RMetS
Royal Meteorological Society

Evaluation of intercomparisons of four different types of model simulating TWP-ICE

Jon Petch,[a*] Adrian Hill,[a] Laura Davies,[b] Ann Fridlind,[c] Christian Jakob,[d] Yanluan Lin,[e]
Shaoecheng Xie[f] and Ping Zhu[g]

[a] *Met Office, Exeter, UK*
[b] *University of Melbourne, Victoria, Australia*
[c] *NASA Goddard Institute for Space Studies, New York, NY, USA*
[d] *School of Mathematics, Monash University, Melbourne, Victoria, Australia*
[e] *Center for Earth System Science, Tsinghua University, Beijing, China*
[f] *Lawrence Livermore National Laboratory, CA, USA*
[g] *Department of Earth and Environment, Florida International University, Miami, FL, USA*

*Correspondence to: J. C. Petch, Met Office, FitzRoy Road, Exeter EX1 3PB, UK. E-mail: jon.petch@metoffice.gov.uk

Four model intercomparisons were run and evaluated using the TWP-ICE field campaign, each involving different types of atmospheric model. Here we highlight what can be learnt from having single-column model (SCM), cloud-resolving model (CRM), global atmosphere model (GAM) and limited-area model (LAM) intercomparisons all based around the same field campaign. We also make recommendations for anyone planning further large multi-model intercomparisons to ensure they are of maximum value to the model development community. CRMs tended to match observations better than other model types, although there were exceptions such as outgoing long-wave radiation. All SCMs grew large temperature and moisture biases and performed worse than other model types for many diagnostics. The GAMs produced a delayed and significantly reduced peak in domain-average rain rate when compared to the observations. While it was shown that this was in part due to the analysis used to drive these models, the LAMs were also driven by this analysis and did not have the problem to the same extent. Based on differences between the models with parametrized convection (SCMs and GAMs) and those without (CRMs and LAMs), we speculate that that having explicit convection helps to constrain liquid water whereas the ice contents are controlled more by the representation of the microphysics.

Key Words: convection; microphysics; numerical modelling

Received 7 January 2013; Revised 7 May 2013; Accepted 17 May 2013; Published online in Wiley Online Library 18 July 2013

1. Introduction

Weather and climate prediction relies on numerical models designed to represent our best understanding of the relevant components of the Earth system. One critical component of both weather and climate prediction systems is the representation of the atmospheric processes, both dynamical and physical. Global atmospheric models (GAMs) represent the whole globe and generally use coarse grid lengths which rely on the representation (or parametrization) of many physical processes whose scales are sub-grid, with convection and clouds being a key example. Regional or limited-area models (LAMs) are a key tool for weather prediction, and are increasingly used in climate research and prediction to dynamically downscale global climate predictions to add better understanding of the regional impacts of climate change (e.g. Kendon *et al.*, 2010). LAMs are an attractive tool because they cover smaller regions and are thus able to use smaller grid lengths for the same computational costs, and this allows a more explicit representation of the local orography as well as convective processes and the associated cloud.

The continuous development and improvement of atmospheric models is of critical importance to the weather and climate community (Randall *et al.*, 2003). Two further modelling systems which are valuable tools supporting the development of LAMs and GAMs are cloud-resolving models (CRMs) and single-column model (SCM) versions of the GAMs (Randall *et al.*, 1996). CRMs are very similar to LAMs in that they also represent a limited area and utilise shorter grid lengths to explicitly resolve key processes. They differ in that they have generally been developed to understand the physical processes of the atmosphere rather than as a prediction system. They are also often run at much higher resolution than LAMs and include more complex and computationally expensive representations of the physical processes such as microphysics. In this article, as is quite often the case, the CRMs also differ from the LAMs in the way they are forced. The LAMs include the real land surface boundaries and use open boundary conditions provided by analysis of global forecasts. In contrast, the CRMs use a uniform ocean surface, employ cyclic boundary conditions and are driven by a uniform forcing consistent with the SCMs. SCMs, while they have their limitations, allow us to isolate the behaviour of GAM parametrizations from dynamical feedbacks and also prove a computationally efficient method for quickly evaluating parametrization changes (e.g. Randall *et al.*, 2003). In this work, they also use a uniform ocean and are essentially driven in the same way as the CRMs.

A framework to test all the models types discussed above and confront these with observations is the intercomparison. An intercomparison is where various models of the same type are run for the same case and their results compared. The benefits of this collaborative activity to model developers go beyond the ability to compare their model with many other models and identify their key deficiencies (as discussed in Petch *et al.*, 2006) because they also bring the community together to jointly discuss and tackle key challenges in model development. The GEWEX (Global Energy and Water Exchanges) project Global Atmospheric System Studies (GASS) acknowledge the importance of these activities and thus focus much of their work on coordinating these activities. This article describes the lessons we can learn from bringing together the results of four model intercomparisons involving GAMs, LAMs, CRMs and SCMs.

The intercomparisons were all based around the Tropical Warm Pool–International Cloud Experiment (TWP-ICE) which took place in and around Darwin, Australia, from 20 January to 13 February 2006. Its focus was to describe the evolution of tropical convection, including the large-scale heat, moisture, and momentum budgets at 3 h time resolution, while at the same time obtaining detailed observations of cloud properties and the impact of the clouds on their environment (May *et al.*, 2008). A field campaign of this kind provides an ideal test-bed for driving and evaluating a range of atmospheric models used in weather and climate research and prediction.

Under the umbrella of GASS and with the support of the US Department of Energy (DOE) Atmospheric System Research (ASR) program, observations made during the TWP-ICE campaign have been used to drive and evaluate multiple models of four different types. The resulting collaboration and articles describing model intercomparisons provide an important reference for various institutions to carry out further experiments which support their model development processes. The articles describing their comparisons using the TWP-ICE data are:

- CRMs: Cloud-resolving models (Fridlind *et al.*, 2012)
- LAMs: Limited-area models used in regional weather and climate prediction (Zhu *et al.*, 2012)
- GAMs: Global atmospheric models for predicting on weather or climate time-scales (Lin *et al.*, 2012)
- SCMs: Single-column models (Davies *et al.*, 2013).

In Fridlind *et al.* (2012), observations made during TWP-ICE were used to perform the most comprehensive evaluation of a cloud-resolving model intercomparison that has ever been carried out. The ability to challenge the models with such a range of observations, particularly those which describe the variability within the CRM domains, highlighted many challenges for both CRM development and for designing the frameworks in which the CRMs are run. Specific conclusions from the article noted that there was a wide spread in the prediction of cloud stratiform fraction and that all models systematically overestimated the areas with strong convective mean precipitation. While precipitation was constrained by the forcing, it was clear that the CRMs differed significantly in their prediction of the precipitation distribution. It was noted there was a large spread in predicted ice water path and, compared to observational estimates, it was overestimated in all models apart from those which were run in two dimensions. However, the existing estimates of uncertainty in ice water path retrievals also require further evaluation, as discussed by Fridlind *et al.* (2012).

Zhu *et al.* (2012) presented the first comparison of convective-scale LAMs carried out within GASS. The models produced realistic large-scale thermodynamic fields when compared to observations, although the locations of precipitation within the domains varied. As with the CRMs, ice water paths differed by large amounts between models. Stratiform cloud fractions showed large spread, especially high ice anvils, which can have large impacts on the radiative properties of the cloud systems. Both the water contents and the ice cloud fractions were seen to vary significantly between models.

In Lin *et al.* (2012), GAMs were compared over the TWP-ICE region and, while the models all captured the large-scale precipitation event seen in the observations, it was delayed by over a day. As with the CRMs, ice water contents had a very large spread (more than an order of magnitude) but it was also clear that in GAMs there was a large spread in liquid water paths. There were enough models involved in the comparison to identify that the models whose convection schemes were more responsive to mid-level moisture performed better during the less active periods.

Davies *et al.* (2013) described the first SCM intercomparison to use ensemble forcing which represented the observational uncertainty and provided a sensitivity study for the SCMs involved. It also included a single 2D and 3D CRM as a reference for the SCMs using the ensemble forcing. It was shown that, while the ensemble mean generally behaved like a single realisation using the mean forcing, this was not the case for all diagnostics or all models. The ensemble forcing was also shown to be of particular value for investigating how different models respond to changes in the forcing.

While the four articles described above each make conclusions relevant to evaluation and improvement of the individual model classes they address, the archive of all the modelling results and observations is also a key output of this project. Individual modelling centres can begin to make use of this for their model development work. The availability of different model types for this case makes this an even more valuable resource. For example, the ability of a weather or climate modelling centre to carry out sensitivity studies using both their SCM and GCM and place this into context by comparing against other models driven in the same way is very valuable. Petch *et al.* (2007) described a comparison using a single GCM, SCM and CRM driven and evaluated using observations made during the Tropical Ocean–Global Atmosphere Coupled Ocean–Atmosphere Response Experiment (TOGA-COARE) as a preparation for an intercomparison involving these model types. However, this work lacked the unique opportunity of having a completed intercomparison to allow us to study the spread and variation of the models. In this ‘multi-model type intercomparison’, we bring together the key findings of each of the separate modelling studies and diagnose some cross-model differences to learn more about both the models themselves and the experimental framework used.

