sources in this study to ensure the framework is generic and applicable to other regions.

2.3.1 Meteorological variables

HYSPLIT provides many important meteorological variables for aerosol properties during the calculation of trajectories. In this framework we include trajectory height, boundary layer height, temperature (at trajectory height) and relative humidity (RH). Many of the driving processes for natural aerosol formation are also photosynthetically driven, therefore the surface net solar radiation from ERA-Interim is included from the trajectory calculation. Time in boundary layer is calculated from the

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

trajectory height compared to the boundary layer height for each trajectory point. Trajectory speed is calculated from the latitude and longitude of the trajectory, using the Haversine formula.

Many near surface processes generating pre-cursor gases and primary aerosol have meteorological drivers. Therefore, we additionally collocated surface meteorological variables: 2m temperature and 10m wind speed from ERA5 reanalysis (Hersbach et al., 2020). The ERA5 data is provided at a 6-hourly resolution on a 0.25-degree grid.

2.3.2 Aerosol sink proxies

It is also essential to consider removal processes during transport; to account for wet removal ERA-Interim reanalysis surface precipitation large scale and convective data (Dee et al., 2011; https://codes.ecmwf.int/grib/param-db/228) is collocated along trajectories. The HYSPLIT calculated precipitation fields are truncated and thus, are poorly representative of light rain and drizzle (Talvinen et al., 2025). Satellite products were considered, however due to data sparsity there is not global coverage particularly at the highest latitudes at the time resolution required, therefore for this study the reanalysis dataset is used and collocated after trajectory calculation. Time dependent weightings are applied during the calculation of summary statistics to account for the time dependent influence of precipitation during transport, discussed in Sect. 2. As in Isokääntä et al. (2022) and Tunved et al. (2013) we calculate 'time in cloud' based on an RH threshold (where RH from the ERA-Interim reanalysis exceeds 94 %) to account for cloud processing in our model.

2.3.3 Aerosol source proxies

Numerous studies have examined source-receptor relationships in different environments, utilising proxies to represent aerosol sources. Time over land has been shown in several studies in the boreal forest to be strongly correlated with aerosol mass (Liao et al., 2014; Petäjä et al., 2022; Räty et al., 2023; Tunved et al., 2006), and has been utilised in previous studies to predict aerosol properties (Isokääntä et al., 2022; Song et al., 2022). To obtain an estimation of time of trajectory over land the General Bathymetric Chart of the Oceans (GEBCO) 2019 product was used (GEBCO Compilation Group, 2019), this provides global coverage of terrain elevation over land and ocean. We use this to create a land mask (land > 0 m a.s.l.) to collocate onto the trajectories. However, from initial analysis this was found to act as a proxy for anthropogenic emissions in some regions therefore we elected to replace this with separate land classes: time over evergreen forest, deciduous forest, shrub, cropland and urban. To implement this separation, land cover maps utilising the United Nations Food and Agriculture Organization's (UN FAO) Land Cover Classification System (LCCS) were collocated along trajectories with the appropriate classes selected for each variable (Table S2). The dataset is based on classifying a baseline land cover map using Medium Resolution Imaging Spectrometer (MERIS) and updated using change detected from System Pour l'Observation de la Terre-Vegetation (SPOT-VGT) (1998 to 2012) and Project for On-Board Autonomy-Vegetation (PROBA-V) and Sentinel-3 OLCI (S3 OLCI) (from 2013). Time over land was then replaced with time over sea – the summation of the trajectory timesteps not classified as land.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

In addition to considering the time over each annual vegetation species classification maps, we seek to further isolate the impact of different aerosol sources such as biogenic volatile organic compound (BVOC) emissions from trees and shrubs and dimethylsulfide (DMS) from phytoplankton so include additional source proxies with a higher temporal resolution as explanatory variables. Leaf area index (LAI) is used here as the proxy for BVOCs which are important sources of secondary organic aerosol, particularly for Boreal regions (e.g. Spracklen et al., 2008). In this study we use effective LAI, derived from normalized top-of-canopy reflectance from the Vegetation (VGT) sensor onboard two satellites covering the data period: PROBA (2013-2020) and SPOT (1998-2014). The LAI product is derived every 10 days with a 20-day composite window, with a spatial resolution of 1km. Chlorophyll-a has been shown to be a good proxy for secondary organic aerosol (O'Dowd et al., 2015) and primary marine organic emission at Mace Head (Rinaldi et al., 2013) and has been used in the parameterisations for primary marine emissions in GCMs (Rinaldi et al., 2013). The GlobColour product was used for to collocate chlorophyll concentrations along trajectories (Fanton d'Andon et al., 2009; Maritorena et al., 2010), this is a combined L3 product from MODIS, MERIS and SeaWiFS (O'Reilly et al., 2000).

Sea ice plays an important role in aerosol processes in Arctic and Antarctic regions, with decreased sea ice extent leading to increases in sea salt aerosol flux and increased biogenic flux (e.g. Struthers et al., 2011; Dall'Osto et al., 2017; Yan et al., 2020). To a lesser extent sea ice has also been found to be a source of aerosol from the sublimation of blowing snow (Frey et al., 2019). The Operational SST and Sea Ice Analysis (OSTIA) sea ice product with a spatial resolution of 0.05 degrees is produced at a daily resolution. This has been averaged to a monthly resolution, to decrease data sparsity from cloudy retrievals (Donlon et al., 2012). A threshold fraction of 0.25 is used to classify a trajectory timestep as spending time over sea ice.

Despite the selection of a pristine site to study natural aerosol processes, we acknowledge that there could be transport of anthropogenic emissions and periodic wildfires to this region, and these variables will be important for other regions. To account for these, anthropogenic emissions are included in the explanatory variables using the Community Emissions Data System (CEDS) gridded emissions (v_2021_04_21) at 0.1° resolution. This dataset is a downscaled version of the 0.5° dataset (Hoesly et al., 2018), using 0.1° proxy data from EDGAR (Janssens-Maenhout et al., 2019). The estimates of global air emissions species at a monthly resolution were summed over all sectors to provide a total anthropogenic emission of BC, NH_3, NOx and SO_2. To investigate the impact of fires in the study, BC and OC emission estimates from the Global Fire Emissions Database (GFED) were included as explanatory variables at two heights to account for lofting. The BC and OC emissions obtained from GFED were at a daily temporal resolution and 0.25° spatial resolution. Whilst not expected to dominate in Antarctica, fires emissions will be of greater importance in other regions, for example the Arctic (e.g. Gramlich et al., 2024; Warneke et al., 2010).

All these variables have been shown in previous studies, as discussed, to have an impact on aerosol properties and data sources have been selected to provide the fullest coverage across the globe and consistent timeseries for this framework. The selected variables and sources are summarised in Table 2, resulting in 35 explanatory variables. The mean was used for meteorological

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

variables, with a weighting (Sect. 2). For surface solar radiation the weighted sum was used, to introduce a measure for the length of the length of daylight hours. For source and sink variables the weighted sum was used to consider the accumulation of source variables during transport using the weight (Eq. (1)), discussed in Sect. 2. The 'time over' mask variables are summarised using the sum to ensure interpretability and facilitate comparison to previous studies.

No.	Variable (unit)	Source	Spatial resolution	Temporal resolution	Reference	Averaging method over trajectory
1	Height (m a.g.l)	Hysplit + Era-Interim	1.0°	6-hourly	(Dee et al., 2011; Stein et al., 2015)	Weighted mean
2	Boundary layer height (m)	Reanalysis (Era-Interim)	1.0°	6-hourly	(Dee et al., 2011)	Weighted mean
3	Temperature (K)	Reanalysis (Era-Interim)	1.0°	6-hourly	(Dee et al., 2011)	Weighted mean
4	Trajectory speed	Reanalysis (Era-Interim)	1.0°	6-hourly	(Dee et al., 2011)	Mean
5, 6, 7	Total surface precipitation (convective and large scale) ($mmhr^{-1}$)	Reanalysis (Era-Interim)	1.0°	6-hourly	(Dee et al., 2011)	Weighted and non-weighted sum and reverse weighted sum.
8	Snowfall ($mmhr^{-1}$)	Reanalysis (Era-Interim)	1.0°	6-hourly	(Dee et al., 2011)	Weighted sum
9	Relative humidity (%)	Reanalysis (Era-Interim)	1.0°	6-hourly	(Dee et al., 2011)	Weighted mean
10	Surface net solar radiation (Wm^{-2})	Reanalysis (Era-Interim)	1.0°	6-hourly	(Dee et al., 2011)	Weighted sum
11	Time in boundary layer (hr)	Reanalysis (Era-Interim)	1.0°	6-hourly	(Dee et al., 2011)	Sum
12	Time in cloud (hr)	Reanalysis (Era-Interim)	1.0°	6-hourly	(Dee et al., 2011)	Sum

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

13	Chlorophyll (ugL^{-1})	Satellite (GLOB-Colour)	4km	Monthly	(Fanton d'Andon et al., 2009; Maritorena et al., 2010)	Weighted sum
14	LAI (m^2m^{-2})	Satellite (SPOT VGT)	1km	Monthly	Copernicus Climate Change Service, Climate Data Store, (2018): Leaf area index and fraction absorbed of photosynthetically active radiation 10-daily gridded data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.7e59b01a (Accessed on 26/03/2021)	Weighted sum
15, 16, 17, 18	10m wind speed (ms^{-1})	Reanalysis (ERA5)	0.25°	6-hourly	(Hersbach et al., 2020)	Mean, std, max and non-weighted mean
19	2m temperature (°C)	Reanalysis (ERA5)	0.25°	6-hourly	(Hersbach et al., 2020)	Weighted mean
20	Time over sea ice (hr)	Satellite (Ostia)	0.05°	Monthly	(Donlon et al., 2012)	Sum
21	Sea ice weighted mean	Satellite (Ostia)	0.05°	Monthly	(Donlon et al., 2012)	Weighted mean
22	Time over evergreen forest (hr)	ESACCI	300m	Yearly	ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). Available at: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf	Sum
23	Time over deciduous forest (hr)	ESACCI	300m	Yearly	ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). Available at: maps.elie.ucl.ac.be/CCI/viewer/d	Sum

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

					ownload/ESACCI-LC-Ph2-PUGv2_2.0.pdf	
24	Time over shrub (hr)	ESACCI	300m	Yearly	ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). Available at: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf	Sum
25	Time over urban (hr)	ESACCI	300m	Yearly	ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). Available at: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf	Sum
26	Time over sea (hr)	GEBCO	0.0042°	N/A	(GEBCO Compilation Group, 2019)	Sum
27, 28, 29, 30	Anthropogenic emissions (BC, SO_2, NO_x, NH_3)	CEDS	0.1°	Monthly	(Hoesly et al., 2018)	Weighted sum
31, 32, 33. 34	Emissions from fires (BC high, BC low, OC high, OC low)	GFEDS	0.25°	Daily	(Giglio et al., 2013; Mu et al., 2011; Van Der Werf et al., 2017)	Weighted sum
35	Arrival hour	N/A	N/A	N/A	N/A	N/A

Table 2: Explanatory variables source, resolution, reference and averaging technique.

2.4 UKESM1 configuration

We compare the performance of our Lagrangian ML modelling framework to aerosol concentration predictions in Eulerian space at the receptor site by the UK Earth System Model (UKESM1.0), hereafter referred to as UKESM. The configuration used is the same atmosphere-only style as for the Atmospheric Model Intercomparison Project (AMIP). The external forcing datasets are consistent with the Coupled Model Intercomparison Project Phase 6 (CMIP6) implementation of UKESM and time-evolving sea surface temperature, sea ice and prescribed marine biogenic emissions are used from a fully coupled model simulation as described in Sellar et al. (2020). The horizontal winds in UKESM are nudged to ERA-Interim (Telford et al., 2008), and the resolution used is 1.875 degrees longitude, 1.25 latitude and 85 vertical levels. The 2-moment modal aerosol

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

scheme in United Kingdom Chemistry and Aerosols (UKCA) model is described in (Mann et al., 2010). We note that UKESM1.0 uses a binary neutral homogeneous H_2SO_4–H_2O nucleation scheme (Vehkamäki et al., 2002) throughout the atmosphere.

We linearly interpolate the log-normal modal aerosol fields (diameter, concentration, and geometric standard deviation) at the receptor site latitude, longitude and trajectory start height, at a 3 hourly resolution. We use the UKESM1 UKCA modal parameters (dry diameters, number concentrations and geometric mean diameters, see Table S3) to calculate the log-normal PNSD (Seinfeld and Pandis, 1998) on the same grid as for the observations described in Sect. 2.1 The size distributions are then averaged to 6-hourly to match the resolution of the observations.

3 Methodology

Explainable machine learning techniques are implemented to predict and then interrogate aerosol properties at the receptor site using the airmass history as described in Sect. 2.3 In this section we describe the overarching framework, data preparation, receptor models used for analysis and finally the regression model setup and interrogation.

3.1 Framework

The airmass history is used to predict aerosol properties at the measurement site, in the framework described by the schematic in Fig. 2. Combining Lagrangian frameworks with regression models, is a relatively new approach. Previous studies have relied on meteorological inputs at the site (Qin et al., 2019; Song et al., 2022), but not the impacts of processes acting during the air-mass on the formation and growth of aerosols during transport. Therefore, we consider a much more comprehensive description of the airmass history, as described in Sect. 2.3. We build XGBoost regression models to predict aerosol concentrations, then interrogate these models with SHAP methods to investigate the dominant processes leading to the measured aerosol properties at the receptor site.

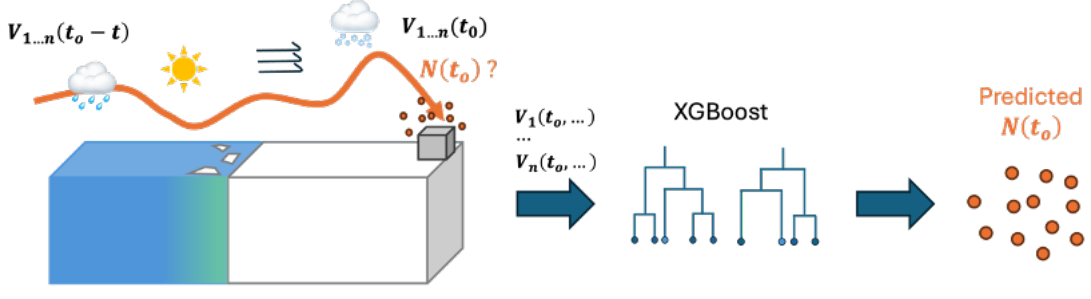


Figure 2: Schematic to describe the framework implemented in this study.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

3.2 Data preparation

To prepare the explanatory variables for the regression modelling, we seek to reduce the dimensionality to avoid model overfitting and increase interpretability by taking summary statistics along trajectories for each variable. These are shown in Table 2.

As this study focuses on building a ML model corresponding to aerosol properties observed near the surface, where the aerosol lifecycle in the atmosphere tends to be on the scale of days (Boucher et al., 2013), a weighting (Eq. (1)) is applied to the summary statistics to account for the decay of feature importance along the trajectory, the graph for this equation over a 240-hour trajectory is shown in Fig. S3. We use an exponential decay based on the trajectory hour (i) but introduce a 'daily' factor to reduce the gradient of the exponent, as typically atmospheric process act on timescales of days (Raes et al., 2000).

$$w_i = e^{-\frac{1}{24}},\ i \in \{1, \dots, N\} \tag{1}$$

We also ensure that the weights are normalised in order to preserve the scale of the explanatory variables and maximise interpretability. For missing data, in order to preserve the scale of the explanatory variables, we remove these weights from the trajectory before any normalisation is applied. For variables such as LAI, ocean can be represented as missing data in the dataset, however in this context there is an absence of the source (e.g. LAI=0), therefore we use the land mask to apply a padding to ocean points, to fill with zeros, before the weighting step. This approach is consistently applied for chlorophyll concentrations and sea ice fraction.

For the weighted summary statistics, we consider a weighted sum (Eq. (2)) for aerosol source and sink proxies, and for meteorological variables we consider the weighted mean (Eq. (3)), for each variable, x, over a trajectory of length N.

$$S = \sum_{i=1}^{N} x_i w_i s \tag{2}$$

$$E = \frac{\sum_{i=1}^{N} x_i w_i s}{N} \tag{3}$$

$$where\ s = \frac{N}{\sum_{i=1}^{N} w_i} \tag{4}$$

A linear weight was also tested (not shown), however the exponential performed better, highlighting importance of accounting for the decay in relevance of variables during transport. Data is all averaged to 6-hourly, using the arithmetic mean, to match the time resolution of the aerosol concentration data. A 6-hourly resolution was selected to remove noise from the hourly aerosol data, whilst maintaining a diurnal cycle, and during initial tests (not shown), this averaging period was found to result

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

in the best model performance. Any data points with missing data in the calculated explanatory or response variables were removed prior to fitting the ML framework. For regression tasks it is common to apply a Box-Cox transform (Eq. (5)) to transform the distribution of the response variable to a normal distribution to improve model performance (Osborne, 2010). Here, lambda is fitted on the whole response variable dataset and the same lambda value used on all data splits.

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{(\lambda)} - 1}{\lambda} \; if \; \lambda \neq 0 \\ \ln(y_i) \; if \; \lambda = 0 \end{cases} \tag{5}$$

Finally, the data must be split into training and test datasets. The implementation of a random test-train split, as implemented in Song et al. (2022), disregards the inherent correlation of timeseries data, therefore the reported skill is likely higher than the model skill on unseen data. Here we select a full year (2017) to use for testing and remove this prior to training to avoid contamination of the test set. The number of data points in each dataset is shown in Table 3.

