Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

Geoscientific Model Development Discussions
Open Access
EGU

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

Geoscientific Model Development Discussions
Open Access
EGU

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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

Geoscientific Model Development Discussions
Open Access
EGU

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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 by applying a comprehensive cloud evaluation to 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*, 2016a) 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) are very similar to each other and to a climatological bias obtained

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

from an AMIP (Atmosphere Model Intercomparison Prroject; Gates, 1992) simulation at N96 resolution (135km). 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 documented by Walters *et al* (2016a). GA7 has recently been frozen and is documented by Walters *et al*

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

(2016b). 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* (2016b). 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 (Zhang and Klein, 2013).

2. A package of changes designed to improve warm rain microphysics, which include a change to the auto-conversion scheme to be based on Khairoutdinov and Kogan (2000), but upscaled to a GCM following Boutle *et al* (2014).

3. Improved cloud ice optical properties and ice particle size distributions following Baran *et al* (2014) and Field *et al* (2007) respectively.

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 largely as a tuning in an earlier configuration 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).

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)).

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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').

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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. 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 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. 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 Webb *et al* (2001), Chepfer *et al* (2008), Haynes *et al* (2007) for the ISCCP, CALIPSO and CloudSat simulators 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 GOCCP closer 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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

effect of a 5 km averaging length from the lidar 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 and 5km averaging lengths are 0.01 (Mace and Zhang, 2014), substantially smaller than the model biases that we detect.

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 20 year means for the AMIP simulations.

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 experi-

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

ments 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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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 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. 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 is the result of the revised numerics of the convection scheme, but the overall altitude of the thin cirrus remains too low. The low altitude 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 and simulated grid-box cloud fraction on the model grid (lower panels of Figure 3), 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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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.

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 (Figure 2d). In this case the excess hydrometeor fraction is completely removed. Examining these drizzle rates in the model, they are very low (typically $<0.005mm/hr$), 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).

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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). Although improved in GA7, the excess cirrus in this region remains.

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.

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 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 composite the observed and modelled CALIPSO cloud profile by daily $500hPa$ vertical velocity (ω_{500})

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

Geoscientific Model Development Discussions
Open Access
EGU

and SST (Figure 5). The excess cirrus in GA6 appears to be a problem across the different large scale vertical velocity regimes, with the bias being largest in regions of strongest ascent, but still present in strong subsidence. GA7 is a clear improvement, although there's now possibly 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. 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, but

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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).

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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, within the cloud head 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).

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 $d^2p/dx^2+d^2p/dy^2 < 0$. All the plots are available in the Supplementary 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 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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

whilst the cloud simulation was reasonably good, 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 in GA7, composites of the top of atmosphere (TOA) radiation biases reveal some issues (Figure 9). The outgoing longwave radiation (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 biases have a structure which is consistent with the radiation errors. Since the cloud amount errors are not large enough to contribute significantly to these SW errors, we suggest that microphysical processes are primarily responsible through incorrect cloud albedos. Although the subject of ongoing research, we believe that the bias for cloud albedos on the cloud-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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

interaction with the clouds.

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 configurations at around 1km, however the CALIPSO profiles suggest the cloud fractions at this level are generally correct. This 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 excess precipitation. 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.

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

399 frequently be attenuated before detecting the lowest cloud layers. Accurate predictions of
400 cloud near the surface are of the highest importance for a number of users of the model,
401 especially aviation. Here we use SYNOP data which are likely to be the most reliable
402 observation type available for this lowest layer whilst having a reasonable global coverage
403 over land. By looking at the lowest 1km, many of the issues associated with the SYNOP
404 data (combining human and automated data and differing observational errors associated
405 with each) may be minimised. In order to confine the analysis to cloud with bases below
406 1km, we use the cloud base height observation and look at frequency of occurrence of cloud
407 bases below 1km. The cloud base height is defined as the height of cloud with coverage of
408 3 oktas or more, hence instances of small cloud coverage are excluded from this analysis.
409 As a consequence, significant model biases in this diagnostic can appear if the observed
410 cloud amount is typically just over 3 oktas and the model cloud fraction is just under (or
411 vice-versa). This appears to be an issue for the UM in parts of the tropics (not shown),
412 however more generally the diagnostic is reflecting errors in the frequency of occurrence
413 of low-base cloud. Based on comparison with the active instruments at higher altitudes,
414 we suspect that biases are more often reflecting errors in the frequency of occurrence of
415 low cloud rather than errors in the cloud base height on any one occasion.

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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 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 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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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 lack of RSW over the Southern Ocean 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 reduc-

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

tion 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.

