Author Comment for:

"The value of remote marine aerosol measurements for constraining radiative forcing uncertainty"

We thank the reviewers for their thoughtful comments on our paper. We have adapted our article in response to many of the helpful comments and suggestions. In particular, we have added detail to the description of our method, including additional SI tables and figures. We have also included an additional table in the main article to show the process-based nature of our constraint on aerosol forcing. Many of our assumptions and their implications are now described in more detail, as suggested.

Response to anonymous reviewer 1:

Line 79. Is this RF or is it ERF?

We make use of 2 PPEs in this study, one which outputs aerosol RF (the AER PPE) and one which outputs aerosol ERF (AER-ATM), as described in the 3rd paragraph of the introduction (line 56 onwards).

These PPEs were designed to complement one another and are described in full in Yoshioka et al. (2019). The AER-ATM PPE used in Regayre et al. (2018) to diagnose ERF includes rapid atmospheric adjustments and perturbations to multiple physical atmosphere parameters, alongside perturbations to aerosol parameters, with horizontal wind fields nudged only above the boundary layer. However, the AER PPE used in Johnson et al. (2019) is nudged throughout the atmosphere to suppress meteorological effects entirely. Hence, no rapid adjustments were included in this PPE meaning that only aerosol RF can be analysed and constrained in this case. In this article we aim to complement and extend the constraint using output from the AER PPE from Johnson, et al. (2019) with remote marine aerosol measurements, so we mainly refer to aerosol RF. We refer to ERF when referring to corresponding results from the AER-ATM PPE.

For additional clarity, we have added "(neglecting rapid adjustments)" to the text describing the Johnson et al. (2019) constraint on around line 75 and more fully describe the efficacy of previous constraints on components of aerosol forcing:

"Previous analysis of HadGEM3 PPEs showed that measurements of the present-day atmosphere in regions affected by anthropogenic emissions help to constrain the uncertainty in aerosol-radiation interaction forcing (RF_{ari}) but not the component due to aerosol-cloud interactions (RF_{aci}). For example, Regayre et al. (2018) showed that top-of-the-atmosphere shortwave radiation flux measurements reduce ERF_{aci} uncertainty by only around 10%, despite the fluxes in the present-day and early-industrial environments sharing multiple causes of uncertainty. Johnson et al. (2019) showed that a much larger dataset of over 9000 (predominantly Northern Hemisphere) aerosol measurements reduced the uncertainty in global, annual mean aerosol RF_{ari} (neglecting rapid adjustments) by 35%, but RF_{aci} uncertainty by only around 7%."

Line 93 N700: Is it wet or dry size. Radius or diameter?

We use dry aerodynamic particle diameters from the Schmale et al. (2019) dataset.

We have clarified this for the reader by changing the text to "*with dry aerodynamic diameter (N_{700}; corresponds to volume equivalent diameter larger than around 500 to 570*

nm; Schmale et al., 2019a)" on line 93 of the main article, line 113 of the SI and in the caption of figure 1.

Line 109-1111 + Figure 1d. While the constrainment of the model parameterisation narrows the range of CCN concentrations and reduces the original model bias, the constrained values has a very low variability compared to the observations. While it is understandable that the combined product should may have lower deviation than both the model and measurements alone. Is it realistic that the constrained variability is so much lower compared to observations?

To clarify, Fig 1d is a map of the standard deviation of the monthly mean CCN concentration. It doesn't show model variability. We are aware that the use of point measurements to constrain monthly mean fields of CCN introduces temporal and spatial uncertainties, but we account for these uncertainties in our implausibility metric (through the Var(R) term in SI equation 1). Although we constrain monthly mean uncertainties in each model gridbox by more than half, the remaining uncertainties at the model gridbox scale are non-negligible and of the remaining uncertainty is the same order of magnitude as the gridbox means.

Figure 2. The N700 number is much lower than than CCN_0.2 Do you have any estimates for N(total sea-salt) to show that the constrained sea-salt emissions increase is indeed the cause of CCN_0.2 and not e.g. the increase in Nss-sulphate. Or the constraining of accumulation mode dry deposition.

We think the reviewer is referring to line 162 of original manuscript, which read: "*the consistency of constraint of $CCN_{0.2}$ and N_{700} towards higher values (Fig. 1) implies that a general scaling of the existing sea spray flux is consistent with the measurements without the need for an additional source of fine-mode, organic-rich particles*.". There are many ways to combined multiple uncertain processes and get approximately the same outcome. Our figure 2 (incorrectly referred to as Fig. 1 in the original manuscript) shows that $CCN_{0.2}$, N_{700} and nss-sulfate concentrations are all constrained to higher values. We relied on preliminary work on understanding the effects of constraining individual measurement types on model parameters to inform our analysis. However, we did not make this evidence available to the reader. We have added a table to the SI (table S3) which provides information about the effect of each measurement type constraint on model parameters. This table shows higher values of the sea spray emission flux scale factor parameter are consistent with $CCN_{0.2}$ measurements.

We introduce table S3 in the SI section "SI Results: Constrained marginal parameter distributions:

"In addition to the constraint achieved by combining remote marine aerosol measurements, table S3 shows the effect of individual measurement type constraints (table S2) on model parameters and how these translate into a combined constraint (Fig. 3)."

We now refer to the new table S3 in the following places.

The Fig. 3 caption, which now reads:

"Fig. 3. Marginal probability distributions for the 26 aerosol parameters after constraint using ACE-SPACE measurements. The density of parameter values in the unconstrained sample are shown as horizontal dashed lines (uniform sampling over

the parameter space). Densities of constrained samples are shown in colour and are scaled so that the maximum densities in the constrained and unconstrained samples are aligned. The 25th, 50th and 75th percentiles of each marginal distribution are shown in the central boxes. Parameter values on the x-axes correspond to values used in the model (Yoshioka et al., 2019, table S3)."

Around line 153 which now reads:

"These joint constraints (see also Fig. S3) suggest the model-measurement comparison is improved when aerosol number concentrations and mass are relatively high."

Around line 160 of the manuscript, which now reads:

"We do not make any assumptions about the composition of these additional summertime sea spray particles. They may be rich in organic material as proposed by Gantt et al. (2011) which would alter the CCN activity of emitted particles. However, the consistency of constraint of $CCN_{0.2}$ and N_{700} towards higher values (Fig. 2, table S3) implies that a general scaling of the existing sea spray flux is consistent with the measurements from December to April, without the need for an additional source of fine-mode, organic-rich particles."

and on line 275 which reads:

"Secondly, even within the considerably reduced volume of multi-dimensional parameter space there still exist many compensating parameter effects (Fig. S3), which limit the constraint on individual parameter ranges (Lee et al., 2016; Regayre et al., 2018). The impact of these compensating effects could be greatly reduced by perturbing uncertain emissions regionally rather than globally as we do here."

Figure 2: Does both model and measurements use the same definition of aerosol size, i.e. the same relative humidity? If the measurements is done at e.g. 80 % relative humidity and the model results use dry radius, the N700 from the model should be lower than the measurements

Yes. We use measured N_{700} concentrations of particles with dry aerodynamic diameters (40% relative humidity at the APS device air intake valve) larger than 700 nm. The volume equivalent diameter of these particles is around 500 to 570 nm. Aerosol concentrations are also calculated using dry particle diameters.

The N700 description from line 93 of the original manuscript now reads: "*number concentrations of particles with dry aerodynamic diameter larger than 700 nm (N700; corresponds to volume equivalent diameter larger than around 500 to 570 nm; Schmale et al., 2019a).*"

Line 216: Adding the NH experiment is reducing the number of constrained model versions to 0.7 % of the total. As this likely give an even more narrow range for the constrained estimate e.g. as in figure 1d. Any comments on the validity of this heavy constrainment given that it is based on a very limited amount measurements?

The constrained sample contains 700 model variants, which is far more than are typically used to quantify model uncertainty or multi-model diversity. We don't agree that the measurements are "very limited". We used over 9000 measurements from Johnson et al. (2019) combined with hundreds of measurements for four aerosol properties covering much of the Southern Ocean. The results is 'valid' in the sense that these are the model variants that are most consistent with this very large set of measurements.

We have contextualised this for the reader on line 216 which now reads:

"Around 700 model variants (0.07%) are observationally plausible in both the Southern Ocean (ACE-SPACE) and Johnson et al. (2019) constraints. Although this is a relatively small percentage of the original sample, 700 observationally-plausible model variants is far more than are typically used to quantify model uncertainty or multi-model diversity (e.g. around 30 for CMIP6)."

SI Line 106. Any estimates for the uncertainty caused by the sampling procedure?

The sentence in question is "*Fig. 1 shows the $CCN_{0.2}$ mean and standard deviation from the unconstrained and constrained model variants*". It's not clear to us what the referee means by "sampling procedure". Our statistical approach densely samples the model's uncertain parameter space. The constraint methodology accounts for multiple sources of uncertainty within the implausibility metric (equation S1; including using an emulator in place of the model). Therefore, the uncertainty in the posterior CCN distribution implicitly accounts for our sampling methodology.

We added a clarification to our emulation description on line 101 of the SI:
"*Some additional uncertainty is caused by emulating (rather than simulating) model output and this uncertainty is incorporated into our model-measurement constraint process (SI Methods: Model-measurement comparisons), despite being much smaller than other sources of uncertainty (Johnson et al., 2019).*"

SI: Wind speed discrepancies. I can not see that the assumption about wind speed discrepancy being unimportant is supported at all by Korhonen et al. On the contrary the main point of Korhonen et al is that even a quite modest increase in wind-speed creates a higher CCN concentration. As the wind speed in the ensembles is said to be lower than the values in ACE-SPACE and even much lower than the climatological values the unconstrained values, the unconstrained sea-salt emissions is expected to be lower than during the campaign and even lower compared with climatological values (potentially relevant if the "NH" added constrainment use retrieved values for AOD). Any deviations for the high wind speeds would be even more deleterious for the constrainment of sea-salt emissions.

The reviewer is correct that the wording of this section was misleading. The reference to Korhonen et al. (2010) has now been removed. We referred to a subtle result in Korhonen et al. (2010) which our description did not make clear. We originally cited this article to point out that there are many factors other than sea spray which affect remote marine cloud condensation nuclei concentrations (52% of the CCN variability according to their research).

Our approach compares in-situ measurements with monthly mean model data. In-situ measurements are inherently more variable because of differences in averaging period, and hence are more likely to include high wind speed events. Indeed, wind speed and N700 measurements from the ACE-SPACE campaign are only weakly correlated. Measured wind speed averages within regions defined by model gridboxes (as used in our comparison to monthly mean model output) are only weakly correlated with N700 measurements (Pearson correlation coefficient of around 0.2).

Our constraint methodology is designed to avoid relying on measurements that are in strong disagreement with the model, because these discrepancies are more likely to be caused by model structural errors. This approach also avoids the use of measurements where nudged meteorology causes large model-measurement discrepancies in variables used for constraint.

We have altered the associated section of the SI (line 208 onwards) to clarify the importance of understanding these wind speed discrepancies and how we prevent measurements with high wind speed discrepancies from affecting our results:

"SI Results: Wind Speed discrepancies
Southern Ocean wind speeds during the ACE-SPACE expedition were often much lower than climatological mean values, but on average were higher than winds in our ensemble (Schmale et al., 2019). We account for the effects of inter-annual variability in the Var(R) term in equation S1. However, monthly mean differences between ERA-Interim wind speeds in the measurement year and the year used in the ensemble are less than 20% along the route taken by the ACE-SPACE campaign vessel (Fig. S4). The modest discrepancy in wind speeds may be important for constraining aerosol concentrations, because sea spray emissions in our model are strongly dependent on wind speeds (Gong, 2003). However, the measured wind speed and N700 values are only weakly correlated (Pearson correlation coefficient of around 0.2) when degraded to the resolution used for comparison with model output.

Our constraint process has in-built functionality that prevents the use of measurements with large model-measurement discrepancies. We tested the robustness of our constraint methodology to the discrepancy in wind speeds by neglecting around 50% of the measurements (those with the largest discrepancies between measured and AER-ATM PPE mean simulated winds) and repeating the constraint. The effects on marginal parameter and aerosol forcing constraints were negligible (not shown). The consistency of constraint, with and without measurements in locations with relatively large model-measurement wind speed discrepancies, suggests the constraint methodology is insensitive to wind speed discrepancies caused by daily wind speed variability and differences in meteorological years between model simulations and measurements."

Response to anonymous reviewer 2:

We have positively responded to many of reviewer 2's suggestions and think this has improved our revised manuscript considerably. However, we think some of the changes suggested by reviewer 2 could mislead the reader, by giving a too simple representation of our constraint process. The reviewer encourages us to emphasise uncertainty in the measurements, yet our focus in this article is on uncertainties in the model-measurement constraint process. Hence, there are some suggestions we have not been able to accommodate.

Fig. 1(b/d), please overlay measured standard deviation as dots, as performed for the average of the measurements (a/c).

We appreciate the motivation behind this suggestion, but think adding these dots would confuse the reader. We overlaid the mean measurements over the mean model field in Fig 1 a) and c) because these values are directly comparable. However, the model standard deviation represents uncertainty in the model parameters while the measurement standard deviation represents temporal and spatial variability, as well as instrument error – they should not be compared. We have not neglected the measurement variability. The implausibility metric used in our model-measurement constraint process includes spatial and temporal representation errors, emulation errors, inter-annual variability and

instrument uncertainty. It would be misleading to compare any one of these with the model parametric uncertainty.

The authors are focussing on natural aerosol. How were any ship measurements influenced by anthropogenic pollution eliminated from the analysis?

The measurements were filtered to ensure that they are free of ship stack contamination. In appendix A of Schmale et al., 2019, we explain our method: "*Equivalent black carbon, trace gases data such as CO and CO2, and the 10 s^{-1} variability of particle number concentrations were used to identify the influence of ship exhaust. Identified exhaust periods are not included here and constitute about 50% of the total data. Size-dependent particle losses in the inlet lines were determined experimentally after the cruise and data are corrected accordingly. Losses were <10% for submicron particles and about 15% for supermicron particles.*".

It is important to note that we don't assume that all the sampled aerosol was natural. The atmosphere may have contained some anthropogenic aerosol from distant sources. The model includes these and several anthropogenic aerosol emission and process parameters were perturbed in our ensemble. These parameters were very modestly constrained, suggesting that the environment is dominated by natural aerosol.

Is each measurement used given equal weighting in constraining the model?

Yes, measurement types are given as equal a weighting as possible in the constraint process.

The description of our constraint process in the original manuscript lacked some important details and relied on the methodology described in Johnson et al. (2019). Our constraint process relies on the use of "implausibility" metrics, which are calculated for each of the one million model variants, for each measurement type at each measurement location. We set implausibility thresholds for each measurement type and also set exceedance tolerances, defined as the number (or percentage) of measurements for which a model variant's output exceeds the specified threshold. The constraint efficacy differs between measurement types and we adjust the tolerance and exceedance threshold values (defined in the SI) for each measurement type to ensure each variable constraint retains approximately the same proportion of the original sample of model variants. The proportion retained by individual measurement type constraints varies from 18% to 30%.

We have enhanced the description of our implausibility threshold and exceedance tolerance value selection process in the SI section "SI Methods: Model-measurement comparisons" and have included two additional tables (tables S1 and S2) for readers interested in the specific values used for each measurement type. The additional text included in the adjusted SI reads:

"We set threshold and tolerance values for each variable distinctly for each month of data. This makes processing the implausibility data more efficient and allows for a degree of automation of the constraint process. We ensure that each measurement type on each leg of the journey (Schmale et al., 2019) affects the combined constraint. This requires quantification of the constraint of individual measurement types on parameter values at multiple combinations of threshold and implausibility exceedance tolerances. We avoid increasing the threshold and/or tolerance values in individual months for each measurement type, if the constraint efficacy of the

measurement would saturate as a result. Otherwise, threshold and tolerances for each month are required to be as similar as possible.

Although our analysis in the main article focusses on a combined measurement constraint, this analysis is informed by individual measurement type constraints. The threshold and exceedance tolerances for individual measurement type constraints are summarised in table S1. Only 0.004% of the one million model variants (40 variants) are retained when these individual constraints are combined. Thus, we relax the threshold and tolerance criteria for each measurement type constraint when combining constraints (table S2)."

SI: The authors state: "The variance terms in the denominator of Eq. (1) are calculated uniquely for each measurement. Following Johnson et al., (2019), we use a measurement uncertainty of 10%". Are the measurement errors for the constraints used in this study homoscedastic or heteroscedastic? Do they correspond with the definition of the implausibility metric (eq. 1, SI)? How does the variability in the measurements compare to the uncertainty chosen (10%)?

We applied heteroscedastic uncertainties for measurement and representation errors for consistency with Johnson et al. (2019). It would have been far simpler to apply homoscedastic uncertainties. We acknowledge that our choice of heteroscedastic errors is a subjective decision. However, as shown in Fig. 2, we reject model variants with the lowest values for each measurement type, which correspond to our lowest instrument error values. If we had used homoscedastic errors, all constraints would have been weaker. Thus, we would have needed to reduce implausibility thresholds and exceedance tolerances to attain the same degree of constraint.

We think the 10% instrument error applied here is an overestimate. This is a cautious approach that allows us to avoid over-constraint based on this set of measurements. Furthermore, there is limited data available to inform our choice of spatial and temporal representation errors. The "variability in the measurements" on short timescales at point locations conflates instrument error with spatial and temporal representation errors, but does not fully encompass any of these. Dedicated measurement campaigns are required to establish robust estimates of these errors. We therefore elected to use 10% of the measured value for instrument error, as well as 20% and 10% respectively for spatial and temporal representation errors, to maintain consistency with Johnson et al. (2019). These errors are typically larger than strictly necessary, which is intentional. Larger uncertainties prevent us from over-constraining the model. Our approach is based on ruling out model variants (and parts of parameter space) that are implausible, rather than on finding all model variants that are plausible. This is a subtle, but important, distinction that shapes our methodology. Even so, using these relatively large model-measurement comparison uncertainties, we are able to rule out the majority of model variants successfully by adjusting our implausibility thresholds and exceedance tolerances.

CCN0.2% and CCN1% are used as observational constrains in the study. The measurement study in which these constraints were taken from measured CCN at more than two supersaturations. Why was a CCN spectra (or measured aerosol size distribution) not used from the observations to provide a tighter constraint on the model?

It is reasonable to assume we would attain a stronger constraint if we had used the full CCN spectra. However, we showed in Johnson et al. (2018) that diverse measurements are more

useful for constraint than additional measurements of a similar nature. Hence, we elected to combine measured concentrations of N_{700} and non-sea-salt sulfate with CCN concentrations at two supersaturatons that provide distinct information about the aerosol size distribution, rather than multiple supersaturations that would provide similar information and constraints.

We actually found that the $CCN_{0.2}$ and $CCN_{1.0}$ measurements provide very similar constraints on the model parameters, so we do not have cause to believe additional CCN measurements at alternative supersaturations would improve the constraint. We have included an additional table (table S3) in the SI showing how individual measurement types constrain the parameters, so that the reader can appreciate the similarity of CCN constraints at different supersaturations.

Please provide more detail on the observations used as constraints in the SI, linking clearly to Fig. 1 in the main article. For example, demonstrate a time-series of one of the observation dots in Fig. 1 graphically, including the variability (bars representing standard deviation), and colour of dotted time-series representing position. Clearly link this graphic to the mathematical construction of the model constraint e.g. implausibility metric in the SI.

The measurements used in our constraint are publically available (https://zenodo.org/communities/spi-ace?page=1&size=20%20) and we reference the link to the dataset in Schmale et al. (2019), as well as here in the appendix on line 293.

It is important to note that our implausibility metric relies on multiple sources of model-measurement comparison uncertainty and the reviewer's request highlights only one (limited) aspect of the uncertainty. However, we think showing how the high time resolutions data is degraded for comparison to model output warrants attention. Therefore, we have included an additional figure in the SI (Fig. S1) which gives an example of this process. We added this figure to the section "SI Methods: Model-measurement comparisons" on around line 132, so that the relevance of degrading measurements for comparison with model output using our implausibility metric is contextualised.

