# Multiyear Evaluations of a Cloud Model Using ARM Data

PETER W. HENDERSON AND ROBERT PINCUS

*Cooperative Institute for Research in Environmental Sciences, University of Colorado, and Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado*

(Manuscript received 19 September 2008, in final form 3 April 2009)

ABSTRACT

This work uses long-term lidar and radar retrievals of the vertical structure of cloud at the Atmospheric Radiation Measurement (ARM) program's Southern Great Plains site to evaluate cloud occurrence in multiyear runs of a cloud system–resolving model in three configurations of varying resolution and sophistication. The model is nudged to remain near the observed thermodynamic state and model fields are processed to mimic the operation of the observing system. The model's skill in predicting cloud occurrence is evaluated using both traditional performance measures that assume ergodicity and probabilistic measures that do not require temporal averaging of the observations.

The model shows considerable skill in predicting cloud occurrence when its thermodynamic state is close to that observed. The overall bias in modeled cloud occurrence is relatively small in all model runs, suggesting that this field is relatively well calibrated. The Brier scores attained by all configurations also suggest considerable model skill. Greater differences in performance are found between seasons than between model configurations during the same season, despite substantial differences between the computational costs of the configurations. Several significant seasonal dependencies are identified, most notably greater conditional bias, but better timing, of boundary layer cloud in winter, and substantially less conditional bias in high cloud during summer.

## 1. Systematic evaluation of cloud-scale models

Cloud-scale models, such as cloud system–resolving models (CSRMs) were originally developed to investigate cloud-scale processes and were used to simulate a relatively narrow range of situations. Their success at simulating observations, especially by comparison with parametric representations taken from global models (Randall et al. 2003), has led them to be used as a bridge between global models and observations (Randall et al. 1996) and to provide a basis for parameterizations in global models (Lock et al. 2000). Multiscale models of the atmosphere take this logic to its extreme and use a CSRM within each grid cell to replace existing parameterizations. As a result, multiscale models use the CSRM in circumstances far beyond those in which they have been evaluated.

Most CSRM evaluations are made using relatively short case-study periods for which detailed observations are available and, because computational cost is not the driving factor, the model may be run at high spatial resolution. When CSRMs are used in global models, however, they are run at much lower resolution, and model performance at these resolutions is not a foregone conclusion. CSRMs used in global models are also subject to a much wider range of atmospheric conditions than are explored in most case studies, and evaluation of the CSRM should logically follow suit.

What observations may be used to evaluate CSRMs over such a wide range of conditions? One possibility is the ground-based measurements made at observatories operated by the Atmospheric Radiation Measurement (ARM) and Cloudnet programs. These sites combine retrievals from upward-pointing active remote sensing instruments (radars and lidars) to produce long-term, high-frequency records of the vertical structure of clouds (e.g., Clothiaux et al. 2000). The main difficulty with using these observations is that they are pointlike, whereas model predictions are defined at the larger spatial scales of model domains.

*Corresponding author address:* Robert Pincus, Cooperative Institute for Research in Environmental Sciences, University of Colorado, NOAA/Earth System Research Laboratory, Physical Sciences Division, 325 Broadway, Boulder, CO 80305–3337.
E-mail: robert.pincus@colorado.edu

DOI: 10.1175/2009JAS2957.1

Pointlike measurements are usually compared to the output of large-scale models by invoking the ergodic hypothesis, which asserts that observations averaged over time are equivalent to the spatial mean of the field. This approach relies heavily on identifying optimal averaging scales because the ergodic hypothesis fails if significant differences exist between the spatial and temporal statistics of cloud in either the model or the observations, making the application of many measures at best difficult and often inappropriate.

An alternative is to apply probabilistic techniques developed to verify ensemble forecasts (Jakob et al. 2004). These techniques are well established in numerical weather prediction but have seldom been applied to cloud models. They are conceptually appealing because they bridge the disparities of scales without reducing the information content of the observations or relying on time averaging. However, it is not clear if this approach has any demonstrable advantage in practice, or even if the measures are sensitive enough to distinguish between the performances of different models.