This article analyses the results of all models used in the TWP-ICE intercomparison to identify what can be learnt from a multi-model type intercomparison. It also identifies and documents the key lessons learnt during this project which should help the planning of any similar work in the future. Section 2 describes the experimental frameworks and models used in these studies and how they differed for the different model types. Section 3 analyses the results of all model types and in section 4 we summarise our findings and make recommendations which should improve any further similar multi-model intercomparisons.

2. Experimental framework

The analysis of all the modelling carried out in this cross-comparison was based around the TWP-ICE field campaign.

Table 1. A summary of the models used in the four separate intercomparison articles.

Model type Reference	LES/CRM Fridlind *et al.* (2012)	LAM Zhu *et al.* (2012)	Global Lin *et al.* (2012)	SCM Davies *et al.* (2013)
Number of models	10	6	9	9
Horizontal domain size	200–300 km^2	400–500 km^2	Global	One column
Analysis area	Domain	Average of grid boxes overlapping with the TWP-ICE variational analysis domain	Average of grid boxes overlapping with the TWP-ICE variational analysis domain	One grid box
Horizontal grid length (km)	0.9–3	1–3	20–250	25–200
Vertical grid length (km) around 500 mb	0.18–0.6	0.3–0.5	0.3–1.0	0.3–1.0
Forecast lead time analysed	Free running for whole period	12–36 h	24–48 h	Free running for whole period
Forcing	Variational analysis	Nested in global models driven by EC analysis	ECMWF analysis	variational analysis
Deep convection	Explicit	Explicit	Parametrized	Parametrized
Shallow convection	Explicit	Mix of BL, shallow schemes numerical/explicit	Parametrized	Parametrized
Cloud fraction scheme	All or nothing	Some all or nothing, some parametrized	Parametrized	Parametrized

The periods simulated were designed to be a balance between covering a broad range of conditions at the site and the increasing computational costs of longer runs. As described in May *et al.* (2008), the Darwin region was influenced by a typical monsoonal circulation during TWP-ICE. It experienced three distinct regimes: active monsoon during 20 to 25 January, suppressed monsoon during 26 January to 3 February, and a monsoon break period during 3 to 13 February 2006. The active monsoon period was characterized by westerly monsoon flow, intensive mesoscale convective systems of mostly oceanic origin, and heavy surface precipitation. During the suppressed monsoon period, clouds were primarily associated with relatively shallow convection accompanied by much lower surface precipitation than in the preceding monsoon period. The break monsoon period was featured by intense afternoon thunderstorms with several squall lines crossing Darwin in the evening and early morning. Due to the high computational cost, the CRM study focused only on the active and suppressed periods and the LAMs were run only for the period 1200 UTC on 22 to 0000 UTC on 26 January. In contrast, both the SCMs and GAMs were run for the entire TWP-ICE period from 0000 UTC on 20 January to 0000 UTC on 13 February.

2.1. The models used

A brief overview of the basic properties of all the models used and the runs carried out in the four intercomparisons are summarised in Table 1. As with any intercomparison project, there is a need to be pragmatic when specifying the design of the experiments. A balance is needed between constraining the components of the experiment such that the comparison is as clean as possible and the time it takes for the participants to adhere to any requirements. This is reflected in the ranges seen in properties such as the details of how the models are forced or how properties such as the land surface are initialised. While there would be benefit in having these set the same across all models, it is often more useful for centres to carry out the experiments with settings relevant to their typical use.

Table 1 shows that there are generally only a small number of models of each type. LAMs in particular only had six different models of which three were variants of WRF. The extent to which we should consider the variants of WRF as different models is quite subjective and through this article we will generally treat all model submissions equally. Full details of all the individual models used are available in the individual comparison articles and these will not be discussed here in any detail.

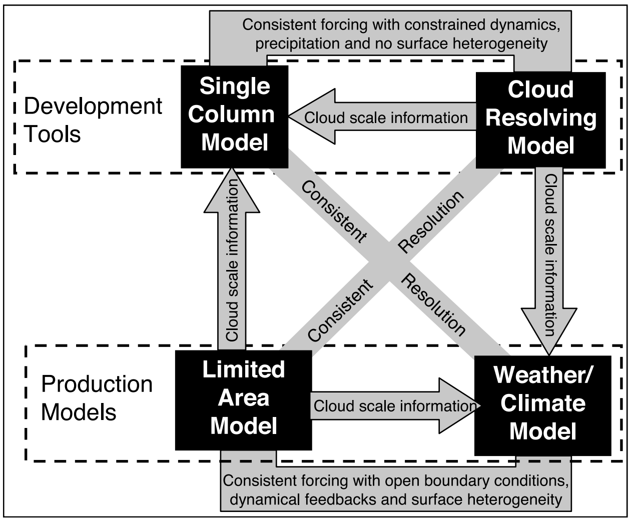

Figure 1. Schematic of the models used in the various comparisons with their relationships. Note that observations can be used in all components. CRMs and LAMs have resolved convection so provide additional information to convective parametrized models. SCMs are a tool to isolate climate model physics and keep consistent forcing.

It is useful to consider the purpose of each model type in this cross-comparison article. Figure 1 depicts some of the key similarities and differences in the models used in this study, along with the way in which they are forced. The CRMs and SCMs are described here as development and research tools, i.e. they are not generally used to make operational predictions of weather or climate but more to understand atmospheric processes and support the development of other models. The CRMs can provide realistic cloud-scale motions to help diagnose how the processes should be parametrized in the larger-scale models. The SCMs allow us to isolate the model physics from the dynamics in a computationally inexpensive tool; in Davies *et al.* (2013), this efficiency has been exploited by carrying out an ensemble of forcing to learn about the behaviour of the physics in the SCMs as a function of the mean state and forcing. While there may be some exceptions, the GAMs and LAMs can be considered operational models as they are used to make predictions of weather and climate, often operationally as part of a national weather or climate service. It is the combination of all these model types which allows us to draw further general conclusions from this study. We are also able to identify key issues we should address when we organise future model intercomparisons involving SCMs, CRMs, GAMs and LAMs.

2.2. *Comparing the different forcing and boundary conditions*

Variational analysis is used to derive the domain-mean large-scale forcing dataset used to drive the SCMs and CRMs. The forcing data have a 10 mb vertical resolution and 3 h temporal resolution (centred in time) and were created using a combination of observations and the ECMWF analysis (Xie *et al.*, 2010). In both the SCM and CRM integrations, the models were initialised only once. The SCMs were then integrated for the entire length of the experiment, with no nudging towards observed profiles. In contrast, as described in Fridlind *et al.* (2012), the CRMs were free running below 16 km, while above 16 km the vapour and temperature profiles were nudged towards observed profiles. Thus, both the CRMs and SCMs simulations of the troposphere were free running for the entire period. There are pros and cons of running the entire period, but free runs are common in CRM and SCM intercomparisons (e.g. Xie *et al.*, 2002; Xu *et al.*, 2002). The SCM article (Davies *et al.*, 2013) focuses on the results from using an ensemble of forcing created using variational analysis with the precipitation perturbed within the observational uncertainties. However, in this article we focus on the deterministic forcing used as the basis for the CRM comparison; this forcing was also used in the SCM article for comparison to the ensemble forcing and was shown to give similar answers to the ensemble mean for most diagnostics.

A notable difference between the deterministic SCM forcing and the CRM forcing was the way in which the variational analysis was used. The horizontal advection term (which is generally much smaller than the vertical) is the same in both methods. However, for the SCM comparison, the vertical velocity from the analysis was used with the model predicted thermodynamic fields to give the vertical advection term to drive the model. In contrast, the CRM comparison used both the thermodynamic fields and vertical velocities from the analysis to derive vertical advection tendencies. Ghan *et al.* (2001) compared these two methods of forcing during the SCM comparison of midlatitude continental convection and it was concluded that there was no systematic dependence on the forcing method employed. Later, however, we will demonstrate that the different forcing method employed in the CRM and SCM intercomparison leads to a difference in the thermodynamic profile and is therefore a limitation of this cross-model comparison and a difference which should be avoided in future work. Hereafter we will refer to these methods as SCM forcing and CRM forcing, although we note that either forcing method can be applied to both SCMs and CRMs. A further difference between the SCM and CRM forcing was that the CRMs were transitioned from free-running below 15 km to nudging of model domain-mean water vapour and potential temperature towards observed domain-mean conditions with a 6 h time-scale above 16 km (specification and discussion in Fridlind *et al.* 2012). This was a pragmatic decision to better maintain a tropopause layer structure consistent with observations while not influencing total surface precipitation relative to a fully free-running simulation. This difference was much less significant for the results discussed in this article.

The GAMs wind, temperature, moisture, and surface pressure were initialized at 0000 UTC daily from the ECMWF operational analysis using the Cloud-Associated Parametrizations Testbed (CAPT) approach (Phillips *et al.*, 2004). Other fields, such as land surface properties (vegetation, soil moisture and temperature) were constrained to be as realistic as possible using a range of methods in the various models. The methods were chosen by the modelling centres themselves as the option they considered the best choice for their modelling system. The second day (24–48 h) of the forecasts was used for the comparisons made in this article. This was chosen to allow spin-up of the models but still keeping the large-scale dynamics close to the analysis. The different LAMs were constrained in somewhat different ways for the dynamics, thermodynamics and other fields such as land surface properties. This was an example where the various centres have typical or operational methods of driving their models and it was reasonable that they were driven using their own methods. However, the wind, temperature, moisture, and surface pressure in all models were essentially initialised and driven at the boundaries either by an analysis or by a short-range forecast initialised by the same ECMWF analysis that was used by the GAMs.