The revised SI text referring to Fig. S1 reads:
"The variance terms in the denominator of Eq. (1) are calculated uniquely for each measurement. Following Johnson et al. (2019), we use an instrument error of 10%, a spatial co-location uncertainty of 20% and a temporal co-location uncertainty of 10%. Fig. S1 shows an example of the degradation of data for comparison with monthly mean model output. Emulator uncertainty is calculated for each model-measurement combination using the error on the predicted mean from the emulator for the model variant. We use residuals in de-trended monthly mean output from a HadGEM-UKCA hindcast simulation over the period of 1980-2009 (Turnock et al., 2015) to estimate the inter-annual variability for each variable across all model gridboxes and months."

The authors use four measurements as a constraint (listed above). Which measurements provided the highest information content for model constraint? I would like to see some discussion on the relative constrain the individual measurement parameters provided o the model. This would help inform future measurement campaigns in this region on key measurement parameters. For example, the authors state (SI): "Non-sea-salt sulfate was calculated by subtracting this fraction from the total particulate sulfate". How much extra

constraint on the parameters (Fig. 3) is provided by using both N700 and Nss-sulfate as constraints, over just one of these

We agree with the reviewer. The efficacy of individual measurement type constraints on model parameters and processes is important and could be used to motivate targeted measurement campaigns. This information could also help identify model development priorities. We have therefore included an additional table in the SI (table S3) that shows the effects of individual measurement type constraints on model parameters. Significant additional effort would be needed to quantify the 2-way and 3-way constraint combinations. Our research focus in this paper is on the benefits of the ACE-SPACE measurements over and above more readily available measurements, so we have only added the requested individual constraints, which we think will satisfy the curiosity of the vast majority of readers. In addition to table S3, we have added the following explanatory text in the SI:

"*In addition to the constraint achieved by combining remote marine aerosol measurements, table S3 shows the effect of individual measurement type constraints (table S2) on model parameters and how these translate into a combined constraint (Fig. 3).*"

The authors provide the unconstrained and constrained model PDFs of the aerosol properties. A uniform prior range is assumed in this method. How does this represent the observations? Please show a PDF of the observed distributions to see if this is a true representation of the ship observations.

We think there has been some confusion. The uniform prior ranges are applied to individual uncertain model parameters (not variables) and are used to densely sample model uncertainty (one million model variants) uniformly across the multi-dimensional parameter space using our statistical emulators. This unconstrained sample results in pdfs of the output variables shown in Fig. 2. We make no assumptions about measurement distributions, except for the model-measurement comparison uncertainties included in equation S1.

The authors have shown how the aerosol parameters are constrained using observations, and subsequently the reduction in forcing uncertainty from the original PPEs. The paper is missing some discussion on the linkage between the constraint of these parameters and forcing. Inclusion of this would be very beneficial to the community. For example, how has average cloud microphysical properties –e.g. cloud droplet concentrations been constrained following the constraints shown in Fig. 2? Do they compare better, or worse with satellite observations in the region? This would help inform whether the constrain on forcing represents a true constraint on the aerosol processes (i.e. is the constraint of CCN by scaling sea salt right for the right reasons, or should the results be presented/interpreted as a tuning...?).

We agree that this is an important consideration, which will help the reader understand that our method leads to an actual constraint on model output, which a typical model tuning approach would not. Our method relies on ruling out implausible model variants (a true constraint), rather than identifying the best model (a tuning process). Therefore, processes

are constrained as the reviewer suggests. We have described these results and added a new table to the main article (table 1). We also include an additional co-author (Daniel P. Grosvenor) who provided cloud droplet number concentration data for analysis. The text at around line 191 now reads:

"Firstly, the magnitude of median RF_{aci} weakens from -1.99 W m^{-2} to -1.88 W m^{-2} (-1.64 to -1.49 W m^{-2} for ERF_{aci}). A weaker forcing is consistent with higher natural aerosol emissions, increased aerosol load and higher cloud droplet number concentrations in the early-industrial period. Table 1 shows that our constraint on natural emission parameters also constrains Southern Ocean cloud droplet number concentrations towards higher values, reducing the credible interval by around 50% and bringing mean values into closer agreement with MODerate Imaging Spectroradiometer (MODIS; Salomonson et al., 1989) instrument data (note that droplet number concentrations were not used to constrain the model). Thus, we conclude that the constraint on aerosol forcing towards weaker values is a genuine constraint and not the result of an arbitrary tuning."

We have also described how we processed the cloud droplet number concentrations in the SI section "SI Methods: Measurements" on line 108 of the SI:

"We present monthly mean and annual cloud droplet number concentrations in table 1 from the model and from satellite data, over the region between 50°S and 60°S. Following Grosvenor et al., (2018), we calculated cloud droplet concentrations from the MODIS (MODerate Imaging Spectroradiometer) Collection 5.1 Joint Level-2 (Aqua satellite) for the year 2008 (to correspond to the meteorological year used in our simulations). Our calculation used cloud optical depth and 3.7 micron effective radius values derived under the adiabatic cloud assumption (essentially, cloud liquid water increases linearly with height, droplet concentrations are constant throughout the cloud and the ratio of volume mean radius to effective radius is constant). We improved the cloud droplet concentration data (Grosvenor et al., 2018b) by excluding 1x1 degree data points for which the maximum sea-ice areal coverage over a moving 2-week window exceeded 0.001%. The sea-ice data used in this process were the daily 1x1 degree version of Cavalieri et al. (2016). As with other data used in our model-measurement comparison, we degraded the cloud droplet number concentration data to the model gridbox and monthly mean spatial and temporal resolutions."

Finally, we state how the cloud droplet number concentration data can be accessed in the "Data availability" section:
"The basis for our cloud droplet number concentration data are available from http://catelogue.ceda.ac.uk/uuid/cf97ccc802d348ec8a3b6f2995dfbbff."

What is the average supersaturation over the Southern Ocean simulated by the model? How does this correspond with the selected value of CCN0.2% as representative for (cloud-active aerosol, SI) in the region?

Cloud supersaturation is not known from measurements. The measurements of CCN at 0.2 and 1.0% supersaturation span a range of likely values. The key point here is that both CCN definitions constrain the model quite similarly (now made clear to the reader in table S3), so it is not vital that we know the actual supersaturation precisely.

The authors make clear that they are targeting parametric uncertainty, and the method does not address model structural uncertainty. However, some of the conclusions presented rely too heavily on the information provided by the parametric uncertainty analysis alone, specifically in the comparison to Revell et al., (2019) (Line 166 and thereafter). The differences in conclusions related to over/underestimation of sea spray aerosol are attributed to a lack of sampling of aerosol processes by Revell et al., 2019. A discussion on the role of structural errors in the model used by the author would be is required. What are the key differences between the model configurations with respect to representation of marine aerosol sources and sinks? What is the relevant contribution to

aerosol mass from secondary vs. primary marine aerosol sources in the two model configurations?

We agree with the reviewer. Many readers will be interested in how repeating our analysis using a model that includes structural developments might affect our results. We encourage that sort of activity. Therefore, we have added detail to our description (in the discussion section) of the importance of interpreting our results in the context of single climate model uncertainty. We have also made it clear that our method neglects structural uncertainties. This adds to our discussion about the need to quantify both single model uncertainty and multi-model diversity in our conclusions. We have not contrasted the primary vs secondary contributions to marine aerosols in the models because this goes well beyond the scope of our article. However, we now highlight some of the structural advances most likely to affect our results in the paragraph starting on line 281 of the original manuscript.

The revised text reads:

"Our results are based on uncertainty in a single climate model. The model is structurally consistent in our experiments, so neglects uncertainty caused by choice of microphysical and atmospheric process representations. Our model also neglects some potentially important sources of remote marine aerosol, such as primary marine organic aerosol (Mulcahy et al., 2020) and methane-sulfonic acid (Schmale et al., 2019; Hodshire, et al., 2019; Revell et al., 2019). Model inter-comparison projects (such as CMIP6) can be used to quantify the diversity of RF (or ERF) output from models, but they lack information about single model uncertainty. Ideally, multi-model ensembles would contain a perturbed parameter component, so that model diversity and single model forcing uncertainty could be quantified simultaneously. But, computational costs prevent many modelling groups from engaging with this important aspect of uncertainty quantification, limiting our shared knowledge about the causes of aerosol forcing uncertainty. Studies like ours that quantify the remaining uncertainty in aerosol forcing and its components after constraint using multiple measurement types fill an important knowledge gap. This knowledge can be used to form a more complete understanding of the importance of historical and near-term aerosol radiative forcing which would reduce the diversity in equilibrium climate sensitivity across models. "

It has been brought to our attention that we misinterpreted the results in Revell et al. (2019), by misreading the seasonal effects described in the article. Revell et al. (2019) showed, using a more recent version of our model and using interactive chemistry, that simulated sea spray aerosol concentrations are higher than observed in Jun-Aug when wind speeds are relatively high. However, in Dec-Feb the model simulates too-low cloud droplet number concentrations and AOD compared with satellite observations. Our article text has been adapted to more accurately represent the consistency of our constraint with the findings of Revell et al. (2019), and to more transparently describe the seasonal specificity of our constraint on sea salt emissions using the Gong (2003) parametrisation.

The adapted text on line 158 reads:

"This suggests that sea spray emissions in our model need to be significantly higher than those calculated using the wind speed dependent Gong (2003) parametrisation in the Southern Hemisphere summer. The higher flux is consistent with Revell et al. (2019), who showed that a more recent version of our model simulates cloud droplet concentrations and aerosol optical depth values that are lower than observed over the Southern Ocean in the Southern Hemisphere summer. However, in the Southern Hemisphere winter Revell et al., (2019) simulated higher aerosol optical depths than observed, which they corrected by reducing the dependence of sea spray emissions on wind speed. Hence, our constraint on sea spray emission fluxes may only be appropriate for Southern Hemisphere summer when wind speeds are relatively low. We do not make any assumptions about the composition of these additional summertime sea spray particles. They may be rich in organic material as proposed by Gantt et al. (2011) which would alter the CCN activity of emitted particles. However, the

consistency of constraint of $CCN_{0.2}$ and N_{700} towards higher values (Fig. 2, table S3) implies that a general scaling of the existing sea spray flux is consistent with the measurements from December to April, without the need for an additional source of fine-mode, organic-rich particles."

Given the use of an older configuration of the model HadGEM by the authors, the results should be presented in light of the latest configuration. Stars showing the values for the parameters overlaid on Fig.3/5 that represent the configuration used by Revell et al., 2019 should be included to aid the reader in understanding differences found between the two studies with regard to sea salt emissions.

This suggested change is no longer pertinent, since our results are in better agreement with Revell et al., (2019) than we initially thought. We have not included the suggested alteration to our figures, because Revell et al. (2019) made structural changes to process representations, including to the Gong (2003) sea salt emission parameterisation and model chemistry. Thus, we believe highlighting parameter values used in a structurally different model would mislead the reader. The effect of including structural changes on model output is described in Revell et al. (2019) and in the model development papers cited within.

How much of the constraints found in Fig.3 are due to compensating parameters across the multi-dimensional marginal probability distributions? For example, what is the relationship between the marginal distributions between dry deposition and sea salt? Could the authors also provide an investigation of the joint marginal histograms between DMS and sea salt emission.

We agree that the joint constraint of key parameters may be of considerable interest to readers and thank the reviewer for the suggestion. In Fig. 13 of Regayre et al. (2018) we used 2-dimensional density plots to highlight the important role model equifinality plays on reducing constraint efficacy when single-model uncertainty is densely sampled. Here, we added figure S3 to the SI to show joint marginal densities of key parameters as suggested.

We introduce the new figure (Fig. S3) in the SI section "SI Results: Constrained marginal parameter distributions":

"Constrained marginal parameter distributions in Fig. 3 and Fig. 5 of the main article tell a one-dimensional story. In Fig. S3, we show the effect of constraint to remote marine aerosol measurements, combined with the constraint from Johnson et al. (2019) on a subset of the marginal 2-dimensional parameter combinations."

We refer to this new figure on around line 153 of our revised article to highlight the consistency of constraint across the parameter space and on line 275 to emphasise how compensating parameter effects limit the efficacy of constraint.

"*These joint constraints (see also Fig. S3) suggest the model-measurement comparison is improved when aerosol number concentrations and mass are relatively high.*"

"*Secondly, even within the considerably reduced volume of multi-dimensional parameter space there still exist many compensating parameter effects (Fig. S3), which limit the constraint on individual parameter ranges (Lee et al., 2016; Regayre et al., 2018).*"

It is stated that the "model-measurement comparison is improved when aerosol

number concentrations and mass are relatively high". Does the model configuration used have the same total sources of aerosol number/mass compared to the configuration of the model used by Revell et al., 2019? This could be included in the SI.

We agree that it is important to make the reader aware of potentially important structural advances that may affect interpretation of our results, which we now do in the final paragraph of the discussion. However, we do not contrast our total aerosol and mass with those in Revell et al. (2019) for two main reasons. Firstly, our results are more consistent with those of Revell et al. (2019) than originally thought. Secondly, our article focuses on the single-model uncertainty constraint. Comparisons between model versions with structurally distinct process representations is beyond the scope of our article. This sort of analysis, based on multiple structural changes, is best presented using experiments designed for that specific purpose.

The revised text reads:

"Our results are based on uncertainty in a single climate model. The model is structurally consistent in our experiments, so neglects uncertainty caused by choice of microphysical and atmospheric process representations. Our model also neglects some potentially important sources of remote marine aerosol, such as primary marine organic aerosol (Mulcahy et al., 2020) and methane-sulfonic acid (Schmale et al., 2019; Hodshire, et al., 2019; Revell et al., 2019). Model inter-comparison projects (such as CMIP6) can be used to quantify the diversity of RF (or ERF) output from models, but they lack information about single model uncertainty. Ideally, multi-model ensembles would contain a perturbed parameter component, so that model diversity and single model forcing uncertainty could be quantified simultaneously. But, computational costs prevent many modelling groups from engaging with this important aspect of uncertainty quantification, limiting our shared knowledge about the causes of aerosol forcing uncertainty. Studies like ours that quantify the remaining uncertainty in aerosol forcing and its components after constraint using multiple measurement types fill an important knowledge gap. This knowledge can be used to form a more complete understanding of the importance of historical and near-term aerosol radiative forcing which would reduce the diversity in equilibrium climate sensitivity across models."

Are there any other potential marine aerosol sources currently missing in the model configuration used by the authors that would increase aerosol number/mass by a similar magnitude than scaling sea salt emissions to 3 times the default value? This requires discussion, in particular in light of the conclusions presented by the study cited for the source of the observations (Schmale et al., 2019) used by the authors, e.g.:
Schmale et al., 2019: "The regions of highest underestimation are close to the coast of Antarctica during leg 2, close to South Africa and around 45_E during leg 1. These regions coincide with the highest concentrations of gaseous MSA… This preliminary model–measurement comparison suggests that the model may be missing an important source of high-latitude CCN."

We have highlighted the potential role of marine organic material, but also stated that consistency of constraint of CCN and N_{700} does not suggest the need for a special source into the accumulation mode. In Schmale et al. (2019) we drew attention to high MSA concentration measurements near the Antarctic coast. DMS emissions themselves were constrained near their central value, so do not appear to be the cause of the biases. We are not aware of any other potential explanations for such large and consistent biases in CCN and N700. The additional text on line 281, which refers to additional sources of aerosol neglected by our experiments is:

"The model is structurally consistent in our experiments, so neglects uncertainty caused by choice of microphysical and atmospheric process representations. *Our model also neglects some potentially important sources of remote marine aerosol, such as primary marine organic aerosol (Mulcahy et al., 2020) and methane-sulfonic acid (Schmale et al., 2019; Hodshire, et al., 2019*; Revell et al., 2019*).*"

SI: The authors state that the wind speed discrepancies do not affect the results presented. This is an important statement that deserves more detailed justification as I currently do not see how this is supported by the data or Korhonen et al., 2010. How do the differences in simulated and observed wind-speeds relate to the scaling of sea salt required to constrain CCN?

Our reference to Korhonen et al. (2010) has now been removed. We agree that wind speeds are important for calculating sea spray emissions and did not intend to mislead the reader. We originally cited Korhonen et al. (2010) to point out that there are many factors other than sea spray which affect remote marine cloud condensation nuclei concentrations (52% of the CCN variability according to their research). Wind speeds measured during the ACESPACE campaign are only weakly correlated with measured N_{700} concentrations. The Pearson correlation coefficient is only 0.2 when degraded to the model gridbox scale used for comparison to model output. Also, our constraint method is designed to avoid the use of measurements where structural model errors are the cause of model-measurement discrepancies. Thus, when we pre-filter the data by removing all measurements where nudged and measured wind speeds differ meaningfully, the constraint on parameters is unaffected. We have adapted this section of the SI to give the reader a better appreciation of why wind speed discrepancies do not affect the constraint.

The revised section "SI Results: Wind Discrepancies" reads:
"Southern Ocean wind speeds during the ACE-SPACE expedition were often much lower than climatological mean values, but on average were higher than winds in our ensemble (Schmale et al., 2019). We account for the effects of inter-annual variability in the Var(R) term in equation S1. However, monthly mean differences between ERA-Interim wind speeds in the measurement year and the year used in the ensemble are less than 20% along the route taken by the ACE-SPACE campaign vessel (Fig. S4). The modest discrepancy in wind speeds may be important for constraining aerosol concentrations, because sea spray emissions in our model are strongly dependent on wind speeds (Gong, 2003). However, the measured wind speed and N700 values are only weakly correlated (Pearson correlation coefficient of around 0.2) when degraded to the resolution used for comparison with model output.

Our constraint process has in-built functionality that prevents the use of measurements with large model-measurement discrepancies. We tested the robustness of our constraint methodology to the discrepancy in wind speeds by neglecting around 50% of the measurements (those with the largest discrepancies between measured and AER-ATM PPE mean simulated winds) and repeating the constraint. The effects on marginal parameter and aerosol forcing constraints were negligible (not shown). The consistency of constraint, with and without measurements in locations with relatively large model-measurement wind speed discrepancies, suggests the constraint methodology is insensitive to wind speed discrepancies caused by daily wind speed variability and differences in meteorological years between model simulations and measurements."

b) SI: The authors nudge the models to 2008 meteorology from reanalysis data. A comparison between the meteorological data between the measurement years and that used in the model simulation should be provided in the SI, comparing both monthly averages and variability.

We have included an additional figure (Fig. S4) to highlight the importance of the discrepancy in meteorology. We relate the difference in meteorology to the inter-annual variability uncertainty term included in our implausibility calculations. Monthly mean data are compared, since this is the reference scale used in our model-measurement comparison. The effects of differences in daily wind speed variability between measurement year and model simulation year are included in the inter-annual variability and temporal and spatial error terms in our implausibility metric. However, we have included an additional figure S1 to exemplify the effect of degrading measurement data to the model gridbox resolution for model-measurement comparison.

Fig. S4 is references in the revised section "SI Results: Wind Discrepancies" which now reads:

"Southern Ocean wind speeds during the ACE-SPACE expedition were often much lower than climatological mean values, but on average were higher than winds in our ensemble (Schmale et al., 2019). We account for the effects of inter-annual variability in the Var(R) term in equation S1. However, monthly mean differences between ERA-Interim wind speeds in the measurement year and the year used in the ensemble are less than 20% along the route taken by the ACE-SPACE campaign vessel (Fig. S4). The modest discrepancy in wind speeds may be important for constraining aerosol concentrations, because sea spray emissions in our model are strongly dependent on wind speeds (Gong, 2003). However, the measured wind speed and N700 values are only weakly correlated (Pearson correlation coefficient of around 0.2) when degraded to the resolution used for comparison with model output."

c) SI "Marginal parameter distributions are constrained consistently when we remove measurements with average wind speed differences larger than 50% of the measured value from the model-measurement comparison." How many results does this effect? Please show a global map where the grid-box colour represents a measure of how often this threshold is exceeded.

