# Response to referees on *"Evaluation of Southern Ocean cloud in the HadGEM3 general circulation model and MERRA-2 reanalysis using ship-based observations"* by Peter Kuma et al.

Peter Kuma[1], Adrian J. McDonald[1], Olaf Morgenstern[2], Simon P. Alexander[3], John J. Cassano[4], Sally Garrett[5], Jamie Halla[5], Sean Hartery[1], Mike J. Harvey[2], Simon Parsons[1], Graeme Plank[1], Vidya Varma[2], Jonny Williams[2]

[1]School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand
[2]National Institute of Water and Atmospheric Research, Wellington, New Zealand
[3]Australian Antarctic Division, Kingston, Australia
[4]Cooperative Institute for Research in Environmental Sciences and Department of Atmospheric and Oceanic Sciences, University of Colorado, Boulder, Colorado, US
[5]New Zealand Defence Force, Wellington, New Zealand

We would like thank the referees for their valuable comments. We have addressed a number of related referee comments by replacing the free-running model GA7.0U/1980-90 with a nudged model run GA7.1N/2015-2018, nudged to observations based on the observatioanl HadISST SST and sea ice dataset and the ERA-Interim reanalysis. The new model performed much better in terms of cloud representation relative to observations, but a significant error in TOA outgoing SW radiation and cloud occurrence representation remains, especially related to low cloud and fog. This is reflected in the revised manuscript.

The referees' comments below are marked in **bold**, followed by authors' response. We supply a latexdiff document which identifies the changes made. Page and line numbers in our response comments refer to the latexdiff document.

Introduction to Figure 10 in the text has been relocated to Results, but for clarity we keep the original numbering of figures, which can be changed in a final revision.

## Anonymous Referee #1

**Review of "Evaluation of Southern Ocean cloud in the HadGEM3 general circulation model and MERRA-2 reanalysis using ship-based observations" by Kuma et al. (acp-2019-201)**

**Summary:**
**The paper investigates cloud cover over the Southern Ocean through comparisons between numerous ship-based measurements and model outputs (including reanalysis). They demonstrate underestimation of low-level cloud cover in the HadGEM3 model and in the MERRA2 reanalysis. They investigate the link between boundary layer thermodynamics and low-level cloud cover and cloud biases. They show that the TOA SW biases are mainly related to places where the coldest near-surface airmasses are (near or below zero). They conclude on the subgrid-scale parameterisations being responsible for misrepresentation of clouds in model rather than boundary layer thermodynamics.**

**Relevance of the paper and overall comment:**
**The paper presents and describe a very valuable dataset of ship-based measurements of low-level cloud over the Southern Ocean, where observations are badly needed to understand the near-surface processes affecting cloud formation and responsible for the cloud/radiative biases in climate models over the SO. To this respect the paper addresses relevant science questions in the scope of ACP. However, it seems to me that more work is needed to achieve ACP standards, in the way the science is presented and discussed (major revision). The dataset deserves better scientific discussion and less vague or speculative comments in several parts of the paper. Figure 5, 7,8 and 10 are very interesting but the analysis and discussion should be better handled. I first list some major comments, and then line by line comments.**

**————————————- Major comments: ————————————-**

**1)**

**The use of different time-periods needs to be much better introduced, justified, and discussed. I don't understand why the author use GCM simulations for the 1980-1989 period in a free-running mode, and then a nudged simulation for the year 2007 (only), while MERRA is used only for the 2015-2018 (the years where ship-based measurements took place). The reader needs much better justification for the choice the authors make to compare different periods. And a discussion on the shortcomings of doing so should appear in the paper. P4-Line 26, the authors say "Limited data availability meant that no nudged runs were available for the period 2015-2018". Is this really the case? And if this is the case, why not having a free-running simulation for this period then? And why is the nudged run over 2007 only? Also, MERRA2 could be used for the 1980 period. MERRA2 is available for >=1980. MERRA could help bridge between the period 1980's/2007's of the GCM outputs and the period of the ship-based measurements (2015-2018). At least using MERRA2 also for the 1980's and 2007 + explaining/discussing the choice for the time periods of the GCM runs would be needed. The best case would be to have GCM runs over 2015-2018. How using different periods for GCM/MERRA2 would affect Figure 5 for instance? And what about Figure 1 and the TOA biases where only the year 2007 is shown?**

The availability of model datasets was indeed limited by organisational capabilities. Especially, a nudged run for the observational period 2015–2018 was not available, therefore the choice of statistical comparison with a decadal simulation of 1980–1990. To address this major comment, we have produced a new nudged run of GA7.1 (UM 11.0) for the observational period. This nudged run was evaluated in the same way as MERRA-2 originally, i.e. in a 1:1 comparison, assuming that weather conditions are comparable between the model and observations. The results based on this run were more accurate than GA7.0U. We have decided to leave out the decadal run of GA7.0U entirely to improve clarity of the manuscript.

Due to this change, a significant part of the Results section (P13L11–P19L3) has been updated. Figure 5 shows significantly different results for GA7.1N (cloud occurrence bias is smaller), but the nature of the error has not changed – low cloud and fog is still underestimated in the model. Figure 6 shows that GA7.1N has lower average bias at 4–9%. Figure 7 shows that the correspondence between min{SLL,LCL} and CBH is still not well represented in the model. Figure 8 has been updated to show min{SLL,LCL} (previously it showed SLL), GA7.1N is matching the observed distribution quite well.

**The authors speak about the years 2016-2018 that had unusually low sea ice extent (p15-Line16): how does this impact the comparisons with other years where sea ice was different?**

The influence of low sea ice concentration in 2016–2018 may be hard to quantify with our data, considering that few of the ship observations are available prior to 2016. We therefore consider the whole ship-based dataset as representing relatively uniform conditions. The new nudged model run (as well as MERRA-2) are based on the observed sea ice concentration, therefore the 1:1 comparison of the models with observations should not be biased.

**Having said that it is possible that the paper could be improved by giving up Figure 1 or 2, while focusing more on the novelty of this work, which is the ship-based measurements of clouds (+thermodynamics), and drop the comparisons to GCMs (mainly because they are different periods. . .) and keep the comparisons to MERRA2, and add the ERAI reanalysis. To this respect ERAI, which is widely used, could also be presented here. How is ERAI doing vs. MERRA2. Say, if ERAI has a contrasting behavior to MERRA2 (ie more like the GCMs) the authors could consider presenting the ship-based measurements + MERRA2 + ERAI over similar periods, and drop the GCM runs, which deal with other periods.**

We did not include ERA-Interim in the new revision. Technically, ERA-Interim does not provide the necessary model fields for running the simulator. However, it would be possible to compare with ERA5, which provides these fields. The nudged run of GA7.1 is nudged to ERA-Interim. Therefore, the dynamical conditions are likely very similar, even though they can still differ in their representation of clouds. We did

not want to leave out the GA model because it is a model which the authors want to improve (the authors participate on development of this model).

We agree that comparison with another reanalysis such as ERA5 would be interesting, but we prefer to limit the scope to just GA7.1 and MERRA-2.

**2)**

**The discussion of ship-based measurements should be better related to the TOA SW bias over all periods where these field measurements were available. Since the authors try to understand what the models are doing wrong, the discussion should make the most of the different measurements period.**

In the updated manuscript we link Figure 5 with the SW radiation bias by introducing a "back-of-the-envelope" calculation showing how the SW radiation bias would change by increasing cloud cover in GA7.1N by 5%, assuming no change in cloud albedo (Table 3; P23L9-14). The error would be reduced by over a half by increasing the cloud cover alone by an amount which is approximately consistent with the results shown in the updated Figure 5 and 6.

**Because maximum insolation occurs in January, a focus is made on this month but Figure 2 clearly shows that March – for instance – can also and still show substantial biases.**

The updated Figure 2 now shows the biases. The bias in MAM (Figure 2h, i) has a very similar spatial pattern as in DJF (Figure 2e, f) and the annual bias (Figure 2b, c) in both GA7.1N and MERRA-2. The magnitude of the bias in MAM is approximately 1/2 of the magnitude in DJF due to lower solar insolation. We think one can therefore reasonably expect the nature of the underlying cloud bias to be similar in DJF and MAM (P13L19-22). Because the focus of the work is on improving the SW radiation bias, we think it is suitable to focus primarily on the months with the greatest bias (DJF) (P19L14-19).

**Since Ship-based measurements are also available in autumn and November-December, it would be very welcome to have also biases like the ones shown in Fig. 2 for the autumn and other summer months. Is the TOA SW bias spatial pattern (and the related comparisons between models) the same during these other months? I suggest Figure 3 to show only biases (the subplots m-p) for summer and autumn. The other maps (a-l) are difficult to read with the blue-shaded colourscale and I am not convinced they need to remain.**

We are have not analysed the November data (AA15 and HMNZS Wellington) to keep focus on the season with the greatest bias (DJF). We also wouldn't have data to cover a substantial part of the September-November (SON) season. The updated Figure 3 addresses the other points.

**Better discussing the cloud cover results in Summer/Fall (Figure 5) in relation to radiation biases in Summer/Fall would improve the overall discussion and conclusions of the paper. This would allow to make the most of the ship-based observations.**

This is now discussed in P19L14-16, P20L4-L20, and by adding a back-of-the-envelope calculation (Table 3; P23L9-14).

**Figure 5 is a great one and it would deserve better discussion in light of the motivations (i.e. the TOA SW biases over the SO and why theses biases are present).**

We have added more discussion in P15L26-P16L3 and P20L14-20. We think the relationship between the results shown in Figure 5 and the SW bias is relatively simple. Because Figure 5 shows the results based on the cloud mask, and the models underestimate the amount of cloud (as detected by the cloud mask), this leads to a proportional underestimate of outgoing SW radiation, unless the albedo of the cloud is overestimated (which is not analysed in our work). The SW radiation bias results in GA7.1N south of 60°S appear to be in line with this (Table 3), unlike MERRA-2 which is overestimating the outgoing radiation,

necessarily by overestimating the cloud albedo (P20L4-13).

We think it is useful to focus on the cloud amount and albedo, as it is important to fix the SW bias for the right physical reasons, i.e. the models simulate the correct amount and type of cloud at the correct altitude, and this is where our work makes a contribution. Other studies have already focused extensively on the analysis of TOA and surface radiation balance and the cloud albedo (through analysing the cloud phase). Our work is therefore complementary to these studies, and this is the advatage of ceilometers is over other instruments, but their ability to tell anything about the phase or the reflected SW radiation is limited.

We are currently working on a follow-up study which will focus more on comparing the individual cloud features observed by lidars and simulated by models. Our preliminary results suggest that GA7.1N is almost unable to simulate observed layers of stratocumulus cloud in the region and fog is either missing in the model or greatly underestimated. Therefore, the boundary layer, convection and large-cloud schemes will likely need to be fixed to be able to generate these types of cloud in the conditions typical in the SO.

**The authors note that GA7.1 reduces the SO SW radiation bias (e.g. in the abstract p1-L9). Figure 5 does not show any cloud from GA7.1 (only GA7.0). Why GA7.1 is performing better? Is this really because of better cloud representation (but we cannot see it from Figure 5)? And if not, what does it say about cloud being the main/only reason of SW radiation bias?**

We do not include GA7.0 in the manuscript any more (largely because the available run wasn't nudged, but also to limit the scope), but the updated Figure 5 shows that GA7.1 is much better than GA7.0 when it comes to simulation of cloud occurrence, and is now also better than MERRA-2. This should also explain much of the improvement in SW radiation in GA7.1 relative to GA7.0.

**Related to 1), what ERAI would give in Figure 3? More like the GCMs or like MERRA2? Perhaps ERAI brings this contrasting behavior that the authors highlight between GCMs and MERRA2, and this would allow to have both observations and reanalyses (only) used over the same period (2015-2018)**

While we agree that including ERA5 (ERAI does not contain the necessary fields) would be intereting, it would significantly expand the scope. The comparison with MERRA-2 already shows some interesting contrasting results, such as the different latitude of the split between the positive and negative SW radiation bias, the overestimation of cloud albedo in MERRA-2 (P20L4-13) and the cloud pase differrences (Figure 9). Perhaps a future study could focus on comparison of different reanalyses (ERA5, MERRA-2, JRA-55) with lidar observations in the region. Reanalyses are a suitable target for this kind of comparison due to having correspondence to the observed weather, while results from GCMs are often not available in a nudged mode, at least not via public repositories such as CMIP5, with the exception of the TAMIP project (GCM hindcasts). We therefore prefer to keep the scope limited to GA7.1 and MERRA-2.

**3)**

**The authors tend too often to rely on previous conclusions from previous papers (e.g. the Bodas-Salcedo et al. ones) to comment on what they find, rather than more thoroughly commenting/discussing their own novel results. The discussion part for instance gives much room to results of previous published study and/or to speculative comments about why GA7.1 is doing better than GA7.0 etc. and how MERRA2 is overcompensating for the low cloud-cover etc. Several sentences using "we cannot conclude. . .", "we cannot make the same conclusion. . . but it seems plausible. . ." "we cannot make substantial conclusions. . ." considerably weaken the discussion (section 4) from the beginning, and hence the paper, while it seems that all the ship-based measurements bring very valuable results (Figure 5, Figure 7, Figure 8) and interesting comparisons to MERRA2, and GCM runs (but cf. my point 1. on the time periods used).**

We have extended the Discussion with a more detailed discussion of the results (P19L30-P20L20, P20L30-P21L10, P22L34-P23L14). We explain in P20L4-13 why the statement about MERRA-2 overestimating the cloud albedo in not speculative but factual based on the results.

**4)**

**The discussion on the effect of sea ice is overlooked while it seems that some discussion could be made from Figure 8 (q,r and w,x). Also, while it seems that 8w is still showing some correlation, Figure 8x shows very different behavior and no attempt is made to comment on this. Given that a lot of the soundings you use (65) were made in 100% sea ice regions, and many of your CBH observations as well (I see the number of points present in your Figure 8x compared to the other similar subplots), I would expect to see more in-depth study of these observations, and this is really missing the present version of the paper in my opinion. For instance, the recent study by Jolly et al. (2018) that you cite showed the influence of different regimes on cloud cover: can the observation in Figure 8x be explain by particular synoptic-scale regimes or just by the sea ice being 100%? And why? Also, that other recent study by Listowski et al. (2018) that you also cite showed that not all low-level clouds anticorrelate with sea ice fraction but only the liquid-bearing ones. Can the behaviour you see in Figure 8x also be explained by clouds being of different sort/phase? The absence of correlation between CBH and min(SLL and LCL) may lead to think that you could be observing clouds advected from other places not related to local atmospheric conditions, that may be different in nature/phase from clouds over open water (you mention some hints towards the detection of supercooled liquid water with one of your instrument, can't you improve the Figure 8 by adding information on the phase, notably for Figure 8x?) In other words, can what you are observing from regions with 100% sea ice be explained by changing synoptic scale regimes, or cloud phase, or other things? Speaking of the sea ice regions, in Figure 7 the very low CBH are identified as being due to fog/very low clouds (p13-L20) and these are the points we also see in Figure 8x. Could this be blowing snow since we are in a 100% sea ice-covered region where snow can accumulate? In relation to cloud phase, in Figure 9 you compare LWP and IWP for GA and MERRA only for a specific year and month (jan. 2007). This does not seem satisfying to conclude for the longer time scales/other periods. Here again the use of different time-periods in the paper is not very welcome (see my point 1.). Do you really need Figure 9? If you really want to go into the cloud phase, using the lidar observations to assess the nature of cloud phase would be welcome. Or, as suggested in 1), perhaps only using MERRA2, ERAI (to contrast with MERRA2?) over 2015-2018 (only) would be a better option rather than using 2007, i.e almost a decade before the 2015-2018 ship-based observations. . .**

We do not discuss the effect of sea ice largely because cloud representation over sea ice makes little difference to the SW radiation bias (the surface is already highly reflective in SW). We consider the results interesting, but outside of the scope of the paper. But, this is potentially a topic for future effort that is being explored in our group.

The 65 radiosonde observations in Figure 8x were likely performed in a similar location in the high-latitude Ross Sea region (70–75°S). This region is likely affected by its proximity to land and not very representative of SO in general. These observations were only marginally related to our focus on SO.

We have updated Figure 9 to show a longer time period, which also better matches with the observations (DJF 2017/2018). We think this figure demonstrates nicely where some of the SW radiation difference between the models comes from. In our analysis we didn't use lidar observations which would be able to distinguish liquid and ice clouds. From one of the voyages we have data from a dual-polarisation lidar MiniMPL (but not from sea ice), which could allow us to perform such an analysis of cloud phase in the future, or provide the data to someone else upon request.

The radiosonde observations on TAN1802 were all performed in ice-free regions, and relatively far from any ice covered regions. Therefore, it is unlikely that the clouds could be advected from ice covered regions. None of the points in Figure 8x (70–75°S) appear in Figure 7 (60–70°S). This choice was made due to the likely effect of land/sea ice on observations performed between 70–75°S, while we intended Figure 7 to represent conditions in the open ocean.

**5)**

**Finally, the authors say that subgrid-scale processes should likely be responsible for the cloud mis-representation in models rather than the boundary layer thermodynamics but it is never said and**

**commented on what these subgrid-scale processes are. Do you mean the microphysics? Other processes? A discussion of what is used in the models regarding these processes would perhaps help to understand what should be improved in priority in the models and why the models are wrong. Using the contrasting behaviors of models to try to pin down the cause of cloud misrepresentation is an interesting method but the authors should provide with more clues in the discussion about what those subgrid-scale processes are and try to spot the main differences in the way the models implement these processes.**

We have added a paragraph in Discussion commenting on which subgrid-scale processes might be responsible (P22L34-P23L14). However, precise identification of the processes (we mean cloud parametrisation, boundary layer parametrisation and convection parametrisation in the Unified Model) is not something we can do without a more extensive analysis and especially running model experiments with modified parametrisation. Microphysics would likely have an effect on the cloud phase and thus SW reflectivity of the clouds, but we consider the cloud occurrence/cover a potentially larger problem than cloud albedo (see the added "back-of-the-envelope" calculation). We plan to focus on this problem in an upcoming paper.

**————————————————- Line by line comments: ————————————————-**
**————————– Abstract ————————–**
**P1-L9 By how much GA7.1 reduces the bias?**

We have removed GA7.0 from the analysis and only compare GA7.1N and MERRA-2.

**P1-L17 The analysis you mention is referring to your Figure 9 and the related comments. They only refer to the period January 2007. . . as mentioned in my major comment 4) this is not satisfying I think. When one reads the abstract it seems that you compare modelling and MERRA2 over the same period as the ship-based measurements, which is not the case. This is misleading.**

In the updated analysis we are comparing the same time period in the models and observations.

**————————————– 1.Introduction ————————————–**
**P3 – L12 : "It was also more. . ." : what does "it" refer to exactly ?**

Replaced by "The clouds were also" (P3L18).

**P3 – L14 : "more likely to have intermediate cloud fraction" This is not clear. What is meant here by "intermediate cloud fraction"?**

"Intermediate cloud fraction" comes directly from Protat et al. (2017). We have clarified by adding "rather than very low or very high cloud fraction" (P3L18-20).

**P3 – L19-20 Please double-check and be more precise here (what "tuning" do you mean?). Kay et al. (2016) changed the threshold temperature below which detrained condensates are ice crystals and not liquid any more. They lower this threshold, allowing for more condensates to remain in the supercooled liquid phase when being detrained. The way the sentence is written suggests that ice crystals only are detrained.**

We have replaced "tuning" with "decreasing" (P3L25).

**P3 – L25 The reference to Jakob (2003) is a bit short or can be removed unless you specify what you mean by "cloud evaluation" regarding this specific study.**

We have removed the sentence.

**P3 – L27-35 Please make a new paragraph and give section numbers to help the reader.**

We have split the paragraph into two and added section numbers (P3L34-P4L8).

**————— 2. Method —————**

**General comment: I would suggest a section 2. Datasets and 3. Method (lidar simulator). As it stands, it seems that this section combines too many different information about the data/methods used in the paper.**

We have split Methods into Datasets, Methods and Spatiotemporal subsets investigated (P4L9-P13L10).

**P4 – L2 As mentioned in my major comments, adding ERAI would be very interesting since this is a widely used reanalysis by the community, and would allow to contrast MERRA2 on same time-periods than ship-based obs.**

Please see the our comments above regarding this point.

**P4 – L9 I wonder whether a small appendix summarizing the main aspects of the lidar simulator would not be needed here, since the reference put is a paper in prep.**

We have added a reference to the website containing technical documentation of the new simulator (https://alcf-lidar.github.io) (P10L25).

**P4 – L16 What is the difference between GA7.0 and GA7.1 ? This would help understand and discuss the better performance of the latter in terms of SW bias (as stated in the abstract).**

We have removed GA7.0 from the analysis and focus solely on GA7.1N and MERRA-2.

**P4 – L 18 As said in the major comment it should be explained why these runs are used. 1980-1989, and then 2007. Why not having runs over more recent periods (as the ship-based measurements).**

We have removed this section. Only one GA model run and one time period are investigated in the updated analysis.

**P4 24 – "Can only be compared statistically" What do you mean? Please clarify.**

We have removed this sentence (GA7.0U is no longer used). The nudged run GA7.1N is now compared 1:1 to observations.

**P4 – L26 "Limited data availability. . ." What do you mean? See my major comment 1) It does not seem that you are saying over which period you analyse MERRA2. This should appear in this section.**

We have removed the sentence on "Limited data availability" (addressed by using a nudged GA7.1). We have added a sentence to the MERRA-2 section regarding the analysed time period (P8L16), which is the same as the period of observations.

**P5 – L15 "downsampled" from what initial resolution?**

In the updated analysis we used the original MERRA-2 resolution without downsampling. We have removed the sentence.

**P6 – L5 "appears largely zonally symmetric" I don't think we can say the pattern of the bias is symmetric, even zonally, but rather that the bias is present across all longitudes, but its magnitude does change zonally.**

We have replaced the statement with "the SO SW radiation bias is present at all longitudes in the SO"

(P6L14-17).

**P6 – L6 "with a notable exception. . ." Precisions not needed in the section presenting the ship measurements. . .**

We have replaced this part by "SO SW radiation bias is present at all longitudes in the SO (...), affected by atmospheric circulation in the SO (...)." (P6L14-17).

**P6 – L8 "Figure 1. . ." This Figure is already mentioned before P5 – L21.**

We have removed the sentence.

**P6 – L20 I am not sure what is meant by "directly reveal the cloud liquid. . .". The strength of using a simulator is to compare the observables and not to rely on all the hypotheses used by inversion routines to retrieve IWC and LWC from lidar observations.**

We have replaced the sentence by: "Due to signal attenuation and noise ceilometers cannot measure clouds obscured by a lower cloud, and therefore cannot be used for 1:1 comparison with model clouds without using a lidar simulator, which accounts for this effect (Chepfer et al., 2008)." (P6L29-P7L3).

**P7 L24 – Please clarify the title e.g. "Geographical areas/domains investigated" or "Domains used for the analysis" Also, having 2.1 as "Datasets" then 2.2 as "Domains" then 2.3 "COSP simulator" is not ideal I think, and I would first present all datasets and tools, and then the domains.**

We have changed the title to "Spatiotemporal subsets investigated" (P12L19). We have also restructured the text as suggested. Subsections of Methods are now split into Datasets, Methods and Spatiotemporal subsets investigated (P4L9-P13L10).

**P8 – L12 The title of this subsection is misleading since you are not using COSP simulator in the end, but your own simulator. Please change the title accordingly.**

We have renamed the section to "Lidar simulator" (P10L1).

**It seems to me you don't need a section 2.3 and you could have everything put in current 2.4.3 where you could at once explain the modeling of the lidar signal along with its processing.**

Some of the same lidar processing steps (such as cloud detection) are applied on the simulated lidar as on the observations. Therefore, the current arrangement make sense from this perspective.