To understand the role of the different forcing from these experiments, it is useful to compare the ECMWF analysis to the variational analysis. Figure 2 shows vertical velocity from the ECMWF analysis (driving the LAMs and GAMs) and the variational analysis (driving the CRMs and SCMs). A key difference between the two forcings is that the strong upward motion in the wet period around 23–25 January has quite a different timing with the negative peak between 23 and 24 January for the variational analysis, but this occurs a whole day later in the ECMWF analysis. The mean vertical velocity from the wet and dry periods (defined in Figure 3(a)) shows that the upward velocities are typically stronger in the ECMWF analysis than they are in the variational analysis. In the dry period, there is also weak ascent in the ECMWF analysis while there is descent in the variational analysis. The implications of these mean and timing differences will be discussed during the evaluation of results from the different models in the next section.

3. Cross-model comparison

It is a challenge to show the vast number of results from various models in a simple set of plots. To do this we mainly focus on showing a selection of basic fields as the mean of each model type and the spread of those models. In plots used here we have focused on the use of the mean and standard deviation between models to describe the spread. It should be stressed that this potentially hides a great deal of information and can lead to somewhat misleading conclusions if more detailed analysis is not carried out. This is particularly true because of the relatively small sample sizes of models of each type (ten CRMs, six LAMs, nine SCMs and nine GAMs). Later we show the specific implication of the use of means and standard deviation for presenting the data.

By combining all models, we are removing the option for an individual centre to identify its own model and development needs. However, this level of detail would not be appropriate for an overview article and also near-impossible to present clearly. Instead we can highlight some basic interesting features from the models as a whole and thus allow any centre to carry out further work needed to review the performance of its model and any required sensitivity studies.

As shown in Table 1, different types of model were run over different sizes of experimental domain. To make an appropriate comparison with the observations and between different types of model, simulations from these four model types were all averaged over the variational analysis domain, which is the same as the TWP-ICE pentagonal sounding array. It should be noted that the actual domain size represented by the GAMs is slightly larger than the sounding array due to the coarse resolutions used in the GAMs. In addition, since different GAMs were run with different horizontal resolutions, model grid points used in the average vary from two for the coarsest model to over one hundred for models at 20 km resolution. As indicated in Lin *et al.* (2012), these differences are small compared to the variations between these GAMs. So we do not expect that they have large impact on our analysis.

3.1. *Evolution of the models*

Time series of some basic fields, such as precipitation, are important to give a basic guide to the evolution of the weather during the simulations, and these are shown in Figures 3 and 4. The period we focus on is 22 January to 3 February. The LAMs were only run for the period 0000 UTC on 23 to 0000 UTC on 26 January and the SCMs and GAMs were run for a longer period

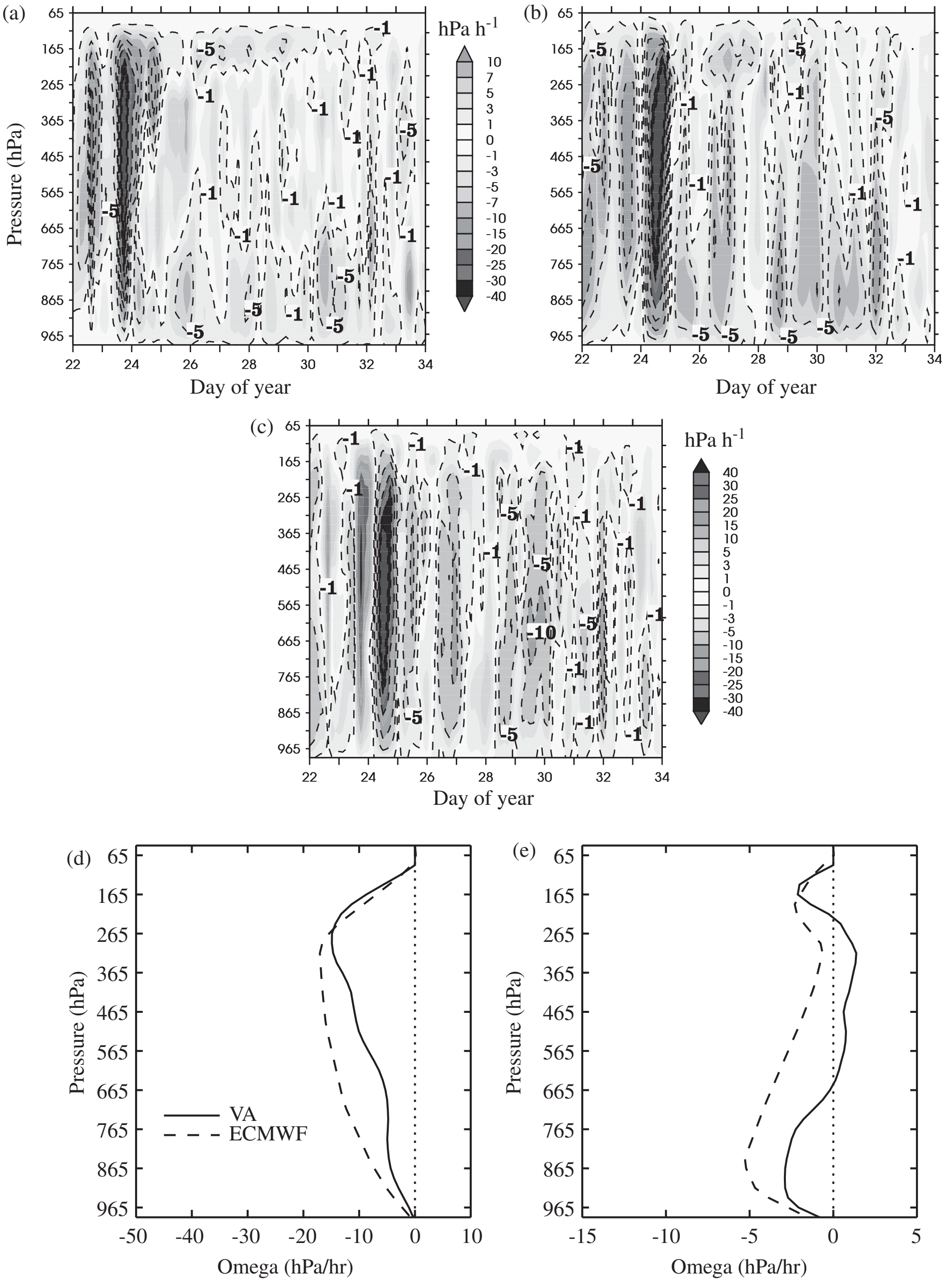

Figure 2. Comparison of the vertical velocity, ω, from the variational analysis (used to drive the CRMs and SCMs) and the ECMWF analysis (used to drive the GAMs and LAMs), showing time–height plots of 6 h mean ω from (a) the variational analysis, (b) the ECMWF analysis, and (c) the difference between the ECMWF analysis and the variational analysis. Also shown is the mean during the (d) wet and (e) dry periods (as defined in Figure 3).

than shown. The mean values use a 6 h averaging period. This was a pragmatic choice, balancing the removal of shorter temporal noise which makes the plot look too busy while maintaining useful information around the temporal variability. The standard deviations of the model spread were also calculated at these 6 h intervals but then averaged to 24 h to further remove noise. Standard deviation is shown normalised by the multi-model mean value. The spread was plotted separately from the mean, since plotting together (as is done in some later plots) led to too much clutter.

Figure 3(a) shows the time series of the multi-model mean surface rain rate along with observations. Also highlighted on this plot are two sub-periods we will describe as ‘wet’ and ‘dry’. The wet period runs from 0000 UTC on 23 to 0000 UTC on 25 January and the dry period from 1200 UTC on 25 to 1200 UTC on 2 February (day of year 33.5). These are chosen to allow us to sample a period of organised convection producing heavy precipitation and then a more suppressed period characterised by mid-level and shallow convection or broken deep convection producing lighter domain-averaged precipitation. The LAM simulations covered only the wet period. One further point to note is that the periods are based on the observed surface precipitation and timing errors in some model types would influence results from this kind of averaging; where relevant this is discussed.

From Figure 3(a) we can see that the CRMs and SCMs produce very similar precipitation rates through the period, as we would expect, due to these models being mostly constrained on the longer time-scales by their forcing (as discussed in Fridlind *et al.*, 2012; Davies *et al.*, 2012). Perhaps of interest is that the highest peak rates are a little lower in the SCMs, something seen much more strongly in the GAMs. This may suggest that the SCM is trying to behave like the GAM but needs dynamical feedback to show the full response, although it could also be related to a low-level dry bias and the way the forcing is applied to the SCMs (discussed later).

