Manuscript prepared for Geosci. Model Dev. Discuss.
with version 4.1 of the LaTeX class copernicus_discussions.cls.
Date: 30 March 2017

Author's response to referee 1 on "A multi-diagnostic approach to cloud evaluation"

K. D. Williams and A. Bodas-Salcedo

Correspondence to: K. D. Williams
(keith.williams@metoffice.gov.uk)

1 Major comments from referee 1

1.1 Referee Comments

My primary concern is on the scientific focus of this study. The title of the paper seems to suggest that the aim of this work is to introduce "a multi-diagnostic approach to cloud evaluation". However, the paper has spent a lot of time on the inter-comparison of the two configurations of the UM model. I have no problem with whichever topic the study is designed to focus on, as both topics have their own values. However, since the study "tries" to cover two topics at a time, the discussions are somewhat lacking in depth. Therefore, the paper reads more like a report.

If the study is designed to focus on introducing a new multi-diagnostic approach, then a thorough introduction of this approach, including the developments of individual diagnostic methods (including necessary technical details), their merits and limitations, their applications in the literature, as well as a quantitative estimate of the uncertainties of these methods, should be fully discussed. The authors have discussed some of the above mentioned aspects, but only to a very limited extent.

If the study is designed to focus on the evaluation of the simulations, then I have real trouble in understanding what have been done in the new configuration. Section 2a provides a general summary of the changes that have been made, but necessary details such as what processes or parameters have been added or changed in the parameterizations are lacking. Also, there is no dedicated case study to investigate the model performance in depth (except a snapshot in Figure 3 and Figure 8). As such, it is very difficult for a reader to appreciate what differences in the simulations can be considered as a real improvement. This is particularly true when considering the presence of new errors in the new configuration for some cloud properties.

1.1.1 Author's response

The aim of the paper is to show how a comprehensive approach to cloud evaluation can be valuable in developing and assessing a new model configuration. In order to achieve this aim we believe we need to show both the multi-diagnostic approach and how it can be used to

assess the performance of a new model configuration against a control. However, in the revised manuscript we address the referee's concern that not enough depth of information is given. We have deliberately taken the approach of, wherever possible, using published methods, hence a technical discussion of these methods already exists in the published literature and we believe it would make the paper too cumbersome to repeat it all. The novel aspect here is that we draw the techniques together for the purpose of assessing a new model configuration as part of the model development process. Hence we have provided more detail of the parametrization changes made in the model development process and the relative merits of the different diagnostic approaches (e.g. why one observational dataset might be chosen over another to look at a particular aspect of the cloud simulation). We feel that this discussion of the diagnostic approach is best placed within the results sections to highlight the point that the chosen approach will vary depending on the particular characteristic being examined.

1.1.2 Manuscript changes

Section 2a has been considerably expanded with a more thorough description of the relevant parametrization changes. Within section 3, where possible, the text attributing changes in the errors to particular parametrization changes has been expanded (e.g. around the warm rain microphysics discussion) to discuss how the parametrization differences lead to the improvement and the physical processes operating. The results of two new simulations have been added to Figure 2 in order to clearly attribute the differences seen to particular parametrization changes.

The description of the observational datasets and, where relevant to the paper, their uncertainties has been expanded in Section 2b and in the results sections. In a number of places, we have enhanced the discussion of the value of the multi-diagnostic approach and the increased process-orientated understanding it can provide (e.g. around the mid-latitude cyclone RSW error and cloud errors over mid-latitude land).

1.2 Referee Comments

My second concern is on the comparison of model simulations against satellite observations (e.g. Figure 7, 9, 10, and similarly supplementary Figure 2-4). Many differences are discussed; however, these is no discussion on their statistical significance. How much of the difference is due to sample errors and how much is due to systematic errors in the model? In my view, a significance test should be applied to the analysis to insure that the differences discussed are meaningful. To do this you can use something simple such as a t-test or more appropriately a Monte Carlo method as applied in Booth et al. (2013).

1.2.1 Author's response

We have now conducted a t-test based on the inter-annual variability of the observations and the models for the figures the referee refers to (and several others where this could be considered an issue). As expected, all the systematic errors discussed in the paper are considerably larger than the inter-annual variability and so remain significant.

1.2.2 Manuscript changes

Figures 2, 4, 5 and 11 have been reprocessed with shading around the line plots to represent 5% significance. For Figures 7, 9 & 10, the region of $<5\%$ significance has been coloured white so that all coloured regions in these plots show statistically significant differences. The significance test is also referred to in section 2b.

2 Specific comments from referee 1

2.1 Comment

Line 62-63: that's fine, but you also have spent a lot of time on inter-comparison of the two configurations of the model.

2.1.1 Response & manuscript change

The purpose of the paper is to show how a comprehensive approach to cloud evaluation can be valuable in developing and assessing a new model configuration. The sentence has been altered in the revised manuscript to indicate this.

2.2 Comment

Line 66: "high", "mid", and "low" clouds need to be defined.

2.2.1 Response & manuscript change

Definitions have been added to the manuscript as low:>680hPa, mid:440hPa–680hPa, high:<440hPa.

2.3 Comment

Line 73: please define "NWP".

2.3.1 Response & manuscript change

This was already defined on line 24.

2.4 Comment

Line 97-117: a summary of the changes is good, but what changes have actually been made? What processes or parameters have been added or modified in the parameterizations? For example, what has been changed in the auto-conversion scheme (line 101)? What does the change do in the new aerosol scheme (line 112)? What does the turbulent scheme do to the production of liquid water (line 110)? You have provided the references, but some necessary details would be appreciated by the readers and would help justify your argument of the model improvement.

2.4.1 Response & manuscript change

This section of the paper is considerably expanded in the revised manuscript with a more detailed explanation of the parametrization changes as the referee requests. It should also be noted that it is intended that the present paper will be included within a GMD special issue which will also include the GA7 model description paper containing a full documentation of all the parametrization changes.

2.5 Comment

Line 145 and 147: CloudSat and CALIPSO provide a "curtain view" of the clouds, which are not really 3-D.

2.5.1 Response & manuscript change

'3D structure' has been replaced with 'hydrometeor profile'

2.6 Comment

Line 180: so how many years are used exactly?

2.6.1 Response & manuscript change

The following has been added to the revised manuscript "25 years for ISCCP, 12 years for CERES-EBAF and 5 years for CloudSat/CALIPSO"

2.7 Comment

Line 194: but you said "3-D" before

2.7.1 Response & manuscript change

The '3D' on line 146 (now removed) referred to CloudSat. Here we are stating that CALIPSO provides the best 2D (latitude/longitude) estimate of total cloud fraction; this doesn't preclude it from having useful information in the vertical as well.

2.8 Comment

Line 212: "this corrected in GA7" should be "this is corrected in GA7".

2.8.1 Response & manuscript change

Revised manuscript has been corrected as reviewer suggests.

2.9 Comment

Line 219: it does appear to be the case in GA6 to me. Please clarify.

2.9.1 Response & manuscript change

Referring to the top left panel of Figure 2b we can see no evidence that the altitude of the cirrus with lower backscattering ratios (3-5) is any higher than the thicker cirrus (backscattering ratios 7-20) - if anything the reverse is true. This is unlike the panels for GA7 and CALIPSO which show the cirrus with the lowest backscattering ratios to be higher. We really can't see how we can make this clearer and request that the referee looks again at the text and figure.

2.10 Comment

Line 224: in this case I see the model produces a lot of mid-top clouds (which seem to have moderate optical depth) whereas you argue earlier (line 191) that the model simulates too little of this type of cloud?

2.10.1 Response & manuscript change

The hydrometeor signal observed is likely to be the spurious large scale precipitation referred to in the discussion of Figure2c and result from thin large scale cloud which has formed in the moist air around the convective system. As they occur under the anvil of a deep-convective system they won't be seen by ISCCP, and most of them may not be seen by CALIPSO either due to full attenuation from the ice cloud above. In contrast, the mid-top cloud which is 'missing' should be visible to CALIPSO (almost certainly it is missing congestus-type cloud). An extra paragraph has been added to the manuscript discussing these points.

2.11 Comment

Line 230: how "cloud fraction" is defined in the simulation and in the observational data set, respectively? Is a direct comparison meaningful? Please clarify.

2.11.1 Response & manuscript change

The radar–lidar product has considerably higher along track resolution (nominally 1.7km) than the model (80km at the equator), hence regridding the combined radar-lidar data onto the model grid gives an observed cloud fraction to a precision of about 2%. The main assumption here is that the along-track cloud fraction is representative of the 2D grid box. Whilst this is a fair assumption when considering a large number of cases which the A-train will cross at random orientations, we acknowledge that there may be an error when considering a single case such as this. However it's unlikely to affect the key model errors discussed in the paper regarding the figure. These points and caveats have been expanded upon in the revised manuscript as the reviewer suggests.

2.12 Comment

Line 238-239: a lot of these "drizzling" clouds in the simulations have a reflectivity below -20 DBZ, which is very, very weak. It seems odd that these clouds are not picked up by the CALIPSO simulator at all.

2.12.1 Response & manuscript change

As we note in the paper, the rates are $<0.005mm/hr$ which is consistent with the very weak signal. The concrete evidence given in the paper is that if large scale rain is not passed to the CloudSat simulator then the signal is removed. As we are below a thick anvil, the CALIPSO simulator signal is likely to have been attenuated and so not see cloud if present. However in GA7, once the spurious precipitation is removed, there is still a cloud signal in the CloudSat simulator just below the threshold of -40dBZ which suggests that the cloud is very thin.

2.13 Comment

Line 257-258 and relevant texts throughout the paper: care should be taken when drawing this conclusion. Previous studies (e.g. Chepfer et al. 2013) have shown that, due to the averaging issue, differences in the zonal cloud fraction retrieved in different CALIPSO products can be quite large (up to a factor of 2 for some regions). It is not unlikely that the GOCCP may have underestimated the cirrus extent.

2.13.1 Response & manuscript change

The reviewer is correct, and GOCCP probably underestimates the amount of cirrus. Chepfer at al. (2013) show that the averaging effect is sensitive to the length of the averaging and is higher for low-level, small-scale broken cloud. For high clouds, the differences between GOCCP and the CALIPSO cloud retrieval used by RL-GEOPROF are dominated by the SR detection threshold. The height-dependent SR detection threshold used in this study increases the sensitivity to high clouds (supplementary Figure 1). For cirrus clouds in the regions shown in Figure 2, the

bias introduced by lack of averaging smaller than 0.05 (Figure 10 in Chepfer at al., 2013). This supports the interpretation that GA6, and to a lesser extent GA7, overestimates cirrus. This discussion has been added to the manuscript.

2.14 Comment

Line 298: this is a fairly big box. While I understand that this is a standard method used in previous studies, I am not convinced that it is appropriate for high-latitude regions, where cyclones (e.g. polar vortex) are generally smaller in size and the distance between individual cyclones can be a lot smaller (compared to mid-latitude cyclones).

2.14.1 Response & manuscript change

Throughout the paper we have tried to use published methodologies. As the referee notes, this is a standard size used in other studies looking at similar latitude bands. We acknowledge that cyclones come in a range of sizes and that this technique will just combine them all and hence smooth the signal (this point has been added to the revised manuscript), however we have tested different box sizes and the results are qualitatively similar and conclusions unchanged.

2.15 Comment

Line 331-322: This is a complicated case, with multiple fronts being diagnosed. Therefore it is hard for me to associate the cloud features discussed in the paper to the synoptic components shown on the MSLP chart. Further information such as latitude and longitude on the discussed cloud fields should help.

2.15.1 Response & manuscript change

Latitude references have been added to the revised manuscript as the reviewer suggests.

2.16 Comment

Line 332-334: I don't understand this sentence.

2.16.1 Response & manuscript change

This and adjacent sentences have been re-written to explain the point more clearly and a reference added as the result is consistent with previous experience.

2.17 Comment

Line 364 and relevant text later in the paper: I disagree. What I have seen is that the large RSW bias is present in some of the cold air side of the cyclone, but almost everywhere in the poleward side of the cyclone! Why? This is, to me, an important issue but no discussion has been made in the paper (or the referenced study). There is a lot of focus on the cold air side of the cyclone, but this is only part of the story revealed by the plot. Also, the bias on the cold air side of the cyclone does not explain the poleward increase of the radiative bias shown in Figure 13.

2.17.1 Response & manuscript change

We don't really understand the referee's distinction between the "cold air" and "poleward" side of the cyclone. On average, the poleward side of the cyclone is a subset of the cold air which also extends around the western side of the cyclone. By "cold air" we are referring to all of the region from the poleward side of the cloud head associated with the typical warm front position of about 3 o'clock on the southern hemisphere composite in figure 9, around the poleward and westward side to the poleward edge of the typical cold front position at 10-11 o'clock. In the revised manuscript we have clarified that cold air includes the poleward side of the cyclone.

2.18 Comment

Line 368: I don't think Figure 10 is necessary. It does not seem to provide any substantially different information than Figure 9.

2.18.1 Response & manuscript change

The key difference is that RSW in Figure 9 will depend on the insolation, hence for the same cloud albedo error, the RSW error will be larger in the summer than winter. Figure 10 shows the in-cloud albedos which do not depend on the insolation (this point has been explicitly added to the revised manuscript). Hence Figure 10 shows the interesting point that the in-cloud albedos are lower in the austral summer compared with the winter which, combined with the stronger insolation, leads to the larger negative RSW bias in Figure 9.

2.19 Comment

Line 369-370: why the cloud amount errors are not large enough to contribute significantly to the SW errors? Please explicate.

2.19.1 Response & manuscript change

We accept that this statement was probably too strong and requires correction which we have done in the revised manuscript. The cloud fraction errors for GA7 in Figure 7 and the supplementary material do not have the same spatial pattern (and are sometimes of the wrong sign) to explain the RSW error.

2.20 Comment

Line 373-374: again, the errors seem to be prevalent in the poleward side of the cyclone, too (which is also the case in the referenced study).

2.20.1 Response & manuscript change

See above response regarding “poleward” and “cold air”.

2.21 Comment

Line 411: why this appears to be an issue for the UM? Please explicate.

2.21.1 Response & manuscript change

Sentence has been revised in the manuscript to “This appears to be an issue for the UM in parts of the tropics where too little shallow cumulus is simulated and typically the model has cloud fractions of <3 oktas whereas fractions over this threshold are often observed and hence a cloud base height assigned”.

2.22 Comment

Line425-426: could some of these excess low clouds actually be precipitation not being detected by the instrument?

2.22.1 Response & manuscript change

Here we are comparing the model cloud (no simulator involved) with SYNOP observations so the presence, or not, of precipitation should be irrelevant.

2.23 Comment

Line 447-448: but now there seems to be too much red (for RSW) in the sub-tropics which was non-existent in HadGEM2-A?

2.23.1 Response & manuscript change

The sentence has be revised to read "The error in the sub-tropical cumulus transition regions of excess RSW has been removed and there is now a smaller negative bias in GA7". We have also reproduced Figure 13 with a revised colour bar to make it clearer that the negative bias in GA7 is smaller in magnitude than the positive bias in HadGEM2-A.

2.24 Comment

Figure 2: (1) you use "equivalent reflectivity factor" in the plot but "reflectivity" in the caption (and the following figures). (2) What do the colour bars show? It should be indicated in the plots. (3) The lowest km should be masked in the CloudSat plot (as done in your following figures).

2.24.1 Response & manuscript change

The term "reflectivity" is now used throughout (including figures). Indication of what the colour bars show has been added to the Figure 2 caption and explanation of how the histograms are formed has been added to section 2. The lowest 1.2km has now been masked for CloudSat (the main instrument to suffer from near-surface issues).

2.25 Comment

Figure 3: What does "CloudSat/CALIPSO" mean in the top left plot while you only show reflectivity?

2.25.1 Response & manuscript change

The caption to the figure explains that "...situations in which the lidar detected cloud but the radar did not being included with a nominal value of -40dBZ (e.g. Mace and Wrenn, 2013)"

2.26 Comment

Figure 5: I don't think including the very cold SST ranges is necessary as they are quite rare and the plots show very similar features (i.e. the first four plots in the bottom panel).

2.26.1 Response & manuscript change

Although relatively rare spatio-temporally, these typically represent the subtropical stratocumulus regions which are widely regarded by the cloud feedback community to be critical for the cloud response to climate change, hence their separation by SST is relevant. In addition, the coldest SST bin shows a significant (i.e. still distinguishable following the significance tests discussed above have been applied) error in GA6 which has been improved in GA7.

2.27 Comment

Figure 8: is 64.32N (the right end of the cross-session plots) over land? I can see the topography-like feature at the surface in the bottom left plot, but why there are clouds produced underneath the surface in the simulations?

2.27.1 Response & manuscript change

Neither cross section is over land. The masked region in the lower-right corner corner of the observed cross-section is due to the reliability mask associated with the RL-GEOPROF product indicating high uncertainty in both CloudSat and CALIPSO for these bins.