**P9 L15 What is this known value of LR? Where does it come from?**

The value is $18.8 \pm 0.8$ sr as stated in the paper referenced in the paragraph (O'Connor et al., 2004). We have added this number in parentheses (P11L14).

**P9 L21-22 Citing Kotthaus et al. at the end of the paragraph falls a bit short and I am wondering if it should not appear earlier in the paragraph with some more explanation about why you refer to this study. Are you using their method? Then please say it.**

We have removed the sentence.

**P10 L7-9 Why do you need to do this? How are these random samples used then?**

Added sentences "The lidar simulator processes each sample individually. The resulting cloud occurrence is calculated as the average of the 10 samples." (P12L7-9).

**To shorten this section 2. I would not define SLL here, rather when it is used for the first time. Also SSL is neither a dataset, nor a tool, rather a variable defined to help with the analysis. Also, is there any past reference using this definition? If yes, please cite relevant paper.**

We have relocated the paragraph introducing SLL to the Results section (Section 5.3) (P16L31-P17L7). The authors are not aware of any references of previous use of SLL or an equivalent metric. We noted the relationship between SLL and CBH as theoretically plausible and later confirmed by joint radiosonde and ceilometer observations (Figure 7).

**————————- 3. Results ————————-**

**P10-L26 to P11 L-16 There are too many statements dealing with observations made on Figure 3 that are actually difficult to see, whereas Figure 4, introduced after, is more helpful to confirm statements made by the authors. Also, as suggested in major comments, I would tend to simplify Figure 3 by showing only the biases and remove all the blue-shaded figures where the biases are difficult to read, especially regarding the statements made by the authors in the main text.**

We have updated Figure 3 to show biases, which should now be more clear.

**P10 – L28 "Lower" than what? And what biases? Please clarify.**

The biases are discussed in Loeb et al., 2018 referenced in the sentence. We have clarified in parentheses (P13L16).

**P10 – L29-30 I would remove the sentence about the "predominantly zonally symmetric pattern" and the "more variable patterns in the tropics", which is not very clear to me.**

We have replaced the part with "relatively zonally symmetric pattern of negative and positive bias" (P13L17-19), as the updated figure now shows the bias more clearly, and removed the part about the tropics, which are not covered in the updated figure.

**P11 – L1 "upwelling and downwelling" what?**

We have removed the sentence. Figure 3 is now zoomed on the SO rather than covering the tropics to better highlight the features.

**P11 – L3 "large differences" between what? I would drop the mentions to the Peninsula and what is happening to the east of it as it is not clear why one would give so much importance to this since the ships did not get there anyway.**

"Large differences" between the models and CERES. We have removed this part of the sentence (P13L23-25).

**P11 – L1 I don't understand the footnote. Also, I am not convinced there is a need to highlight a particular day in the present paper.**

We have removed the footnote. The day picked in Figure 3 and Figure 9 is now 1 January 2018. We think that it makes sense to keep the daily plots due to the stark difference in TOA outgoing SW radiation between CERES and the models visible on the daily means, consistent with the statistical results. We have changed the scale and colormap of the plots to better highlight the difference.

**P11 – L4 One cannot really see this "greater reflectivity".**

This should be visible in the updated Figure 3. We have replaced the sentence with "The region on the eastern side of the Antarctic Peninsula shows the greatest negative bias in the models (Figure 3b, c, e, f)". (P13L25-27).

**P11 – L13 "With some individual cloud systems being too bright". I am not sure this should remain in the text. Again, I think all the consideration about the blue-shaded maps in Figure 3 (but biases maps should be kept) should be removed and Figure 4 should be used instead.**

We have replaced the blue-white colormap with a grayscale colormap on a smaller scale of values and smaller span of latitudes. The differences between the models and CERES in the updated Figure 3 should be more obvious now. We have also removed the part of the sentence "with some individual cloud systems being too bright" (P14L2-6).

**P11 – L21 "cyclical". Rather "seasonal"?**

Replaced with "seasonal" (P14L13, P14L16).

**P11 – L20 What is meant by "likely a secondary modulating factor". Please be more explicit. A modulating factor for what?**

Modulating factor for SW radiation. We have clarified: "modulating factor of the TOA outgoing SW radiation" (P14L14-16).

**P11- L26 "These panels also justify why. . ." Not needed.**

We have removed the sentence.

**P12 L4-6: The two sentences fall a bit short. Also, they would be in better place in the discussion part, with more explanations. ". . ..in the GA7.1 model": so what?**

We have removed this paragraph (we no longer compare GA7.0 and GA7.1).

**P12 L2 Figure 5 is very interesting and rich, and more analysis should be provided also regarding similarities or differences between summer results and autumns results. (Please consider adding letter to designate specific subplots of Figure 5). Also it seems that obs and model agree more where the statistics is larger (more days), can't you say something about that? Isn't it possible that at other time/places the larger disagreement between model/obs is partly due to smaller statistics of observations? This relates to my major comments that more analysis and discussion are really needed on this plot.**

The results from the updated Figure 3 suggest that there is little difference in the TOA outgoing SW radiation bias between the austral summer (DJF) and autumn (MAM) (the geographical pattern is very similar, except for the magnitude modulated by the incoming solar radiation). We therefore expect the bias has the same underlying cause in both DJF and MAM. MAM is also much less important in terms of fixing the SW bias in models due to much lower solar insolation in the season. Also Figure 5 does not indicate that there is a significant difference in cloud occurrence between DJF and MAM.

We think that greater number of days in Figure 5 does not necessary imply much greater weight of the result due to time correlation of weather patterns and correlation with sea ice concentration around Antarctica. Therefore, we partially consider the subplots of Figure 5 as independent "snapshots" each with the same weight (solid lines in Figure 6), in addition to calculating the weighted averages (dashed lines in Figure 6).

We have added labels to Figure 5.

**P12-L10 What period is used for MERRA2 here?**

This has been addressed by removing the statistical comparison with GA7.0U and stating the time period in Methods (P8L16).

**P12 L12. As mentioned in the major comment. Why can the authors trust comparisons between simulation of the 1980s period and the 2015-2018. This should be much better introduced/justified.**

We have removed the sentence (addressed by using a nudged model).

**P12 L19-20 how much higher?**

We have removed this sentence. Schuddeboom et al. (2018) evaluated GA7.0 which had a much greater bias than now evaluated GA7.1.

**P12 L20-21 "Due to the zonal. . .of the whole SO" Could suit the discussion part. Not needed here.**

We have removed the sentence.

**P12 L27-34 I would drop Figure 6 and give only numbers. It saves a Figure.**

We would like to keep Figure 6 as it gives a good visual summary of Figure 5. We are also using the numbers derived from this Figure in the abstract, and in the "back-of-the-envelope" calculation added in the revised manuscript.

**Also what bothers me is that GA7.1 is said to be better from nudged simulations but, in the end, only GA7.0 is presented here, because of the decadal run being only available with GA7.0. This is again a shortcoming of accepting to work with so many different time periods for different simulations.**

This has been addressed by using GA7.1N/2015–2018 instead of GA7.0U/1980-1990.

**P13 – L1-5 Have the authors consider to use satellite data, or to rely on previous publications to try to assess how the comparison to models is biased by extinction of the ceilometer signal into the lowest thick clouds? At least this should be discussed in the discussion part. This is not the case now.**

The lidar simulator accounts for signal attenuation. Therefore, the comparison with the models is not biased by extinction.

**P13 L10-11 Is the extraction made above the lat/lon of the balloon launch or does it follow the radiosonde trajectory? I guess it is the latter but you may want to clarify this in the text.**

It is the former. The resolution of the models is generally not high enough to make a difference. The balloon trajectory length was on average 58 km on the TAN1802 voyage, and the higher altitudes when the balloon was further away from the ship would likely not affect the analysis, which mostly found differences in clouds in the lowest part of the troposphere. We have clarified the text (P16L23-30).

**P13L11 Can you make a subplot for each of the dataset? It is difficult like this to spot differing behaviours between coloured markers.**

We have increased the size of the markers to make them easier to distinguish.

**P13 – L14 What relationship?**

Added "observed and modelled relationship" (P17L11-12). CBH and min{SLL,LCL} and the axes of the scatter plot. The relationship is between these two coordinates of the points.

**P13 – L19 How large?**

We have added quantification in parentheses (P17L12-14).

**P13- L23 how weaker?**

This is now quantified by adding two subplots in Figure 7 for SLL (Figure 7c) and LCL (Figure 7d). We have added text "weaker relationship than min{SLL,LCL}: 26% and 31% of observed profiles have CBH within 100 m of SLL and LCL, respectively (Figure 7c, d)." (P17L19-20).

**P13 – L25-27 The fact that LTS is not a good indicator should be discussed in the discussion part and I don't think it is the case for now. This relates to my major comment 3) where I suggest that more emphasis should be given in the discussion to all results obtained from these novel ship-based measurements.**

We think Figure 7 demonstrates relatively well why min{SLL,LCL} is a better predictor for CBH than LTS. The correlation coefficient in OBS for min{SLL,LCL} (Figure 7a) is 0.4, and for LTS (Figure 7b) -0.2. The main difference, however, is that min{SLL,LCL} is very close to CBH (within 100 m) in 40% of cases. LTS cannot be used as a 1:1 predictor for CBH due to having different units (K vs. m), and the correlation is not strong enough for a linear model. The updated Figure 7 now also shows the graphs based SLL and LCL as predictors, both of which show an inferior relationship with CBH.

**P13 – L28-34. Figure 8 is introduced, but then some general statement are made about synoptic scale forcing. It would be much better, for the reader, to stick to the Figure.**

The general statement explains why we are showing Figure 8. Therefore, we would like to keep it.

**P14 – L1 "As can be seen. . .where there is no sea ice". What can be said about Figure8a and b where there is no cloud but at the same time GA7.0 is not in agreement at all with observations? Also what is the unit in the x-axis of subplots Figure 8a-f and Figure 8m-r? In Figure 8g-l and s-x, you are not showing the modelled dots, only observations. I would have expected to see the model outputs as well. Or is it not useful here?**

Figure 8a, b now show much better agreement with GA7.1N (as opposed to GA7.0U), most likely due to the nudging. We have added units to the axes.

We have removed the scatter plots in Figure 8 (Figure 9 shows similar information more clearly).

**P14L3-4 "There is no substantial difference between. . ." This is not true for Figure 8a and b. . . which present non-sea ice cases. This should be discussed. "Plausible effect"? What do you mean?**

GA7.1N now shows a much better agreement in Figure 8.

By "plausible effect" we mean the theoretical expectation of how min{SLL,LCL} relates to convection as explained in the introduction of the quantity. We have removed this part of the sentence (P18L3-6).

**P14 L8 – What is meant by subgrid-scale processes? Please be more specific.**

We have commented on the subgrid-scale processes by adding a paragraph in Discussion (P23L4-14).

**P14 - L2 I am not sure about this subsection. I struggle with having it only focusing on January 2007. Since the novelty of the paper is the ship-based measurements I am not sure having this part here is relevant, especially that it is only about comparing Jan 2007 for two models. Plus, the GA7.0 one is not the one used in Figure 5, but the nudged one, and it is not clear what period is used for MERRA2. Why not showing also GA7.1 since it is spotted as reducing the SW biases (because of the modelling of larger supercooled LWP?)**

This has been partially remedied by analysing all models for the same time period as the observations.

In the updated analysis GA7.1 shows much better match with observations in terms of cloud occurrence (Figure 5). Therefore, it is expected that the improvement of TOA outgoing SW radiation bias over GA7.0 can be largely attributed to the improvement of cloud cover representation rather than improved super-cooled LWP.

**P14L26-30 "We should note. . ." These are comments for the discussion part, but even so, these considerations are also and already mentioned in other places of the paper and remain very general and a bit speculative. I am not sure these zonal plots deserve a separate section, also because of these time period issues mentioned above.**

We have removed the statement.

The zonal plots now show the same time period as the rest of the analysis.

**————————————- 4. Discussion ————————————-**

**In general the discussion should be more focused on your results at least in the beginning and spend less time on explaining previous works. Figure 10 (which is interesting indeed) should come earlier in the discussion. Also, you don't seem to do discuss Figure 10b, but only Figure 10a. Sentences like (P15-L18-21) "Combined. . ." are a bit speculative and more room should be rather given to discussing the results obtained from ship-based measurements, ie. Figure 5, Figure 7 and Figure 8, and 10. And then make the link to the TOA SW bias issue and relate it, possibly, to the LWP as modelled (cf. Figure 9 – if still considered relevant in a revised version).**

We have relocated introduction of Figure 10 to Results (P15L1-7) and replaced Figure 10b with equivalent plot based on MERRA-2.

The statement P15-L18-21 (in the original manuscript) was not speculative in the context of the results – we have shown that the cloud cover is underestimated in MERRA-2 (Figure 5), and the only way the model can overestimate the total (all-sky) TOA outgoing SW radiation (Figure 3, 4) at the affected latitudes and time of year is by overestimating cloud albedo. We have clarified this point in the Discussion (P20L4-13).

**P15-L34 to P16 L4. This is too much about other study, not enough discussing your results. Figure 10 comes after that and this is not appropriate. Also – as an example of additional discussion element – are the ship-based observations, which show larger discrepancies from MERRA2, in places where the near-surface temperature is the coldest? In other words, can you relate Figure 10 with your cloud results, instead of only speaking of the SW bias?Also, why is Figure 10 only showing the year 2007? Why not showing the decadal simulation, and the MERRA2 outputs as well (during the ship-based measurements)? What do they say? How is it consistent or not with the cloud simulations in these models? These sorts of analysis/discussions are really missing in the paper, in my opinion.**

We have added a subplot showing MERRA-2 and extended the time period to January 2018. We have relocated introduction of Figure 10 to Results (P15L1-7).

**P17 L9 "Because sea ice is an important factor. . .": What is meant by "secondary effect on cloud cover"? It seems to me you have the opportunity to say something about the effect of sea ice on very low clouds (and specifically the ones missed by satellites) – e.g. your Figure 8x – but you are not exploring this in the paper. This goes along with my major comments that not enough efforts are made to discuss the very interesting observations you have from ship over three years and in sea-ice free/covered regions.**

The focus of the paper is on improving SW radiation bias in GCMs. Even though the difference between ice-free and sea ice cases is interesting, clouds over sea ice covered regions have relatively small impact

on the SW radiation (the ice covered surface is already highly reflective, and presence or absence of clouds makes little difference). Therefore we prefer to limit the scope of the paper mostly on the ice-free regions. This is an area that might be completed in future studies within the group.

**———————— 5. Conclusion ————————**

**In the conclusion only you speak again about the subgrid-scale processes without specifying them. This should be a paragraph on its own in the discussion part, trying at least to understand how the various models are doing different in parameterising these processes. This would give more perspective to the present work I think.**

We have added a paragraph in Discussion commenting on this problem (P22L34-P23L8).

**————————- Figures. ————————-**

**Figure 2 If you still want to keep all the model results (provided you better justify your method – see my major comments) then you should add the time-periods for the simulations you use, and for the observations, so that one immediately knows you are using different times for comparisons (and that this is then discussed in the text).**

This has been addressed by leaving out GA7.0U/1980-1990 and instead comparing GA7.1N, MERRA-2 and observations over the same time period.

**Figure 3 As I said before, one struggles to see features with a single colour-shaded scale. As suggested I would keep only the plots showing the biases, and for summer and autumn (as these are seasons investigated with ship measurements).**

We now show bias in DJF and MAM (2015–2018) with a latitude range of 45–75°S. We have changed the colormap, which should better highlight the differences.

**Figure 4 The horizontal line indicates the "0" value for the bias (red curve). Please make it red (and thicker, or dashed).**

We have made the line dashed, thicker and red.

**Figure 5 What period is used for MERRA2? Why not also showing the nudged runs with the better (according to what you say) version GA7.1U.**

This has been addressed by using GA7.1 nudged for 2015–2018. GA7.1U was not available in our original analysis.

**Figure 6 Not sure this figure is needed. See my comment in the relevant section.**

We use results from this figure in the abstract to quantify the cloud cover bias, and to perform a "back-of-the-envelope" calculation added to the revised manuscript. Therefore, we think the figure adds valuable summary information.

**Figure 7 This would be better to separate the dataset in different subplots to see the different behaviours.**

We have increased the size of the markers to make them easier to distinguish.

**Figure 8 What are the x-axis units in the subplots a-f and m-r? The markers in the g-l and s-x subplots are quite small. Can you either make them larger or increase the size of the subplots.**

We have added x-axis units and increased scatter plot markers.

**Figure9 What are the contour values?**

We have added contour level values.

## Anonymous Referee #2

**Review Kuma et al: ' Evaluation of Southern Ocean cloud in the HadGEM3 general circulation model and MERRA-2 reanalysis using ship-based observations' ( MS No.: acp-2019-201) The authors conducted analysis of three model datasets by focusing on the Southern Ocean to understand errors in models in the shortwave (SW) radiative flux at the top-of-the-atmosphere, using ship observational dataset as well as satellite observations to understand the errors. They found that GA7 runs and MERRA-2 runs have the opposite bias in the outgoing SW flux (underestimate in GA7, overestimate in MERRA-2) over the southward latitude of 55S. They compared their cloud amounts with the ship observations and showed that both models underestimate their cloud amounts. They also conducted nudged-runs and showed that there is a big difference in cloud liquid water amount in these models, concluded that the main source of the difference in their SW bias is from the difference in their cloud properties, which are determined by the sub-grid cloud parameterizations. The shortwave bias over the Southern Ocean tends to be a common problem in climate models. This is a nice piece of work which contributes to improve our understanding of the representations of clouds over the region. However, current manuscript misses some information for their logic to convince readers, hence the key message remains unclear. I suggest this paper to be published after a minor revision.**

**Main comments: Although GA7 runs and MERRA-2 runs have the opposite bias in the outgoing SW flux over the southward latitude of 55S, both HadGEM3 GA7 and MERRA 2 underestimate cloud amount. In Discussion section, the authors mentioned that models may fail to represent fog or low cloud which are generated by convection which are induced from subzero airmass from polar regions over warm water. What our community is keen to know is whether we can improve the representations of such clouds in GCM or we should seriously start thinking of using cloud resolving model or GCM. Whether/how much the underestimate of the cloud amount improves in their nudged runs will provide a clue for it. The authors should add a figure which shows cloud amounts in free run and nudged runs.**

To address the major comments of Referee 1, we have replaced the free running model with a nudged model (GA7.1N), and Figure 5 now compares the nudged model, MERRA-2 and observations. Based on the updated results, we think it is possible to fix the parametrisation schemes to generate more low cloud (of the right type) and fog in the conditions typical in the region. GA7.1N is underestimating the cloud cover by about 4–9%, but this is still enough to cause a relatively large bias in SW radiation (Table 3).

We are currently performing a more detailed analysis of the lidar backscatter vs. a nudged model with the aim of getting the model to simulate the missing types of cloud, especially layers of stratocumulus and fog on certain days. This should be the subject of a paper in preparation. We think it is possible to fix the parametrisation schemes in GCMs, which have likely been tuned for other regions globally and neglected the SO due to the lack of ground-based observations (satellite observations do not identify these types of cloud correctly if obscured by a higher-level cloud).

**The authors showed that main difference in SW radiative flux bias over the Southern Ocean between HadGEM3 GA7 runs and MERRA 2 runs is cloud water amount. This shows a big impact of subgrid cloud parameterizations on radiation. Please check subgrid cloud parameterizations in GA7 and MERRA2 then discuss which parameterization could potentially cause the difference in radiative flux. Since the authors showed the opposing sign of the SW CRE south and north of 55S in GA7.1, it would be useful to apply the same analysis (comparison to the ship observations, analysis of the nudged runs) to the region of the north of 55S, confirm whether the smaller error is because**

**of the (less worse) representations of the cloud amount over the region.**

We have added a paragraph in Discussion which comments on the possible subgrid-scale parametrisation schemes in GA7.1N responsible for the bias (P23L4-8). A concrete identification of the problem will likely require experimenting with the parametrisation schemes to achieve a better match with the observed cloud occurrence profiles. The observed SW radiation bias is likely a combined effect of underestimation of cloud cover and overestimation of cloud albedo, resulting in the latitudinal gradient of bias, which is positive north of about 55°S (65°S) in GA7.1N (MERRA-2) and negative south of this latitude.

We already compare a large number of geographical subsets (5x6) and therefore by adding more we would make the plots even more complicated. The latitude of 55–60°S already covers a region of the positive bias in GA7.1N in the Ross Sea sector (Figure 3e). Figure 5a1, a2, a3 compare this region between the observations and GA7.1N and show that GA7.1N and observations had cloud cover of almost 100%, but with differences in fog/low cloud simulation (the model didn't simulate the correct altitude of cloud). Because the error in cloud cover in this region was so small, this points to cloud albedo overestimation as the reason for the positive SW radiation bias, even though Figure 5a1, a2, a3 are over a relatively small number of days (4.4).

**Minor comments: Discussion: the beginning (L1-10) was difficult to read, because the authors mention the opposing sign of the SW CRE south and north of 55S in GA7.1, but then solely talk about the results over the south of 55S.**

We have changed this part of Discussion also to address multiple comments of Referee 1 (P19L5-).

**Figure 6: Clarify what is the weight for the weighted average.**

We have clarified that the weight is the number of days the ship spent in the spatiotemporal subset.

**Figure 8: add grid values to the Frequency axis**

We have added units to the x-axis in Figure 8.

**P11-l1: 'upwelling and downwelling' Where are regions of upwelling and downwelling radiative flux? If the authors are talking about large scale circulation, these should be 'ascent and descent'.**

We have removed the part of the sentence (Figure 3 is now focused on the SO only) (P13L17-19).

**P11-l4: I cannot see the results described about models. And the contrast between western and eastern sides of the Antarctic Peninsula contradicts to the following description 'The zonal symmetry. . ..'**

We have updated Figure 3 to show biases and increased the contrast and scale of the plots. We have replaced "zonally symmetric" with "relatively zonally symmetric" (P13L17).

**P11-l14: Figure 3p?**
**P11-l32: 'consistently positive': negative in Sep-Dec in 60S-70S**
**P11-l33: 'also lower than GA7.0 and GA7.1': not necessarily in GA7.1**

We have updated this part to account for the updated Figure 3.

**P13-l14: Did you define SLL and LCL? (Super liquid level and lifting condensation level?) How did you define SLL?**

We use the traditional definition of the lifting condensation level. We now define the "SST lifting level" (SLL) in the Results section before the first use of the quantity (P16L31-P17L7).

**P13-l22: Give a speculation why min(SLL, LCL) is better correlated with CBH than SLL/LCL individually.**

We have added a section in Discussion which details why we think the relationship of CBH with min{SLL, LCL} is better than SLL/LCL (P20L30-P21L3).

**P14-l5: Provide a figure or reference about SLL in GA7.0 is higher than observed.**

In the updated Figure 8 we plot distribution of min{SLL,LCL} instead of SLL (to be more consistent with the rest of the analysis). The updated description of the figure comments on the observed and modelled distribution (P17L26-P18L10). GA7.1N represents the observed distribution relatively well.

**P14-l16: Fig 9. It is not clear why the authors create these plots over two different backgrounds.**

We prefer to show both fields due to their effect on cloud (potential temperature through convection and relative humidity through condensation).

**P14-l18: Fig 9. Not clear. Different colors should be used for different levels to show this.**

We have lowered the value of the lowest contour to 12 $gm^{-3}$, which means some cloud ice contours are now visible on the MERRA-2 plots.

**P14-l29: cloud cover a reduce ..': typo?**

We have removed this sentence due to the overall change of a dataset.

**Fig 5: Why did the authors exclude 50S-55S for the plots?**

We did not include 50–55°S in order to keep the paper relatively focused. The radiosonde observations on TAN1802 and NBP1704 voyages were only available south of 60°S, which limited some of the plots.