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

Geoscientific Model Development Discussions
Open Access
EGU

References

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.

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.

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.

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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

544 ——, ——, ——, 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.

545 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.

546 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.

547 ——, 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.

548 ——, 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.

549 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.

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

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

552 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.

553 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.

554 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.

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

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

557 ——, 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.

558 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.

559 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**(), 9429–9444. doi:10.1175/JCLI-D-12-00207.1.

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

561 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.

562 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.

563 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.

564 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.

565 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.

566 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.

567 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.

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

569 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

570 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.

571 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.

572 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, 2016a: 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.

573 Walters, D. N., M. E. Brooks, D. Copsey, and P. E. Earnshaw, 2016b: The Met Office Unified Model Global Atmosphere 7.0 and JULES Global Land 6.0 configurations. *In preparation.*

574 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.

575 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.

576 ——, 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

577 ——, 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.

578 ——, 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.

579 ——, ——, ——, 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.

580 ——, 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.

581 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.

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

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

584 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

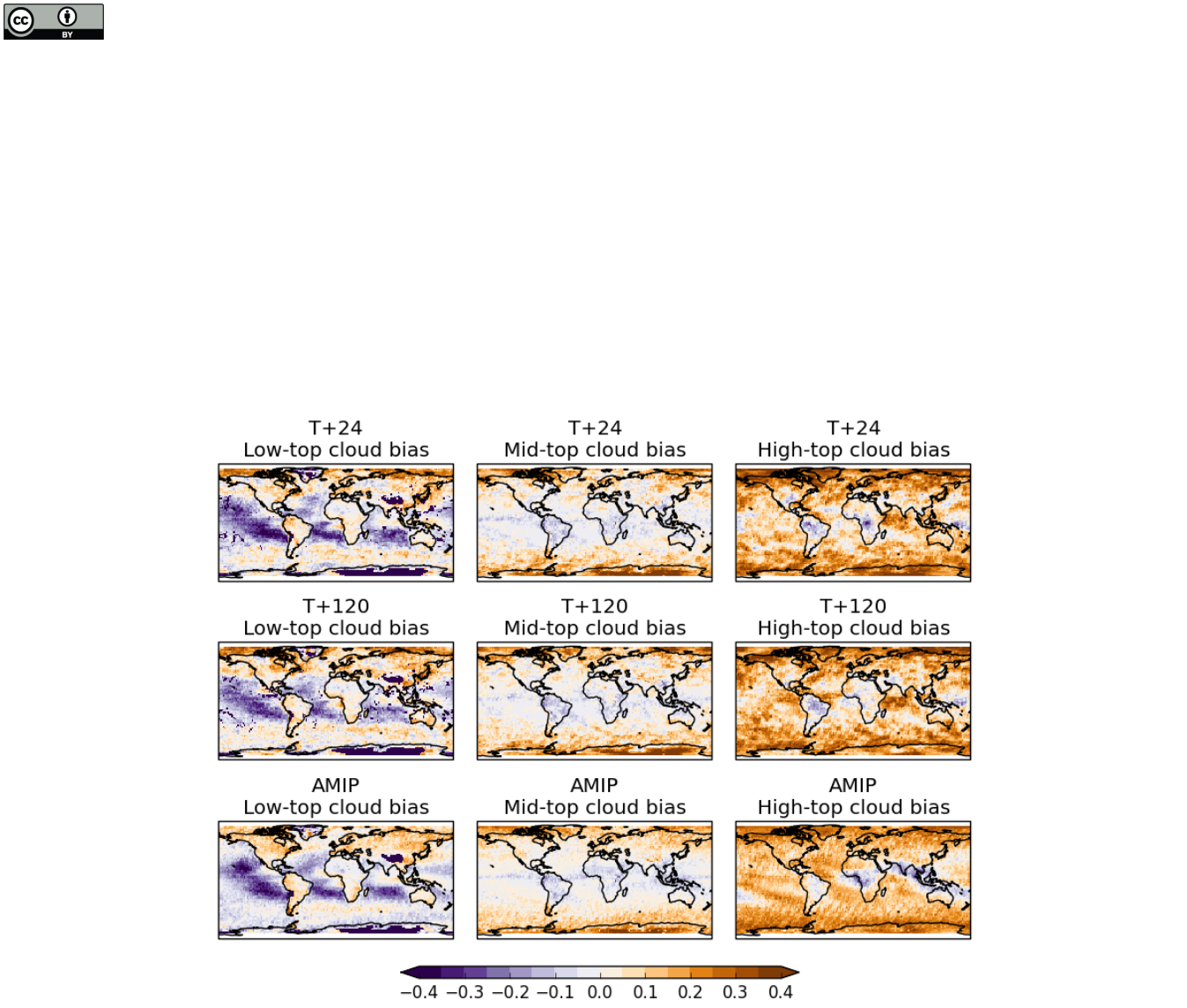

Figure 1: Absolute bias (model field minus observed field) in GA6 configuration of the UM for low (left) mid (centre) and high (right) fractional cloud cover against the GCM Orientated CALIPSO Cloud Product (GOCCP), using the CALIPSO simulator in COSP (see Section b. Top and middle rows are mean biases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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

