Response to Reviewer 2

We would like to thank this Reviewer for their input and appreciate their careful comments. We have addressed each of their comments below.

Major issue 1: additional context needed on significance of model-observation differences

- *At present, the manuscript text presents a large number of percentage differences between observed and simulated SO aerosol properties. The only context for the significance of these properties comes from the figure shading around 25th-75th percentiles for the control simulation (not of the other 7 simulations) and observations. From only this information, it is very challenging for the reader to evaluate whether these percentage differences are significant and to assess if the first part of the manuscript title ("The ACCESS-AM2 climate model strongly underestimates aerosol concentration in the Southern Ocean") is justified. I outline below a few different sources of uncertainty that are missing or only mentioned qualitatively in the present text.*

We thank the Reviewer for their concern in this respect. We agree that more context would be desirable for this work, however, adding this information comes at a cost of many more figures and tables, or increasing complexity of figures (hence only showing the percentiles for the observations and control).

We have now included on Figures 2-4 the 25th-75th percentiles for each of the voyage data points, and the annual means for the station data. We have also added this range for the PMO+H22 simulations seasonal cycle. We hope that this provides more context.

- *Point measurements vs grid cell means: The manuscript does not appear to address at all that it is comparing observed point (or near point when integration time is a few minutes) measurements to grid cell mean simulated measurements. This could induce large biases depending on the shape of the sub-grid scale distribution of aerosols, especially for the coastal sites. Some context on the sub-grid scale distribution could be gained from plotting distributions of measurements near in time and space and assessing skewness.*

We have considered this problem, which as the Reviewer suggests could induce biases especially near coastal sites. As stated on line 324, we linearly interpolated the daily mean ship location to the model grid. For the open ocean and a moving platform, we have deemed this satisfactory. For the land-based measurements, Macquarie Island, is not resolved by the model's land fraction, so any coastal interaction will not be resolved by default.

For Kennaook/Cape Grim, the standard practice in the Australian community doing model comparisons to the station is to take an oceanic grid box upwind from the station in order to account for the coastal interference in the model, which is usually removed from the observations via the baseline filtering (as done in this work). This method has not been published and remains to be tested vigorously. Nevertheless, we did perform this analysis, but found that the performance of the model was worse. This is discussed on line 316 onwards.

At Syowa, terrestrial influence is not of a concern for this work (but is of future interest) and so no filtering was applied to isolate oceanic only airmasses, and no adjustments to the grid boxes were made.

This is in all to say, comparing point based observations to a coarse modelled product is never bulletproof, but we are satisfied that our methods can provide results that are representative of the modelled biases. Assessing skewness based on nearby gridboxes would not yield information about sub-grid scale distribution, as this latter is strongly influenced by processes (e.g. convection) that are not explicitly resolved in our simulations. Lack of resolution of finer scale processes is a necessary compromise of global climate models.

- *Instrument and instrument simulator uncertainty: The manuscript is missing context on instrument uncertainty. Further, for the CCN comparison, the manuscript should provide uncertainty estimates due to interchanging a super-saturation-based threshold for the observations and a size-based threshold for the model. This translation will be uncertain due to assumptions needed to translate particle activation supersaturation to particle activation size (as is somewhat mentioned*

presently but not made quantitative) and due to imprecision in supersaturation control in the instruments. Ideally this uncertainty could be propagated through to the reported model results, but if this is too difficult, sensitivity tests of different radius thresholds should be reported. This latter test seems straight forward from the code in process_aer_along_ship.ipynb. Additionally, some of campaigns (e.g., MICRE) have both 0.5% supersaturation aerosol count and aerosol size distribution measurements. A comparison of the two CCN methods can be made directly on the observations and its error assessed.

We have not included a full description of each instruments uncertainty/sensitivity in this work as they are well documented in the prior papers (eg. Humphries et al 2021a). We will make this more clear in Section 3.3. Because the measurements were taken on different instruments at different times, propagating a consistent uncertainty for all the observations is difficult.

Line 247: 'Most of these observations have been described, collated, quality controlled, harmonised and published in Humphries et al. (2023), providing the first seasonal and latitudinal description of Southern Ocean aerosol properties. The Humphries et al. (2023) paper provides the ideal basis from which to perform a modelling evaluation. '

For the CCN comparison, yes, uncertainty is unavoidable when translating between a supersaturation and a size-based threshold. We have now included more text around this in Section 3.4.1.

Line 336: 'The model does not provide CCN diagnostics at a specific supersaturation, but provides CCN at selected dry diameters. The modelled size distribution can be used to calculate CCN at any particular activation diameters.'