Here we use both traditional and probabilistic methods to evaluate the performance of a CSRM in predicting cloud occurrence under a wide range of atmospheric conditions. We consider three model configurations of varying spatial resolution and sophistication to quantify trade-offs between model performance and computational cost. We focus on forecasts of cloud occurrence and its mean field (cloud fraction) as opposed to continuous fields (e.g., liquid water content) in order to reduce observational uncertainties. We also map the model forecasts to the observations by accounting for instrument sensitivity and other observational artifacts.

Section 2 gives details of the CSRM runs, including the nudging used to keep the model near the observed thermodynamic state. Section 3 describes the methods that we use to make modeled and observed cloud data more comparable. Section 4 compares various measures used to evaluate model performance; these methods are applied to model predictions using three configurations in section 5. Section 6 discusses the results and offers some possible explanations for the main findings.

## 2. Multiyear simulations using three model configurations

We evaluate the performance of the System for Atmospheric Modeling (SAM; Khairoutdinov and Randall 2003) at ARM's Southern Great Plains (SGP) site. SAM is the CSRM component of the multiscale form of the National Center for Atmospheric Research (NCAR) Community Atmosphere Model (CAM); this combination is known as the ''super-parameterized'' CAM (SP-CAM; Khairoutdinov et al. 2005).

We run the model for three years (1999–2001) using estimates of the surface fluxes of latent and sensible heat, large-scale advective tendencies of temperature and moisture, and horizontal winds derived using variational analysis (Zhang et al. 2001; Xie et al. 2004). The forcing data do not provide estimates of condensate advection.

We run SAM in three configurations: (i) a standard configuration like that used in SP-CAM, namely, a 32-column 2D domain, oriented east–west with $\Delta x =$ 4 km and 28 vertical levels; (ii) a much higher spatial resolution ($\Delta x = \Delta y = 500$ m) 128 × 128-column, 3D domain with 64 levels, which is capable of better-resolved dynamics and is large enough to simulate deep convection; and (iii) a configuration with the same low spatial resolution of the standard model, but which includes an intermediate prognostic high-order closure (IPHOC) of turbulence to improve the representation of shallow cumulus and its transition to deep convection (Cheng and Xu 2008). IPHOC treats subgrid-scale transport and, in particular, allows for subgrid-scale fractional cloudiness.

All configurations have cyclic boundary conditions and use a stretched vertical grid, such that the spacing between levels increases with height and is typically 100–500 m in the planetary boundary layer (PBL). The domain extends to ~28 km. Newtonian damping is applied to the upper ⅓ of the model domain to suppress gravity waves (Khairoutdinov and Randall 2003).

Instantaneous model output is collected hourly, producing $\sim 2.6 \times 10^4$ values at each model level in each column of the domain. More frequent sampling is not warranted: observations of cloud occurrence have autocorrelation values of >0.6 below 11 km and >0.7 in the PBL at time lags of an hour.

Our goal is to assess the model's ability to predict cloud occurrence when the thermodynamic state is correctly specified. In free model runs, however, significant biases in thermodynamic fields appear within a matter of weeks, even when the model is forced with observed fluxes and advective tendencies (Khairoutdinov and Randall 2003). To keep these biases manageable, we nudge the model's temperature field toward the observations with time scale $\tau_T$ chosen on the basis of 15 long model runs covering a wide range of $\tau_T$. Our selection of $\tau_T = 24$ h balances the magnitude of the temperature errors with the need to keep the nonphysical nudging term as small as possible compared to the advective tendency (Ghan et al. 1999). Winds are nudged on a 2-h time scale (e.g., Khairoutdinov and Randall 2003); this accounts for the absence of a pressure gradient term in the model's momentum equation. Water vapor is not nudged because it is already heavily constrained by surface precipitation.

## 3. Reducing the gap between models and observations

We evaluate model predictions of cloud fraction using instantaneous binary observations of cloud occurrence from the ARM program's Active Remotely Sensed Cloud Locations (ARSCL) lidar and radar cloud-boundaries product (Clothiaux et al. 2000). The observations are quite dense in the vertical and each model layer contains multiple observations. We verify forecasts using observations closest to the model levels; in the absence of systematic correlation between clouds in nature and the location of the model grid levels in the vertical, this is equivalent to averaging the observed binary field over model layers and rounding to 0 or 1.