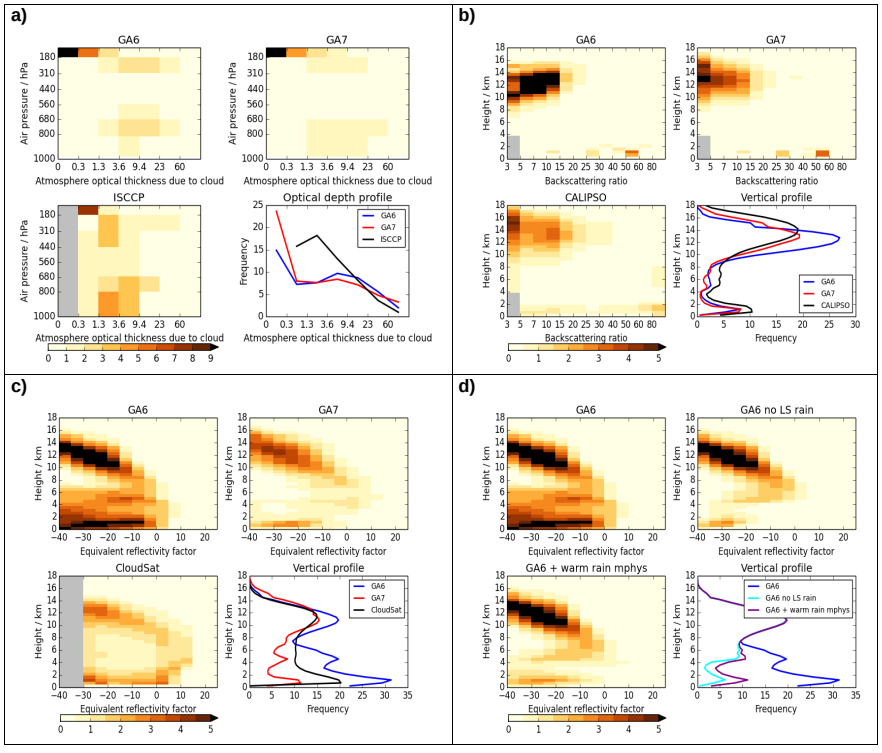

Figure 2: Tropical multi-annual mean observed and GA6 & GA7 simulated satellite data summaries. a) ISCCP cloud-top pressure–optical depth joint frequency histograms. Lower right panel is a single optical depth frequency histogram (i.e. the joint histograms have been summed across cloud top pressure bins). 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. 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