Manuscript prepared for Geosci. Model Dev. Discuss.
with version 4.1 of the LaTeX class copernicus_discussions.cls.
Date: 30 March 2017

Author's response to referee 2 on "A multi-diagnostic approach to cloud evaluation"

K. D. Williams and A. Bodas-Salcedo

Correspondence to: K. D. Williams
(keith.williams@metoffice.gov.uk)

1 Major comments from referee 2

1.1 Referee Comments

The paper mainly serves as an example of how model evaluation against available satellite and ground-based observations of cloud properties might be performed, including the use of techniques to account for uncertainties or biases in satellite retrievals (using simulators), techniques to isolate specific cloud or dynamical regimes (using compositing), and techniques to isolate the climatological or systematic biases in the model from short-timescale processes (using hindcasts). While this is a useful contribution, the paper leaves much to be desired in terms of physical interpretation, attribution, and discussion of identified biases, and instead focuses primarily on listing the identified biases.

Additionally, there is little (if any) discussion of uncertainties in the observational products used, or of the uncertainties in the comparisons between the model fields and those observations. In light of these shortcomings, I would recommend major revisions to the paper, in particular to dive somewhat deeper into identifying physical processes responsible for the identified biases in the model in terms of the model formulation

1.1.1 Author's response

We have included further detail regarding changes made to the parametrizations between GA6 and GA7 and have provided further evidence for attributing identified changes in cloud errors to particular model improvements. It should also be noted that it is intended that the present paper will be part of a GMD special issue which will also include the GA7 model description paper, so a complete description of the model will be readily available. We wish to avoid including speculation in the paper, however to the extent possible, in the revised manuscript we have attempted to link the errors to known model issues.

Observational uncertainties are complex. They depend on the details of the scene being observed (e.g. cloud size, height), illumination conditions, etc. Therefore, a full description of observational uncertainty is not possible within the scope of this paper. We have opted for

bringing in information on observational uncertainty when appropriate within the discussion of the results (see response to specific comments).

1.1.2 Manuscript changes

Section 2a has been considerably expanded with a more thorough description of the relevant parametrization changes. Within section 3, where possible, the text attributing changes in errors to particular parametrization changes has been expanded (e.g. around the warm rain microphysics discussion) to discuss how the parametrization differences lead to the improvement and the physical processes operating (see answers to specific comments). We have also added text to draw together the results from different diagnostic techniques to provide greater process understanding of the errors (e.g. around the mid-latitude cyclone RSW error).

The results of two new simulations have been added to Figure 2 in order to clearly attribute the differences seen to particular parametrization changes.

In the revised manuscript, we provide greater discussion of the uncertainties in the observational products where they are relevant to the paper (e.g. the differences between GOCCP and the CALIPSO cloud retrieval used by RL-GEOPROF). The revised manuscript also includes estimates of significance associated with sampling error to the figures.

2 Specific comments from referee 2

2.1 Comment

140: A definition of low, mid, and high cloud categories should be provided here (i.e., what are the altitude bounds for each category?). A short description of how these histograms are produced would also be useful to the reader here, in addition to providing the reference provided (i.e., cloud occurrence in each category is defined as that which exceeds a minimum backscatter ratio of ??).

2.1.1 Response & manuscript change

Definitions have been added to the manuscript as low:>680hPa, mid:440hPa–680hPa, high:<440hPa along with a description of the histogram as the referee suggests.

2.2 Comment

153: A brief explanation of the approach for each simulator would be helpful here (i.e., the ISCCP simulator emulates the way the retrieval infers cloud top pressure by estimating brightness temperature...).

2.2.1 Response & manuscript change

A brief description of each simulator has been added as the referee suggests.

2.3 Comment

156-165: The addition of this diagnostic the combines the CALIPSO and CloudSat hydrometeor occurrence is fantastic, but this description and discussion of the implementation is not nearly sufficient. A much more thorough description of the algorithm should be provided. The rationale for the choice of thresholds used seems somewhat incomplete as well, and it would be nice to see the comparison between GOCCP and RL-GEOPROF referred to on line 159. On line 160 it is suggested that the cloud detection algorithms differ between that used in COSP and that in RL-GEOPROF, but the nature of this difference is not explicitly stated and probably should be. Overall, some discussion of the uncertainties and sensitivities to the formulation of this new diagnostic should probably be provided to justify its use in the model evaluation. This could potentially be a significant contribution of this paper.

2.3.1 Response & manuscript change

As the referee requests, this paragraph has been completely re-written and expanded to provide a more detailed description of the diagnostic along with justification of the choices made.

2.4 Comment

212-215: This is a nice result, and it would be worth expanding on the cause for the difference in cirrus between GA6 and GA7. In particular, some justification for the claim that the largest difference is due to the reduction in the rate of cirrus spreading could be shown, such as a figure showing the cirrus amount in GA7 with and without the adjusted cirrus spreading parameterization. I do not think the formulation of the cirrus spreading parameterization, or the changes made to improve the simulation, have been documented well enough in the manuscript. This result showing the decrease in cirrus and better agreement with both CALIPSO and CloudSat is a nice validation of the improvement in the simulation due to these changes, and would go nicely with a more thorough explanation of what is going on here.

2.4.1 Response & manuscript change

Figure 2b now has two additional simulations added to it, one of which is GA6 but with the cirrus spreading rate reduced to the GA7 value to demonstrate the impact as the referee suggests. Discussion of this is expended where the figure is referred to in section 3 and the description of the cirrus spreading change in section 2a has been expanded to include the origin of the parametrization, how it is working and the justification for reducing this parameter.

2.5 Comment

221-222: How do we know that the revised numerics are responsible for the improvement in GA7? What specifically changed in the formulation of the model?

2.5.1 Response & manuscript change

Figure 2b now has two additional simulations added to it, one of which is GA6 but using the 6A convection scheme (revised numerics) to demonstrate that this is responsible for the increase in altitude of the cirrus. The description of the 6A convection scheme has been considerably expanded in section 2a with a list of the changes made to the formulation, however the increase is cirrus height is very much an outcome - it's not clear why these changes have this effect (other than the numerics are more accurate).

2.6 Comment

230: How is the "grid-box cloud fraction" being calculated? I am somewhat confused as to how this is produced alongside the profiles of reflectivity shown in the top panel. Is cloud fraction simply being aggregated onto a coarser grid from the reflectivity, calculated as the fraction within the coarser bins above some reflectivity threshold?

2.6.1 Response & manuscript change

Yes, the combined radar–lidar product has considerably higher along track resolution (nominally 1.7km) than the model (80km at the equator), hence regridding the combined radar-lidar data onto the model grid gives an observed cloud fraction to a precision of about 2%. This has been made clear in the revised manuscript.

2.7 Comment

232-236: What does this imply about the model formulation (the cloud parameterizations)?

2.7.1 Response & manuscript change

The following has been added to the manuscript "This is likely due to too little condensate being detrained at these altitudes, with what there is being either the result of convection going

slightly deeper on occasional timesteps or, more likely, some of the condensate being advected vertically having been detrained below."

2.8 Comment

242: Add a note here that the drizzle rates cited are not shown here.

2.8.1 Response & manuscript change

Added in the revised manuscript as the referee suggests.

2.9 Comment

247-250: This is a nice demonstration of the impact of the new microphysics package, but this is lacking a discussion of the mechanisms for the improvement, and should be accompanied by a description of the changes.

2.9.1 Response & manuscript change

The description of the warm rain microphysics scheme has been expanded in section 2a. We have also added the following in section 3 where the attribution of the change to the warm rain microphysics package is discussed "Within this package, the change to use the Khairoutdinov and Kogan (2000) scheme reduces auto-conversion rates by a factor of around 100 compared with the scheme used in GA6. These rates would be too low without the Boutle et al. (2014) GCM upscaling, however even after this correction, the auto-conversion rates remain around 10 times small than GA6 which accounts for the removal of the spurious drizzle."

2.10 Comment

258: Could the increase in cirrus here be explained by excessive advection of the cirrus outflow, or again maybe something to do with the cirrus spreading parameterization referred to earlier? What is responsible for the improvement in GA7?

2.10.1 Response & manuscript change

The improvement in GA7 is due to the cirrus spreading change and this has now been added to the manuscript. The upper tropospheric wind errors are not large enough for the bias to be attributable to excessive advection, hence we retain the suggestion in the text as "possibly due to errors in microphysical processes, or macrophysical fields (such as relative humidity being too high)."

2.11 Comment

261-270: This discussion does not contain much substance, and inclusion of the ISCCP comparison seems to almost be an afterthought. This either needs a more complete treatment of the sources of differences, or consider cutting from the manuscript to make room for some of the more fleshed out analysis, such as the discussion of improvements in thin cirrus.

2.11.1 Response & manuscript change

As we describe in the text, accurate simulation of cloud in this region is believed to be particularly important in determining the global cloud feedback under climate change. For this reason, the excellent simulation of stratocumulus amount is worth showing, however it hasn't changed much between the two configurations shown, hence the brevity of the paragraph. We have expanded the discussion around the ISCCP comparison since it highlights one of the key outstanding errors which remain, namely that in many regions low cloud remains too reflective. We have added "Consistent with this, comparison against a number of observational datasets indicates that the cloud effective radius simulated by the model is too low in many regions, including in subtropical stratocumulus, and is indicative of the aerosol cloud indirect effect being too strong."

2.12 Comment

278-279: This statement could use evidence or a citation to back it up.

2.12.1 Response & manuscript change

This was based on personal experience. Whilst we believe it correct, we do not have a reference and have therefore removed the statement from the revised manuscript.

2.13 Comment

281-286: This could be better tied in with the discussion of cirrus above. In general though the results from this figure are not very compelling and do not seem to add much to the discussion. It is also not clear to me from Figure 5 that cirrus is overestimated in GA6. The most apparent biases in this figure are the altitude bias in the location of the cirrus maximum in GA6, and an overall underestimation of cirrus in GA7.

2.13.1 Response & manuscript change

The manuscript has been re-worded to link back to the discussion of the tropics as a whole. We now highlight the cirrus height increase and refer to the cirrus amount as a change rather than a universal improvement. This variance in whether the change is an improvement or detrement across the regimes highlights the importance of this figure in providing information over what was in the tropical mean analysis in Figure 2 - a point which has been added to the manuscript.

2.14 Comment

287-290: These conclusions are difficult to draw from Figure 5 as shown due to the scales of the axes used. If boundary layer cloud is the focus of this figure, it would be better to show just the boundary layer for the lower panel (SST composites), and on a cloud fraction scale that allows the reader to actually see the differences between the different curves.

2.14.1 Response & manuscript change

As the referee suggests, the lower panel of figure 5 has been re-drawn to just show the lowest few km and the cloud fraction scale adjusted to make it easier to view the differences.

2.15 Comment

340: I realize this is explained in the cited manuscript, but at least a simple explanation of the equation tested should be given here.

2.15.1 Response & manuscript change

The present manuscript has been revised to indicate that the change made to the equation when testing for anticyclones is identifying a local maxima in surface pressure rather than a local minima.

2.16 Comment

352: ”Reasonably good” is awkward language to use here. I would suggest replacing with something like ”while the cloud simulation was in reasonable agreement with observations”.

2.16.1 Response & manuscript change

Sentence changed in the revised manuscript as reviewer suggests.

2.17 Comment

356: Again ”reasonably good” is awkward here.

2.17.1 Response & manuscript change

Sentence changed to "Despite the cloud amount composites showing cloud fraction errors of less than 0.15 (and often less than 0.05) in GA7...".

2.18 Comment

368-369: Elaborate on how these biases are consistent with the radiation errors.

2.18.1 Response & manuscript change

Description expanded in revised manuscript to highlight that in regions of positive albedo bias in Figure 10, there is a positive RSW bias in Figure 9 and vice-versa. However, the in-cloud albedos in Figure 10 do not depend on the insolation hence for the same cloud albedo error, the RSW error will be larger in the summer than winter.

2.19 Comment

385-389: This is an excellent example of the utility of using multiple observations in the evaluation strategy. This would be a good point to emphasize, and perhaps use as a jumping off point for a more elaborate investigation of the source of these differences (multi-layered cloud vs excess precipitation) than is given in the sentences to follow.

2.19.1 Response & manuscript change

We have highlighted that this is a good example of the utility of using multiple instruments. We have also expanded the discussion explaining why we can't rule out either the shielded low cloud or precipitation options at this stage (our suspicion is that both may contribute).

2.20 Comment

401: Why is SYNOP data the most reliable here?

2.20.1 Response & manuscript change

Sentence expanded in the revised manuscript to discuss the problems of viewing the lowest levels from space and that an upward pointing ceilometer or human observer is likely to be at their most accurate for low cloud bases.

2.21 Comment

403-405: Need evidence or references to back this up.

2.21.1 Response & manuscript change

Reference to Mittermaier (2012) added.

2.22 Comment

410: How is an okta defined in the context of the model?

2.22.1 Response & manuscript change

This is simply a cloud fraction of 1/8 (0.125). This has been added to the revised manuscript.

2.23 Comment

439: What caused the reduction in the cold bias in GA7?

2.23.1 Response & manuscript change

This was mainly due to the introduction of the 6A convection scheme. This has been added to the manuscript, however it is beyond the scope of this paper to discuss these non-cloud related impacts of the model changes and instead a reference given to Walters et al. (2017) who discuss this further.

2.24 Comment

447-450: I am not sure I entirely agree with these conclusions. The reflected shortwave biases around the subtropical cumulus transitions seem to have reversed in sign between HadGEM2 and GA7, but the magnitudes do not seem to be universally reduced. Perhaps I am looking at the wrong part of the figure though, so maybe a box or symbol on the figure indicating the region where the improvement is evident would be appropriate. The underestimate in reflective shortwave over the Southern Ocean also does not appear to be significantly reduced.

2.24.1 Response & manuscript change

The sentence has be revised to read "The error in the sub-tropical cumulus transition regions of excess RSW has been removed and there is now a smaller negative bias in GA7. The lack of RSW over the Southern Ocean has been reduced by a third and...". We have also reproduced Figure 13 with a revised colour bar to make it easier to quantify the changes e.g. that the negative bias in the transition region in GA7 is smaller in magnitude compared with the positive bias in HadGEM2-A.

2.25 Comment

482-485: This seems to really be a key point of the paper: to demonstrate that the multi-diagnostic approach used reduces the possibility of drawing the wrong conclusions. This is hinted to at points in the paper, but I think this could be drawn together a little better here, perhaps by recounting the points in the preceding analysis that illustrate this (such as the contrast in the comparisons between CloudSat and CALIPSO that demonstrate errors due specifically to thin cirrus, or to excess precipitation as opposed to cloud errors).

2.25.1 Response & manuscript change

The discussion has been expanded here using a number of examples including the ones the referee suggests.

A multi-diagnostic approach to cloud evaluation

Keith D. Williams* and Alejandro Bodas-Salcedo

Met Office, Exeter, UK

*Corresponding author address: Keith Williams, Met Office, FitzRoy Road, Exeter, EX1 3PB, UK.

Email: keith.williams@metoffice.gov.uk Tel: +44 (0)1392 886905 Fax: +44 (0)1392 885681

March 31, 2017

Abstract

Most studies evaluating cloud in general circulation models present new diagnostic techniques or observational datasets, or apply a limited set of existing diagnostics to a number of models. In this study, we use a range of diagnostic techniques and observational datasets to provide a thorough evaluation of cloud, such as might be carried out during a model development process. The methodology is illustrated by analysing two configurations of the Met Office Unified Model - the currently operational configuration at the time of undertaking the study (Global Atmosphere 6, GA6), and the configuration which will underpin the United Kingdom's Earth System Model for CMIP6 (Coupled Model Intercomparison Project 6) (GA7).

By undertaking a more comprehensive analysis which includes compositing techniques, comparing against a set of quite different observational instruments and evaluating the model across a range of timescales, the risks of drawing the wrong conclusions due to compensating model errors are minimised and a more accurate overall picture of model performance can be drawn.

Overall the two configurations analysed perform well, especially in terms of cloud amount. GA6 has excessive thin cirrus which is removed in GA7. The primary remaining errors in both configurations are the in-cloud albedos which are too high in most northern hemisphere cloud types and sub-tropical stratocumulus, whilst the stratocumulus on the cold air side of southern hemisphere cyclones has in-cloud albedo's which are too low.

1 Introduction

The accurate simulation of cloud in general circulation models (GCMs) is of considerable importance across all timescales. At numerical weather prediction (NWP) timescales of a few days or less, cloud amount as a forecast product is of direct relevance to a number of users (e.g. aviation, solar farms, etc.) and affects forecasts of other variables through its radiative impact on the surface temperature and the effects of diabatic heating on the large scale circulation. On climate timescales, the radiative feedback from cloud on the global energy budget remains one of the largest uncertainties in determining the global climate sensitivity (Flato *et al*, 2013).