Line 347: 'The calculation for estimating a critical diameter for a given supersaturation is imperfect and does increase the uncertainty of our results. We have also tested the critical diameter at CCN50 (which more closely matches observed aerosol populations at 0.5% super saturation (Fossum et al. 2018), which showed marginally better results near Kennaook/Cape Grim (no longer overestimated), and marginally worse results elsewhere. However, our critical diameter calculations were consistent in reporting 40nm as the cut-off for the modelled aerosol size distribution. '

The Reviewer is correct that aerosol size distribution measurements are published for some of the voyages (eg. Alroe et al. 2020 https://doi.org/10.5194/acp-20-8047-2020). In fact size distributions are available for most of the voyages, but still require processing. This is an area of high priority for us, and we currently have a student working on harmonising all size distribution data for all Australian SO voyages, providing over 10 years of data. For this reason we have not included the preliminary analysis we did on this here as we hope to perform an extensive evaluation in the near future.

- *Variability of non-control model runs: The manuscript should comment in some form on the variability (25th-75th percentile) of the non-control simulations. Ideally this could be visualized in Figures 3-4 (though this might be too visually confusing) or the supplemental.*

The reviewer is correct that we have not shown the percentiles for the sensitivity simulations as they made the figures unreadable. We have now added them for the voyage data and for the annual station data. For the seasonal cycles, we have added just the PMO+H22 simulation (to that of the observations and control) to retain readability. We note that the range of results can also be inferred from the standard deviation which is also reported in the text.

- *Exhaust corrections: Some context on the magnitude of these corrections on the observed N10 and CCN at different sites/ships should be given as well as the uncertainty introduced by these corrections.*

I would hesitate to call the exhaust filtering a correction. The authors do not change any of the data, rather than eliminate data the appears to be contaminated by ship exhaust, which by nature, will not be found in a model. To be representative, we also eliminate those days from the model data. This approach does reduce our sample size, in some campaigns by over 75%, but this is why a large scale study such as this, using multiple campaigns is essential to build up the statistics. Much more detail

about the filtering can be found in Humphries et al. 2019.

Specific comments

- *Abstract lines 13-14: "Our results indicate significant problems in the model's microphysical processes and with over tuning." Given how the manuscript is currently written, I believe this sentence should be rephrased to reflect that this is the authors' opinion. This statement is one possible interpretation of the manuscript's results. I see only one paragraph of the present manuscript referencing the tuning issue and only tangentially. (See Lines 606-611 for more detail.)*

We have revised end of the abstract to read:

Line 13: ' This significantly improved CCN in the marine regions, but resulted in detrimental impacts on the region's radiation budget, indicating that drastically improving the Southern Ocean's CCN budget may lead to poorer outcomes for the global climate.'

- *Abstract line 14: "We suggest this needs to be addressed in a holistic way." I only see part of one sentence explicitly supporting this abstract statement: "... [this] points to a need to consider model development in this space as an entire system rather than individual components." I think this should have slightly more discussion to be mentioned in the abstract.*

We have removed this sentence.

- *Lines 105-106: "By performing these evaluations [of N10 and CCN], the model biases associated with aerosol-cloud-radiation interactions around the Southern Ocean and Antarctic can be better understood and the degree of uncertainty reduced." Another sentence or two connecting why these aerosol properties specifically (especially N10) would accomplish this would make the introduction a lot stronger. (Just a suggestion.)*

We have edited this sentence to now read:

Line 117: 'By performing these evaluations of N10 and CCN, we can gain a better understanding of the modelled aerosol population and it's biases. Examining the population at two different sizes can give us insight as to how different species may be impact both the overall aerosol population as well as that of cloud-relevant size and the growth that occurs to this size. With this knowledge, we can outline how the model biases associated with aerosol-cloud-radiation interactions around the Southern Ocean and Antarctic can be better understood and the degree of uncertainty reduced.'

- *Section 3.2: The current model configurations tested (Table 1) all focus on aerosol sources. What about aerosol sinks? Even if it is out of scope to test these components of the aerosol scheme, too, the text should at least mention how the model represents aerosol sinks (e.g., coagulation, rain out) and discuss how these processes' parameterization might contribute to model-observation discrepancies.*

This is a good point and something that we have discussed within our group. On line 134 we do mention these processes within the model - coagulation, deposition, cloud processing, however these processes are not discussed in terms of their impact on the aerosol.

We have now added a paragraph in the introduction which discusses these processes, however, as the Reviewer notes, testing these parameterisations is out of scope within this work.

Line 91: 'Aerosol sinks, and how they are modelled, are also a key source of uncertainty. Aerosol can be removed from the atmosphere via dry deposition or wet deposition. Dry deposition is difficult to measure and evaluate. Regayre et al. (2020), after applying Southern Ocean observational constraints to a perturbed parameter ensemble, find that it is likely that a scaling factor for the accumulation mode dry deposition velocity in the Unified Model needs to be lower than the default value. Other observational studies have indicated that wet deposition (rain after coalescence of cloud droplets) is an important control of CCN variability in the Southern Ocean, particularly in relation to shallow convection

(Alinejadtabrizi et al. 2024) and stratocumulus (Kang et al. 2022). Given the tendency for models to produce too much light rain (Stephans et al. 2010), it has been suggested that wet deposition may be overestimated in models (Kang et al. 2022).'