We agree that 50–55°S might be an interesting addition in Figure 5. We can extend this figure before a final revision of the manuscript if the referees think it useful in the updated analysis.

**Fig 8: The authors did not analyze model results in other latitudes where clouds shows the opposite bias (in 50S-55S).**

We could not provide plots in Figure 8 north of 60°S due to radiosonde observations only available south of this latitude.

**P15-l10-11: I cannot follow the logic here.**

We have added a paragraph in Discussion explaining this point (P20L4-13). The logic is that the effect of clouds on reflected SW radiation is the product of cloud cover (the cloudy fraction of the sky) and cloud albedo (reflectivity of the cloud). We have shown that cloud cover is underestimated in the models, while at the same time MERRA-2 overestimates the reflected SW radiation. Therefore, if the first factor of the product (cloud cover) is underestimated, the second factor (cloud albedo) must be overestimated to get overestimated reflected SW radiation.

**P16 l23: Is it possible to add the definitions of supercooled liquid in GA7.0 and MERRA-2?**

We consider any cloud liquid at air temperature below zero supercooled. We have clarified this at the first mention of the term in the Introduction (P3L8). We have also added a statement in Figure 9 caption clarifying that all cloud liquid in the plot is supercooled.

**P17 l11: Is this a result from the nudged run or from other studies?**

We have removed the sentence. In the updated analysis we are comparing with the nudged run which uses sea ice concentration prescribed from satellite observations.

# Evaluation of Southern Ocean cloud in the HadGEM3 general circulation model and MERRA-2 reanalysis using ship-based observations

Peter Kuma[1], Adrian J. McDonald[1], Olaf Morgenstern[2], Simon P. Alexander[3], John J. Cassano[4], Sally Garrett[5], Jamie Halla[5], Sean Hartery[1], Mike J. Harvey[2], Simon Parsons[1], Graeme Plank[1], Vidya Varma[2], and Jonny Williams[2]

[1]School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand
[2]National Institute of Water and Atmospheric Research, Wellington, New Zealand
[3]Australian Antarctic Division, Kingston, Australia
[4]Cooperative Institute for Research in Environmental Sciences and Department of Atmospheric and Oceanic Sciences, University of Colorado, Boulder, Colorado, US
[5]New Zealand Defence Force, Wellington, New Zealand

**Correspondence:** Peter Kuma (pku33@uclive.ac.nz)

**Abstract.** Southern Ocean (SO) shortwave (SW) radiation biases are a common problem in contemporary general circulation models (GCMs), with most models exhibiting a tendency to absorb too much incoming SW radiation. These biases have been attributed to deficiencies in the representation of clouds during the austral summer months, either due to cloud cover or cloud ~~optical thickness~~ albedo being too low. The problem has been the focus of many studies, most of which utilised satellite datasets for model evaluation. We use multi-year ship based observations and the CERES spaceborne radiation budget measurements to contrast cloud representation and SW radiation in the atmospheric component Global Atmosphere (GA) version ~~7.0 and~~ 7.1 of the HadGEM3 GCM and the MERRA-2 reanalysis. We find that ~~MERRA-2 is biased in the opposite direction to GA (reflects too much SW radiation). In addition,~~ the prevailing bias is negative in GA7.1 and positive in MERRA-2. GA7.1 performs better than MERRA-2 ~~performs better~~ in terms of absolute SW bias~~than nudged runs of GA7.0 and GA7.1 in the 60–70°S latitude band.~~. Significant errors of up to 21 $\mathrm{Wm}^{-2}$ (GA7.1~~reduces the SO SW radiation biases relative to GA7.0, but significant errors remain at up to 20~~) and 39 $\mathrm{Wm}^{-2}$ ~~between 60 and 70°S~~ (MERRA-2) are present in both models in the austral summer~~months~~. Using ship-based ceilometer observations, we find low cloud below 2 $\mathrm{km}$ to be predominant in the Ross Sea and the Indian Ocean ~~sector~~ sectors of the SO. Utilising a novel surface lidar simulator developed for this study, derived from an existing COSP-ACTSIM spaceborne lidar simulator, we find that GA7~~.0~~.1 and MERRA-2 both underestimate low cloud and fog occurrence relative to the ship observations on average by ~~18–25% on average , though the cloud cover in~~ 4–9% (GA7.1) and 18% (MERRA-2~~is closer to observations by about 7%~~). Based on radiosonde observations, we also find the low cloud to be strongly linked to boundary-layer atmospheric stability and the sea surface temperature. GA7~~.0~~.1 and MERRA-2 ~~agree well with observations in terms of boundary-layer stability, suggesting that subgrid-scale parametrisations do not generate enough cloud in response to the thermodynamic profile of the atmosphere and the surface temperature. Our analysis shows~~ do not represent the observed relationship between boundary layer stability and clouds well. We find that MERRA-2 has a

much greater proportion of cloud liquid water in the SO in ~~January~~ austral summer than GA7~~.0~~.1, a likely key contributor to the difference in ~~SW radiation. We show that boundary-layer stability and relative humidity fields are very similar in GA7.0 and MERRA-2, and unlikely to be the cause of the different cloud representation, suggesting that~~ the SW radiation bias. Our results suggest that subgrid-scale ~~parametrisations are responsible for the difference between the models~~processes (cloud and boundary layer parametrisations) are responsible for the bias, and that in GA7.1 a major part of the SW radiation bias can be explained by cloud cover underestimation, relative to underestimation of cloud albedo.

## 1 Introduction

Clouds are considered one of the largest sources of uncertainty in estimating global climate sensitivity (Boucher et al., 2013; Flato et al., 2013; Bony et al., 2015). Clouds over oceans are especially important for determining the radiation budget due to the low albedo of the sea surface compared to land. Over the Southern Ocean (SO), cloud cover is very high at over 80%, with boundary-layer clouds being particularly common (Mace et al., 2009). Excess downward shortwave (SW) radiation in general circulation models (GCMs), with a bias over the SO of up to 30 $\mathrm{Wm}^{-2}$, is a problem well-documented by Trenberth and Fasullo (2010) and Hyder et al. (2018), and has been the subject of many studies. Bodas-Salcedo et al. (2014) evaluated the SW bias in a number of GCMs and found that a strong SW bias is a very common feature, leading to increased sea surface temperature (SST) in the SO and corresponding biases in the storm track position. Trenberth and Fasullo (2010) note that a poor representation of clouds might lead to unrealistic climate change projections in the Southern Hemisphere. The SW bias has also been linked to large-scale model problems such as the double-Intertropical Convergence Zone (Hwang and Frierson, 2013), biases in the position of the midlatitude jet (Ceppi et al., 2012) and errors in the meridional energy transport (Mason et al., 2014). Bodas-Salcedo et al. (2012) studied the SO SW bias in the context of the Global Atmosphere (GA) 2.0 and 3.0 models and found that mid-topped and stratocumulus clouds are the dominant contributors to the bias.

Due to its extent and magnitude, the SW radiation bias is believed to limit accuracy of the models, especially for modelling the Southern Hemisphere climate. A model based on the Hadley Centre Global Environmental Model version 3 (HadGEM3) is currently used in New Zealand for assessing future climate (Williams et al., 2016). In this paper we evaluate ~~two versions of~~ the atmospheric component of HadGEM3, GA7~~.0 and GA7~~.1 (Walters et al., 2017) and the reanalysis Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2) using observations collected in the SO on a number of voyages. ~~The main objective of this study is to evaluate SO cloud in GA7.0 and GA7.1 based on ship-based remote sensing and in situ observations.~~ Ship-based atmospheric observations in the SO provide a unique view of the atmosphere not available via any other means. Boundary layer observations by satellite instruments are limited by the presence of an almost continuous cloud cover, potentially obscuring the view of low level clouds. The frequently used active instruments CloudSat (Stephens et al., 2002) and Cloud–Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) (Winker et al., 2010) are both of limited use when observing low level, thick or multi-layer cloud: CloudSat is affected by surface clutter below approximately 1.2 km (Marchand et al., 2008) and the CALIPSO lidar signal cannot pass through thick cloud. Likewise, passive instruments and datasets such as the Moderate Resolution Imaging Spectroradiometer (MODIS) (Salomonson et al., 2002) and the

International Satellite Cloud Climatology Project (ISCCP) (Rossow and Schiffer, 1999) can only observe radiation scattered or emitted from the cloud top of optically thick clouds. Therefore, one can accurately identify the cloud top height or cloud top pressure with satellite instruments, but not always the cloud base height (CBH) or the vertical profile of cloud, although there has been some recent progress on deriving CBH statistically from CALIPSO measurements (Mülmenstädt et al., 2018). Ship-based measurements therefore provide valuable extra information.

Multiple explanations of the SW radiation bias have been proposed: cloud underestimation in the cold sectors of cyclones (Bodas-Salcedo et al., 2014), cloud–aerosol interaction (Vergara-Temprado et al., 2018), cloud homogeneity representation (Loveridge and Davies, 2018), lack of supercooled liquid (cloud liquid at air temperature below 0 °C) (Kay et al., 2016; Bodas-Salcedo et al., 2016) and the "too few, too bright" problem (Nam et al., 2012; Klein et al., 2013; Wall et al., 2017). Each model can exhibit the bias for a different set of reasons, and results from one model evaluation therefore do not necessarily explain biases in all other models (Mason et al., 2015). The use of SO voyage data for atmospheric model evaluation is not new, and has recently been used by Sato et al. (2018) to evaluate the impact of SO radiosonde observations on the accuracy of weather forecasting models. Klekociuk et al. (2018) contrasted SO cloud observations with the ECMWF Interim reanalysis (ERA-Interim) and the Antarctic Mesoscale Prediction System–Weather Research and Forecasting Model (AMPS-WRF), and found that these models underestimate the coverage of the predominantly low cloud. Protat et al. (2017) compared ship-based 95 GHz cloud radar measurements at 43–48°S in March 2015 with the Australian Community Climate and Earth-System Simulator (ACCESS) NWP model, a model related to HadGEM3, and found low cloud peaking at 80% cloud cover, which was underestimated in the model. ~~It was~~ The clouds were also more spread out vertically (especially due to "multilayer" situations defined as co-occurrence of cloud below and above 3 km) and more likely to have intermediate cloud fraction ~~in the model~~rather than very low or very high cloud fraction. Previous studies have documented that supercooled liquid is often present in the SO cloud in the austral summer months (Morrison et al., 2011; Huang et al., 2012; Chubb et al., 2013; Huang et al., 2016; Bodas-Salcedo et al., 2016; Jolly et al., 2018) and is linked to SO SW radiation biases in GCMs, which underestimate the amount of supercooled liquid in clouds in favour of ice. Warm clouds generally reflect more SW radiation than cold clouds containing the same amount of water (Vergara-Temprado et al., 2018). In particular, Kay et al. (2016) reported a successful reduction of SO absorbed SW radiation in the Community Atmosphere Model version 5 (CAM5) by ~~tuning~~ decreasing the shallow convection ice detrainment temperature and thereby increasing the amount of supercooled liquid cloud.

Two common techniques used for model cloud evaluation have been cloud regimes (Williams and Webb, 2009; Haynes et al., 2011; Mason et al., 2014, 2015; McDonald et al., 2016; Jin et al., 2017; McDonald and Parsons, 2018; Schuddeboom et al., 2018, 2019) and cyclone compositing (Bodas-Salcedo et al., 2012; Williams et al., 2013; Bodas-Salcedo et al., 2014, 2016; Williams and Bodas-Salcedo, 2017), both of which link the SW radiation bias to specific cloud regimes and cyclone sectors. ~~Jakob (2003) discusses different methods of cloud evaluation.~~ We use simple statistical techniques, rather than sophisticated classification or machine learning algorithms, the advantage of which is easier interpretation for the purpose of model development.

We first assess the magnitude of the Top of Atmosphere (TOA) SO SW radiation bias in ~~the GA7.0 and~~ a nudged run of GA7.1 ~~models and the Modern-Era Retrospective analysis for Research and Applications, version 2~~ ("GA7.1N") and MERRA-

2 ) reanalysis with respect to the Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) and CERES Synoptic (SYN) products (Section 5.1). This allows us to identify the underlying magnitude of the SW bias and how this might change based on the ship track sampling pattern. We then evaluate cloud occurrence in GA7.0 .1N and MERRA-2 relative to the SO ceilometer observations and compare SO radiosonde observations with pseudo-radiosonde profiles derived from the models (Sections 5.2 and 5.3). Lastly, we look at zonal plots of potential temperature, humidity, cloud liquid and ice content in GA7.0 .1N and MERRA-2 to show how these models differ in their atmospheric stability and representation of clouds (Section 5.4). Our aim is to identify how differences between GA7.0 .1N and MERRA-2 can explain the TOA SW outgoing SW radiation bias, assuming misrepresentation of clouds is the major contributor to the bias.

## 2 MethodsDatasets

We used an observational dataset of ceilometer and radiosonde data comprising multiple SO voyages (Figure 1Section 2.1), GA7.0 and GA7.1N atmospheric model simulations (Walters et al., 2017) (Section 2.2) and the MERRA-2 reanalysis (Gelaro et al., 2017) (Section 2.3). Later in the text, we will refer to GA7.0, GA7.1N and MERRA-2 together as "the models", even though MERRA-2 is more specifically a reanalysis. CERES satellite observations (Wielicki et al., 1996) were also used as a reference for TOA outgoing SW radiation and an National Snow and Ice Data Center (NSIDC) satellite-based dataset (Maslanik and Stroeve, 1999) was used as an auxiliary dataset for identifying sea ice. CFMIP Observation Simulator Package (COSP) (Bodas-Salcedo et al., 2011), a set of instrument simulators developed by the Cloud Feedback Model Intercomparison Project (CFMIP), was extended with a surface lidar simulator and used to produce virtual lidar measurements from model fields (Kuma et al., 2019). Resampling, noise reduction and cloud detection were performed on observational and (where applicable) model lidar data in a consistent way to reduce structural uncertainty (see Section 3.2). The schematic in Figure 2 shows the processing pipeline utilised in this study.

### 2.1 Datasets

#### 2.0.1 HadGEM3

HadGEM3 (Walters et al., 2017) is a general circulation model developed by the UK Met Office and the Unified Model Partnership. It is used either in a free-running mode or "nudging" (Telford et al., 2008) – relaxing winds and potential temperature towards the ERA-Interim reanalysis (Dee et al., 2011). The Met Office Global Atmosphere 7.0 and 7.1 (GA7.0 and GA7.1, respectively) is the atmospheric component of HadGEM3 (Walters et al., 2017).

The following runs were used in our analysis:

- 1980–89 run of GA7.0 free-running ("GA7.0U/1980-1989").
- 2007 run of GA7.0 nudged ("GA7.0N/2007").
- 2007 run of GA7.1 nudged ("GA7.1N/2007").

The model runs used the HadISST sea surface temperature dataset (Rayner et al., 2003) as lateral boundary conditions. The nudged simulations represent atmospheric dynamics as determined by observations. In the free-running model, atmospheric dynamics can only be compared statistically with observations or reanalyses. The model was run on a 1.875° × 1.25° (longitude × latitude) "N96" resolution grid, which corresponds to a horizontal resolution of about 100×140 at 60°S and 85 vertical levels. The model output was provided as instantaneous fields sampled every 6 hours. Limited data availability meant that no nudged runs were available for the period of 2015–2018 when the ship observations are available. Therefore, we used the decadal free-running simulation to compare cloud representation statistically.

### 2.0.1 MERRA-2

Modern-Era Retrospective analysis for Research and Applications (MERRA-2) is a reanalysis provided by the NASA Global Modelling and Assimilation Office (Gelaro et al., 2017). The reanalysis was chosen for its contrasting results of TOA shortwave radiation bias in the SO compared to GA7.0 and GA7.1. Its bias is positive rather than negative, when CERES is used as a reference.

We used the following products (Bosilovich et al., 2015):

- 1-hourly average Radiation Diagnostics (product "M2T1NXRAD.5.12.4")
- 3-hourly instantaneous Assimilated Meteorological Fields (product "M2I3NVASM.5.12.4")
- 1-hourly instantaneous Single-Level Diagnostics (product "M2I1NXASM.5.12.4")
- 3-hourly average Assimilated Meteorological Fields (product "M2T3NVASM.5.12.4")
- 1-hourly average Single Level Diagnostics (product "M2T1NXSLV.5.12.4")

We used the "Radiation Diagnostics" in TOA SW radiation evaluation (Section 5.1), the instantaneous "Assimilate Meteorological Fields" and "Single-Level Diagnostics" products to generate simulated ceilometer profiles and pseudo-radiosoundings (Section 5.2 and 5.3), and the average "Assimilate Meteorological Fields" and "Single-Level Diagnostics" to generate zonal plane plots of thermodynamic and cloud fields (Section 5.4). Before running the COSP simulator, we downsampled the grid resolution to a 1° × 1° grid in order to reduce the computational demands. The 4-dimensional MERRA-2 fields were provided on pressure and model levels. For our analysis we chose to use the model-level products (72 levels) due to their higher vertical resolution compared to pressure-level products.

### 2.0.1 ~~Ship observations~~

## 2.1 Ship observations

We use ship-based ceilometer and radiosonde observations made in the SO on 5 ~~separate~~ voyages between 2015 and 2018 (Table 1 and Figure 1):[1]

- 2015 TAN1502 voyage of the NIWA ship RV *Tangaroa* from Wellington, New Zealand to the Ross Sea.
- 2015–2016 voyages (V1–V3) of the Australian Antarctic Division (AAD) icebreaker *Aurora Australis* from Hobart, Australia to Mawson, Davis, Casey and Macquarie Island ("AA15")
- 2016 Royal New Zealand Navy (RNZN) ship HMNZS *Wellington* voyages ("HMNZSW16").
- 2017 NBP1704 voyage of the NSF icebreaker RV *Nathaniel B. Palmer* from Lyttelton, New Zealand to the Ross Sea.
- 2018 TAN1802 voyage of RV *Tangaroa* from Wellington to the Ross Sea (Hartery et al., 2019).

Together, these voyages cover latitudes between 41 and 78°S and the months of November to June inclusive. A total of 298 days of observations were collected. Geographically, the voyages mostly cover the Ross Sea sector of the SO, with only AA15 covering the Indian Ocean sector (Figure 1). This sampling emphasises the Ross Sea sector over other parts of the SO, although the SO SW radiation bias ~~appears largely zonally symmetric~~ is present at all longitudes in the SO (Section 5.1), ~~with a notable exception of the eastern side of the Antarctic Peninsula, as is the~~ affected by atmospheric circulation in the SO (Jones and Simmonds, 1993; Sinclair, 1994, 1995; Simmonds and Keay, 2000; Simmonds et al., 2003; Simmonds, 2003; Hoskins and Hodges, 2005; Hodges et al., 2011)~~, which should allow these results to be extrapolated over the whole of SO at the affected latitudes. Figure 1 shows the tracks of the voyages used in this study~~. The voyage observations were performed using a range of instruments (described below). Table 2 details which instruments were deployed on each voyage.

The primary instruments were the Lufft CHM 15k and Vaisala CL51 ceilometers. A ceilometer is an instrument which typically uses a single-wavelength laser to emit pulses vertically into the atmosphere and measures subsequent backscatter resolved on a large number of vertical levels based on the timing of the retrieved signal (Emeis, 2010). Depending on the wavelength, the emitted signal interacts with cloud droplets, ice crystals and precipitation by Mie scattering, and to a lesser extent with aerosol and atmospheric gases by Rayleigh scattering (Bohren and Huffman, 2008). The signal is quickly attenuated in thick cloud and therefore it is normally not possible to observe mid and high level parts of such a cloud, or a multi-layer cloud. The main derived quantity determined from the backscatter is CBH, but it is also possible to apply a cloud detection algorithm to determine cloud occurrence by height. The range-normalised signal is affected by noise which increases with the square of range. A major source of noise is solar radiation which causes a diurnal variation in noise levels (Kotthaus et al., 2016). Due to ~~noise and signal attenuation , the cloud profile retrieved by a ceilometer does not directly reveal the~~

[1]The voyage name pattern is a 2–6 character ship name followed by a 2 digit year and a 2 digit sequence number. TANxxxx and NBPxxxx are official voyage names, while HMNZSW16 and AA15 are names made for the purpose of this study.

~~cloudliquid and ice mixing ratios in an atmospheric model output, and~~ signal attenuation and noise ceilometers cannot measure clouds obscured by a lower cloud, and therefore cannot be used for 1:1 comparison with model clouds without using a lidar simulator~~has to be used to account for these effects~~ , which accounts for this effect (Chepfer et al., 2008). The Lufft CHM 15k ceilometer operates in the near-infrared spectrum at 1064 nm, measuring lidar backscatter up to a maximum height of 15 km, producing 1024 regularly spaced bins (about 15 m resolution). The sampling rate of the instrument is 2 s. The Vaisala CL51 ceilometer operates in the near-infrared spectrum at 910 nm. The sampling rate of the instrument is 2 s and range is 7.7 km, producing 770 regularly spaced bins (10 m resolution).

Radiosonde observations were performed on the TAN1802 and NBP1704 voyages south of 60°S. Temperature, pressure, relative humidity and GNSS coordinates (from which wind speed and direction are derived) were retrieved to altitudes of about 10–20 km, terminated by a loss of radio communication or balloon burst.

On the TAN1802 voyage we used ~~the~~ iMet-1 ABx radiosondes, measuring pressure, air temperature, relative humidity and GNSS coordinates of the sonde (from which wind speed and direction are derived). The sondes were launched three times per day at about 8:00, 12:00 and 20:00 UTC on 100 g Kaymont weather balloons. They reached a typical altitude of 10–20 km, and then terminated by balloon burst or loss of radio communication. We used 10 s resolution profiles generated by the vendor-supplied iMetOS-II control software for further processing. ~~We also had access to automatic~~

Automatic weather station (AWS) data ~~from some of the voyages (RV *Tangaroa* and RV *Nathaniel B. Palmer*)~~were available on the TAN1502, TAN1802 and NBP1704 voyages. These included variables such as air temperature, pressure, sea surface temperature, wind speed and wind direction. Voyage track coordinates were obtained from the ships' Global Navigation Satellite System (GNSS) receivers.

#### 2.1.1 ~~CERES~~

### 2.2 HadGEM3

HadGEM3 (Walters et al., 2017) is a general circulation model developed by the UK Met Office and the Unified Model Partnership. It can be used in a "nudging" (Telford et al., 2008) mode, in which winds and potential temperature are relaxed towards the ERA-Interim reanalysis (Dee et al., 2011). The Met Office Global Atmosphere 7.1 (GA7.1) is the atmospheric component of HadGEM3 (Walters et al., 2017), based on the Unified Model (UM) version 11.0.

The model runs used the HadISST sea surface temperature dataset (Rayner et al., 2003) as lateral boundary conditions. The nudged simulations represent atmospheric dynamics as determined by observations. The model was run on a 1.875°×1.25° (longitude × latitude) "N96" resolution grid, which corresponds to a horizontal resolution of about 100×140 km at 60°S and 85 vertical levels. The model output fields were sampled every 6 hours (instantaneous) and daily (mean). In our analysis we used a nudged run of GA7.1 ("GA7.1N") between years 2015 and 2018, corresponding to the ship observations.

### 2.3 MERRA-2

Modern-Era Retrospective analysis for Research and Applications (MERRA-2) is a reanalysis provided by the NASA Global Modelling and Assimilation Office (Gelaro et al., 2017). The reanalysis was chosen for its contrasting results of TOA outgoing SW radiation bias in the SO compared to GA7.1. As shown later (Figure 3), its bias is positive rather than negative, when CERES is used as a reference.