Station	Total dataset	Training dataset	Test dataset
Trollhaugen	6497	5149	1348

Table 3: The number of datapoints in each dataset for explanatory and response variables, at the 6-hourly resolution.

3.3 Receptor models

Five types of receptor model are used in this study to inform our understanding a priori of aerosol processes associated with the aerosol receptor station and evaluate model performance. Transport frequency receptor models are used to investigate the spatial distribution of the airmass history for each site. We define this over a grid, with the frequency of transport within a given grid cell defined as:

$$F_{ij} = \frac{100}{T} \sum_{l=1}^{M} v_{ijl} \tag{6}$$

where F_{ij} is the frequency of visits for a grid cell, T is the total number of trajectories, $v_{ijl}$the number of unique visits by a trajectory, l, in the grid cell, i,j, and M is the total number of trajectories with endpoints in the i,jth grid cell.

Concentration weighted trajectory (CWT) receptor models are used to investigate the potential dominant sources of aerosols for each receptor site used in this study. These are defined over the same grid as for the transport frequency models using the definition as defined in Hsu et al. (2003):

$$C_{i,j} = \frac{1}{\sum_{l=1}^{M} \tau_{ijl}} \sum_{l=1}^{M} C_l \tau_{ijl} \tag{7}$$

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

where C_l is the concentration associated with trajectory l, τ_{ijl} is the total number of trajectories endpoints in the grid cell (i,j) associated with the C_l sample. We additionally build receptor error model to investigate regions of error in the models:

$$e_{i,j} = \frac{1}{\sum_{l=1}^{M} \tau_{ijl}} \sum_{l=1}^{M} e_l \tau_{ijl} \tag{8}$$

where e_l is the error in predicted concentration ($e_l = C_{(l,modelled)} - C_{(l,observed)}$).

To highlight dominant regions of error we weight the error functions by the frequency of visits per grid cell to produce the weighted absolute error (W) receptor model:

$$W_{i,j} = \frac{100}{T} \sum_{l=1}^{M} v_{ijl} \frac{1}{\sum_{l=1}^{M} \tau_{ijl}} \sum_{l=1}^{M} e_l \tau_{ijl} \tag{9}$$

For variables such as height a collocated variable trajectory (CVT) receptor model is used. CVT receptor models, similarly to the meteorological weighted trajectory models in (Pernov et al., 2024), represent an average of a collocated variable at a point in space during the transport of the airmass to the measurement site. Generally, this will closely reflect the monthly mean distribution for the collocated values and take into account the spatial variability in meteorological conditions and air mass transport pathways (Pernov et al., 2024). The CVT receptor model is described by Eq. (10):

$$V_{i,j} = \frac{1}{\sum_{l=1}^{M} \tau_{ijl}} \sum_{l=1}^{M} V_{ijl} \tau_{ijl} \tag{10}$$

where $V_{(i,j,l)}$ is the variable value associated with trajectory point, τ_{ijl}, in the grid cell i,j is the total number of trajectories endpoints in the grid cell (i,j).

A Lambert azimuthal equal-area projection is used for the receptor models in this study to ensure that the defined regularly spaced grid results in grid cells with equal area. The receptor models are build using the hexbin function from cartopy (Elson et al., 2022), but it must be noted that there is some approximation of equal area based on the 'best fit' of hexagons over the domain by introduced by the algorithm. However, the bias introduced by this approximation is negligible, considering geographic coordinate grids. Note that for these maps we do not apply a mask to remove low trajectory counts, in order to facilitate direct comparison to the SHAP results.

3.4 Regression model and SHAP model interrogation technique

The regression model used in this study was XGBoost (eXtreme Gradient Boosting) (Chen & Guestrin, 2016), an additive tree-based model, built from decision trees. XGBoost was selected as it frequently outperforms Random Forest models on

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

imbalanced datasets. For the model setup, the mean squared error was used as the loss function and hyperparameter tuning was performed using Tree-based Parzen Estimators (TPE) implemented by the hyperopt package (Bergstra et al., 2013), results from which are described in Table S7. Hyperparameter tuning was conducted for 300 iterations, using the negative R^2 score as the minimisation objective. The models developed in this study are hereafter referred to as the 'ML models'.

SHAP is used in this study for model interrogation, estimating Shapley values to explain the contribution of each feature to a prediction of an instance (Lundberg et al., 2017). KernelSHAP calculates the contribution of features by sampling from the marginal distribution of the dataset to represent absent features. However, it is important to note that the use of the marginal expectation in KernalSHAP to estimate the contribution of a feature ignores the dependence structure of the features so if there are high correlations between variables this can break those dependencies. The more recently developed path-dependent TreeSHAP calculates the expectation $f(S)$ using the conditional expectation based on the structure of the trees, therefore not breaking the correlations between features (Lundberg et al., 2018). As the features of the regression model used in this study demonstrate high correlations, it was paramount to implement a model interrogation technique that did not break the dependencies between variables. However, a caveat of TreeSHAP is that correlated features can result in a SHAP value different from zero, therefore correlation between features must be considered in the interpretation of TreeSHAP results. Additionally, a correlated feature used higher in the tree will be given more importance than a correlated feature used lower in the trees.

SHAP is employed in this study using path-dependent TreeSHAP from the python SHAP package (Lundberg et al., 2020). SHAP rankings were used to perform recursive feature selection during the model build: removing all features with zero importance then recursively removing the lowest ranked features until model performance degrades. The TreeSHAP rankings were used here as the standard feature importance is based on a shuffling algorithm so assumes independence of features, the results of feature selection are described in Table S6.

SHAP is used for model interrogation, to investigate the relationships leading to predictions of the ML model in Sect. 4.4. As well as considering the SHAP-feature relationships from the SHAP analysis we consider the spatial distribution, to visualise this, we employ the CWT framework (Sect. 3.2) to the SHAP results, where each SHAP value for a measurement at the site corresponds to a trajectory. Note that the SHAP analysis is conducted on the regression models which predict Box-Cox transformed aerosol concentrations, thus the SHAP values are on the magnitude of the transformed data. The Box-Cox transformation is monotonic, therefore the SHAP analysis holds for the aerosol concentrations, so the SHAP results are discussed in relation to the aerosol number concentrations throughout. Additionally, it is important to note that as the Box-Cox transformations are performed on each concentration dataset separately, the magnitudes of the SHAP values cannot be compared between the two models. Thus, in the analysis of the results the rankings of the variables between the models, and the magnitudes only are considered within each ML model analysis.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

4 Results and discussion

In this section we present the results of this study. Starting by demonstrating the seasonal cycle of aerosol at Trollhaugen. Next, we present the results of the regression models for each site: evaluating first the ML model performance in a Eulerian perspective, comparing to UKESM, and finally in a Lagrangian perspective. Then we present the results of the SHAP analysis utilising the ML models for both aerosol particle number concentration size ranges.

4.1 Seasonal cycle at each site

Figure 3 shows the average seasonal cycle for the period considered in the study for each measurement site. Trollhaugen is a pristine site; highlighted by the extremely low concentrations and variability in the austral winter. There is a very distinct seasonal cycle for both the accumulation and Aitken concentrations (Fig. 3). The shape of the distributions is very similar, however in the austral summer concentrations in the Aitken mode are almost double that of the accumulation mode (Fig. 3a). This is consistent with results for the seasonal cycle across Antarctic aerosol measurement sites (Fiebig et al., 2014; Lachlan-Cope et al., 2020; Rose et al., 2021). A strong seasonal cycle in total particle number concentration has been found to be a prominent feature across Antarctic in previous studies with Austral Summer being found to be up to 20–100 times greater than during the winter (e.g. Fiebig et al., 2014; Ito, 1993; Shaw, 1979; Weller et al., 2011), and has been found to be more pronounced at measurements sites on the upper plateau rather than the coastal sites which are more influenced by sea salt concentrations (Lachlan-Cope et al., 2020). Trollhaugen is located between the Antarctic plateau and the coast but still demonstrates a strong seasonal cycle in number concentrations for both modes.

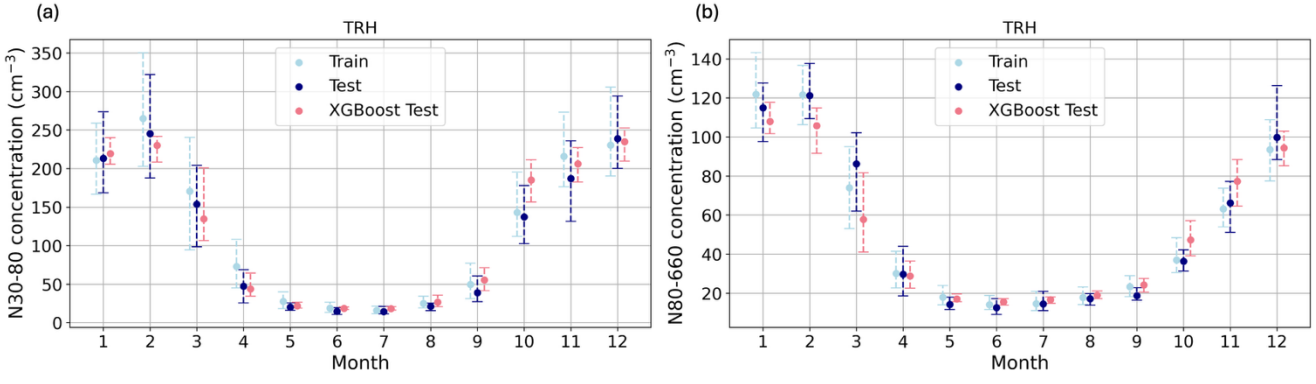


Figure 3: The seasonal cycle of concentrations (cm^{-3}) at TRH (Trollhaugen) for (a) 30-80nm and b) 80-660nm for the observations for the years used for model training (2014-2018 inc., excluding 2017) (light blue, left shift), the observations for the year used to test the model (navy, centred) and the ML model (XGBoost) predictions for the test year (2017) (pink, right shift). The monthly median is indicated by the central point and the bars show the 25^{th}-75^{th} percentiles.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

4.2 Model evaluation

The ML regression models developed in this study for each site and aerosol size range accurately replicate the seasonal cycle of aerosol concentrations (Fig. 3). Comparing to UKESM, ML models better replicate the average aerosol concentrations (Fig. 4). Notably, we see significant underprediction by UKESM of accumulation mode aerosol particle number concentrations during the Austral summer at Trollhaugen (Fig. 4b), by a factor of 2. It has long been shown that UKESM underpredicts aerosol concentrations in the Antarctic (Mulcahy et al., 2020) and as it is the accumulation mode particles that have the most significant impact on cloud properties, this underprediction could have significant ramification for climate predictions in this highly climatically significant region. The ML model also better replicates the variance of the observations for the accumulation mode, particularly for March (Fig. 4b). For N30-80nm the ML model consistently outperforms UKESM at this monthly scale, apart from October and November when the range is better captured by UKESM (Fig. 4a). Furthermore, when we consider the overall skill of the ML models on a 6-hourly resolution, as measured by the coefficient of determination (Table 4), we see that the ML models significantly outperform UKESM. These results give us confidence in the considerable improvement our ML predictions offer, and therefore, we can use these ML models to inform decisions to improve the parameterisation of aerosol processes in pristine regions.

We utilise UKESM as a benchmark in this study, however it is important to note that UKESM is tuned to be globally representative, compared to the site-specific ML models. Nevertheless, we have demonstrated current biases in UKESM as we aim for insights for potential limitations of UKESM aerosol process representation in Antarctica. Future work will extend the framework to additional sites, enabling insights across the globe.

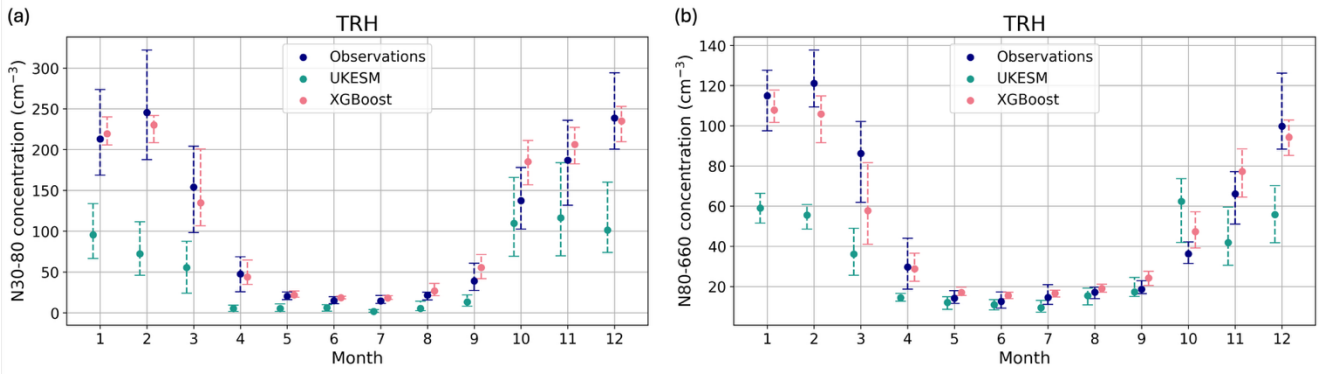


Figure 4: The seasonal cycle of concentrations (cm^{-3}) TRH (Trollhaugen) for the test year for (a) 30-80nm and (b) 80-660nm. The monthly median is indicated by the central point and the error bars show the 25th-75th percentiles for observations (blue, centred), ML model (XGBoost) results from this study (pink, right shift) and UKESM (turquoise, left shift).

The ML models' limitations lie in the ability to predict the highest concentrations, likely due to the sparsity of these in the training dataset and the model setup (Figs. S4 and S5). Inherently the choice of mean squared error as the learning objective

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

for the regression task does not optimise the model to predict extremes (when there are relatively few such examples), however in this study the objective is to identify the prevailing processes represented at each site, hence the selection of the mean squared error and the use of seasonal average plots for evaluation here. We note that this is particularly true for the Aitken mode: in both the test and training datasets, the ML model was not able to replicate the highest concentrations (Fig. S4), which could be associated with short-term high concentrations from new particle formation (NPF) events. Some of these extremes, particularly those in the winter, could be due to remaining contamination of the underlying PNSD data by instrument errors and station generator contamination. However, overall, the ML models demonstrate a relatively high predictive performance (Table 4), especially on a seasonal scale (Fig. 3), which allows us to have confidence in relationships represented by the model and implement model interrogation techniques to investigate these.

Model simulated quantity	***Trollhaugen N30-80nm***	***Trollhaugen N80-660nm***
ML model (Cross-validation average)	0.65	0.68
ML model (test year)	0.72	0.75
UKESM (test year)	0.15	0.19

Table 4: Coefficient of determination, R^2, of the models for each size range at each site for the average across the cross-validation years for the ML models, the score for the test year for the ML models and the score for the test year data for UKESM compared to observations. The cross-validation average is calculated as the mean of the R^2 scores during k-fold cross-validation (without the test year).

4.3 Receptor models

To understand potential sources regions of aerosol particle number concentration at Trollhaugen and evaluate ML model performance in a Lagrangian framework we now utilise the receptor models described in Sect. 3.3.

We consider the transport history built from the backwards trajectories at each site. At Trollhaugen, in Fig. 5a, we see the clear pattern of the Antarctic easterly winds with transport mainly coming from the coast to the East, similarly, to findings for previous studies at the nearby Halley station (Lachlan-Cope et al., 2020; Paglione et al., 2024). We also consider the height of trajectories above ground level during transport to the measurement site in Fig. 5b. For the dominant transport pathway around the continent, this follows the relatively low orography of the coastline, near ground level. Over the Weddell Sea trajectories are high compared to transport from the Bellingshausen Sea where trajectories are much closer to ground (or sea) level, which would likely lead to more of a dominance of marine sources from this region, similarly to the results of Fiebig et al. (2014).

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

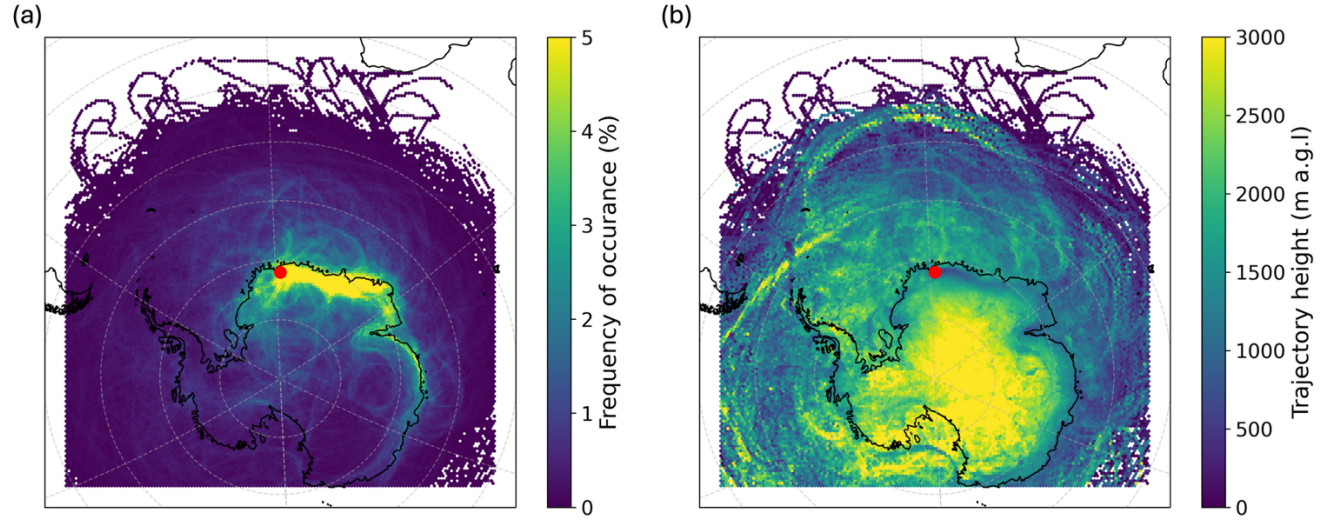


Figure 5: (a) Transport frequency receptor model and (b) CVT receptor model (Eq. 10) for height of trajectories (above ground level) calculated using ERA-Interim meteorology for the full test year at Trollhaugen (2017).