The GAMs have a significant delay (of 1 to 2 days) and a reduced peak (by over 50%) in the large rain event on 23 January but have similar amounts of rain during the latter part of the period. The LAMs also show a reduced peak rain event, and heavier than observed rain after the peak. However, the timing of the peak event in the LAMs is much closer to the timing of the observed peak than we see in the GAMs. As discussed

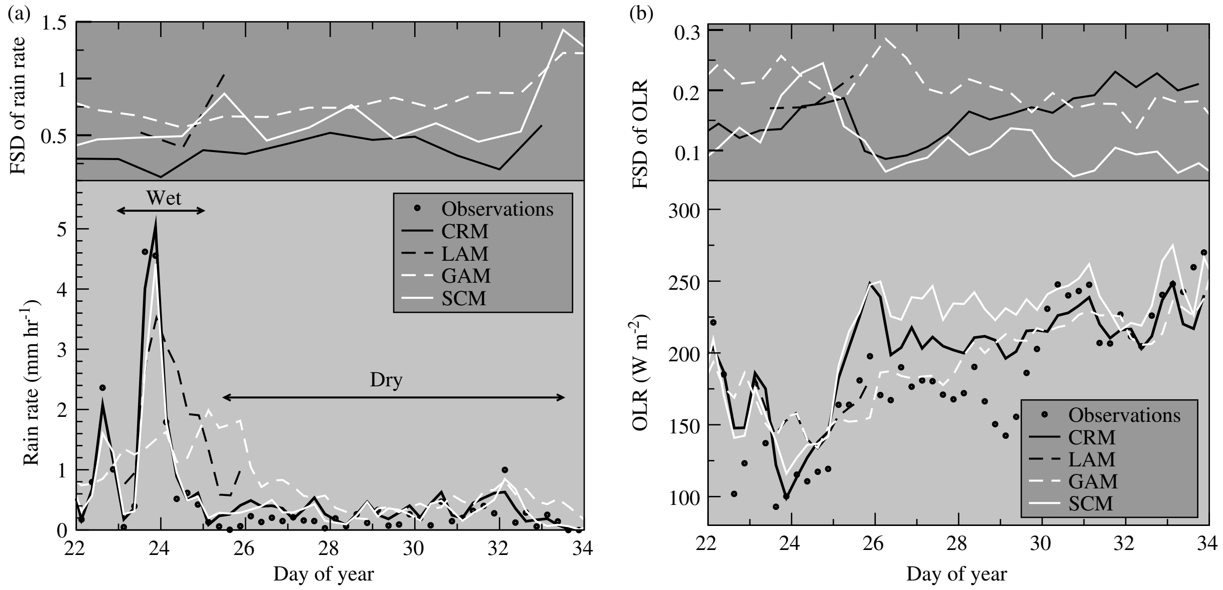

Figure 3. Time series of mean and fractional standard deviation (FSD, defined as standard deviation normalised by the mean value) of (a) rain rate and (b) outgoing long-wave radition (OLR). The time axis shows the day of the year at 0000 UTC.

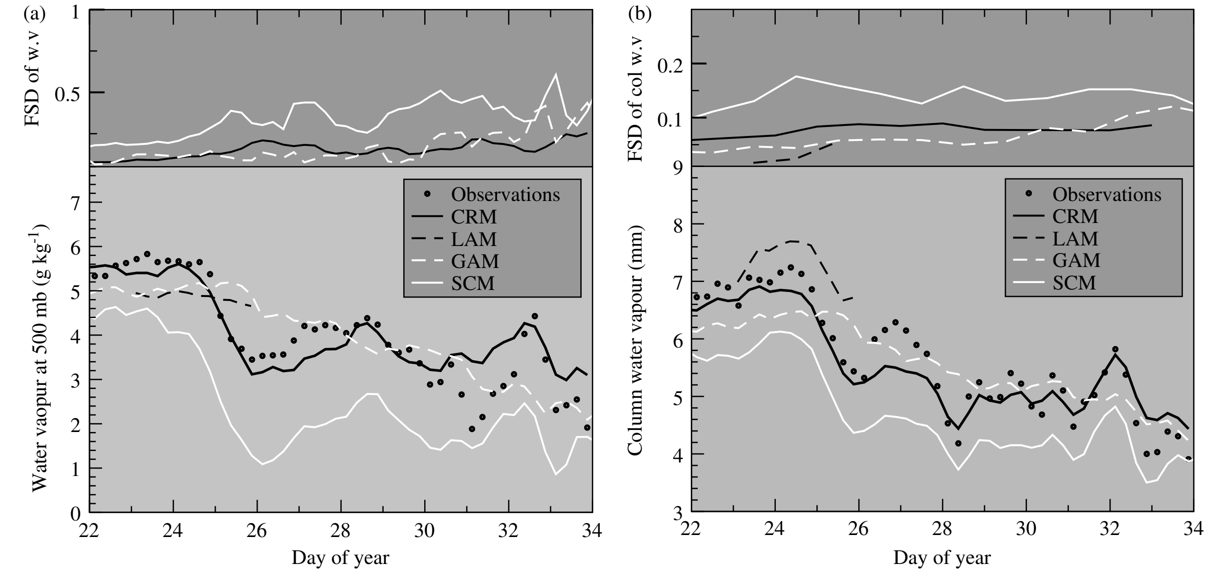

Figure 4. As Figure 3, but for (a) water vapour at 500 mb and (b) column water vapour.

in Zhu *et al.* (2013), the extended period of intense rain in the LAMs is the result of differences in mesoscale organisation in the inner model domain. For example, some models maintained the cyclone in the domain for longer than observed, which resulted in an increase in the period of intense precipitation. This issue is not seen in the CRMs and SCMs because these model types were constrained by the applied large-scale forcing. As the LAMs and GAMs are both run from the ECMWF analysis, this suggests that the large delay and reduction in the peak rain in the GAMs is, at least in part, due to the physics or dynamics in these models and not simply an issue with the forcing, which in Figure 2 was shown to have around a 1 day delay compared to the variational analysis. Issues with GAMs raining too frequently, producing too much lighter rain and not enough heavy rain events, has been noted before (Sun *et al.*, 2006; Wilcox and Donner, 2007; Stephens *et al.*, 2010) and this may be a useful case to study this. We also speculate here that the delay in the ECMWF analysis compared to the observationally based variational analysis may be due to the physics of the ECMWF model. The ECMWF physics will influence their analysis whereas the variational analysis uses precipitation observations directly to modify the divergent wind field. The SCM results highlight that this issue is not well studied in an SCM due to the need for a dynamical feedback.

The spread in the precipitation between models is shown in the top panel of Figure 3(a). The SCMs have larger spread than the CRMs, suggesting that, although precipitation is dynamically constrained on longer timescales, different SCMs can still produce quite different rain rates on a 6 h time-scale. The GAMs produce the largest spread between models on average, although it is clear that the LAMs vary a lot in their precipitation fields on 25 January as the large rain event moves away from the domain.

Figure 3(b) shows the time series and spread of outgoing long-wave radiation (OLR). The GAMs and LAMs lack of very intense precipitation, which is reproduced in the CRMs and SCMs, is also clear in the OLR with both model types missing the dip around 24 January. The CRM and SCM have larger than observed OLR between 25 and 30 January, suggesting a lack of cirrus cloud that was clearly seen in the observations and captured by the forcing data (Xie *et al.*, 2010). While the GAMs look better during the first half of this period, this is probably associated with the delay in the main convective event rather than them having a better cirrus representation. Overall, it appears that the GAMs and CRMs perform better than the SCMs. When we look at the spread of the models for OLR (Figure 3(b), top) it is clear that, although the GAMs do well in the multi-model mean, there is much larger spread than that exhibited by the CRMs and SCMs.

Figure 4 shows the mean and fractional standard deviation of (a) vapour on the model levels closest to 500 mb and (b) column integrated water vapour. In Figure 4 we see that in general the means of the CRMs compare well to observations. The good agreement between CRMs and observations in Figure 4(a) may suggest that the resolved convection is doing a good job of moistening the mid-troposphere. The GAMs and LAMs also capture the general water vapour trends, but seem to miss the steep drop in both column vapour and vapour at 500 mb around the heavy rain event. This is consistent with them delaying and reducing the peak in the precipitation during these events. We note here that we would expect the GAMs to remain reasonably close the observations as the analysis focuses on the 24–48 h period of the forecast, so long-term biases are not able to grow. On the other hand, the SCMs have already generated an obvious dry bias in both the column vapour and the mid-tropospheric vapour content at 500 mb by 22 January. This dry bias grows through the first half of the run. The dry bias at 500 mb in the SCMs may be a consequence of the convection schemes not responding to free tropospheric humidity and thus not producing an appropriate amount of mid-depth convection; this issue is reviewed in DelGenio *et al.* (2012). The causes of the overall dry bias exhibited by the SCMs are discussed in more detail in section 3.2. The spread in 500 mb water vapour and column vapour (Figure 4(a, b), top) in the GAMs and CRMs is similar through most of the period. The SCMs exhibit a larger spread than the other models which is in part because the mean is smaller but there is more absolute spread. This could be consistent with the fact that some convection schemes are more capable of moistening the free troposphere (as discussed in Lin *et al.*, 2012) and do not have such large biases.