We used the following products (Bosilovich et al., 2015):

- 1-hourly average Radiation Diagnostics (product "M2T1NXRAD.5.12.4")
- 3-hourly instantaneous Assimilated Meteorological Fields (product "M2I3NVASM.5.12.4")
- 1-hourly instantaneous Single-Level Diagnostics (product "M2I1NXASM.5.12.4")
- 3-hourly average Assimilated Meteorological Fields (product "M2T3NVASM.5.12.4")
- 1-hourly average Single Level Diagnostics (product "M2T1NXSLV.5.12.4")

We used the "Radiation Diagnostics" in TOA outgoing SW radiation evaluation (Section 5.1), the instantaneous "Assimilate Meteorological Fields" and "Single-Level Diagnostics" products to generate simulated ceilometer profiles and pseudo-radiosonde profiles (Section 5.2 and 5.3), and the average "Assimilate Meteorological Fields" and "Single-Level Diagnostics" to generate zonal plane plots of thermodynamic and cloud fields (Section 5.4). The 4-dimensional MERRA-2 fields were provided on pressure and model levels. For our analysis we chose to use the model-level products (72 levels) due to their higher vertical resolution compared to pressure-level products. The analysed time period of MERRA-2 data was 2015–2018.

### 2.4 CERES

The Clouds and the Earth's Radiant Energy System (CERES) is a set of low Earth orbit (LEO) satellite instruments and a dataset of SW and longwave (LW) radiation observations (Loeb et al., 2018; Doelling et al., 2016). The CERES instruments (called FM1 to FM6) provide a continuous record of observations since the first deployment on the Tropical Rainfall Measuring Mission (TRMM) satellite in 1997 (Simpson et al., 1996), and have been flown on Terra, Aqua (Parkinson, 2003), the Suomi NPOESS Preparatory Project (Suomi NPP) and Joint Polar Satellite System-1 (JPSS-1) (Goldberg et al., 2013) satellites since. Currently CERES is considered the best available global Earth radiation datasets, and is often used as the primary dataset for GCM tuning and validation (Schmidt et al., 2017; Hourdin et al., 2017). We used the following CERES products in our analysis:

- CERES SYN1deg-Day Edition 4A (configuration code ~~401405~~406406 and 407406) product of daily average radiation ("CERES SYN").
- CERES EBAF-TOA Edition ~~4.0~~ 4.1 (CERES_EBAF_Ed4~~.0~~.1) product of monthly energy-balanced average radiation ("CERES EBAF").

Due to the sun-synchronous orbits of the LEO satellite platforms, the Flight Model (FM) instruments of CERES do not capture the full diurnal variation of radiation. The EBAF and and SYN1deg products are adjusted for diurnal variation by using 1-hourly geostationary satellite observations between 60°S and 60°N, and use an algorithm to account for changing solar zenith angle and diurnal land heating. The CERES EBAF-TOA Edition ~~4.0~~ 4.1 product is a Level 3B product, which means it has been globally balanced by ocean heat measurements using the Argo network (Roemmich and Team, 2009).

#### 2.4.1 ~~NSIDC sea ice concentration~~

### 2.5 NSIDC sea ice concentration

We used the Near-Real-Time Defense Meteorological Satellite Program (DMPS) Special Sensor Microwave Imager/Sounder (SSMIS) Daily Polar Gridded Sea Ice Concentrations, Version 1 product (NSIDC-0081) (Maslanik and Stroeve, 1999) provided by the National Snow and Ice Data Center (NSIDC) to classify observations into those affected and unaffected by sea ice. The sea ice concentration product has a resolution of 25 $\times$ 25 km. We used a cutoff value of 15% of sea ice concentration for the binary classification of sea ice, in line with previous studies (Comiso and Nishio, 2008).

### 2.6 ~~Domains~~

~~Because our observational dataset does not span the entire geographical area of the SO or all months of the year, and the atmospheric conditions in the SO are geographically variable, we subset our datasets into a number of geographical regions by latitude and time periods by season. The three geographical regions identified are 55–60°S, 60–65°S and 65–70°S and the time periods are austral summer, months December–January–February (DJF) and autumn months March–April–May (MAM). Although we have a substantial quantity of data taken at latitudes south of 70°S, we do not use them here, as they would likely be affected by circulation induced by land near the Ross Sea (Coggins et al., 2014), and therefore may not be representative of the SO in general. This decision builds on the analysis detailed in Jolly et al. (2018) which shows a significant gradient in cloud properties between the Ross Ice Shelf and the Ross Sea and strong influences associated with synoptic conditions. Likewise, we would have to exclude land areas, which have very different atmospheric climatologies.~~

## 3 Methods

~~There is likely temporal variability present within the austral summer and austral autumn periods, but we decided to limit the number of temporal classes to maintain a reasonable quantity of observations in each class in this analysis. The magnitude of the SO TOA SW radiation bias is primarily modulated by incoming solar radiation, which is the highest in the austral summer period. The voyages do not uniformly cover all geographical regions or time periods, with the largest number of observations in the Ross Sea sector south of New Zealand (TAN1802, TAN1502, HMNZSW16, NBP1704), followed by the Indian Ocean sector south of Western Australia (AA15) . Temporally, the voyage observations mostly cover summer to late summer/autumn months of the year. When subsetting model data, we sample along voyage tracks (geographically and temporally), and in~~

### 3.1 Lidar simulator

CFMIP Observation Simulator Package (COSP) (Bodas-Salcedo et al., 2011), a set of instrument simulators developed by the ~~case of the free-running GA7.0 simulation, we compare 10 years of model data statistically, and the same time period relative to the start of the year~~Cloud Feedback Model Intercomparison Project (CFMIP), was extended with a surface lidar simulator and used to produce virtual lidar measurements from model fields (Kuma et al., 2019). Resampling, noise reduction and cloud detection were performed on observational and (where applicable) model lidar data in a consistent way to reduce structural uncertainty (see Section 3.2). The schematic in Figure 2 shows the processing pipeline utilised in this study.

### 3.2 ~~COSP simulator~~

COSP was originally developed as a satellite simulator package whose aim is to produce virtual satellite (and more recently ground-based) observations from atmospheric model fields in order to improve comparisons of model output with observations (Bodas-Salcedo et al., 2011). This approach is required because physical quantities derived from satellite observations generally do not directly correspond to model fields. COSP accounts for the limited view of the satellite instrument by calculating radiative transfer through the atmosphere, i.e. attenuation by hydrometeors and air molecules and backscattering. COSP comprises multiple instrument simulators, such as MODIS, ISCCP, MISR, CALIPSO and CloudSat. It has been used extensively by previous studies of model cloud, for example by Kay et al. (2012), Franklin et al. (2013), Klein et al. (2013), Williams and Bodas-Salcedo (2017), Jin et al. (2017) , and Schuddeboom et al. (2018). COSP is planned to be used in the upcoming Coupled Model Intercomparison Project Phase 6 (CMIP6) (Webb et al., 2017).

For our analysis, we have developed a ground-based lidar simulator ~~based on the COSP CALIPSO~~ by modifying the COSP ACTSIM spaceborne lidar simulator (Chiriaco et al., 2006) (see the Code and data availability section at the end of the document). This required reversing of the vertical layers, as the surface lidar looks from the surface up rather than down from space to the surface, and changing the radiation wavelength affecting Mie scattering by cloud droplets and Rayleigh scattering by air molecules. In this paper we present only a brief description of the surface lidar simulator, with a more complete description planned in an upcoming paper. ~~These changes will be contributed to the upstream COSP project, or made publicly available , so that the scientific community can reuse the surface lidar simulator in the future~~The new simulator is made available as part of the Automatic Lidar and Ceilometer Framework (ALCF) at https://alcf-lidar.github.io.

The recently introduced COSP version 2 (Swales et al., 2018) added support for a surface lidar simulator, although we believe our implementation, developed before COSPv2 was available, is more complete in the present context due to its treatment of Mie scattering at wavelengths other than 532 nm (the wavelength of the CALIPSO lidar). Previously, a surface lidar simulator based on COSP has been used by Chiriaco et al. (2018) and Bastin et al. (2018). A ground-based radar simulator in COSP has also recently been implemented (Zhang et al., 2018).

The surface lidar simulator takes model cloud liquid and ice mixing ratios, cloud fraction and thermodynamic profiles as the input, and calculates vertical profiles of attenuated backscatter. This can be done either by running the simulator "online" within the model code or "offline" on the model output. We used the offline approach in our analysis.

### 3.2 Lidar data processing

Lidar data in this study came from two different instruments: Lufft CHM 15k and Vaisala CL51 ceilometers and the lidar simulator. These instruments use different output formats, wavelengths, sampling rates and range bins, as previously noted. Backscatter and derived fields such as CBH are provided in the firmware generated data products, but the backscatter is uncalibrated and the derived fields such as cloud detection are based on instrument-dependent algorithms. Therefore, we performed consistent subsampling, noise reduction and cloud detection on data from both instruments, and applied the same methods to the lidar simulator output. As part of the processing we developed a publicly available tool called cl2nc ("CL to NetCDF") for converting the Vaisala CL51 ceilometer data format to NetCDF (see the Code and data availability section at the end of the document).

#### 3.2.1 Calibration

The backscatter profiles produced by the Lufft CHM 15k and Vaisala CL51 ceilometers are not calibrated to physical units, even though they are expressed in $\mathrm{m}^{-1}\mathrm{sr}^{-1}$. To calibrate these backscatter fields we used the method described by O'Connor et al. (2004). This method uses the lidar ratio (LR) to calculate a calibration factor based on a known value of the LR in fully scattering cloudy scenes ($18.8 \pm 0.8$ sr), such as thick stratocumulus clouds, which are common over the SO. We applied this technique by using visually identified scenes and choosing a calibration factor which achieves the known value. Due to the nature of the conditions (LR can be highly variable even in thick cloud scenes), the calibration is likely accurate to only about 50% of the backscatter value. We do not expect this to have a serious impact on the accuracy of cloud detection completed in this study, largely because the predominantly low cloud tends to cause backscatter orders of magnitude greater than clear air, and because of the very large differences in cloud occurrence between the observations and models. ~~Kotthaus et al. (2016) provide a detailed description of backscatter retrieval by Vaisala ceilometers.~~

#### 3.2.2 Subsampling, noise removal and cloud detection

In order to simplify further processing and increase the signal-to-noise ratio, we subsampled the ceilometer observations at a sampling rate of 5 minutes by averaging multiple profiles, and vertically averaging on regularly spaced 50 $\mathrm{m}$ bins. We expect that in most cases cloud was almost constant on this time and vertical scale, and therefore we were not averaging together different cloud types or clear and cloudy profiles. At the same time as subsampling, we performed noise removal by estimating the noise distribution (mean and standard deviation) based on returns in the uppermost range bins (i.e. 300 samples over 5 min when sampling rate was 2 $\mathrm{s}$), and subtracting the range-scaled noise mean from the backscatter. We then used the range-scaled noise standard deviation ($\sigma$) for cloud detection: a bin was considered cloudy if the calibrated backscatter minus $3\sigma$ exceeded $20\times10^{-6}$ $\mathrm{m}^{-1}\mathrm{sr}^{-1}$. This threshold was chosen subjectively so that cloud was visually well separated from other features, such as boundary-layer aerosol and noise on backscatter profile plots. The same threshold was used on both the observations and output from the COSP surface lidar simulator and thus should cause little bias.

### 3.2.3 Model lidar data processing

We used the same sampling rate (5 min) and model levels as range bins on the surface lidar simulator output. For each vertical profile we used model data at the same location as the ship and the same time relative to the start of the year. Model data were selected using nearest-neighbour interpolation. The model resolution is lower than the distance travelled by the ship in 5 minutes, therefore the same model data were used multiple times to generate consecutive profiles. However, we also used the SCOPS (Webb et al., 2001) subcolumn generator included in COSP to generate 10 random samples of cloud for each profile based on cloud fraction and the maximum/random cloud overlap assumption (Bodas-Salcedo, 2010). The lidar simulator processes each sample individually. The resulting cloud occurrence is calculated as the average of the 10 samples. The lidar simulator does not generate noise, and therefore we did not perform any noise removal on the simulated profiles, but we used the same threshold of $20\times10^{-6}$ $\mathrm{m}^{-1}\mathrm{sr}^{-1}$ and vertical bins of 50 m for detecting cloud (as used on the observations). For the MERRA-2 cloud occurrence analysis, we applied the lidar simulator on the 3-hourly instantaneous Assimilated Meteorological Fields (M2I3NVASM.5.12.4) product~~subsampled to a 1×1 degree global horizontal grid~~.

### 3.3 ~~SST lifting level~~

~~In our analysis we used a metric "SST lifting level" (SLL) derived from SST and boundary-layer atmospheric potential temperature (measured by radiosondes or simulated by a model). We define SLL as the level to which an air parcel with the same temperature as SST, rising from the sea surface, would rise adiabatically by buoyancy. That is, it is the level closest to the surface at which potential temperature is equal to SST, provided the air parcel is permitted to rise to this level by buoyancy (otherwise the air parcel does not rise and SLL is 0 m). This metric is applicable in sea ice-free conditions in~~

## 4 Spatiotemporal subsets investigated

Because our observational dataset does not span the entire geographical area of the SO and all months of the year, and the atmospheric conditions in the SO are geographically variable, we subset the datasets into a number of geographical regions by latitude and time periods by season. The geographical regions investigated are 50–75°S by 5 degrees of latitude, and the temporal periods investigated are austral summer of December, January, February (DJF) and autumn months of March, April, May (MAM).

We do not use data from 70–75°S and 50–55°S in all parts of the analysis. The data from 70–75°S are likely affected by circulation induced by land near the Ross Sea (Coggins et al., 2014), and therefore may not be representative of the SO in general. This decision builds on the analysis detailed in Jolly et al. (2018) which shows a significant gradient in cloud properties between the Ross Ice Shelf and the Ross Sea and strong influences associated with synoptic conditions. The data from 50–55°S were relatively sparse (the ships spent relatively little time passing through this latitudes). Radiosonde observations were only available south of 60°S.

There is likely temporal variability present within the DJF and MAM time periods, but we decided to limit the number of temporal subsets to maintain a reasonable quantity of observations in each subset. The magnitude of the SO TOA outgoing SW radiation bias is primarily modulated by incoming solar radiation, which is the ~~SO, when cold Antarctic air is warmed by the open sea surface and is lifted by buoyancy until it reaches a limit imposed by the atmospheric stability of the atmosphere. Together with the lifting condensation level (LCL)we found SLL to be a useful metric for evaluation of boundary-layer CBH. Apart from SST and LCL, we also evaluate cloud with respect to lower tropospheric stability (LTS) (Klein and Hartmann, 1993)~~ highest in DJF. The voyages do not uniformly cover all geographical regions or time periods, with the largest number of observations in the Ross Sea sector south of New Zealand (TAN1802, TAN1502, HMNZSW16, NBP1704), followed by the Indian Ocean sector south of Western Australia (AA15). Temporally, the voyage observations mostly cover summer to autumn months of the year.

## 5 Results

### 5.1 Shortwave radiation balance

Figure 3 shows ~~reflected TOA~~ TOA outgoing SW radiation in CERES, GA7~~.0, GA7~~.1 and MERRA-2. We present this panel plot in order to evaluate how well GA7~~.0, GA7~~.1N and MERRA-2 are performing in terms of ~~the~~ SW radiation bias in the SO relative to CERES. This analysis assumes that CERES is a good observational reference, although it is affected by ~~biases~~ errors of lower order of magnitude ~~(Loeb et al., 2018)~~(2.5 $Wm^{-2}$ "regional monthly uncertainty" (Loeb et al., 2018, sec. 4a.) ). The plots reveal ~~a predominantly~~ relatively zonally symmetric pattern of ~~reflectivity in the SO on the yearly~~ negative and positive bias on the annual (Figure 3~~a–d) and monthly scales~~ b, c) and seasonal (Figure 3~~e–h) , with more variable patterns in the tropics related to regions of upwelling and downwelling. We chose 19 January 2007~~ e, f, h, i) time scales. GA7.1N shows predominantly negiative bias, while MERRA-2 shows predominantly positive bias. The annual average is dominanted by the bias in DJF due to the relatively strong incoming solar radiation in DJF. The bias displays very similar geographical pattern on the annual scale, DJF and MAM. The bias is much lower in MAM compared to DJF due to lower incoming solar radiation.

We chose 1 January 2018 as a representative day in ~~January~~ DJF to show the daily ~~pattern~~[2]scale. On the daily scale (Figure 3~~i–l; 19 January 2007)~~j, k, l), the patterns are closely linked to synoptic features~~, with close inspection displaying particularly large differences in the TOA SW radiation near the Antarctic Peninsula~~. The region on the eastern side of the Antarctic Peninsula shows ~~a greater reflectivity in CERES (Figure 3e), but not in any of the models(Figure 3f–h). The zonal symmetry of the annual and monthly means (Figure 3a–h) suggests~~ the greatest negative bias in the models. The relatively zonally symmetric annual and seasonal means suggest that there is not a significant need for subsetting by longitude, and that latitude averages can be very useful in identifying the key features of the SW radiation biases. The daily synoptic features are generally well-correlated between CERES and the models~~(Figure 3i–l)~~, which is expected in nudged model runs and reanalyses. ~~The highest reflectivity is generally associated with frontal regions and extratropical/polar cyclones, although cloud-associated reflectivity~~

[2]~~A choice made for convenience during the analysis due to overlap with existing Transpose-AMIP HadGEM2-A hindcasts.~~

~~is present throughout the SO. Examination shows that~~ MERRA-2 has greater ~~upwelling TOA SW radiation~~ TOA outgoing SW radiation than GA7.1N on all three time ~~scales~~ periods presented here. Considering that cloud is the dominant factor affecting SW radiation in the SO, this can only be associated with either cloud cover which is too high, or ~~clouds which are too bright, and our analysis of cloud occurrence (Section 5.2) supports the latter~~cloud albedo which is too high. GA7~~.0 and GA7~~.1~~are less reflective than CERES between~~ N reflects too little SW radiation south of 60°S and ~~70~~too much north of 60°S (Figure 3~~m, n), with some individual cloud systems being too bright (Figure 3j, k~~b, e, h). MERRA-2 ~~is also much more reflective on the January monthly mean at all latitudes between 55°S and 70~~reflects too much SW radiation in most of the SO except for coastal regions of Antarctica (approx. 65–70°S~~(Figure 3o) . The opposing sign of the SO~~) and the eastern side of the Antarctic Peninsula. The opposite sign of SW radiation bias in GA7~~.0 and GA7~~.1N compared to MERRA-2 suggests that contrasting the two models could be useful ~~in~~ for uncovering the cause of the ~~SO SW radiation biases~~bias.

Figure 4 shows line plots of zonal mean reflected SW radiation and bias relative to CERES by month in multiple latitude bands between ~~55~~ 50 and 70°S, with the southernmost band 65–70°S limited to 180–80°W to avoid covering land areas in Antarctica. The annual cycle follows the expected ~~cyclical~~ seasonal pattern modulated by varying incoming solar radiation with maxima of reflected radiation in December and maxima of bias in December and January. The Antarctic sea ice extent, at its minimum in February and peaking in September, is also likely a secondary modulating factor of the TOA outgoing SW radiation at higher latitudes. The models represent the ~~cyclical~~ seasonal pattern well, but differ substantially during the periods of peak incoming solar radiation. ~~Inspection of the~~ The GA7~~.0~~ .1N model (Figure 4b, ~~f, j, n) shows a largely negative bias in the SO, increasing with latitude and reaching -38 between 65 and 70°S in January. Between 50 and 55°S (Figure 4b) however, the~~ e, h, k) exhibits bias ranging from -21 to +11 $Wm^{-2}$. The bias is positive ~~at its peak, reaching 5 and overall is close to zero throughout the year. This is important because previous studies of SO cloud often do not discern different latitudes, partly due to the limited availability of surface and in situ cloud observations in the SO. These panels also justify why it is important to do spatial subsetting by latitude when analysing the SO SW bias in models. The GA7.1 model (Figure 4c, g, k, o) exhibits lower bias than GA7.0 at all latitude bands except for 50–55~~north of 55°S ~~(Figure 4c), where the positive bias is greater, peaking at 10 and is fairly constant throughout the year. The likely explanation for this feature is that GA7.1 is reflecting more SW radiation in the SO, which reduces the bias where it is negative , but increases the bias where it is already positive. Overall, GA7.1 improved the peak bias from -38 to -20 at 65–70~~and negative south of this latitude, with the greatest absolute bias between 60 and 65°S~~in January~~. MERRA-2 displays a clearly different bias ~~than GA7.0 and~~ from GA7.1~~(d, h, l, p~~N, ranging from -12 to 39 $Wm^{-2}$ (Figure 4c, f, i, l). The ~~SW bias is consistently positive , i.e. too much SW radiation is reflected at all latitudes between 50 and 70~~peak SW bias in MERRA-2 is positive for latitudes north of 65°S and negative south of this this latitude. The absolute ~~value of the bias is also lower~~ bias in MERRA-2 is much larger than in GA7~~.0 and GA7~~.1~~between~~ N north of 60~~and 70~~°S ~~(15 in MERRA-2 vs. -20 in~~ and similar to GA7~~.0)~~.1N south of this latitude. Therefore, the MERRA-2 results are valuable for contrasting with GA7~~.0 and GA7~~.1. ~~In the low latitude SO (50–60°S), however, MERRA-2 performs more poorly than GA7.0 and GA7.1, showing bias of about 30~~ The strong latitudinal variation of the TOA outgoing SW radiation bias is important to take into consideration. Previous studies of SO clouds often did not discern different latitudes.

~~To summarise, we find that GA7.1 is an improvement over GA7.0 with respect to the~~ Figure 10 shows scatter plot of the TOA outgoing SW radiation bias in ~~the SO,~~ GA7.1N and MERRA-2 ~~is superior to~~ as a function of near-surface air temperature and relative humidity between 55 and 70°S in January 2018. The bias is predominantly negative in GA7.1~~at latitudes poleward of 60~~N and positive MERRA-2. There is a strong cluster of negative bias at temperature around 0 °~~S. However, due to compensating biases commonly present in the models it might not be a superior representation of reality. Schuddeboom et al. (2019) evaluate compensating errors in the~~ C in GA7.1~~model~~N and -2 °C in MERRA-2, and a cluster of positive bias at higher temperatures. This is consistent with the latitudinal dependence of bias in both models shown above.

### 5.2 Cloud occurrence in model and observations

To understand how clouds contribute to the SW bias, we examine cloud cover and cloud occurrence as a function of height in the models and observations. Figure 5 shows cloud occurrence profiles derived from ceilometer observations on different voyages and GA7~~.0U~~.1N and MERRA-2 model output derived via the COSP surface lidar simulator, ~~as a function of~~ in subsets by latitude and season. ~~The seasons cover the austral summer and late summer/autumn months. The comparison with GA7.0 is completed statistically, i.e. 10 individual years of free-running GA7.0 simulation data between 1980 and 1989 are compared with observations and reanalysis between 2015 and 2018.~~ Most notably, the observed cloud cover is consistently very high in the observations (80–100%) for all periods and latitude bands examined and greater than 90% in most of ~~these~~ the subsets. This finding differs substantially from the modelled cloud cover ~~(~~derived via the surface lidar simulator~~)~~, which ranges between ~~17 and 92~~69 and 100% in GA7~~.0~~.1N, and is about ~~25~~4-9% lower than observations across the subsets. ~~Cloud~~ The cloud cover in MERRA-2 is also ~~generally lower than observations, but higher than~~ lower than observed and much lower than in GA7~~.0, spanning 49–95%.~~.1N, spanning 51–95%. Only in 4 subsets is the cloud cover greater in GA7.1N than observed, and only in 1 subset is the cloud cover greater in MERRA-2 than observed (out of 21 subsets). Our analysis therefore shows that cloud cover is underestimated in both GA7~~.0~~.1N and MERRA-2 in the evaluated geographical regions and seasons~~evaluated here. This shows a similar bias to previous analysis by Schuddeboom et al. (2018) who compared COSP derived cloud cover from the GA7.0 model with MODIS satellite observations and found much higher cloud cover in the observations. Due to the high zonal symmetry of the SW radiation in the SO, shown in Figure 3, and the magnitude of the cloud cover bias in the model, these results are likely representative of the whole SO.~~ .