We have added the requested detail to the SI text. However, at the resolution used for model-measurement comparison, the correlation between measured wind speed and N700 is near-zero (Pearson correlation coefficient of 0.2). Furthermore, our constraint methodology is insensitive to large model-measurement discrepancies caused by model structural errors. This feature was our motivation for including an SI section on wind speed discrepancies, but was inadequately described in our original manuscript. Thus, we have not included the suggested figure, which could confuse the reader by leading them to assume the N700 measurements in these locations are less reliable than they are. Instead, we have refined the "SI Results: Wind Discrepancies" text:

"Southern Ocean wind speeds during the ACE-SPACE expedition were often much lower than climatological mean values, but on average were higher than winds in our ensemble (Schmale et al., 2019). We account for the effects of inter-annual variability in the Var(R) term in equation S1. However, monthly mean differences between ERA-Interim wind speeds in the measurement year and the year used in the ensemble are less than 20% along the route taken by the ACE-SPACE campaign vessel (Fig. S4). The modest discrepancy in wind speeds may be important for constraining aerosol concentrations, because sea spray emissions in our model are strongly dependent on wind speeds (Gong, 2003). However, the measured wind speed and N700 values are only weakly correlated (Pearson correlation coefficient of around 0.2) when degraded to the resolution used for comparison with model output.

Our constraint process has in-built functionality that prevents the use of measurements with large model-measurement discrepancies. We tested the robustness of our constraint methodology to the discrepancy in wind speeds by neglecting around 50% of the measurements (those with the largest discrepancies between measured

and AER-ATM PPE mean simulated winds) and repeating the constraint. The effects on marginal parameter and aerosol forcing constraints were negligible (not shown). The consistency of constraint, with and without measurements in locations with relatively large model-measurement wind speed discrepancies, suggests the constraint methodology is insensitive to wind speed discrepancies caused by daily wind speed variability and differences in meteorological years between model simulations and measurements."

Line 178: "The constraint on the scaled DMS emission flux is two-sided, 179 reducing the credible range of DMS emission scaling from 0.5 to 2.0 down to 0.54 to 1.9." Could the authors please make clear what in the figure 0.54/1.9 corresponds to.

We have now defined DMS as dimethylsulfide on line 152:

"Several other parameters (related to cloud droplet pH, dimethylsulfide (DMS) emissions and wet deposition) are more modestly constrained."

We have also clarified that DMS is an aerosol precursor, and that this parameter is a scale factor on the default emissions (originally referred to as a scaling). We have now included references to the schemes used, since other modelling groups may use different seawater concentrations and/or emission flux representations.

Revised text on line 178 now reads:

"Other parameters are more modestly constrained. The constraint on the aerosol precursor DMS emission flux scale factor is two-sided, reducing the credible range of DMS emission scalings from 0.5 to 2.0 down to 0.54 to 1.9. This constraint suggests the default surface sea water concentration (Kettle and Andreae, 2000) and emission parameterisation (Nightingale, et al., 2000) are consistent with measurements (including aerosol sulfate) and do not benefit from being scaled. Furthermore, ACE-SPACE measurements are consistent with less-efficient aerosol scavenging (55% likelihood of Rain_Frac, the parameter that controls the fractional area of the cloudy part of model grid boxes where rain occurs, being below the unconstrained median value 0.5) and less aqueous phase sulfate production (pH of cloud droplets has a 62% likelihood of being lower than the unconstrained median value). These combined constraints suggest, in agreement with sea spray and deposition parameter constraints, higher aerosol number and mass concentrations are consistent with measurements."

SI, Line 95: Grammar - "pdfs with centralised tendencies will by heavily weighted". Change by to be.

Done.

SI, Line 63: "We make use of the ATM and AER-ATM perturbed parameter ensembles (PPEs)". Following this the authors refer only to AER and AER-ATM. Should this read: "We make use of the AER and AER-ATM"?

Yes, this has been corrected.

Fig. 2: Should y-axis density not be labelled 0-1? Or are these not normalised marginal densities.

Our description of these figures was inadequate. The purpose of this figure is to contrast the shape of the probability densities for the unconstrained and constrained sets of model variants. These are not normalised marginal densities. The density curve for each sample of model variants (unconstrained and constrained) is scaled such that the area under the curve integrates to one. This means that the densities can be compared visually on the same

figure. The values on the y-axis are not helpful (or needed) for comparing the shape of probability density curves of the different samples, and have therefore not been included in the figure.

We have added an extra sentence to the caption of figure 2 to make the scaling clear.

"*Densities for each sample of model variants are scaled so that the area under the curve integrates to one."#*

The value of remote marine aerosol measurements for constraining radiative forcing uncertainty

Leighton A. Regayre[1], Julia Schmale[2,3], Jill S. Johnson[1], Christian Tatzelt[4], Andrea Baccarini[2], Silvia Henning[4], Masaru Yoshioka[1], Frank Stratmann[4], Martin Gysel-Beer[2]–, Daniel P. Grosvenor[1,5] and Ken S. Carslaw[1]

[1]Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds, LS2 9JT, UK
[2]Paul Scherrer Institute, Laboratory of Atmospheric Chemistry, Villigen, Switzerland
[3]École Polytechnique Fédéderale de Lausanne, Lausanne, Switzerland
[4]Leibniz Institute for Tropospheric Research, Leipzig, Germany
[5]National Centre for Atmospheric Science, Leeds, UK

Correspondence to: Leighton Regayre (L.A.Regayre@leeds.ac.uk)
Correspondence related to measurements to: Julia Schmale (julia.schmale@psi.ch)

Abstract. Aerosol measurements over the Southern Ocean are used to constrain aerosol-cloud interaction radiative forcing (RF_{aci}) uncertainty in a global climate model. ~~Aerosol f~~Forcing uncertainty is quantified using one million climate model variants that sample the uncertainty in nearly 30 model parameters. ~~Ship-based measurements of cloud condensation nuclei, particle number concentrations and sulfate mass concentrations from the Antarctic Circumnavigation Expedition: Study of Preindustrial-like Aerosols and Their Climate Effects (ACE-SPACE) are used to identify observationally implausible variants and thereby reduce the spread in the simulated forcing. Southern Ocean m~~Measurements of cloud condensation nuclei and other aerosol properties from an Antarctic circumnavigation expedition strongly constrain natural aerosol emissions: default sea spray emissions ~~in the model~~ need to be increased by around a factor of 3 to be consistent with measurements. ~~Aerosol f~~Forcing uncertainty is reduced by around 7% using this set of several hundred~~ese~~ measurements, which is comparable to the 8% reduction achieved using ~~an~~ a diverse and extensive set of over 9000 predominantly Northern Hemisphere measurements. ~~The radiative forcing due to aerosol-cloud interactions (RF_{aci}) is constrained to -2.61 to -1.10 W m^{-2} (95% confidence) and the effective radiative forcing from aerosol-cloud interactions (ERF_{aci}) is constrained to -2.43 to -0.54 W m^{-2}.~~ When Southern Ocean and Northern Hemisphere measurements are combined, ~~the~~ uncertainty in RF_{aci} is reduced by 21% and the strongest 20% of forcing values are ruled out as implausible. In this combined constraint ~~the,~~ observationally plausible RF_{aci} is around 0.17 W m^{-2} weaker (less negative) with 95% credible values ranging from -2.51 to -1.17 W m^{-2} ~~and from~~ (standard deviation -2.18 to -1.46 W m^{-2}) ~~when using one standard deviation to quantify the uncertainty~~. The Southern Ocean and Northern Hemisphere measurement datasets are complementary because they constrain different processes. These results highlight the value of remote marine aerosol measurements.

1 Introduction

The uncertainty in the magnitude of the effective radiative forcing caused by aerosol-cloud interactions (ERF_{aci}) due to changing emissions over the industrial period is around twice that for CO_2 (Stocker et al., 2013). It is

essential to reduce this uncertainty if global climate models are to be used to robustly predict near-term changes in climate (Andreae et al., 2005, Myhre et al., 2013, Collins et al., 2013, Tett et al., 2013, Seinfeld et al., 2016).

Aerosol forcing uncertainty has persisted in climate models since the 1990s partly because there are no measurements covering the industrial period that can be used to directly constrain simulations of long-term changes in aerosol and cloud properties (Gryspeerdt et al., 2017; McCoy et al., 2017). Estimates of aerosol forcing over the industrial period therefore rely on models that have been evaluated against measurements made in the present-day atmosphere. However, it is known that the aerosol forcing (in particular the component caused by aerosol-cloud interactions) depends sensitively on the state of aerosols in the pre-industrial period (Carslaw et al., 2013; Wilcox et al. 2015) when natural aerosols were dominant (Carslaw et al., 2017). Observations of natural aerosols in the present-day atmosphere are therefore expected to help constrain the simulated forcing unless there have been significant changes in natural aerosol processes over the industrial period, for which there is little evidence (Carslaw et al., 2010).

In this paper we address the questions: i) To what extent can measurements of aerosols in pristine (natural) environments help to constrain model simulations and thereby reduce the large uncertainty in aerosol forcing? ii) What is the relative importance of measurements in remote and polluted environments for constraining the forcing uncertainty? It is known that the abundance of natural aerosols affects the magnitude of forcing in a model (Spracklen and Rap, 2013; Carslaw et al., 2013). However, to assess the effect on the *uncertainty* in forcing it is necessary to explore how the spread of predictions of a set of models changes when constrained by measurements. The 5th Coupled Model Intercomparison Project is inadequate for this purpose because of insufficient aerosol diagnostics (Wilcox et al., 2015). Here we use large perturbed parameter ensembles (PPEs) of the UK Hadley Centre General Environment Model HadGEM3 (Hewitt et al, 2011). The PPEs were created by systematically perturbing numerous model parameters related to natural and anthropogenic emissions and physical processes (Yoshioka et al., 2019). The simulated aerosol forcings have uncertainty ranges that exceed those of multi-model ensembles (Yoshioka et al., 2019; Johnson et al., 2019). Instantaneous radiative forcing (RF) is quantified using the 26-parameter AER PPE in which just aerosol-related parameters were varied, and the effective radiative forcing (ERF) is quantified using the 27-parameter AER-ATM PPE in which aerosol and physical atmosphere parameters were varied (Yoshioka et al., 2019). We use these PPEs to quantify how the constraint provided by pristine aerosol measurements affects the spread of aerosol forcings simulated by the ensembles.

Previous analysis of HadGEM3 PPEs showed that measurements of the present-day atmosphere in regions affected by anthropogenic emissions help to constrain ~~have limited impact on~~ the uncertainty in ~~simulated~~ aerosol-radiation interaction forcing (RF_{ari}) but not the component due to aerosol-cloud interactions (RF_{aci}). For example, Regayre et al.~~,~~ (2018) showed that top-of-the-atmosphere shortwave radiation flux measurements reduce ERF_{aci} uncertainty by only around 10%, despite the fluxes in the present-day and early-industrial environments sharing multiple causes of uncertainty. Johnson et al. (2019) showed that a much larger dataset of over 9000 (predominantly Northern Hemisphere) aerosol measurements ~~constrained the~~reduced the uncertainty in global, annual mean aerosol RF_{ari} (neglecting rapid adjustments) ~~uncertainty~~ by 35%, but RF_{aci} uncertainty by only around ~~8~~7%. These measurements reduce the uncertainty in a small number of parameters related to anthropogenic emissions and aerosol processing in polluted environments. However, important causes of uncertainty in RF_{aci}, such as natural aerosol emission fluxes, were largely unconstrained.

The Southern Ocean is one of the few regions on Earth (along with some boreal forests) in which the same processes are expected to affect cloud-active aerosol concentrations in the present-day and early-industrial atmospheres (Hamilton et al., 2014). In this study we make use of aerosol measurements from the Antarctic Circumnavigation Expedition: Study of Preindustrial-like Aerosols and Their Climate Effects (ACE-SPACE) campaign (Schmale et al., 2019). They offer a unique opportunity to constrain the early-industrial aspects of aerosol forcing uncertainty because the Southern Ocean is a source of natural aerosols that are relevant at the global scale and remains largely unaffected by anthropogenic aerosol and precursor emissions.

We use near-surface measurements of cloud condensation nuclei concentrations at 0.2% and 1.0% supersaturations ($CCN_{0.2}$ and $CCN_{1.0}$; Tatzelt et al., 2019), as well as mass concentrations of non-sea-salt sulfate particles with dry aerodynamic diameters less than 10 µm~~in PM_{10}~~ and number concentrations of particles with dry aerodynamic diameter larger than 700 nm (N_{700}; corresponds to volume equivalent diameter larger than around 500 to 570 nm; Schmale et al., 2019a). The measurements are compared to output from 1 million variants of the HadGEM3 model that sample combinations of parameter settings in the model. These model variants are used to represent aerosol forcing uncertainty in our model using probability density functions (pdfs) and were generated by sampling from Gaussian Process emulators that were trained on the PPE model outputs

(see SI Methods). Model variants that were judged to be observationally implausible against the measurements were rejected, resulting in a set of plausible variants from which the uncertainty in aerosol forcing could be computed (see SI Methods). In the results shown below, we retained approximately 3% of model variants (following Johnson et al., 2019) that best match all four measured aerosol properties.

2 Results

Fig. 1 shows the $CCN_{0.2}$ mean and standard deviation from the unconstrained and constrained model variants to exemplify the effect of constraint on model output. The mean concentrations in the unconstrained sample are much smaller than measured concentrations. However, the range of $CCN_{0.2}$ values in the unconstrained sample spans the measurements in most locations (Fig. 1b). The measurement constraint increases $CCN_{0.2}$ concentrations (more than double the unconstrained mean in many locations; Fig. 1c) and greatly reduces the $CCN_{0.2}$ uncertainty (by more than half everywhere to less than 50 cm^{-3}; Fig. 1d).

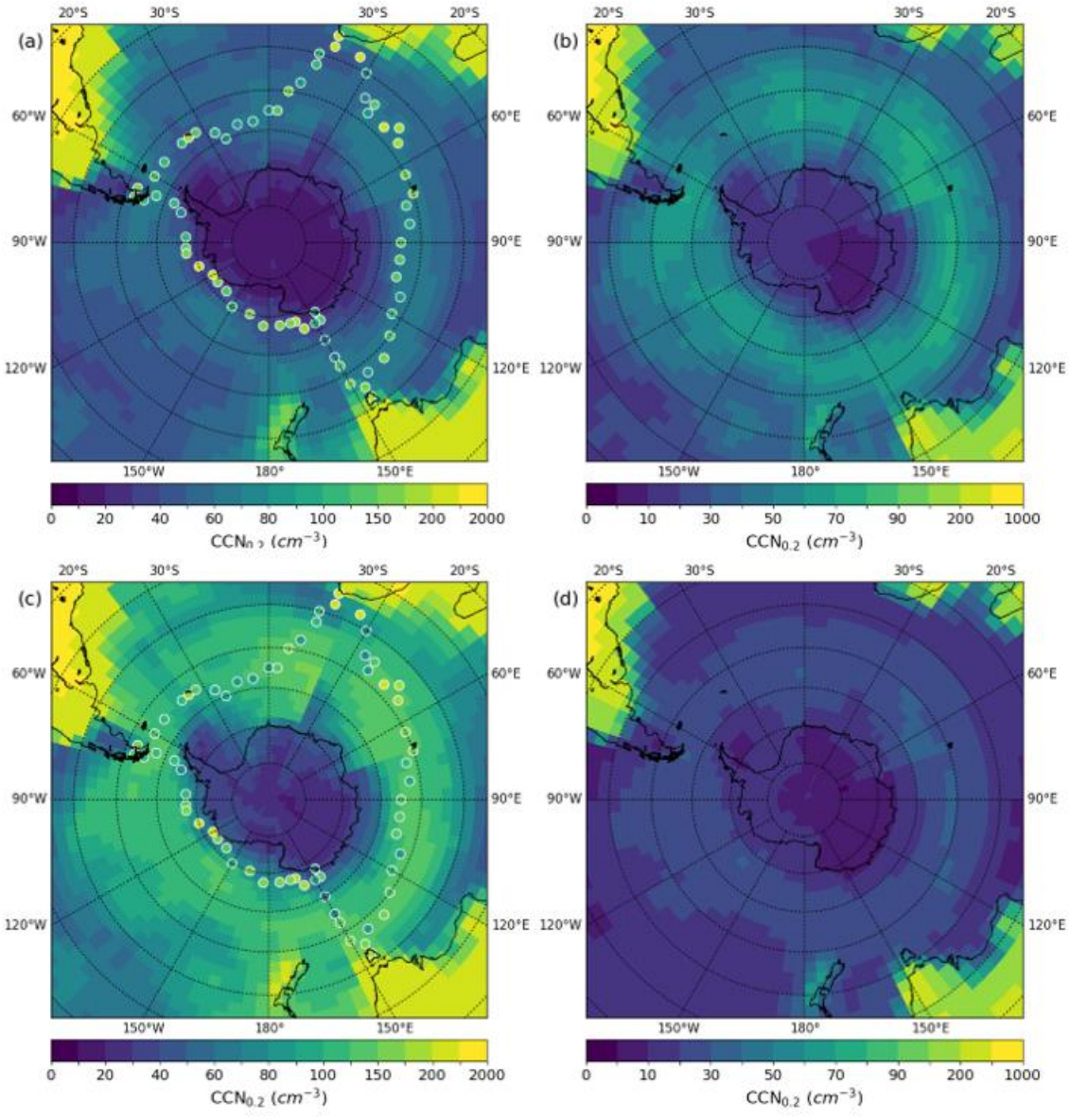

Fig. 1. a,c) Mean and b,d) standard deviation of $CCN_{0.2}$ concentrations from the a,b) unconstrained sample and c,d) the sample constrained using concentration measurements of $CCN_{0.2}$, $CCN_{1.0}$, non-sea-salt sulfate and particles with dry

aerodynamic diameters ~~numbers~~ larger than 700 nm. Measured $CCN_{0.2}$ values are plotted as dots. Means and standard deviations were calculated using samples taken from emulators trained using monthly mean values. December to March sample values were combined based on longitudinal agreement with measurements.

Fig. 2 shows pdfs of the output from the model for the four variables used as constraints, calculated as means over the locations where measurements were taken. The constraint reduces the uncertainty in all measurement types (narrower pdfs) and the central tendency of the pdfs is closer to the regional mean of measurements after constraint. Rejecting around 97% of model variants as implausible compared to measurements greatly improves the model-measurement comparison.

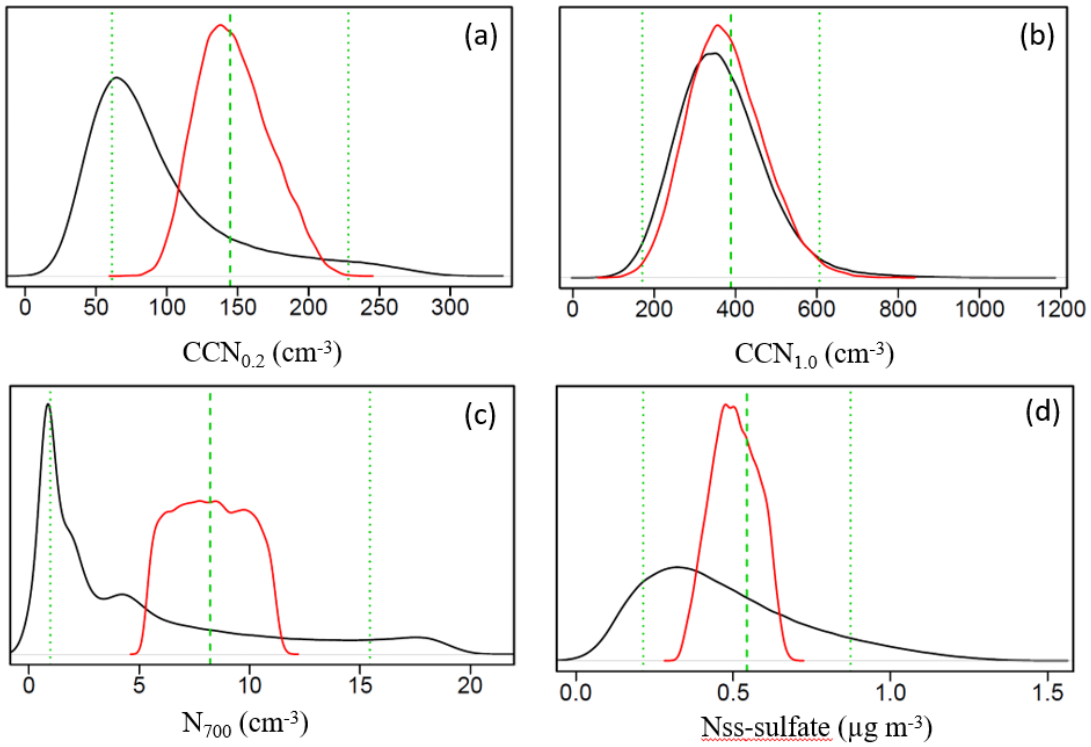

Fig. 2. Unconstrained (black) and observationally constrained (red) pdfs of aerosol properties: a) $CCN_{0.2\%}$, b) $CCN_{1.0\%}$, c) N_{700} and d) aerosol sulfate. The pdfs were calculated at locations where measurements were used for constraint across the months December to March. Densities for each sample of model variants are scaled so that the area under the curve integrates to one. The green dashed line shows the median of the measurements and the dotted green lines show the approximate uncertainty ranges due to multiple model-measurement comparison uncertainties that were accounted for in the constraint (See SI Methods).