The CWT receptor models (Eq. (7)) for the observed and predicted aerosol number concentrations are shown for the two size ranges considered in this study: 30-80nm (Fig. 6a-d) and 80-660nm (Fig. 6e-h) across the test year for each receptor site. These results highlight the potential source regions of aerosol for each size range. We use this as an evaluation tool, building a replica map using the ML predictions. Linking the bias of the ML model at each time point to the respective averaged trajectories, in the methodology of a CWT model (Eq. (8)), we can utilise an error map to highlight source regions where the ML model performance could be improved. The error plots have been used throughout the study in model development to assess the skill of the models and also highlight potential methods for improvement and potential missing processes. Shown in Fig. 6 are the CWT plots for the observations, ML model predictions and bias at each site.

For Trollhaugen the high concentrations for both size ranges are associated with the dominant transport pathway, anticlockwise around the continent, as well as the Southern Ocean (Figs 6a and 6e). The potential dominant sources over the Southern Ocean would likely be associated with sea spray and secondary marine organics (Lachlan-Cope et al., 2020; Paglione et al., 2024). The ML model replicates the potential source regions very accurately for both size ranges (Figs 6b and 6f), this is also highlighted by the error figures (Figs 6c and 6g) where the regions of high error are edge cases, associated with fewer trajectories (Fig. 5a). The weighted error demonstrates overprediction for the Aitken mode predictions around the coast in the region of highest trajectory frequency (Fig. 6d). Whereas, for the accumulation mode model, when weighting by transport, there is a clear path of transport around the coast associated with underprediction in the model: this path corresponds to the highest observed concentrations, which the model is not able to replicate. The pathway around the coast associated with underprediction of the accumulation mode in the model is a region of low-level transport (Fig. 5b). We suggest that the underprediction in the ML model along the coast could be due to a lack of representation of seabird colonies. Seabirds have been shown to be significant sources of nitrates and ammonia at in the Antarctic (e.g. Boyer et al., 2025; Brean et al., 2025;

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Dall'Osto et al., 2022) and colonies are located all around the continent coast (e.g. Riddick et al., 2012), however we do not currently have a representation of these in the model.

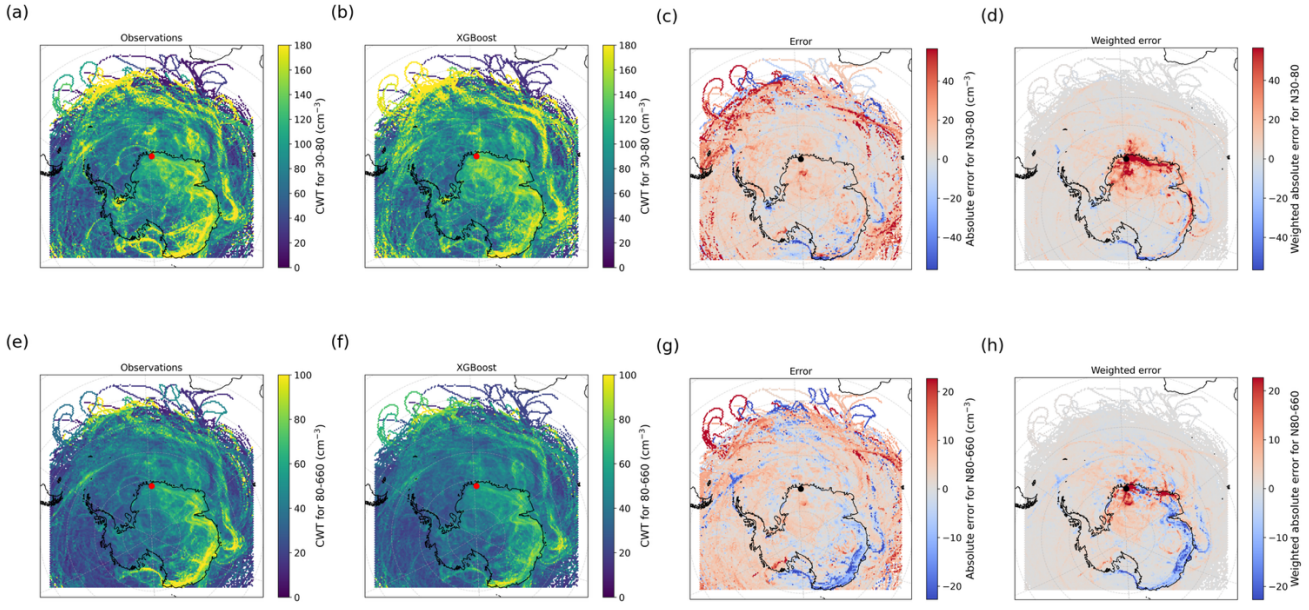


Figure 6: Concentration weight trajectory maps for Trollhaugen test year using ERA-Interim trajectories with: a) Observation concentrations for 30-80nm, b) ML model (XGBoost) predicted concentrations for 30-80nm c) the absolute error between observation and ML model predicted concentration for 30-80nm, d) Observation concentrations for 80-660nm, e) ML model predicted concentrations for 80-660nm f) the absolute error between observation and ML model predicted concentration for 80-660nm.

4.4 ML Model interrogation

Trollhaugen is a pristine environment and clearly dominated by natural processes that exhibit pronounced seasonal variation (Fig. 3). This is also demonstrated in the SHAP analysis (Figs 7a and 7b) where the majority of the top features have distinct seasonal cycles, represented by surface solar radiation, temperature, and time over sea ice (Fig. S6). Other studies at Antarctic sites have found a distinct austral summer and winter in aerosol properties, indicating the importance of solar intensity and temperature in the formation and growth of particles in the region (Kim et al., 2019). We find similar results and find that increased observed aerosol concentrations are associated with higher levels of surface solar radiation, higher temperatures, higher sea surface temperatures, and decreased time over sea ice in the summer months (Figs. 7, 8 and 9), associated with more open ocean, increased biological activity and therefore emission of primary marine aerosol (PMA) and DMS contributing to the aerosol burden.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

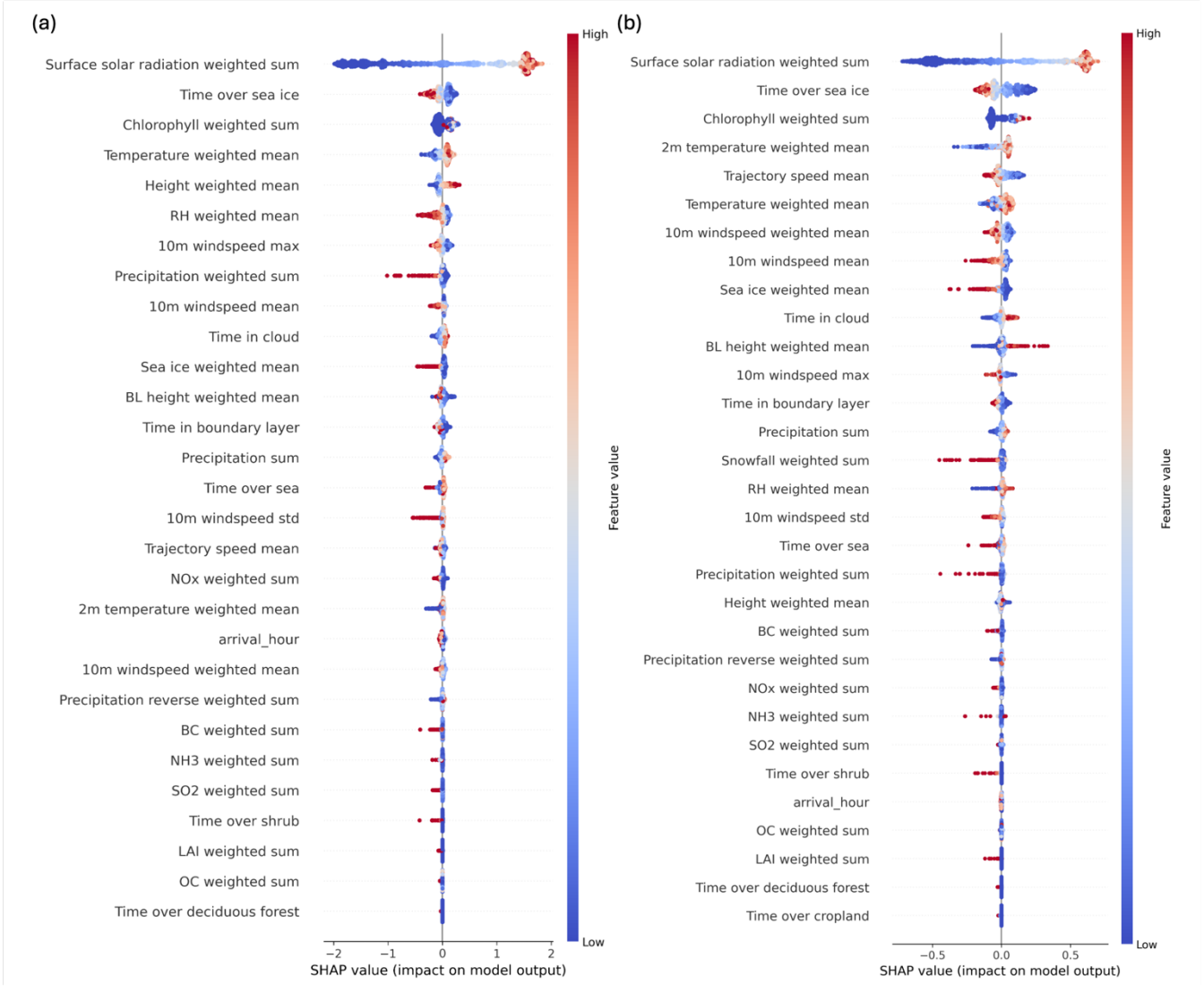


Figure 7: SHAP beeswarm for Trollhaugen (a) Aitken mode concentrations and (b) accumulation mode concentrations. Plots are ordered by the rank of each feature as determined by the TreeSHAP analysis. The points are coloured by the feature value corresponding to the data point for the SHAP value, and the 'violin' shape of the distribution represents the density of points.

As the SHAP ranking are based on the absolute mean of the SHAP values that can result in very close values for several features, the rank of a feature can change with different model runs, due to the stochastic nature of the model. Therefore, we do not focus heavily on the specific rank value of each feature in our analysis, rather the relationship between SHAP and feature values and the relative magnitude of the SHAP values for each model. We test the interventional TreeSHAP approach, which uses the training data as the background dataset to fill each missing feature, to ensure the robustness of our results shown

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

in Sect. S2.7. We see only slightly changes in SHAP rankings of features with the use of interventional SHAP (Fig. S10), and not at all in the relationships between SHAP values and features (Fig. S11), giving us confidence in the relationships found using the path-dependent method.

Key features have been identified through the investigation of SHAP rankings, to further investigate the processes represented by these features the SHAP-feature distributions, as well as spatial and covariate relationships, are analysed for the highest-ranking features in the following sections.

4.4.1 Surface solar radiation

For both size ranges we see a strong positive linear correlation between integrated surface solar radiation and SHAP for model prediction concentrations up to 50,000 W/m^2 accumulated over the 10-day trajectory period, corresponding to an approximate average of 210 W/m^2. At radiation levels exceeding the threshold there is an asymptotic plateau with no further increase in predicted aerosol concentration associated with increased solar insolation (Figs 8 and 9). A study by Fiebig et al. (2014) found that particle volume was linearly correlated with integral insolation in the Antarctic and suggested that photooxidative production was limited by photooxidative capacity, not the availability of precursor gases.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

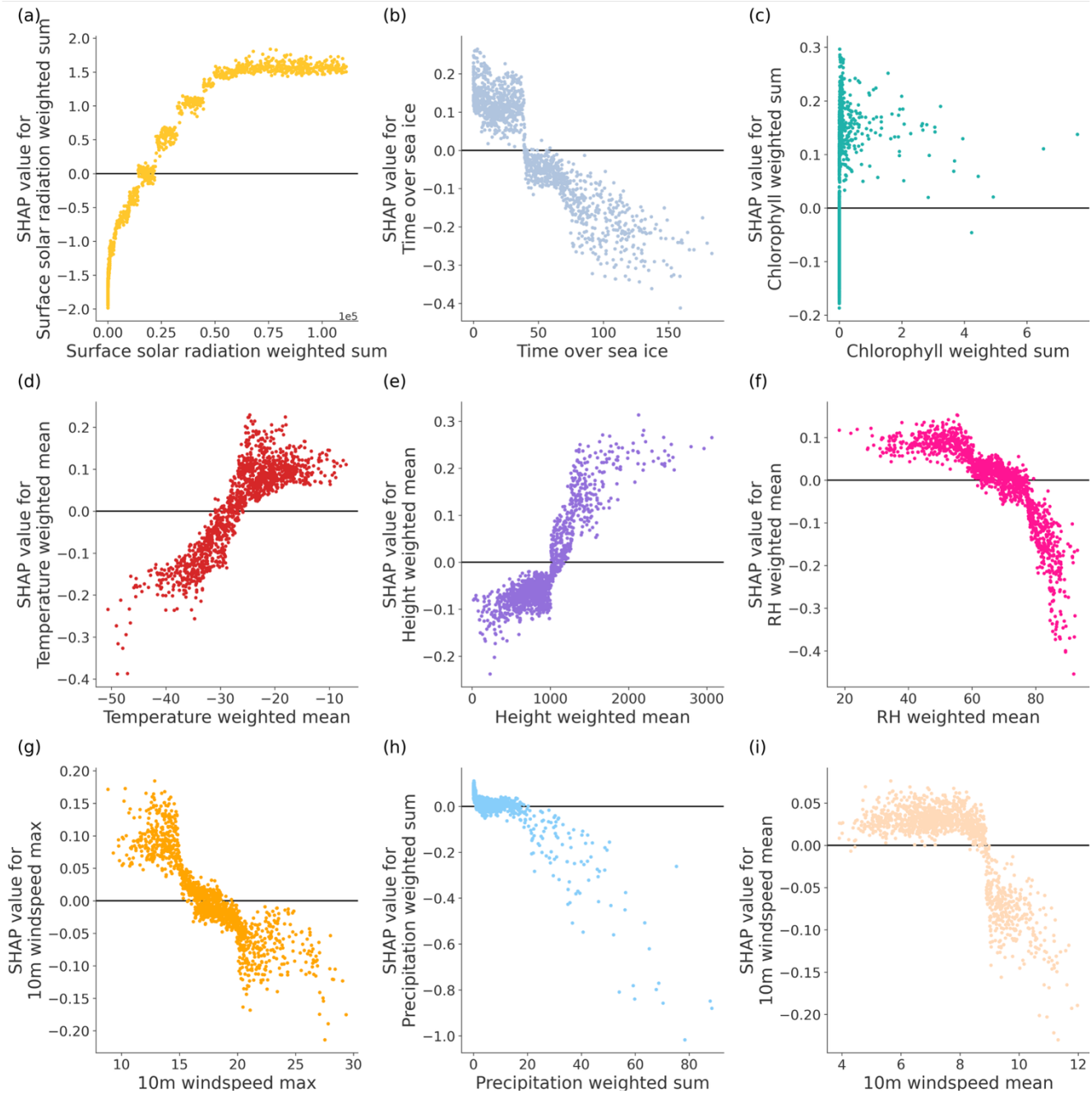
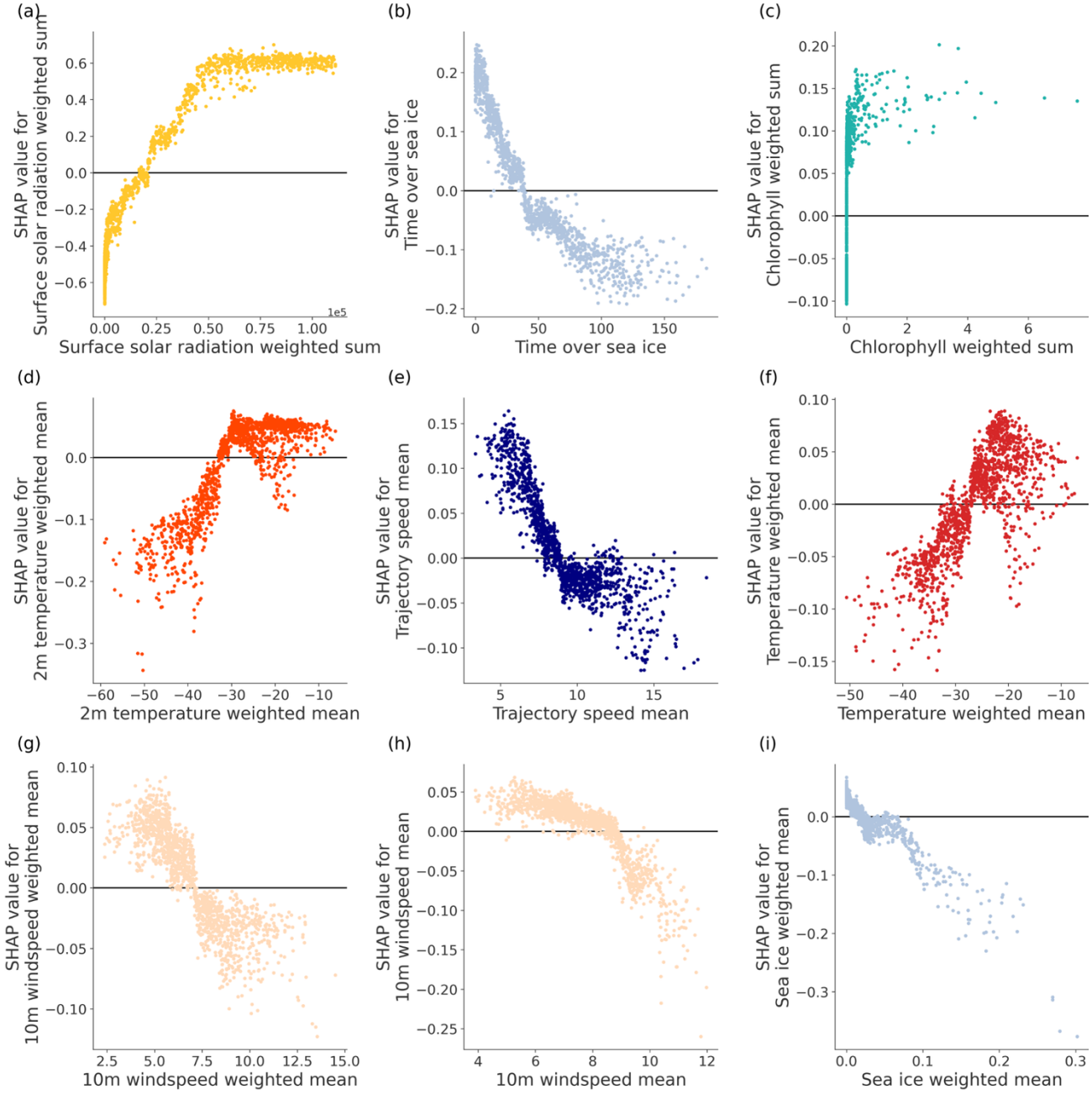