Examination of the vertical distributions in Figure 5 shows that ~~the observations indicate~~ observations have a strong predominance of cloud below 2 km ~~, peaking below 1~~ and peaking below 500 km in most subsets, including a substantial amount of surface-level fog in some subsets. In contrast, GA7~~.0 simulates cloud~~ .1N and MERRA-2 simulate clouds at a higher altitude, peaking at about ~~1 . Further analysis shows that the MERRA-2 vertical distribution appears more consistent with the observations than GA7.0, often peaking below 1 , but overall having lower cloud cover at the peak altitude than the observations~~500 m and generally the peak is higher than in observed clouds. Especially, clouds below 500 m and fog appear to be lacking in the models.

The subsets in Figure 5 are derived from uneven length of ship observations (1.0–28.9 days) due to the limited availability of data. The longer subsets (Figure 5a4, b4, c2, c4, f1) appear marginally more consistent between the models and observations in terms of the cloud ocurrence profile, but the cloud cover is still markedly underestimated.

Figure 6 shows the model subsets of Figure 5 as points by their cloud cover bias relative to observations. It can be seen that GA7.~~0U~~ .1N underestimates cloud cover by about ~~25~~4% and MERRA-2 by ~~18~~16% when non-weighted averages are considered, and ~~both models underestimate this amount by about 18% when weighted averages are considered. This difference is caused mostly by an outlier: AA15 DJF 65–70°S (28.2 days) , which exhibits 80% cloud cover in observations and 90% in~~ by 9% (GA7.~~0U. Neither the averages nor weighted averages , however, should be accepted uncritically, as they group together different latitudes and statistically correlated weather situations. The same weather can persist for several days, and therefore measurements taken during a continous period of time are statistically correlated, whereas measurements on different voyages represent statistically independent samples.~~1N) and 18% (MERRA-2) when weighted averages are considered.

Due to the nature of the lidar measurements, ~~mid to high level cloud~~ middle to high clouds may be obscured by low ~~level cloud, because~~ clouds, as the laser signal is quickly attenuated by thick cloud. Therefore, the lack of ~~cloud~~ clouds above 2 km in the plots does not imply that ~~there is no cloud at these heights.~~ no clouds are present. The lidar simulator, however, ensures unbiased 1:1 comparison with observations by accounting for the signal attenuation.

The results demonstrate the value of surface cloud measurements in the SO relative to satellite measurements such as CloudSat and CALIPSO, which would likely provide a biased sample of these clouds because of "ground clutter" and ~~obscuring higher level cloud~~obscuration by higher-level clouds, respectively (Alexander and Protat, 2018).

### 5.3 Radiosonde observations

~~Radiosonde observations were performed on the~~ We use radiosonde measurements performed on TAN1802 and NBP1704 ~~voyages. Temperature, pressure, relative humidity and GNSS coordinates (from which wind speed and direction are derived) were retrieved to altitudes of 10–20 , terminated by a loss of radio communication or balloon burst. We use these data to evaluate boundary-layer~~ to evaluate boundary layer properties and correlate them with ~~cloud as~~ clouds observed by a ceilometer. We compare the observations with "pseudo-radiosonde" profiles extracted from model fields at the same location and time~~of the year. Figure 7 shows the relationship between CBH and the minimum of SSL and LCL ("min{SLL,LCL}" ) as a scatter plot based on a merged dataset from TAN1802 and NBP1704 voyages, and the corresponding points from GA7.0 and MERRA-2. We choose to evaluate min{SLL,LCL} as a predictor instead of either SSL or LCL individually for the following reasons. This relationship becomes quite notable when examining the individual voyage radiosonde profiles (not presented here). If SLL is higher than LCL, an air parcel~~ . The location is based on the GNSS coordinates of the ship at the time of the balloon launch (the ballon trajectory length was generally not long enough to span multiple model grid cells in the lower troposphere).

We define a new quantity "SST lifting level" (SLL) derived from SST and boundary layer atmospheric potential temperature, defined as the level to which an air parcel with the same temperature as SST, rising from the sea surface, would rise adiabatically by buoyancy. That is, it is the level closest to the surface at which potential temperature is equal to SST, provided the air parcel is permitted to rise to this level by buoyancy (otherwise the air parcel does not rise and SLL is 0 m). This quantitiy is applicable

in sea ice-free conditions in the SO, when cold Antarctic air is warmed by the ~~sea surface rises by buoyancy past LCL, at which point water vapour starts to condensate (assuming enough cloud condensation nuclei are present at 100% saturation) , forming cloud with CBH equal to LCL. If SLLis lower than LCL, the air parcel rises to SLL~~ open sea surface and is lifted by buoyancy until it reaches a limit imposed by the atmospheric stability of the atmosphere. Alongside the lifting condensation level (LCL) we found SLL to be a useful quantity for evaluation of CBH. The authors are not aware of any previous use of SLL, ~~where air lifted from the sea surface eventually accumulates, potentially forming cloud if enough moisture is transported from the sea surface~~but this definition is supported by observations (see below).

Apart from SLL and LCL, we also use the lower tropospheric stability (LTS) (Klein and Hartmann, 1993). LTS is defined as the difference between potential temperature at 700 hPa and sea level pressure (Klein and Hartmann, 1993). It has been used in multiple previous studies (Williams et al., 2006; Franklin et al., 2013; Williams et al., 2013; Naud et al., 2014).

Figure 7 shows the observed and modelled relationship between CBH and the minimum of SLL and LCL ("min{SLL,LCL}"), LTS, SLL and LCL. A large fraction of the observed points (OBS) in Figure 7a ~~lies~~ lie close to the origin (40% in the first 100 m in observations, vs. 26% and 17% in GA7.1N and MERRA-2, respectively), which suggests that near zero min{SLL,LCL} is a good indicator of fog or very low cloud~~. The~~ , a relationship not well-represented in the models. The remaining observed points show a close equivalence between min{SLL,LCL} and CBH, while the models do not ~~seem to represent this relationship~~ represent this equivalence well. The histogram in Figure 7a ~~shows that about 40~~reveals that about 42% of observed profiles have CBH within 100 m of min{SLL,LCL}, while only about ~~20% of MERRA-2 profiles and 15% of~~ 28% of GA7~~.0U/1980-89~~.1N and 21% of MERRA-2 profiles do. ~~Using SSL~~

Using SLL or LCL as a predictor for CBH individually resulted in a weaker relationship than min{SLL,LCL}~~(not presented here).~~ : 25% and 31% of OBS profiles have CBH within 100 m of SLL and LCL, respectively (Figure 7c, d). This suggests that min{SLL,LCL} is more strongly related to CBH than SLL or LCL individually. Figure 7b shows ~~the same points~~ CBH as a function of LTS~~, defined as the difference between potential temperature at 700 hPa and sea level pressure (Klein and Hartmann, 1993) , and used in previous studies (Williams et al., 2006; Franklin et al., 2013; Williams et al., 2013; Naud et al., 2014)~~. LTS does not display a good predictive ability for CBH in this dataset, with the exception of very stable profiles (LTS > 15 K), when observed CBH was below 250 m in all but one case.

Figure 8 shows the distribution of ~~SSL as derived from radiosonde observations and model fields, and scatter plots of CBH vs.~~ min{SLL,LCL} ~~as in Figure 7. The purpose of the panel plot is to evaluate the relationship between local boundary-layer thermodynamics and cloud occurrence. In the absence of a synoptic-scale forcing and geographical features, one can expect clouds in the boundary layer to be well correlated with the local thermodynamic profile in the boundary layer. Due to the very persistent cloud cover observed in the SO in summer months (close to 100%), as shown by the cloud occurrence analysis in Figure 5, we might expect that conditions in the SO are such that an almost continuous cloud formation takes place or that cloud persists even in the absence of synoptic forcing. We hypothesise that the models underestimate cloud cover in these quiescent conditions. As can be seen in the scatter plots in Figure 8, there is a strong correspondence between min{LCL, SLL} and CBH in cases where there is no sea ice. Because of the observed close link between SLL and CBH, we examined whether the models may be misrepresenting SLL. As can be seen in the SLL vertical distribution panels~~ derived from radiosonde observations and

model fields. In observations, the quantity almost consistently peaks near the ground and reaches up to 1.5 km in ice-free cases (Figure 8a1–a5, b4). GA7.1N represents this distribution relatively well. This is not the case with MERRA-2, which is less likely to peak near the ground and the distribution is less consistent with observations. The sea-ice cases (Figure 8a–f, m–r) , there is no substantial difference between the models and radiosonde observations in non-sea ice cases, in which SLL has a plausible effect on cloud, even though b5, b6) show markedly different observed distribution of the quantity, with peak at about 300 m. GA7.0 simulates a slightly higher SLL than observed . We conclude that SLL difference is likely not a cause for the underestimated cloud cover in the models. SLL is a function of SST and the boundary layer potential temperature profile, we therefore expect both fields to be well simulated in the boundary layer of the models, and by ruling out this potential cause, the alternative explanation – that subgrid-scale model processes are an important factor in the underestimation of cloud cover – is more likely.1N and MERRA-2 represent the distribution over sea ice relatively poorely.

### 5.4 Zonal plane comparison of GA7.0 .1N and MERRA-2

In order to better understand the differences in the SW radiation bias between GA7.0 .1N and MERRA-2, we inspect zonal plane plots of cloud occurrence and thermodynamic fields of both models in January and on a specific day the models in DJF 2017/18 and 1 January 2018 (Figure 9). The GA7.0 model is a nudged run, which ensures a general correspondence between synoptic features in the model, the reality and the MERRA-2 reanalysis, i.e. both columns of Figure 9 show the same synoptic features. The figure shows monthly figure shows seasonal and daily average cloud liquid and ice mixing ratio contours (a monthly average in January 2007 and a daily average on 19 January 2007, respectively) plotted over two different backgrounds – potential temperature and relative humidity (RH). The daily average plot plots (Figure 9c, d) shows show a very pronounced difference between the cloud liquid amount between the two models, with MERRA-2 simulating a much greater amount of cloud liquid. In contrast, GA7.0 simulates clouds .1N simulates cloud with ice, which are nearly absent in MERRA-2 at the chosen contour levels. The liquid content is generally concentrated near SLL as observed in Figure 8, and therefore at the top of the surface coupled boundary layer, whereas the ice content in MERRA-2, but much less so in GA7.0 .1N, where SLL is often at 0 m. The cloud ice in GA7.1N generally has significantly greater vertical extent than the cloud liquid. These differences are also present in the monthly average on the seasonal scale (Figure 9a, b). Relatively small differences in the background potential temperature and SLL between the two models fields suggest that the cloud differences are not explained by these fields. Moreover, the fields appear fairly consistent The difference in potential temperature between the models , suggesting that the synoptic state of the atmosphere is not responsible for the cloud differences. There isis relatively small. GA7.1N, however, a pronounced difference in relative humidity shows a slightly higher potential temperature. The RH field is very different between GA7.0 .1N and MERRA-2in the mid to high troposphere, quite clearly visible on the monthly average plots , with MERRA-2 simulating higher RH by about 10%.

Pehaps most interestngly, the vertically integrated liquid and ice content (Figure 9e, f) . This bias does not seem to be present in the boundary layer and it is therefore not a likely explanation for the cloud bias. We should note that it is not obvious from this analysis whether i, j) is very different between the models. Both models simulate almost the same liquid + ice total, but the phase composition of cloud in GA7.0 or .1N is majority liquid, while in MERRA-2 are closer to reality, even though larger

amount of simulated cloud is preferable with respect to reducing the model SW radiation biases. Model cloud liquid and ice which is more spread out horizontally, while holding the same totalamount of water, would increase the cloud cover a reduce the cloud opacity, and lead to a better correspondence with our observationsit is almost entirely ice.

## 6 Discussion

The TOA outgoing SW radiation assessment showed that shows that the models exhibit monthly average biases of up to -38 39 $\mathrm{Wm}^{-2}$ (GA7.0, 65–70MERRA-2, 50–55°S in JanuaryDecember), and that these biases have a significant latitudinal dependency, leading to opposing signs of the with the opposite sign of bias between different latitude bands. This conclusion is also supported by In GA7.1N the bias is predominantly negative, while in MERRA-2 the bias is predominanly positive. Similar pattern of bias is present both models. The bias is positive north of 55°S (65°S) in GA7.1N (MERRA-2) and negative south of this latitude. This finding is consistent with Schuddeboom et al. (2019), who observed opposing sign of the opposite sign of SW cloud radiative effect (CRE) south and north of 55°S in GA7.1. We found GA7.0/GA7.1 and MERRA-2 to be biased in the opposite direction, with GA reflecting too little SW radiation in the high latitude SO, while MERRA-2 reflects too much SW radiation in the SO

Very simlar geographical pattern of bias is present in DJF and MAM, suggesting that similar cloud biases are present in both seasons. This is also supported by Figure 5, which does not display a significant difference in observed cloud occurrence and bias in the models between DJF and MAM. Consistent with the maximum of incoming solar radiation, January was December and January were found to be the month months with the greatest absolute bias in models. For this reason, improving model cloud biases in austral summer months is the models. Therefore, fixing the representation of clouds in the SO in these months is relatively more important than in other monthswith respect to the SW radiation bias. Cloud representation differences are expected to be the strongest factor in modulating the TOA SW radiation, and this can happen either via cloud cover or cloud opacity effects, or both simultaneously. Therefore, we conclude that .

Figure 10 suggests that the bias correlates not only with latitude, but also with near-surface air temperature. The negative bias is strongly clustered around 0 °C in GA7.0/7.1 simulates too little cloud cover, but we cannot conclude whether it is too opaque or too transparent, and .1N, and -2 °C in MERRA-2simulates too little cloud cover and too opaque cloud. The cloud occurrence analysis , and positive bias is predominantly correlated with higher temperature.

The ship-based lidar cloud occurrence revealed close to 100% cloud cover as measured by a ceilometer on a number of voyages. This seems to be the case across different latitudes in the austral summer and autumn, even though the results are limited by region (the Ross Sea and Indian Ocean sectors)and the relatively brief passage of the ships through some of these regions. The 2016–2018 voyages may have been affected by the unusually low sea ice extent (discussed below) , which can have a significant effect on cloud (Frey et al., 2018; Taylor et al., 2015). We found in multiple subsets. Subsetting allowed us to identify whether the cloud cover is substantially different by latitude and season, and also sample independent weather situations (it is expected that cloud occurrence profiles are highly correlated over several days due to persistance of synoptic situations). The subsets show a relatively consistent cloud occurrence profile peaking below 500 m, and almost zero above

2 km (possibly also due to obscuration of lidar signal by lower clouds). The models generally do not reproduce this profile well. Apart from underestimating the total cloud cover, the peak of cloud occurrence in the models is higher than observed. Improving the cloud profile representation in the models is likely key for improving the SW radiation bias.

The effect of clouds on SW radiation is the product of cloud cover (the fraction of the sky containing clouds) and cloud albedo (the fraction of SW radiation reflected by the clouds). With our ship-based lidar observations we measured cloud cover (total, and cloud cover as a function of height), while we did not measure cloud albedo. The cloud cover was almost consistently underestimated in both GA7.~~0 underestimates totalcloud cover by a relatively large amount (nearly 25%), with~~ .1N and MERRA-2 ~~underestimating total cloud cover by a lesser extent. Combined with the overestimation of the TOA SW radiation, we concluded that~~ across all latitudes. At the same time, the satellite observations show that MERRA-2 reflects too much all-sky SW radiation. Therefore, the cloud albedo in MERRA-2 must be ~~overestimating cloud opacity to an extent which overcompensates for~~ too high in order to cause too much all-sky SW radiation reflection despite the lack of cloud cover. ~~We cannot make the same conclusion about~~ This effect is visible on the daily scale in Figure 3j–l, where the individual clouds in MERRA-2 appear significantly brighter than on satellite observations.

Remarkably, the observed and simulated cloud ocurrence profiles do not appear to be significantly different between the DJF and MAM seasons or different latitude bands between 55 and 70°S (Figure 5). This is in contrast with the SW radiation bias analysis, which showed a strong latitudinal gradient of the TOA outgoing SW radiation bias in the models (Figure 3, 4). Based on the the presented results a plausible explanation for the SW radiation bias could be overestimation of cloud albedo north of about 55°S (65°S) in GA7.~~0, but it seems plausible that strongly underestimated cloud cover alone can explain the TOA~~ .1N (MERRA-2) causing positive TOA outgoing SW radiation bias north of this latitude and underestimation of cloud cover over the whole SO causing negative TOA outgoing SW radiation bias ~~relative to CERES~~south of this latitude.

~~During the TAN1802 voyage~~ In the ship observations we found a notable correspondence between CBH, SLL and LCL. Boundary layer thermodynamics, determining the lifting levels, is a plausible driver of cloud formation in the absence of other forcing. We examined SLL in models and radiosonde observations, and found differences which are likely too small to explain the cloud occurrence differences between the models and ceilometer observations. Bodas-Salcedo et al. (2012), in their analysis of an earlier version of the GA model (GA3.0) using cyclone composites also noted that biases in thermodynamics are not likely to explain the SW radiation bias, but may still play a significant role. The presence of positive TOA outgoing SW radiation bias in the SO between 50 and 55°S in GA7.1, which ~~has an opposing sign to the bias in the high latitude~~~~SO~~contrasts with the negative bias south of the latitude, is important because it places a limit on the applicability of other studies which used SO observational data from regions north of 55°S (Lang et al., 2018).

In Section 5.3 we show that min{SLL,LCL} has a stronger equivalence to CBH than SLL, LCL individually or LTS. This relationship becomes quite notable when examining the individual voyage radiosonde profiles (not presented here). We hypothesise that the theoretical reason for this relationship is the following. When SLL is higher than LCL, an air parcel warmed by the sea surface to temperature close to SST rises by buoyancy past LCL to a level with the equivalent potential temperature. The water vapour starts to condensate at LCL (assuming enough cloud condensation nuclei are present at 100% saturation), forming cloud with CBH equal to LCL. If SLL is lower than LCL, the air parcel rises to the level of equivalent potential

temperature, where air lifted from the sea surface eventually accumulates, potentially forming cloud if enough moisture is transported from the sea surface. The models do not represent the observed relationship well, and improving this relationship may be one way of improving the cloud simulation.

Considering the strong observed relationship between min{SLL,LCL} and CBH (CBH tends to occur at the same level as min{SLL,LCL}), we evaluated the distribution of min{SLL,LCL} in the models in comparison with radiosonde observations (Figure 8). We found that GA7.1N represents this distribution relatively well in sea-ice-free cases, while MERRA-2 represents this distribution relatively poorly. MERRA-2, however, tends to underestimate the distribution of min{SLL,LCL} near the ground. This may be the reason for the underestimation of very low cloud and fog in this model identified in the comparison with lidar observations. Therefore, improving the distribution of the quantity in MERRA-2 may lead to improvement of low cloud simulation.

It is interesting to contrast our results with previous studies which used cyclone compositing for the TOA SW radiation bias evaluation in GCMs. We cannot make substantial conclusions from our results on how much of the model bias is attributable to cyclones. It appears, however, that the cloud cover and cloud liquid and ice mixing ratio bias in GA7~~.0~~ .1N is systematic rather than isolated to cyclonic activities due to its relative consistency across spatiotemporal subsets in the high latitude SO. This does not rule out even greater biases related to cyclonic sectors. Specifically, Bodas-Salcedo et al. (2014) evaluated a large set of models, including HadGEM2-A, a predecessor model to HadGEM3, likely affected by similar biases, and found that about 80% of grid cells south of 55°S could be classified as affected by a cyclone, and that these grid cells were responsible for the majority of the total SW radiation bias. Moreover, their cyclone compositing showed that the bias in HadGEM2-A was largely negative in the cold quadrants, and near zero in the warm quadrants. Their results also indicate a strong contrast in SW bias south and north of 55°S, similar to the result we found in GA7~~.0 and GA7~~.1N. We think these results can be reconciled with our study by assuming that the model has a particular difficulty in representing cloud in situations when near-surface air temperature is lower than the SST. In these regions the heat flux is from the ocean to the atmosphere is positive, which in the austral summer predominantly occur south of 55°S and in the cold sectors of cyclones. The cloud representation when near-surface air temperature is greater than SST is relatively more accurate, this case occurring predominantly north of 55°S and in the warm sector of cyclones. ~~To evaluate the viability of this explanation we plotted the daily average~~ As shown in Figure 10, the negative TOA outgoing SW radiation bias in the ~~GA7.0N/2007 grid cells as a function of near-surface air temperature and near-surface relative humidity between 40°S and 70°S on 19 January 2007 (Figure 10). The grid cells with strong negative bias are visibly clustered between -2 and +2 , whereas at higher temperatures the bias tends to be more equally distributed between positive and negative values~~models is clustered at zero and sub-zero temperatures. This suggests a possible explanation that subzero air mass advecting from Antarctica or from sea ice covered areas over warm water could be inducing ~~convection and steam fog or low cloud~~ low cloud and fog, and this process is not well represented ~~by the model. Therefore the cloud biases in HadGEM2-A and HadGEM3 may not be linked to cyclonic activity as such, but secondarily through their impact on near-surface air temperature and its difference from SST~~in the models.

~~Supercooled liquid was not a focus of this study, but we can note a number of things.~~ Previous studies have documented that supercooled liquid is often present in the SO cloud in summer months. We cannot substantially add to these findings with

our observations, although preliminary analysis of a polarising lidar ~~(~~Sigma Space MiniMPL ~~)~~ profiles from the TAN1802 voyage suggests supercooled liquid was commonly present in the ubiquitous stratocumulus cloud. The side-by-side comparison of cloud liquid and ice mixing ratios on the zonal plane (Figure 9) suggests that models can differ significantly in their representation of cloud phase, with GA7~~.0 having~~ .1N simulating markedly less supercooled liquid than MERRA-2~~(in January). Notwithstanding the cloud phase, the major problem of both models appears to be the lack of cloud cover compared to observations~~. This is the most likely the explanation for the overestimation of TOA outgoing SW radiation in MERRA-2, despite the underestimated cloud cover in this model. If cloud cover is increased in ~~the model~~MERRA-2 to better match with the lidar observations, the cloud ~~opacity may also need~~ albedo would have to be lowered to obtain a ~~good~~ reasonable match of TOA outgoing SW radiation with CERES~~observations. We know that MERRA-2 overestimates cloud opacity, and GA7 may also be overestimating cloud opacity in order to partially compensate for the lack of cloud cover.~~ .

~~In our results there is some indication that sea ice has an impact on the cloud base height (Figure 8x), but it is not easily separated from the possible effect of the geographical location (high latitude Ross Sea region) and the time of the year (MAM)~~The 2016–2018 voyages may have been affected by the unusually low sea ice extent (discussed below), which can have a significant effect on cloud (Frey et al., 2018; Taylor et al., 2015). The modulating effect of sea ice on cloud in the SO has previously been shown by Listowski et al. (2018) and there is an apparent difference in cloud between the Ross Sea and Ross Ice Shelf as shown by Jolly et al. (2018), with cloud over the ice shelf having smaller cloud cover, a greater amount of altostratus cloud and a smaller amount of deep convective cloud. The sea ice and ice shelves block transport of heat and moisture to the atmosphere. Their low thermal conductivity and high albedo mean the surface can cool to very low temperature and thus have an effect on the radiation balance of the atmosphere. We did not focus on sea ice conditions, since one can expect the effect of cloud biases on the SW radiation bias over sea ice to be small – the ice surface is already highly reflective in the SW, and the presence of cloud has little impact on the grid cell SW reflectivity (the SW albedo of cloud is similar to sea ice, depending on the sea ice concentration).