Traditionally, the evaluation of cloud has been limited to quantities which were perceived to be of interest to the end user such as ground-based observations of total cloud amount (Mittermaier, 2012), or top-of-atmosphere cloud radiative forcing (CRF) (e.g. Gleckler *et al*, 2008). However, compensating errors within GCMs can result in a model performing well on such a limited set of metrics, despite the processes within the model being in error. A classic example is the simulation of subtropical stratocumulus, for which many GCMs simulate too little cloud cover, but the cloud which is simulated is too bright, the two errors compensating to result in a reasonable CRF (e.g. Williams *et al*, 2003; Nam *et al*, 2012).

Over recent years, a range of process-orientated diagnostic techniques have been developed which composite the data according to other large-scale variables, with the intention of reducing the chances of a model appearing to perform well due to compensating errors. Compositing variables have, amongst others, included: large scale vertical velocity, (Bony *et al*, 2004); various measures of lower tropospheric stability, (Klein and Hartmann, 1993; Williams *et al*, 2006; Myers and Norris, 2015); position relative to cyclone centre, (Klein

and Jakob, 1999; Govekar *et al*, 2011) and cloud regime (Williams and Tselioudis, 2007).

In addition to model errors, there are errors in the observational datasets and how they are used for GCM evaluation. For example, the 'total cloud amount' obtained from ground-based ceilometers will be underestimated since they typically cannot detect the highest clouds. When these issues are known, they can be mitigated by sampling the model in a consistent manner to the observations (e.g. in this case, only considering model clouds up to the maximum height the ceilometer can detect). For cloud evaluation against satellite data, increasing use is being made of satellite simulators which aim to emulate the observations by carrying out a consistent retrieval on the model. A number of satellite simulators have been brought together in the CFMIP (Cloud Feedback Model Intercomparison Project) Observational Simulator Package (COSP; Bodas-Salcedo *et al*, 2011) which has now been included in many GCMs.

Arguably the best way to minimise issues around compensating model errors, observational error and model-observation comparison issues, is to routinely evaluate cloud in GCMs against a wide range of different observational datasets, using simulators where appropriate and using a range of diagnostic techniques in order to gain a consistent picture of model biases. In this study, we illustrate ~~the approach~~ how the approach can be used for model development by applying a comprehensive cloud evaluation to two configurations of the Met Office Unified Model (UM).

Cloud errors in the UM, possibly more than any other variable, are very similar across timescales and horizontal resolutions (Williams and Brooks, 2008). Figure 1 shows the bias in high, mid and low cloud in the Global Atmosphere 6 (GA6; Walters *et al*, 2016) configuration of the UM against CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation). It can be seen that the day 1 and day 5 forecast biases at N320

resolution (40km in mid-latitudes) are very similar to each other and to a climatological bias obtained from an AMIP (Atmosphere Model Intercomparison ~~Prroject~~Project; Gates, 1992) simulation at N96 resolution (135km in mid-latitudes). This means that we can make use of each timescale in our analysis to its strengths and the conclusions should be applicable across the systems. Although the UM is being used (a model which is routinely assessed for both NWP and climate work), we consider the cross-timescale approach a key aspect of the comprehensive evaluation. The initialised hindcasts provide case studies where model biases can be investigated in detail for particular meteorological events, in situations where the large scale dynamics remain close to those observed. In contrast, the longer climate simulations provide characterisation and statistics of the systematic errors. For those GCMs which are typically only used for a limited set of timescales, the AMIP (Gates, 1992) and Transpose-AMIP (Williams *et al*, 2013) experimental designs allow the possibility of this cross-timescale evaluation.

In the next section we provide details of the models, experiments and observational data subsequently presented. We then evaluate the cloud simulation in the model over the tropics, mid-latitude storm tracks and mid-latitude land in sections 3, 4 & 5 respectively. The overall impact of the cloud on the global radiation balance is then discussed in section 6. We summarise in Section 7.

2 Models and observational datasets

a Models and experimental design

Two configurations of the UM are used in this study. GA6 has been operational in all global model systems at the Met Office since 15th July 2014 and is fully docu-

mented by Walters *et al* (2016). GA7 has recently been frozen and is documented by ~~Walters *et al* (2017)~~Walters *et al* (2017). It is intended that GA7 will form the physical atmosphere model used by the United Kingdom Earth System Model 1 (UKESM1) which will be submitted to CMIP6 (Coupled Model Intercomparison Project 6).

There are numerous physical parametrization changes between GA6 and GA7 which are detailed in ~~Walters *et al* (2017)~~Walters *et al* (2017). Those of most relevance for this study are:

1. The introduction of a scheme to allow the turbulent fluxes within the boundary layer capping inversion to be resolved and for clouds ('forced cumulus') to form within it. The height of the top of the capping inversion is diagnosed using an energetic argument based on ?Beare (2008) which is applied to the convection diagnosis parcel. Within the undulations of the capping inversion, if the parcel doesn't reach it's level of free convection then forced cumulus may form. A cloud fraction profile is parametrized from the (Zhang and Klein, 2013) ~~.—~~data and inhomogeneously forced into the cloud fraction profile between the lifting condensation level and inversion top. The in-cloud water content is taken from the adiabatic parcel ascent in the cumulus diagnosis.

2. A package of changes designed to improve warm rain microphysics~~, which include~~ . This includes a change to the auto-conversion scheme to be based on Khairoutdinov and Kogan (2000) ~~, but~~ which was developed from a bin resolved microphysics scheme, and so closely correspond to best estimates of what these process rates should be. They are upscaled to a GCM following Boutle *et al* (2014). Because microphysical process rates are nonlinear, calculating the process rate from in-cloud mean quantities (as is done in GA6) can lead to large biases in the process rate in low

resolution GCMs where the sub-grid variability is significant. This parametrization corrects the process rates for the presence of sub-grid variability, based on parametrizations of the sub-grid variability derived from aircraft, CloudSat (Stephens *et al*, 2002) and CloudNet-ARM (Atmosphere Radiation Measurement) site observations.

3. Improved cloud ice optical properties and ice particle size distributions (PSD) following Baran *et al* (2014) and Field *et al* (2007) respectively. The new PSD is an empirical fit that is better supported by observations and in GA7 is used consistently between the microphysics and radiation schemes.

4. Reduced rate of cirrus spreading by two orders of magnitude. The cirrus spreading was a simple parametrization intended to account for the spreading of cirrus through shear as it falls~~, however it was included~~. It uses the model wind shear between successive layers to spread the ice as it falls at a rate controlled by a tunable parameter. It was included, largely as a tuning of outgoing longwave radiation (OLR), in an earlier configuration (GA4; Walters *et al*, 2014) and it is desirable to reduce the effect until the scheme is developed on firmer physical grounds.

5. Addition of the turbulent production of liquid water in mixed-phase clouds following Field *et al* (2014). An exactly soluble stochastic model is used to describe sub-grid relative humidity fluctuations. The probability density function (PDF) of the fluctuations in a model grid-box depends on the turbulent local state based on the boundary layer turbulent kinetic energy and on any pre-existing ice cloud. Increments to liquid water cloud prognostic fields are diagnosed from the PDF. This increases the liquid water contents and volume fractions of liquid cloud. A temperature threshold restricts the scheme to regions below 0 Celsius.

6. A change to the aerosol scheme from CLASSIC (Coupled Large-Scale Aerosol Simulator for Studies In Climate; (Bellouin *et al*, 2011)) to GLOMAP-mode (Global Model of Aerosol Processes modal aerosol scheme; (Mann *et al*, 2010)). GLOMAP-mode models the aerosol number, size distribution, composition and optical properties from a detailed, physically-based treatment of aerosol microphysics and chemistry. The scheme simulates speciated aerosol mass and number in 4 variable-size soluble modes to cover different aerosol size ranges (nucleation, aitken, accumulation and coarse modes) as well as an insoluble aitken mode. The prognostic aerosol species represented by GLOMAP-mode are sulphate, black carbon, organic carbon and sea salt. Cloud condensation nuclei are activated into cloud droplets using the Activate aerosol activation scheme based on Abdul-Razzak and Ghan (2000).

7. Although only small changes have been made to the scientific basis of the convection scheme, the numerics of the scheme have been re-written (the so called '6A convection scheme'). This is described in Walters *et al* (2017), but the key points are:

   - Three iterations rather than one iteration is used to solve the implicit equations for the potential temperature of the detrained mass and the residual plume in the calculation of the forced detrainment.
   - Three rather than two iterations are used in determining the potential temperature at saturation after lifting the the parcel from one level to the next under dry ascent. The evaporation of parcel condensate is now also allowed if the parcel becomes sub-saturated after entrainment and the dry ascent.
   - The ascent in the 6A scheme will terminate when the mass flux falls below 5% of its value at cloud base, which replaces the previous arbitrary small value.

- The convection scheme will introduce small errors in the conservation of energy and water. These are now corrected locally to ensure that the column integral of these quantities is the same after the call to convection as they were before, replacing the previous global correction.

For each configuration, two types of experiment have been conducted, both being standard tests used within the model development cycle for proposed changes to the UM. These are a 20 year (1988-2007) AMIP experiment run at a horizontal resolution of N96 (135km in mid-latitude), and a set of 24 independent 5-day NWP hindcasts spread between December 2010 and August 2012, run at N320 (40km in mid-latitude) and initialised from European Centre for Medium range Weather Forecasts (ECMWF) analyses. ECMWF rather than Met Office analyses are used for case study tests within the model development cycle so as not to favour the performance of the control model which may have had the UM data assimilation system tuned towards it. This also makes the hindcasts consistent with the standard Transpose-AMIP experiment (Williams *et al*, 2013), except for the specific dates run.

b Observational datasets and simulators

We make use of a variety of observational datasets. The International Satellite Cloud Climatology Project (ISCCP) D1 product (Rossow and Schiffer, 1999) uses passive radiometer data from geostationary and polar orbiting satellites to produce 3-hourly histograms of cloud fraction on a 2.5^o grid in seven cloud top pressure and six optical depth bins. CALIOP (Cloud-Aerosol Lidar with Orthogonal Polarization) is a cloud lidar on the CALIPSO platform (Winker *et al*, 2010), which is part of the NASA A-train satellite constellation. It uses a nadir pointing instrument with a beam diameter of 70m at

the earth's surface and produces footprints every 333m in the along-track direction. We use the GCM-orientated CALIPSO cloud product (Chepfer *et al*, 2010) which contains histograms of cloud amount in joint height–backscatter ratio bins and well as total cloud amount in standard low ~~, mid and high categories.~~ (>680hPa), mid (440hPa–680hPa) and high (<440hPa) categories. The histograms are formed by assigning the cloud occurrence in each height and backscattering ration category with a minimum backscattering ratio of 3. The percentage occurrence in each bin is then determined. CloudSat (Stephens *et al*, 2002), also on the A-train, is a 94GHz cloud radar which pulses a sample volume of 480m in the vertical and a cross-track resolution of 1.4km. We use the CloudSat 2B geometrical profile (2B-GEOPROF) (Marchand *et al*, 2008) product which includes histograms of hydrometeor frequency in joint height–radar reflectivity bins. The complementary nature of the CloudSat and CALIPSO in terms of the ~~3D structure~~ hydrometeor profile provided by the radar and detection of very thin clouds by the lidar, and their co-location on the A-train mean that they may be combined to produce a 'best estimate' ~~3D hydrometeor fraction~~ hydrometeor fraction through the depth of the atmosphere column. This has been done by Mace and Zhang (2014) in the form of the radar-lidar geometrical profile (RL-GEOPROF) product. In this study we use revision 4 (R04) of RL-GEOPROF.

All of the above have a simulator within COSP (Bodas-Salcedo *et al*, 2011) in order to produce comparable diagnostics from the model by emulating the satellite retrieval. The simulators are described by Klein and Jakob (1999)/Webb *et al* (2001), Chepfer *et al* (2008), Haynes *et al* (2007) for the ISCCP, CALIPSO and CloudSat simulators respectively. The ISCCP simulator uses a perfect optical depth retrieval, taking into account the subgrid variability of cloud condensate used in the model's radiative transfer model. The cloud top pressure is based on a simple estimation of the 10.5 micron

brightness temperature, which is then mapped onto the temperature profile as a function of pressure. The CALIPSO and CloudSat simulators are forward models of the attenuated backscattering ratio at 532nm, and reflectivity at 94GHz, respectively.

COSP version 1.4 is used in this study, which does not include a diagnostic of combined radar-lidar cloud fraction. In order to compare model clouds against RL-GEOPROF, a new diagnostic that combines CALIPSO scattering ratio and CloudSat reflectivities has been developed. ~~For this diagnostic , the scattering ratio cloud detection threshold has been lowered to 3, from its standard value of 5. This value has been chosen because it brings the observational estimates of cloud fraction from GOCCPcloser to RL-GEOPROF in the midle and upper troposphere. Since RL-GEOPROF uses a cloud detection algorithm that differs from the one used in COSP there are effects that this diagnostic neglects, the main one being the effect of a~~ The new diagnostic is a simple combined cloud mask. Each volume in each sub-column is flagged as cloudy if the CALIPSO scattering ratio (SR) is above the detection threshold (SR$\geq$3.0) or the CloudSat reflectivity is greater than -30dBZ. Then the cloud fraction at each level is calculated as the ratio of cloudy volumes divided by the total number of volumes.

The cloud identification of the GCM Orientated CALIPSO Cloud Product (GOCCP) is performed at the nominal horizontal resolution (330m below 8km, and 1km above 8km). At that resolution, the instrument noise level is high. In order to minimise false positives due to noise, GOCCP uses a very conservative scattering ratio threshold (SR=5~~km averaging length from the lidar~~). The CALIPSO cloud mask used in the RL-GEOPROF product uses a 5km spatial averaging to increase the signal-to-noise ratio and allow the detection of thinner clouds. Chepfer *et al* (2013) show that the implicit SR detection threshold in the CALIPSO cloud mask used in ~~the~~ RL-GEOPROF ~~cloud~~

~~detection. We do not think this is a problem for our analysis since the differences in the RL-GEOPROF cloud fractions between 1km~~ ranges between 1 and ~~5km averaging lengths are 0.01 , substantially smaller than the model biases that we detect~~3. We have therefore reduced the SR threshold from 5 to 3 in COSP in order to represent a diagnostic that is more comparable to the RL-GEOPROF cloud mask. A value of 3 is chosen because it is one of the boundaries used by GOCCP to construct height-SR histograms. Supplementary material Figure 1 shows the impact of reducing the SR threshold in the vertical profile of cloud fraction over the tropical belt.

Evaluation of the top-of-atmosphere radiative fluxes are made against CERES-EBAF (Clouds and the Earth's Radiant Energy System–Energy Balanced and Filled) dataset (Loeb *et al*, 2009). We also make use of synoptic surface observation (SYNOP) data (WMO, 2008). Mittermaier (2012) discuss some of the issues around using these data for cloud verification. We consider the most significant for evaluation of model biases are the differences in the maximum altitude at which automated ceilometers used by different countries can detect cloud, which in turn differ from human observers. In this study we just use cloud base height information in situations where the cloud base is below 1km. It is in these situations that the SYNOP observations should be the most consistent and reliable.

Compositing techniques are employed to provide a more process-orientated cloud evaluation. In all cases, the data used to composite the observed cloud fields ($500hPa$ vertical velocity, pressure at mean sea level, etc.) are from ERA-I (ECMWF Interim Re-analyses; Dee *et al*, 2011). Composites using daily mean data are formed from 5 year datasets. Other multi-annual mean plots are formed from all of the complete years of data available for the observational datasets ~~and~~ (25 years for ISCCP, 12 years for CERES-EBAF and 5

years for CloudSat/CALIPSO) and 20 year means for the AMIP simulations. We perform a Student's t-test based inter-annual variability of the data available to determine the 5% significance of model–model and model–observational differences. These have been added to figures in the paper, however in general the inter-annual variability is small compared to the differences discussed.

3 Tropical cloud evaluation

Tropics-wide (20^oN-20^oS) multi-annual average frequency histograms for ISCCP, CALIPSO and CloudSat, together with the outputs from COSP for GA6 and GA7 AMIP experiments are shown in Figure 2a-c. Taking ISCCP first (Figure 2a), retrievals from passive instruments provide a cloud top view. Compared with the newer active instruments, the vertical resolution is poor and there are issues with the height assignment under certain conditions (Mace and Wrenn, 2013). Nevertheless, the optical depth information from ISCCP remains valuable for optical depths greater than approximately 1.0, hence an optical depth frequency profile is also shown. Both GA6 and GA7 tend to simulate too little cloud with intermediate optical thicknesses (1.0-10.0) and slightly too much optically thick cloud. Referring back to the full histograms, this bias appears to be the case for both high and low-top cloud.

Arguably, CALIPSO provides the best global picture of total 2D cloud cover since, unlike the other instruments considered here, it can detect thin sub-visual cirrus. The vertical resolution is good, hence in Figure 2b, as well as providing the full histograms, we collapse along the backscattering ratio axis to provide a vertical profile of cloud frequency. In doing this, for altitudes below 4km we only consider backscattering ratios greater than 5 due to the potential contamination from aerosols in the boundary layer, however

above 4km backscattering ratios as low as 3 are included so as to account for very thin cirrus. This choice of the vertical profile of backscattering ratio threshold also gives a profile which most closely matches the CALIPSO cloud detection product used within the RL-GEOPROF dataset (Supplementary material Figure 1). The lidar does become attenuated in the presence of thick ice cloud, and is attenuated quickly in the presence of liquid cloud, hence this profile remains largely a cloud-top view.