Figure 8: SHAP dependence plots for the Trollhaugen N30-80nm model. Plots are ordered by the rank of each feature as determined by the TreeSHAP analysis and the top 9 ranked features are shown. Colours are associated with each variable.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025



555

Figure 9: 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.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

4.4.2. Contribution of marine sources

Time over sea ice is one of the most important features for both N30-N80 and N80-N660 ML models showing a consistent negative SHAP-feature relationship (Figs 7-9). As found with numerous previous studies, with less sea ice, in the warmer months, Antarctic sites have increased aerosol concentrations associated with more contribution from the ocean (e.g. Lachlan-Cope et al., 2020; Paglione et al., 2024). Accumulated chlorophyll is also ranked highly for SHAP of both aerosol size range models (Fig. 7) and demonstrates a very consistent logarithmic relationship (Figs 8d and 9c). Chlorophyll in the model acts as a proxy for both secondary organic marine aerosol and primary marine organic aerosol (PMOA) and the high ranking suggests that these sources dominate aerosol prediction for this location. UKESM significantly underestimates the accumulation mode in the Austral summer, which could be associated with underestimation of PMOA or DMS concentrations (Mulcahy et al., 2020).

4.4.3 Generation and loss mechanisms associated with windspeed

The negative relationships between 10m wind speed features and SHAP (Figs 7-9) suggest that concentrations of Aitken and accumulation mode particles are not dominated by the primary production aerosol from breaking waves sea spray (driven by 10m wind speed). A recent study at Halley found strikingly low concentrations of sea salt aerosol and found that the aerosol concentrations (PM1) were dominated by secondary sources (Paglione et al., 2024), we conclude that this is highly likely to be the same for the Aitken mode at Trollhaugen from the SHAP results and the similarity in the position of the measurement sites on the Antarctic coast.

Sanchez et al., (2021) found similar relationships in the North Atlantic when considering the correlation between $N(D_p < 100$ nm), $N(D_p > 100$ nm) and 5-day airmass history sea surface wind speed and suggested, while seemingly counterintuitive, the inverse correlation they found between $N(D_p > 100$ nm) and sea surface windspeed was likely driven by enhanced PMA at higher wind speeds, that results in a larger condensation sink. PMA, while not expect to contribute significantly to particle number concentrations, was found to contribute to significantly to the total particle surface area, and thus the condensation sink. Therefore, the elevated total particle surface area from PMA at higher surface windspeeds could reduce the likelihood of occurrence of NPF (Cainey & Harvey, 2002; Yoon & Brimblecombe, 2002). We see this reflected in the combined SHAP plot for chlorophyll and 10m windspeed for the accumulation mode model: for larger accumulated chlorophyll concentrations, higher values of SHAP for the same chlorophyll concentrations, are associated with lower wind speed values (Fig. 10a). If a model is overestimating PMA, as UKESM has been found to in the Southern Ocean (Revell et al., 2019; Venugopal et al., 2025), then N80 and therefore CCN could be reduced due to high condensation sink caused by PMA.

PMA are more suspectable to deposition, so we could also be seeing the impact of dry deposition on a PMA contribution to accumulation mode particles (Sanchez et al. 2021). However, the gradient of the windspeed SHAP relationship is lower for

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

colder 2m temperatures (Fig. 10b), highlighting the dominance of this mechanism for warmer temperatures, when the contribution of secondary mechanisms from DMS will be higher.

Trajectory speed is an important variable to investigate growth of aerosol along trajectories; slower transport facilitates longer timescales for growth of particles during descent from free troposphere (Clarke et al., 1998) and transport in the boundary layer (Covert et al., 1996). It is shown to be very important for predictions for the N80-660nm ML model compared to the N30-80nm ML model, highlighting the importance of the role of growth during transport into the accumulation mode size range. Interrogation of the explanatory variables reveals that the trajectory speed mean is highly correlated with 10m windspeed mean (0.8) (Fig. S7) and demonstrate similar SHAP relationships (Fig. 10d), which suggests that trajectory speed mean could be representing the same wind driven processes, and due to the caveats of TreeSHAP discussed in Sect. 3.4 we cannot separate the importance of these features.

Similarly, to the higher correlation found in the Sanchez et al. (2021) study for $N(D_p > 100nm)$ and windspeed compared to $N(D_p < 100$ nm), we find that windspeed and trajectory speed features play a larger role in the accumulation mode model compared to the Aitken mode (Fig. 7). This could suggest that NPF in the free troposphere, which is independent of PMA in the boundary layer, is contributing more significantly to the Aitken mode concentrations, rather than NPF in the boundary layer.

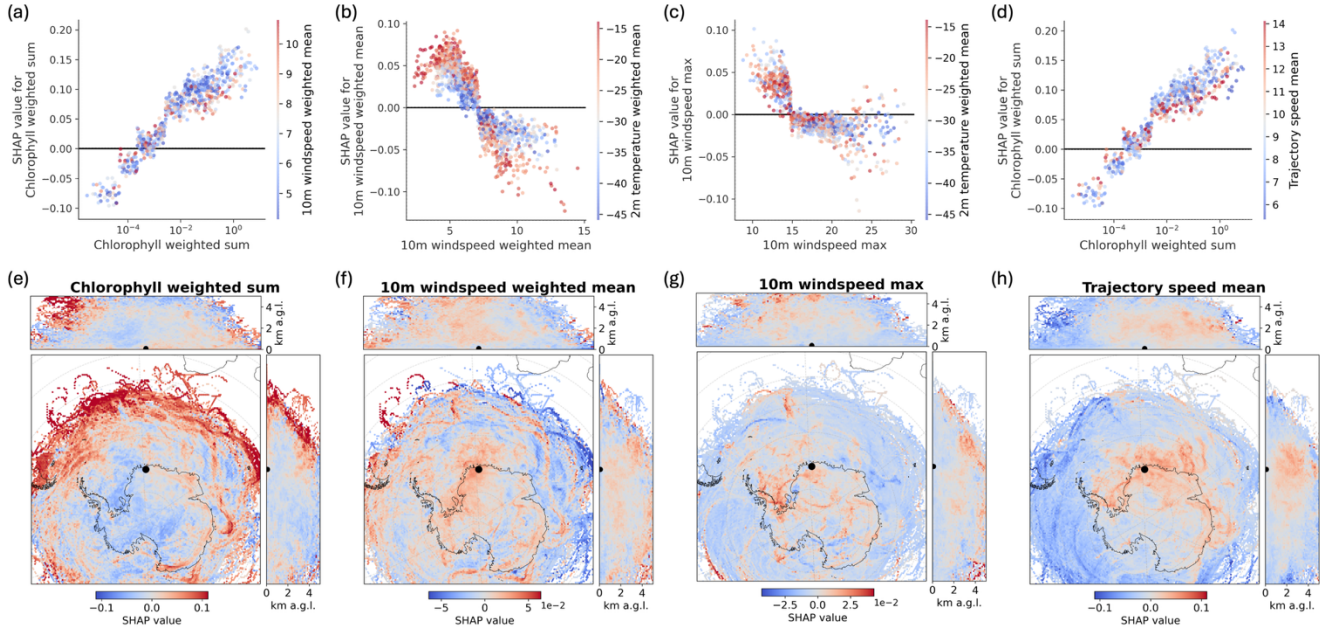


Figure 10: SHAP dependence plots at Trollhaugen with the corresponding values for a second variable indicated by the colour of points: (a) chlorophyll weighted mean with 10m windspeed weighted mean (b) 10m windspeed mean with 2m temperature weighted mean, (c) 10m windspeed max with 2m temperature weighted mean and (d) chlorophyll weighted mean with trajectory speed mean.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

SHAP spatial distributions using the CWT receptor model framework, with the vertical distribution shown in the side panels for (e) chlorophyll weighted sum, (f) 10m windspeed weighted mean, (g) 2m temperature weighted mean and (h) trajectory speed mean. Note that (a) and (d) use a log scale on the x-axis. All figures are for the N80-660nm model.

Although the dominant relationships with the windspeed features are negative, there are a few outliers of positive SHAP values for the highest maximum windspeeds for the accumulation mode model (Fig. 10c) above 20 ms^{-1}, which could be indicative of sea spray production in gusty conditions. Blowing snow has also been found to be an important source of aerosol in polar regions (Frey et al., 2019; Lachlan-Cope et al., 2020; Yang et al., 2019). First direct observations of blowing snow events were recorded during a storm with wind speed maxima between 15 and 25 ms^{-1} (Frey et al., 2019), the upper end of which corresponds to the values for which we see positive SHAP. However, there are only a few outliers associated with positive SHAP at high windspeed (Fig. 10c); thus, it is not possible conclude model representation of this source. Perhaps due to the extreme concentrations associated with blowing snow events, comparative to the low concentrations in the winter (Frey et al., 2019) and the sparsity of events in the dataset (Fig. S5), our model is not able to replicate these.

4.4.4 Role of vertical transport pathways and boundary layer structure

Average trajectory height (above ground level) plays an important role in the contribution to aerosol concentration prediction for the Aitken mode size range (Fig. 7a). With transport from aloft, SHAP values are higher, indicating a high aerosol concentration prediction. This is associated with transport from the free troposphere above the continental plateau, highlighted as the region of high SHAP in Fig. 11e.

Lachlan-Cope et al. (2020) proposed two mechanisms of NPF leading to aerosol concentrations at the Halley measurement site (a coastal site 940 km away from Trollhaugen) with airmasses arriving from marginal sea ice zone and those arriving from the free troposphere above the Antarctic plateau but could not conclude the relative importance of each mechanism. The results of Lachlan-Cope et al. (2020) pointed to secondary aerosol processes in sea ice regions and open-ocean water the regions, noting that these regions are not only sources of gaseous precursors but also of NPF. In the first mechanism, formation and growth occurs in the marine boundary layer, whereas in the second mechanism, precursor gases are lofted into the free troposphere, where nucleation and growth occur. Particles are brought down again by the Antarctic drainage flow (James, 1989) and then transported to the Antarctic coastal stations from the continent by the katabatic winds, in a mechanism first proposed by Ito (1993).

The relationship between trajectory height and SHAP values is much stronger for the Aitken mode and ranks much higher compared to the accumulation model SHAP results (Fig. 7). From the absolute mean SHAP value, height ranks higher

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

compared to parameters representative of boundary layer transport, indicating that this is due to the relative contribution of the NPF mechanisms, with transport from the free troposphere dominating the aerosol burden for the Aitken mode.

To investigate potential mechanisms leading to the contribution of Aitken mode particles from transport aloft, we investigate the seasonality and spatial pattern associated with the trajectory height relationship, as the second proposed mechanism is associated with DMS emission. Height is highly correlated with 2m temperature mean with a correlation coefficient of -0.6 (Fig. S7), and we see that for the covariate SHAP relationship, high SHAP values for trajectory height are mostly associated with low 2m temperatures (Fig. 11b).

Considering the seasonal cycle of the SHAP values, we see that height is positively contributing to predictions during the transition between Austral Summer and Winter (March, September and October) (Fig. 11c). From the spatial distribution for these months (Fig. 11e) we see high SHAP for trajectory height with transport from Bellingshausen and Ross Sea regions, and to a lesser extent the Weddell Sea, suggesting lofting of precursor gases into the free troposphere. Whilst the trajectory height also contributes to predictions during the Austral winter, the strongest positive influence is during March, September and October, suggesting the importance of photooxidative processes in contributions from aloft to Aitken mode concentrations. For summer (January, February, November and December), where the trajectory height contributes much less to model predictions (Fig. 11c), the spatial transport pattern is markedly different and more constrained, with the majority of transport coming from Southern Ocean and East of the Antarctic continent (Fig. 11d). There are outliers which have strong positive SHAP associated with transport from aloft (Fig. 11c), which can be clearly seen on the SHAP spatial maps, associated with transport from the free troposphere, likely from long range transport, and perhaps some contribution from the Lazarev and Ross Seas (Fig. 11d). In order to investigate these mechanisms further, longer back-trajectories would be required to investigate the contribution from long-range transport to the Aitken mode burden at Trollhaugen. However, longer trajectories are associated with increased uncertainty (Engström & Magnusson, 2009).

Sedimentation from polar stratospheric clouds (PSCs) have been suggested in previous studies to contribute significantly to the nitrate aerosol burden at coastal Antarctic sites (Frey et al., 2009; Savarino et al., 2007; Traversi et al., 2014), which could be associated with the high contribution to the Aitken mode from aloft. The formation and contribution from PSCs has been suggested to peak during the winter and early summer (Frey et al., 2009; Savarino et al., 2007; Wagenbach et al., 1998), which could coincide with the peak contribution from the feature 'height weighted mean' to the prediction of Aitken mode concentrations, however, this contrasts with the additional peak contribution during March (Fig. 11c). Chemical speciation at the in-situ measurement site could enable separation of the potential sources from aloft that contribute to the Aitken mode.

For the accumulation mode, whilst playing a much smaller role in model predictions, with low SHAP values, the SHAP results show an interesting bimodal relationship (Fig. 11a): for warmer 2m temperatures we see a mostly negative relationship with average trajectory height up to around 1000m above ground level, and for average heights above 1000m, associated with colder

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

temperatures we see a positive relationship. This demonstrates a clear change in the dominant sources of aerosol throughout the seasons: local PMA transported in the boundary layer during the summer and transport from aloft during the winter. Cloud processing and condensation growth within the marine boundary layer have been found to be important for increasing accumulation mode concentrations in the Southern Ocean (McCoy et al., 2021), therefore the negative relationship for higher temperatures could be associated with the growth of NPF particles into the accumulation size range during higher biological activity in the warmer months.

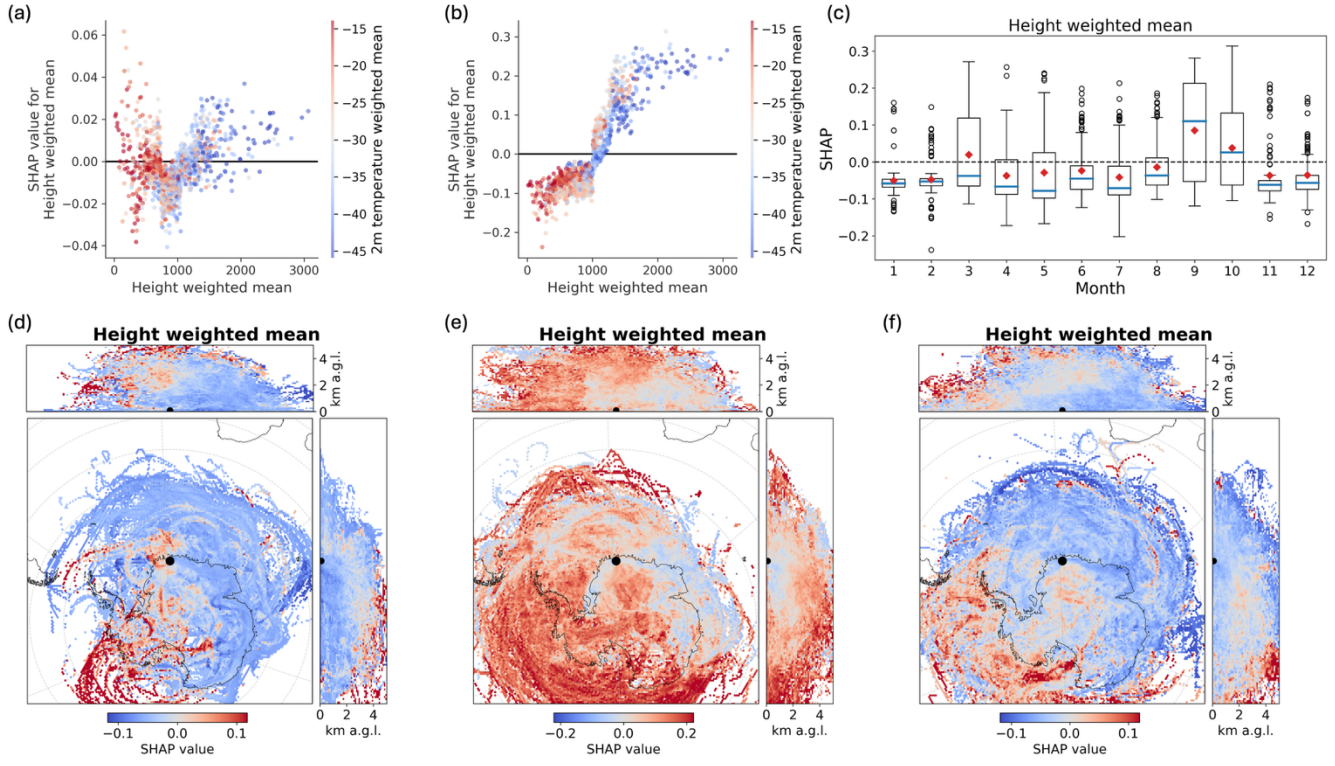


Figure 11: SHAP value dependence plots for height weighted mean with the corresponding 2m temperature values indicated by the colour of points for (a) the N80-660nm mode and (b) the N30-80nm model. (c) The boxplots for the SHAP value for height weighted mean for the N30-80nm model for each month. SHAP spatial distributions for height weighted mean, using the CWT receptor model framework, with the vertical distribution shown in the side panels for (d) January, February, November and December, (e) March, September and October and (f) April-August (inclusive) for the N30-80nm model.