The Antarctic sea ice extent has undergone a rapid decrease starting in the spring of 2016 after about a decade of slightly increasing extent (Turner et al., 2017; Stuecker et al., 2017; Doddridge and Marshall, 2017; Kusahara et al., 2018; Schlosser et al., 2018; Ludescher et al., 2018). The sea ice extent due to this decrease was found to be the lowest on observational record since 1979, and the Ross Sea was particularly affected by this anomaly. The unusually low sea ice extent likely affected atmospheric observations made on the voyages presented in this study, e.g. the TAN1802 voyage in February and March 2018 to the Ross Sea experienced no sea ice during the entire voyage. Because sea ice is an important factor influencing the atmospheric boundary-layer stability and radiation balance, a significant secondary effect on cloud cover, cloud phase and opacity is expected. Sea ice is, however, not expected to be responsible for the SO SW radiation bias in models, because the bias is present even when sea ice concentration is prescribed from satellite observations. ~~In our analysis, this may have an effect on comparison of cloud occurrencein the free-running~~ Given that few of the ship-based observations were collected before 2016, we cannot reliably estimate how the anomalous sea ice extent affected our results.

In our results we found that even when model atmospheric dynamics is prescribed based on past observations, the TOA outgoing SW radiation bias is large and cloud occurrence, especially of low cloud and fog, is underestimated. CBH is found

to be strongly linked to the boundary layer thermodynamics, and this link does not seem to be well represented in GA7.0U relative to observations taken between 2016 and 2018. The representation of cloud in the nudged run and .1N and MERRA-2, however, should be comparable with observations without being affected by the sea ice anomaly due to having prescribed sea icebased on the satellite record. We therefore expect that cloud and boundary layer parametrisations (as part of subgrid scale processes in the models) are responsible for this bias. We have identified parts of the GA7.1N model most likely responsible: the large-scale cloud scheme, the PC2 scheme (Wilson et al., 2008a, b) and the boundary layer scheme. A future study should focus on these schemes to identify the parts responsible for the bias. In particular, the model should improve simulation of very low cloud and fog and achieve a closer match between the lifting levels and CBH (Figure 7a).

In Table 3 we present a simple calculation how the GA7.1N peak TOA outgoing SW radiation bias would change if the cloud cover were increased by 5% (as suggested by Figure 6), assuming the cloud albedo does not change. This correction would explain 51–111% of the bias depending on the latitude. The remaining part of the bias must be attributed to cloud albedo. One way this could be improved is by increasing the supercooled liquid fraction, or by increasing the total cloud water (liquid + ice) path. Therefore, our results suggest that in GA7.1N underestimation of cloud cover is responsible for the majority of the negative TOA outgoing SW radiation bias, relative to underestimation of cloud albedo.

## 7 Conclusions

We analysed 4 years of observational SO ship data, and contrasted them with a decade of free-running GA7.0 simulation, one year of nudged and free-running GA7.0 and nudged run of the GA7.1 simulation, and the GCM, and MERRA-2 reanalysis. We used satellite observations of the Earth radiation budget to assess the TOA outgoing SW radiation bias in the SO in the three models. We examined the total cloud cover and vertical distribution of cloud as measured by ceilometers and simulated by a ceilometer simulator based on the model data. We also compared SO radiosonde observations from two voyages with virtual radiosonde pseudo-radiosonde profiles from the models in order to assess boundary-layer boundary layer stability and the correlation between cloud base and atmospheric lifting levels. We also compared model fields of cloud liquid and ice content, potential temperature and relative humidity in a zonal plane analysis across the SO in order to contrast cloud and thermodynamics simulated by GA7.0 .1N and MERRA-2.

Despite improvements, the The SO SW radiation bias remains significant in the is significant in GA7.1atmospheric modelN and MERRA-2, and tends to be positive in the northern parts of the SO and negative in the southern parts of the SO in both models. MERRA-2 shows greater absolute bias than GA7.1N. SO ship-based lidar and radiosonde observations are a valuable tool for model cloud evaluation, considering the amount of low cloud in this region which is likely poorly sampled by satellite instruments due to possible obscuration by higher overlapping cloud. The main findings of this study are that multi-year ship-based observations:

- corroborate satellite-based evidence of underestimated cloud cover, with both GA7.0 .1N and MERRA-2 underestimating cloud cover by 18–25% on average by about 4–9% (GA7.1N) and 18% (MERRA-2),

- show that low cloud below 2 km is almost continuous in the SO in summer months in sea ice-free conditions, and not well represented in the models,
- indicate that ~~boundary-layer~~ boundary layer thermodynamics is a strong driver of cloud in the SO, ~~but~~ and this relationship is not well represented in the models,
- suggest that subgrid-scale processes in situations when near-surface atmospheric temperature is lower or close to SST are responsible for the cloud misrepresentation.

Future studies of SO cloud representation in the GA model could focus on specific details of the model subgrid-scale cloud processes (such as the large scale cloud, boundary layer and convection schemes), and how their tuning impacts cloud occurrence distributions compared to the ship observations. The stark difference between GA7~~.0~~ .1N and MERRA-2 cloud liquid and ice content also remains to be explained, and could provide valuable insight for improving the SO SW radiation bias in the model and the reanalysis.

*Code and data availability.* The original COSP version 1 simulator is open source and available publicly at https://github.com/CFMIP/COSPv1. The modified COSP version 1 simulator including the ground-based lidar simulator used in this study is open source and available at https://alcf-lidar.github.io. The cl2nc software for converting Vaisala CL51 data to NetCDF is available at https://github.com/peterkuma/cl2nc. The CERES EBAF and SYN1deg products are available publicly from the CERES website: https://ceres.larc.nasa.gov/. The Near-Real-Time DMPS SSMIS Daily Polar Gridded Sea Ice Concentrations product is available publicly from the NSIDC website: https://nsidc.org/data/nsidc-0081. The Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST) is available publicly from the Met Office website: https://www.metoffice.gov.uk/hadobs/hadisst/. The MERRA-2 data are available publicly from the MERRA-2 website: https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/. The ship-based observations dataset as well as all processing code is available on request from the authors.

*Author contributions.* Peter Kuma participated on methodology development, voyage observations, data analysis, writing and reviewing of the manuscript. Adrian McDonald participated on conceptualisation, funding acquisition, methodology development, voyage observations, data analysis, writing and reviewing of the manuscript. Olaf Morgenstern participated on model development, methodology development, data analysis, writing and reviewing of the manuscript. Simon Alexander, John Cassano, Jamie Halla, Sean Hartery, Sally Garrett, Mike Harvey, Simon Parsons, and Graeme Plank participated on voyage observations and reviewing of the manuscript. Vidya Varma and Jonny Williams participated on model development and reviewing of the manuscript.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* We would like to thank everyone who participated on obtaining the Southern Ocean voyage observations, especially Kelly Schick and Peter Guest for performing ceilometer and radiosonde measurements on RV *Nathaniel B. Palmer*; the Royal New Zealand Navy for ceilometer and radar measurements on HMNZS *Wellington*; Alex Schuddeboom for deployment of instruments on RV *Nathaniel B. Palmer*, the crew of the TAN1502, *Aurora Australis* V1–V3 2015/16, HMNZS *Wellington*, NBP1704 and TAN1802 voyages. Logistical and technical support for the ceilometer observations made aboard *Aurora Australis* during the summer of 2015/16 were provided as part of the Australian Antarctic Science project 4292. We acknowledge the Met Office for use of the MetUM, and for providing the HadGEM3 model. We acknowledge NASA-GMAO and ECMWF for the MERRA-2 and ERA-Interim reanalyses, respectively. In this analysis we used publicly available satellite datasets provided by NASA and NSIDC. The CERES data were obtained from the NASA Langley Research Center CERES ordering tool (https://ceres.larc.nasa.gov). We wish to acknowledge the contribution of the NeSI high-performance computing facilities to the results of this research. New Zealand's national facilities are provided by the NZ eScience Infrastructure and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme (https://www.nesi.org.nz). We would like to acknowledge the financial support that made this work possible provided by the Deep South National Science Challenge via the "Clouds and Aerosols" project. We acknowledge the software tools Python, R (R Core Team, 2018), numpy (Oliphant, 2006), scipy (Jones et al., 2001–), matplotlib (Hunter, 2007), Climate Data Operators (CDO) (Schulzweida, 2018) and parallel (Tange et al., 2011), which we used in our data analysis.

## References


Alexander, S. and Protat, A.: Cloud properties observed from the surface and by satellite at the northern edge of the Southern Ocean, Journal of Geophysical Research: Atmospheres, 123, 443–456, 2018.

Bastin, S., Chiriaco, M., and Drobinski, P.: Control of radiation and evaporation on temperature variability in a WRF regional climate simulation: comparison with colocated long term ground based observations near Paris, Climate dynamics, 51, 985–1003, 2018.

Bodas-Salcedo, A.: COSP user's manual: Version 1.3.1, 2010.

Bodas-Salcedo, A., Webb, M., Bony, S., Chepfer, H., Dufresne, J.-L., Klein, S., Zhang, Y., Marchand, R., Haynes, J., Pincus, R., et al.: COSP: Satellite simulation software for model assessment, Bulletin of the American Meteorological Society, 92, 1023–1043, 2011.

Bodas-Salcedo, A., Williams, K., Field, P., and Lock, A.: The surface downwelling solar radiation surplus over the Southern Ocean in the Met Office model: The role of midlatitude cyclone clouds, Journal of Climate, 25, 7467–7486, 2012.

Bodas-Salcedo, A., Williams, K. D., Ringer, M. A., Beau, I., Cole, J. N., Dufresne, J.-L., Koshiro, T., Stevens, B., Wang, Z., and Yokohata, T.: Origins of the solar radiation biases over the Southern Ocean in CFMIP2 models, Journal of Climate, 27, 41–56, 2014.

Bodas-Salcedo, A., Hill, P., Furtado, K., Williams, K., Field, P., Manners, J., Hyder, P., and Kato, S.: Large contribution of supercooled liquid clouds to the solar radiation budget of the Southern Ocean, Journal of Climate, 29, 4213–4228, 2016.

Bohren, C. F. and Huffman, D. R.: Absorption and scattering of light by small particles, John Wiley & Sons, 2008.

Bony, S., Stevens, B., Frierson, D. M., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T. G., Sherwood, S. C., Siebesma, A. P., Sobel, A. H., et al.: Clouds, circulation and climate sensitivity, Nature Geoscience, 8, 261, 2015.

Bosilovich, M., Lucchesi, R., and Suarez, M.: MERRA-2: File specification, 2015.

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann, U., et al.: Clouds and aerosols, in: Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 571–657, Cambridge University Press, 2013.

Ceppi, P., Hwang, Y.-T., Frierson, D. M., and Hartmann, D. L.: Southern Hemisphere jet latitude biases in CMIP5 models linked to shortwave cloud forcing, Geophysical Research Letters, 39, 2012.

Chepfer, H., Bony, S., Winker, D., Chiriaco, M., Dufresne, J.-L., and Sèze, G.: Use of CALIPSO lidar observations to evaluate the cloudiness simulated by a climate model, Geophysical Research Letters, 35, 2008.

Chiriaco, M., Vautard, R., Chepfer, H., Haeffelin, M., Dudhia, J., Wanherdrick, Y., Morille, Y., and Protat, A.: The ability of MM5 to simulate ice clouds: Systematic comparison between simulated and measured fluxes and lidar/radar profiles at the SIRTA atmospheric observatory, Monthly weather review, 134, 897–918, 2006.

Chiriaco, M., Dupont, J.-C., Bastin, S., Badosa, J., Lopez, J., Haeffelin, M., Chepfer, H., and Guzman, R.: ReOBS: a new approach to synthesize long-term multi-variable dataset and application to the SIRTA supersite, Earth System Science Data, 10, 919, 2018.

Chubb, T. H., Jensen, J. B., Siems, S. T., and Manton, M. J.: In situ observations of supercooled liquid clouds over the Southern Ocean during the HIAPER Pole-to-Pole Observation campaigns, Geophysical Research Letters, 40, 5280–5285, 2013.

Coggins, J. H., McDonald, A. J., and Jolly, B.: Synoptic climatology of the Ross Ice Shelf and Ross Sea region of Antarctica: k-means clustering and validation, International journal of climatology, 34, 2330–2348, 2014.

Comiso, J. C. and Nishio, F.: Trends in the sea ice cover using enhanced and compatible AMSR-E, SSM/I, and SMMR data, Journal of Geophysical Research: Oceans, 113, 2008.

Dee, D. P., Uppala, S. M., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the royal meteorological society, 137, 553–597, 2011.

Doddridge, E. W. and Marshall, J.: Modulation of the seasonal cycle of Antarctic sea ice extent related to the Southern Annular Mode, Geophysical Research Letters, 44, 9761–9768, 2017.

Doelling, D. R., Haney, C. O., Scarino, B. R., Gopalan, A., and Bhatt, R.: Improvements to the geostationary visible imager ray-matching calibration algorithm for CERES Edition 4, Journal of Atmospheric and Oceanic Technology, 33, 2679–2698, 2016.

Emeis, S.: Surface-based remote sensing of the atmospheric boundary layer, vol. 40, Springer Science & Business Media, 2010.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., et al.: Evaluation of climate models, 2013.

Franklin, C. N., Sun, Z., Bi, D., Dix, M., Yan, H., and Bodas-Salcedo, A.: Evaluation of clouds in ACCESS using the satellite simulator package COSP: Regime-sorted tropical cloud properties, Journal of Geophysical Research: Atmospheres, 118, 6663–6679, 2013.

Frey, W., Morrison, A., Kay, J., Guzman, R., and Chepfer, H.: The combined influence of observed Southern Ocean clouds and sea ice on top-of-atmosphere albedo, Journal of Geophysical Research: Atmospheres, 123, 4461–4475, 2018.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., et al.: The modern-era retrospective analysis for research and applications, version 2 (MERRA-2), Journal of Climate, 30, 5419–5454, 2017.

Goldberg, M. D., Kilcoyne, H., Cikanek, H., and Mehta, A.: Joint Polar Satellite System: The United States next generation civilian polar-orbiting environmental satellite system, Journal of Geophysical Research: Atmospheres, 118, 13–463, 2013.

Hartery, S., Kuma, P., McGregor, J., Marriner, A., Sellegri, K., Saint-Macary, A., Law, C., von Hobem, M., Kremser, S., Lennartz, S., Archer, S., DeMott, P., Hill, T., Querel, R., Brailsford, G., Geddes, A., Parsons, S., McDonald, A., and Harvey, M.: Atmospheric Measurements During the Antarctic and Southern Ocean Marine Environment and Ecosystem Study (ASOMEES), in preparation, 2019.

Haynes, J. M., Jakob, C., Rossow, W. B., Tselioudis, G., and Brown, J.: Major characteristics of Southern Ocean cloud regimes and their effects on the energy budget, Journal of Climate, 24, 5061–5080, 2011.

Hodges, K. I., Lee, R. W., and Bengtsson, L.: A comparison of extratropical cyclones in recent reanalyses ERA-Interim, NASA MERRA, NCEP CFSR, and JRA-25, Journal of Climate, 24, 4888–4906, 2011.

Hoskins, B. J. and Hodges, K. I.: A new perspective on Southern Hemisphere storm tracks, Journal of Climate, 18, 4108–4129, 2005.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., et al.: The art and science of climate model tuning, Bulletin of the American Meteorological Society, 98, 589–602, 2017.

Huang, Y., Siems, S. T., Manton, M. J., Protat, A., and Delanoë, J.: A study on the low-altitude clouds over the Southern Ocean using the DARDAR-MASK, Journal of Geophysical Research: Atmospheres, 117, 2012.

Huang, Y., Siems, S. T., Manton, M. J., Rosenfeld, D., Marchand, R., McFarquhar, G. M., and Protat, A.: What is the role of sea surface temperature in modulating cloud and precipitation properties over the Southern Ocean?, Journal of Climate, 29, 7453–7476, 2016.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing In Science & Engineering, 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

Hwang, Y.-T. and Frierson, D. M.: Link between the double-Intertropical Convergence Zone problem and cloud biases over the Southern Ocean, Proceedings of the National Academy of Sciences, 110, 4935–4940, 2013.

Hyder, P., Edwards, J. M., Allan, R. P., Hewitt, H. T., Bracegirdle, T. J., Gregory, J. M., Wood, R. A., Meijers, A. J., Mulcahy, J., Field, P., et al.: Critical Southern Ocean climate model biases traced to atmospheric model cloud errors, Nature communications, 9, 2018.

Jakob, C.: An improved strategy for the evaluation of cloud parameterizations in GCMs, Bulletin of the American Meteorological Society, 84, 1387–1402, 2003.

Jin, D., Oreopoulos, L., and Lee, D.: Regime-based evaluation of cloudiness in CMIP5 models, Climate dynamics, 48, 89–112, 2017.

Jolly, B., Kuma, P., McDonald, A., and Parsons, S.: An analysis of the cloud environment over the Ross Sea and Ross Ice Shelf using CloudSat/CALIPSO satellite observations: the importance of synoptic forcing, Atmospheric Chemistry and Physics, 18, 9723–9739, 2018.

Jones, D. A. and Simmonds, I.: A climatology of Southern Hemisphere extratropical cyclones, Climate Dynamics, 9, 131–145, 1993.

Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python, http://www.scipy.org, accessed 23 November 2018, 2001–.

Kay, J., Hillman, B., Klein, S., Zhang, Y., Medeiros, B., Pincus, R., Gettelman, A., Eaton, B., Boyle, J., Marchand, R., et al.: Exposing global cloud biases in the Community Atmosphere Model (CAM) using satellite observations and their corresponding instrument simulators, Journal of Climate, 25, 5190–5207, 2012.

Kay, J. E., Wall, C., Yettella, V., Medeiros, B., Hannay, C., Caldwell, P., and Bitz, C.: Global climate impacts of fixing the Southern Ocean shortwave radiation bias in the Community Earth System Model (CESM), Journal of Climate, 29, 4617–4636, 2016.

Klein, S. A. and Hartmann, D. L.: The seasonal cycle of low stratiform clouds, Journal of Climate, 6, 1587–1606, 1993.

Klein, S. A., Zhang, Y., Zelinka, M. D., Pincus, R., Boyle, J., and Gleckler, P. J.: Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator, Journal of Geophysical Research: Atmospheres, 118, 1329–1342, 2013.

Klekociuk, A., French, W., Alexander, S., Kuma, P., and McDonald, A.: The state of the atmosphere in the 2016 southern Kerguelen Axis campaign region, Deep-Sea Research Part II: Topical Studies in Oceanography, accepted for publication, 2018.

Kotthaus, S., O'Connor, E., Münkel, C., Charlton-Perez, C., Haeffelin, M., Gabey, A. M., and Grimmond, C. S. B.: Recommendations for processing atmospheric attenuated backscatter profiles from Vaisala CL31 ceilometers, Atmospheric Measurement Techniques, 9, 3769–3791, 2016.

Kuma, P., McDonald, A., and Morgenstern, O.: Ground-based lidar simulator framework for comparing models and observations, in preparation, 2019.

Kusahara, K., Reid, P., Williams, G. D., Massom, R., and Hasumi, H.: An ocean-sea ice model study of the unprecedented Antarctic sea ice minimum in 2016, Environmental Research Letters, 13, 084 020, 2018.

Lang, F., Huang, Y., Siems, S., and Manton, M.: Characteristics of the Marine Atmospheric Boundary Layer Over the Southern Ocean in Response to the Synoptic Forcing, Journal of Geophysical Research: Atmospheres, 123, 7799–7820, 2018.

Listowski, C., Delanoë, J., Kirchgaessner, A., Lachlan-Cope, T., and King, J.: Antarctic clouds, supercooled liquid water and mixed-phase investigated with DARDAR: geographical and seasonal variations, Atmospheric Chemistry and Physics Discussions, 2018, 1–52, https://doi.org/10.5194/acp-2018-1222, https://www.atmos-chem-phys-discuss.net/acp-2018-1222/, 2018.

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., Liang, L., Mitrescu, C., Rose, F. G., and Kato, S.: Clouds and the earth's radiant energy system (CERES) energy balanced and filled (EBAF) top-of-atmosphere (TOA) edition-4.0 data product, Journal of Climate, 31, 895–918, 2018.

Loveridge, J. and Davies, R.: Cloud Heterogeneity in the Marine Midlatitudes: Observations by MISR and Related Errors in GCMs, in preparation, 2018.

Ludescher, J., Yuan, N., and Bunde, A.: Detecting the statistical significance of the trends in the Antarctic sea ice extent: an indication for a turning point, Climate Dynamics, https://doi.org/10.1007/s00382-018-4579-3, https://doi.org/10.1007/s00382-018-4579-3, 2018.

Mace, G. G., Zhang, Q., Vaughan, M., Marchand, R., Stephens, G., Trepte, C., and Winker, D.: A description of hydrometeor layer occurrence statistics derived from the first year of merged Cloudsat and CALIPSO data, Journal of Geophysical Research: Atmospheres, 114, 2009.

Marchand, R., Mace, G. G., Ackerman, T., and Stephens, G.: Hydrometeor detection using CloudSat—An Earth-orbiting 94-GHz cloud radar, Journal of Atmospheric and Oceanic Technology, 25, 519–533, 2008.

Maslanik, J. and Stroeve, J.: Near-Real-Time DMSP SSMIS Daily Polar Gridded Sea Ice Concentrations, Version 1, accessed October 2018, doi: https://doi.org/10.5067/U8C09DWVX9LM, 1999.

Mason, S., Jakob, C., Protat, A., and Delanoë, J.: Characterizing observed midtopped cloud regimes associated with Southern Ocean shortwave radiation biases, Journal of Climate, 27, 6189–6203, 2014.

Mason, S., Fletcher, J. K., Haynes, J. M., Franklin, C., Protat, A., and Jakob, C.: A hybrid cloud regime methodology used to evaluate Southern Ocean cloud and shortwave radiation errors in ACCESS, Journal of Climate, 28, 6001–6018, 2015.

McDonald, A. and Parsons, S.: A Comparison of Cloud Classification Methodologies: Differences Between Cloud and Dynamical Regimes, Journal of Geophysical Research: Atmospheres, 123, 11–173, 2018.

McDonald, A. J., Cassano, J. J., Jolly, B., Parsons, S., and Schuddeboom, A.: An automated satellite cloud classification scheme using self-organizing maps: Alternative ISCCP weather states, Journal of Geophysical Research: Atmospheres, 121, 2016.

Morrison, A. E., Siems, S. T., and Manton, M. J.: A three-year climatology of cloud-top phase over the Southern Ocean and North Pacific, Journal of Climate, 24, 2405–2418, 2011.

Mülmenstädt, J., Sourdeval, O., Henderson, D. S., L'Ecuyer, T. S., Unglaub, C., Jungandreas, L., Böhm, C., Russell, L. M., and Quaas, J.: Using CALIOP to estimate cloud-field base height and its uncertainty: the Cloud Base Altitude Spatial Extrapolator (CBASE) algorithm and dataset, Earth System Science Data, 10, 2279–2293, https://doi.org/10.5194/essd-10-2279-2018, https://www.earth-syst-sci-data.net/10/2279/2018/, 2018.

Nam, C., Bony, S., Dufresne, J.-L., and Chepfer, H.: The 'too few, too bright'tropical low-cloud problem in CMIP5 models, Geophysical Research Letters, 39, 2012.

Naud, C. M., Booth, J. F., and Del Genio, A. D.: Evaluation of ERA-Interim and MERRA cloudiness in the Southern Ocean, Journal of Climate, 27, 2109–2124, 2014.

O'Connor, E. J., Illingworth, A. J., and Hogan, R. J.: A technique for autocalibration of cloud lidar, Journal of Atmospheric and Oceanic Technology, 21, 777–786, 2004.

Oliphant, T. E.: A guide to NumPy, vol. 1, Trelgol Publishing USA, 2006.

Parkinson, C. L.: Aqua: An Earth-observing satellite mission to examine water and other climate variables, IEEE Transactions on Geoscience and Remote Sensing, 41, 173–183, 2003.