Although the CloudSat radar is not sensitive to sub-visual cirrus, it uniquely provides a full 3 dimensional view of the cloud, only becoming attenuated in moderate and heavy rain. Despite the name, it should be noted that CloudSat is sensitive to precipitation as well as cloud. As for CALIPSO, in Figure 2c we provide a vertical profile of hydrometeor frequency in addition to the full height–radar reflectivity histograms from CloudSat.

Comparing the models with CALIPSO and CloudSat (Figure 2b&c), GA6 clearly has excess amounts of cirrus and this is corrected in GA7. A number of physical improvements included in GA7 have changed the amount of cirrus including the new ice particle size distribution and revised ice optics, however the largest decrease in cirrus has come from the reduction in the rate of cirrus spreading associated with wind shear as the ice falls between successive model levels. This is clear from the orange line on the profile plot of Figure 2b which is a simulation identical to GA6 (the blue line) but with the cirrus spreading reduced to the value used in GA7. The altitude of the cirrus is also too low compared with CALIPSO, but this bias doesn't appear to exist when comparing with CloudSat, which indicates that the issue is associated with very thin cirrus. The CALIPSO histograms indicate that as the cloud thins to the lowest backscattering ratios, the altitude of the cloud should increase, however this does not appear to be the case in GA6. In GA7 the altitude–backscatter ratio relationship is improved such that the highest cloud has

the lowest backscattering ratios. This ~~improvement~~ slight increase in the altitude of the cirrus is the result of the revised numerics of the convection scheme~~, but the~~. This can be seen from the cyan line in Figure 2b which is a simulation identical to GA6 (the blue line) but with the convection using the 6A scheme (revised numerics). Despite this slight increase in height, the overall altitude of the thin cirrus ~~remains too low.~~ still remains below that observed by CALIPSO.

The low altitude cirrus bias can be examined in more detail in a case study using a short-range hindcast (Figure 3). In this example (which is typical of other convective cases examined), the A-train overflew a convective system over the South China Sea. The top panels of Figure 3 show the observed and GA6 simulated radar reflectivities. Data from CALIPSO have been added in locations where the lidar was detecting cloud which was not detected by the radar. It can be seen that the model is able to simulate thin cloud in the upper levels of the convective system right up to the observed altitudes of around 16km. ~~However, if we compare the observed~~ The nominal along-track resolution of the RL-GEOPROF product is 1.7km, so if a threshold of -40dBZ is used for cloud identification and it is regridded onto the model grid, which is 80km near the equator, then an observed cloud fraction over a model grid-box can be estimated. This assumes that the along-track cloud fraction is representative of the 2D grid box. Whilst this is a fair assumption when considering a large number of cases which the A-train will cross at random orientations, there may be an error when considering a single case such as this. The observed and simulated grid-box cloud fraction on the model grid ~~(~~are shown in the lower panels of Figure 3~~), large~~. Large cloud fractions occur up to the top of the convective system in the observations, whereas they reduce quickly above 14km in the model. So it appears that the lack of the highest thin cirrus is primarily because the

fractional coverage of grid-boxes is too small in situations where some cloud is present, rather than there being too many completely clear grid boxes at these altitudes. This is likely due to too little condensate being detrained at these altitudes, with what there is being either the result of convection going slightly deeper on occasional timesteps or, more likely, some of the condensate being advected vertically having been detrained below.

Moving down in altitude, Figure 2b suggests the models have too little mid and low top cloud in GA6, whereas Figure 2c may be interpreted as GA6 having considerably too much. However, the excess hydrometeor frequency at lower levels in GA6 is entirely due to excess drizzle in the model rather than cloud. This can be demonstrated by re-running GA6 but not passing the large scale precipitation field to the CloudSat simulator (cyan line in Figure 2d). In this case the excess hydrometeor fraction is completely removed. Examining these drizzle rates in the model, they are very low (typically $<0.005mm/hr$, not shown), possibly explaining why this model defect had not been spotted before, and again showing the benefit of carrying out evaluation against multiple datasets. This anomalous drizzle is corrected in GA7 to leave the hydrometeor fraction slightly too small at low levels (Figure 2c), which is believed to by mainly due to a lack of heavy convective rain (region of the histogram with radar reflectivities >0). The improvement in drizzle in GA7 is entirely due to the warm rain microphysics package, which can be demonstrated if GA6 is run again (all fields passed to the simulator) with just the GA7 change to the warm rain microphysics applied (Figure 2d). Within this package, the change to use the Khairoutdinov and Kogan (2000) scheme reduces auto-conversion rates by a factor of around 100 compared with the scheme in GA6. These rates would be too low without the Boutle *et al* (2014) GCM upscaling, however even after this correction, the auto-conversion rates remain around 10 times small than GA6 which accounts for the

removal of the spurious drizzle.

Figure 3 shows, under the cirrus shield in GA6, an extensive region of high hydrometeor fraction and reflectivities of the order -10dBZ, between the surface and 7km which is absent in the observed transect. This is consistent with the region of the histogram in Figure 2c where there is spurious large-scale rain. It is likely that large scale cloud is forming in the moist air around the convective system and that it is undergoing auto-conversion, showing up as a strong signal in the CloudSat simulator. In GA7 (not shown) this precipitation signal is removed with just a cloud signal remaining at around -40dBZ. It should be noted that this is in a region largely attenuated for CALIPSO as it is below the cirrus shield and so doesn't contribute to the 'missing' mid-top cloud which is believed to more cumulus congestus rather than large-scale.

Tropical low cloud can be more easily assessed if regions are examined in which deep convection is rare/non-existent. Considering a region of the tropical Pacific dominated by trade cumulus and comparing with CALIPSO (Figure 4), GA6 appears to have too little cloud. The forced shallow cumulus scheme improves the amount of shallow cumulus at heights of around 1km, although there looks to be a secondary peak in low cloud around 2km which is absent in both configurations of the model. The region does receive some thin cirrus outflow from nearby deep convective regions, however the amounts are far too large in the model. This indicates that the cirrus lifetime is too great, possibly due to errors in microphysical processes, or macrophysical fields (such as relative humidity being too high). Although improved in GA7 due to the reduced cirrus spreading rate, the excess cirrus in this region remains. Chepfer *et al* (2013) show that the averaging effect is sensitive to the length of the averaging and is higher for low-level, small-scale broken cloud. For high clouds, the differences between GOCCP and the CALIPSO cloud

retrieval used by RL-GEOPROF are dominated by the SR detection threshold. The height-dependent SR detection threshold used in this study increases the sensitivity to high clouds (supplementary Figure 1). For cirrus clouds in the regions shown in Figure 4, the bias introduced by lack of averaging smaller than 0.05 (Figure 10 in Chepfer *et al*, 2013), supporting the interpretation that the cirrus amounts simulated by the models are excessive in this region.

Over the past couple of decades, a key focus of model development in the UM in relation to clouds has been on improving the simulation of subtropical stratocumulus due to its importance in determining the global cloud feedback under climate change (e.g. Bony and Dufresne, 2005). Many models have too little cloud in this region, with what there is being too bright (Nam *et al*, 2012). A number of improvements in previous configurations have resulted in the cloud amounts being in very good agreement with CALIPSO (Figure 4), although the low cloud amounts are reduced slightly in GA7 as a result of the change in the aerosol scheme to GLOMAP-mode. Compared with ISCCP, GA7 has considerably too little moderately reflective cloud in this region, but slightly too much optically thick cloud indicating that what cloud there is remains too reflective. Consistent with this, comparison against a number of observational datasets indicates that the cloud effective radius simulated by the model is too low in many regions, including subtropical stratocumulus (not shown), and is indicative of the aerosol cloud indirect effect being too strong (Walters *et al*, 2017).

Compositing cloud data by large scale variables is a useful way of summarising the tropical cloud structures across different meteorological situations. The most common are to composite against $500hPa$ vertical velocity (Bony *et al*, 2004) and a measure of lower tropospheric stability. A number of measures for the latter have been proposed (e.g.

Klein and Hartmann, 1993; Williams *et al*, 2006; Wood and Bretherton, 2006), however ~~because~~ here we simply use the spacial variation in sea surface temperature (SST) ~~is greater than the free tropospheric temperature, and the ocean provides an unlimited moisture source for humidity, the variation of boundary layer cloud with SST provides many of the features seen with more complex measures of lower tropospheric stability.Here we~~ (e.g. Williams *et al*, 2003). We composite the observed and modelled CALIPSO cloud profile by daily $500hPa$ vertical velocity (ω_{500}) and SST (Figure 5). The ~~excess cirrus in GA6 appears to be a problem across the different~~ desirable increase in altitude of the cirrus, discussed above for the tropics as a whole, can be seen in all the large scale vertical velocity regimes. The reduced cirrus amount in Figure 2b is also reflected in Figure 5, with the ~~bias being largest~~ largest reduction in regions of strongest ascent~~,~~ ~~but still present in strong subsidence. GA7 is a clear improvement, although there 's now possibly~~ . However there now appears to be too little cirrus in weakly ascending regimes ~~. The lack of mid-level cloud (with tops between 4 and 8km), is a bias in both models in regions of large scale ascent.~~ in GA7. This separation by regime therefore gives useful insights on where there might be compensating errors in the tropical mean picture provided by Figure 2.

The SST composites appear to better separate the stratocumulus regions at the coldest end as these bins clearly show higher fractions of boundary layer cloud. There is slightly too little low cloud in a number of the SST and ω_{500} composite bins, whilst there looks to be too much stratocumulus in the coldest SST bin. However in general, low-top cloud amounts appear to be reasonably well simulated.

4 Cloud evaluation in the mid-latitude storm tracks

The weather over the mid-latitude oceans is characterised by the passage of synoptic systems. Since the cloud structures change on a daily basis, compositing of climatological data is essential. Here we follow Govekar *et al* (2011) to analyse RL-GEOPROF cloud data around a composite cyclone, using the cyclone compositing technique of Field and Wood (2007). Cyclone centres are identified from daily ERA-I PMSL (pressure at mean sea level) data over the northern hemisphere oceans (35^oN-70^o) and the RL-GEOPROF data extracted for a 30^o latitude by 60^o longitude box centred on the cyclone. All the cyclones from 5 years worth of daily December-January-February (DJF) data are then averaged to form a composite cyclone. In order to visualise the composite, Figure 6 shows several sections through the 3 dimensional composite. The top panels are horizontal sections in the boundary layer (1.7km) and upper troposphere (6km) with the mean PMSL contoured. The positions of frontal features will vary with time and between systems, and the size of cyclones varies which also smooths the composite, but on average it would be expected that fronts would occupy the south-east quadrant with a cloud head wrapping around the north of the cyclone (Field and Wood, 2007). This can be seen as higher cloud fractions in these locations in the section at 6km, whilst the boundary layer hydrometeor fraction appears more symmetrical around the cyclone with a maximum near the centre. The lower panels on Figure 6 are vertical sections across the composite to the south and to the east of the centre, with the contours indicating the average vertical velocity from ERA-I (dashed indicates ascent). The east-west cross section at 4^o south of the centre has large-scale descent in the cold air on the left of the plot with cloud largely confined to the boundary layer. Moving to the east, there is a change to large scale ascent and higher cloud fractions throughout the troposphere as we cross the composite warm conveyor belt.

The north-south section shows similar strong ascent and high cloud fractions in the cloud head just to the north of the surface cyclone centre, but also an indication of a secondary maximum at the southern end (-5^o, 2km to -12^o, 6km) where the section will sometimes pass through a trailing cold front.

The same compositing methodology can be applied to the model with a simulated RL-GEOPROF product from the CloudSat and CALIPSO simulators. The difference between the modelled and observed composite cyclones can be calculated (Figure 7). Both model configurations have excess hydrometeor frequency in the boundary layer around the cyclone. This is slightly improved in GA7 with the largest bias confined to the western periphery of the cyclone. GA6 also has considerably too much cirrus on the rearward side of the frontal regions. The excess cirrus is completely removed in GA7 through the reduced cirrus spreading rate such that cloud amount biases in the free troposphere around the GA7 composite cyclone are very small.

A case study again provides a useful illustration of the excess cirrus in GA6 (Figure 8). In this example the A-train passed over a mature depression in a very similar section to the lower-right panel of the cyclone composite in Figure 6. Given this is a forecast with a greater than 1 day lead time, the simulated positions of the frontal features are very good. The main bias is the width of the cloud associated with the warm conveyor belt being too large, especially visible for the trailing cold front ~~. Hence in this case, the bias is grid-boxes which ought to be clear are cloud covered rather than the fractional coverage of partly cloudy boxes being too high. Indeed~~at around 44^{o}N. Examining the cloud fraction on the model grid, there are instances on the edges of the fronts where the observations suggest clear sky but the model simulates partially cloud grid-boxes. In contrast, within the cloud head around 60^{o}N there is an indication that the model too readily breaks up

the cloud when the grid box should be completely covered~~(similar to the highest cirrus in the tropical case)~~. This tendency for the model to too often simulate partially cloudy grid-boxes rather than 0% or 100% is consistent with previous experience with the UM (e.g. Mittermaier, 2012) and may relate to a critical relatively humidity still be used to initially form/decay cloud when the grid box is 0%/100% cloud covered respectively.

The same cyclone compositing methodology has been carried out over the northern hemisphere oceans for June-July-August (JJA) and for the summer and winter seasons in the southern hemisphere (40^oS–70^oS). We have also composited anticyclones using the same cyclone settings as Field and Wood (2007), but testing for $d^2p/dx^2 + d^2p/dy^2 < 0$ in order to identify a local maxima in surface pressure rather than a local minima. All the plots are available in the Supplementary ~~material~~ Material and show a broadly similar picture of excess cloud in the free troposphere and boundary layer in GA6, the former being essentially fixed and the latter improved in GA7. The GA6 cirrus biases in anticyclones are smaller than cyclones, but the boundary layer issues are more comparable. The cyclone composite for the Southern Hemisphere summer now suggests slightly too little mid-level (2-5km) cloud on the cold air side (poleward and westward side) of the cyclone in GA7 (Supplementary Material Figure 2). This may be associated with a lack of congestus cloud here which is a long-standing problem, but was being masked in GA6 through the excess cirrus throughout the free troposphere. Govekar *et al* (2011) provided an evaluation of cyclone composite cloud amounts over the Southern Ocean in an earlier configuration of the UM (Australian Community Climate and Earth System Simulator, ACCESS1.3). They concluded that ~~whilst~~ while the cloud simulation was ~~reasonably good~~in reasonable agreement with observations, the large scale vertical velocity was poor and they cautioned that there may be a compensating error in the cloud simulation. In

both GA6 and GA7, the vertical velocities in the cyclone composites compare well with ERA-I (e.g. Figure 7), hence this issue is no longer of concern.

Despite the cloud amount composites ~~being reasonably good, especially~~ showing cloud fraction errors of less than 0.15 (and often less than 0.05) in GA7, composites of the top of atmosphere (TOA) radiation biases reveal some issues (Figure 9). The ~~outgoing longwave radiation (OLR)~~ OLR is slightly too low across the cyclone composites which is believed to generally reflect a slight tropospheric cold bias in the model. However, the main issue is in the reflected shortwave (RSW). Unsurprisingly, this error is larger in the summer season in each hemisphere when the insolation is greatest. The northern hemisphere has excess RSW across the cyclone composite, and particularly in regions of the composite with more cloud. In contrast the southern hemisphere has a large deficit of RSW on the cold air side of the cyclone, a common bias in climate models (Bodas-Salcedo *et al*, 2014). The northern hemisphere being too reflective can also be seen in the anticyclone composites (Supplementary material Figure 4), but the southern hemisphere error seems mainly confined to the cyclone composite.

Figure 10 shows composite cyclone in-cloud albedo biases against ISCCP. ~~These~~ In contrast to the RSW, the in-cloud albedo does not depend on the insolation and so a cloud microphysical error affecting the albedo which is present throughout the year will appear the same in the DJF and JJA plots. However, these albedo biases have a structure which is consistent with the radiation errors ~~. Since the~~ e.g. the fact that the negative RSW bias on the poleward side of the southern hemisphere cyclone is larger in DJF than JJA is partly due to there being a larger albedo error in the austral summer rather than just the insolation being higher. In the northern hemisphere, the DJF in-cloud albedo has the largest positive bias in the south-west quadrant of the composite cyclone, which is

where there is the largest positive bias in RSW; whereas in JJA, the in-cloud albedo bias is more in the central and south-east side, again consistent with the RSW error. Unlike the in-cloud albedo errors, the cloud amount errors ~~are not large enough to contribute significantly to these SW errors , we~~ in Figure 7 and the Supplementary Material appear not well correlated spatially with the RSW errors around the composite cyclone. We therefore suggest that microphysical processes are primarily responsible for the SW errors through incorrect cloud albedos. This is a good example of the value of the compositing technique for understanding the likely cause of radiation errors. Although the subject of ongoing research, we believe that the ~~bias for cloud albedos on the cloud-air~~ negative in-cloud albedo bias on the cold-air side of the southern hemisphere cyclone ~~to be too low~~ is due to a lack of super-cooled liquid water (Bodas-Salcedo *et al*, 2016), whereas the northern hemisphere bias is thought to be associated with issues around the simulation of aerosols and their interaction with the clouds, particularly the strong cloud–aerosol interaction noted earlier.