After constraint, the remaining model variants inhabit specific parts of the 26-dimensional parameter uncertainty space used to quantify the model uncertainty. We explore the effect of constraints on parameter values using 1-dimensional marginal probability distributions (described in detail in Johnson et al., 2019) – see Fig. 3 and Fig. ~~S1~~ S2 for equivalent AER-ATM results. The magnitude of the marginal probability distribution after constraint reflects the number of ways in which a particular value of a parameter can be combined with settings of all the other parameters to produce an observationally plausible model. The white space in the marginal pdfs shows where parameter value density has decreased.

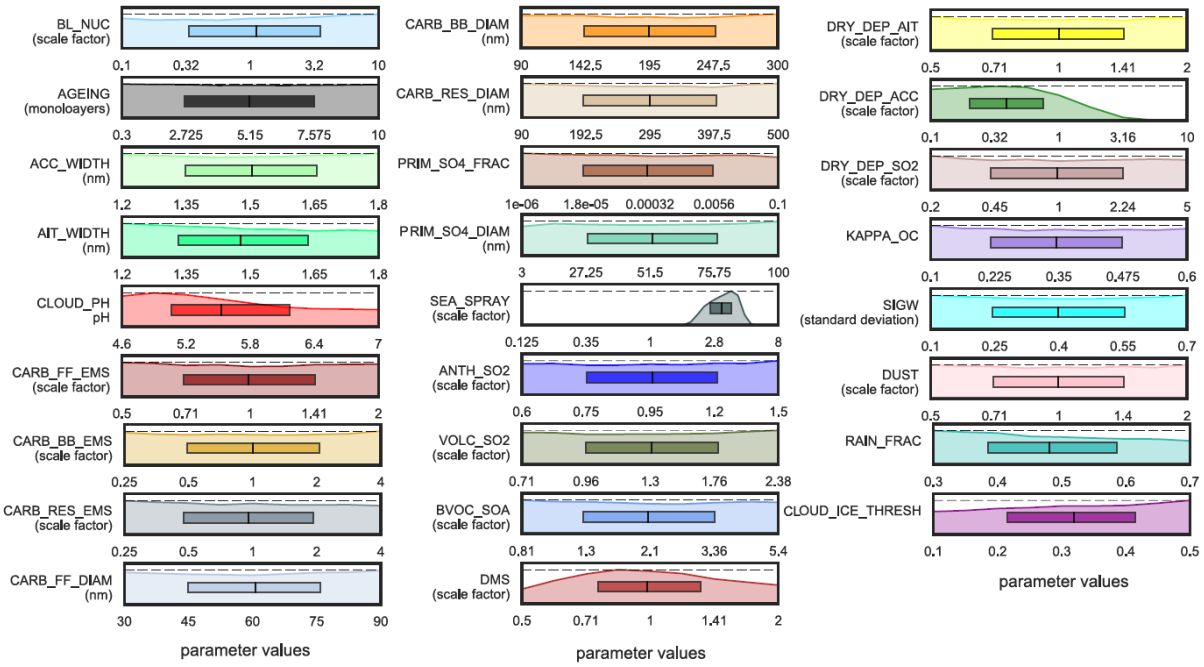

Fig. 3. Marginal probability distributions for the 26 aerosol parameters after constraint using ACE-SPACE measurements. The density of parameter values in the unconstrained sample are shown as horizontal dashed lines (uniform sampling over the parameter space). Densities of constrained samples are shown in colour and are scaled so that the maximum densities in the constrained and unconstrained samples are aligned. The 25th, 50th and 75th percentiles of each marginal distribution are shown in the central boxes. Parameter values on the x-axes correspond to values used in the model (Yoshioka et al., 2019, table S3).

The relative simplicity of aerosol emissions and processes over the Southern Ocean (compared to polluted continental regions) means that measurements can be used to tightly constrain uncertainty in the associated parameters. Two parameters~~,~~ (sea spray emissions and dry deposition velocity~~,~~) are tightly constrained such that some parameter values are ruled out as implausible even when combined with uncertainties in all other parameters. Several other parameters (related to cloud droplet pH, dimethylsulfide (DMS) emissions and wet deposition) are more modestly constrained. These joint constraints (see also Fig. S3) suggest the model-measurement comparison is improved when aerosol number concentrations and mass are relatively high.

Sea spray emissions are tightly constrained to be around 3 times larger than the default model value. Observationally plausible values of the sea spray scaling parameter range from around 1.6 to 5.1 and all other values (including the default emission calculated in the model) are ruled out as implausible. This suggests that sea spray emissions in our model need to be significantly higher than those calculated using the wind speed dependent Gong (2003) parametrisation in the Southern Hemisphere summer. ~~-~~The higher flux is consistent with Revell et al. (2019), who showed that a more recent version of our model simulates cloud droplet concentrations and aerosol optical depth values that are lower than observed over the Southern Ocean in the Southern Hemisphere summer. However, in the Southern Hemisphere winter Revell et al., (2019) simulated higher aerosol optical depths than observed, which they corrected by reducing the dependence of sea spray emissions on wind speed. Hence, our constraint on sea spray emission fluxes may only be appropriate for Southern Hemisphere summer when wind speeds are relatively low. We do not make any assumptions about the composition of these additional summertime sea spray particles. They may be rich in organic material as proposed by Gantt et al.~~,~~ (2011) which would alter the CCN activity of emitted particles. However, the consistency of constraint of $CCN_{0.2}$ and N_{700} towards higher values (Fig. 2~~1~~, table S3) implies that a general scaling of the existing sea spray flux is consistent with the measurements from December to April, without the need for an additional source of fine-mode, organic-rich particles.

~~A better understanding of these conflicting results could be achieved using a multi-model experiment that sampled a range of atmospheric process representations.~~

~~We do not make any assumptions about the composition of these additional sea spray particles. They may be rich in organic material as proposed by Gantt et al., (2011) which would alter the CCN activity of emitted~~

~~particles. However, the consistency of constraint of $CCN_{0.2}$ and N_{700} towards higher values (Fig. 1) implies that a general scaling of the existing sea spray flux is consistent with the measurements without the need for an additional source of fine-mode, organic-rich particles.~~

~~These results conflict with the findings of Revell et al. (2019) who suggest the relatively simple wind speed dependent nature of the Gong (2003) parametrisation produces too much sea spray aerosol over the Southern Ocean from December to February. If Revell et al. (2019) had sampled a wider range of processes (such as deposition) as we have here, our results might be brought into agreement. A better understanding of these conflicting results could be achieved using a multi-model experiment that sampled a range of atmospheric process representations.~~

The dry deposition velocity of accumulation mode aerosols (Dry_Dep_Acc) has an 84% likelihood of being lower than the default model value after applying the constraint. Furthermore, deposition velocities larger than around 3 times the default value are effectively ruled out. This constraint is consistent with the higher aerosol concentrations implied by constraint of the sea spray emission parameter.

Other parameters are more modestly constrained. The constraint on the ~~scaled~~ aerosol precursor DMS emission flux scale factor is two-sided, reducing the credible range of DMS emission scalings from 0.5 to 2.0 down to 0.54 to 1.9. This constraint suggests the default surface sea water concentration~~emission inventory~~ (Kettle and Andreae, 2000) and emission parameterisation (Nightingale, et al., 2000) ~~is~~ are consistent with measurements (including aerosol sulfate) and do ~~esn't~~ not benefit from being scaled. Furthermore, ACE-SPACE measurements are consistent with ~~less~~ less-efficient aerosol scavenging (55% likelihood of Rain_Frac, the parameter that controls the fractional area~~proportion~~ of the cloudy part of model grid boxes where rain occurs, being below the unconstrained median value 0.5) and less aqueous phase sulfate production (pH of cloud droplets has a 62% likelihood of being lower than the unconstrained median value). These combined constraints suggest, in agreement with sea spray and deposition parameter constraints, higher aerosol number and mass concentrations are consistent with measurements.

The effects of measurement constraint on pdfs of RF_{aci} and ERF_{aci} are shown in Fig. 4. Removing implausible model variants has reduced the uncertainty in several parameters including natural aerosol emission fluxes, which translates into a reduction in RF_{aci} uncertainty (Carslaw et al., 2013). The measurement constraints have two important effects on aerosol forcing. Firstly, the magnitude of median RF_{aci} weakens from -1.99 W m^{-2} to -1.88 W m^{-2} (-1.64 to -1.49 W m^{-2} for ERF_{aci}). A weaker forcing is consistent with higher natural aerosol emissions ~~and~~, increased aerosol load and higher cloud droplet number concentrations in the early-industrial period. Table 1 shows that our constraint on natural emission parameters also constrains Southern Ocean cloud droplet number concentrations towards higher values, reducing the credible interval by around 50% and bringing mean values into closer agreement with MODerate Imaging Spectroradiometer (MODIS; Salomonson et al., 1989) instrument data (note that droplet number concentrations were not used to constrain the model). Thus, we conclude that the constraint on aerosol forcing towards weaker values is a genuine constraint and not the result of an arbitrary tuning. Secondly, the constrained forcing pdfs are approximately symmetric but have shorter tails (lower kurtosis). This suggests the constraints are selectively ruling out model variants that are outliers. The 95% credible range of RF_{aci} values is reduced by around 9% (from -2.84 to -1.15 W m^{-2} down to -2.64 to -1.10 W m^{-2}) and around 9% for ERF_{aci} (from -2.69 to -0.62 W m^{-2} down to -2.43 to -0.54 W m^{-2}). The consistency of forcing constraint across two distinct PPEs suggests the results are insensitive to differences in meteorology, parameters perturbed in the PPEs, and the inclusion of rapid atmospheric adjustments. These results are also insensitive to additional constraint to ensure energy balance at the top of the atmosphere (Fig. S5~~2~~).

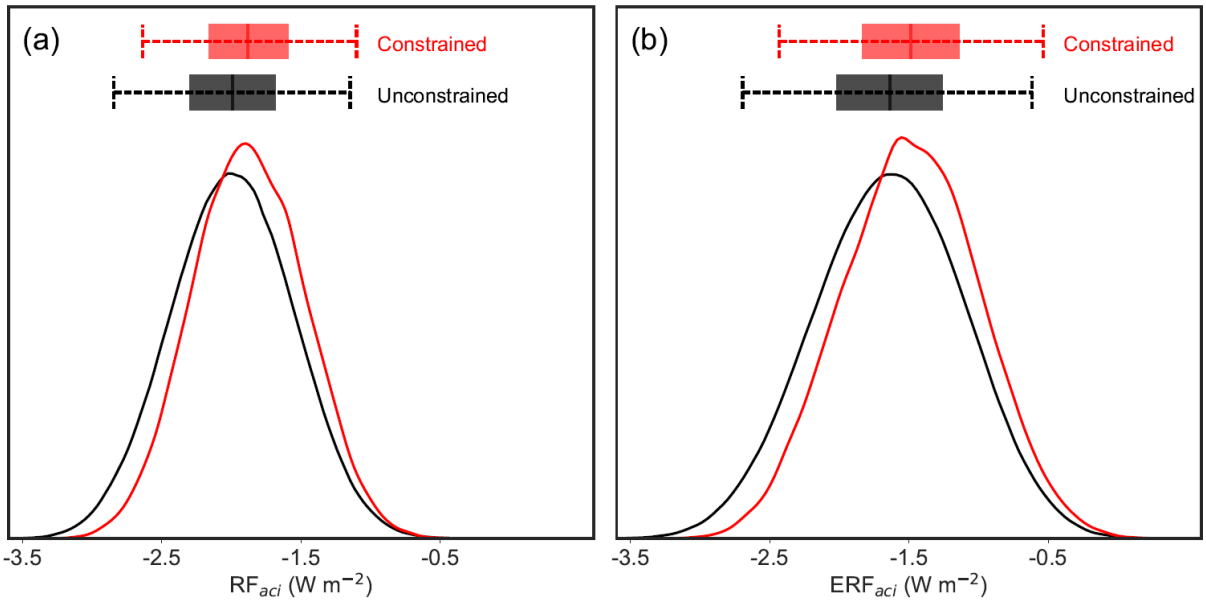

Fig. 4. Probability distributions of a) RF_{aci} and b) ERF_{aci}. The distributions of the unconstrained sample of one million model variants from statistical emulators of each PPE are in black. Red lines show the distributions after constraint using ACE-SPACE measurements (around 3% of the unconstrained sample). The 25th, 50th and 75th percentiles of each sample are shown as shaded boxes and dashed lines span the 2.5th and 97.5th percentiles.

Table 1. Annual and monthly mean cloud drop number concentrations over the Southern Ocean (over the region between 50°S and 60°S at around 1km altitude above sea level) in the original unconstrained sample and the sample of model variants constrained to ACESPACE campaign measurements. Mean values and 95% credible interval values are shown for each sample, with interquartile ranges in brackets. For comparison, we show cloud drop concentrations calculated from MODIS instrument data following Grosvenor et al., (2018) for the year 2008 (SI Methods: Measurements).

	Annual	December	January	February	March	April
MODIS (cm^{-3})	73	89	91	90	82	63
Unconstrained mean (cm^{-3})	38	39	39	41	42	39
Unconstrained credible interval (cm^{-3})	7-125 (112)	8-115 (103)	8-117 (109)	7-122 (115)	7-129 (122)	7-118 (111)
Constrained mean (cm^{-3})	66	67	69	72	76	70
Constrained credible interval (cm^{-3})	41-96 (55)	43-96 (53)	44-99 (55)	45-105 (60)	47-111 (64)	44-101 (57)

Johnson et al. (2019) reduced the global, annual mean RF_{aci} uncertainty by constraining multiple anthropogenic emission and model process parameters (as well as some natural aerosol parameters) using over 9000 predominantly Northern Hemisphere measurements of aerosol optical depth, $PM_{2.5}$, particle number concentrations and mass concentrations of organic carbon and sulfate. We used the same methodology as Johnson et al. (2019) to rule out implausible model variants from the same original sample of one million model variants, so we can readily combine these constraints. Around 700 model variants (0.07%) are observationally plausible in both the Southern Ocean (ACE-SPACE) and Johnson et al. (2019) constraints. Although this is a relatively small percentage of the original sample, 700 observationally-plausible model variants is far more than are typically used to quantify model uncertainty or multi-model diversity (e.g. around 30 for CMIP6). The marginal parameter pdfs from this 700-member sample are shown in Fig. 5. Because Johnson et al. (2019) studied only the AER PPE (from which RF_{aci} can be computed) we are unable to explore the effect of the combined constraint on ERF_{aci}.

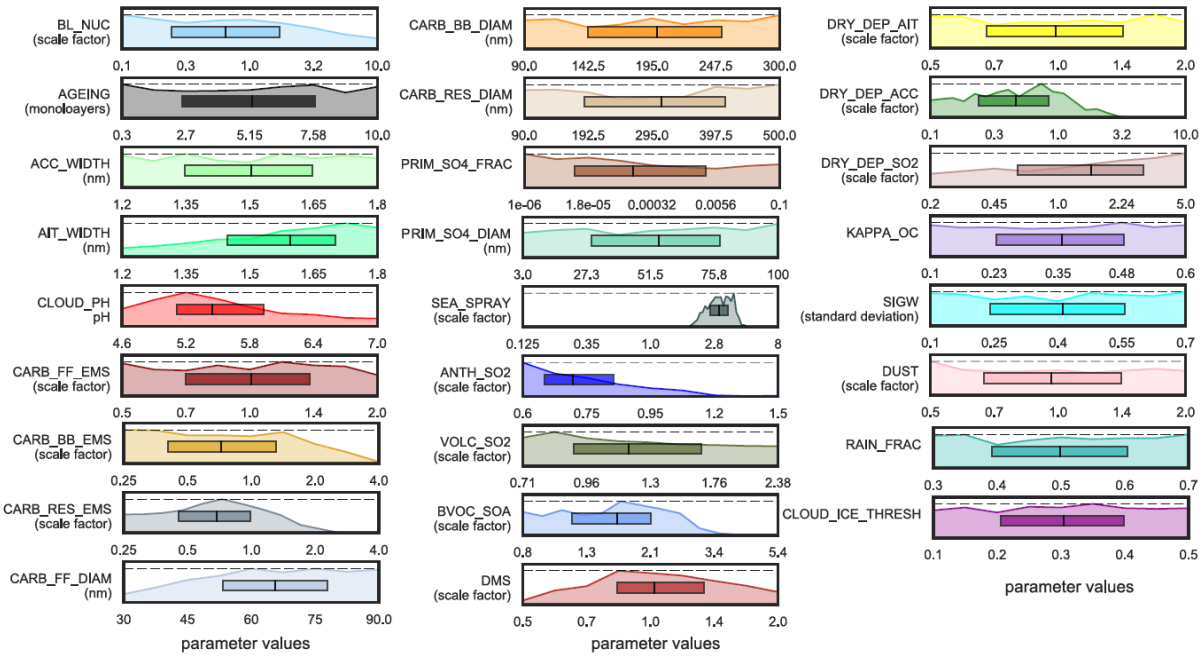

Fig. 5. Marginal probability distributions for the 26 aerosol parameters after constraint using around 250 Southern Ocean measurements and more than 9000 aerosol measurements in Johnson et al. (2019). Plotting features of this figure are identical to Fig. 3.

The two measurement datasets constrain distinct groups of parameters. There are a few cases where the same parameters are constrained by both datasets and in these cases the parameter values are constrained consistently (e.g. cloud droplet pH) or more strongly through ACE-SPACE (e.g. sea spray emissions). The complementary nature of these constraints means that the combined constraint marginal parameter pdfs (Fig. 5) are remarkably similar to those in our Fig. 3e (for sea spray and DMS emission fluxes, as well as deposition and pH parameters) and in figure 6 of Johnson et al. (2019) for other parameters.

The Johnson et al. (2019) constraint reduced the RF_{aci} uncertainty by around 6% and our ACE-SPACE measurement constraint reduced the uncertainty by around 9%. However, the RF_{aci} uncertainty is reduced by around 21% (Fig. 6a) after applying both constraints, meaning the combined constraint is stronger than the sum of individual constraints.