### *a. Mapping modeled cloud to observed cloud*

We use ''instrument simulators'' (e.g., Klein and Jakob 1999; Bodas-Salcedo et al. 2008) to convert the condensate fields predicted by SAM into clouds masks that would be observed by ARSCL. At each time step we predict the radar reflectivity and lidar signal that would be measured in each column of the model domain, replicate the logic employed by ARSCL to discriminate clouds and precipitation, and form a cloud mask comparable to the observations.

We simulate the reflectivities observed by ARM's 35-GHz millimeter cloud radar (MMCR) using the QuickBeam radar simulator (Haynes et al. 2007). QuickBeam uses profiles of temperature, relative humidity, and the mixing ratio of each hydrometeor species to compute the radar reflectivity that would be observed by a ground-based instrument, accounting for attenuation of the radar beam by atmospheric gases and hydrometeors. Radar reflectivity is computed for every column in the model domain at each hour using the instantaneous fields. Calculations employ precomputed look-up tables in place of exact Mie calculations; this introduces reflectivity errors typically less than 2 dB (Haynes et al. 2007). The radar cloud mask is constructed by comparing the simulated reflectivity in each column to the instrument detection threshold $\mathrm{dB_{lim}} = \log_{10}(z^{20}) - 50$, for $z$ in km (P. Kollias 2007, personal communication). When the reflectivity exceeds this threshold we say that the radar would have detected cloud.

ARSCL uses optical lidar retrievals to detect thin clouds and to distinguish falling precipitation from clouds (Clothiaux et al. 2000). We approximate this process by computing the lidar extinction coefficient $k_{\mathrm{ext}}$ and using this to determine when a lidar signal would be detected and when it would be attenuated. SAM uses bulk microphysics schemes and so does not predict particle sizes. However, it does use CAM's radiation scheme, which specifies the effective droplet radius over land as a function of temperature. We use this value of $r_e$ to infer the droplet number density $N$ from local cloud liquid and cloud ice concentrations. Assuming a scattering efficiency of 2 gives $k_{\mathrm{ext}} \approx 2\pi r_e^2 N$, and we define cloud occurrence (according to the lidar) whenever $k_{\mathrm{ext}} > 0$. At any given altitude, the beam is assumed to be fully extinguished whenever the optical depth exceeds 2.

The IPHOC scheme used in one of our model runs predicts, among other quantities, the subgrid-scale distribution of total water content $p(q_t)$ within each of SAM's grid cells and a corresponding subgrid-scale cloud fraction $p_c = p(q_c > 0)$. We treat subgrid-scale cloudiness using a single-sample version of the Monte Carlo techniques used in global models (e.g., Klein and Jakob 1999; Räisänen et al. 2004). For each model column at each time we generate a random number $r \in [0, 1]$. Cells within the column are clear where $r < 1 - p_c$ and cloudy otherwise; cloudy cell condensate amounts are scaled by $1/p_c$ to preserve the cell mean. This approach uses the maximum overlap assumption, and results using random overlap are essentially the same. IPHOC does not treat precipitation, so we define a probability of precipitation using the greatest $p_c$ above each level to which it is vertically connected.

The lidar and radar cloud masks are merged to produce a final mask. This takes the value of the lidar mask when and where the lidar is known not to have been attenuated; otherwise, the value of the radar mask is used. The domain mean of this mask, computed at each model level at each observation time, is the probability of cloud $p$ (or cloud fraction) that we evaluate against the dichotomous observations.

### *b. Sensitivity to assumptions*

Radar reflectivity can be strongly influenced by the choice of drop size distribution (DSD). Our assumptions are consistent with those used elsewhere in SAM: we use the cloud drop and ice crystal size from the radiation scheme and ensure that precipitating hydrometeors follow the exponential Marshall–Palmer distribution (see Marchand et al. 2009). We assume exponential and lognormal DSDs for cloud ice and liquid, respectively. The QuickBeam radar simulator further assumes that all condensed species are spherical with density dependent on diameter only.