Protat, A., Schulz, E., Rikus, L., Sun, Z., Xiao, Y., and Keywood, M.: Shipborne observations of the radiative effect of Southern Ocean clouds, Journal of Geophysical Research: Atmospheres, 122, 318–328, 2017.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2018.

Rayner, N., Parker, D., Folland, C., Horton, E., Alexander, L., and Rowell, D.: The global sea-ice and sea surface temperature (HadISST) data sets, J. Geophys. Res, 2003.

Roemmich, D. and Team, A. S.: Argo: the challenge of continuing 10 years of progress, Oceanography, 22, 46–55, 2009.

Rossow, W. B. and Schiffer, R. A.: Advances in understanding clouds from ISCCP, Bulletin of the American Meteorological Society, 80, 2261–2288, 1999.

Salomonson, V. V., Barnes, W., Xiong, J., Kempler, S., and Masuoka, E.: An overview of the Earth Observing System MODIS instrument and associated data systems performance, in: Geoscience and Remote Sensing Symposium, 2002. IGARSS'02. 2002 IEEE International, vol. 2, pp. 1174–1176, IEEE, 2002.

Sato, K., Inoue, J., Alexander, S. P., McFarquhar, G., and Yamazaki, A.: Improved Reanalysis and Prediction of Atmospheric Fields Over the Southern Ocean Using Campaign-Based Radiosonde Observations, Geophysical Research Letters, 2018.

Schlosser, E., Haumann, F. A., and Raphael, M. N.: Atmospheric influences on the anomalous 2016 Antarctic sea ice decay, Cryosphere, 12, 1103–1119, 2018.

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., Molod, A., Neale, R. B., and Saha, S.: Practice and philosophy of climate model tuning across six US modeling centers, Geoscientific Model Development, 10, 3207–3223, 2017.

Schuddeboom, A., McDonald, A. J., Morgenstern, O., Harvey, M., and Parsons, S.: Regional Regime-Based Evaluation of Present-Day General Circulation Model Cloud Simulations Using Self-Organizing Maps, Journal of Geophysical Research: Atmospheres, 123, 4259–4272, 2018.

Schuddeboom, A., Varma, V., A, M., Morgenstern, O., Harvey, M., Parsons, S., Field, P., and Furtado, K.: Cluster-based Evaluation of Model Compensating Errors: A Case Study of Cloud Radiative Effect in the Southern Ocean, submitted to Journal of Geophysical Research: Atmospheres, 2019.

Schulzweida, U.: CDO User Guide Version 1.9.5, 2018.

Simmonds, I.: Modes of atmospheric variability over the Southern Ocean, Journal of Geophysical Research: Oceans, 108, 2003.

Simmonds, I. and Keay, K.: Mean Southern Hemisphere extratropical cyclone behavior in the 40-year NCEP–NCAR reanalysis, Journal of Climate, 13, 873–885, 2000.

Simmonds, I., Keay, K., and Lim, E.-P.: Synoptic activity in the seas around Antarctica, Monthly Weather Review, 131, 272–288, 2003.

Simpson, J., Kummerow, C., Tao, W.-K., and Adler, R. F.: On the tropical rainfall measuring mission (TRMM), Meteorology and Atmospheric physics, 60, 19–36, 1996.

Sinclair, M. R.: An objective cyclone climatology for the Southern Hemisphere, Monthly Weather Review, 122, 2239–2256, 1994.

Sinclair, M. R.: A climatology of cyclogenesis for the Southern Hemisphere, Monthly Weather Review, 123, 1601–1619, 1995.

Stephens, G. L., Vane, D. G., Boain, R. J., Mace, G. G., Sassen, K., Wang, Z., Illingworth, A. J., O'connor, E. J., Rossow, W. B., Durden, S. L., et al.: The CloudSat mission and the A-Train: A new dimension of space-based observations of clouds and precipitation, Bulletin of the American Meteorological Society, 83, 1771–1790, 2002.

Stuecker, M. F., Bitz, C. M., and Armour, K. C.: Conditions leading to the unprecedented low Antarctic sea ice extent during the 2016 austral spring season, Geophysical Research Letters, 44, 9008–9019, 2017.

Swales, D. J., Pincus, R., and Bodas-Salcedo, A.: The Cloud Feedback Model Intercomparison Project Observational Simulator Package: Version 2, Geoscientific Model Development, 11, 77–81, 2018.

Tange, O. et al.: Gnu parallel-the command-line power tool, The USENIX Magazine, 36, 42–47, 2011.

Taylor, P. C., Kato, S., Xu, K.-M., and Cai, M.: Covariance between Arctic sea ice and clouds within atmospheric state regimes at the satellite footprint level, Journal of Geophysical Research: Atmospheres, 120, 12 656–12 678, 2015.

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

Trenberth, K. E. and Fasullo, J. T.: Simulation of present-day and twenty-first-century energy budgets of the southern oceans, Journal of Climate, 23, 440–454, 2010.

Turner, J., Phillips, T., Marshall, G. J., Hosking, J. S., Pope, J. O., Bracegirdle, T. J., and Deb, P.: Unprecedented springtime retreat of Antarctic sea ice in 2016, Geophysical Research Letters, 44, 6868–6875, 2017.

Vergara-Temprado, J., Miltenberger, A. K., Furtado, K., Grosvenor, D. P., Shipway, B. J., Hill, A. A., Wilkinson, J. M., Field, P. R., Murray, B. J., and Carslaw, K. S.: Strong control of Southern Ocean cloud reflectivity by ice-nucleating particles, Proceedings of the National Academy of Sciences, 115, 2687–2692, 2018.

Wall, C. J., Hartmann, D. L., and Ma, P.-L.: Instantaneous linkages between clouds and large-scale meteorology over the Southern Ocean in observations and a climate model, Journal of Climate, 30, 9455–9474, 2017.

Walters, D., Baran, A., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., Furtado, K., Hill, P., Lock, A., Manners, J., Morcrette, C., Mulcahy, J., Sanchez, C., Smith, C., Stratton, R., Tennant, W., Tomassini, L., Van Weverberg, K., Vosper, S., Willett, M., Browse, J., Bushell, A., Dalvi, M., Essery, R., Gedney, N., Hardiman, S., Johnson, B., Johnson, C., Jones, A., Mann, G., Milton, S., Rumbold, H., Sellar, A., Ujiie, M., Whitall, M., Williams, K., , and Zerroukat, M.: The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations, Geosci. Model Dev. Discuss., in review, 2017.

Webb, M., Senior, C., Bony, S., and Morcrette, J.-J.: Combining ERBE and ISCCP data to assess clouds in the Hadley Centre, ECMWF and LMD atmospheric climate models, Climate Dynamics, 17, 905–922, 2001.

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., Chepfer, H., Douville, H., Good, P., Kay, J. E., et al.: The Cloud Feedback Model Intercomparison Project (CFMIP) contribution to CMIP6, Geoscientific Model Development, 10, 359–384, 2017.

Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee III, R. B., Smith, G. L., and Cooper, J. E.: Clouds and the Earth's Radiant Energy System (CERES): An earth observing system experiment, Bulletin of the American Meteorological Society, 77, 853–868, 1996.

Williams, J., Morgenstern, O., Varma, V., Behrens, E., Hayek, W., Oliver, H., Dean, S., Mullan, B., and Frame, D.: Development of the New Zealand Earth System Model: NZESM, Weather and Climate, 36, 25–44, 2016.

Williams, K. and Webb, M.: A quantitative performance assessment of cloud regimes in climate models, Climate dynamics, 33, 141–157, 2009.

Williams, K., Ringer, M., Senior, C., Webb, M., McAvaney, B., Andronova, N., Bony, S., Dufresne, J.-L., Emori, S., Gudgel, R., et al.: Evaluation of a component of the cloud response to climate change in an intercomparison of climate models, Climate Dynamics, 26, 145–165, 2006.

Williams, K., Bodas-Salcedo, A., Déqué, M., Fermepin, S., Medeiros, B., Watanabe, M., Jakob, C., Klein, S., Senior, C., and Williamson, D.: The Transpose-AMIP II experiment and its application to the understanding of Southern Ocean cloud biases in climate models, Journal of Climate, 26, 3258–3274, 2013.

Williams, K. D. and Bodas-Salcedo, A.: A multi-diagnostic approach to cloud evaluation, J. Geophys. Res., submitted, 10, 2017.

Wilson, D. R., Bushell, A. C., Kerr-Munslow, A. M., Price, J. D., and Morcrette, C. J.: PC2: A prognostic cloud fraction and condensation scheme. I: Scheme description, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 134, 2093–2107, 2008a.

Wilson, D. R., Bushell, A. C., Kerr-Munslow, A. M., Price, J. D., Morcrette, C. J., and Bodas-Salcedo, A.: PC2: A prognostic cloud fraction and condensation scheme. II: Climate model simulations, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 134, 2109–2125, 2008b.

Winker, D., Pelon, J., Coakley Jr, J., Ackerman, S., Charlson, R., Colarco, P., Flamant, P., Fu, Q., Hoff, R., Kittaka, C., et al.: The CALIPSO mission: A global 3D view of aerosols and clouds, Bulletin of the American Meteorological Society, 91, 1211–1230, 2010.

Zhang, Y., Xie, S., Klein, S. A., Marchand, R., Kollias, P., Clothiaux, E. E., Lin, W., Johnson, K., Swales, D., Bodas-Salcedo, A., et al.: The ARM Cloud Radar Simulator for Global Climate Models: Bridging Field Data and Climate Models, Bulletin of the American Meteorological Society, 99, 21–26, 2018.

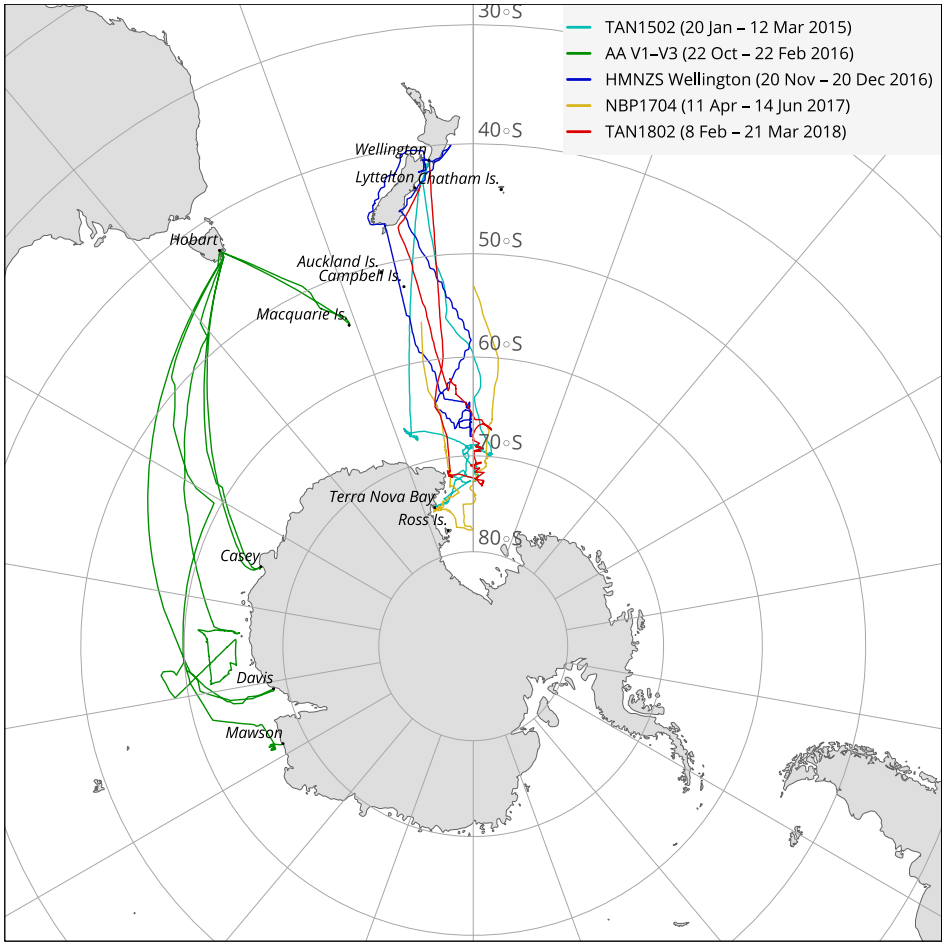


**Figure 1.** Map showing tracks of voyages used in this study. The ship observational dataset comprises 5 voyages between 2015 and 2018, spanning months from November to June and latitudes between 40°S and 78°S, of which data between 50°S and 70°S are used in this study.

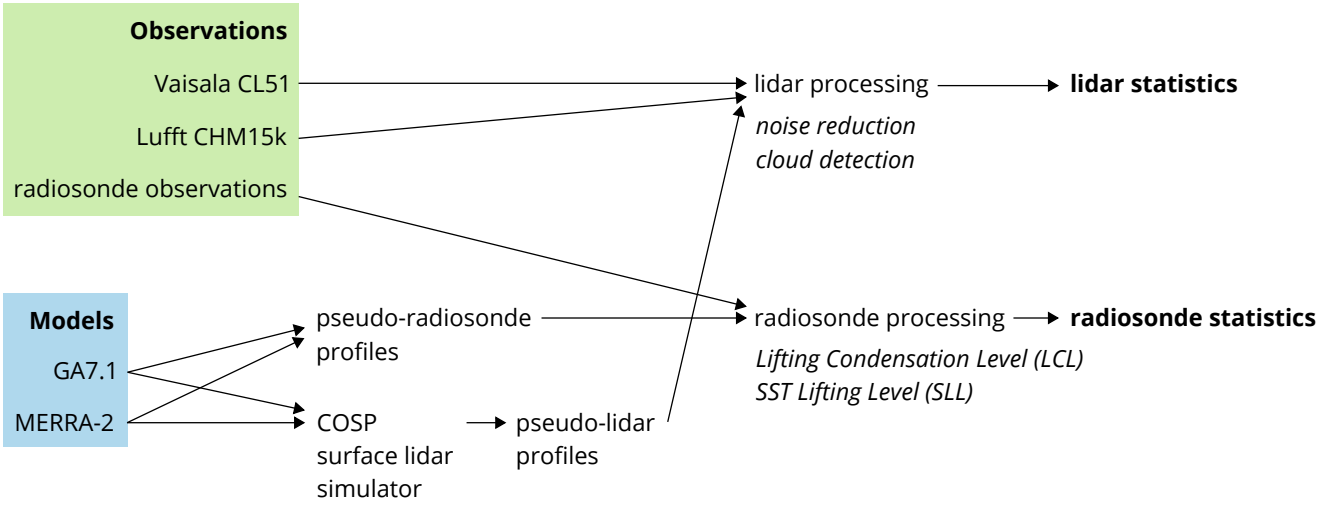


**Figure 2.** Schematic of the processing pipeline utilised in this study to produce lidar and radiosonde statistics from observations and model data.

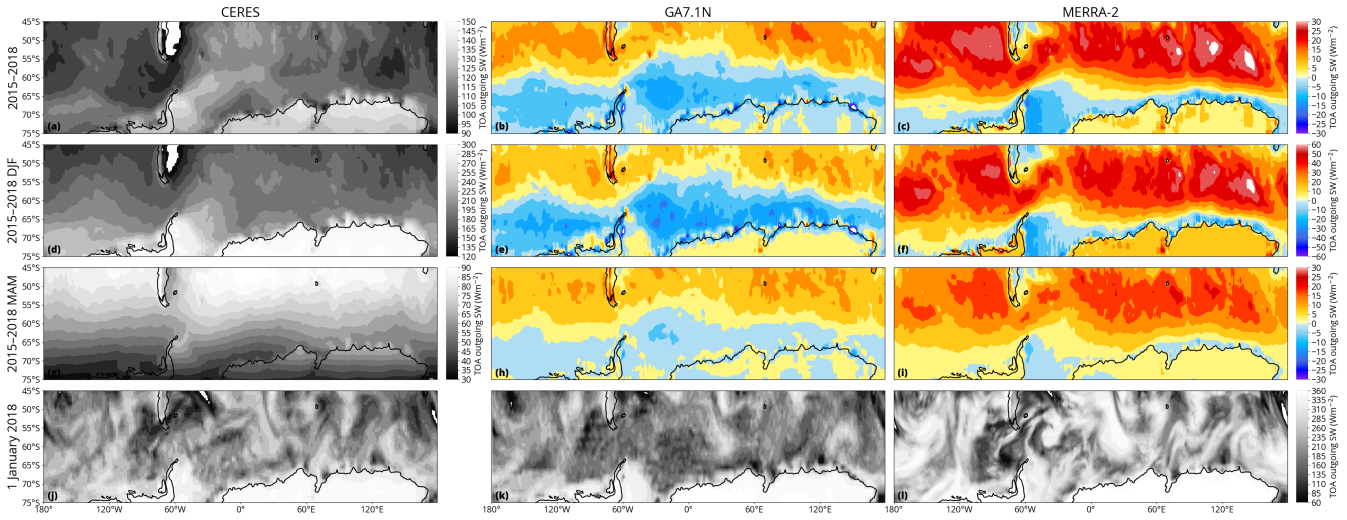


**Figure 3.** Geographical distribution of the TOA outgoing SW ~~upwelling~~ radiation in CERES, GA7.1N and ~~multiple models~~MERRA-2. The plots show global all sky SW radiation as ~~a yearly~~ annual (~~2007~~2015–2018; **~~a~~a–~~d~~**d–c), ~~monthly~~ seasonal (~~January 2007~~2015–2018 DJF, MAM; **~~e~~e–~~h~~**h–i) ~~,~~ and daily (~~19~~ 1 January ~~2007~~2018; **~~i~~i–~~l~~**l–l) ~~average and monthly average~~ mean. The blue–red colormap shows bias relative to CERES (~~January 2007;~~ **~~n~~n–~~p~~**p). ~~Highlighted are multiple latitude bands (55~~b, ~~60~~c, ~~65~~e, ~~70°S). In~~ **~~n~~n–~~p~~**f, ~~positive (red~~h, i)~~values indicate that the model overestimates reflected SW radiation~~, while ~~negative~~ the grayscale colormap shows absolute values (~~blue~~a, d, g, j, k, l)~~values indicate that the model underestimates reflected SW radiation relative to CERES~~.

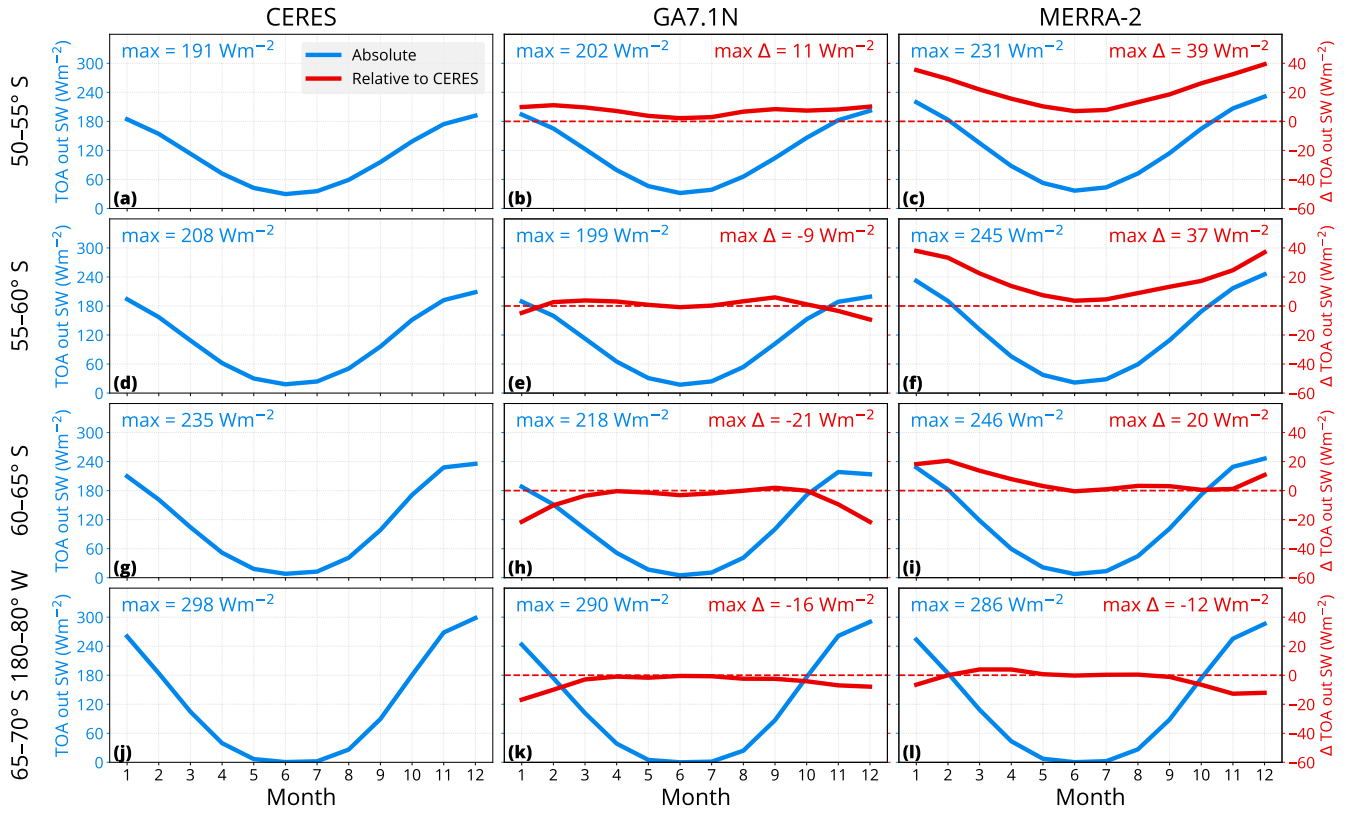


**Figure 4.** Zonal means of the TOA outgoing SW ~~upwelling~~ radiation in CERES, GA7.1N and ~~multiple models~~ MERRA-2 during the ~~year 2007~~ years 2015–2018 in ~~several~~ 5-degree latitude bands between 50 and 70°S. The plots show ~~time series of~~ monthly zonal mean TOA outgoing SW ~~upwelling~~ radiation (blue) and its difference relative to CERES (red) as a function of month. Shown are also the maxima ~~of the SW radiation~~ ("max") and ~~its~~ the difference from CERES ("max Δ").