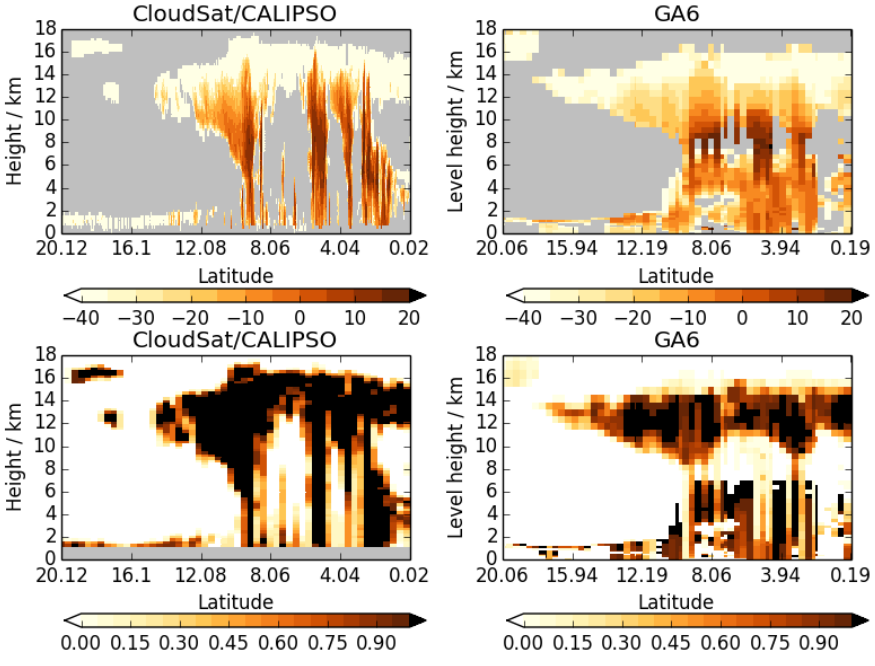

Figure 3: Case study of a GA6 6 hour forecast verifying at 18:00UTC on 17th December 2010 for an A-train pass over the South China Sea. Top: the observed and simulated radar reflectivities (dBZ) with situations in which the lidar detected cloud but the radar did not being included with a nominal value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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

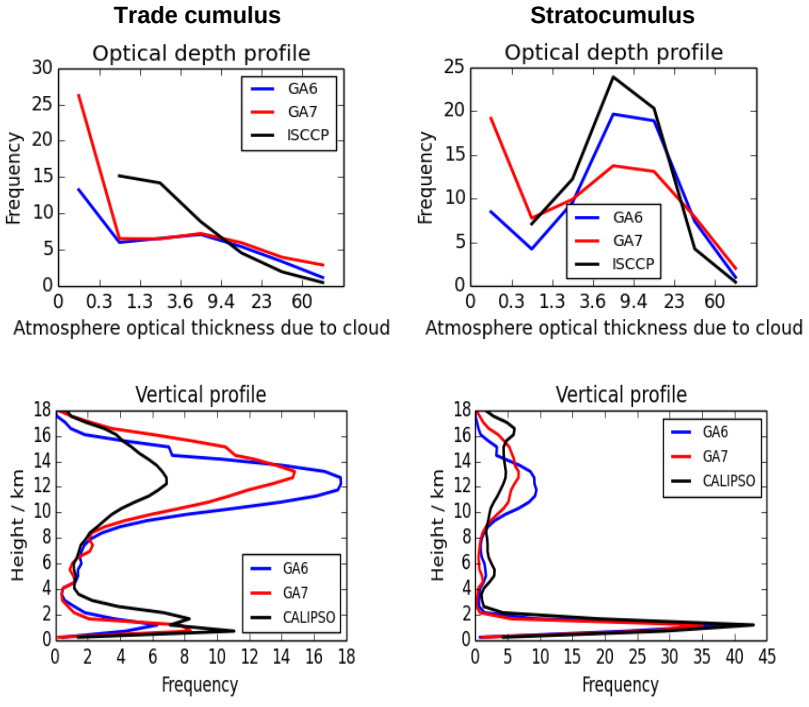

Figure 4: Observed and simulated multi-annual mean ISCCP optical depth frequency histograms (top) and CALIPSO height frequency histograms (bottom) for a trade cumulus region (130-160^{o}W, 0-20^{o}S, left) and stratocumulus region (80-90^{o}W, 0-20^{o}S, right).

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

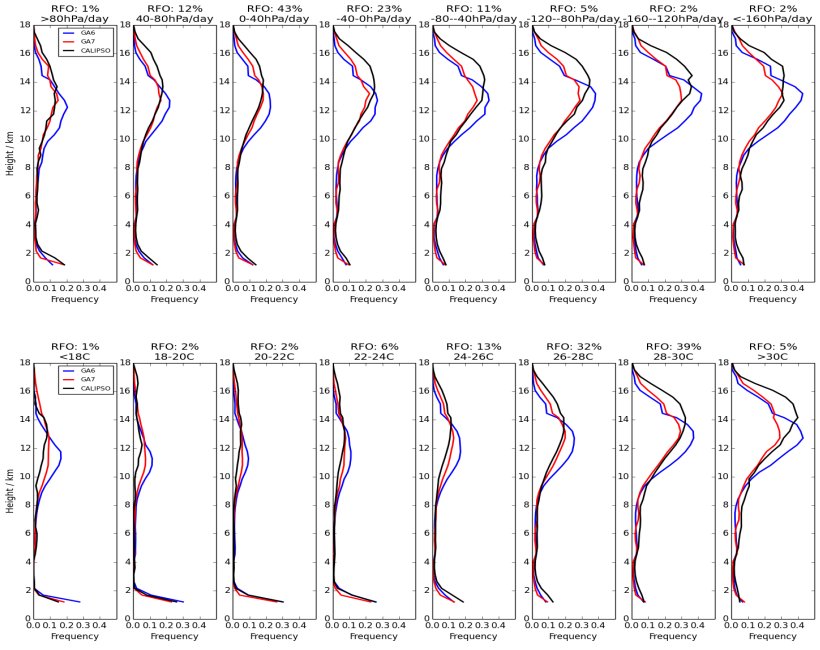

Figure 5: Observed and simulated CALIPSO height frequency histograms composited by daily ω_{500} (top) and SST (bottom) over the tropics (20°N–20°S). The range and relative frequency of occurrence (RFO) are shown at the top of each bin. Negative ω_{500} indicates ascent.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