5 Cloud evaluation over mid-latitude land

Much of the northern hemisphere mid-latitudes are land covered and here we composite the RL-GEOPROF hydrometeor fraction and CALIPSO cloud fraction, along with their simulated equivalents, by ω_{500}. We illustrate the results for DJF (Figure 11), although JJA is qualitatively similar. The excess cirrus issue in GA6 can again be seen and this is removed in GA7. For some of the regimes, it looks as though there may be now too little cirrus in GA7, although these are the relatively less populated regimes of strongest ascent and strongest subsidence.

There appears to be a significant excess of hydrometeor fraction in both model config-

urations at around 1km, however the CALIPSO profiles suggest the cloud fractions at this level are generally correct. This exemplifies the utility of of using multiple observation types and indicates that the excess hydrometeor in the RL-GEOPROF comparison is either low cloud in situations where there is thick high cloud above, ~~or is~~ and/or excess precipitation. Although a detailed investigation is yet to be carried out, it is suspected that both may be contributing. Case study analysis in the vicinity of the UK in February 2015 has identified a few occasions with spurious drizzle/light rain falling from stratocumulus (not shown). Unlike the warm drizzle cases in the tropics which were improved by changes to the auto-conversion scheme in GA7, these mid-latitude winter cases have frozen cloud tops. It is possible that the microphysical errors leading to excess drizzle in frozen stratocumulus seen in the case study are a general issue contributing to the bias in Figure 11. However, low cloud is frequently simulated by the model over land areas in the winter and given that a cirrus shield is present on many occasions, it is quite possible that excess low cloud is also being simulated but shielded from the CALIPSO simulator.

The active satellite instruments provide an invaluable global picture of the three dimensional cloud structure through most of the troposphere, however the radar can be contaminated with ground clutter in the lowest few hundred metres, and the lidar will frequently be attenuated before detecting the lowest cloud layers. Accurate predictions of cloud near the surface are of the highest importance for a number of users of the model, especially aviation. Here we use SYNOP data which, whilst having a reasonable global coverage over land, are likely to be the most reliable observation type available for this lowest layer~~whilst having a reasonable global coverage over land.~~ . They avoid the ground clutter issues of remote sensing from space and an upward pointing ceilometer or human observer looking from the ground is likely to achieve higher accuracy for low cloud bases

as they avoid the problem of attenuation from cloud above. By looking at the lowest 1km, many of the issues associated with the SYNOP data (combining human and automated data and differing observational errors associated with each) ~~may be minimised~~ will be minimised (Mittermaier, 2012). In order to confine the analysis to cloud with bases below 1km, we use the cloud base height observation and look at frequency of occurrence of cloud bases below 1km. The cloud base height is defined as the height of cloud with coverage of 3 oktas or more, hence instances of small cloud coverage are excluded from this analysis. As a consequence, significant model biases in this diagnostic can appear if the observed cloud amount is typically just over 3 oktas and the model cloud fraction is just under (or vice-versa). This appears to be an issue for the UM in parts of the tropics ~~(not shown) ,~~ ~~however more~~ where too little shallow cumulus is simulated and typically the model has cloud fractions of <3 oktas (i.e. grid box fraction of <0.375) whereas fractions over this threshold are often observed and hence a cloud base height assigned. More generally the diagnostic is reflecting errors in the frequency of occurrence of low-base cloud. Based on comparison with the active instruments at higher altitudes, we suspect that biases are more often reflecting errors in the frequency of occurrence of low cloud rather than errors in the cloud base height on any one occasion.

Figure 12 shows the day 1 bias in the frequency of occurrence of cloud base height for one year of data since GA6 became operational. Note that here the term 'bias' uses the definition of the the international Joint Working Group for Forecast Verification Research as being (hits + false alarms)/(hits + misses) (http://www.cawcr.gov.au/projects/verification/), so a value of 1.0 would indicate no model bias. In order to visualise the station density more clearly, we show a section over Europe which illustrates the key points of the mid-latitude land regions in general. Over most of the area the model performs well and

is essentially unbiased. Its performance over the UK is comparable to a 1.5km convective permitting configuration of the UM which is run operationally over the region (not shown). However over areas of notable orography, such as the Alps, there appears to be excess low cloud in the model. In contrast, around some of the coasts (especially France and Italy) there is too little low cloud. Further work is required to ~~indentify~~ identify the cause of these errors.

6 Global cloud radiative effects

Traditionally the primary evaluation of clouds in climate models was through an assessment of their impact on the TOA radiation budget. However, as discussed in the introduction, this could hide compensating errors which might result in an incorrect cloud radiative response to climate change. We suggest instead that this assessment should be towards the end of a wider cloud evaluation, such as that presented above, feeding into the model development process.

The GA6 and GA7 bias in TOA RSW and OLR is shown in Figure 13. Generally the biases are reasonably similar with some local improvements (e.g. in RSW over India and the equatorial Indian Ocean) and local detriments (e.g. in OLR over the Maritime Continent). A widespread bias for the free troposphere to be too cold in GA6 has been slightly improved in GA7 (mainly due to the introduction of the 6A convection scheme (Walters *et al*, 2017)) which largely accounts for the general increase in OLR in the newer model. Given that GA7 will be the physical model underpinning the UK submission to CMIP6, it is useful to compare back to HadGEM2-A (Hadley Centre Global Environmental Model 2 - Atmosphere; Martin *et al*, 2011) which was the CMIP5 submission. It should be noted that HadGEM2-A is a comparatively old model with some 7 years of continuous

model development having occurred between this and GA6, hence the differences in the radiation budget are much larger. It can be seen that GA7 is a considerable improvement on HadGEM2, especially for the RSW. The error in the sub-tropical cumulus transition regions of excess RSW has been ~~eliminated, whilst the~~ removed and there is now a smaller negative bias in GA7. The lack of RSW over the Southern Ocean has been reduced by a third and RSW & OLR biases over the Maritime Continent have been significantly improved.

Metrics are often used to summarise the overall performance of the model. There are few such metrics in the literature for NWP–seasonal cloud prediction applications, however a number have been proposed for aspects of the cloud simulation which are likely to be important for the radiative response of cloud to climate change (e.g. Pincus *et al*, 2008; Klein *et al*, 2013; Myers and Norris, 2015). Here we illustrate the calculation of metrics as the final step in the evaluation process by presenting the present day Cloud Regime Error Metric (CREMpd) of Williams and Webb (2009). This metric assesses the ability of the models to simulate primary cloud regimes (as determined by the daily mean cloud cover, optical depth and cloud top height) with the correct frequency of occurrence and radiative properties. Here we modify one aspect of the Williams and Webb (2009) approach by using the newer global regimes proposed by Tselioudis *et al* (2013) instead of calculating the tropics, extra-tropics and snow/ice covered regions separately. Figure 14 shows the CREMpd for GA6, GA7 and all the CMIP5 models for which the required data are available, with zero being a perfect score compared with the observations. GA6 is comparable with the previous HadGEM2-A model as being among the better performing models on this metric, with GA7 performing slightly worse but still competitive with other CMIP5 models. Having a climate change application focus, CREMpd is very sensitive to

the accuracy of the simulation of clouds with the strongest net radiative effect, namely stratocumulus. Consequently GA7 is penalised compared with GA6 for the overall reduction in the albedo of sub-tropical stratocumulus (Figure 4). In contrast, the metric has limited acknowledgment of the large improvements in the amount of cirrus in GA7 since the radiative effect of this, largely sub-visual cloud, is small.

7 Summary and discussion

In this study we have attempted to convey a more thorough evaluation of cloud than has traditionally been undertaken as part of a model development process. Our experience has been that using a limited set of diagnostics and/or observational datasets can result in compensating errors. An example is the rate of cirrus spreading which was part of a change introduced in GA4 (Walters *et al*, 2014), but at the time we were not routinely evaluating against CALIPSO. We have now discovered that this was producing excessive amounts of sub-visual cirrus and this has been corrected in GA7. The ability to compare the models with multiple satellite datasets using COSP, combined with a variety of compositing techniques has permitted a detailed, process-orientated evaluation to be undertaken. We find that the use of multiple datasets and diagnostic techniques to draw a consistent picture of model errors is likely to reduce the risk of drawing the wrong conclusions and more accurately focus future model development. Examples include the comparisons between CloudSat and CALIPSO that demonstrate errors due specifically to thin cirrus, or to excess precipitation as opposed to cloud error; the use of cyclone composites of cloud amount, in-cloud albedo and radiative fluxes to show, through similar spatial patterns, that the error in the RSW is likely due to errors in the in-cloud albedo rather than cloud amount; the use of surface-based observations for the lowest atmospheric layers where

remote sensing from space becomes problematic; etc.

The combination of CloudSat and CALIPSO provides a unique three dimensional observational dataset of hydrometeor frequency through much of the atmosphere. We find that some care is required in its use for model evaluation in terms of separating cloud and precipitation, and the ability to perform multiple simulations passing different fields to the simulator can be valuable. Despite being an older satellite dataset, the optical depth information from ISCCP remains extremely valuable for model evaluation purposes. Evaluation of very low cloud (<1km) remains a challenge, especially when thicker cloud exists above. We have made use of the SYNOP data which have reasonable coverage over land and, for cloud at these altitudes, may be regarded as fairly reliable. The thresholds and variables available in the SYNOP data do limit the evaluation though.

A key part of our evaluation process is the cross-timescale assessment which enables the statistical robustness of the climate simulations to be combined with more detailed analysis of case studies in NWP hindcasts to understand the model errors at the process level. Although many centres don't routinely run simulations across these timescales, the AMIP and Transpose-AMIP experiments proposed by the Working Group Numerical Experimentation (WGNE) provide a relatively simple methodology enabling all centres to benefit from this approach.

GA6 generally performs well given the critical examination presented here. The main errors are:

1. A considerable excess of thin, often sub-visual, cirrus erroneously extending from thicker cirrus clouds which ought to be present. This has been essentially fixed in GA7.

2. In-cloud albedo is too high in tropical and extra-tropical stratocumulus, except on

the cold air side of cyclones in the Southern hemisphere where they are too low.

3. A slight excess of boundary layer hydrometeor fraction over the mid-latitudes which is suspected to be a combination of excess cloud and drizzle.

Apart from errors in external driving factors such as the location and timing of convection and synoptic systems, item 2 in the list above is the main cloud error affecting the mean radiation bias.

Although we have attempted the most comprehensive assessment possible in the time available, the task is inevitably open ended. The main omissions which we would have liked to address are an evaluation of the diurnal cycle of clouds globally and cloud over high latitude regions. Sea ice and snow cover are likely to be quite sensitive to cloud and this is a region which has generally received little detailed systematic cloud evaluation. Use of data from additional instruments such as ground-based cloud radar and lidar, and from the Multi-angle Imaging Spectro-Radiometer (MISR) satellite instrument would also be valuable additions in future studies.

Code availability

The UM is available for use under licence. A number of research organisations and national meteorological services use the UM in collaboration with the Met Office to undertake basic atmospheric process research, produce forecasts, develop the UM code and build and evaluate Earth system models. For further information on how to apply for a licence see http://www.metoffice.gov.uk/research/collaboration/um-collaboration. Versions 8.6 (for GA6) and 10.3 (for GA7) of the source code are used in this paper.

Acknowledgments

This work was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). We thank Cyril Morcrette, Paul Field and David Walters for useful discussions throughout the study.

References

Abdul-Razzak, H., and S. J. Ghan, 2000: A parameterization of aerosol activation: 2. multiple aerosol types. *J. Geophys. Res.*, **105**, 6837–6844. doi:10.1029/1999JD901161.

Baran, A. J., R. Cotton, K. Furtado, S. Havemann, L.-C. Labonnote, F. Marenco, A. Smith, and J.-C. Thelen, 2014: A self-consistent scattering model for cirrus. II: The high and low frequencies. *Q. J. R. Meteorol. Soc.*, **140**(), 1039–1057. doi:10.1002/qj.2193.

Beare, R. J., 2008: The role of shear in the morning transition boundary layer. *Boundary-Layer Meteorol.*, **129**, 395–410. doi:10.1007/s10546-008-9324-8.

Bellouin, N., J. Rae, A. Jones, C. Johnson, J. Haywood, and O. Boucher, 2011: Aerosol forcing in the Climate Model Intercomparison Project (CMIP5) simulations by HadGEM2-ES and the role of ammonium nitrate. *J. Geophys. Res.*, **116**(D20). doi:10.1029/2011jd016074.

Bodas-Salcedo, A., M. J. Webb, S. Bony, H. Chepfer, J. L. Dufresne, S. Klein, Y. Zhang, R. Marchand, J. M. Haynes, R. Pincus, and V. O. John, 2011: COSP: satellite simulation software for model assessment. *Bull. Am. Meteorol. Soc.*, **92**(8), 1023–1043. doi:10.1175/2011BAMS2856.1.

——, K. D. Williams, M. A. Ringer, I. Beau, J. N. S. Cole, J.-L. Dufresne, T. Koshiro, B. Stevens, Z. Wang, and T. Yokohata, 2014: Origins of the solar radiation biases over the Southern Ocean in CFMIP2 models. *J. Climate*, **27**, 41–56. doi:10.1175/JCLI-D-13-00169.1.

——, P. G. Hill, K. Furtado, K. D. Williams, P. R. Field, J. C. Manners, P. Hyder, and S. Kato, 2016: Large contribution of supercooled liquid clouds to the solar radiation budget of the Southern Ocean. *J. Climate*, **29**(11), 4213–4228. doi:10.1175/JCLI-D-15-0564.1.

Bony, S., and J. L. Dufresne, 2005: Marine boundary layer clouds at the heart of cloud feedback uncertainties in climate models. *Geophys. Res. Lett.*, **32**(20), L20806.

——, J.-L. Dufresne, H. Le Treut, J.-J. Morcrette, and C. A. Senior, 2004: On dynamic and thermodynamic components of cloud changes. *Clim. Dyn.*, **22**, 71–86. doi:10.1007/s00382-003-0369-6.

719 Boutle, I. A., S. J. Abel, P. G. Hill, and C. J. Morcrette, 2014: Spatial variability of liquid cloud and rain: observations and microphysical effects. *Q. J. R. Meteorol. Soc.*, **140**, 583–594. doi:10.1002/qj.2140.

720 Chepfer, H., S. Bony, D. Winker, M. Chiriaco, J.-L. Dufresne, and G. Sèze, 2008: Use of CALIPSO lidar observations to evaluate the cloudiness simulated by a climate model. *Geophys. Res. Lett.*, **35**, L15704. doi:10.1029/2008GL034207.

721 ——, ——, ——, G. Cesana, J.-L. Dufresne, P. Minnis, C. J. Stubenrauch, and S. Zeng, 2010: The GCM Oriented Calipso Cloud Product (CALIPSO-GOCCP). *J. Geophys. Res.*, **115**, D00H16. doi:10.1029/2009JD012251.

722 ——, G. Cesana, D. Winker, B. Getzewich, M. Vaughan, and Z. Liu, 2013: Comparison of two different cloud climatologies derived from CALIOP-attenuated backscattered measurements (Level 1): The CALIPSO-ST and the CALIPSO-GOCCP. *J. Atmos. Oceanic Technol.*, **30**, 725–744. doi:10.1175/JTECH-D-12-00057.1.

723 Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hölm, L. Isaksen, P. Kallberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J. J. Morcrette, P. K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J. N. Thêpaut, and F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.*, **137**(656), 553–597. doi:10.1002/qj.828.

724 Field, P. R., and R. Wood, 2007: Precipitation and cloud structure in midlatitude cyclones. *J. Climate*, **20**(2), 233–254. doi:10.1175/JCLI3998.1.

725 ——, A. J. Heymsfield, and A. Bansemer, 2007: Snow size distribution parameterization for midlatitude and tropical ice clouds. *J. Atmos. Sci.*, **64**, 4346–4365. doi:10.1175/2007JAS2344.1.

726 ——, A. A. Hill, K. Furtado, and A. Korolev, 2014: Mixed-phase clouds in a turbulent environment. Part 2: Analytic treatment. *Q. J. R. Meteorol. Soc.*, **140**(), 870–880. doi:10.1002/qj.2175.

727 Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen, 2013: Evalutation of climate models. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, Eds., Cambridge University Press, 741–866. doi:10.1017/CBO9781107415324.020.

728 Gates, W., 1992: The atmospheric model intercomparison project. *Bull. Am. Meteorol. Soc.*, **73**, 1962–1970.

729 Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104. doi:10.1029/2007JD008972.