Boundary layer height plays an important role for accumulation mode concentration predictions and Aitken mode concentration predictions demonstrated by the SHAP relationships in Fig. 7, but the contrasting relationships highlight the difference in dominating processes contributing to prediction of each mode. With increased boundary layer depth, we see lower SHAP values for the Aitken mode model (Fig. 12a), relating to dilution aerosols in a larger volume of air, thus lower

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

concentrations at the measurement site. Additionally, there will be dilution of pre-existing aerosol emissions in a larger volume of air, therefore less growth through condensation and coagulation. A deeper boundary layer also results in additional transport to the upper troposphere, for gas-phase species of low solubility such as DMS, which are not scavenged and thus are entrained more easily into the free troposphere (Zheng et al., 2021) compared to PMA. For the accumulation mode there is a contrasting SHAP-feature relationship: increased boundary layer depths and warmer 2m temperatures are associated with higher concentrations of accumulation mode particles (Fig. 12c), perhaps associated with the increased lifetime of particles, the mechanisms of which warrant investigation in future work.

Time in boundary layer shows similar relationships for both Aitken and accumulation mode, with longer proportions of the trajectory spent in the boundary layer contributing negatively to the prediction of aerosol concentration. This contribution is much more pronounced for cooler months compared to summer months (Figs 12 b and d), likely due to the greater contribution of aerosol from aerosols aloft during winter.

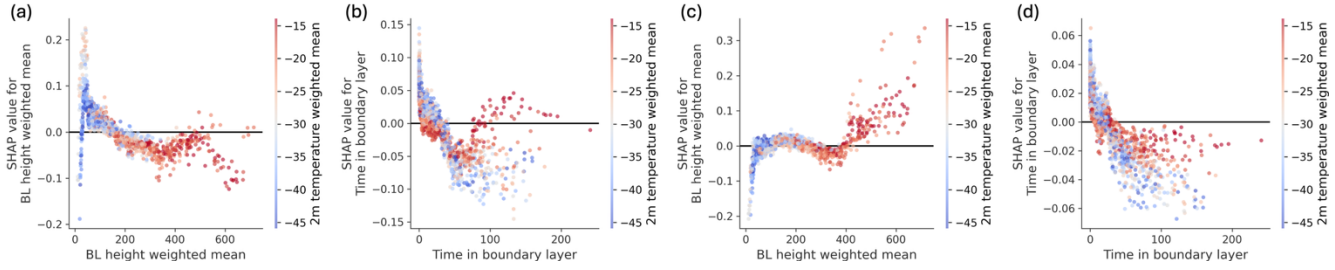


Figure 12: SHAP dependence plots with the corresponding values for 2m temperature weighted mean indicated by the colour of points: (a and c) BL height weighted mean, (b and d) Time in boundary layer for (a-b) the N30-80nm model and (c-d) the N80-660nm model.

4.4.5 Role of cloud processing and relative humidity during transport

The contrasting relationships between RH and SHAP value for the two size range models (Figs 13a and 13c) highlights the importance of RH for growth into the accumulation mode, leading to a reduced concentration in the Aitken mode.

'Time in cloud' is not a weighted variable and therefore, represents the total experienced cloud and averaged relative humidity during the 10-day back-trajectory and shows consistent relationships for both model size ranges (Figs 13b and 13d). With increased time in cloud during transport, there is increased SHAP for both accumulation and Aitken mode models. This suggests the role of cloud processing, resulting in populations of larger sized aerosol after removal of aerosol populations during transport, similarly to precipitation.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

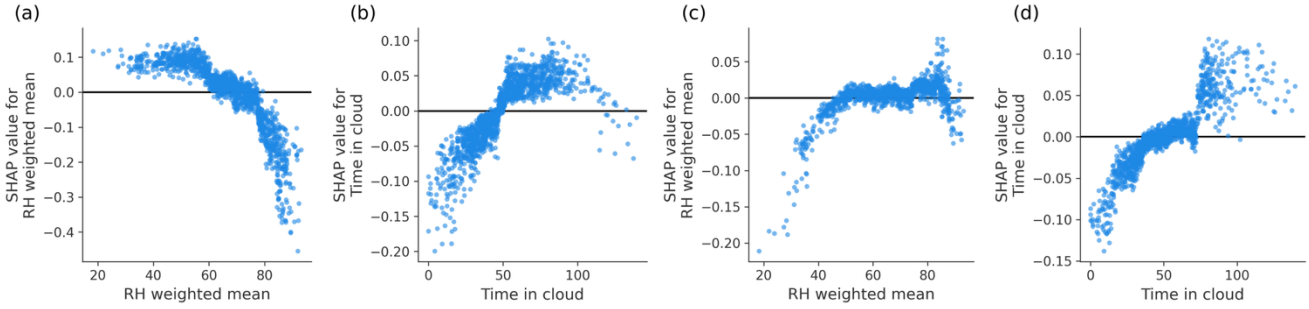


Figure 13: SHAP dependence plots for (a, c) RH weighted mean and (b, d) Time in cloud. Where (a and b) and for the 30-80nm model and (c and d) for the 80-660nm model.

4.4.6 Precipitation experienced during transport

Precipitation is shown to act as a sink for aerosol at Trollhaugen when we consider the weighted sum: higher values of weighted accumulated precipitation are associated with lower SHAP values for precipitation (Figs 14a and 14c). This highlights the role of precipitation scavenging as a dominant aerosol loss process, particularly for precipitation closer to the measurement site. However, considering the accumulated precipitation with no weighting, allows us to consider the impact of precipitation further away from the site. For both size ranges we see a strong positive relationship with SHAP (Figs 14b and 14d), linked in previous studies to a reduction in condensation sink, providing preferential conditions for new particle formation (Andronache, 2004; Ueda et al., 2016), and therefore growth into the size ranges considered in this study. For the weighted sum, only a few predictions are strongly affected by the wet removal, the positive relationship with non-weighted sum is consistently stronger and dominates the feature importance for the Aitken mode (Fig. 7). The positive region of SHAP is over the Southern Ocean, associated with trajectories travelling close to the surface for both size range models (Figs 5b, 14f and 14h). The duality of the relationship with precipitation highlights the importance of considering the timing of precipitation during transport.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

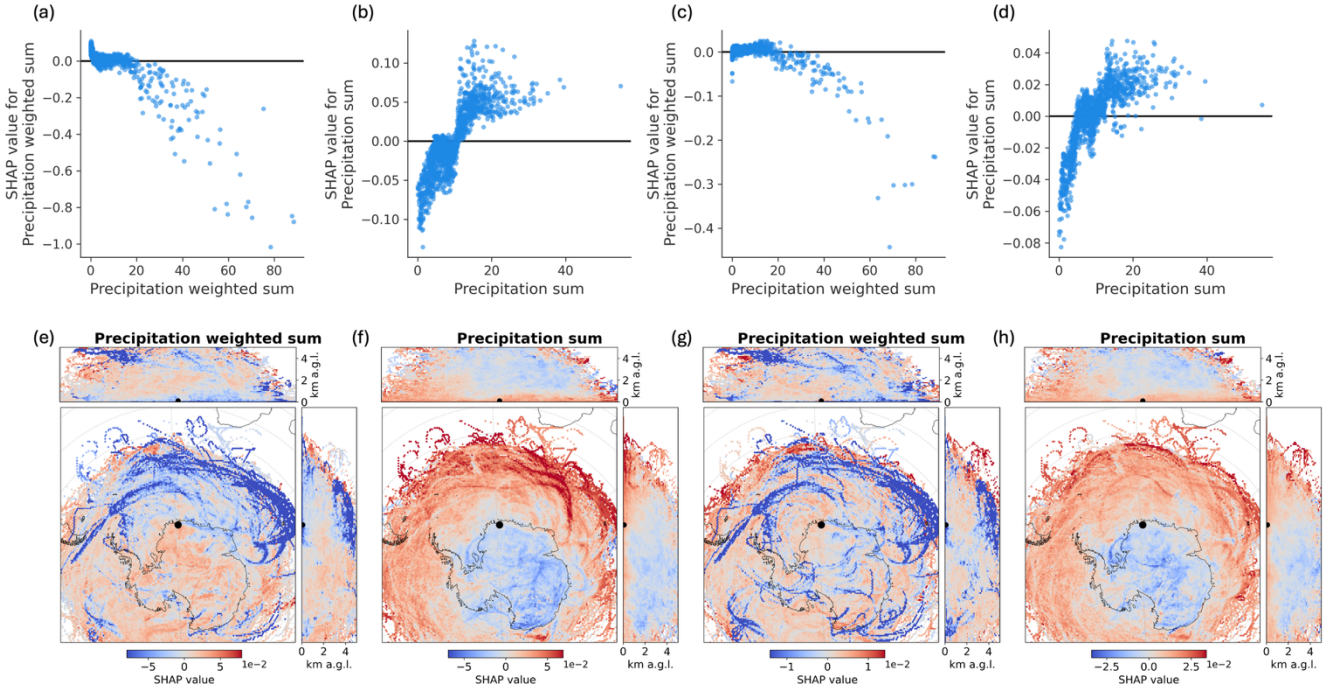


Figure 14: SHAP dependence plots for (a, c) precipitation weighted sum and (b,d) precipitation sum. Where (a and b) and for the 30-80nm model and (c and d) for the 80-660nm model. SHAP spatial distributions using the CWT receptor model framework, with the vertical distribution shown in the side panels for (e, g) precipitation weighted sum and (f,h) precipitation sum. Where (e and f) and for the 30-80nm model and (g and h) for the 80-660nm model.

It is worth noting that for the Antarctic most precipitation will be snow, demonstrated by the extremely high correlation between snowfall weighted sum and precipitation weighted sum (Fig. S7), so these are not analysed separately in this study, however for other environments it could be important to distinguish precipitation types.

4.5 Testing the generality of the framework

To demonstrate the generality of the newly developed ML framework we provide proof of concept results for two additional stations representative of two distinctly different aerosol environments. A detailed assessment of the processes driving aerosol properties in these environments will be the focus of a future study. Here we demonstrate the performance of the ML model for the accumulation size range for a site in the boreal forest and a maritime continental site. The same methodology was applied to regress aerosol concentrations calculated from in-situ PNSD measurements from Värriö (VAR) SMEAR I measurement station (67.767°N, 29.583°E, 390m a.s.l.), 120km north of the Arctic circle (Hari et al., 1994) and Mace Head (53.3267º N, 9.9046º W, 8m a.s.l.) on the coast of Ireland (O'Dowd et al., 1998). Back trajectories were calculated using the trajectory release locations described in Table 5 and the study was conducted over the periods described in Table 5, with 2012 used as the test year.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

740

Station	Description	Years included in study (inclusive)	Diameter range included (nm)	Trajectory release location			Trajectory length (hrs)
				Latitude	Longitude	Height m A.G.L.	
Värriö	Boreal forest	2009-2016 (excl. 2014)	30-790	-72.012	2.535	150	240
Mace Head	Maritime continental	2009-2018 (excl. 2013-2016)	30-470	53.327	-9.9040	100	240

Table 5: The receptor sites with the site description, years included in the study, diameter range and the trajectory release locations.

The Eulerian evaluation results are shown here (Fig. 15), demonstrating the performance of the ML models at each site for the test year (2012). Both the models are able to accurately replicate the seasonal cycles for each site and the variance. However, for some of the highest concentrations (January and February for Värriö, and March for Mace Head) the ML models are not
745 able to replicate the observed range. Overall, the models are able to accurately predict the observed concentrations, thus demonstrating the generality of the framework to different environments, which will be the subject of future studies.

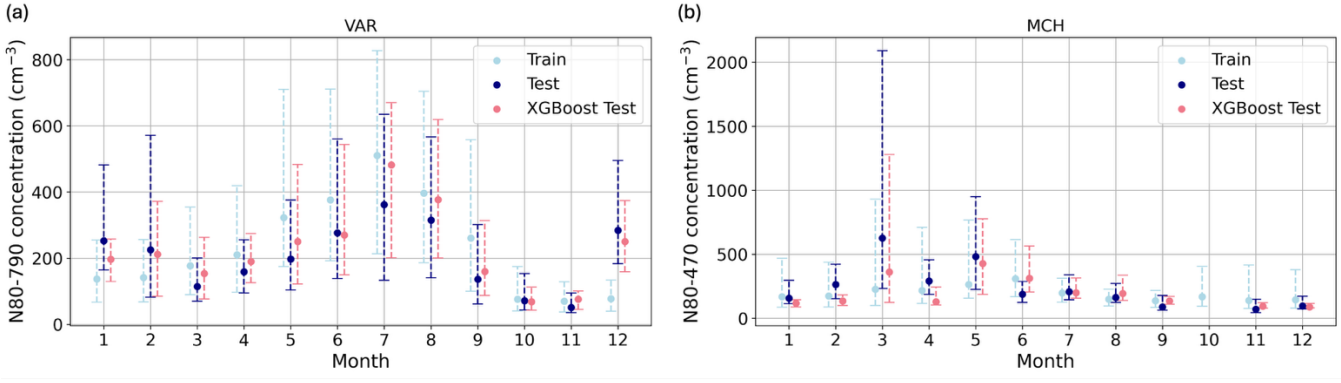


Figure 15: The seasonal cycle of concentrations (cm^{-3}) at (a) VAR (Värriö) for 80-790nm and (b) MCH (Mace Head) for 80-470nm for the observations for the years used for model training (light blue, left shift), the observations for the year used to test the model
750 **(navy, centred) and the ML model (XGBoost) predictions for the test year (pink, right shift).**

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

5 Conclusions

Our objectives were to (1) identify the seasonal cycle of aerosol number concentration and the dominant source regions for the measurement site, in order to (2) build and evaluate a model for accurate predictions of aerosol concentrations from airmass history, (3) elucidate the relationships between meteorological parameters, source proxies and aerosol concentrations, in order to (4) identify dominant source and sink processes controlling the number concentration for Aitken and accumulation modes in the Antarctic. These aims were realised in this study through the development and interrogation of Lagrangian based explainable AI regression models.

This study has demonstrated the power of the explainable machine learning framework to predict aerosol concentrations from airmass histories, yielding representative models of distinct regions. The regression models built for each size range at the case study site have been shown to accurately replicate the seasonal cycle and potential source regions, and generally exhibited relative strong predictive power. However, for extremes in aerosol concentrations, the limits in the models' predictive capabilities were demonstrated. This limitation was particularly evident for the Aitken mode size range model which is strongly influenced by NPF. Future work could explore additional model learning objectives and architectures, however, the sparsity of data of the extreme cases limits the ability to build a highly effective model for these events in this framework. Despite extensive additional filtering efforts, there also remains the potential for contamination of the underlying PNSD data by instrument errors and generator contamination at the site, this could be contributing to some of the extremes in the timeseries of aerosol concentration data.