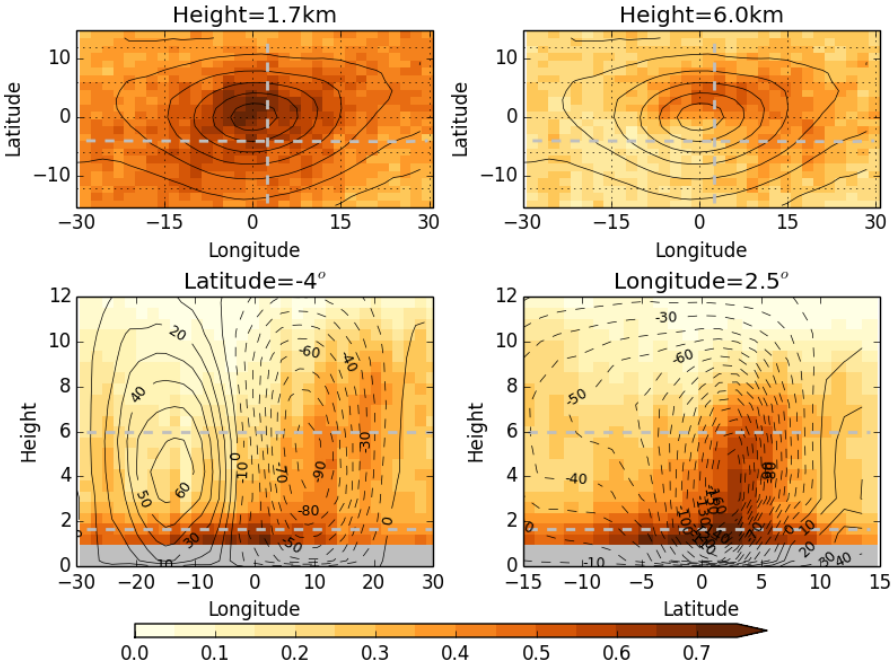

Figure 6: Distribution of average observed hydrometeor (cloud plus precipitation) fraction (colours) around a composite of ERA-I cyclones over northern hemisphere oceans for 5 years of DJF daily data. Top row shows horizontal sections through the composite cyclone at 1.7 & 6km with the mean PMSL contoured 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).

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

Geoscientific Model Development Discussions
Open Access
EGU

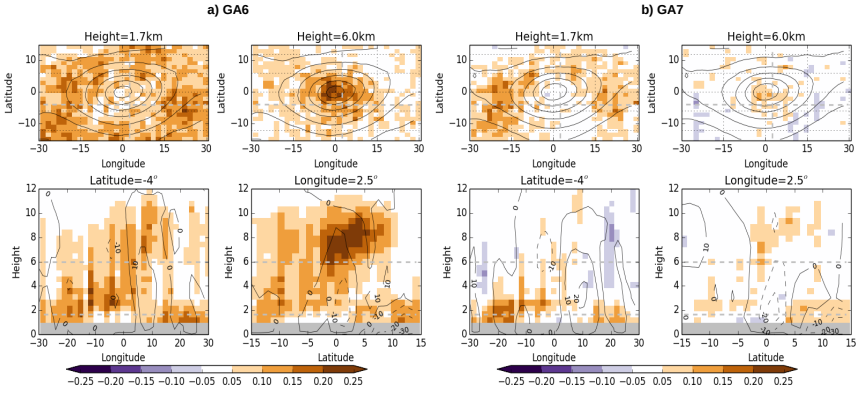

Figure 7: Cloud fraction absolute bias (model field minus observed field) (colours) for composite cyclones. Produced as per Figure 6 for a) GA6 b) GA7 and the observed composite then subtracted. Black contours in top plots are the model mean PMSL and in the lower plots are the bias in vertical velocity.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