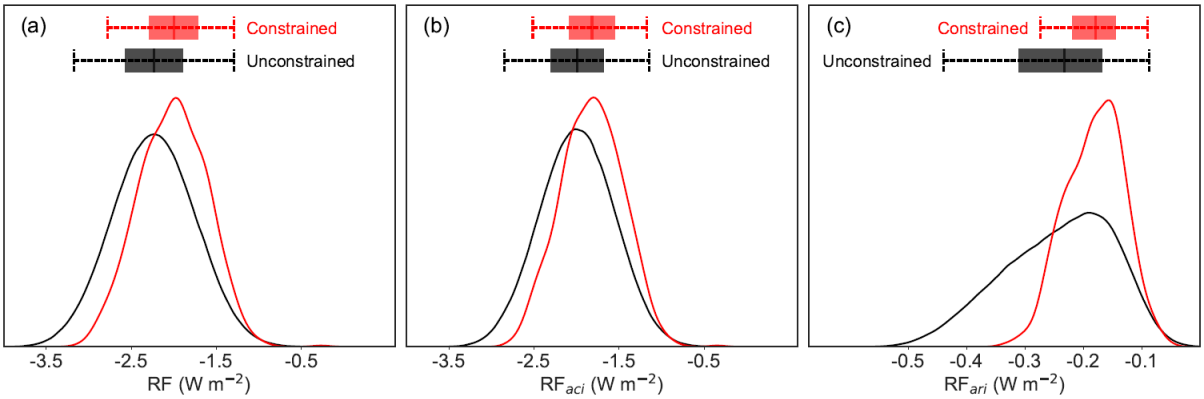

Fig. 6. Probability distributions of a) RF, b) RF_{aci} and c) RF_{ari} from the unconstrained (black line) and constrained (red line) samples of model variants. The constrained sample includes model variants that agree with our ACE-SPACE measurement constraint and the Johnson et al. (2019) constraint. Plotting features are identical to Fig. 4.

The Johnson et al. (2019) constraint strengthened the RF_{aci} by around 0.3 W m^{-2} (more negative) because the largest sea spray emission flux scaling and largest new particle formation rates were ruled out (Fig. 6 in Johnson

et al., 2019). Our ACE-SPACE constraint rules out the same large sea spray emission fluxes, but also rules out all emission flux scale factors lower than around 1.6 (Fig. 3), which increases the baseline aerosol concentration in the early-industrial atmosphere. The ACE-SPACE measurements also constrain several other parameters that collectively ~~weaken RF$_{aci}$~~ weaken the median RF$_{aci}$ by around 0.18 W m^{-2}. Therefore, using the combined measurement dataset, the ~~highest and lowest~~strongest RF$_{aci}$ values have been ruled out as implausible and the credible range of observationally plausible RF$_{aci}$ values is reduced to around -2.51 to -1.17 W m^{-2} (-2.18 to -1.46 W m^{-2} , when using one standard deviation to quantify the uncertainty). Uncertainty in RF$_{ari}$ is reduced by around 48% with observationally plausible values ranging from -0.27 to -0.09 W m^{-2} (-0.23 to -0.13 W m^{-2}, when using one standard deviation), because the strongest RF$_{ari}$ values are ruled out as observationally implausible.

3 Discussion

Our results show, as hypothesised from previous sensitivity analyses, that remote marine measurements are valuable for constraining the natural aerosol state of the atmosphere (Carslaw et al., 2013; Regayre et al., 2014; Regayre et al., 2018). Remote marine aerosol measurements provide new information about plausible model behaviour because they are closely related to model emissions and processes that measurements in polluted environments do not constrain.

For the first time we have achieved a meaningful reduction of 21% in the RF$_{aci}$ uncertainty by constraining the aerosol properties in the model. The reduction in forcing uncertainty can still be improved by considering the following: Firstly, using measurements of cloud properties and cloud-aerosol relations, as well as measurements associated with primary sulfate and carbonaceous particle emission sizes, could constrain model parameters that ~~there are several~~ cause~~s of~~ RF$_{aci}$ uncertainty ~~that~~ but are not constrained by a combination of Northern Hemisphere and pristine Southern Ocean measurements. ~~Identifying measurements associated with primary particle emission diameters (BB_diam and Prim_SO4_diam), Aitken mode aerosol removal rates (Dry_Dep_Ait) and model process parameters related to cloud droplet activation (Kappa_OC, Ait_width, Sig_W) and using them as additional constraints should further reduce the forcing uncertainty.~~ Secondly, even within the considerably reduced volume of multi-dimensional parameter space there still exist many compensating parameter effects (Fig. S3), which limit the constraint on individual parameter ranges (Lee et al., 2016; Regayre et al., 2018). The impact of these compensating effects could be greatly reduced by perturbing uncertain emissions regionally rather than globally as we do here.

Our results are based on uncertainty in a single climate model. The model is structurally consistent in our experiments, so neglects uncertainty caused by choice of microphysical and atmospheric process representations. Our model also neglects some potentially important sources of remote marine aerosol, such as primary marine organic aerosol (Mulcahy et al., 2020) and methane-sulfonic acid (Schmale et al., 2019; Hodshire, et al., 2019; Revell et al., 2019). Model inter-comparison projects (such as CMIP6) can be used to quantify the diversity of RF (or ERF) output from models, but they lack information about single model uncertainty. Ideally, multi-model ensembles would contain a perturbed parameter component, so that model diversity and single model forcing uncertainty could be quantified simultaneously. But,~~but the~~ computational cost~~s~~ prevent~~s~~ many modelling groups from engaging with this important aspect of uncertainty quantification, limiting our shared knowledge about the causes of aerosol forcing uncertainty. Studies like ours that quantify the remaining uncertainty in aerosol forcing and its components after constraint using multiple measurement types fill an important knowledge gap. This knowledge can be used to form a more complete understanding of the importance of historical and near-term aerosol radiative forcing which would reduce the diversity in equilibrium climate sensitivity across models.

Data availability

The ACE-SPACE data are accessible from: https://zenodo.org/communities/spi-ace. The basis for our cloud droplet number concentration data are available from http://catelogue.ceda.ac.uk/uuid/cf97ccc802d348ec8a3b6f2995dfbbff. Simulation output data ~~in~~ for both AER

and AER-ATM PPEs are available on the JASMIN data infrastructure (http://www.jasmin.ac.uk). Some of the climate-relevant fields are derived and stored in netCDF files (.nc) containing data for all ensemble members and made available as a community research tool as described in Yoshioka et al. (2019). Model data and analysis code can be made available from the corresponding author upon request.

Author Contribution

LR applied the statistical methodology and generated results. LR and MY created the PPEs. LR and JJ designed the experiments and elicited probability density functions of all aerosol parameters. KC and MY participated in the formal elicitation process. JS, AB, MG, CT, SH and FS collected and processed the ACE-SPACE measurements. DG processed the cloud droplet number concentration data. LR, KS, JS and JJ analysed the results. LR and KS wrote the manuscript with contributions from all authors.

Competing Interests

Author KC is an executive editor of ACP.

Acknowledgements

We acknowledge funding from NERC under grants AEROS, ACID-PRUF, GASSP and A-CURE (NE/G006172/1, NE/I020059/1, NE/J024252/1 and NE/P013406/1) and the European Union ACTRIS-2 project under grant 262254. This work and its contributors (LR, JJ and KC) were supported by the UK-China Research ~~\&~~ Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund. MY and KS received funding from the National Centre for Atmospheric Science (NCAS), one of the UK Natural Environment Research Council (NERC) research centres via the ACSIS long-term science programme on the Atlantic climate system. LR was funded by a Natural Environment Research Council (NERC) Doctoral Training Grant, and a CASE studentship with the UK Met Office Hadley Centre. KC ~~is currently~~was a Royal Society Wolfson Merit Award holder during this research. ACE-SPACE, JS, SH and AB received funding from EPFL, the Swiss Polar Institute and Ferring Pharmaceuticals. ACE-SPACE was carried out with additional support from the European FP7 project BACCHUS (grant agreement no. 49603445). CT was supported by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the priority programme "Antarctic Research with comparative investigations in Arctic ice areas" SPP 1158 (grant STR 453/12-1). AB received funding from the Swiss National Science Foundation grant No. 200021_169090. DG was funded by the NERC-funded ACSIS programme via NCAS. This work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk). ARCHER project allocations n02-chem, n02-NEJ024252, n02-FREEPPE and the Leadership Project allocation n02-CCPPE were used to perform sensitivity tests and create the ensembles. We thank Andre Welti and Markus Hartmann for CCN measurement support provided during the ACE-SPACE campaign.

References

Andreae, M. O., Jones, C. D., and Cox, P. M.: Strong present-day aerosol cooling implies a hot future, Nat., 435, 1187–1190, doi:10.1038/nature03671, 2005.

Bennartz, R. and Rausch, J.: Global and regional estimates of warm cloud droplet concentration based on 13 years of AQUA-MODIS observations, Atmos. Chem. Phys., 17, 9815-9836, doi:10.5194/acp-17-9815-2017, 2017.

Carslaw, K. S., Boucher, O., Spracklen, D. V., Mann, G. W., Rae, J. G. L., Woodward, S., and Kulmala, M.: A review of natural aerosol interactions and feedbacks within the Earth system, Atmos. Chem. Phys., 10, 1701–1737, 10.5194/acp-10-1701-2010, 2010.

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., Regayre, L. A., and Pierce, J. R.: Large contribution of natural aerosols to uncertainty in indirect forcing, Nat., 503, 67–71, doi:10.1038/nature12674, 2013.

Carslaw, K. S., Gordon, H., Hamilton, D. S., Johnson, J. S., Regayre, L. A., and Yoshioka, M.: Aerosols in the pre-industrial atmosphere, Curr. Clim. Change Rep., 3, 1–15, doi:10.101007/s40641-017-0061-2, 2017.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J. L., Fichefet, D., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A. J., and Wehner, M.: Long-term Climate Change: Projections Commitments and Irreversibility, 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, edited by Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Gantt, B., Meskhidze, N., Facchini, M.C., Rinali, M., Ceburnis, D. and O'Dowd, C.D.: Wind speed dependent size-resolved parameterization for the organic mass fraction of sea spray aerosol, Atmos. Chem. Phys., 11, 8777-8790, doi:10.5194/acp-11-8777-2011, 2011.

Gong, S. L.: A parameterization of sea-salt aerosol source function for sub- and super-micron particles, Glob. Biogeochem. Cyc., 17, 1097, doi:10.1029/2003GB002079, 2003.

Grosvenor, D.P., Sourdeval, O., Zuidema, P., Ackerman, A., Alexandrov, M.D., Bennartz, R., Boers, R., Cairns, B., Chiu, J.C., Christensen, M., Deneke, H., Diamond, M., Feingold, G., Fridlind, A., Hünerbein, A., Knist, C., Kollias, P., Marshak, A., McCoy, D., Merk, D., Painemal, D., Rausch, J., Rosenfeld, D., Russchenberg, H., Seifert, P., Sinclair, K., Stier, P., van Diedenhoven, B., Wendisch, M.,, Werner, F., Wood, R., Zhang, Z. and Quaas, J.: Remote sensing of droplet number concentration in warm clouds: A review of the current state of knowledge and perspectives, Rev. Geophys., 56, 409-453, doi:10.1029/2017RG000593, 2018.

Gryspeerdt, E., Quaas, J., Ferrachat, S., Gettelman, A., Ghan, S., Lohmann, U., Morrison, H., Neubauer, D., Partridge, D. G., Stier, P., Takemura, T., Wang, H., Wang, M., and Zhang, K.: Constraining the instantaneous aerosol influence on cloud albedo, Proc. Natl. Acad. Sci., 114, 4899–4904, doi:10.1073/pnas.1617765114, 2017.

Hamilton, D. S., Lee, L. A., Pringle, K. J., Reddington, C. L. S., Spracklen, D. V., and Carslaw, K. S.: Occurrence of pristine aerosol on a polluted planet, Proc. Natl. Acad. Sci., 111, 18 466–18 471, doi:10.1073/pnas.1415440111, 2014.

Hewitt, H. T., Copsey, D., Culverwell, I. D., Harris, C. M., Hill, R. S. R., Keen, A. B., McLaren, A. J., and Hunke, E. C.: Design and implementation of the infrastructure of HadGEM3: the next-generation Met Office climate modelling system, Geosci. Mod. Dev., 4,223–253, doi:10.5194/gmd-4-223-2011, 2011.

Hodshire, A. L., Campuzano-Jost, P., Kodros, J. K., Croft, B., Nault, B. A., Schroder, J. C., Jimenez, J. L. and Pierce, J. R.: The potential role of methanesulfonic acid (MSA) in aerosol formation and growth and the associated radiative forcings, Atmos. Chem. Phys., 19, 3137-316-, doi:10.5194/acp-19-3137-2019, 2019.

Johnson, J. S., Regayre, L. A., Yoshioka, M., Pringle, K. J., Turnock, S. T., Browse, J., and Carslaw, K. S.: Robust observational constraint of processes and emissions in a climate model and the effect on aerosol radiative forcing, Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2019-834, in review, 2019.

Kettle, A. J. and Andreae, M. O.: Flux of dimethylsulfide from the oceans. A comparison of updated data sets and flux models, J. Geophys. Res., 105 (D22), 26793-26808, doi:10.1029/2000JD900252, 2000.

Lee, L. A., Reddington, C. L., and Carslaw, K. S.: On the relationship between aerosol model uncertainty and radiative forcing uncertainty, Proc. Natl. Acad. Sci., 113, 5820–5827, doi:10.1073/pnas.1507050113, 2016.

Madry, W. L., Toon, O. B. and O'Dowd, C. D.: Modeled optical thickness of sea-salt aerosol, J. Geophys. Res., 116, doi:10.1029/2010jd014691, 2011.

McCoy, D. T., Bender, F. A., Mohrmann, J. K. C., Hartmann, D. L., Wood, R., and Grosvenor, D. P.: The global aerosol-cloud first indirect effect estimated using MODIS, MERRA, and AeroCom, J. Geophy. Res. Atm., 122, 1779–1796, doi:10.1002/2016JD026141, 2017.

Myhre, G., Shindell, D., Bréon, F. M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J. F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., and Zhang, H.: Anthropogenic and Natural Radiative Forcing, 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, edited by Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Mulcahy, J., Johnson, C., Jones, C. G., Povey, A. C., Scott, C. E., Sellar, A., Turnock, S. T., Woodhouse, M. T., Abraham, N. L., Andrews, M. B., Bellouin, N., Browse, J., Carslaw, K. S., Dalvi, M., Folberth, G. A., Glover, M., Grosvenor, D., Hardacre, C., Hill, R., Johnson, B., Jones, A., Kipling, Z., Mann, G., Mollard, J., O'Connor, F. M., Palmieri, J., Reddington, C., Rumbold, S. T., Richardson, M., Schutgens, N. A. J., 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. Discuss., doi:10.5194/gmd-2019-357, 2020.

Nightingale, P. D., Liss, P. S. and Schlosser, P.: Measurements of air-sea gas transfer during an open ocean algal bloom. Geophys. Res. Lett., 27, 2117–2120, doi:10.1029/2000GL011541, 2000.

Regayre, L. A., Pringle, K. J., Booth, B. B. B., Lee, L. A., Mann, G. W., Browse, J., Woodhouse, M. T., Rap, A., Reddington, C. L. S., and Carslaw, K. S.: Uncertainty in the magnitude of aerosol-cloud radiative forcing over recent decades, Geophys. Res. Lett., 41, 9040–9049, doi:10.1002/2014GL062029, 2014.

Regayre, L. A., Johnson, J. S., Yoshioka, M., Pringle, K. J., H.Sexton, D. M., Booth, B. B. B., Lee, L. A., Bellouin, N., and Carslaw, K. S.: Aerosols and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF, ACP, 18, 9975–10 006, doi:10.5194/acp-18-9975-2018, 2018.

Revell, L. E., Kremser, S., M. Harvey, S. H., 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. ~~Discuss.~~, 19, doi:10.5194/acp-19-15447-2019~~10.5194/acp-2019-629~~, 2019.

Salomonson, V. V., Barnes, W. L., Maymon, P. W., Montgomery, H. E. and Ostrow, H.: MODIS: advanced facility instrument for studies of the Earth as a system, in: IEEE Transactions on geoscience and remote sensing, 27, 2, 145-153, doi:10.1109/36.20292, 1989.

Schmale, J., Baccarini, A., Thurnherr, I., Henning, S., Efraim, A., Regayre, L. A., Bolas, C., Hartmann, M.,Welti, A., Lehtipalo, K., Aemisegger, F., Tatzelt, C., Landwehr, S., Modini, R., Tummon, F., Johnson, J. S., Harris, N., Schnaiter, M., Toffoli, A., Derkani, M., Bukowiecki, N., Stratmann, F., Dommen, J., Baltensperger, U., Wernli, H., Rosenfeld, D., Gysel-Beer, M., and Carslaw, K.: Overview of the Antarctic Circumnavigation Expedition: Study of Preindustrial-like Aerosols and Their Climate Effects (ACE-SPACE), Bull. Amer. Meteorol. Soc., ~~Early Online Release~~100, 11, 2260-2283, doi:10.1175/BAMS-D-18-0187.1, 2019.

Schmale, J., Henning, S., Tummon, F., Hartmann, M., Baccarini, A., Welti, A., Lehtipalo, K., Tatzelt, C., Landwehr, S. and Gysel-Beer, M.: Course mode aerosol particle size distribution collected in the Southern Ocean in the austral summer of 2016/2017, during the Antarctic Circumnavigation Expedition, Version 1.0 Dataset, doi:10.5281/zenodo.2636709, 2019a.

Seinfeld, J. H., Bretherton, C., Carslaw, K. S., Coe, H., DeMott, P. J., Dunlea, E. J., Feingold, G., Ghan, S., Guenther, A. B., Kahn, R., Kraucunas, I., Kreidenweis, S. M., Molina, M. J., Nenes, A., Penner, J. E., Prather, K. A., Ramanathan, V., Ramaswamy, V., Rasch, P. J., Ravishankara, A. R., Rosenfeld, D., Stephens, G., and Wood, R.: Improving our fundamental understanding of the role of aerosol-cloud interactions in the climate system, Proc. Natl. Acad. Sci., 113, 5781–5790, doi:10.1073/pnas.1514043113, 2016.

Spracklen, D. V. and Rap, A.: Natural aerosol-climate feedbacks suppressed by anthropogenic aerosol, Geophys. Res. Lett., 40, 5316–5319, doi:10.1002/2013GL057966, 2013.

Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M.: Summary for Policymakers, 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, edited by Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Tatzelt, C., Henning, S., Tummon, F., Hartmann, M., Baccarini, A., Welti, A., Lehtipalo, K. and Schmale, J.: Cloud Condensation Nuclei number concentrations over the Southern Ocean during the austral summer of 2016/2017, Version 1.0 Data set, doi:10.5281/zenodo.2636765, 2019.

Tett, S. F. B., Rowlands, D. J., Mineter, M. J., and Cartis, C.: Can Top-Of-Atmosphere Radiation Measurements Constrain Climate Predictions? Part II: Climate Sensitivity, J. Climate., 26, 9367–9383, doi:10.1175/JCLI-D-12-00596.1, 2013.

Wilcox, L. J., Highwood, E. J., Booth, B. B. B., and Carslaw, K. S.: Quantifying sources of inter-model diversity in the cloud albedo effect, Geophys. Res. Lett., 42, 1568–1575, doi:10.1002/2015GL063301, 2015.

Yoshioka, M., Regayre, L. A., Pringle, K. J., Johnson, J. S., Mann, G. W., Partridge, D., Stier, P., Kipling, Z., Bellouin, N., Sexton, D. M. H ., Lister, G. M. S., Browse, J., Booth, B. B. B., Johnson, C. E., Johnson, B., Mollard, J. D. P., and Carslaw, K. S.: Ensembles of global climate model variants for the quantification and constraint of uncertainty in aerosols and their radiative forcing, J. Adv. Model. Earth Syst., 11, 3728-3754~~Early Online Release~~, doi:10.1029/2019MS001628, 2019.