730 Govekar, P. D., C. Jakob, M. J. Reeder, and J. Haynes, 2011: The three-dimensional distribution of clouds around Southern Hemisphere extratropical cyclones. *Geophys. Res. Lett.*, **38**, L21805. doi:10.1029/2011GL049091.

731 Haynes, J. M., R. T. Marchand, Z. Luo, A. Bodas-Salcedo, and G. L. Stephens, 2007: A multi-purpose radar simulation package: Quickbeam. *Bull. Am. Meteorol. Soc.*, **88**(11), 1723–1727. doi:10.1175/BAMS-88-11-1723.

732 Khairoutdinov, M. F., and Y. L. Kogan, 2000: A new cloud physics parameterization in a large-eddy simulation model of marine stratocumulus. *Mon. Weather Rev.*, **128**(), 229–243.

733 Klein, S. A., and D. L. Hartmann, 1993: The seasonal cycle of low stratiform clouds. *J. Climate*, **6**(8), 1587–1606.

734 ——, and C. Jakob, 1999: Validation and sensitivities of frontal clouds simulated by the ECMWF model. *Mon. Weather Rev.*, **127**(10), 2514–2531.

735 ——, Y. Zhang, M. D. Zelinka, R. Pincus, J. Boyle, and P. J. Gleckler, 2013: Are climate model simulations of clouds improving? An evaluation using the isccp simulator. *J. Geophys. Res.*, **118**, 1329–1342. doi:10.1002/jgrd.50141.

736 Loeb, N. G., B. A. Wielicki, D. R. Doelling, S. Kato, T. Wong, G. L. Smith, D. F. Keyes, and N. Manalo-Smith, 2009: Toward optimal closure of the Earth's top-of-atmosphere radiation budget. *J. Climate*, **22**(3), 748–766. doi:10.1175/2008JCLI2637.1.

737 Mace, G. G., and F. J. Wrenn, 2013: Evaluation of the hydrometeor layers in the east and west Pacific within ISCCP cloud-top pressure–optical depth bins using merged CloudSat and CALIPSO data. *J. Climate*, **26**(23), 9429–9444. doi:10.1175/JCLI-D-12-00207.1.

738 ——, and Q. Zhang, 2014: The CloudSat radar-lidar geometrical profile product (RL-GeoProf): Updates, improvements, and selected results. *J. Geophys. Res.*, **119**(15), 9441–9462. doi:10.1002/2013JD021374.

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

740 Marchand, R., G. G. Mace, T. Ackerman, and G. L. Stephens, 2008: Hydrometeor detection using Cloudsat - an Earth-Orbiting 94-GHz cloud radar. *J. Atmos. Oceanic Technol.*, **25**, 519–533. doi:10.1175/2007JTECHA1006.1.

741 Martin, G. M., N. Bellouin, W. J. Collins, I. D. Culverwell, P. R. Halloran, S. C. Hardiman, T. J. Hinton, C. D. Jones, and others, 2011: The HadGEM2 family of Met Office Unified Model climate configurations. *Geosci. Model Devel.*, **4**, 723–757. doi:10.5194/gmd-4-723-2011.

742 Mittermaier, M., 2012: A critical assessment of surface cloud observations and their use for verifying cloud forecasts. *Q. J. R. Meteorol. Soc.*, **138**(), 1794–1807. doi:10.1002/qj.1918.

743 Myers, T. A., and J. R. Norris, 2015: On the relationships between subtropical clouds in meteorology in observations and CMIP3 and CMIP5 models. *J. Climate*, **28**(), 2945–2967. doi:10.1175/JCLI-D-14-00475.1.

744 Nam, C., S. Bony, J.-L. Dufresne, and H. Chepfer, 2012: The 'too few, too bright' tropical low-cloud problem in CMIP5 models. *Geophys. Res. Lett.*, **39**(L21801). doi:10.1029/2012GL053421.

745 Pincus, R., C. P. Batstone, R. J. Patrick-Hofmann, K. E. Taylor, and P. E. Gleckler, 2008: Evaluating the present-day simulation of clouds, precipitation and radiation in climate models. *J. Geophys. Res.*, **133**(D14209). doi:10.1029/2007JD009334.

746 Rossow, W. B., and R. A. Schiffer, 1999: Advances in understanding clouds from ISCCP. *Bull. Am. Meteorol. Soc.*, **80**, 2261–2287.

747 Stephens, G. L., D. G. Vane, R. J. Boain, G. G. Mace, K. Sassen, Z. Wang, A. J. Illingworth, E. J. O'Connor, W. B. Rossow, S. L. Durden, S. D. Miller, R. T. Austin, A. Benedetti, C. Mitrescu, and The CloudSat Science Team, 2002: The CloudSat mission and the A-Train. *Bull. Am. Meteorol. Soc.*, **83**, 1771–1790.

748 Tselioudis, G., W. Rossow, Y. Zhang, and D. Konsta, 2013: Global weather states and their properties from passive and active satellite cloud retrievals. *J. Climate*, **26**, 7734–7746. doi:10.1175/JCLI-D-13-00024.1.

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

750 Walters, D., M. Brooks, I. Boutle, T. Melvin, R. Stratton, S. Vosper, H. Wells, K. Williams, N. Wood, T. Allen, A. Bushell, D. Copsey, P. Earnshaw, J. Edwards, M. Gross, S. Hardiman, C. Harris, J. Heming, N. Klingaman, R. Levine, J. Manners, G. Martin, S. Milton, M. Mittermaier, C. Morcrette, T. Riddick, M. Roberts, C. Sanchez, P. Selwood, A. Stirling, C. Smith, D. Suri, W. Tennant, P. L. Vidale, J. Wilkinson, M. Willett, S. Woolnough, and P. Xavier, 2016: The Met Office Unified Model Global Atmosphere 6.0/6.1 and JULES Global Land 6.0/6.1 configurations. *Geosci. Model Devel.* doi:10.5194/gmd-2016-194.

751 Walters, D. N., A. Baran, I. Boutle, M. E. Brooks, K. Furtado, P. Hill, A. Lock, J. Manners, C. Morcrette, J. Mulchay, C. Sanchez, C. Smith, R. Stratton, W. Tennant, K. Van Weverberg, S. Vosper, H. Ashton, R. Essery, N. Gedney, K. D. Williams, and M. Zerroukat, 2017: The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations. *In preparation.*

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

753 Williams, K. D., and M. E. Brooks, 2008: Initial tendencies of cloud regimes in the Met Office Unified Model. *J. Climate*, **21**(4), 833–840. doi:10.1175/2007JCLI1900.1.

754 ——, and G. Tselioudis, 2007: GCM intercomparison of global cloud regimes: Present-day evaluation and climate change response. *Clim. Dyn.*, **29**, 231–250. doi:10.1007/s00382-007-0232-2.

755 ——, and M. J. Webb, 2009: A quantitative performance assessment of cloud regimes in climate models. *Clim. Dyn.*, **33**(1), 141–157. doi:10.1007/s00382-008-0443-1.

756 ——, M. A. Ringer, and C. A. Senior, 2003: Evaluating the cloud response to climate change and current climate variability. *Clim. Dyn.*, **20**, 705–721. doi:10.1007/s00382-002-0303-3.

757 ——, ——, ——, M. J. Webb, B. J. McAvaney, N. Andronova, S. Bony, J.-L. Dufresne, S. Emori, R. Gudgel, T. Knutson, B. Li, K. Lo, I. Musat, J. Wegner, A. Slingo, and J. F. B. Mitchell, 2006: Evaluation of a component of the cloud response to climate change in an intercomparison of climate models. *Clim. Dyn.*, **26**, 145–165. doi:10.1007/s00382-005-0067-7.

758 ——, A. Bodas-Salcedo, M. Deque, S. Fermepin, B. Medeiros, M. Watanabe, C. Jakob, S. A. Klein, C. A. Senior, and D. L. Williamson, 2013: The Transpose-AMIP II experiment and its application to the understanding of Southern Ocean cloud biases in climate models. *J. Climate*, **26**, 3258–3274. doi:10.1175/JCLI-D-12-00429.1.

759 Winker, D. M., J. Pelon, J. A. Coakley Jr, S. A. Ackerman, R. J. Charlson, P. R. Colarco, P. Flamant, Q. Fu, R. M. Hoff, C. Kittaka, T. L. Kubar, H. L. Treut, M. P. McCormick, G. Mégie, L. Poole, K. Powell, C. Trepte, M. A. Vaughan, and B. A. Wielicki, 2010: The CALIPSO mission: A global 3D view of aerosols and clouds. *Bull. Am. Meteorol. Soc.*, **91**(9), 1211–1229. doi:10.1175/2010BAMS3009.1.

760 WMO, 2008: Guide to meteorological instruments and methods of observation. Technical report WMO-8, World Meteorological Organisation, Geneva.

761 Wood, R., and C. S. Bretherton, 2006: On the relationship between stratiform low cloud cover and lower tropospheric stability. *J. Climate*, **19**, 6425–6432.

762 Zhang, Y., and S. A. Klein, 2013: Factors controlling the vertical extent of fair-weather shallow cumulus clouds over land: Investigation of diurnal-cycle observations collected at the ARM Southern Great Plains site. *J. Atmos. Sci.*, **70**(), 1297–1315. doi:10.1175/JAS-D-12-0131.1.

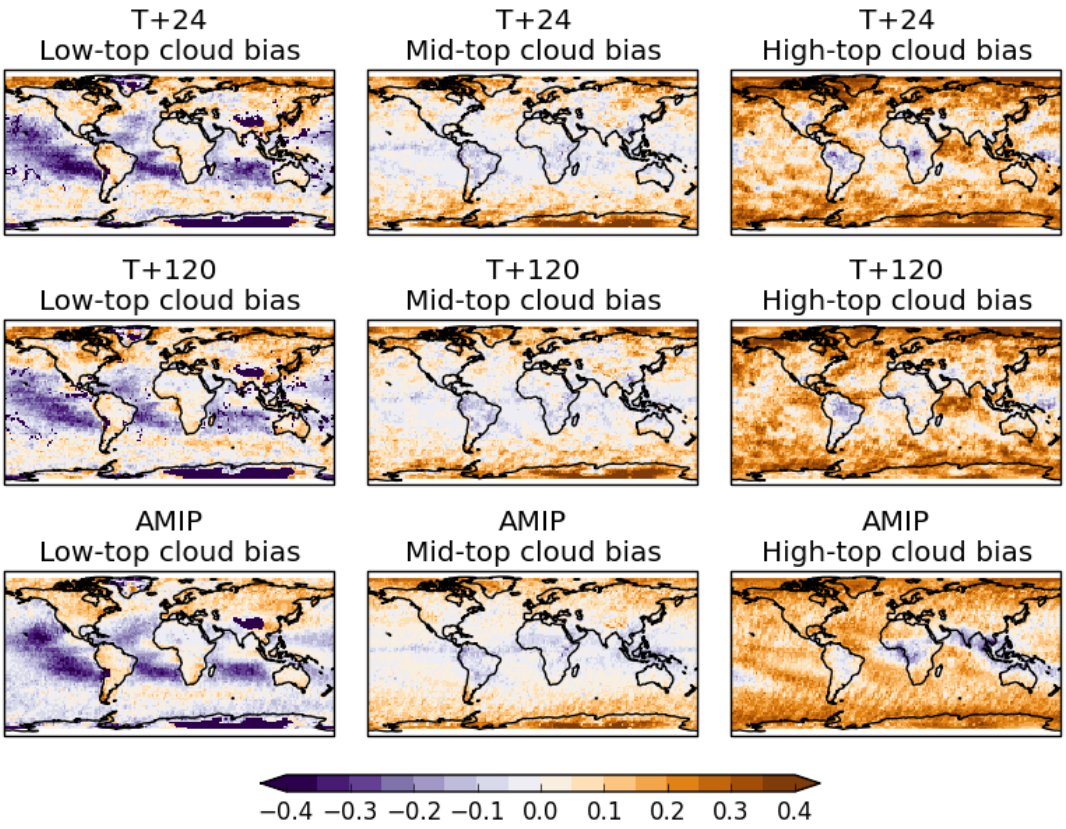

Figure 1: Absolute bias (model field minus observed field) in GA6 configuration of the UM for low (left) mid (centre) and high (right) fractional cloud cover against the GCM Orientated CALIPSO Cloud Product (GOCCP), using the CALIPSO simulator in COSP (see Section b. Top and middle rows are mean biases at day 1 and day 5 averaged across all the NWP hindcasts at N320 (40km in mid-latitudes) resolution. The bottom row is the bias in the AMIP climatology at N96 (135km in mid-latitudes) resolution.

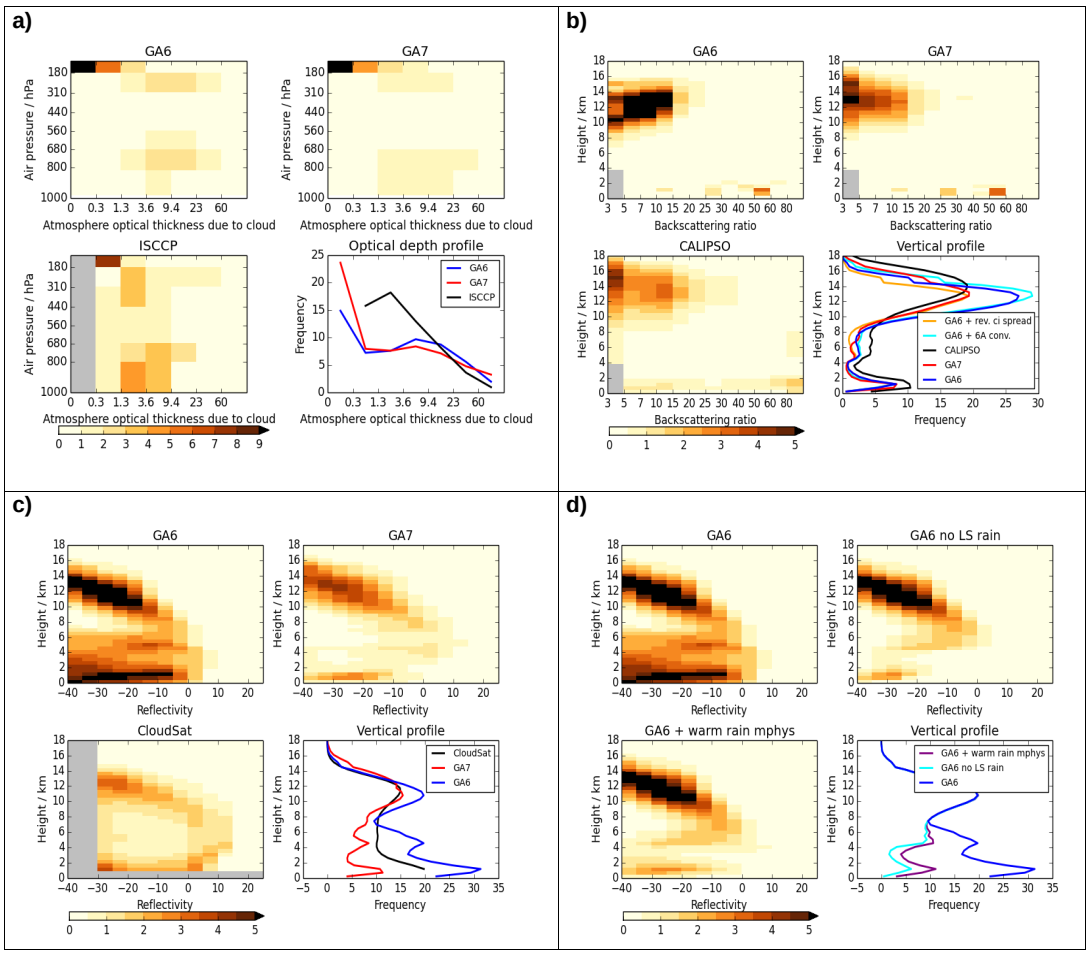

Figure 2: Tropical multi-annual mean observed and GA6 & GA7 simulated satellite data summaries. a) ISCCP cloud-top pressure–optical depth joint frequency histograms. Lower right panel is a single optical depth frequency histogram (i.e. the joint histograms have been summed across cloud top pressure bins). The threshold optical depth for detection by ISCCP is believed to be approximately 0.3, hence the masking of the lowest bin in the observed histogram. b) CALIPSO height–backscattering ratio joint frequency histograms. Lower right panel is a single height frequency histogram (i.e. the joint histograms have been summed across backscattering ratio bins). Within the boundary layer, backscattering ratios <5 are likely to be due to aerosols (see Supplementary Material Figure 1) and hence are masked. The lower right panel also shows frequency profiles for GA6 with the cirrus spreading reduced to GA7 values, and GA6 but with the 6A convection scheme used. c) CloudSat height–radar reflectivity (dBZ) joint frequency histograms. Lower right panel is a single height frequency histogram (i.e. the joint histograms have been summed across reflectivity bins). d) As c) but showing GA6, GA6 without large-scale rain being passed to the simulator, and showing GA6 plus the warm rain microphysics package which is included in GA7. Colour scale for the histograms show frequency of occurrence of cloud/hydrometeor in the bin (%). Shading around the line plots has been added to reflect significance bounds, however this is often less than the thickness of the plotted lines.