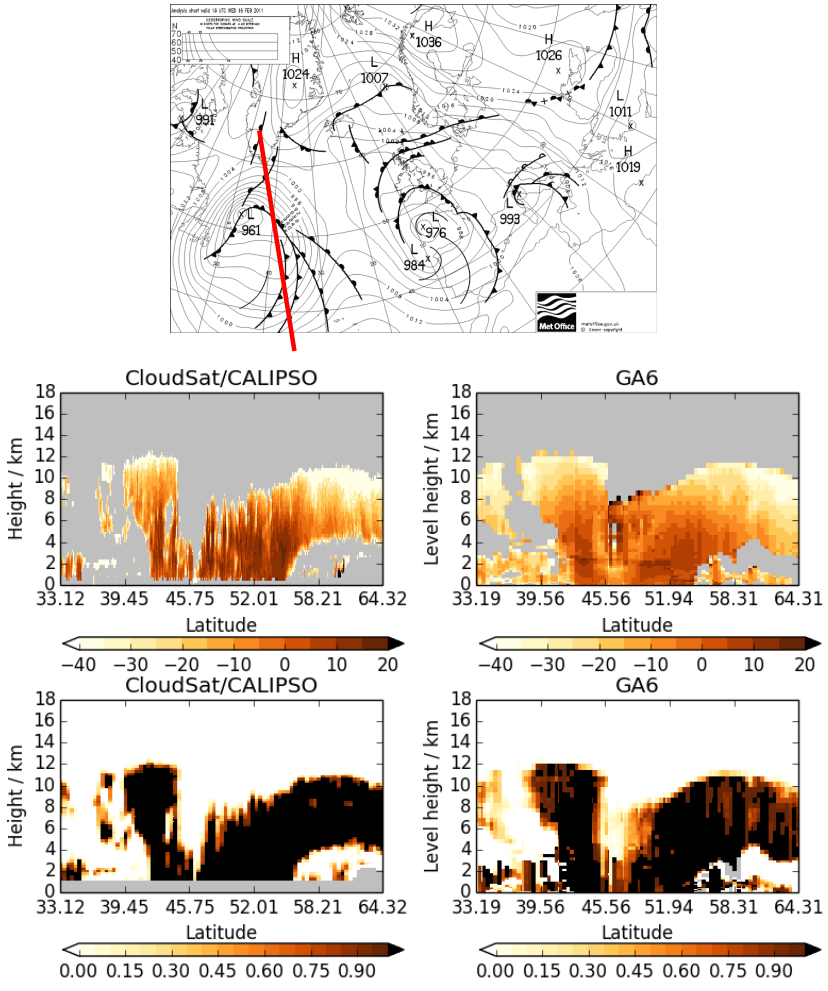

Figure 8: Case study of a GA6 27 hour forecast verifying at 15:00UTC on 16th February 2011 for an A-train pass over the North Atlantic as shown by the red line on the synoptic analysis. Top: the observed and simulated radar reflectivities (dBZ) with situations in which the lidar detected cloud but the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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

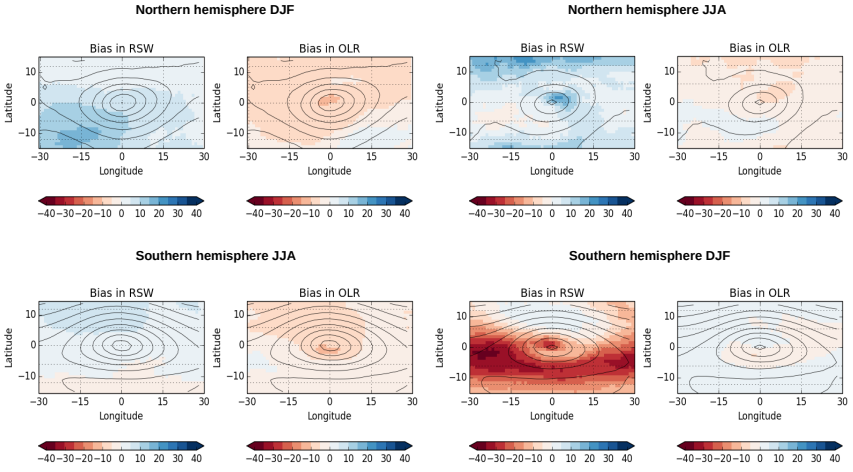

Figure 9: Cyclone composite GA7 mean bias in RSW and OLR (Wm^{-2}) against CERES-EBAF (colours). Black contours are GA7 PMSL. Northern and Southern hemisphere composites are shown for the respective winter (left) and summer (right) seasons.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