Overall, Figure 3 shows the GAMs and CRMs simulate OLR better than the SCMs, presumably because the GAMs and CRMs do not exhibit the large dry bias and associated lack of mid- to high-level cloud seen in the SCMs. GAMs tend to produce the largest spread in OLR. This suggests that modelling centres should focus their attention on OLR and in particular its links with the behaviour of their convection schemes during both the wet and dry period. An SCM may be a useful tool for this; however, care must be taken since Figure 4 demonstrates large water vapour biases seen in this model type for these simulations, which have a strong influence on the OLR.

3.2. *The wet and dry periods*

Figure 5 shows the profiles of temperature and water vapour mixing ratio for the models and the ECMWF analysis differenced from the variational analysis averaged over the wet and dry periods (as indicated in Figure 3(a)). While not shown here, we note that the temperature and water vapour profiles taken directly from the radiosondes were a good match to the variational analysis, so the differences plotted are a true bias from observations. The first point to note from Figure 5 is that, particularly for temperature but also for water vapour, there are very large biases in the mean of the SCMs. The multi-model spread is large, but in the wet period there is no overlap between a standard deviation from the mean and the other models. As CRMs do not exhibit these same biases but are forced in a similar way, the biases are either related to the physics in the SCMs or to the different ways in which the forcing was applied in the two model types.

To investigate whether the method of forcing is leading to differences in the CRMs and SCMs, we carried out a sensitivity study using the Met Office CRM. Figure 6 shows the temperature and moisture bias from the wet period using the two methods for forcing the CRMs and SCMs. It is clear that the standard CRM forcing method and the SCM forcing method do lead to significantly different biases. In the lowest few kilometres, the temperature and water vapour bias in the Met Office CRM looks much more like the multi-model mean SCM bias, when the SCM forcing method is applied. However, above this the Met Office CRM appears to behave a little more like other CRMs, suggesting that the main influence of the forcing is in the lowest 5 km. This can be understood if we think about the convergence of water vapour with the two forcing methods, noting that during the wet periods there is a convergence of water vapour through much of the lower troposphere due to the vertical motion (Figure 2). In the CRM forcing method, the water vapour convergence is prescribed by the observations. However, in the SCM method of forcing the water vapour convergence is determined by the predicted water vapour in the model. Application of the SCM forcing to the CRM results in a 30% reduction in water vapour convergence due to the bias. As the SCMs and CRMs have a negative bias in water vapour in the lower troposphere, the convergent term of the forcing will give a positive feedback on this bias with the SCM style of forcing. Table 2 summarises the convergent term of the forcing feedback on the average bias.

Table 2. A summary of the feedback of the SCM-type forcing on the model bias.

Situation	Model bias	Feedback with SCM forcing
Convergence	Negative bias	Increased negative bias
Convergence	Positive bias	Increased positive bias
Divergence	Negative bias	Reduced negative bias
Divergence	Positive bias	Reduced positive bias

The sensitivity to the forcing method (Figure 6) suggests that the SCMs general cold/dry bias in the mid to upper troposphere (Figure 5) is not entirely an artifact of the forcing because the bias is much smaller in the CRM using identical forcing. Also, while we would expect the low-level dry bias also to contribute to such a bias (less moisture for latent heating in convection), this low-level dry bias is also present in the CRMs but the upper-level bias is much smaller. Therefore, this cold/dry bias may well be of interest to model developers as it is also evident in the full GAMs (although with a smaller magnitude for reasons discussed later).

The multi-model spread (of each model type) (Figure 5) is consistent with time series. The SCMs have a larger spread than the CRMs presumably because convection is parametrized in significantly different ways in the different SCMs. The GAMs have a larger spread than the LAMs for the same reason. The GAMs and LAMs both have lower spread than the CRMs and SCMs for two main reasons. Firstly, they are only 24–48 h into the forecast so very large biases do not have time to grow. Secondly, it is typical for large-scale dynamical feedbacks in the GAMs to prevent biases of the magnitude of those seen in the SCMs from developing.

Figure 7 shows profiles of the mean and spread of cloud fractions from the different models. During the wet period there are notable differences in the means, with the CRMs producing less cloud fraction below 5 km and more above. The GAMs and LAMs also have lower fractions than the SCMs. This may suggest there are issues around the forcing during this period leading to these differences but, given the very large multi-model spread, care should be taken into reading too much into this. It is also worth noting that the cloud fraction profiles are consistent with the OLR presented in Figure 3, namely, the CRMs produce the largest cloud fraction in Figure 7(a) and this corresponds to the lowest OLR during the wet period. Likewise the LAMs and GAMs simulate lowest cloud fractions, which is consistent with the largest OLR during the wet period. The mean of the cloud fractions during the dry period agree reasonably well across model types, particularly above the freezing level. However, again given the large spread of all model types, this would seem an area for all modellers to focus further attention on. We note here that the CRMs cloud fraction is produced from resolved cloud-scale motion whereas the SCMs and GAMs will all have a parametrization scheme to represent this. Above the freezing level, CRMs have as large a spread in cloud fraction as the SCMs and GAMs and this should be a focus for those who develop and evaluate CRMs. Fridlind *et al.* (2012)

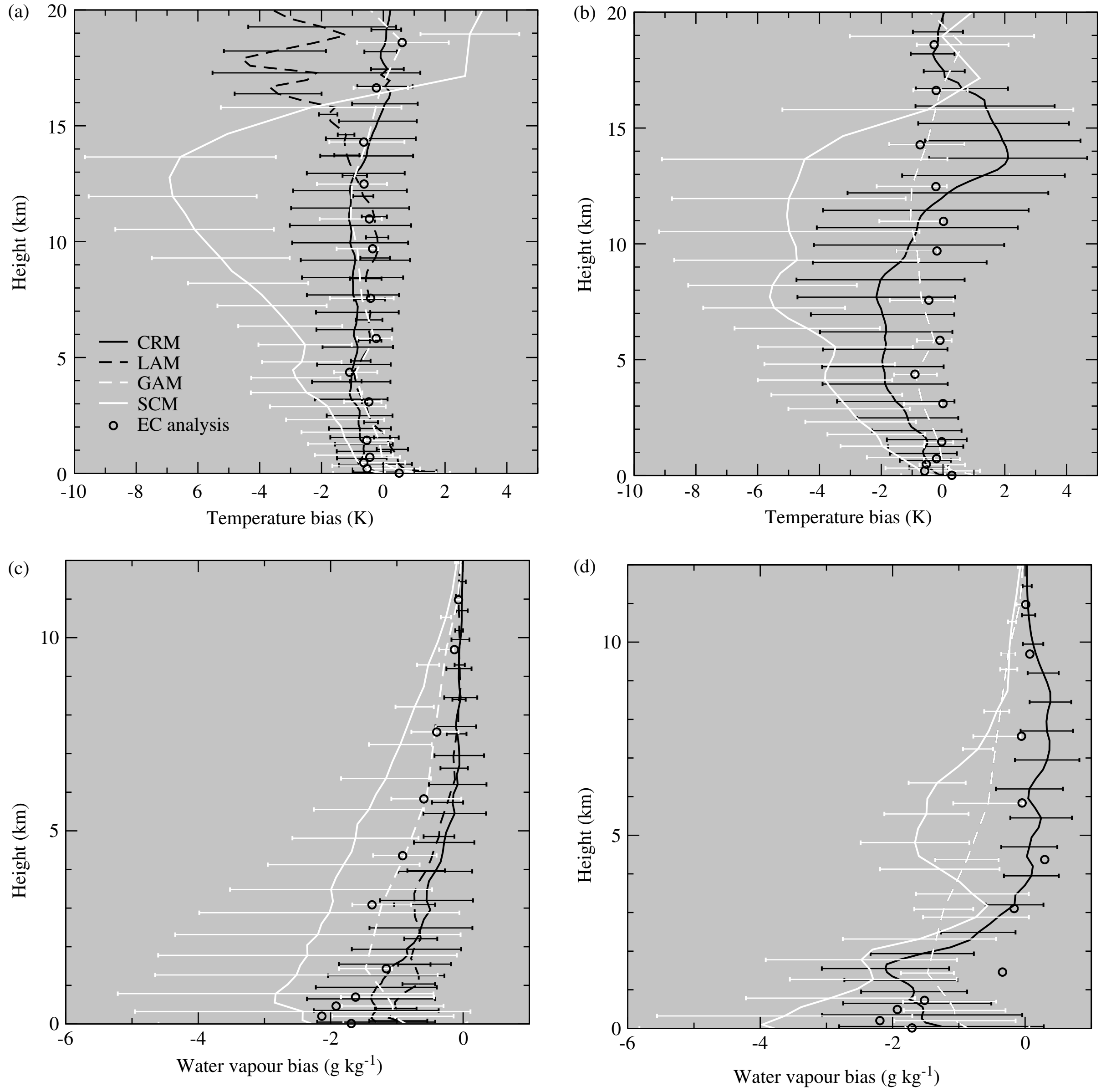

Figure 5. Mean temperature and water vapour mixing ratio biases (multi-model mean differenced from the variational analysis) and the multi-model spread shown as a standard deviation each side of the mean: temperature bias for (a) the wet period and (b) the dry period, and water vapour bias for (c) the wet period and (d) the dry period. Also included is the mean ECMWF analysis.