For the first time we identify dominating processes by simultaneously considering most processes and their interactions, finding distinct dominant processes for each size range, summarised in Fig. 16. Aerosol concentrations at Trollhaugen are dominated by marine natural sources and well as transport from the free troposphere. The contribution from the free troposphere dominates aerosol burden of the Aitken mode in the transition periods between summer and winter, compared to a larger contribution in the summer from local marine sources from transport in the boundary layer. Longer trajectories coupled with additional variables such as height cluster in future studies could enable further confirmation of the sources of tropospheric aerosol contributing to the Aitken concentrations. For the accumulation mode, local marine sources dominate the number concentration for the warmer months, but transport from aloft is the dominant contributor to the low concentrations measured during the winter. PMA plays a greater role as condensation sink for Trollhaugen accumulation mode concentration rather than a source, with NPF resulting in higher concentrations during the warmer months. This study has highlighted the importance of considering transport and airmass history during aerosol studies and the importance of timing for variables such as precipitation.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

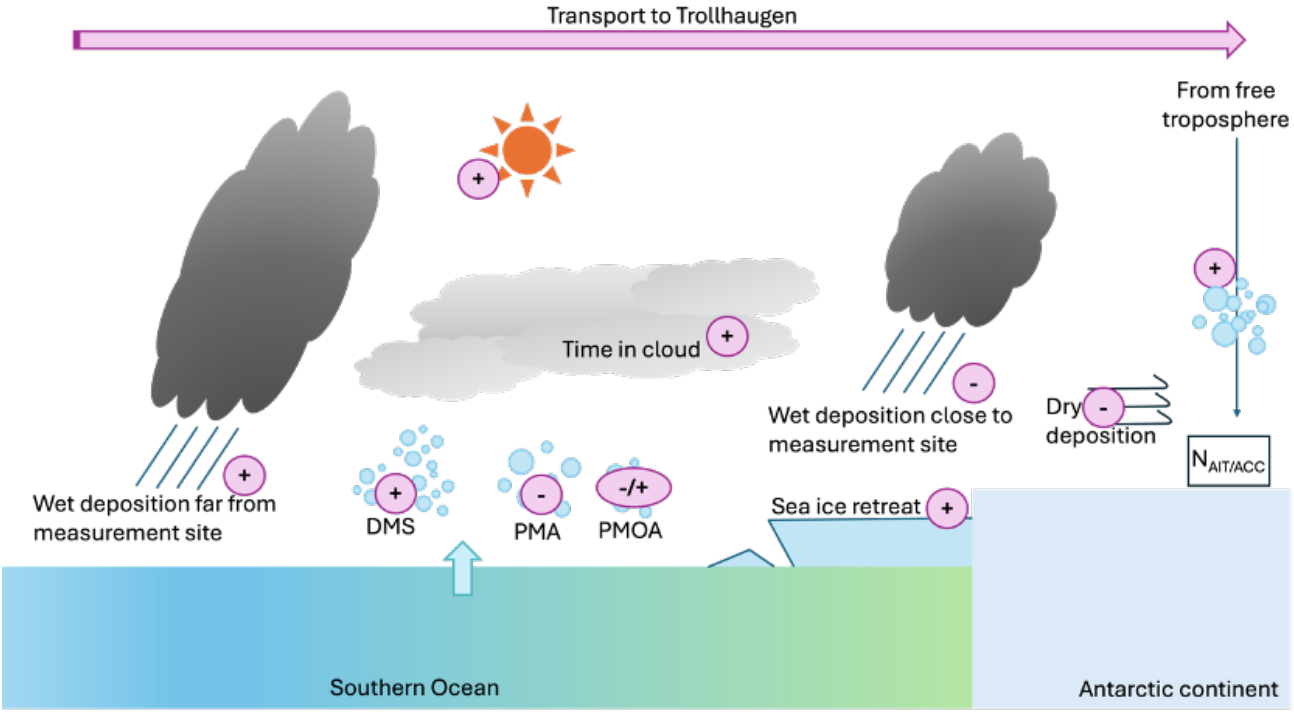


Figure 16: Schematic of the processes during transport controlling aerosol lifecycle and thus the aerosol particle number concentrations measured at Trollhaugen.

In future work, the driving meteorology, which contributes the most to the uncertainty in trajectories, could be replaced with ERA5, which is a higher resolution reanalysis product. Furthermore, an ensemble of trajectories (with small perturbations to the underlying meteorological grid in the three dimensions) could be used to reduce uncertainty in the trajectories. However, these steps would significantly add to the computational burden, particularly for the calculation of trajectories and the collocation of explanatory variables.

This analysis was not able to discern the influence of blowing snow which is known to be a significant source of SSA in the Antarctic, additional variables such as terrestrial ice and snow depth would allow for further exploration of the importance of this source at Trollhaugen. Additional sources, such as ammonia and nitrates from natural terrestrial sources such as sea bird colonies have previously been shown to significantly contribute to aerosol populations in the Antarctic (Boyer et al., 2025; Brean et al., 2025). There is not currently a clear representation of these sources using the current explanatory variables, to explore the addition of a representation of seabird colonies, seabird tracking datasets could be tested (such as https://www.seabirdtracking.org/resources/). Chemical speciation at Trollhaugen would be useful to provide more certainty and separation of aerosol sources alongside the current framework.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

In future studies, combining datasets from multiple in-situ stations could enable provide more data coverage for data sparse regions such as the Antarctic, similarly to Pernov et al. (2024) which built Pan-Arctic models, enabling a more holistic view of an environment. This would also allow for exploring the generality of a Pan-Antarctic model by transfer learning (by testing on a site not included in the training dataset). Moreover, expanding the dataset through the addition of measurement sites would allow for investigation into interannual relationships and differences in dominant processes across an environment.

By comparing to UKESM as a benchmark we can highlight potential areas for model development at key locations. In future work upon expansion of the framework to further sites, consistent sources of model bias can be identified, to improve representation of natural aerosol processes in GCMs tuned for best global representation.

Using a holistic approach to aerosol studies allows us to explore the relative importance of different aerosol processes and the interactions between them. This study has significantly contributed to the understanding of aerosol processes in a pristine environment in which GCMs have been shown to exhibit significant biases, thus is a key environment to reduce uncertainty in climate modelling. The framework developed in this study has the flexibility to easily adapt to different aerosol sources and processes through the choice of explanatory variables and can be extended to any site with aerosol size distribution measurements. Future studies will explore further the generalisability of the framework by applying the framework to additional environments and perform a transparent, process-based evaluation of GCMs using trajectories derived from GCM meteorological data combined with GCM output for explanatory and target variables in a consistent manner.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Data availability

Field data (particle number size distributions) are freely available from the EBAS database at http://ebas.nilu.no/ (last access: 21st October 2021; NILU, 2021), full citations provided in Table S4. Satellite and reanalysis products are freely available from the databases described in Table S5.

The datasets for the framework used in this study will be made available upon publication.

Code availability

Data analysis was conducted in Python (version 3.9.16), and colour maps for the figures considering colour vision deficiencies were inspired by Crameri et al., (2020).

Python scripts used for the analysis and plotting will be made available upon publication.

Author contribution

DGP, EKD and JEF conceptualized the idea of the study, the machine learning (ML) framework and simulations. EKD designed the ML framework and setup (with support from DGP and JEF), prepared the ML input data, performed the ML simulations, analysed the simulation outputs and made the figures, with contributions from DGP, JEF and AS. EKD, DGP, JEF and JH selected and designed the explanatory and target variables. EKD wrote the machine learning code with support from JEF and DGP. PK wrote the Lagrangian modelling code which EKD adapted and reconfigured for use with the machine learning framework. DGP, AS and ET designed and configured the UKESM simulations and ET performed the UKESM simulations. EKD processed the UKESM output with support from DGP, AS, ET and PK. ET formatted the Mace Head SMPS data. The manuscript was written by EKD with contribution from DGP and JEF. All co-authors commented, edited and gave feedback on the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

We thank technical and scientific staff from the Trollhaugen surface station. We acknowledge use of the Monsoon2 system, a collaborative facility supplied under the Joint Weather and Climate Research Programme, a strategic partnership between the UK Met Office and the Natural Environment Research Council. We also thank all the people responsible for the development of UKESM.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

The authors would like to thank Duncan Watson-Parris for his valuable input and discussions on machine learning techniques during the development of this work. D.G.P. would like to extend personal thanks to Ben Johnson and Andy Jones, who provided support for the configuration of the UKESM simulations performed as part the AeroCom GCM Trajectory experiment on which these simulations are based and to Zak Kipling who supported the calculation of trajectories from ERA-Interim data. E.K.D. would also like to thank Ben Johnson for his help with the processing of GFED emission data.

Data used in this study were accessed from EBAS (https://ebas.nilu.no) hosted by NILU. Specifically, the use included data affiliated with the frameworks: ACTRIS, GAW-WDCA, EMEP.

Financial support

The work has been supported by a doctoral training grant awarded as part of the UKRI AI Centre for Doctoral Training in Environmental Intelligence (UKRI grant no. EP/S022074/1) and the UK Met Office through a CASE PhD studentship. DGP has received support from NERC project (grant no. NE/T006331/1) and Horizon Europe programme via project CleanCloud (Clouds and climate transitioning to post-fossil aerosol regime), grant agreement ID: 101137639. ET has received support from NERC GW4+ (grant no. NE/L002434/1). AS was supported by the Met Office Hadley Centre Climate Programme funded by DSIT.

The ACTRIS project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654109.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

References

Andronache, C.: Precipitation removal of ultrafine aerosol particles from the atmospheric boundary layer, J. Geophys. Res.: Atmos., 109, 16–23, https://doi.org/10.1029/2003JD004050, 2004.

Arjunan Nair, A. and Yu, F.: Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements, Atmos. Chem. Phys., 20, 12853–12869, https://doi.org/10.5194/ACP-20-12853-2020, 2020.

Asmi, E., Kivekäs, N., Kerminen, V.-M., Komppula, M., Hyvärinen, A.-P., Hatakka, J., Viisanen, Y., and Lihavainen, H.: Secondary new particle formation in Northern Finland Pallas site between the years 2000 and 2010, Atmos. Chem. Phys, 11, 12959–12972, https://doi.org/10.5194/acp-11-12959-2011, 2011.

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., Boucher, O., Carslaw, K. S., Christensen, M., Daniau, A. L., Dufresne, J. L., Feingold, G., Fiedler, S., Forster, P., Gettelman, A., Haywood, J. M., Lohmann, U., Malavelle, F., Mauritsen, T., McCoy, D. T., Myhre, G., Mülmenstädt, J., Neubauer, D., Possner, A., Rugenstein, M., Sato, Y., Schulz, M., Schwartz, S. E., Sourdeval, O., Storelvmo, T., Toll, V., Winker, D., and Stevens, B.: Bounding Global Aerosol Radiative Forcing of Climate Change, Rev. Geophys., 58, e2019RG000660, https://doi.org/10.1029/2019RG000660, 2020.

Bergstra, J., Yamins, D., and Cox, D.: Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures, in Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 17-19th June 2013, 115-123, 2013.

Blichner, S. M., Yli-Juuti, T., Mielonen, T., Pöhlker, C., Holopainen, E., Heikkinen, L., Mohr, C., Artaxo, P., Carbone, S., Meller, B. B., Quaresma Dias-Júnior, C., Kulmala, M., Petäjä, T., Scott, C. E., Svenhag, C., Nieradzik, L., Sporre, M., Partridge, D. G., Tovazzi, E., Virtanen, A., Kokkola, H., and Riipinen, I.: Process-evaluation of forest aerosol-cloud-climate feedback shows clear evidence from observations and large uncertainty in models, Nat. Comms. 2024 15:1, 15, 1–12, https://doi.org/10.1038/s41467-024-45001-y, 2024.

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S., Sherwood, S., Stevens, B., Zhang, X., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P.: Clouds and Aerosols. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 2013.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Bowman, K. P., Lin, J. C., Stohl, A., Draxler, R., Konopka, P., Andrews, A., and Brunner, D.: Input Data Requirements for Lagrangian Trajectory Models, Bull. Am. Meteorol. Soc., 94, 1051–1058, https://doi.org/10.1175/BAMS-D-12-00076.1, 2013.

Brean, J., Beddows, D. C. S., Asmi, E., Virkkula, A., Quéléver, L. L. J., Sipilä, M., Van Den Heuvel, F., Lachlan-Cope, T., Jones, A., Frey, M., Lupi, A., Park, J., Yoon, Y. J., Weller, R., Marincovich, G. L., Mulena, G. C., Harrison, R. M., and Dall'Osto, M.: Multiple eco-regions contribute to the seasonal cycle of Antarctic aerosol size distributions, Atmos. Chem. Phys., 25, 1145–1162, https://doi.org/10.5194/ACP-25-1145-2025, 2025.

Cainey, J. M. and Harvey, M.: Dimethylsulfide, a limited contributor to new particle formation in the clean marine boundary layer, Geophys. Res. Lett., 29, 32–1, https://doi.org/10.1029/2001GL014439, 2002.

Carslaw, K. S., Lee, L. A., Reddington, C. L., Pringle, K. J., Rap, A., Forster, P. M., Mann, G. W., Spracklen, D. V., Woodhouse, M. T., Regayre, L. A., and Pierce, J. R.: Large contribution of natural aerosols to uncertainty in indirect forcing. Nature, 503, 67–71, https://doi.org/10.1038/nature12674, 2013.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Chen, Y., Haywood, J., Wang, Y., Malavelle, F., Jordan, G., Partridge, D., Fieldsend, J., De Leeuw, J., Schmidt, A., Cho, N., Oreopoulos, L., Platnick, S., Grosvenor, D., Field, P., and Lohmann, U.: Machine learning reveals climate forcing from aerosols is dominated by increased cloud cover, Nat. Geo., 15:8, 15, 609–614, https://doi.org/10.1038/s41561-022-00991-6, 2022.

Chen, Y., Haywood, J., Wang, Y., Malavelle, F., Jordan, G., Peace, A., Partridge, D. G., Cho, N., Oreopoulos, L., Grosvenor, D., Field, P., Allan, R. P., and Lohmann, U.: Substantial cooling effect from aerosol-induced increase in tropical marine cloud cover, Nat. Geo. 17:5, 17, 404–410, https://doi.org/10.1038/s41561-024-01427-z, 2024.

Chen, Y.-C., Li, J.-L. F., Lee, W.-L., Stowell, J. D., Bi, J., Al-Hamdan, M. Z., Geng, G., Meng, X., He, K., and Liu, Y.: Random forest models for PM2.5 speciation concentrations using MISR fractional AODs, Environ. Res. Lett., 15, 034056, https://doi.org/10.1088/1748-9326/AB76DF, 2020.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Dal Maso, M., Sogacheva, L., Aalto, P. P., Riipinen, I., Komppula, M., Tunved, P., Korhonen, L., Suur-Uski, V., Hirsikko, A., Kurtén, T., Kerminen, V.-M., Lihavainen, H., Viisanen, Y., Hansson, H.-C., and Kulmala, M.,: Aerosol size distribution measurements at four Nordic field stations: identification, analysis and trajectory analysis of new particle formation bursts, Tellus B: Chem. Phys. Meteorol., 59(3), 350–361. https://doi.org/10.1111/j.1600-0889.2007.00267.x, 2007.

Dall'Osto, M., Beddows, D. C. S., Tunved, P., Krejci, R., Ström, J., Hansson, H. C., Yoon, Y. J., Park, K. T., Becagli, S., Udisti, R., Onasch, T., Ódowd, C. D., Simó, R., and Harrison, R. M.: Arctic sea ice melt leads to atmospheric new particle formation, Sci. Rep. 2017 7:1, 7, 1–10, https://doi.org/10.1038/s41598-017-03328-1, 2017.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. R. Meteorolog. Soc., 137, 553–597, https://doi.org/10.1002/QJ.828, 2011.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., and Wimmer, W.: The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system, Remote Sens. Environ., 116, 140–158, https://doi.org/10.1016/J.RSE.2010.10.017, 2012.

Elson, P., Sales de Andrade, E., Lucas, G., May, R., Hattersley, R., Campbell, E., Dawson, A., Raynaud, S., scmc72, Little, B., Snow, A. D., Donkers, K., Blay, B., Killick, P., Wilson, N., Peglar, P., lbdreyer, Andrew, Szymaniak, J., Berchet, A., Bosley, C., Davis, L., Filipe, Krasting, J., Bradbury, M., Kirkham, D., stephenworsley10, Clément, Caria, G., and Hedley, M.: SciTools/cartopy: v0.20.2 [code], https://doi.org/10.5281/ZENODO.5842769, 2022.

Engström, A. and Magnusson, L.: Estimating trajectory uncertainties due to flow dependent errors in the atmospheric analysis, Atmos. Chem. Phys., 9, 8857–8867, https://doi.org/10.5194/acp-9-8857-2009, 2009.

Fanton d'Andon, O., Mangin, A., Lavender, S., Antoine, D., Maritorena, S., Morel, A., and Barrot, G.: GlobColour - the European Service for Ocean Colour, in: Proceedings of the 2009 IEEE International Geoscience & Remote Sensing Symposium, 2009.

Fiddes, S. L., Mallet, M. D., Protat, A., Woodhouse, M. T., Alexander, S. P., and Furtado, K.: A machine learning approach for evaluating Southern Ocean cloud radiative biases in a global atmosphere model, Geosci. Model Dev., 17, 2641–2662, https://doi.org/10.5194/GMD-17-2641-2024, 2024.

Fiebig, M., Hirdman, D., Lunder, C. R., Ogren, J. A., Solberg, S., Stohl, A., and Thompson, R. L.: Atmospheric Chemistry and Physics Annual cycle of Antarctic baseline aerosol: controlled by photooxidation-limited aerosol formation, Atmos. Chem. Phys, 14, 3083–3093, https://doi.org/10.5194/acp-14-3083-2014, 2014.

Frey, M., Norris, S., Brooks, I., Anderson, P., Nishimura, K., Yang, X., Jones, A., Nerentorp Mastromonaco, M., Jones, D., and Wolff, E.: First direct observation of sea salt aerosol production from blowing snow above sea ice, Atmos. Chem. Phys., 1–53, https://doi.org/10.5194/acp-2019-259, 2019.

Frey, M., Savarino, J., Morin, S., Erbland, J., & Martins, J. M. F.: Photolysis imprint in the nitrate stable isotope signal in snow and atmosphere of East Antarctica and implications for reactive nitrogen cycling. Atmos. Chem. Phys, 9, 8681–8696. https://doi.org/10.5194/acp-9-8681-2009, 2009.

Fritsch, F. N., & Butland, J.: A Method for Constructing Local Monotone Piecewise Cubic Interpolants. SIAM J. Sci. Comput., 5(2), 300–304. https://doi.org/10.1137/0905021, 1984.

Gao, X. and Li, W.: A graph-based LSTM model for PM2.5 forecasting, Atmos. Pollut. Res., 12, https://doi.org/10.1016/j.apr.2021.101150, 2021.