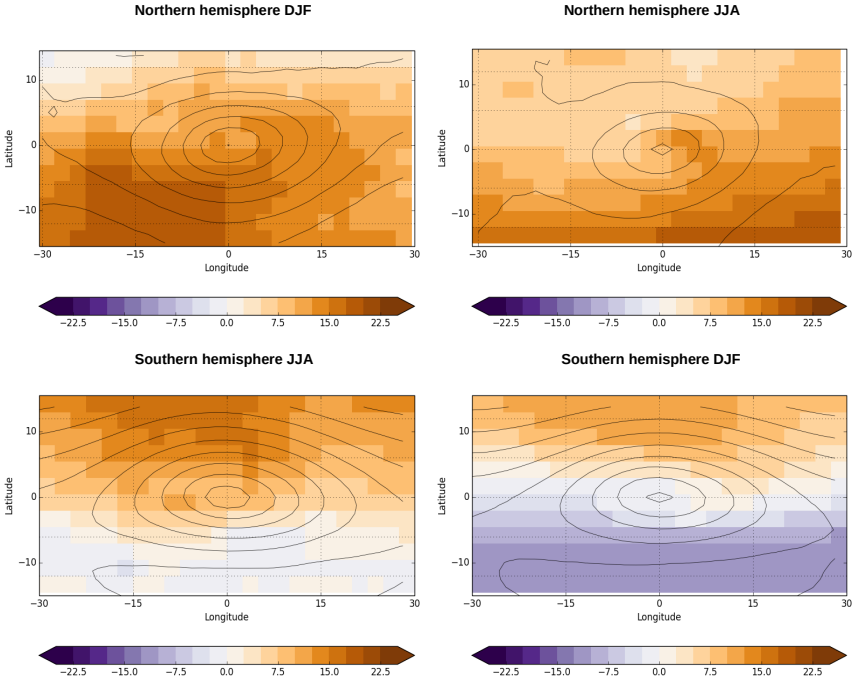

Figure 10: Cyclone composite GA7 mean bias in in-cloud albedo (%) against ISCCP (colours). Black contours are GA7 PMSL.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

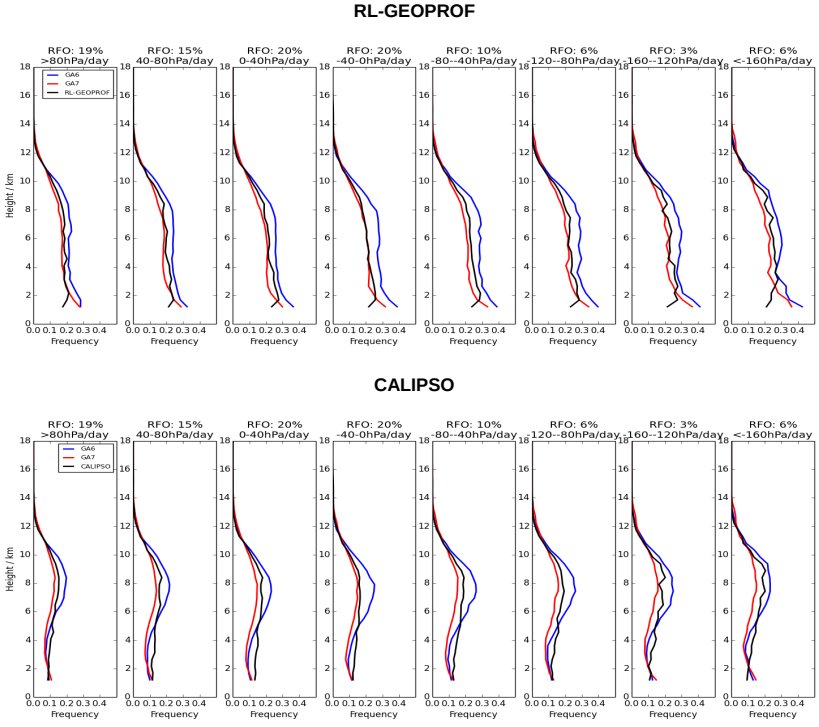

Figure 11: Observed and simulated RL-GEOPROF and CALIPSO height frequency histograms composited by daily ω_{500} over northern hemisphere land (polewards of 20^oN) during DJF. The range and relative frequency of occurrence (RFO) are shown at the top of each bin.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