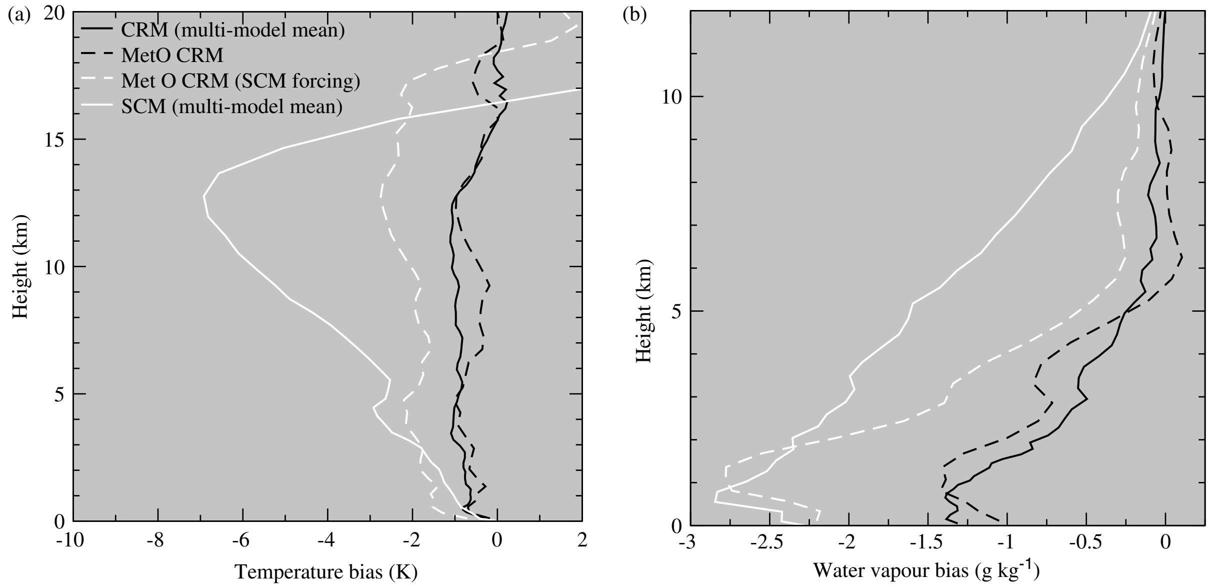

Figure 6. The sensitivity of the two forcing methods on (a) the temperature bias and (b) the moisture bias during the wet period using the Met Office CRM. Also included are the CRM and SCM multi-model means from Figure 5.

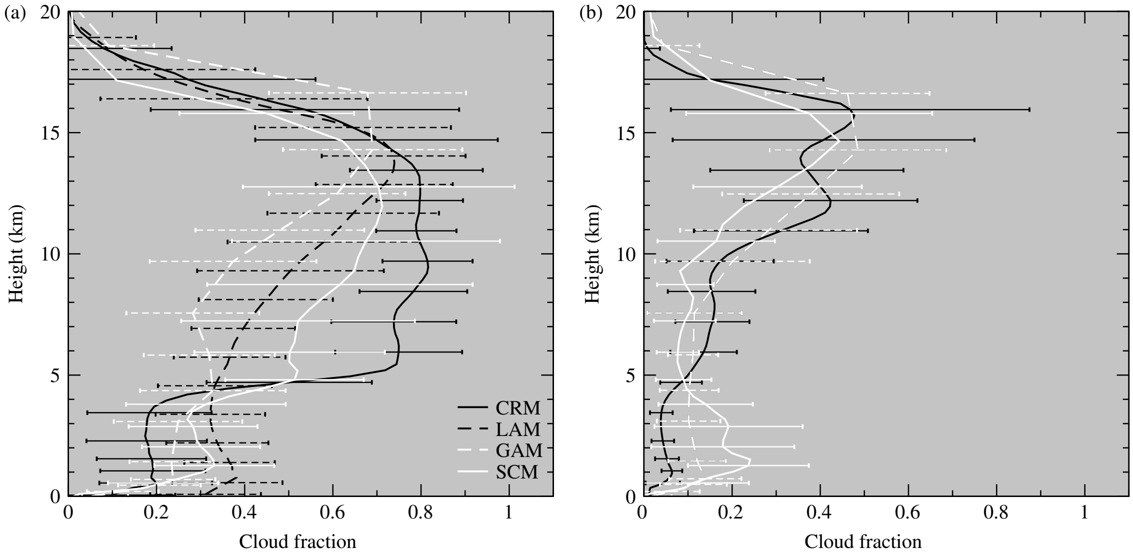

Figure 7. Domain-mean cloud fraction profiles and spread (shown as a standard deviation each side of the mean) for the (a) wet and (b) dry periods. In the CRM/LAMs, a point is considered cloudy if it has a water content greater than 10^{-3} g kg^{-1}.

suggest that CRM cloud fraction differences are attributable in part to differing ice nucleation schemes.

Figure 8 shows profiles of the domain mean and spread of ice and liquid water content. The liquid water shows that there is much lower spread in CRMs and LAMs and the means agree well with each other in the wet period (when the LAMs were run). This suggests that, for models which resolve cloud-scale motions, there is general agreement on the amount of cloud water which should be produced. On the other hand, the ice is a different story. In the wet period there are both large differences between the mean profiles for the model types and also large spread of ice for all model types. In the mean, perhaps of particular note are the low ice contents in the GAMs during this period; most of the SCMs had similarly low ice contents, although the mean does not show this and this is discussed later.

A potential reason for the differences in the mean ice contents in the CRMs and GAMs is related to what defines the ice water content. In the CRMs any solid hydrometeors (including snow and graupel) are included in Figure 8, whereas many GAMs and SCMs do not represent these species explicitly and therefore do not report them. While this difference may be a simple diagnostic issue, there are also potential modelling issues to consider. In particular, the snow and, to a lesser extent, graupel is important for radiative transfer (e.g. Petch, 1998) and if ignored in the GAMs there is likely to be compensating tuning to add in this missing cloud. While the CRM comparison did not request a breakdown of the ice into separate categories (owing to a lack of observation constraints for separate components), we do have this information from some models. As an example, Figure 9 shows the role of including all precipitating hydrometeors into the calculations of water content and cloud fraction for the Met Office CRM. While this impact will depend strongly on the microphysics scheme, it does highlight the need for significant care when we compare water contents and cloud fractions in convective situations.

Yet another issue to deal with when diagnosing and comparing water contents and cloud fraction is the role of the convection scheme in GAMs and SCMs. While convective schemes may detrain condensate into the large scale, they also have implied water content which are often not diagnosed and reported, not used in the radiation schemes, or both. We believe this should also be a focus for future comparisons.

3.3. *Presentational issues*

It was noted in the discussion about ice content that most of the SCMs had lower ice contents than the mean shown and that they were more similar to the GAMs. To highlight this, Figure 10 shows the ice content profiles from all the SCMs for the dry period; the wet period is not included but showed the same feature to a slightly lesser extent. It is clear from Figure 10 that there is a single outlier from other models and this is making a large contribution to the mean and spread. It would be possible to remove this outlier from the data, but there is not a good reason to do this and we noted that CRMs and LAMs had significantly higher ice contents than most of the SCMs. Another alternative would be to plot the median and interquartile range as shown on the figure. While this would be an entirely valid option, this is another way of downweighting outliers, and with relatively small samples (6–10 models) it may be preferable to plot both and ensure the true story is presented in any reporting of the results. For other issues discussed in this article, the averaging used in the plots did not influence the conclusions and therefore we have plotted mean values and standard deviations from this mean. However, when modelling centres use the data from this comparison, they should be aware of these issues.

4. Summary and discussion

In this article we have presented some basic fields from four intercomparisons of different model types, all simulating the TWP-ICE field campaign. This is the first example where there has been a coordinated effort to have CRMs, SCMs, GAMs and LAMs all evaluated for the same case. The large variety of observations and the high temporal sampling of atmospheric conditions, along with the availability of four separate model intercomparisons, make this a very good study for those developing their regional and global atmospheric models. The analysis in this article focused on comparing the multi-model means and the multi-model spread for each model type. When needed, we also performed some additional sensitivity studies using a single model. Conclusions from our analysis fall broadly into three categories. Firstly, there are the issues and lessons learnt about the design of multi-model type intercomparison experiments with recommendations for improvements in the future. Secondly, we highlight areas where most (or all) of the models of a given type are seen to perform poorly or have large spread, and therefore which require attention by model developers. Thirdly, we highlight where care needs to be taken when analysing and plotting model diagnostics.