SUPPLEMENTARY INFORMATION FOR:

The value of remote marine aerosol measurements for constraining radiative forcing uncertainty

Leighton A. Regayre[1], Julia Schmale[2,3], Jill S. Johnson[1], Christian Tatzelt[34], Andrea Baccarini[2], Silvia Henning[34], Masaru Yoshioka[1], Frank Stratmann[34], Martin Gysel-Beer[2], Daniel P. Grosvenor[1,5] and Ken S. Carslaw[1]

[1]Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds, LS2 9JT, UK
[2]Paul Scherrer Institute, Laboratory of Atmospheric Chemistry, Villigen, Switzerland
[3]École Polytechnique Fédéderale de Lausanne, Lausanne, Switzerland
[43]Leibniz Institute for Tropospheric Research, Leipzig, Germany
[5]National Centre for Atmospheric Science, Leeds, UK

Correspondence to: Leighton Regayre (L.A.Regayre@leeds.ac.uk)
Correspondence related to measurements to: Julia Schmale (julia.schmale@psi.ch)

SI Methods

SI Methods: Model Version

We use the Global Atmosphere 4 (GA 4.0; Walters et al., 2014) configuration of the Hadley Centre General Environment Model version 3 (HadGEM3; Hewitt et al., 2011), which incorporates the UK Chemistry and Aerosol (UKCA) model at version 8.4 of the UK Met Office's Unified Model (UM). UKCA simulates trace gas chemistry and the evolution of the aerosol particle size distribution and chemical composition using the GLObal Model of Aerosol Processes (GLOMAP-mode; Mann et al., 2010) and a whole-atmosphere chemistry scheme (Morgenstern et al., 2009; O'Connor et al., 2014). The model has a horizontal resolution of 1.25x1.875 degrees and 85 vertical levels. The aerosol size distribution is defined by seven log-normal modes: one soluble nucleation mode as well as soluble and insoluble Aitken, accumulation and coarse modes. The aerosol chemical components are sulfate, sea salt, black carbon (BC), organic carbon (OC) and dust. Secondary organic aerosol (SOA) material is produced from the first stage oxidation products of biogenic monoterpenes under the assumption of zero vapour pressure and is combined with primary particulate organic matter after kinetic condensation. Use of the GLOMAP model to simulate aerosol size and composition changes reduces Southern Ocean radiative biases in HadGEM3 (Bodas-Salcedo et al., 2019).

GLOMAP simulates new particle formation, coagulation, gas-to-particle transfer, cloud processing and deposition of gases and aerosols. The activation of aerosols into cloud droplets is calculated using globally prescribed distributions of sub-grid vertical velocities (West et al. 2014) and the removal of cloud droplets by autoconversion to rain is calculated by the host model. Aerosols are also removed by impaction scavenging of falling raindrops according to the collocation of clouds and precipitation (Lebsock et al., 2013; Boutle et al., 2014). Aerosol water uptake efficiency is determined by kappa-Kohler theory (Petters and Kreidenweis, 2007) using composition-dependent hygroscopicity factors.

We prescribe anthropogenic emissions using the emission inventory prepared for the Atmospheric Chemistry and Climate Model Inter-comparison Project (ACCMIP) and also prescribed in some of the CMIP Phase 5 experiments. Present-day carbonaceous aerosol emissions were prescribed using a ten year average of 2002 to 2011 monthly mean data from the Global Fire and Emissions Database (GFED3; van der Werf et al., 2010) and according to Lamarque et al. (2010) for 1850. We prescribe volcanic SO_2 ~~SO2~~ emissions for continuously emitting and sporadically erupting volcanoes (Andres et al., 1998) and for explosive volcanic eruptions (Halmer et al., 2002). Surface ocean dimethyl-sulfide concentrations are prescribed using Kettle and Andreae (2000) and emitted into the atmosphere using a surface wind speed dependent parametrisation (following Nightingale et al., 2000). Sea spray is emitted into the atmosphere using the Gong (2003) surface wind speed dependent parametrisation.

Several modifications were made to version 8.4 of UKCA to overcome known structural deficiencies in the model. An organically-mediated boundary layer nucleation parametrisation (Metzger et al., 2010) was included so that remote marine and early-industrial aerosol concentrations were not unrealistically low in the model. We also added a parametrisation for ice crystal suppression of precipitation known to bring remote marine aerosol concentrations in line with measurements (Browse et al., 2012). Dust in the base model is calculated using the CLASSIC bin scheme (Woodward et al., 2001), which we replaced in our model version so that dust is emitted using the GLOMAP modal scheme. This means interactions between dust and other aerosols are explicitly simulated. We better resolve the optical properties of aerosols across wavelengths by improving the resolution of the default look-up tables. Finally, we made minor adjustments to some process parametrisations so that parameter values could be perturbed globally. All changes to the model are described fully in Yoshioka et al. (2019).

SI Methods: Perturbed Parameter Ensembles

We make use of the ~~ATM-~~AER and AER-ATM perturbed parameter ensembles (PPEs) described in Yoshioka et al. (2019). Results in the main article make use of the AER PPE except for the quantification of aerosol ERF and its components. These two PPEs were designed to provide complementary insights into causes of uncertainty in the climate system. The 235 member AER PPE samples uncertainties in a set of 26 aerosol parameters, whilst the 191 member AER-ATM PPE samples uncertainties in 18 aerosol and 9 physical atmosphere parameters related to clouds, radiation and moisture. The effects of rapid atmospheric adjustments to aerosols are not included in AER, but are included in AER-ATM (although they have a relatively minor impact on aerosol forcing in this model (e.g. Mulcahy et al., 2018). Therefore, ERF is calculated for the AER-ATM PPE and combined (in the "SI Results: Additional constraint to achieve radiative balance" section) with the CERES top-of-the-atmosphere constraint employed in Regayre et al. (2018), whilst RF is calculated for the AER PPE and combined (in the main article) with the predominantly Northern Hemisphere aerosol constraint employed in Johnson et al. (2019).

Both PPEs were nudged towards European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim reanalyses. Nudging means that pairs of simulations have near-identical synoptic-scale features, which enables the effects of parameter perturbations to be quantified using single-year simulations, although the magnitude of forcing will vary with the chosen year (Yoshioka et al., 2019; Fiedler et al., 2019). We nudge well above the Earth's surface in order to strike a balance between the computational cost of perturbing multiple parameters and the computational saving of using prescribed meteorology to overcome internal variability (Zhang et al., 2016). In the AER-ATM PPE only horizontal winds above the boundary layer (around 2km) for the year 2006 were prescribed, whilst in AER, horizontal winds and temperatures for 2008 were prescribed above around 1km. In each PPE the model was allowed to respond to parameter perturbations (a spin-up period) prior to simulating the data used here. Despite these differences, results in the main article are consistent across the PPEs.

SI Methods: Sampling and uncertainty

We sample uncertainty in model output using uniform pdfs across each parameter range. The uncertainty in individual parameters could be sampled in a more informed manner. For example, Yoshioka et al. (2019) used expert elicited information about likely parameter values to create parameter pdfs, which were used by Bellouin et al. (2019) and Watson-Paris et al. (2020~~submitted~~) to sample uncertainty in aerosol forcing uncertainty. The additional information provided by expert elicited parameter pdfs is invaluable for quantifying the causes of model uncertainty (e.g. Regayre et al., 2018) because the choice of pdfs affects the contributions to variance in model output. However, in nearly 30 dimensions, samples of combined parameter values using multiple pdfs with centralised tendencies will ~~by~~ be heavily weighted towards the centre of the parameter space. Since our intention in this article is to sample the range of model behaviour in response to the full spectrum of uncertain parameter combinations prior to constraint using measurements, we use uniform pdfs with maximum and minimum values from the expert elicited ranges.

A set of around 200 model variants that make up the PPEs are much too small to allow statistical analysis of model performance across nearly 30 dimensions of parameter space. We therefore use output from the PPEs to train Gaussian Process emulators (e.g. Lee et al., 2012), which define how the model outputs vary continuously over the parameter space. Some additional uncertainty is caused by emulating (rather than simulating) model

output and this uncertainty is incorporated into our model-measurement constraint process (SI Methods: Model-measurement comparisons), despite being much smaller than other sources of uncertainty (Johnson et al., 2019). We sample Monte Carlo points from the emulated parameter space to produce the set of one million model variants.

SI Methods: Measurements

Measurements were collected during the ACE-SPACE campaign between December 2016 and March 2017. The measurement methodology is explained in Schmale et al. (2019) as well as in the metadata of the datasets cited below. We constrain the model uncertainty using near-surface measurements of cloud condensation nuclei concentrations at 0.2% and 1.0% supersaturations ($CCN_{0.2}$ and $CCN_{1.0}$; Tatzelt et al., 2019), as well as number concentrations of particles with dry aerodynamic diameter larger than 700 nm (N_{700}; corresponds to volume equivalent diameter larger than around 500 to 570 nm; Schmale et al., 2019a) and mass concentrations of non-sea-salt sulfate in PM_{10}. We compare simulated and measured $CCN_{0.2}$ concentrations because cloud-active aerosol concentrations are fundamentally important for RF_{aci}. We use $CCN_{1.0}$ measurements to challenge the model's ability to reproduce concentrations of relatively small aerosols that only activate to form cloud droplets at very high supersaturations. We target the highly uncertain sea spray emission flux scaling parameter by comparing concentrations of N_{700} to simulated concentrations of sea spray aerosol, approximated using our model's soluble accumulation and coarse mode aerosol concentrations (Mann et al., 2010). This is not a like-for-like comparison because our soluble accumulation mode includes aerosols with dry diameter larger than 100 nm (Mann et al., 2010; rather than around 500 to 570 nm). Additionally, our soluble accumulation and coarse modes include negligible contributions from sulfate, primary organic matter and aged carbonaceous and dust particles. However, ~~, under the assumption that~~ over the Southern Ocean~~,~~ we think it is safe to assume that sea spray is the predominant (if not only) source of relatively large aerosols. ~~-~~Finally, we compare ~~non-~~non-~~sea-~~sea-salt sulfate concentrations (which omit primary sulfate in sea spray aerosol) in order to constrain the uncertainty in the emission flux of dimethyl-sulphide from the ocean surface. The sea salt fraction of sulfate was calculated using sodium as a tracer for the enrichment of sea salt in the aerosol phase (Sander et al., 2003). Non-sea-salt sulfate was calculated by subtracting this fraction from the total particulate sulfate as detected from ~~PM10~~ PM_{10} filters.

Data for all variables were averaged for comparison with monthly mean model values by taking the mean of all data points that were collected at locations corresponding to positions within model gridboxes. This spatial and temporal degradation introduces representation errors that we account for using our model-measurement comparison (next section). However, the reduction in data volume makes the model-measurement comparison over one million model variants tractable.

We present monthly mean and annual cloud droplet number concentrations in table 1 from the model and from satellite data, over the region between 50ºS and 60ºS. Following Grosvenor et al., (2018), we calculated cloud droplet concentrations from the MODIS (MODerate Imaging Spectroradiometer) Collection 5.1 Joint Level-2 (Aqua satellite) for the year 2008 (to correspond to the meteorological year used in our simulations). Our calculation used cloud optical depth and 3.7 micron effective radius values derived under the adiabatic cloud assumption (essentially, cloud liquid water increases linearly with height, droplet concentrations are constant throughout the cloud and the ratio of volume mean radius to effective radius is constant). We improved the cloud droplet concentration data (Grosvenor et al., 2018b) by excluding 1x1 degree data points for which the maximum sea-ice areal coverage over a moving 2-week window exceeded 0.001%. The sea-ice data used in this process were the daily 1x1 degree version of Cavalieri et al. (2016). As with other data used in our model-measurement comparison, we degraded the cloud droplet number concentration data to the model gridbox and monthly mean spatial and temporal resolutions.

SI Methods: Model-measurement comparisons

Our constraint approach follows Johnson et al. (2019) and involves comparing output from model variants (parameter combinations) to a set of measurements and ruling out variants that are judged to be implausible. This method uses the statistical methodology of history matching, which has been effectively applied to complex models in a range of fields (Craig et al., 1997; Williamson et al., 2013; McNeall et al., 2016; Rodrigues et al., 2017 and Andrianakis et al., 2017). We account for emulator uncertainty, measurement uncertainty (instrument error) and representativeness uncertainties (caused by spatial and temporal mismatches in resolution and sampling between model and measurements). We do not include potential structural errors (e.g. from missing processes) in our constraint approach because such errors cannot be robustly quantified a priori.

For each measurement we calculate a 'measure of implausibility' for each of the one million model variants, calculated as the model-measurement difference standardised by the combined emulator, measurement and representativeness uncertainties. Using this 'implausibility measure' we can identify implausible model variants and rule ~~our~~ out implausible parts of parameter space via the combination of the 'closeness' of the measurement and model output, and the size of the related uncertainties. The 'implausibility metric' is defined as:

$$I(\boldsymbol{x}) = \frac{|\, M - O \,|}{\sqrt{[Var(M) + Var(O) + Var(R)]}}, \quad (1)$$

where *M* is the model variant output and *O* is the observed value (the measurement). In the denominator *Var(M)* is the variance in the model estimate (caused by emulator uncertainty), *Var(O)* is the variance in the measurement (i.e., instrument or retrieval uncertainty) and the representativeness error, *Var(R)*, is the variance associated with comparing model output to measurements at different spatial (Schutgens et al., 2016a; Weigum et al. 2016, Schutgens et al., 2017) and temporal (Schutgens et al., 2016b; Schutgens et al., 2017) resolutions. We compare the 2016-17 measurements to the models nudged towards 2008 meteorology for AER and 2006~~8~~ meteorology for AER-ATM because the measurements were not collected when the PPE was created. The *Var(R)* term therefore includes additional uncertainty due to inter-annual variability. According to the definition of the implausibility measure, model variants will not be ruled out if either the model-measurement difference is small or the uncertainty in the denominator is large. In other words, we retain model variants that are skilful and model variants whose skill cannot be adequately determined because the model-measurement comparison uncertainties are too large.

The variance terms in the denominator of Eq. (1) are calculated uniquely for each measurement. Following Johnson et al.~~,~~ (2019), we use an instrument error ~~measurement uncertainty~~ of 10%, a spatial co-location uncertainty of 20% and a temporal co-location uncertainty of ~~20~~10%. Fig. S1 shows an example of the degradation of data for comparison with monthly mean model output. Emulator uncertainty is calculated for each model-measurement combination using the error on the predicted mean from the emulator for the model variant. We use residuals in de-trended monthly mean output from a HadGEM-UKCA hindcast simulation over the period of 1980-2009 (Turnock et al., 2015) to estimate the inter-annual variability for each variable across all model gridboxes and months.

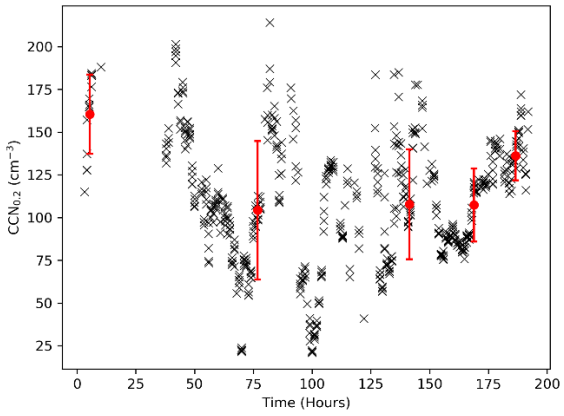

Fig S1: Measured $CCN_{0.2}$ values between the 3rd and 10th January 2017, after filtering for possible ship stack contamination. The ACE-SPACE vessel transited through 5 model gridboxes during this period. We average all measurements collected in locations, over one or more days, within each model gridbox, for comparison with monthly mean model output. These average values and one standard deviation of the measurement data are shown in red at the central time for each measurement subset. From left to right, these values correspond to the five model gridboxes in Fig. 1 between around 60°E and 90°E, at the following latitude and longitudes: 1) 49.5°S, 65.5°E, 2) 49.5°S, 69.5°E, 3) 54°S, 77°E, 4) 54°S, 84.5°E and 5) 56.5°S, 92°E.

We calculate implausibility values for each of the one million model variants for every measurement. Deciding which model variants to retain would be trivial were we comparing the sample output to a single measurement. We would sequentially rule out the variant with the highest implausibility metric until some small fraction of the original sample remained. However, our task is more complex. We need to rule out model variants based on multiple implausibility metrics that are distinct for each measurement location and measurement type.

A variant may compare well with a measurement type in one location and poorly in another because spatial and temporal features in the measurement data (e.g. changing aerosol sources) mean each measurement could provide different information about the plausibility of the models. To avoid prematurely ruling out model variants based on a few poor comparisons, we only rule out variants if their implausibility exceeds a defined threshold for more than a tolerable fraction of measurements. We choose threshold and tolerance values with a goal of retaining around 3% of the original sample. The subjective choice of 3% retention determines the results to some extent. Retaining a much smaller percentage of the model variants could potentially over-constrain the model. However, retaining a larger proportion risks weakening the constraint and retaining addition implausible variants.

We set threshold and tolerance values for each variable distinctly for each month of data. This makes processing the implausibility data more efficient and allows for a degree of automation of the constraint process. We ensure that each measurement type on each leg of the journey (Schmale et al., 2019) affects the combined constraint. This requires quantification of the constraint of individual measurement types on parameter values at multiple combinations of threshold and implausibility exceedance tolerances. We avoid increasing the threshold and/or tolerance values in individual months for each measurement type, if the constraint efficacy of the measurement would saturate as a result. Otherwise, threshold and tolerances for each month are required to be as similar as possible.

Although our analysis in the main article focusses on a combined measurement constraint, this analysis is informed by individual measurement type constraints. The threshold and exceedance tolerances for individual measurement type constraints are summarised in table S1. Only 0.004% of the one million model variants (40 variants) are retained when these individual constraints are combined. Thus, we relax the threshold and tolerance criteria for each measurement type constraint when combining constraints (table S2).

Table S1: Individual measurement type constraint threshold values and exceedance tolerance values for December to April, as well as the percentage of the one million member sample retained by each constraint. Exceedance tolerances values are percentages of the number of measurements in each month.

	$CCN_{0.2}$	$CCN_{1.0}$	Nss-sulfate	N_{700}
Implausibility Threshold	3.5	3.5	3.5	3.5
Exceedance tolerance (%) Dec-Apr	15,15,20,20,10	2,2,2,5,2	15,20,20,15	20,20,25,20,20
Percentage retained	3.3	3.0	6.2	3.0

Table S2: Threshold values and exceedance tolerance values for December to April, as well as the percentage of the one million member sample retained by each constraint. Exceedance tolerances values are percentages of the number of measurements in each month. These constraints are combined to retain around 3% of the one million member sample of model variants, as described in the main article.

	$CCN_{0.2}$	$CCN_{1.0}$	Nss-sulfate	N_{700}
Implausibility Threshold	4.5	4.5	4.0	4.5
Exceedance tolerance (%) Dec-Apr	30,30,30,30,10	25,30,30,15,5	20,20,20,15	25,25,25,30,25
Percentage retained	20.6	18.1	29.9	24.2

SI Results

SI Results: Constrained marginal parameter distributions

In Fig. 3 of the main article we show the marginal probability distributions for the 26 parameters in the AER PPE. These marginal distributions show the effect of measurement constraint on individual parameter likelihoods. ~~Rejecting the majority of the model variants reduces the maximum density, so m~~Marginal densities for the constrained sample are scaled such that the tops of the constrained and unconstrained pdfs are aligned. Similar parameter constraints are found when constraining the AER-ATM PPE using the same constraint

process and original set of measurements (Fig. ~~S1~~S2). In addition to parameters that are perturbed in both PPEs, we show the effect of measurement constraint on the few physical atmosphere parameters (Rad_Mcica_Sigma and Fac_Qsat) that are constrained by our process as well as additional aerosol parameters that were perturbed in AER-ATM (BC_RI and OC_RI).