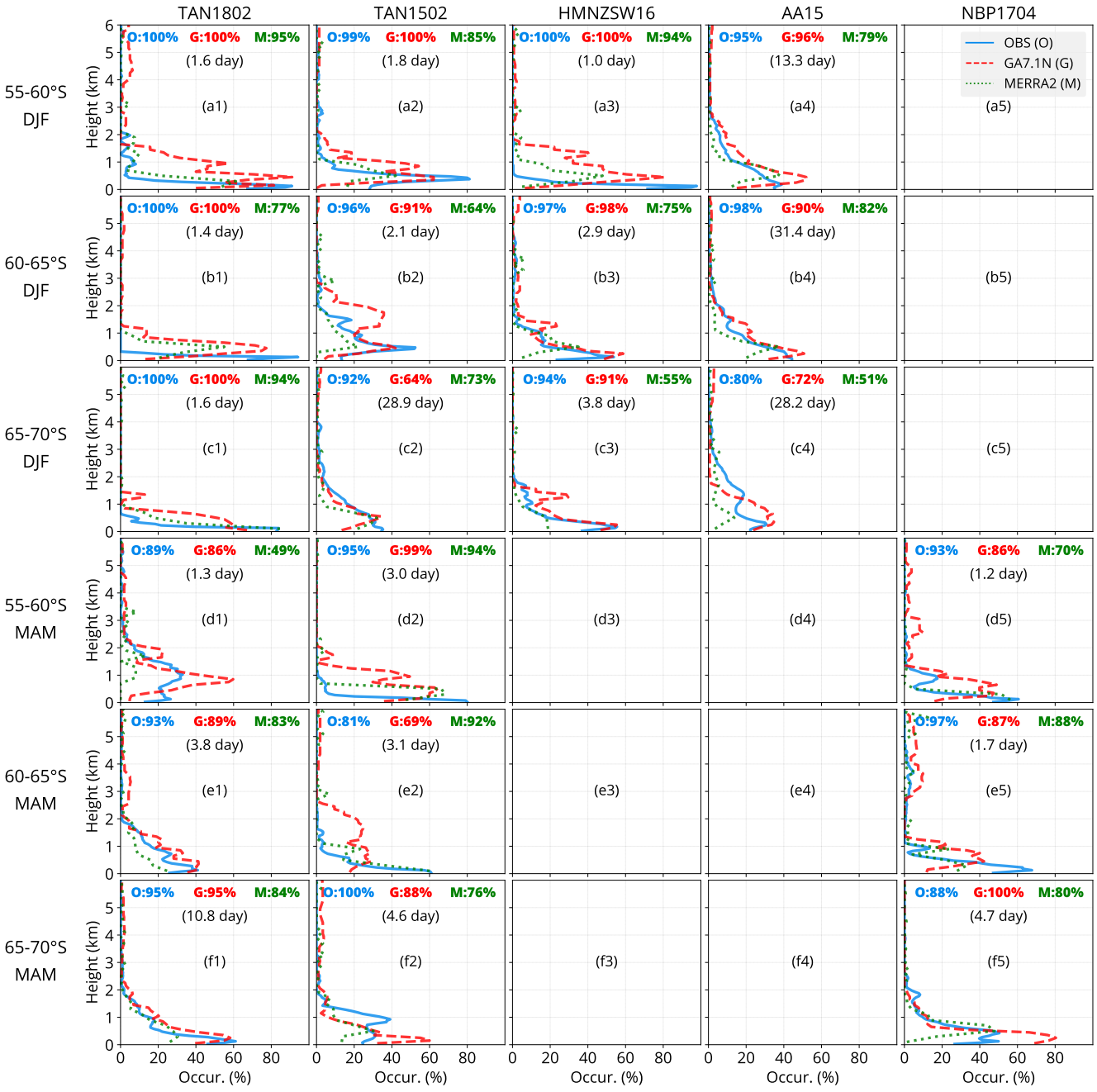


**Figure 5.** Cloud occurrence frequency as a function of height derived from ceilometer ~~backscatter~~ observations (OBS) and model fields (GA7~~.0U~~ .1N and MERRA-2). The observational and model data were subsetted by latitude and season (DJF~~/December-January-February~~, MAM~~/March-April-May~~) along the voyage track. The numbers at the top of each panel show total (vertically integrated) cloud cover and the number of days the ship spent passing through the spatiotemporal subset. ~~A 1-$\sigma$ confidence band calculated from a set of 10 years of the free-running GA7.0 simulation is indicated by a semi-transparent red band.~~ The height in the plots is limited to 6 km. There was no significant amount of cloud detected above this level.

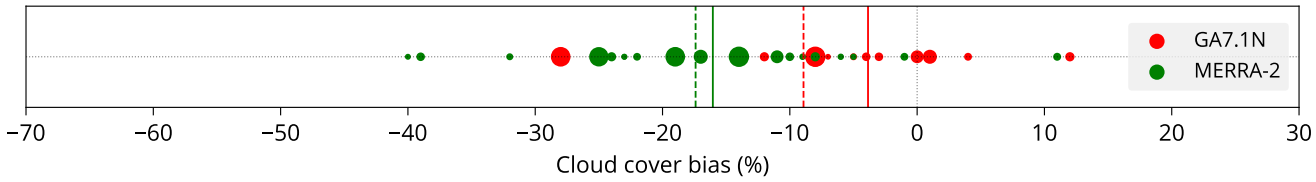


**Figure 6.** Cloud cover bias in models relative to observations. The points represent subsets as in Figure 5. The size of the circles is proportional to the number of days of observations in the subset. The solid lines are averages, and dashed lines are averages weighted by the number of days the ship spent passing through the spatiotemporal subset.

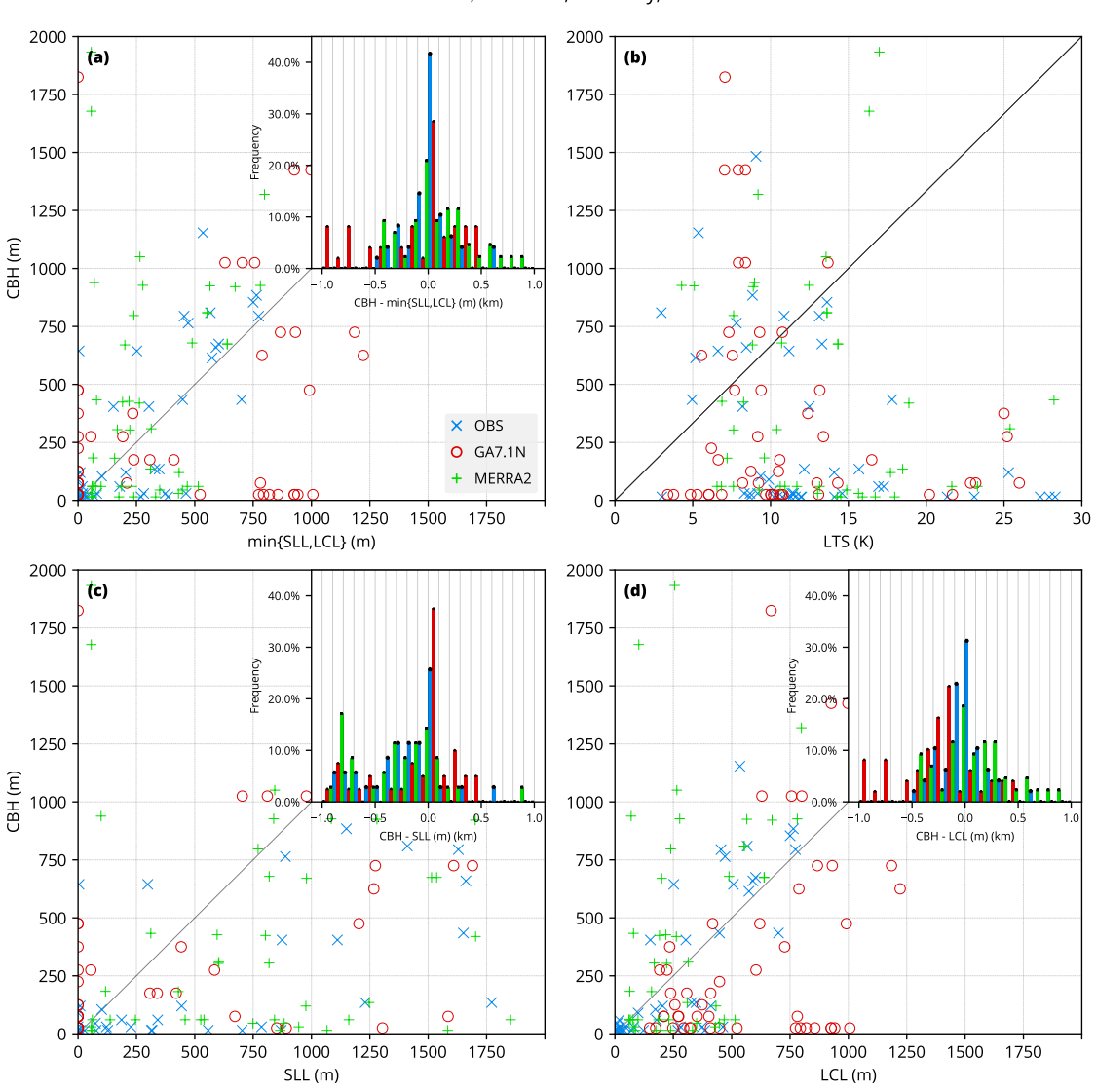


**Figure 7.** Scatter plots of radiosonde measurements on the TAN1802 and NBP1704 voyages between ~~December~~ February and May ~~(inclusive)~~ and 60–70°S latitude. Corresponding profiles from GA7~~.0U/1980-89~~ .1N and MERRA-2 are selected, i.e. having the same geographical coordinates and the same time of the year. Each point on the scatter plots represents a radiosonde ~~measurement~~profile. The plots compare three datasets: observations (OBS), GA7~~.0U/1980-89~~ .1N and MERRA-2. The ~~points of GA7.0U/1980-89 (free-running) are selected randomly from years 1980 to 1989 of the simulation. The~~ radiosonde ~~measurements~~ observations are matched with ceilometer (OBS) and COSP-based CBH (GA7~~.0U/1980-89~~ .1N and MERRA-2). ~~(a)~~ (a) shows the points as a function of min{SLL, LCL} and CBH. The inset histogram shows distribution of the difference of CBH and min{SLL, LCL} in bins of 100 m, where each bin contains three bars for the three datasets. ~~(b)~~ ~~shows~~ (b, c, d) show the points as a function of LTS, SLL and ~~CBH~~LCL, respectively.

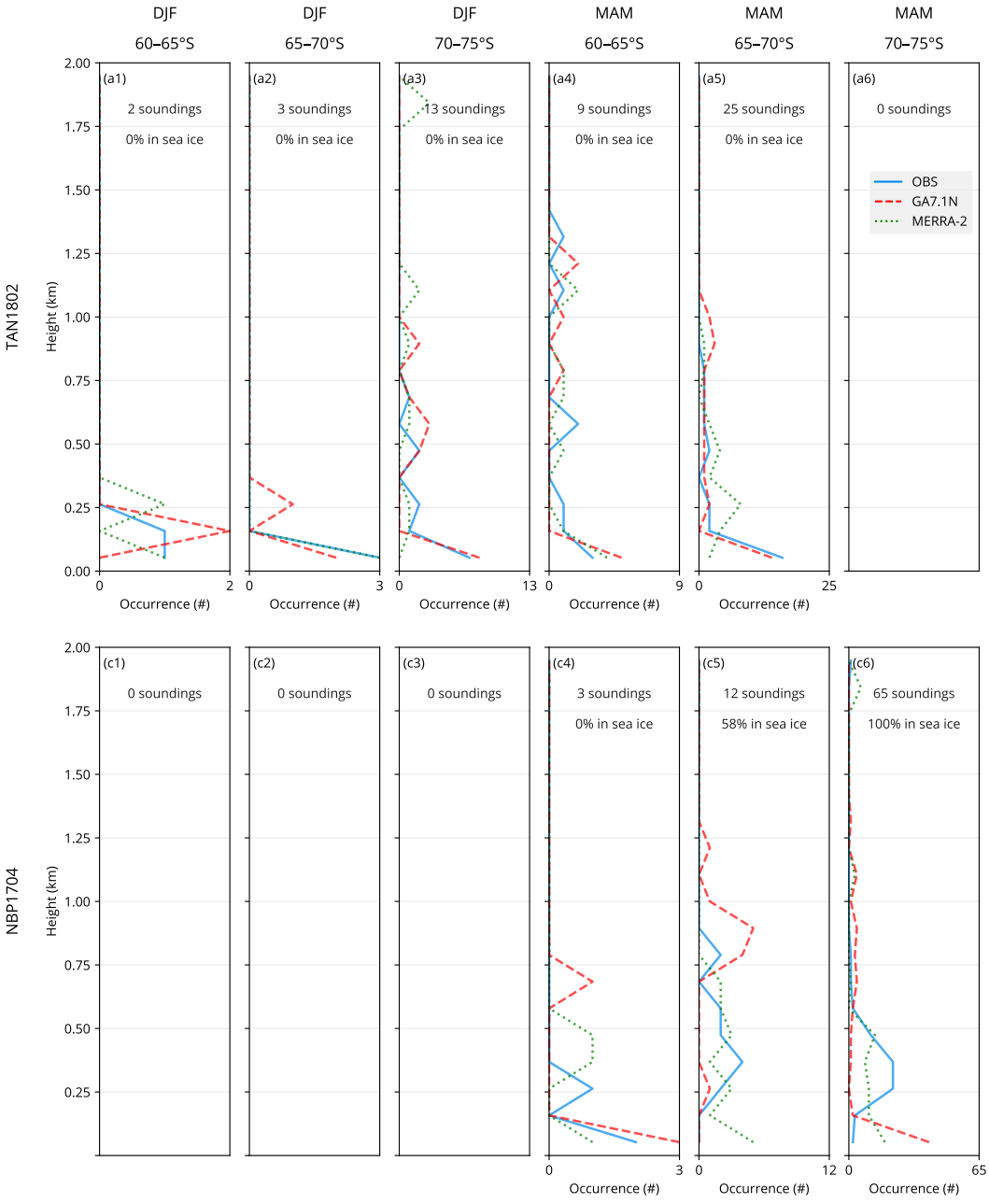


**Figure 8.** ~~SLL distribution panel plot. The plots show histograms of SSL as a function~~ Histogram of ~~pressure~~ min{SLL,LCL} derived from radiosonde ~~measurements~~ observations (OBS) on TAN1802 and ~~model fields (GA7.0U~~NBP1704, ~~MERRA-2) (a–f, m–r)~~ and ~~scatter plots of cloud base height (CBH) vs~~the equivalent profiles in GA7.1N and MERRA-2. ~~minimum of SLL~~ Shown are subsets by latitude between 60 and ~~LCL corresponding to the plots above (g–l, s–x)~~75°S and seasons DJF and MAM. The numbers at the top of each panel indicate the number of ~~soundings~~ profiles which make up the histogram and the percentage of sea ice cases ~~as~~ determined ~~by a~~ from NSIDC ~~satellite derived~~ satellite-derived sea ice concentration~~product. The histograms and scatter plots are binned by season and latitude (column) and voyage (row)~~.

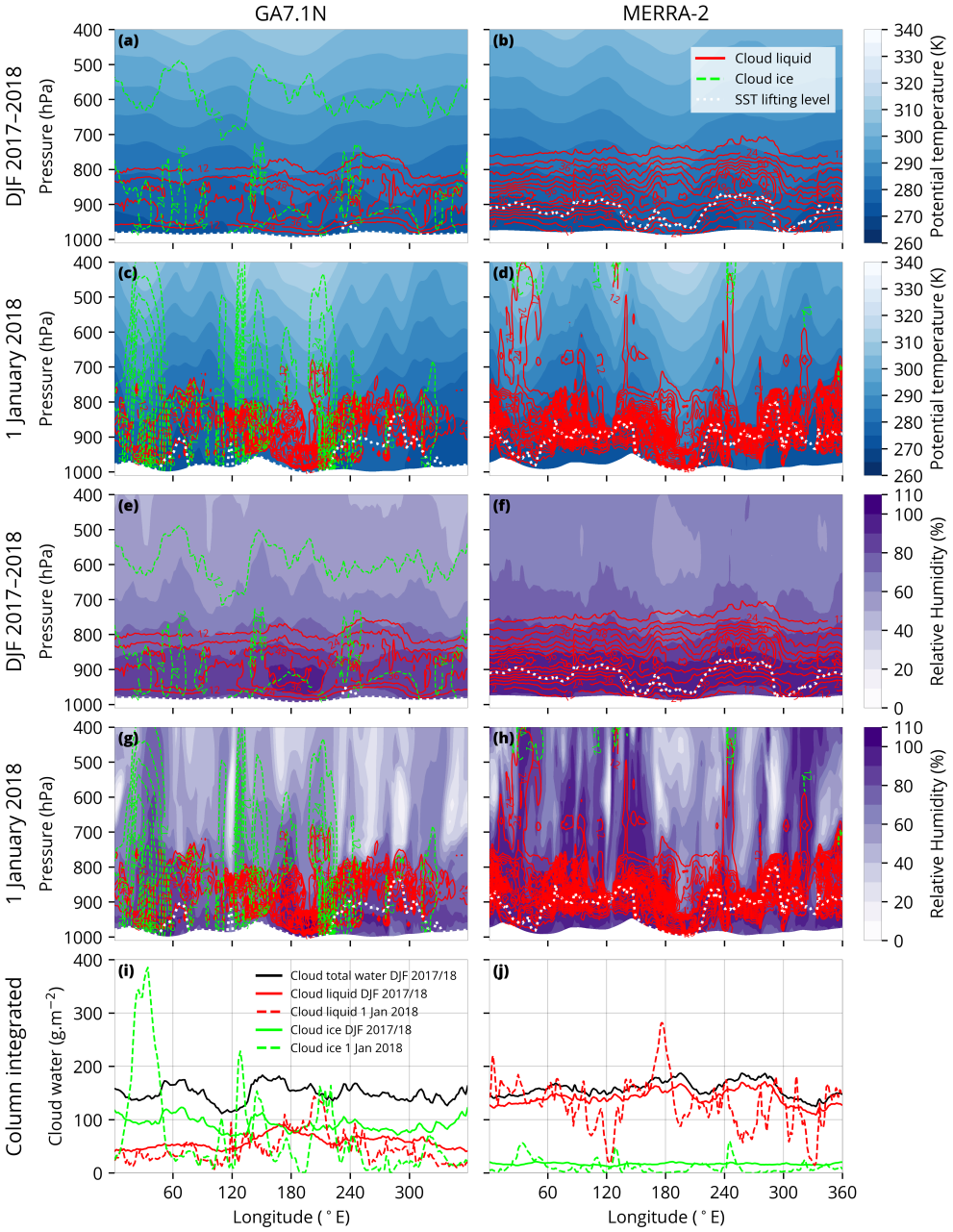


**Figure 9.** Zonal plane plot of cloud liquid and ice mixing ratios in GA7~~.0~~.1N and MERRA-2 at 60°S. The cloud liquid and ice mixing ratios are plotted as contours on top of the potential temperature fields ~~**(a–d)**~~ (a–d) and relative humidity fields ~~**(e–h)**~~(e–h). SLL is indicated by a white line. ~~**(a), (b), (e), (f)**~~ (a, b, e, f) show a ~~monthly~~ seasonal average in ~~January 2007~~ DJF 2017/2018 and ~~**(c), (d), (g), (h)**~~ (c, d, g, h) show a daily average on ~~19~~ 1 January ~~2007. **(i), (j)**~~ 2018. (i, j) show the column-integrated values of cloud liquid and ice water as a function of longitude corresponding to the plots above~~, January 2007 ("monthly") and 19 January 2007~~. All liquid shown in the plots is supercooled (~~"daily"~~air temperature is less than 0 °C everywhere).

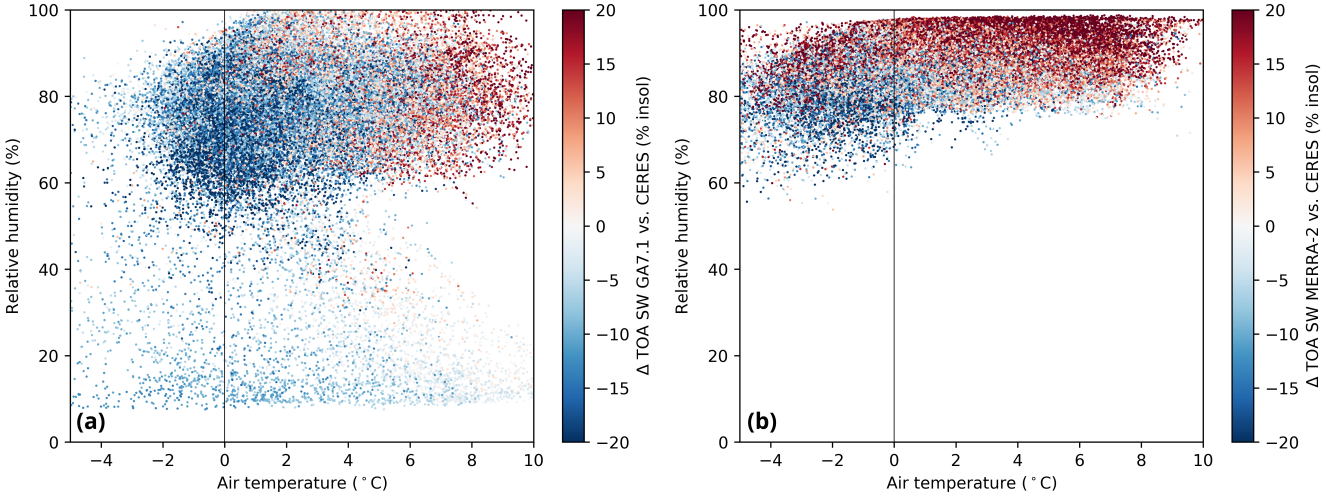


**Figure 10.** Scatter ~~**(a)** and density **(b)**~~ plot of SW radiation bias in (a) GA7~~.0~~.1N ~~/2007~~ and (b) MERRA-2 grid cells between ~~40~~55°S and 70°S ~~as~~ in January 2018. Each point represents a daily average ~~on 19 January 2007. **(a)** shows~~ of SW radiation bias as a function of near-surface air temperature and near-surface relative humidity. ~~**(b)** shows the density of points as a function of near-surface air temperature and the SW radiation bias.~~ The bias is expressed as a percentage of the incoming solar radiation in the grid cell. ~~Each point represents~~ The points are a ~~single model grid cell. -2 and +2 air temperature is marked by dashed lines~~random sample of 100000 points.

**Table 1.** Table of voyages. The table lists voyages analysed in this study. Listed is the voyage name (Voyage), which is the official name of the voyage or an abbreviation for the purpose of this study, ship name (Ship), organisation (Org.), start and end dates of the voyage (Start, End), number of days spent at sea (Days), target region of the SO (Region), maximum and minimum geographical coordinates of the voyage track (Lat., Lon.).

| **Voyage** | **Ship** | **Org.** | **Start** | **End** | **Days** | **Region** | **Lat.** | **Lon.** |
|---|---|---|---|---|---|---|---|---|
| TAN1502 | RV *Tangaroa* | NIWA | 2015-01-20 | 2015-03-12 | 51 | Ross Sea | 41°S–75°S | 162°E–174°W |
| TAN1802 | RV *Tangaroa* | NIWA | 2018-02-08 | 2018-03-21 | 41 | Ross Sea | 41°S–74°S | 170°E–175°W |
| HMNZSW16 | HMNZS *Wellington* | RNZN | 2016-11-20 | 2016-12-20 | 20 | Ross Sea | 36°S–68°S | 166°E–180°E |
| NBP1704 | RV *Nathaniel B. Palmer* | NSF | 2017-04-11 | 2017-06-13 | 63 | Ross Sea | 53°S–78°S | 163°E–174°W |
| AA15 (AA V1–V3) | *Aurora Australis* | AAD | 2015-10-22 | 2016-02-22 | 123 | Indian O. sector | 42°S–69°S | 62°E–160°E |

**Table 2.** Table of deployments. The table cells indicate if data from a given instrument (row) was available from a voyage (column).

| Instrument/Voyage | AA15 | TAN1502 | HMNZSW16 | NBP1704 | TAN1802 |
|---|---|---|---|---|---|
| Lufft CHM 15k | | | ✓ | ✓ | ✓ |
| Vaisala CL51 | ✓ | ✓ | | | |
| iMet radiosondes | | | | | ✓ |
| Radiosondes (other) | | | | ✓ | |

**Table 3.** A table showing a "back-of-the-envelope" calculation how the GA7.1N peak TOA outgoing SW radiation bias (Figure 4) would change if the cloud cover were increased by 5% (Figure 6), asssuming the cloud albedo does not change. The "corrected" TOA outgoing SW radiation is calculated by multiplying the original value by 1.05.

| Latitude | TOA out. SW at max. $\Delta$ ($Wm^{-2}$) | Max. $\Delta$ TOA out. SW ($Wm^{-2}$) | Corrected Max. $\Delta$ TOA out. SW ($Wm^{-2}$) | Explained error |
|---|---|---|---|---|
| 55–60°S | 199 | -9 | 0.95 | 111% |
| 60–65°S | 214 | -21 | -10.3 | 51% |
| 65–70°S | 243 | -16 | 3.85 | 76% |