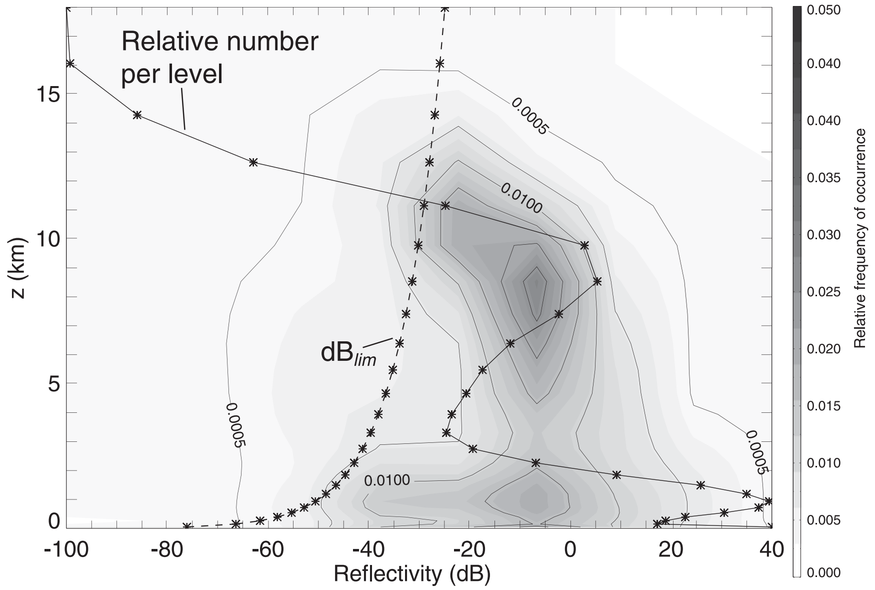

FIG. 1. The distribution of radar reflectivities (normalized by the total number of values $>-100$ dB), which we simulate with QuickBeam using cloud liquid, cloud ice, rain, snow, and graupel from all columns of a 2D 3-yr model run. The detection threshold dB$_{\text{lim}}$ (dashed curve with asterisks) does not pass through the most populated regions of the distribution. The relative number per model level (solid curve with asterisks) is normalized by the number occurring at the most populated level, which is close to 1 km.

These assumptions do not impact the results we show below. The underlying reason for this can be inferred from Fig. 1, which shows a probability distribution function (PDF) of $\sim 2.7 \times 10^6$ of the $\sim 2.4 \times 10^7$ reflectivities simulated for a 2D model run. The remaining values, all of which are $<-100$ dB, are assumed clear, a priori. Shown for comparison is the radar detection threshold dB$_{\text{lim}}$. With very few exceptions, the simulated reflectivities computed from the CSRM fields are either much greater than or much less than the detection threshold.

We have performed similar calculations using other plausible DSD assumptions (e.g., modified gamma cloud liquid and exponential ice as a function of temperature alone; Haynes et al. 2007). Although this changes some of the simulated reflectivities, most differences occur in values that are far away from the threshold, so our binary masks are not sensitive to them. At almost all altitudes, $<0.1\%$ of the binary decisions differ from those using the original DSD assumptions, with a maximum of between 1%–5% above 10 km in one test.

## 4. Measures of model performance

Performance metrics can sometimes be broken down into components estimating different aspects of a model's predictive ability. These are referred to as attributes (Murphy 1993) and in this work we make use of four, as follows: Bias is the correspondence between the mean forecast and mean observation. Reliability (REL) is the correspondence between the conditional mean observation and the conditioning predictions and so measures conditional bias; REL reduces to the usual (overall) bias in the absence of any conditioning. Resolution (RES) is the ability to resolve observed events into subsets with characteristically different outcomes. For a binary field this is exactly equivalent to skill in predicting the timing of events. Uncertainty (UNC) is the variance of the observations and is therefore independent of the model; large values of UNC make skillful forecasts difficult.

The RES of short-term weather forecast models is typically an order of magnitude less than their REL (Stanski et al. 1989). Unlike RES, REL can be improved with a posteriori calibration, wherein forecast probability values are statistically relabeled to improve correspondence with the observed field (Atger 2003). For this reason RES is regarded as the more intrinsic measure of model performance, but reducing REL is usually given priority.

### a. *Mean squared error and Brier's probability score*

A traditional measure of a model's performance in predicting a continuous variable $x$ is the mean squared error (MSE). Although this can be defined using expectations of a general estimator, in practice, it is more often defined for continuous variables only and broken down into terms of bias and random error as $\text{MSE} = \text{var}(p - x) + (\overline{p} - \overline{x})^2$, where $p$ denotes prediction, the overbar denotes the temporal mean, and var is the variance.