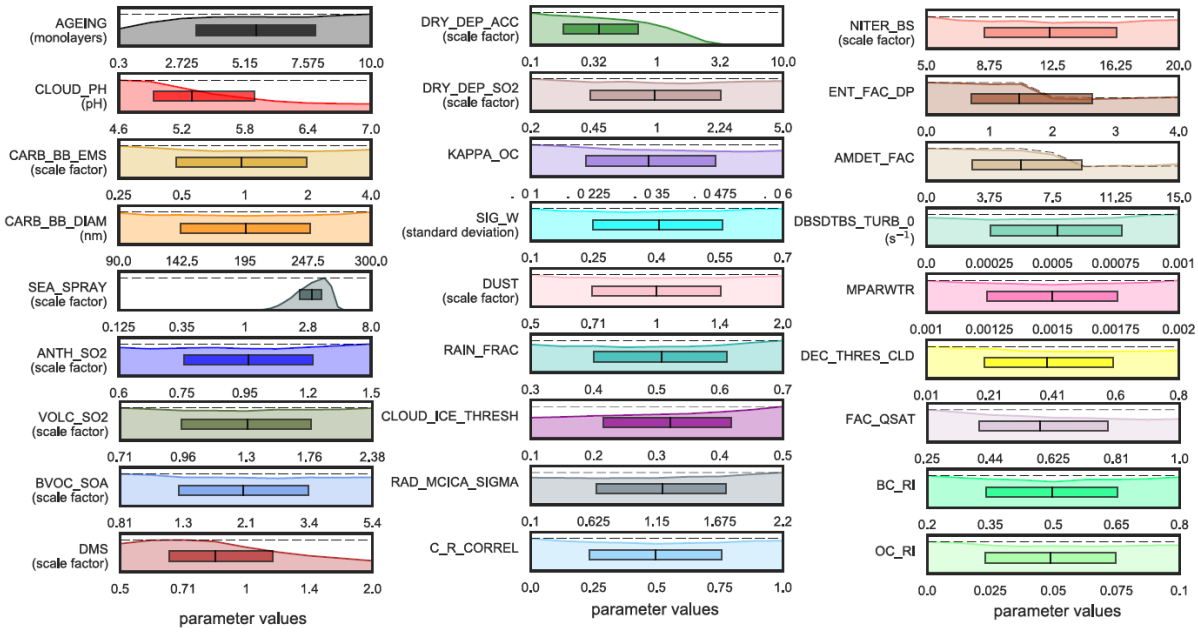

Fig. ~~S1~~S2. Marginal probability distributions for aerosol and physical atmosphere parameters from the AER-ATM PPE after constraint. The density of parameter values in the unconstrained sample are shown as dashed lines. Densities of constrained samples are shown in colour. The 25th, 50th and 75th percentiles of each marginal distribution are shown in the central boxes. Parameter values on the x-axes correspond to values used in the model (Yoshioka et al., 2019).

In addition to the constraint achieved by combining remote marine aerosol measurements, table S3 shows the effect of individual measurement type constraints (table S2) on model parameters and how these translate into a combined constraint (Fig. 3).

Table S3. Ranges and inter-quartile ranges of marginal parameter distributions from individual constraints using measured concentrations of $CCN_{0.2}$, $CCN_{1.0}$, non-sea-salt sulfate and N_{700}, as well as for the combined constraint. These individual constraints are those described in table S2 and were combined to constrain the model and make Fig. 3. Values are marked in bold where the individual measurement type constraint moves the range, 25th or 75th percentile closer towards the range or percentiles of the combined constraint than other measurement types, relative to the unconstrained values.

Parameter Name	Unconstrained	$CCN_{0.2}$	$CCN_{1.0}$	Non-sea-salt sulfate	N_{700}	Combined
BL_Nuc	0.1,10.0 [0.3,3.2]	0.1,10.0 [0.3,**3.5**]	0.1,10.0 [0.3,3.0]	0.1,10.0 [0.3,3.3]	0.1,10.0 [0.3,3.2]	0.1,10.0 [0.3,3.5]
Ageing	0.3,10.0 [2.7,7.6]	0.3,10.0 [3.0,7.9]	0.3,10.0 [2.5,7.5]	0.3,10.0 [2.7,7.6]	0.3,10.0 [2.6,7.5]	0.3,10.0 [2.7,7.6]
Acc_ Width	1.2,1.8 [1.4,1.6]	1.2,1.8 [**1.3**,1.7]	1.2,1.8 [1.4,1.7]	1.2,1.8 [1.4,1.7]	1.2,1.8 [**1.3**,1.7]	1.2,1.8 [1.3,1.7]
Ait_Width	1.2,1.8 [1.3,1.6]	1.2,1.8 [1.3,1.7]	1.2,1.8 [1.3,1.6]	1.2,1.8 [1.3,1.7]	1.2,1.8 [1.3,1.7]	1.2,1.8 [1.3,1.6]
Cloud_pH	4.6,7.0 [5.2,6.4]	4.6,7.0 [**5.1**,6.4]	4.6,7.0 [**5.1**,**6.2**]	4.6,7.0 [5.2,6.4]	4.6,7.0 [5.2,6.4]	4.6,7.0 [5.1,6.2]
Carb_FF_ Ems	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]
Carb_BB_ Ems	0.25,4.00 [0.50,2.00]	0.25,4.00 [0.52,**2.16**]	0.25,4.00 [**0.48**,2.01]	0.25,4.00 [0.50,2.01]	0.25,4.00 [0.49,2.03]	0.25,4.00 [0.49,2.06]
Carb_Res_ Ems	0.25,4.00 [0.50,2.00]	0.25,4.00 [**0.45**,**1.78**]	0.25,4.00 [0.48,2.02]	0.25,4.00 [0.49,2.00]	0.25,4.00 [0.50,2.02]	0.25,4.00 [0.48,1.94]
Carb_FF_	30,90	30,90	30,90	30,90	30,90	30,90

Diam	[45,75]	[45,**76**]	[44,75]	[45,75]	[45,75]	[45,76]
Carb_BB_ Diam	90,300 [143,248]	90,300 [141,**250**]	90,300 [**140**,249]	90,300 [142,248]	90,300 [141,248]	90,300 [141,249]
Carb_Res_ Diam	90,500 [193,398]	90,500 [193,**404**]	90,500 [**190**,399]	90,500 [192,400]	90,500 [193,400]	90,500 [189,400]
Prim_SO4_ Frac	1.0e-6,1.0e-1 [1.8e-5,5.6e-3]	1.0e-6,1.0e-1 [1.7e-5,6.5e-3]	1.0e-6,1.0e-1 [**1.3e-5**,**4.2e-3**]	1.0e-6,1.0e-1 [1.7e-5,5.6e-3]	1.0e-6,1.0e-1 [1.6e-5,6.0e-3]	1.0e-6,1.0e-1 [1.6e-5,5.2e-3]
Prim_SO4_ Diam	3,100 [27,76]	3,100 [26,**75**]	3,100 [**29**,**78**]	3,100 [27,76]	3,100 [26,77]	3,100 [28,77]
Sea_ Spray	0.1,8.0 [0.4,2.8]	1.5,8.0 [2.7,3.8]	**1.9**,8.0 [**3.8**,**5.7**]	0.1,8.0 [0.3,2.8]	1.5,**5.2** [2.5,3.6]	1.6,5.1 [2.6,3.7]
Anth_SO2	0.6,1.5 [0.8,1.2]	0.6,1.5 [0.8,1.2]	0.6,1.5 [0.7,1.2]	0.6,1.5 [0.8,1.2]	0.6,1.5 [0.8,1.2]	0.6,1.5 [0.8,1.2]
Volc_SO2	0.7,2.4 [1.0,1.8]	0.7,2.4 [1.0,1.8]	0.7,2.4 [1.0,1.8]	0.7,2.4 [1.0,1.8]	0.7,2.4 [1.0,1.8]	0.7,2.4 [1.0,1.8]
BVOC_ SOA	0.8,5.4 [1.3,3.4]	0.8,5.4 [1.3,3.5]	0.8,5.4 [1.4,3.5]	0.8,5.4 [1.3,3.4]	0.8,5.4 [1.3,3.4]	0.8,5.4 [1.3,3.4]
DMS	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.5]	0.5,2.0 [0.7,1.4]	0.5,2.0 [**0.8**,1.5]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.8,1.3]
Dry_Dep_ Ait	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.3]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]
Dry_Dep_ Acc	0.1,10.0 [0.3,3.2]	0.1,9.3 [**0.2**,**0.9**]	0.1,**6.7** [**0.2**,1.0]	0.1,10.0 [0.3,1.9]	0.1,10.0 [0.3,3.2]	0.1,6.4 [0.2,0.8]
Dry_Dep_ SO2	0.2,5.0 [0.4,2.2]	0.2,5.0 [0.4,2.2]	0.2,5.0 [0.4,2.4]	0.2,5.0 [0.4,2.2]	0.2,5.0 [0.4,2.2]	0.2,5.0 [0.4,2.2]
Kappa_ OC	0.1,0.6 [0.2,0.5]	0.1,0.6 [0.2,0.5]	0.1,0.6 [0.2,0.5]	0.1,0.6 [0.2,0.5]	0.1,0.6 [0.2,0.5]	0.1,0.6 [0.2,0.5]
Sig_W	0.1,0.7 [0.3,0.5]	0.1,0.7 [0.2,0.6]	0.1,0.7 [0.2,0.6]	0.1,0.7 [0.2,0.6]	0.1,0.7 [0.2,0.6]	0.1,0.7 [0.2,0.6]
Dust	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]	0.5,2.0 [0.7,1.4]
Rain_Frac	0.3,0.7 [0.4,0.6]	0.3,0.7 [0.4,0.6]	0.3,0.7 [0.4,0.6]	0.3,0.7 [0.4,0.6]	0.3,0.7 [0.4,0.6]	0.3,0.7 [0.4,0.6]
Cloud_Ice_ Thresh	0.1,0.5 [0.2,0.4]	0.1,0.5 [0.2,**0.3**]	0.1,0.5 [0.2,0.4]	0.1,0.5 [0.2,0.4]	0.1,0.5 [0.2,0.4]	0.1,0.5 [0.2,0.4]

Constrained marginal parameter distributions in Fig. 3 and Fig. 5 of the main article ~~are~~ tell a one-dimensional story. In Fig. S3, we show the effect of constraint to remote marine aerosol measurements, combined with the constraint from Johnson et al. (2019) on a subset of the marginal 2-dimensional parameter combinations.

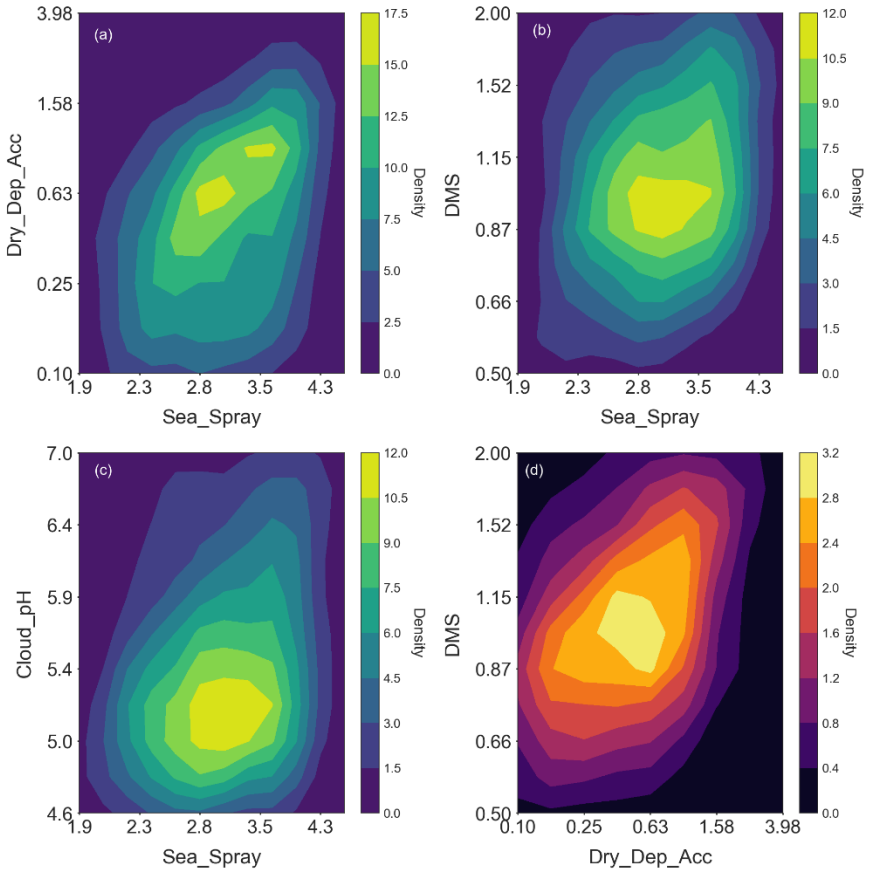

Fig. S3. Two-dimensional marginal probability density distributions for a) sea spray emission flux scale factor (Sea_Spray) and the Accumulation aerosol mode dry deposition velocity scale factor (Dry_Dep_Acc), b) sea spray emission flux scale factor and dimethylsulfide surface water concentration scale factor (DMS), c) sea spray emission flux scale factor and cloud droplet pH (Cloud_pH), and d) Accumulation aerosol mode dry deposition velocity scale factor and dimethylsulfide surface water concentration scale factor. Individual parameter ranges are plotted according to their constrained values (table S3), not the full range of values used in the original sample of model variants as shown in Fig. 3, Fig. 5 and Fig. S2.

SI Results: Wind Speed discrepancies

Southern Ocean wind speeds during the ACE-SPACE expedition were often much lower than climatological mean values, but on average were higher than winds in our ensemble (Schmale et al., 2019). We account for the effects of inter-annual variability in the Var(R) term in equation S1. However, monthly mean differences between ERA-Interim wind speeds in the measurement year and the year used in the ensemble are less than 20% along the route taken by the ACE-SPACE campaign vessel (Fig. S4). The modest discrepancy in wind speeds may be important for constraining aerosol concentrations, because sea spray emissions in our model are strongly dependent on wind speeds (Gong, 2003). However, the measured wind speed and N700 values are only weakly correlated (Pearson correlation coefficient of around 0.2) when degraded to the resolution used for comparison with model output.

Our constraint process has in-built functionality that prevents the use of measurements with large model-measurement discrepancies. We tested the robustness of our constraint methodology to the discrepancy in wind speeds by neglecting around 50% of the measurements (those with the largest discrepancies between measured and AER-ATM PPE mean simulated winds) and repeating the constraint. The effects on marginal parameter and aerosol forcing constraints were negligible (not shown). The consistency of constraint, with and without measurements in locations with relatively large model-measurement wind speed discrepancies, suggests the constraint methodology is insensitive to wind speed discrepancies caused by daily wind speed variability and differences in meteorological years between model simulations and measurements.

~~Wind speed discrepancies do not affect our results, possibly because differences in the resulting wave heights cause compensating effects between sea spray emission fluxes and the removal rate of aerosols by the ocean (Korhonnen et al., 2010). Marginal parameter distributions are constrained consistently when we remove measurements with average wind speed differences larger than 50% of the measured value from the model-measurement comparison~~

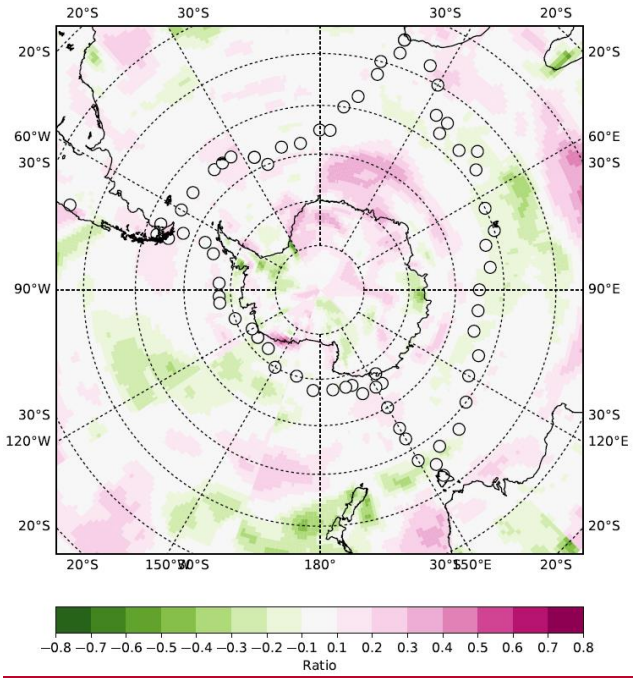

Fig. S4. Ratio of ERA-Interim wind speed differences (between measurement and simulated years) to the measurement year. Monthly mean winds from 2006 (matching the AER PPE) were subtracted from monthly mean winds for December 2016 to April 2017 (matching the ACE-SPACE campaign) to calculate the differences. The map is an assimilation of data between months, where data is presented at each location for months corresponding to the timing of the ACE-SPACE measurement campaign.

~~, possibly because differences in the resulting wave heights cause compensating effects between sea spray emission fluxes and the removal rate of aerosols by the ocean (Korhonnen et al., 2010). Marginal parameter distributions are constrained consistently when we remove measurements with average wind speed differences larger than 50% of the measured value from the model-measurement comparison.~~

SI Results: Additional constraint to achieve radiative balance

We additionally test the effect of ruling out model variants that differ from the Clouds and the Earth's Radiant Energy System (CERES; Loeb et al., 2009) measurement of global, annual mean top-of-the-atmosphere outgoing shortwave radiative flux of ~~98.3~~ 98.3 W m^{-2} by more than 0.25 W m^{-2}, which was the constraint ~~used~~ applied in Regayre et al.~~,~~ (2018). The constraint on ERF using the CERES-derived top-of-the-atmosphere fluxes in addition to the ACE-SPACE measurement dataset weakens the reduction in aerosol ERF from 8% to 7%. Fig. ~~S2~~ S5 (for comparison with Fig. 4a) shows the effect of this additional constraint on aerosol ERF. Retaining only model variants that agree with top-of-the-atmosphere radiative flux measurements does not noticeably affect the constraint on aerosol ERF (as shown in Regayre et al., 2018). Furthermore, the marginal parameter pdfs are unaffected by the additional constraint (not shown).

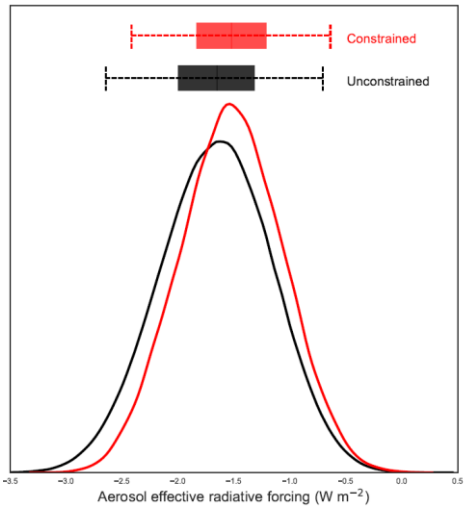

Fig. S~~2~~5. Probability distribution of ERF_{aci} from the AER-ATM PPE. Values from the unconstrained sample of one million model variants are in black. Red lines show the values constrained by ACE-SPACE measurements and additionally constrained using CERES top-of-the-atmosphere measurements. Plotting features are identical to Fig. 4.

SI References:

Adrianakis, I., Vernon, I., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G.: History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation, J. Roy. Stat. Soc. C-App., 66, 717–740, doi:10.1111/rssc.12198, 2017.

Andreae, M. O., Jones, C. D., and Cox, P. M.: Strong present-day aerosol cooling implies a hot future, Nat., 435, 1187–1190, doi:10.1038/nature03671, 2005.

Andres, R. J. and Kasgnok, A. D.: A time-averaged inventory of subaerial volcanic sulfur emissions, J. Geo. Res., 103, 25 251–25 262, doi:10.1029/98JD02091, 1998.

Bellouin, N. Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Paris, D., Boucher, O., Carslaw, K.C., Christensen, M., Daniau, A.-L., Dufresne, J.-L., Feingold, G., Fiedler, S., Forster, P., Gettleman, A., Haywood, J.M., Lohmann, U., Malavelle, F., Mauritsen, T., McCoy, D.T., Mhyre, 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, doi:10.1029/2019RG000660, 2019.

Bodas-Salcedo, A., Mulcahy, J. P., Andrews, T., Williams, K. D., Ringer, M. A., Field, P. R., and Elsaesser, G. S.: Strong dependence of atmospheric feedbacks on mixed-phase microphysics and aerosol-cloud interactions in HadGEM3, J. Adv. Model. Earth Syst., 11, 1735–1758, doi:10.1029/2019MS001688, 2019.

Boutle, I. A., Abel, S. J., Hill, P. G., and Morcrette, C. J.: Spatial variability of liquid cloud and rain: observations and microphysical effects, Quart. J. Roy. Meteor. Soc., 140, 585–594, doi:10.1002/qj.2140, 2014.

Browse, J., Carslaw, K. S., Arnold, S. R., Pringle, K. J., and Boucher, O.: The scavenging processes controlling the seasonal cycle in Arctic sulphate and black carbon aerosol, Atmos. Chem. Phys., 12, 6775–6798, doi:10.5194/acp-12-6775-2012, 2012.