- *Lines 299-301: "Some evaluation of the model[']s meteorology has been carried out, but is not shown in this work. It was found to be satisfactory, which is in line with our expectations due to nudging." Can this be presented in a supplemental? With such minimal description, it is not clear what is "satisfactory."*

We have not presented this work for the sake of brevity. Overall the model performed well which was expected given the nudging. We have now removed reference to this work so that we do not confuse readers and to not take away from the main focus of this paper.

- *Line 316-317: "At KCG, we have used the exact model gridbox that the station is located in, as choosing a gridbox to the south- west of the station resulted in poorer performance." This sentence seems to suggest at other locations the exact model gridbox was not used? If this is true, this conflicts with my understanding of Section 3.4.2 for a stationary location.*

We have edited this sentence to clarify that we are using the exact model grid box everywhere.

Line 353: ' The linearly interpolated model gridbox for each location (see Section 3.4.2) was used to perform the comparisons. This is also true for, as choosing a gridbox to the south-west of the station, as is often standard practice for this location, resulted in poorer performance.'

- *Section 3.4.2: Do ships cross multiple grid cells in a day? (Is the daily average lat-lon location problematic?)*

While we have not calculated the exact change in position for each ship each day to say with certainty, we would say that it is highly likely that a ship can cross more than one grid box in one day. This is especially the case for the resupply voyages where the ship may be travelling at full speed (approximately 12 nautical miles an hour) for up to a week to get to the Antarctic bases. A ship doing this can cover more than 150km in one day - the approximate size of a grid box. We have used a linear interpolation of the point location to extract the model data to account for instances when the average lat lon location is at the edge of a grid box.

- *Figures 3, 4: It is unclear exactly what "the monthly and annual median concentrations of N10/CCN" in the captions means, especially when compared to the methods description. Does this mean "monthly and annual median of mean daily concentrations of N10/CCN" or "median of monthly and annual concentrations of N10/CCN"? I assume the former from the methods, but this is not clear from the text. I am not sure whether the colors used are colorblind safe.*

You are correct that is is the former and we have made this more clear in the captions. We have passed all our figures through a colour blindness simulator to make the best efforts at presenting figures that are colour blind safe. This is also why we have used different line styles to aid visual identification.
Figure 3 caption: 'The monthly and annual median concentrations of daily mean N10 for at...'

- *Figures 6, 8: It would be helpful to compare the different model configurations' TOA SW radiative flux directly to the CERES product instead of plotting line contours where the control bias changes signs. Even just adding one panel with the zonal means would be helpful to support the manuscript's conclusions about which configurations improve the model relative to observations. I find it very difficult to get this important information out of the current plots. (Just a suggestion.)*

We have included Table 2 for to provide a quantitive view of the model changes compared to CERES. For this reason we won't add an additional panel.

- *Figure 7: Since LWP isn't compared to an observational product and is only mentioned as a key controller of TOA SW flux (already shown), this figure and its associated text don't seem necessary to the main manuscript text and could go in a supplement.*

Yes, this is something we discussed in depth as to whether we include or not, as I for one did not want this part of the paper to take away from the main findings of the aerosol evaluation (and also that we already have several papers published looking at the LWP biases). In the end we compromised and left this plot, but didn't go as far as to show the other cloud properties. We note that Reviewer 1 has argued for further cloud property plots, so in compromise again, we will leave the plots as they are.

- *Lines 511-512: "We note that we have evaluated other cloud properties, and the aerosol direct effect via clear sky radiation, but for brevity will not discuss them here." This seems unnecessary to mention if these evaluations aren't even roughly summarized and aren't included in a supplemental.*

We have removed this sentence.

- *Lines 606-611: This paragraph seems to motivate the second half of the title and the end of the abstract. It is not clear to me that these are unique conclusions from the methods presented and analysis shown. There is some discussion of why the experiments are probing what they are probing, but their implementations still contain such large uncertainties that it is unclear if scattered improvements in the simulated-observed CCN comparison demonstrates "improvement of the physical representation" of aerosols. This is why I suggest toning down the abstract above. Since the title includes "could," I think it is acceptably couched.*

Thank-you for this feedback. We have altered the abstract as you suggested. We have also edited the abstract to not overstate our results, removing this specific line.

Technical comments

- *Title: The antecedent for "it" is not clear from the title alone. The antecedent could be either "ACCESS-AM2 climate model" or "aerosol concentrations."*

We have changed the title to 'The ACCESS-AM2 climate model strongly underestimates aerosol concentration in the Southern Ocean; improving the aerosol scheme could be problematic for the modelled climate system'

- *All of the manuscript's equations are presented at the end of paragraphs and are not interwoven into the text.*

We will leave it to the copy editors to decide on final placement of equations.