General Bathymetric Chart of the Oceans (GEBCO) Compilation Group: GEBCO 2019 Grid, 2019.

Geiss, A., Ma, P. L., Singh, B., and Hardin, J. C.: Emulating aerosol optics with randomly generated neural networks, Geosci. Model Dev., 16, 2355–2370, https://doi.org/10.5194/gmd-16-2355-2023, 2023.

Giglio, L., Randerson, J. T., and Van Der Werf, G. R.: Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4), J. Geophys. Res. Biogeosci., 118, 317–328, https://doi.org/10.1002/JGRG.20042, 2013.

Gramlich, Y., Siegel, K., Haslett, S. L., Cremer, R. S., Lunder, C., Kommula, S. M., Buchholz, A., Yttri, K. E., Chen, G., Krejci, R., Zieger, P., Virtanen, A., Riipinen, I., and Mohr, C.: Impact of Biomass Burning on Arctic Aerosol Composition, ACS Earth Space Chem., 8, 920–936, DOI: 10.1021/acsearthspacechem.3c00187, 2024.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Gryspeerdt, E., Povey, A. C., Grainger, R. G., Hasekamp, O., Christina Hsu, N., Mulcahy, J. P., Sayer, A. M., and Sorooshian, A.: Uncertainty in aerosol-cloud radiative forcing is driven by clean conditions, Atmos. Chem. Phys., 23, 4115–4122, https://doi.org/10.5194/ACP-23-4115-2023, 2023.

Hamilton, D. S., Lee, L. A., Pringle, K. J., Reddington, C. L., Spracklen, D. V., and Carslaw, K. S.: Occurrence of pristine aerosol environments on a polluted planet, Proc. Natl. Acad. Sci. U.S.A., 111, 18466–18471, https://doi.org/10.1073/pnas.1415440111, 2014.

Hansen, G., Aspmo, K., Berg, T., Edvardsen, K., Fiebig, M., Kallenborn, R., Krognes, T., Lunder, C., Stebel, K., Schmidbauer, N., Solberg, S., and Yttri, K. E.: Atmospheric monitoring at the Norwegian Antarctic station Troll: measurement programme and first results, Polar Res., 28, 353–363, https://doi.org/10.3402/POLAR.V28I3.6142, 2009.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. R. Meteorolog. Soc, 146, 1999–2049, https://doi.org/10.1002/QJ.3803, 2020.

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., Seibert, J. J., Vu, L., Andres, R. J., Bolt, R. M., Bond, T. C., Dawidowski, L., Kholod, N., Kurokawa, J. I., Li, M., Liu, L., Lu, Z., Moura, M. C. P., O'Rourke, P. R., and Zhang, Q.: Historical (1750-2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS), Geosci. Model Dev., 11, 369–408, https://doi.org/10.5194/GMD-11-369-2018, 2018.

Hsu, Y. K., Holsen, T. M., and Hopke, P. K.: Comparison of hybrid receptor models to locate PCB sources in Chicago, Atmos. Environ., 37, 545–562, https://doi.org/10.1016/S1352-2310(02)00886-5, 2003.

Isokääntä, S., Kim, P., Mikkonen, S., Kühn, T., Kokkola, H., Yli-Juuti, T., Heikkinen, L., Luoma, K., Petäjä, T., Kipling, Z., Partridge, D., and Virtanen, A.: The effect of clouds and precipitation on the aerosol concentrations and composition in a boreal forest environment, Atmos. Chem. Phys., 22, 11823–11843, https://doi.org/10.5194/ACP-22-11823-2022, 2022.

Ito, T.: Size distribution of Antarctic submicron aerosols, Tellus B, 45, 145–159, https://doi.org/10.1034/J.1600-0889.1993.T01-1-00007.X, 1993.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

James, I. N.: The influence of Antarctica on the flow in the Southern Hemisphere troposphere, In: Heywood, R.B., (ed.) University research in Antarctica. Proceedings of British Antarctic Survey Antarctic Special Topic Scheme Symposium, Cambridge, UK, 9-10 November 1988, 23-34, 1989.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., Bergamaschi, P., Pagliari, V., Olivier, J. G. J., Peters, J. A. H. W., Van Aardenne, J. A., Monni, S., Doering, U., Roxana Petrescu, A. M., Solazzo, E., and Oreggioni, G. D.: EDGAR v4.3.2 Global Atlas of the three major greenhouse gas emissions for the period 1970-2012, Earth Syst. Sci. Data, 11, 959–1002, https://doi.org/10.5194/ESSD-11-959-2019, 2019.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., et al.: EDGAR v4.3.2 Global Atlas of the three major greenhouse gas emissions for the period 1970-2012. Earth System Science Data, 11(3), 959–1002. https://doi.org/10.5194/ESSD-11-959-2019, 2019.

Jaxa-Rozen, M. and Kwakkel, J.: Tree-based ensemble methods for sensitivity analysis of environmental models: A performance comparison with Sobol and Morris techniques, Environ. Modell. Software, 107, 245–266, https://doi.org/10.1016/J.ENVSOFT.2018.06.011, 2018.

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Zhang, X., and Sachdeva, S.: Evaluation of Different Machine Learning Approaches to Forecasting PM 2.5 Mass Concentrations, Aerosol Air Qual. Res., 19, 1400–1410, https://doi.org/10.4209/aaqr.2018.12.0450, 2019.

Karlsson, L., Baccarini, A., Duplessis, P., Baumgardner, D., Brooks, I. M., Chang, R. Y. W., Dada, L., Dällenbach, K. R., Heikkinen, L., Krejci, R., Leaitch, W. R., Leck, C., Partridge, D. G., Salter, M. E., Wernli, H., Wheeler, M. J., Schmale, J., and Zieger, P.: Physical and Chemical Properties of Cloud Droplet Residuals and Aerosol Particles During the Arctic Ocean 2018 Expedition, J. Geophys. Res.: Atmos., 127, e2021JD036383, https://doi.org/10.1029/2021JD036383, 2022.

Kerminen, V.-M., Kerminen, V.-M., Paramonov, M., Anttila, T., Riipinen, I., Fountoukis, C., Korhonen, H., Asmi, E., Laakso, L., Lihavainen, H., Swietlicki, E., Svenningsson, B., Asmi, A., Pandis, S. N., Kulmala, M., and Petäjä, T.: Cloud condensation nuclei production associated with atmospheric nucleation: a synthesis based on existing literature and new results, Atmos. Chem. Phys., 12, 12037–12059, https://doi.org/10.5194/acp-12-12037-2012, 2012.

Khadir, T., Riipinen, I., Talvinen, S., Heslin-Rees, D., Pöhlker, C., Rizzo, L., Machado, L. A. T., Franco, M. A., Kremper, L. A., Artaxo, P., Petäjä, T., Kulmala, M., Tunved, P., Ekman, A. M. L., Krejci, R., and Virtanen, A.: Sink, Source or Something

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

In-Between? Net Effects of Precipitation on Aerosol Particle Populations, Geophys. Res. Lett., 50, https://doi.org/10.1029/2023GL104325, 2023.

Kim, J., Jun Yoon, Y., Gim, Y., Hee Choi, J., Jin Kang, H., Park, K. T., Park, J., and Yong Lee, B.: New particle formation events observed at King Sejong Station, Antarctic Peninsula - Part 1: Physical characteristics and contribution to cloud condensation nuclei, Atmos. Chem. Phys., 19, 7583–7594, https://doi.org/10.5194/ACP-19-7583-2019, 2019.

Kim, P., Partridge, D., and Haywood, J.: Constraining the model representation of the aerosol life cycle in relation to sources and sinks., EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-21948, https://doi.org/10.5194/EGUSPHERE-EGU2020-21948, 2020.

Kulkarni, P., Sreekanth, V., Upadhya, A. R., and Gautam, H. C.: Which model to choose? Performance comparison of statistical and machine learning models in predicting PM2.5 from high-resolution satellite aerosol optical depth, Atmos. Environ., 282, 119164, https://doi.org/10.1016/J.ATMOSENV.2022.119164, 2022.

Kumar, P., Vogel, H., Bruckert, J., Muth, L. J., Gholam, &, and Hoshyaripour, A.: MieAI: a neural network for calculating optical properties of internally mixed aerosol in atmospheric models, npj Clim. Atmos. Sci. 2024 7:1, 7, 1–13, https://doi.org/10.1038/s41612-024-00652-y, 2024.

Lachlan-Cope, T., C. S. Beddows, D., Brough, N., E. Jones, A., M. Harrison, R., Lupi, A., Jun Yoon, Y., Virkkula, A., and Dallosto, M.: On the annual variability of Antarctic aerosol size distributions at Halley Research Station, Atmos. Chem. Phys., 20, 4461–4476, https://doi.org/10.5194/ACP-20-4461-2020, 2020.

Laj, P., Myhre, C. L., Riffault, V., Amiridis, V., Fuchs, H., Eleftheriadis, K., Petäjä, T., Salameh, T., Kivekäs, N., Juurola, E., Saponaro, G., Philippin, S., Cornacchia, C., Arboledas, L. A., Baars, H., Claude, A., De Mazière, M., Dils, B., Dufresne, M., Evangeliou, N., Favez, O., Fiebig, M., Haeffelin, M., Herrmann, H., Höhler, K., Illmann, N., Kreuter, A., Ludewig, E., Marinou, E., Möhler, O., Mona, L., Murberg, L. E., Nicolae, D., Novelli, A., O'Connor, E., Ohneiser, K., Altieri, R. M. P., Picquet-Varrault, B., van Pinxteren, D., Pospichal, B., Putaud, J. P., Reimann, S., Siomos, N., Stachlewska, I., Tillmann, R., Voudouri, K. A., Wandinger, U., Wiedensohler, A., Apituley, A., Comerón, A., Gysel-Beer, M., Mihalopoulos, N., Nikolova, N., Pietruczuk, A., Sauvage, S., Sciare, J., Skov, H., Svendby, T., Swietlicki, E., Tonev, D., Vaughan, G., Zdimal, V., Baltensperger, U., Doussin, J. F., Kulmala, M., Pappalardo, G., Sundet, S. S., and Vana, M.: Aerosol, Clouds and Trace Gases Research Infrastructure (ACTRIS): The European Research Infrastructure Supporting Atmospheric Science, Bull. Am. Meteorol. Soc., 105, E1098–E1136, https://doi.org/10.1175/BAMS-D-23-0064.1, 2024.

Liao, L., Kerminen, V. M., Boy, M., Kulmala, M., and Dal Maso, M.: Temperature influence on the natural aerosol budget over boreal forests, Atmos. Chem. Phys., 14, 8295–8308, https://doi.org/10.5194/acp-14-8295-2014, 2014.

Lund Myhre, C., Svendby, T., Hermansen, O., Lunder, C., Platt, S. M., Fiebig, M., Fjaeraa, A. M., Hansen, G., Schmidbauer, N., and Krognes, T.: Monitoring of greenhouse gases and aerosols at Svalbard and Birkenes in 2019, Annual report, NILU, 2020.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4-9 December 2017, 4768–4777, 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, arXiv preprint arXiv:1802.03888, 2018.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. I.: From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence 2020 2:1, 2, 56–67, https://doi.org/10.1038/s42256-019-0138-9, 2020.

Mann, G. W., Carslaw, K. S., Spracklen, D. V, Ridley, D. A., Manktelow, P. T., Chipperfield, M. P., Pickering, S. J., and Johnson, C. E.: Geoscientific Model Development Description and evaluation of GLOMAP-mode: a modal global aerosol microphysics model for the UKCA composition-climate model, Geosci. Model Dev., 3, 519–551, https://doi.org/10.5194/gmd-3-519-2010, 2010.

Maritorena, S., d'Andon, O. H. F., Mangin, A., and Siegel, D. A.: Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues, Remote Sens. Environ., 114, 1791–1804, https://doi.org/10.1016/J.RSE.2010.04.002, 2010.

McCoy, I. L., McCoy, D. T., Wood, R., Regayre, L., Watson-Parris, D., Grosvenor, D. P., Mulcahy, J. P., Hu, Y., Bender, F. A. M., Field, P. R., Carslaw, K. S., and Gordon, H.: The hemispheric contrast in cloud microphysical properties constrains aerosol forcing, Proc. Natl. Acad. Sci. U.S.A., 117, 18998–19006, https://doi.org/10.1073/PNAS.1922502117, 2020.

McCoy, I. L., Bretherton, C. S., Wood, R., Twohy, C. H., Gettelman, A., Bardeen, C. G., and Toohey, D. W.: Influences of Recent Particle Formation on Southern Ocean Aerosol Variability and Low Cloud Properties, J. Geophys. Res.: Atmos., 126, https://doi.org/10.1029/2020JD033529, 2021.

Mu, M., Randerson, J. T., Van Der Werf, G. R., Giglio, L., Kasibhatla, P., Morton, D., Collatz, G. J., Defries, R. S., Hyer, E. J., Prins, E. M., Griffith, D. W. T., Wunch, D., Toon, G. C., Sherlock, V., and Wennberg, P. O.: Daily and 3-hourly variability in global fire emissions and consequences for atmospheric model predictions of carbon monoxide, J. Geophys. Res.: Atmos., 116, 24303, https://doi.org/10.1029/2011JD016245, 2011.

Mulcahy, J., Johnson, C., Jones, C., Povey, A., Scott, C., Sellar, A., Turnock, S., Woodhouse, M., Abraham, N. L., Andrews, M., Bellouin, N., Browse, J., Carslaw, K., Dalvi, M., Folberth, G., Glover, M., Grosvenor, D., Hardacre, C., Hill, R., Johnson, B., Jones, A., Kipling, Z., Mann, G., Mollard, J., O'Connor, F., Palmieri, J., Reddington, C., Rumbold, S., Richardson, M., Schutgens, N., Stier, P., Stringer, M., Tang, Y., Walton, J., Woodward, S., and Yool, A.: Description and evaluation of aerosol in UKESM1 and HadGEM3-GC3.1 CMIP6 historical simulations, Geosci. Model Dev. 1–59, https://doi.org/10.5194/gmd-2019-357, 2020.

O'Dowd, C., Ceburnis, D., Ovadnevaite, J., Bialek, J., Stengel, D. B., Zacharias, M., Nitschke, U., Connan, S., Rinaldi, M., Fuzzi, S., Decesari, S., Cristina Facchini, M., Marullo, S., Santoleri, R., Dell'anno, A., Corinaldesi, C., Tangherlini, M., and Danovaro, R.: Connecting marine productivity to sea-spray via nanoscale biological processes: Phytoplankton Dance or Death Disco?, Sci. Rep. 2015 5:1, 5, 1–11, https://doi.org/10.1038/srep14883, 2015.

O'Reilly, J. E., Maritorena, S., O'Brien, M., Siegel, D., Toole, D., Menzies, D., Smith, R., Mueller, J., Mitchell, G., Kahru, M., Chavez, F., Strutton, P., Cota, G., Hooker, S., McClain, C., Carder, K., Muller-Karger, F., Harding, L., Magnuson, A., Phinney, D., Moore, G., Aiken, J., Arrigo, K., Letelier, R., and Culver, M.: Ocean color chlorophyll a algorithms for SeaWiFS, OC2, and OC4: Version 4, in: SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3, NASA Tech. Memo., vol. 11, edited by: Hooker, S. B. and Firestone, E. R., NASA Goddard Space Flight Center, Greenbelt, 9–23, 2000.

Osborne, J. W.: Improving your data transformations: Applying the Box-Cox transformation, Practical Assessment, Research, and Evaluation, 15(1): 12, https://doi.org/10.7275/QBPC-GK17, 2010.

Paasonen, P., Asmi, A., Petäjä, T., Kajos, M. K., Äijälä, M., Junninen, H., Holst, T., Abbatt, J. P. D., Arneth, A., Birmili, W., Van Der Gon, H. D., Hamed, A., Hoffer, A., Laakso, L., Laaksonen, A., Richard Leaitch, W., Plass-Dülmer, C., Pryor, S. C., Räisänen, P., Swietlicki, E., Wiedensohler, A., Worsnop, D. R., Kerminen, V. M., and Kulmala, M.: Warming-induced increase in aerosol number concentration likely to moderate climate change, Nat. Geosci. 2013 6:6, 6, 438–442, https://doi.org/10.1038/ngeo1800, 2013.

Paglione, M., Beddows, D. C. S., Jones, A., Lachlan-Cope, T., Rinaldi, M., Decesari, S., Manarini, F., Russo, M., Mansour, K., Harrison, R. M., Mazzanti, A., Tagliavini, E., and Dall'Osto, M.: Simultaneous organic aerosol source apportionment at

two Antarctic sites reveals large-scale and ecoregion-specific components, Atmos. Chem. Phys., 24, 6305–6322, https://doi.org/10.5194/ACP-24-6305-2024, 2024.

Pernov, J. B., Beddows, D., Thomas, D. C., Dall´Osto, M., Harrison, R. M., Schmale, J., Skov, H., and Massling, A.: Increased aerosol concentrations in the High Arctic attributable to changing atmospheric transport patterns, npj Clim. Atmos. Sci. 2022 5:1, 5, 1–13, https://doi.org/10.1038/s41612-022-00286-y, 2022.

Pernov, J. B., Harris, E., Volpi, M., Baumgartner, T., Hohermuth, B., Empa, S. H., Aeberhard, W. H., Becagli, S., Quinn, P. K., Traversi, R., Upchurch, L. M., and Schmale, J.: Pan-Arctic Methanesulfonic Acid Aerosol: Source regions, atmospheric drivers, and future projections, npj Clim. Atmos. Sci. 7, 166 https://doi.org/10.21203/RS.3.RS-3976619/V1, 2024.