Where possible the models were forced in a similar way. However, pragmatic choices were made, and, because it was not considered a key issue when the case was designed, the SCMs and CRMs applied their vertical forcing term differently. While both used the same observationally derived vertical velocity, SCMs were

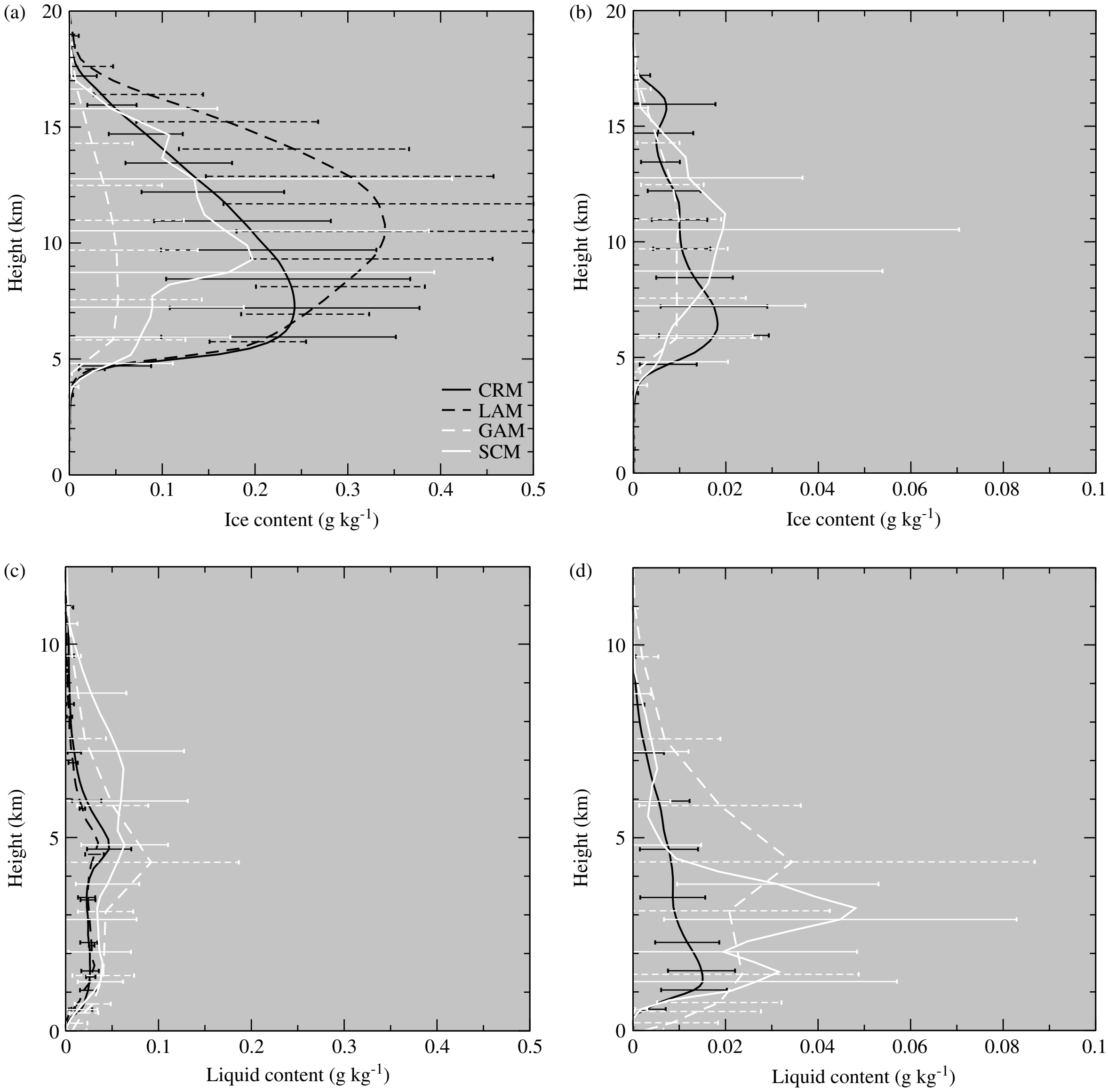

Figure 8. Domain-mean cloud water content profiles and spread (shown as a standard deviation each side of the mean). Included are ice mixing ratio for (a) the wet period and (b) the dry period, and liquid water mixing ratio for (c) the wet period and (d) the dry period. Note that the wet period has an x-axis range five times larger than the dry period.

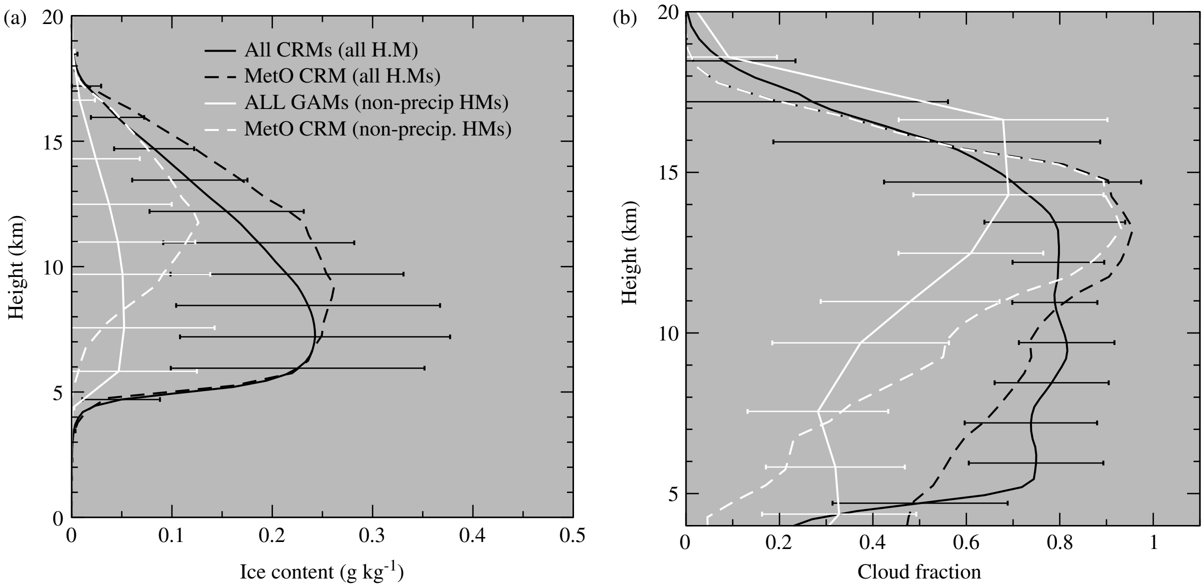

Figure 9. (a) solid water content and (b) fractional cloud cover for the wet period. Included are the profiles and spread from the CRMs and GAMs and a sensitivity study with the Met Office CRM where precipitating hydrometeors are included in the calculation.

forced by advecting their predicted temperature and moisture whereas CRMs were forced by advecting the observed temperature and moisture. This was shown to be important for our results with the SCMs growing a dry bias when there was convergence. While there is no obviously correct method for forcing SCMs and CRMs, it is important for future model comparisons such as this to use the same method (or both). The general strengths and weaknesses of both methods of forcing SCMs and CRMs used here are worth highlighting. The methods try to seperate the biases due to the physics from those due to the dynamics; however, such methods stop us from seeing how physical errors interact with the large scale. This is a reason why SCMs on their own cannot be used as a tool for developing parametrisations and why other methods such as weak temperature gradient have been employed in some studies (e.g. Sobel *et al.*, 2001). Also shown in this work is that the SCMs (and CRMs to some extent) generate large biases in their temperature or moisture profiles which are not seen in the GAMs or LAMs which used a series of short-range forecasts. As these biases may well influence many other physical aspects of the models, it is our recommendation that the SCMs are run as a series of short-range forecasts (as we do with the GAMs). As SCMs are computationally inexpensive to run, it should not be a problem for this to be done in addition to the longer free runs.

We also highlighted two key differences between the variational analysis and the ECMWF analysis. Firstly, there was stronger upward motion in the ECMWF analysis, although this was not linked with any specific differences between model types in our analysis. Secondly, the strongly forced rain event was of the order of a day later in the ECMWF analysis and this led to an expected delay in the peak rain produced in the LAMs and GAMs. Interestingly, the GAMs delayed the peak rain rate by a further day when compared to the LAMs and produced a weakened peak. As a reduced range of precipitation rates is a typical feature in many climate models (Stephens *et al.* 2010), we suggest that this may be a useful test case to study this despite the fact that there is already a signal for this in the driving analysis. The SCMs also produced a reduced peak in precipitation when compared to the CRMs.

A key finding of this study was that all model types had a lower tropospheric dry bias for this case and that the ECMWF analysis itself had a significant dry bias, particularly in the lowest levels. We speculated that, as all the GAMs had a tendency to produce a large dry bias, it was therefore the model contribution to the ECMWF analysis which led to bias in the analysis itself. It is possible that the dry bias which was seen most strongly in the SCMs and the GAMs could be the cause of the reduced precipitation intensity in the GAMs and the SCMs since they will have a reduced source of moisture for producing the precipitation.

This study also highlighted that there remains a great deal of uncertainty in ice microphysics across models. There was essentially a large spread in ice contents for all four model types which, given they all typically use similar bulk microphysical schemes, suggests that this remains an area for model developers on which to focus their attention. It also means there are no reference models for this kind of experiment, so observations to constrain the models remain a critical requirement. Liquid cloud was a somewhat different story. In the models which had explicit convection or cloud-scale dynamics (LAMs and CRMs), there was good agreement and small multi-model spread in liquid water profiles. However, for models with parametrized convection, there were notable differences from the CRMs and LAMs and large multi-model spread. The focus of those developing convection schemes tends to be on their impacts on the vertical transport of heat and water vapour, and on surface precipitation. The results shown here suggest that there also needs to be a focus on the clouds produced.