Carslaw, K. S., Boucher, O., Spracklen, D. V., Mann, G. W., Rae, J. G. L., Woodward, S., and Kulmala, M.: A review of natural aerosol interactions and feedbacks within the Earth system, Atmos. Chem. Phys., 10, 1701–1737, 10.5194/acp-10-1701-2010, 2010.

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., Regayre, L. A., and Pierce, J. R.: Large contribution of natural aerosols to uncertainty in indirect forcing, Nat., 503, 67–71, doi:10.1038/nature12674, 2013.

Carslaw, K. S., Gordon, H., Hamilton, D. S., Johnson, J. S., Regayre, L. A., and Yoshioka, M.: Aerosols in the pre-industrial atmosphere, Curr. Clim. Change Rep., 3, 1–15, doi:10.101007/s40641-017-0061-2, 2017.

Cavalieri, D. J., Parkinson, C. L., Gloersen, P. and Zwally, H. J.: Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS passive microwave data, Version 1., NASA National Snow and Ice Data Center Distributed Active Archive Centerm Boulder, Colorado, USA., Accessed 22nd December 2016, https://doi.org/10.5067/8gq8lzqvl0vl, 1996.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J. L., Fichefet, D., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A. J., and Wehner, M.: Long-term Climate Change: Projections Commitments and Irreversibility, 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, edited by Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Craig, P. S., Goldstein,M., Seheult, A. H., and Smith, J. A.: Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments, in: Case Studies in Bayesian Statistics. Lecture Notes in Statistics, edited by Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P., and N.D., N. D. S., Springer, New York, NY, doi:10.1007/978-1-4612-2290-3_2, 1997.

Fiedler, S., Kinne, S., Katty Huang, W. T., Räisänen, P., O'Donnell, D., Bellouin, N., Stier, P., Merikanto, J., van Noije, T., Carslaw, K. S., Makkonen, R., and Lohman, U.: Anthropogenic aerosol forcing – insights from multi-estimates from aerosol-climate models with reduced complexity, Atmos. Chem. Phys., 19, TBA, doi:10.5194/acp-2018-639, 2019.

Gong, S. L.: A parameterization of sea-salt aerosol source function for sub- and super-micron particles, Glob. Biogeochem. Cyc., 17, 1097, doi:10.1029/2003GB002079, 2003.

Grosvenor, D.P., Sourdeval, O., Zuidema, P., Ackerman, A., Alexandrov, M.D., Bennartz, R., Boers, R., Cairns, B., Chiu, J.C., Christensen, M., Deneke, H., Diamond, M., Feingold, G., Fridlind, A., Hünerbein, A., Knist, C., Kollias, P., Marshak, A., McCoy, D., Merk, D., Painemal, D., Rausch, J., Rosenfeld, D., Russchenberg, H., Seifert, P., Sinclair, K., Stier, P., van Diedenhoven, B., Wendisch, M.,, Werner, F., Wood, R., Zhang, Z. and Quaas, J.: Remote sensing of droplet number concentration in warm clouds: A review of the current state of knowledge and perspectives, Rev. Geophys., 56, 409-453, doi:10.1029/2017RG000593, 2018.

Grosvenor, D. P. and Wood, R.: Daily MODIS (MODerate Imaging Spectroradiometer) derived cloud droplet number concentration global dataset for 2003-2015., Centre for Env. Data Anal., http://catelogue.ceda.ac.uk/uuid/cf97ccc802d348ec8a3b6f2995dfbbff, 2018b.

Gryspeerdt, E., Quaas, J., Ferrachat, S., Gettelman, A., Ghan, S., Lohmann, U., Morrison, H., Neubauer, D., Partridge, D. G., Stier, P., Takemura, T., Wang, H., Wang, M., and Zhang, K.: Constraining the instantaneous aerosol influence on cloud albedo, Proc. Natl. Acad. Sci., 114, 4899–4904, doi:10.1073/pnas.1617765114, 2017.

Halmer, M., Schmincke, H. U., and Graf, H. F.: The annual volcanic gas input into the atmosphere, in particular into the stratosphere: a global data-set for the past 100 years, J. Volcanol. Geotherm. Res., 115, 511–528, doi:10.1016/S0377-0273(01)00318-3, 2002.

Hamilton, D. S., Lee, L. A., Pringle, K. J., Reddington, C. L. S., Spracklen, D. V., and Carslaw, K. S.: Occurrence of pristine aerosol on a polluted planet, Proc. Natl. Acad. Sci., 111, 18 466–18 471, doi:10.1073/pnas.1415440111, 2014.

Hewitt, H. T., Copsey, D., Culverwell, I. D., Harris, C. M., Hill, R. S. R., Keen, A. B., McLaren, A. J., and Hunke, E. C.: Design and implementation of the infrastructure of HadGEM3: the next-generation Met Office climate modelling system, Geosci. Mod. Dev., 4, 223–253, doi:10.5194/gmd-4-223-2011, 2011.

Johnson, J. S., Regayre, L. A., Yoshioka, M., Pringle, K. J., Turnock, S. T., Browse, J., and Carslaw, K. S.: Robust observational constraint of processes and emissions in a climate model and the effect on aerosol radiative forcing, Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2019-834, in review, 2019.

Kelle, A. J. and Andreae, M. O.: Flux of dimethylsulphide from the oceans: A comparison of updated data sets and flux models, J. Geo. Res. Atmos., 105, 26 793–26 808, doi:10.1029/2000JD900252, 2000.

~~Korhonen, H., Carslaw, K. S., Forster, P., Mikkonen, S., Gordon, N. D., and Kokkola, H.: Aerosol climate feedback due to decadal increases in Southern Hemisphere wind speeds, Geophys. Res. Lett., 37, L02805-6, doi:10.1029/2009GL041320, 2010.~~

Lamarque, J. F., Bond, T. C., Eyring, V., Granier, C., Heli, A., Kilmont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M. G., Shindell, D., Smith, S. J., Stehfest, E., Van Aardenne, J., Cooper, O. R., Kainuma, M., Mahowald, N., McConnell, J. R., Naik, V., Riahi, K., and van Vuuren, D. P.: Historical (1850-2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application, Atmos. Chem. Phys., 10, 7017–7039, doi:10.5194/acp-10-7017-2010, 2010.

Lebsock, M., Morrison, H., and Gettleman, A.: Mircophysical implications of cloud-precipitation covariance derived from satellite remote sensing, J. Geo. Res. Atmos., 118, 6521–6533, doi:10.1001/jgrd.50347, 2013.

Lee, L. A., Carslaw, K. S., Pringle, K. J., and Mann, G. W.: Mapping the uncertainty in global CCN using emulation, Atmos. Chem. Phys., 12, 9739–9751, doi:10.5194/acp-12-9739-2012, 2012.

Lee, L. A., Reddington, C. L., and Carslaw, K. S.: On the relationship between aerosol model uncertainty and radiative forcing uncertainty, Proc. Natl. Acad. Sci., 113, 5820–5827, doi:10.1073/pnas.1507050113, 2016.

Loeb, N. G.,Wielicki, B. A., Doelling, D. R., Smith, G. L., Keyes, D. F., Kato, S., Manalo-Smith, N., and Wong, T.: Toward Optimal Closure of the Earth's Top-of-Atmosphere Radiation Budget, J. Climate., 22, 748–766, doi:10.1175/2008JCLI2637.1, 2009.

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.: Description and evaluation of GLOMAP-mode aerosol microphysics model for the UKCA composition-climate model, Geosci. Mod. Dev., 3, 519–551, doi:10.5194/gmd-3-519-2010, 2010.

McCoy, D. T., Bender, F. A., Mohrmann, J. K. C., Hartmann, D. L., Wood, R., and Grosvenor, D. P.: The global aerosol-cloud first indirect effect estimated using MODIS, MERRA, and AeroCom, JGRA, 122, 1779–1796, doi:10.1002/2016JD026141, 2017.

McNeall, D., Williams, J., Booth, B. B. B., Betts, R., Challenor, P., Wiltshire, A., and D.Sexton: The impact of structural error on parameter constraint in a climate model, Ear. Sys. Dyn., 7, 917–935, doi:10.5194/esd-7-917-2016, 2016.

Metzger, A., Verheggen, B., Dommen, J., Duplissy, J., Prevot, A. S. H.,Weingartner, E., Riipinen, I., Kulmala, M., Spracklen, D. V., Carslaw, K. S., and Baltensperger, U.: Evidence for the role of organics in aerosol particle formation under atmospheric conditions, Proc. Natl. Acad. Sci., 107, 6646–6651, doi:10.1073/pnas.0911330107, 2010.

Morgenstern, O., Braesicke, P., O'Connor, F. M., Bushell, A. C., Johnson, C. E., Osprey, S. M., and Pyle, J. A.: Evaluation of the new UKCA climate-composition model – Part 1: The stratosphere, Geosci. Mod. Dev., 2, 43–57, doi:10.5194/gmd-2-43-2009, 2009.

Mulcahy, J., Jones, C., Sellar, A., Johnson, B., Boutle, I. A., Jones, A., Andrews, T., Rumbold, S. T., Mollard, J., Bellouin, N., Johnson, C. E., Williams, K. D., Grosvenor, D. P., and McCoy, D. T.: Improved aerosol processes and effective radiative forcing in HadGEM3 and UKESM1, J. Adv. Model. Earth Syst., 10, 2786—2805, doi:10.1029/2018MS001464, 2018.

Myhre, G., Shindell, D., Bréon, F. M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J. F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., and Zhang, H.: Anthropogenic and Natural Radiative Forcing, 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, edited by Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Nightingale, P. D., Liss, P. S., and Schlosser, P.: Measurements of air-sea gas transfer during an open ocean algal bloom, Geophys. Res. Lett., 27, 2117–2120, doi:10.1029/2000GL011541, 2000.

O'Connor, F. M., Johnson, C. E., Morgenstern, O., Abraham, N. L., Braesicke, P., Dalvi, M., Folberth, G. A., Sanderson, M. G., Telford, P. J., Voulgarakis, A., Young, P. J., Zeng, G., Collins,W. J., and Pyle, J. A.:

Evaluation of the new UKCA climate-composition model-Part 2: The troposphere, Geosci. Mod. Dev., 7, doi:10.5194/gmd-7-41-2014, 2014.

Petters, M. D. and Kreidenweis, S. M.: A single parameter representation of hygroscopic growth and cloud condensation nucleus activity, Atmos. Chem. Phys., 7, 1961–1971, doi:10.5194/acp-7-1961-2007, 2007.

Regayre, L. A., Pringle, K. J., Booth, B. B. B., Lee, L. A., Mann, G. W., Browse, J., Woodhouse, M. T., Rap, A., Reddington, C. L. S., and Carslaw, K. S.: Uncertainty in the magnitude of aerosol-cloud radiative forcing over recent decades, Geophys. Res. Lett., 41, 9040–9049, doi:10.1002/2014GL062029, 2014.

Regayre, L. A., Johnson, J. S., Yoshioka, M., Pringle, K. J., H.Sexton, D. M., Booth, B. B. B., Lee, L. A., Bellouin. N., and Carslaw, K. S.: Aerosols and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF, ACP, 18, 9975–10 006, doi:10.5194/acp-18-9975-2018, 2018.

Rodrigues, L. F. S., Vernon, I., and Bower, R.: Constraints on galaxy formation models from the galaxy stellar mass function and its evolution, Mon. Not. R. Astron. Soc., 466, 2418–2435, doi:10.1093/mnras/stw3269, 2017.

Sander, R., Keene, W. C., Pszenny, A. A. P., Arimoto, R., Ayers, G. P., Baboukas, E., Cainey, J. M., Crutzen, P. J., Duce, R. A., Hönninger, G., Huebert, B. J., Maenhaut, W., Mihalopoulos, N., Turekian, V. C., and Van Dingenen, R.: Inorganic bromine in the marine boundary layer: a critical review, Atmos. Chem. Phys., 3, 1301–1336, doi:10.5194/acp-3-1301-2003, 2003.

Schmale, J., Baccarini, A., Thurnherr, I., Henning, S., Efraim, A., Regayre, L. A., Bolas, C., Hartmann, M.,Welti, A., Lehtipalo, K., Aemisegger, F., Tatzelt, C., Landwehr, S., Modini, R., Tummon, F., Johnson, J. S., Harris, N., Schnaiter, M., Toffoli, A., Derkani, M., Bukowiecki, N., Stratmann, F., Dommen, J., Baltensperger, U., Wernli, H., Rosenfeld, D., Gysel-Beer, M., and Carslaw, K.: Overview of the Antarctic Circumnavigation Expedition: Study of Preindustrial-like Aerosols and Their Climate Effects (ACE-SPACE), Bull. Amer. Meteorol. Soc., 100, 11, 2260-2283~~Early Online Release~~, doi:10.1175/BAMS-D-18-0187.1, 2019.

Schmale, J., Henning, S., Tummon, F., Hartmann, M., Baccarini, A., Welti, A., Lehtipalo, K., Tatzelt, C., Landwehr, S. and Gysel-Beer, M.: Course mode aerosol particle size distribution collected in the Southern Ocean in the austral summer of 2016/2017, during the Antarctic Circumnavigation Expedition, Version 1.0 Dataset, doi:10.5281/zenodo.2636709, 2019a.

Schutgens, N., Gryspeerdt, E., Weigum, N., Tsyro, S., Goto, D., Schulz, M., and Stier, P.: Will a perfect model agree with perfect observations? The impact of spatial sampling, Atmos. Chem. Phys., 16, 6335–6353, doi:10.5194/acp-16-6335-2016, 2016a.

Schutgens, N., Partridge, D., and Stier, P.: The importance of temporal collocation for the evaluation of aerosol models with observations, Atmos. Chem. Phys., 16, 1065–1079, doi:10.5194/acp-16-1065-2016, 2016b.

Schutgens, N., Tsyro, S., Gryspeerdt, E., Goto, D., Weigum, N., Schulz, M., and Stier, P.: On the spatio-temporal representativeness of observations, Atmos. Chem. Phys., 17, 9761–9780, doi:10.5194/acp-17-9761-2017, 2017.

Seinfeld, J. H., Bretherton, C., Carslaw, K. S., Coe, H., DeMott, P. J., Dunlea, E. J., Feingold, G., Ghan, S., Guenther, A. B., Kahn, R., Kraucunas, I., Kreidenweis, S. M., Molina, M. J., Nenes, A., Penner, J. E., Prather, K. A., Ramanathan, V., Ramaswamy, V., Rasch, P. J., Ravishankara, A. R., Rosenfeld, D., Stephens, G., and Wood, R.: Improving our fundamental understanding of the role of aerosol-cloud interactions in the climate system, Proc. Natl. Acad. Sci., 113, 5781–5790, doi:10.1073/pnas.1514043113, 2016.

Spracklen, D. V. and Rap, A.: Natural aerosol-climate feedbacks suppressed by anthropogenic aerosol, Geophys. Res. Lett., 40, 5316–5319, doi:10.1002/2013GL057966, 2013.

Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M.: Summary for Policymakers, 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, edited by Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Tatzelt, C., Henning, S., Tummon, F., Hartmann, M., Baccarini, A., Welti, A., Lehtipalo, K. and Schmale, J.: Cloud Condensation Nuclei number concentrations over the Southern Ocean during the austral summer of 2016/2017, Version 1.0 Data set, doi:10.5281/zenodo.2636765, 2019.

Tett, S. F. B., Rowlands, D. J., Mineter, M. J., and Cartis, C.: Can Top-Of-Atmosphere Radiation Measurements Constrain Climate Predictions? Part II: Climate Sensitivity, J. Climate., 26, 9367–9383, doi:10.1175/JCLI-D-12-00596.1, 2013.

Turnock, S. T., Spracklen, D. V., Carslaw, K. S., Mann, G. W., Woodhouse, M. T., Forster, P. M., Haywood, J., Johnson, C. E., Dalvi, M., Bellouin, N., and Sanchez-Lorenzo, A.: Modelled and observed changes in aerosols and surface solar radiation over Europe between 1960 and 2009, Atmos. Chem. Phys., 15, 13 457—-13 513, doi:10.5194/acpd-15-13457-2015, 2015.

van der Werf, G. R., Randerson, J. T., Giglio, L., Collatz, G. J., Mu, M., Kasibhatla, P. S., Norton, D. C., DeFries, R. S., Jin, Y., and van Leeuwen, T. T.: Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997-2009), Atmos. Chem. Phys., 10, 11 707–11 735, doi:10.5194/acp-10-11707-2010, 2010.

Walters, D. N., Williams, K. D., Boutle, I. A., Bushell, A. C., Edwards, J. M., Field, P. R., Lock, A. P., Morcrette, C. J., Stratton, R. A., Wilkinson, J. M., Willett, M. R., Bellouin, N., Bodas-Solcedo, A., Brooks, M. E., Copsey, D., Earnshaw, P. D., Hardiman, S. C., Harris, C. M., Levine, R. C., MacLachlan, C., Manners, J. C., Matin, G. M., Milton, S. F., Palmer, M. D., Roberts, M. J., Rosriguez, J. M., Tennant, W. J., and Vidale, P. L.: The Met Office Unified Model Global Atmosphere 4.0 and JULES Global Land 4.0 configurations, Geosci. Mod. Dev., 7, 361–386, doi:10.5194/gmd-7-361-2014, 2014.

Watson-Paris, D., Bellouin, N., Deaconu, L., Schutgens, N., Yoshioka, M., Regayre, L.A., Pringle, K.J., Johnson, J.S., Smith, C.S., Carslaw, K.S. and Stier, P.:Constraining uncertainty in aerosol direct forcing, Geophys. Res. Lett., 47, 9, e2020GL087141, doi:10.1029/2020GL087141 ~~submitted~~, ~~2019~~2020.

Weigum, N., Schutgens, N., and Stier, P.: Effect of aerosol subgrid variability on aerosol optical depth and cloud condensation nuclei: implications for global aerosol modelling, Atmos. Chem. Phys., 16, 13 619–13 639, doi:10.5194/acp-16-13619-2016, 2016.

West, R. E. L., Stier, P., Jones, A., Johnson, C. E., Mann, G. W., Bellouin, N., Partridge, D. G., and Kipling, Z.: The importance of vertical velocity variability for estimates of the indircet aerosol effects, Atmos. Chem. Phys., 14, 6369–6393, doi:10.5194/acp-14-6369-2014, 2014.

Wilcox, L. J., Highwood, E. J., Booth, B. B. B., and Carslaw, K. S.: Quantifying sources of inter-model diversity in the cloud albedo effect, Geophys. Res. Lett., 42, 1568–1575, doi:10.1002/2015GL063301, 2015.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Clim. Dyn., 41, 1703–1729, doi:10.1007/s00382-013-1896-4, 2013.

Woodward, S.: Modeling the atmospheric life cycle and radiative impact of mineral dust in the Hadley Centre climate model, J. Geo. Res., 106, 18 155–18 166, doi:10.1029/2000JD900795, 2001.

Yoshioka, M., Regayre, L. A., Pringle, K. J., Johnson, J. S., Mann, G. W., Partridge, D., Stier, P., Kipling, Z., Bellouin, N., Sexton, D. M. H ., Lister, G. M. S., Browse, J., Booth, B. B. B., Johnson, C. E., Johnson, B., Mollard, J. D. P., and Carslaw, K. S.: Ensembles of global climate model variants for the quantification and constraint of uncertainty in aerosols and their radiative forcing, J. Adv. Model. Earth Syst., ~~Early Online Release~~11, 3728-3754, doi:10.1029/2019MS001628, 2019.

Zhang, S., Wang, M., Ghan, S., Ding, A., Wang, H., Zhang, K., Neubauer, D., Lohmann, U., Ferrachat, S., Takeamura, T., Gettleman, A., Morrison, H., Lee, Y. H., Shindell, D. T., Partridge, D. G., Stier, P., Kipling, Z., and Fu, C.: On the characteristics of aerosol indirect effect based on dynamic regimes in climate models, Atmos. Chem. Phys., 16, 2765–2783, doi:10.5194/acp-16-2765-2016, 2016.