Petäjä, T., Tabakova, K., Manninen, A., Ezhova, E., O'Connor, E., Moisseev, D., Sinclair, V. A., Backman, J., Levula, J., Luoma, K., Virkkula, A., Paramonov, M., Räty, M., Äijälä, M., Heikkinen, L., Ehn, M., Sipilä, M., Yli-Juuti, T., Virtanen, A., Ritsche, M., Hickmon, N., Pulik, G., Rosenfeld, D., Worsnop, D. R., Bäck, J., Kulmala, M., and Kerminen, V. M.: Influence of biogenic emissions from boreal forests on aerosol–cloud interactions, Nat. Geosci., 15, 42–47, https://doi.org/10.1038/S41561-021-00876-0, 2022.

Qin, D., Yu, J., Zou, G., Yong, R., Zhao, Q., and Zhang, B.: A Novel Combined Prediction Scheme Based on CNN and LSTM for Urban PM2.5 Concentration, IEEE Access, 7, 20050–20059, https://doi.org/10.1109/ACCESS.2019.2897028, 2019.

Qiu, Y., Feng, J., Zhang, Z., Zhao, X., Li, Z., Ma, Z., Liu, R., and Zhu, J.: Regional aerosol forecasts based on deep learning and numerical weather prediction, npj Clim. Atmos. Sci. 2023 6:1, 6, 1–12, https://doi.org/10.1038/s41612-023-00397-0, 2023.

Raes, F., Van Dingenen, R., Vignati, E., Wilson, J., Putaud, J. P., Seinfeld, J. H., and Adams, P.: Formation and cycling of aerosols in the global troposphere, Atmos. Environ., 34, 4215–4240, https://doi.org/10.1016/S1352-2310(00)00239-9, 2000.

Räty, M., Sogacheva, L., Keskinen, H. M., Kerminen, V. M., Nieminen, T., Petäjä, T., Ezhova, E., and Kulmala, M.: Dynamics of aerosol, humidity, and clouds in air masses travelling over Fennoscandian boreal forests, Atmos. Chem. Phys., 23, 3779–3798, https://doi.org/10.5194/ACP-23-3779-2023, 2023.

Regayre, L. A., Schmale, J., Johnson, J. S., Tatzelt, C., Baccarini, A., Henning, S., Yoshioka, M., Stratmann, F., Gysel-Beer, M., Grosvenor, D. P., and Carslaw, K. S.: The value of remote marine aerosol measurements for constraining radiative forcing uncertainty, Atmos. Chem. Phys., 20, 10063–10072, https://doi.org/10.5194/ACP-20-10063-2020, 2020.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Revell, L. E., Kremser, S., Hartery, S., Harvey, M., Mulcahy, J. P., Williams, J., Morgenstern, O., Mcdonald, A. J., Varma, V., Bird, L., and Schuddeboom, A.: The sensitivity of Southern Ocean aerosols and cloud microphysics to sea spray and sulfate aerosol production in the HadGEM3-GA7.1 chemistry-climate model, Atmos. Chem. Phys., 19, 15447–15466, https://doi.org/10.5194/ACP-19-15447-2019, 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C.: "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, August 13-17, 2016, 1135–1144, https://doi.org/10.1145/2939672.2939778, 2016.

Riddick, S. N., Dragosits, U., Blackall, T. D., Daunt, F., Wanless, S., & Sutton, M. A.: The global distribution of ammonia emissions from seabird colonies. Atmos. Environ., 55, 319–327. https://doi.org/10.1016/J.ATMOSENV.2012.02.052, 2012.

Rinaldi, M., Fuzzi, S., Decesari, S., Marullo, S., Santoleri, R., Provenzale, A., Von Hardenberg, J., Ceburnis, D., Vaishya, A., O'Dowd, C. D., and Facchini, M. C.: Is chlorophyll-a the best surrogate for organic matter enrichment in submicron primary marine aerosol?, J. Geophys. Res.: Atmos., 118, 4964–4973, https://doi.org/10.1002/JGRD.50417, 2013.

Rose, C., Collaud Coen, M., Andrews, E., Lin, Y., Bossert, I., Lund Myhre, C., Tuch, T., Wiedensohler, A., Fiebig, M., Aalto, P., Alastuey, A., Alonso-Blanco, E., Andrade, M., Artinano, B., Arsov, T., Baltensperger, U., Bastian, S., Bath, O., Beukes, J. P., Brem, B. T., Bukowiecki, N., Casquero-Vera, J. A., Conil, S., Eleftheriadis, K., Favez, O., Flentje, H., Gini, M. I., Gomez-Moreno, F. J., Gysel-Beer, M., Hallar, A. G., Kalapov, I., Kalivitis, N., Kasper-Giebl, A., Keywood, M., Kim, J. E., Kim, S. W., Kristensson, A., Kulmala, M., Lihavainen, H., Lin, N. H., Lyamani, H., Marinoni, A., Martins Dos Santos, S., Mayol-Bracero, O. L., Meinhardt, F., Merkel, M., Metzger, J. M., Mihalopoulos, N., Ondracek, J., Pandolfi, M., Perez, N., Petaja, T., Petit, J. E., Picard, D., Pichon, J. M., Pont, V., Putaud, J. P., Reisen, F., Sellegri, K., Sharma, S., Schauer, G., Sheridan, P., Sherman, J. P., Schwerin, A., Sohmer, R., Sorribas, M., Sun, J., Tulet, P., Vakkari, V., Van Zyl, P. G., Velarde, F., Villani, P., Vratolis, S., Wagner, Z., Wang, S. H., Weinhold, K., Weller, R., Yela, M., Zdimal, V., and Laj, P.: Seasonality of the particle number concentration and size distribution: A global analysis retrieved from the network of Global Atmosphere Watch (GAW) near-surface observatories, Atmos. Chem. Phys., 21, 17185–17223, https://doi.org/10.5194/ACP-21-17185-2021, 2021.

Sanchez, K. J., Zhang, B., Liu, H., Saliba, G., Chen, C. L., Lewis, S. L., Russell, L. M., Shook, M. A., Crosbie, E. C., Ziemba, L. D., Brown, M. D., Shingler, T. J., Robinson, C. E., Wiggins, E. B., Thornhill, K. L., Winstead, E. L., Jordan, C., Quinn, P. K., Bates, T. S., Porter, J., Bell, T. G., Saltzman, E. S., Behrenfeld, M. J., and Moore, R. H.: Linking marine phytoplankton emissions, meteorological processes, and downwind particle properties with FLEXPART, Atmos. Chem. Phys., 21, 831–851, https://doi.org/10.5194/ACP-21-831-2021, 2021.

Savarino, J., Kaiser, J., Morin, S., Sigman, D. M., & Thiemens, M. H.: Nitrogen and oxygen isotopic constraints on the origin of atmospheric nitrate in coastal Antarctica. Atmos. Chem. Phys., 7, 1925–1945. https://doi.org/https://doi.org/10.5194/acp-7-1925-2007, 2007.

Schmale, J., Baccarini, A., Thurnherr, I., Henning, S., Efraim, A., Regayre, L., Bolas, C., Hartmann, M., Welti, A., Lehtipalo, K., Aemisegger, F., Tatzelt, C., Landwehr, S., Modini, R. L., Tummon, F., Johnson, J. S., Harris, N., Schnaiter, M., Toffoli, A., Derkani, M., Bukowiecki, N., Stratmann, F., Dommen, J., Sperger, U. B., Wernli, H., Rosenfeld, D., Gysel-Beer, M., and Carslaw, K. S.: Overview of the Antarctic Circumnavigation Expedition: Study of Preindustrial-like Aerosols and Their Climate Effects (ACE-SPACE), Bull. Am. Meteorol. Soc., 100, 2260–2283, https://doi.org/10.1175/BAMS-D-18-0187.1, 2019.

Schmale, J., Sharma, S., Decesari, S., Pernov, J., Massling, A., Hansson, H. C., Von Salzen, K., Skov, H., Andrews, E., Quinn, P. K., Upchurch, L. M., Eleftheriadis, K., Traversi, R., Gilardoni, S., Mazzola, M., Laing, J., and Hopke, P.: Pan-Arctic seasonal cycles and long-term trends of aerosol properties from 10 observatories, Atmos. Chem. Phys., 22, 3067–3096, https://doi.org/10.5194/ACP-22-3067-2022, 2022.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics: From air pollution to climate change, First edition, Whiley, New York, 1998.

Sellar, A. A., Walton, J., Jones, C. G., Wood, R., Abraham, N. L., Andrejczuk, M., et al.: Implementation of U.K. Earth System Models for CMIP6. J. Adv. Model. Earth Syst., 12(4), https://doi.org/10.1029/2019MS001946

Shaw, G. E.: Considerations on the origin and properties of the Antarctic aerosol, Rev. Geophys., 17, 1983–1998, https://doi.org/10.1029/RG017I008P01983, 1979.

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., von der Heydt, A. S., Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein, M., Schmidt, G. A., Tokarska, K. B., and Zelinka, M. D.: An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence, Rev. Geophys., 58, https://doi.org/10.1029/2019RG000678, 2020.

Sogacheva, L., Dal Maso, M., Kerminen, V.-M., and Kulmala, M.: Probability of nucleation events and aerosol particle concentration in different air mass types arriving at Hyytiälä, southern Finland, based on back trajectories analysis., Boreal Environ. Res., 479–491, 2005.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Song, C., Becagli, S., Beddows, D. C. S., Brean, J., Browse, J., Dai, Q., Dall'Osto, M., Ferracci, V., Harrison, R. M., Harris, N., Li, W., Jones, A. E., Kirchgäßner, A., Kramawijaya, A. G., Kurganskiy, A., Lupi, A., Mazzola, M., Severi, M., Traversi, R., and Shi, Z.: Understanding Sources and Drivers of Size-Resolved Aerosol in the High Arctic Islands of Svalbard Using a Receptor Model Coupled with Machine Learning, Environ. Sci. Technol, 56, 11189–11198, https://doi.org/10.1021/acs.est.1c07796, 2022.

Spracklen, D. V, Bonn, B., and Carslaw, K. S.: Boreal forests, aerosols and the impacts on clouds and climate, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., 366, 4613–4626, https://doi.org/10.1098/rsta.2008.0201, 2008.

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D., and Ngan, F.: Noaa's Hysplit Atmospheric 1460 Transport and Dispersion Modeling System, Bull. Am. Meteorol. Soc., 96, 2059–2077, https://doi.org/10.1175/BAMS-D-14-00110.1, 2015.

Stohl, A.: Computation, accuracy and applications of trajectories—A review and bibliography, Atmos. Environ., 32, 947–966, https://doi.org/10.1016/S1352-2310(97)00457-3, 1998.

Struthers, H., Ekman, A. M. L., Glantz, P., Iversen, T., Kirkevåg, A., Mårtensson, E. M., Seland, Ø., and Nilsson, E. D.: The effect of sea ice loss on sea salt aerosol concentrations and the radiative balance in the Arctic, Atmos. Chem. Phys, 11, 3459–3477, https://doi.org/10.5194/acp-11-3459-2011, 2011.

Talvinen, S., Kim, P., Tovazzi, E., Holopainen, E., Cremer, R., Kühn, T., Kokkola, H., Kipling, Z., Neubauer, D., Teixeira, J. C., Sellar, A., Watson-Parris, D., Yang, Y., Zhu, J., Krishnan, S., Virtanen, A., and Partridge, D. G.: Towards an improved understanding of the impact of clouds and precipitation on the representation of aerosols over the Boreal Forest in GCMs, [preprint] https://doi.org/10.5194/EGUSPHERE-2025-721, 2025.

Telford, P. J., Braesicke, P., Morgenstern, O., and Pyle, J. A.: Atmospheric Chemistry and Physics Technical Note: Description and assessment of a nudged version of the new dynamics Unified Model, Atmos. Chem. Phys, 8, 1701–1712, 2008.

Tunved, P., Hansson, H. C., Kerminen, V. M., Ström, J., Dal Maso, M., Lihavainen, H., Viisanen, Y., Aalto, P. P., Komppula, M., and Kulmala, M.: High natural aerosol loading over boreal forests, Science (1979), 312, 261–263, https://doi.org/10.1126/science.1123052, 2006.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Tunved, P., Ström, J., and Krejci, R.: Arctic aerosol life cycle: Linking aerosol size distributions observed between 2000 and 2010 with air mass transport and precipitation at Zeppelin station, Ny-Ålesund, Svalbard, Atmos. Chem. Phys., 13, 3643–3660, https://doi.org/10.5194/ACP-13-3643-2013, 2013.

Ueda, S., Miura, K., Kawata, R., Furutani, H., Uematsu, M., Omori, Y., and Tanimoto, H.: Number–size distribution of aerosol particles and new particle formation events in tropical and subtropical Pacific Oceans, Atmos. Environ., 142, 324–339, https://doi.org/10.1016/J.ATMOSENV.2016.07.055, 2016.

Venugopal, A. U., Bhatti, Y. A., Morgenstern, O., Williams, J., Edkins, N., Hardacre, C., Jones, A., and Revell, L. E.: Constraining the Uncertainty Associated With Sea Salt Aerosol Parameterizations in Global Models Using Nudged UKESM1-AMIP Simulations, J. Geophys. Res.: Atmos., 130, e2024JD041643, https://doi.org/10.1029/2024JD041643, 2025.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. Pietro, Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G. L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., et al.: SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods 2020 17:3, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Wagenbach, D., Legrand, M., Fischer, H., Pichlmayer, F., & Wolff, E. W.: Atmospheric near-surface nitrate at coastal Antarctic sites. J. Geo. Res.: Atmos., 103(3339), 11007–11020. https://doi.org/10.1029/97JD03364, 1998.

Warneke, C., Froyd, K. D., Brioude, J., Bahreini, R., Brock, C. A., Cozic, J., De Gouw, J. A., Fahey, D. W., Ferrare, R., Holloway, J. S., Middlebrook, A. M., Miller, L., Montzka, S., Schwarz, J. P., Sodemann, H., Spackman, J. R., and Stohl, A.: An important contribution to springtime Arctic aerosol from biomass burning in Russia, Geophys. Res. Lett., 37, 1801, https://doi.org/10.1029/2009GL041816, 2010.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Watson-Parris, D., Sutherland, S., Christensen, M., Caterini, A., Sejdinovic, D., and Stier, P.: Detecting anthropogenic cloud perturbations with deep learning, Climate Change: How Can AI Help? Workshop at 2019 International Conference on Machine Learning, Long Beach, CA, USA, 14 September 2019, https://doi.org/10.48550/arXiv.1911.13061, 2019.

Weller, R., Minikin, A., Wagenbach, D., and Dreiling, V.: Characterization of the inter-annual, seasonal, and diurnal variations of condensation particle concentrations at Neumayer, Antarctica, Atmos Chem Phys, 11, 13243–13257, https://doi.org/10.5194/ACP-11-13243-2011, 2011.

Van Der Werf, G. R., Randerson, J. T., Giglio, L., Van Leeuwen, T. T., Chen, Y., Rogers, B. M., Mu, M., Van Marle, M. J. E., Morton, D. C., Collatz, G. J., Yokelson, R. J., and Kasibhatla, P. S.: Global fire emissions estimates during 1997-2016, Earth Syst Sci Data, 9, 697–720, https://doi.org/10.5194/ESSD-9-697-2017, 2017.

Xiao, F., Yang, M., Fan, H., Fan, G., & Al-qaness, M. A. A.: An improved deep learning model for predicting daily PM2.5 concentration. Sci. Rep., 10(1), 1–11. https://doi.org/10.1038/s41598-020-77757-w, 2020.

Yan, J., Jung, J., Lin, Q., Zhang, M., Xu, S., & Zhao, S.: Effect of sea ice retreat on marine aerosol emissions in the Southern Ocean, Antarctica. Sci. Total Environ., 745, 140773. https://doi.org/10.1016/J.SCITOTENV.2020.140773, 2020.

Yang, X., Frey, M. M., Rhodes, R. H., Norris, S. J., Brooks, I. M., Anderson, P. S., Nishimura, K., Jones, A. E., and Wolff, E. W.: Sea salt aerosol production via sublimating wind-blown saline snow particles over sea ice: Parameterizations and relevant microphysical mechanisms, Atmos Chem Phys, 19, 8407–8424, https://doi.org/10.5194/ACP-19-8407-2019, 2019.

Yazdi, M. D., Kuang, Z., Dimakopoulou, K., Barratt, B., Suel, E., Amini, H., et al.: Predicting Fine Particulate Matter (PM2.5) in the Greater London Area: An Ensemble Approach using Machine Learning Methods. Remote Sensing 2020, 12(6), 914. https://doi.org/10.3390/RS12060914, 2020.

Yoon, Y. J. and Brimblecombe, P.: Modelling the contribution of sea salt and dimethyl sulfide derived aerosol to marine CCN, Atmos Chem Phys, 2, 17–30, https://doi.org/10.5194/ACP-2-17-2002, 2002.

Zhao, J., Deng, F., Cai, Y., and Chen, J.: Long short-term memory - Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction, Chemosphere, 220, 486–492, https://doi.org/10.1016/j.chemosphere.2018.12.128, 2019.

https://doi.org/10.5194/egusphere-2025-4298
Preprint. Discussion started: 6 November 2025

Zheng, G., Wang, Y., Wood, R., Jensen, M. P., Kuang, C., McCoy, I. L., Matthews, A., Mei, F., Tomlinson, J. M., Shilling, J. E., Zawadowicz, M. A., Crosbie, E., Moore, R., Ziemba, L., Andreae, M. O., and Wang, J.: New particle formation in the remote marine boundary layer, Nature Communications 2021 12:1, 12, 1–10, https://doi.org/10.1038/s41467-020-20773-1, 2021.