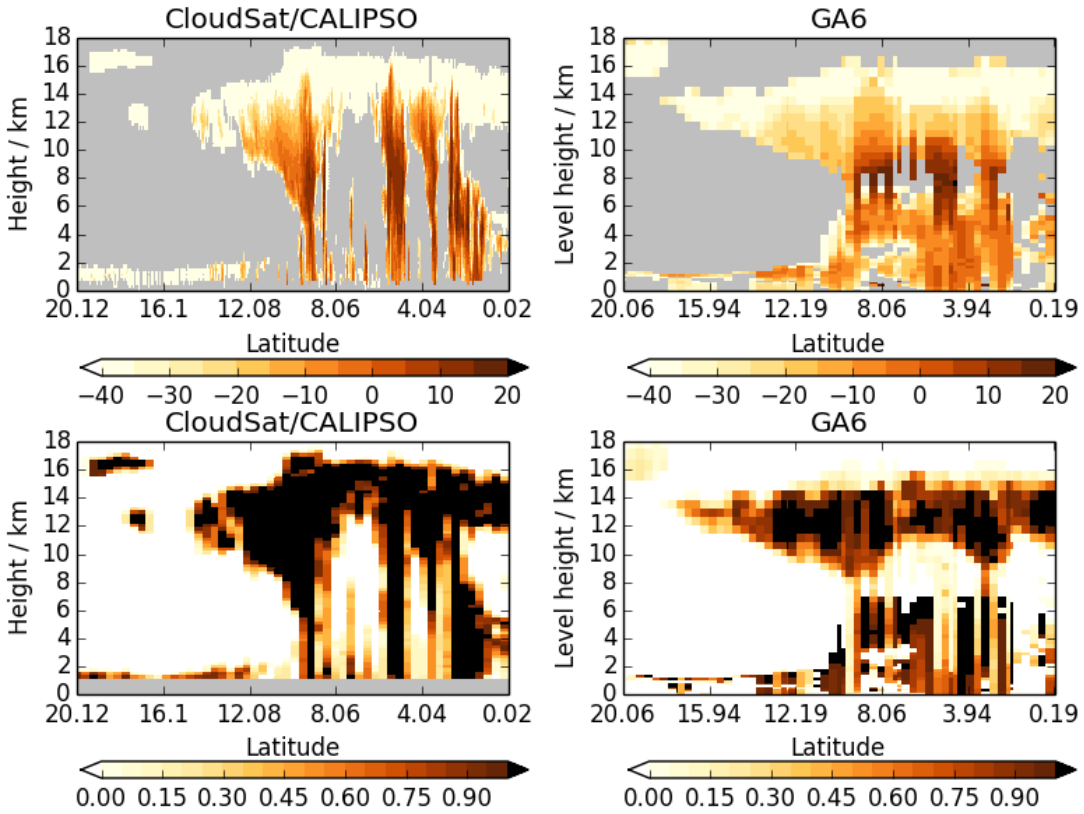

Figure 3: Case study of a GA6 6 hour forecast verifying at 18:00UTC on 17th December 2010 for an A-train pass over the South China Sea. Top: the observed and simulated radar reflectivities (dBZ) with situations in which the lidar detected cloud but the radar did not being included with a nominal value of -40dBZ (e.g. Mace and Wrenn, 2013). Bottom: observed and simulated cloud fraction on the model grid.

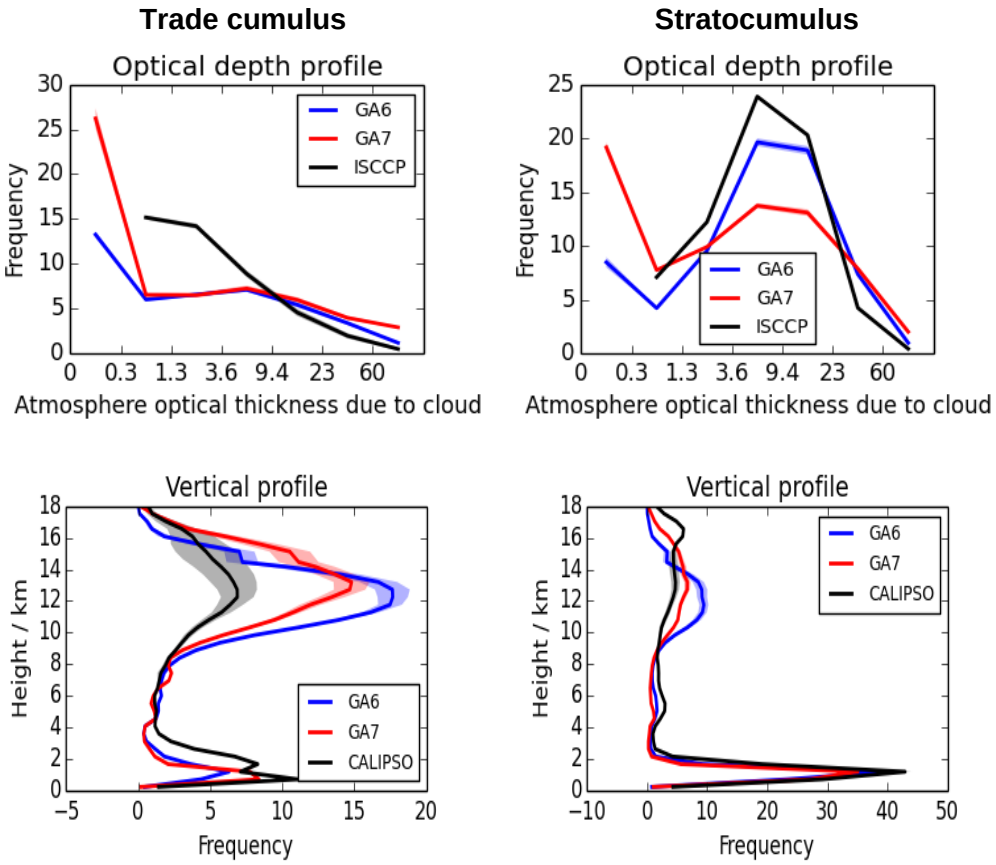

Figure 4: Observed and simulated multi-annual mean ISCCP optical depth frequency histograms (top) and CALIPSO height frequency histograms (bottom) for a trade cumulus region (130-160^{o}W, 0-20^{o}S, left) and stratocumulus region (80-90^{o}W, 0-20^{o}S, right). Shading around the line plots has been added to reflect significance bounds, however this is sometimes less than the thickness of the plotted lines.

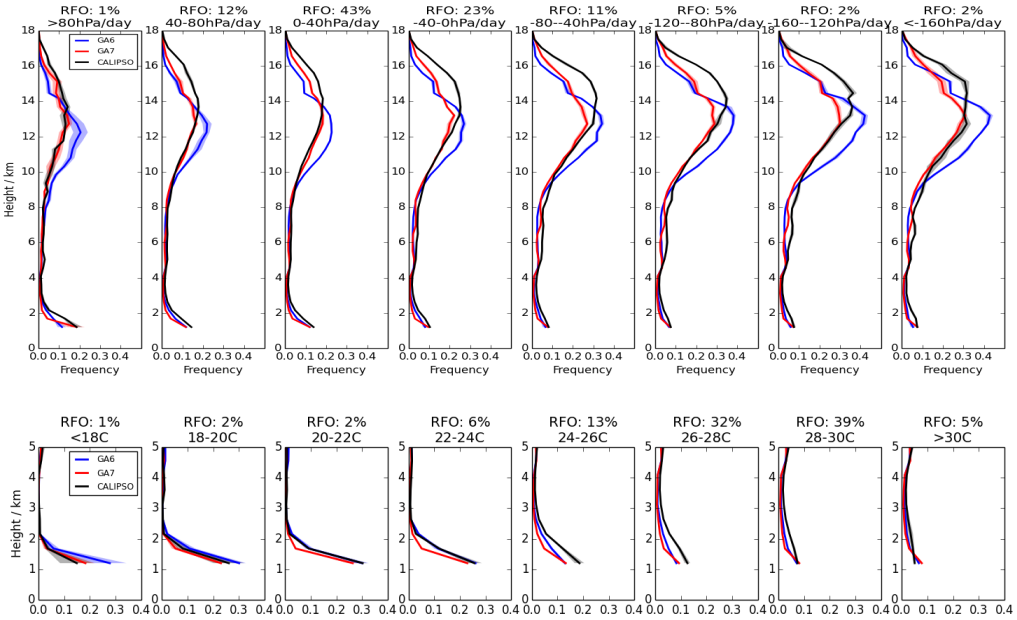

Figure 5: Observed and simulated CALIPSO height frequency histograms composited by daily ω_{500} (top) and SST (bottom) over the tropics (20^oN–20^oS). Only the region below 5km is shown in the lower plot to focus on low cloud. The range and relative frequency of occurrence (RFO) are shown at the top of each bin. Negative ω_{500} indicates ascent. Shading around the line plots has been added to reflect significance bounds, however this is sometimes less than the thickness of the plotted lines.

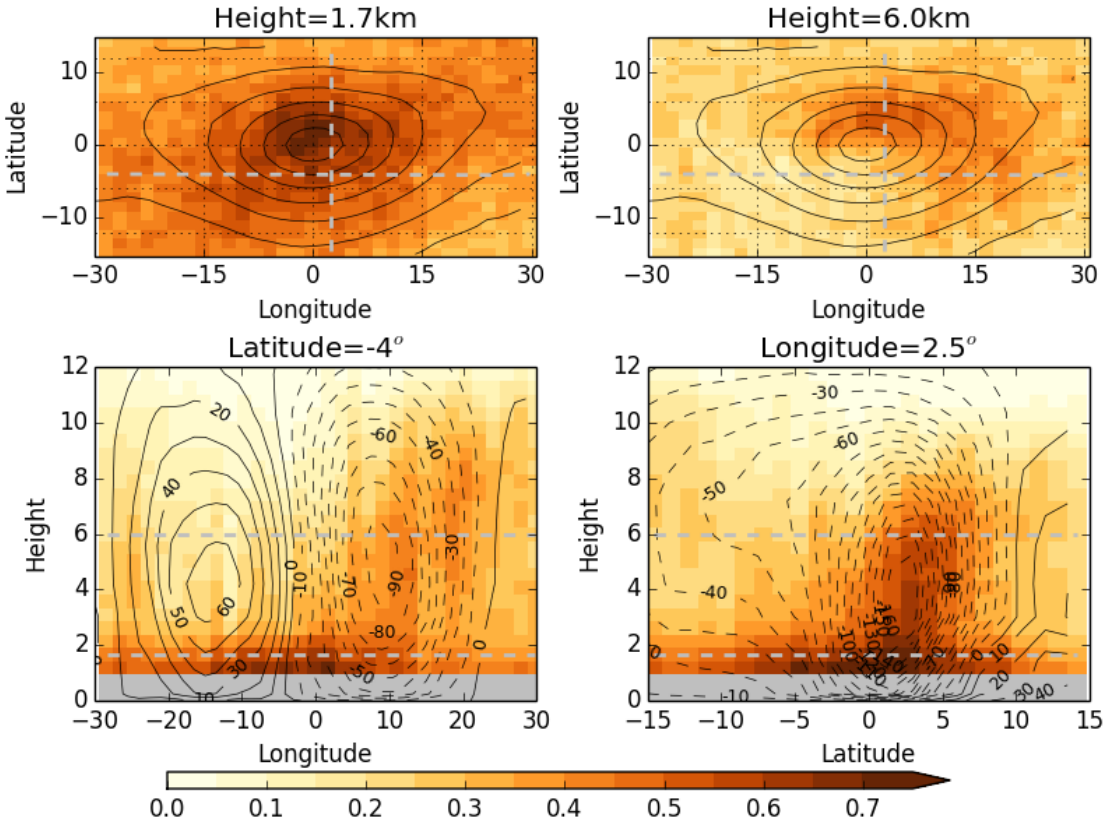

Figure 6: Distribution of average observed hydrometeor (cloud plus precipitation) fraction (colours) around a composite of ERA-I cyclones over northern hemisphere oceans for 5 years of DJF daily data. Top row shows horizontal sections through the composite cyclone at 1.7 & 6km with the mean PMSL contoured at 4hPa intervals. Bottom row shows vertical sections along the grey dashed lines shown in the top plots. Contours on the lower plots are mean vertical velocity from ERA-I (hPa/day; negative values indicate ascent and these contours are dashed).

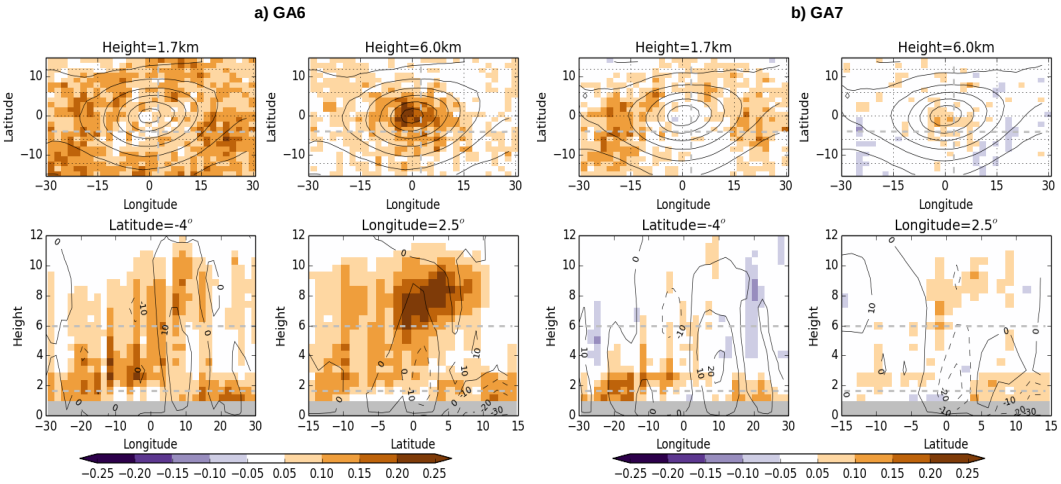

Figure 7: Cloud fraction absolute bias (model field minus observed field) (colours) for composite cyclones. Produced as per Figure 6 for a) GA6 b) GA7 and the observed composite then subtracted. Black contours in top plots are the model mean PMSL and in the lower plots are the bias in vertical velocity. Student's t-test based on internnual variability show that errors greater than 0.05 are significant.

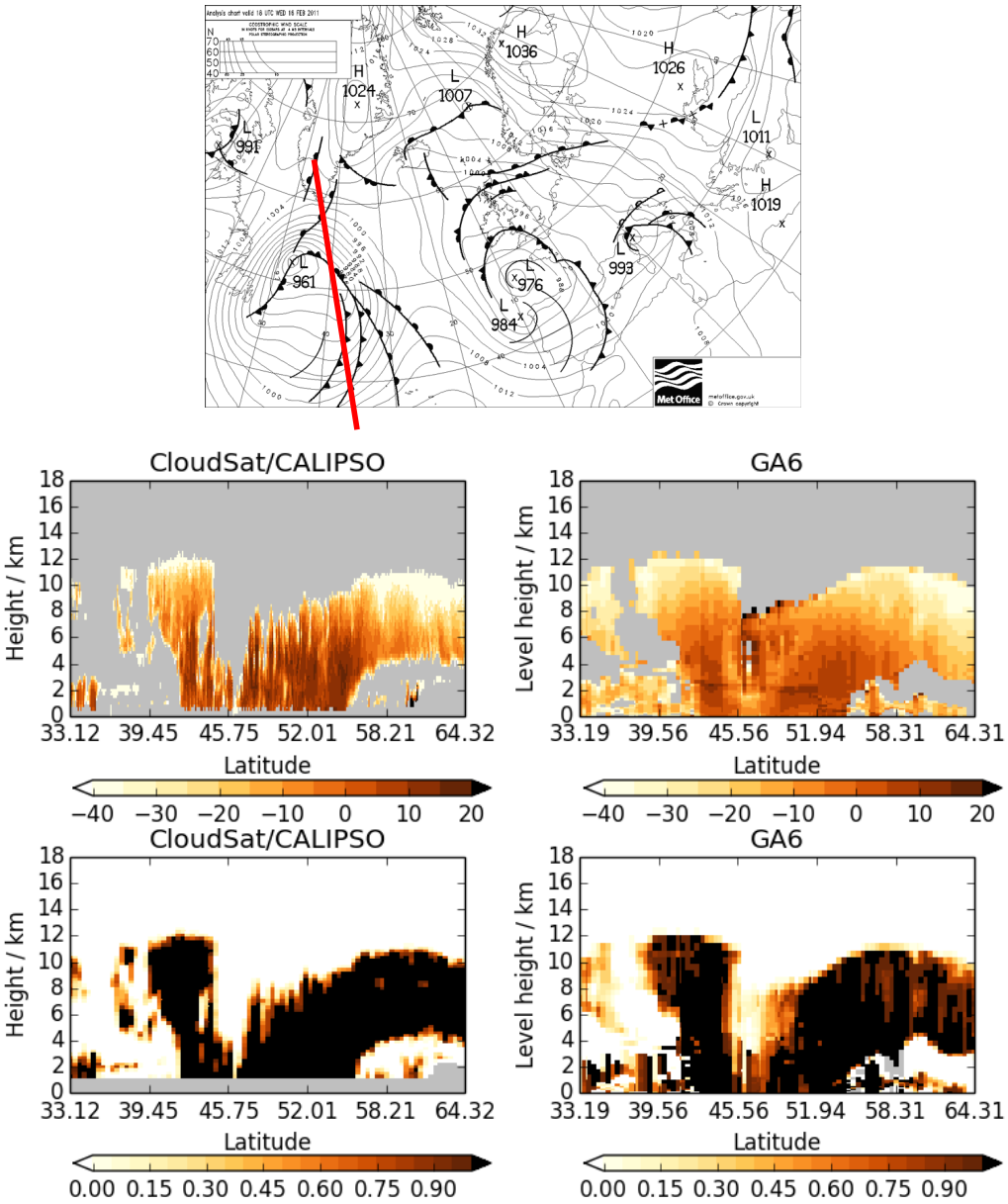

Figure 8: Case study of a GA6 27 hour forecast verifying at 15:00UTC on 16th February 2011 for an A-train pass over the North Atlantic as shown by the red line on the synoptic analysis. Top: the observed and simulated radar reflectivities (dBZ) with situations in which the lidar detected cloud but the radar did not being included with a nominal value of -40dBZ. Bottom: observed and simulated cloud fraction on the model horizontal grid.