The challenges of comparing clouds and microphysical properties across model types were also highlighted in this paper. In particular, we noted that bulk microphysical schemes make different assumptions about how each category (e.g. snow, ice and graupel) is defined. However, often only the ice is prognostic and reported in many SCMs and GAMs, i.e. a precipitating ice hydrometeor content is not always diagnosed or reported. This makes a comparison with CRMs and LAMs difficult. An appropriate comparison may be to use the water content as seen in the radiation scheme, but it is quite possible that many models do not consider precipitating ice despite its significant contribution to optical depth (Petch, 1998). Therefore, to really understand how clouds compare across models of all types, we need to be very specific about the species in the model. While forward modelling and simulators can help (and this was used to compare CRMs with radar in Fridlind *et al.*, 2012), model developers would benefit from some clear comparisons of different hydrometeor types, clearly defining what they are, and how they influence radiative transfer. We acknowledge that it may be premature to focus solely on ice hydrometeor type, given that there is a lack of observational constraint. However, it is useful to understand how different treatments of the microphysics and their application within radiation schemes vary between models and to identify the impact of these differences. We therefore have a further recommendation for future multi-model type intercomparisons to clearly diagnose all hydrometeor types in their models separately and to define how they interact with radiation.

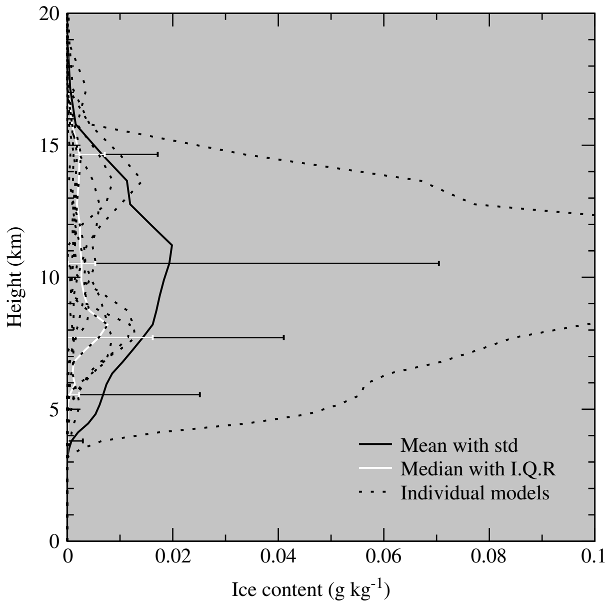

Figure 10. SCM ice contents averaged for the dry period.

In summary, the TWP-ICE field campaign and the intercomparisons of four different model types provide an extremely valuable resource for those developing models. This article (along with the four individual intercomparison articles) highlights some interesting features which this experiment can be used to study further, but there are likely to be many more. We have also made recommendations for some changes to the forcing for those using this case for their model development, as well as various recommendations for those involved in coordinating future multi-model type intercomparisons.

Acknowledgements

A. Fridlind was supported by the NASA Radiation Sciences Program and the US DOE Office of Science, Office of Biological and Environmental Research, through contracts DE-AI02-06ER64173 and DE-SC0006712. The contributions of S. Xie to this work were performed under the auspices of the US Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research by Lawrence Livermore National Laboratory under contract no. DE-AC52-07NA27344 and supported by the Atmospheric Radiation Measurement Program of the Office of Science at the US Department of Energy. Ping Zhu wishes to acknowledge his support by the DOE ASR program under grant DE-FG02-09ER64737. Yanluan Lin was supported by the Office of Science (BER), US Department of Energy. Laura Davies and Christian Jakob were supported by the US DOE ASR program under grant DE-FG02-09ER64742.

References

Davies L, Jakob C, Keane RJ, Whitall MA, Plant RS, Lin Y, Wang W, Wolf A, Del Genio AD, Larson VE, Nielsen BJ, Liu X, Shi X, Song X, Zhang G, Komori T, Hill AA, Petch JC, Hume T, Singh M, Cheung K. 2013. A single-column model ensemble approach applied to the TWP-ICE experiment. *J. Geophys. Res.* submitted.

Del Genio AD. 2012. Representing the sensitivity of convective cloud systems to tropospheric humidity in general circulation models. *Surv. Geophys.* **33**: 637–656. DOI: 10.1007/s10712-011-9148-9.

Fridlind AM, Ackerman AS, Chaboureau J-P, Fan J, Grabowski WW, Hill AA, Jones TR, Khaiyer MM, Liu G, Minnis P, Morrison H, Nguyen L, Park S, Petch JC, Pinty J-P, Schumacher C, Shipway BJ, Varble AC, Wu X, Xie S, Zhang M. 2012. A comparison of TWP-ICE observational data with cloud-resolving model results. *J. Geophys. Res.* **117**: D05204, DOI: 10.1029/2011JD016595.

Kendon LJ, Rowell D, Jones RG. 2010. Mechanisms and reliability of future projected changes in daily precipitation. *Clim. Dyn.* **35**: 489–509.

Lin Y, Donner LJ, Petch JC, Bechtold P, Boyle JS, Klein SA, Komori T, Wapler K, Willett M, Xie X, Zhao M, Xie S, McFarlane SA, Schumacher C. 2012. TWP-ICE global atmospheric model intercomparison: Convection responsiveness and resolution impact. *J. Geophys. Res.* **117**: D09111, DOI: 10.1029/2011JD017018.

May PT, Mather JH, Vaughan G, Bower KN, Jakob C, McFarquhar GM, Mace GG. 2008. The Tropical Warm Pool International Cloud Experiment. *Bull. Am. Meteorol. Soc.* **89**: 629–645.

Petch JC. 1998. Improved radiative transfer calculations from information provided by bulk microphysical schemes. *J. Atmos. Sci.* **55**: 1846–1858.

Petch JC, Willett M, Wong RY, Woolnough SJ. 2007. Modelling suppressed and active convection. Comparing a numerical weather prediction, cloud-resolving and single-column model. *Q. J. R. Meteorol. Soc.* **133**: 1087–1100.

Phillips TJ, Potter GL, Williamson DL, Cederwall RT, Boyle JS, Fiorino M, Hnilo JJ, Olson JG, Xie S, Yio JJ. 2004. Evaluating parameterizations in general circulation models: Climate simulation meets weather prediction. *Bull. Amer. Meteorol. Soc.* **85**: 1903–1915.

Randall DA, Xu K-M, Somerville RJC, Iacobellis S. 1996. Single-column models and cloud ensemble models as links between observations and climate models. *J. Climate* **9**: 1683–1697.

Randall DA, Khairoutdinov M, Arakawa A, Grabowski WW. 2003. Breaking the cloud-parameterization deadlock. *Bull. Amer. Meteorol. Soc.* **84**: 1547–1564.

Sobel AH, Nilsson J, Polvani LM. 2001. The weak temperature gradient approximation and balanced tropical moisture waves. *J. Atmos. Sci.* **58**: 3650–3665.

Stephens GL, L'Ecuyer T, Forbes R, Gettlemen A, Golaz JC, Bodas-Salcedo A, Suzuki K, Gabriel P, Haynes J. 2010. Dreary state of precipitation in global models. *J. Geophys. Res.* **115**: D24211, DOI: 10.1029/2010JD014532.

Sun Y, Solomon S, Dai A, Portmann RW. 2006. How often does it rain? *J. Climate* **19**: 916–934.

Wilcox EM, Donner LJ. 2007. The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric General Circulation Model. *J. Climate* **20**: 53–69.

Xie S, Xu K-M, Cederwall RT, Bechtold P, Del Genio AD, Klein SA, Cripe DG, Ghan SJ, Gregory D, Iacobellis SF, Krueger SK, Lohmann U, Petch JC, Randall DA, Rotstayn LD, Somerville RCJ, Sud YC, Von Salzen K, Walker GK, Wolf A, Yio JJ, Zhang GJ, Zhang M. 2002. Intercomparison and evaluation of cumulus parametrizations under summertime midlatitude continental conditions. *Q. J. R. Meteorol. Soc.* **128**: 1095–1135.

Xie S, Hume T, Jakob C, Klein SA, McCoy RB, Zhang M. 2010. Observed large-scale structures and diabatic heating and drying profiles during TWP-ICE. *J. Climate* **23**: 57–79.

Xu K-M, Cederwall RT, Donner LJ, Grabowski WW, Guichard F, Johnson DE, Khairoutdinov M, Krueger SK, Petch JC, Randall DA, Seman CJ, Tao W-K, Wang D, Xie SC, Yio JJ, Zhang M-H. 2002. An intercomparison of cloud-resolving models with the Atmospheric Radiation Measurement summer 1997 Intensive Observation Period data. *Q. J. R. Meteorol. Soc.* **128**: 593–624.

Zhu P, Dudhia J, Field PR, Fridlind A, Varble A, Zipser E, Petch JC, Chen M, Zhu Z. 2012. A limited-area model (LAM) intercomparison study of a TWP-ICE active monsoon mesoscale convective event. *J. Geophys. Res.* **117**: D11208, DOI: 10.1029/2011JD016447.