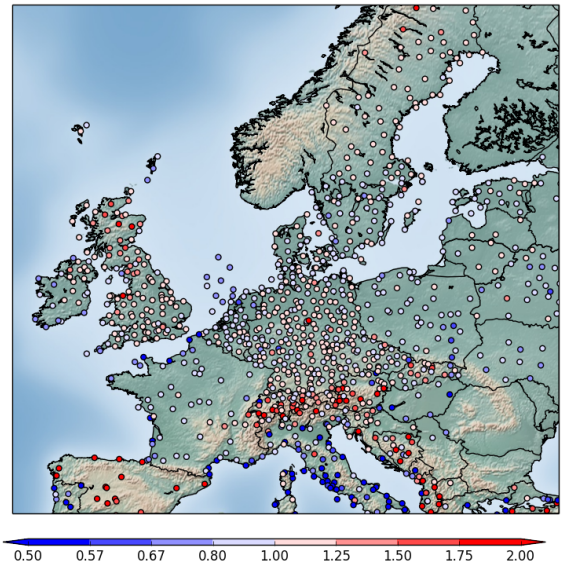

Figure 12: Frequency bias ((hits + false alarms)/(hits + misses)) of cloud base height <1km for cloud fraction ≥ 3 oktas in GA6 against surface station data. The mean bias of 6-hourly forecasts between 16th July 2014 and 15th July 2015 at a 24 hour forecast lead time are shown.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

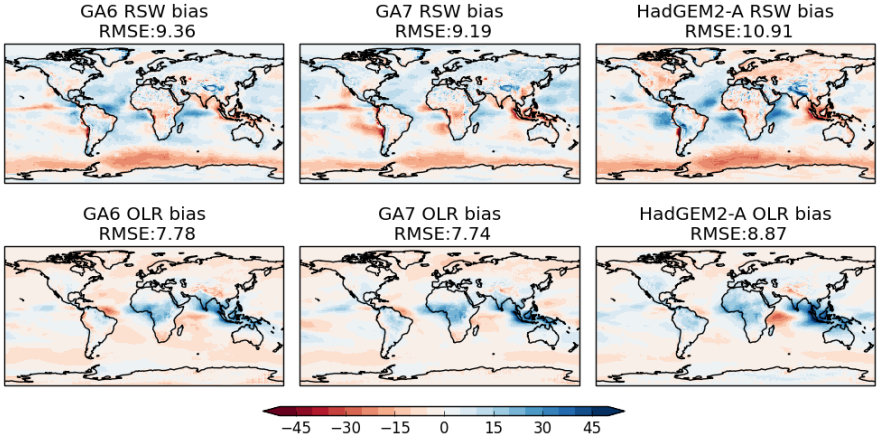

Figure 13: Multi-annual mean bias in RSW (top) and OLR (bottom) (Wm^{-2}) against CERES-EBAF for GA6, GA7 and HadGEM2-A. The spatial root-mean-square error (RMSE) is shown at the top of each panel.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-295, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 3 January 2017

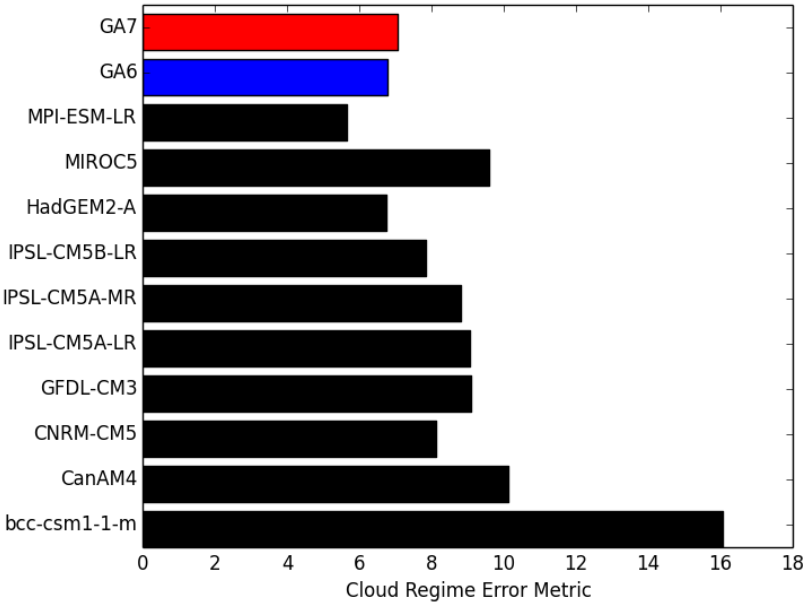

Figure 14: Cloud Regime Error Metric ($CREM_{pd}$) from Williams and Webb (2009) for the global cloud regimes of Tselioudis *et al* (2013) calculated for GA6 (blue), GA7 (red) and all of the CMIP5 models which have the required diagnostics available (black). Zero represents a perfect score with respect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