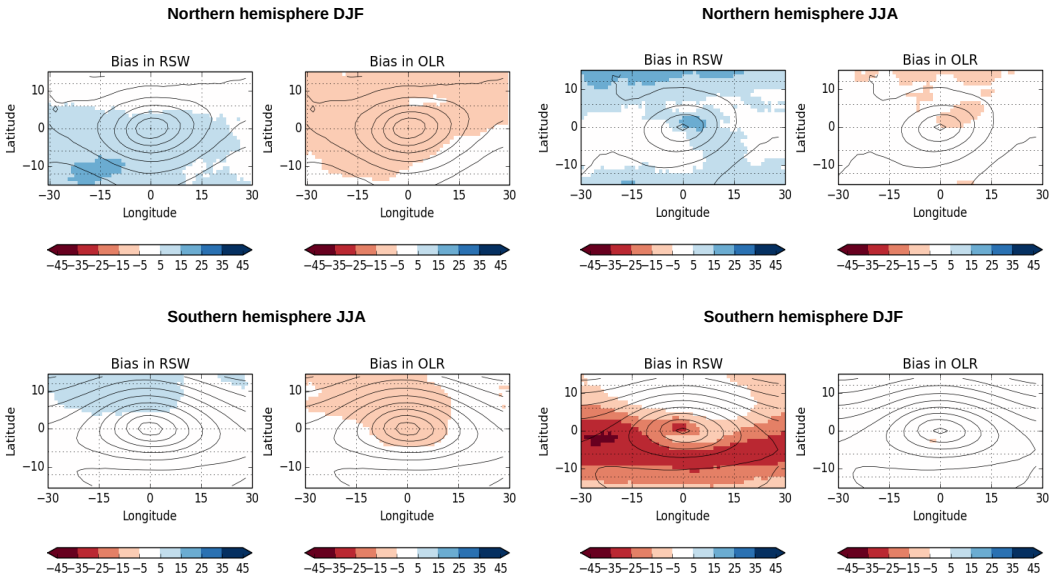

Figure 9: Cyclone composite GA7 mean bias in RSW and OLR (Wm^{-2}) against CERES-EBAF (colours). Black contours are GA7 PMSL. Northern and Southern hemisphere composites are shown for the respective winter (left) and summer (right) seasons. Student's t-test based on internnual variability show that errors greater than $5Wm^{-2}$ are significant.

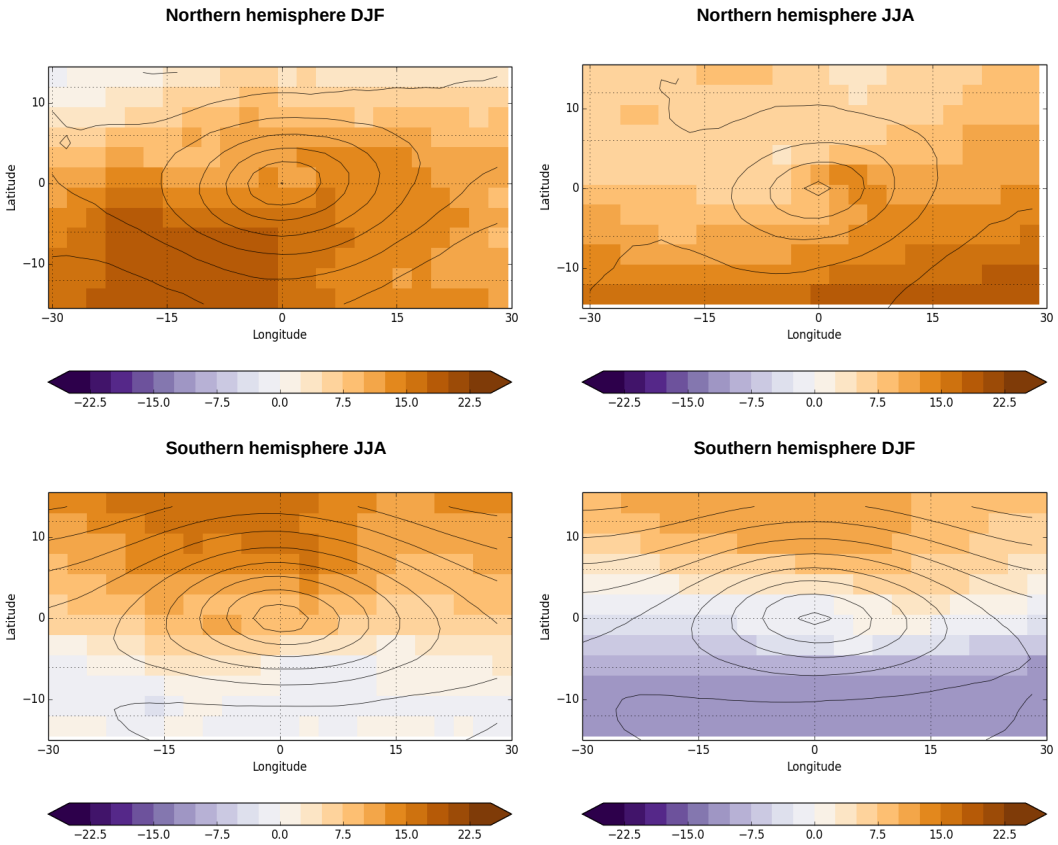

Figure 10: Cyclone composite GA7 mean bias in in-cloud albedo (%) against ISCCP (colours). Black contours are GA7 PMSL.

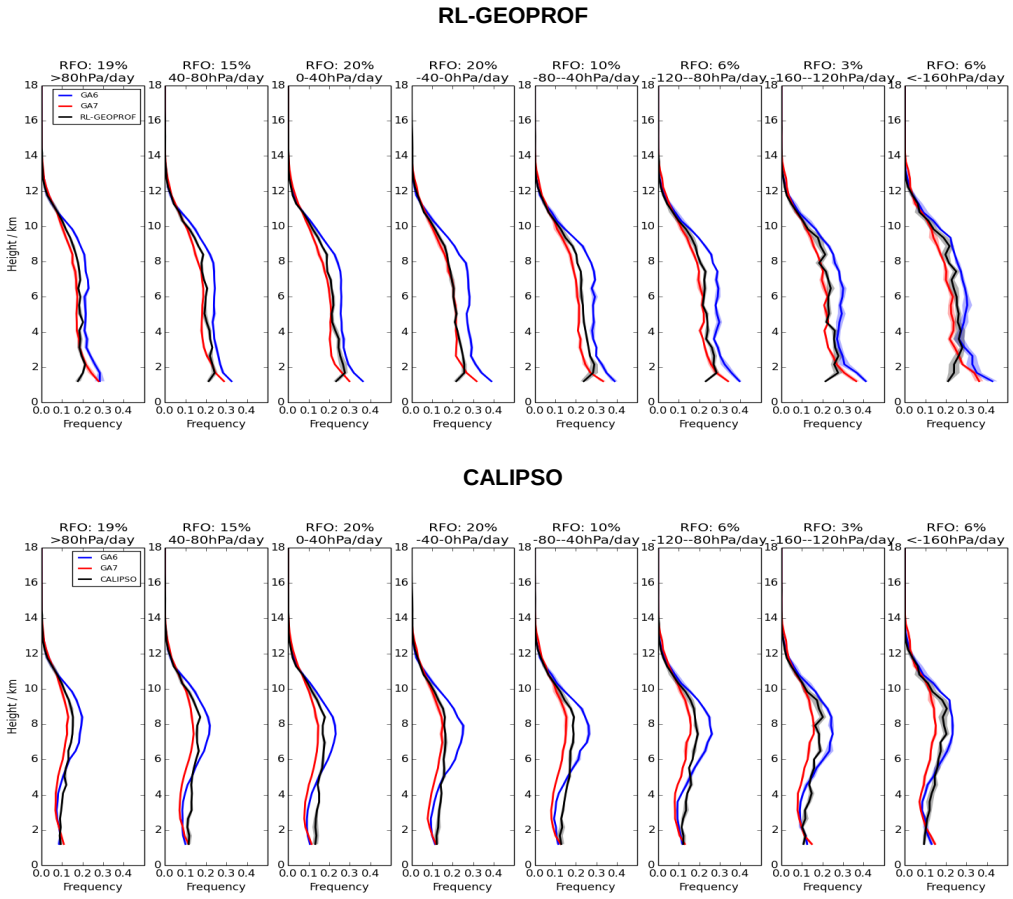

Figure 11: Observed and simulated RL-GEOPROF and CALIPSO height frequency histograms composited by daily ω_{500} over northern hemisphere land (polewards of 20^oN) during DJF. The range and relative frequency of occurrence (RFO) are shown at the top of each bin. Shading around the line plots has been added to reflect significance bounds, however this is sometimes less than the thickness of the plotted lines.

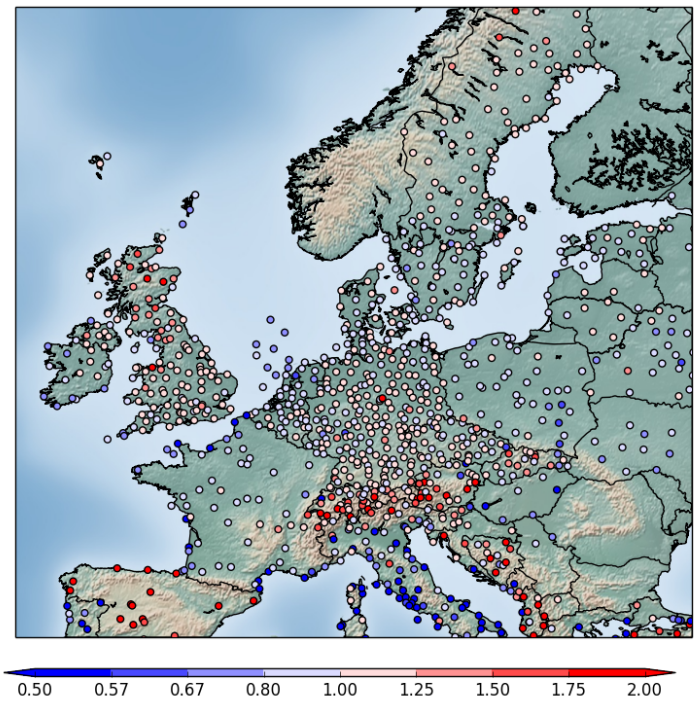

Figure 12: Frequency bias ((hits + false alarms)/(hits + misses)) of cloud base height <1km for cloud fraction $\geq$ 3 oktas in GA6 against surface station data. The mean bias of 6-hourly forecasts between 16th July 2014 and 15th July 2015 at a 24 hour forecast lead time are shown.

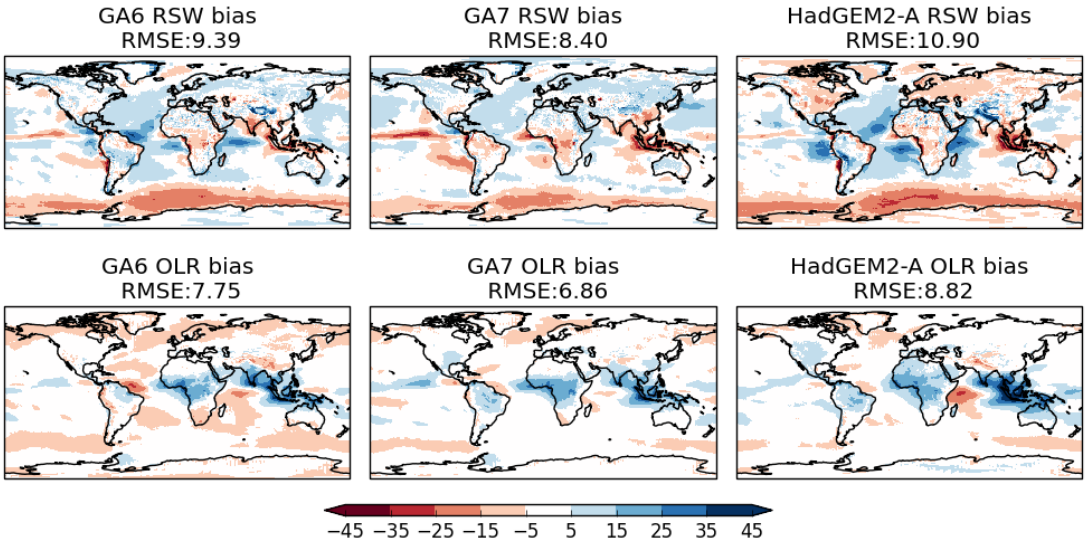

Figure 13: Multi-annual mean bias in RSW (top) and OLR (bottom) (Wm^{-2}) against CERES-EBAF for GA6, GA7 and HadGEM2-A. The spatial root-mean-square error (RMSE) is shown at the top of each panel.

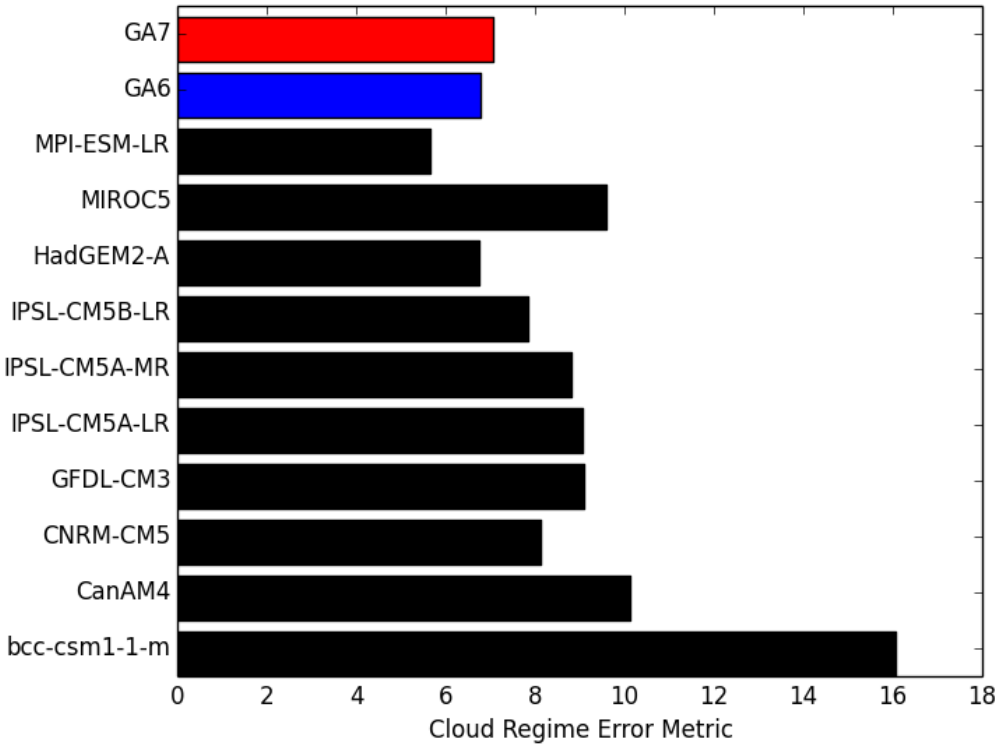

Figure 14: Cloud Regime Error Metric (CREM_{pd}) from Williams and Webb (2009) for the global cloud regimes of Tselioudis *et al* (2013) calculated for GA6 (blue), GA7 (red) and all of the CMIP5 models which have the required diagnostics available (black). Zero represents a perfect score with respect to the ISCCP observations.