How can we compute the MSE of a predicted continuous variable where the individual observations $o$, like ARSCL, are dichotomous (i.e., 1 if the event occurs and 0 if it does not)? The usual solution is to transform the observations into a continuous field by invoking the ergodic hypothesis and averaging the observations over some time period, sampled by $m$ observations. This makes sense if the temporal (observational) and spatial (model) statistics are approximately equivalent; it also requires the identification of an appropriate choice of $m$. Our results below are based on hourly averages of observations made every 10 s, chosen to match the hourly time scale on which model data is reported; so we use $m = 360$.

Alternatively, we can use the instantaneous, dichotomous observations to calculate the Brier score $b$:

$$b = \frac{1}{N}\sum_{i}^{N}(p_i - o_i)^2. \tag{1}$$

The Brier score is used frequently in the verification of operational numerical weather prediction models. Although they are often discussed separately, $b$ can be defined as a limiting case of the MSE. As we decrease the averaging period, so that $m \rightarrow 1$, we find that the mean over this period $x_i \rightarrow o_i$, the $i$th instantaneous observation; that is,

$$\lim_{m\rightarrow 1}(\text{MSE}) = \frac{1}{N}\sum_{i}^{N}(p_i - o_i)^2 = b, \tag{2}$$

where $p_i$ is the probability of the original dichotomous event occurring.

In this work we exploit a decomposition of $b$ into components measuring key attributes of model performance, namely $b = \text{REL} - \text{RES} + \text{UNC}$. More formally, by dividing the probability range [0, 1] into $K$ probability classes (bins), we can write

$$b = \frac{1}{N}\sum_{k}^{K}n_k[(p_k - \overline{o}_k)^2 - (\overline{o}_k - \overline{o})^2] + \overline{o}(1 - \overline{o}), \tag{3}$$

where $\overline{o}_k$ is the mean observed frequency of occurrence for the $k$th class, containing $n_k$ events, and $\overline{o}$ is the observed climatology; other decompositions are also possible (Murphy 1996). The best possible value is $b = 0$, wherein REL = 0 and RES = UNC. However, this perfect score is only attainable by a model predicting (correct) extreme probabilities of 0 or 1, making the predictions deterministic; we can think of this as the asymptotic limit of a probabilistic model making increasingly confident predictions (Toth 2003).

Values of $b$ typically lie in the range [0.10, 0.25] for numerical weather model forecasts and scores $>0.3$ (in most cases) represent poor predictions; scores for forecasts of rare events tend to be better and will usually be $<0.10$ (Stanski et al. 1989). High skill is implied by $b <$ UNC.

### b. *Scores for a wide range of nudging time scales*

To give us some idea of how much variation to expect in the skill scores for the CSRM, we compute the bias, random error, MSE, and Brier score and its components for a set of 2D model runs using various nudging time scales $\tau_T$. The results, shown in Fig. 2, indicate the range of scores that might be expected among simulations with a range of skills.

The similarity between the MSE of the temporal cloud fraction and $b$ is very evident; both scores have almost identical variation with $z$ and $\tau_T$. This means that averaging the observations over time produces similar results to averaging over the model's domain (i.e., the ergodic hypothesis holds in these simulations). In turn, this validates our use of the MSE and its components here.

Above 9 km, the mean bias grows rapidly with $\tau_T$ and correlates with a monotonically increasing negative model bias in $T$ (not shown), although the MSE continues to be dominated by its random component $\sigma^2$. The mean bias, REL, and $\sigma$ all increase with $\tau_T$, but timing, as indicated by RES, is similarly poor for all $\tau_T$. At lower altitudes, RES decreases steadily with $\tau_T$ and demonstrates a larger range than REL.

## 5. Model skill in three configurations with varying computational cost

Here, we compare the performances of the 2D, 2D + IPHOC, and 3D model configurations. The 3D run is approximately a factor of 1000 more computationally expensive than the standard 2D run, and the IPHOC run around a factor of 4. One might reasonably imagine the scores of these configurations to reflect this. Figure 3 shows the aggregate scores and Fig. 4 shows those for events restricted to the periods April–September and October–March, which we refer to as summer and winter, respectively.

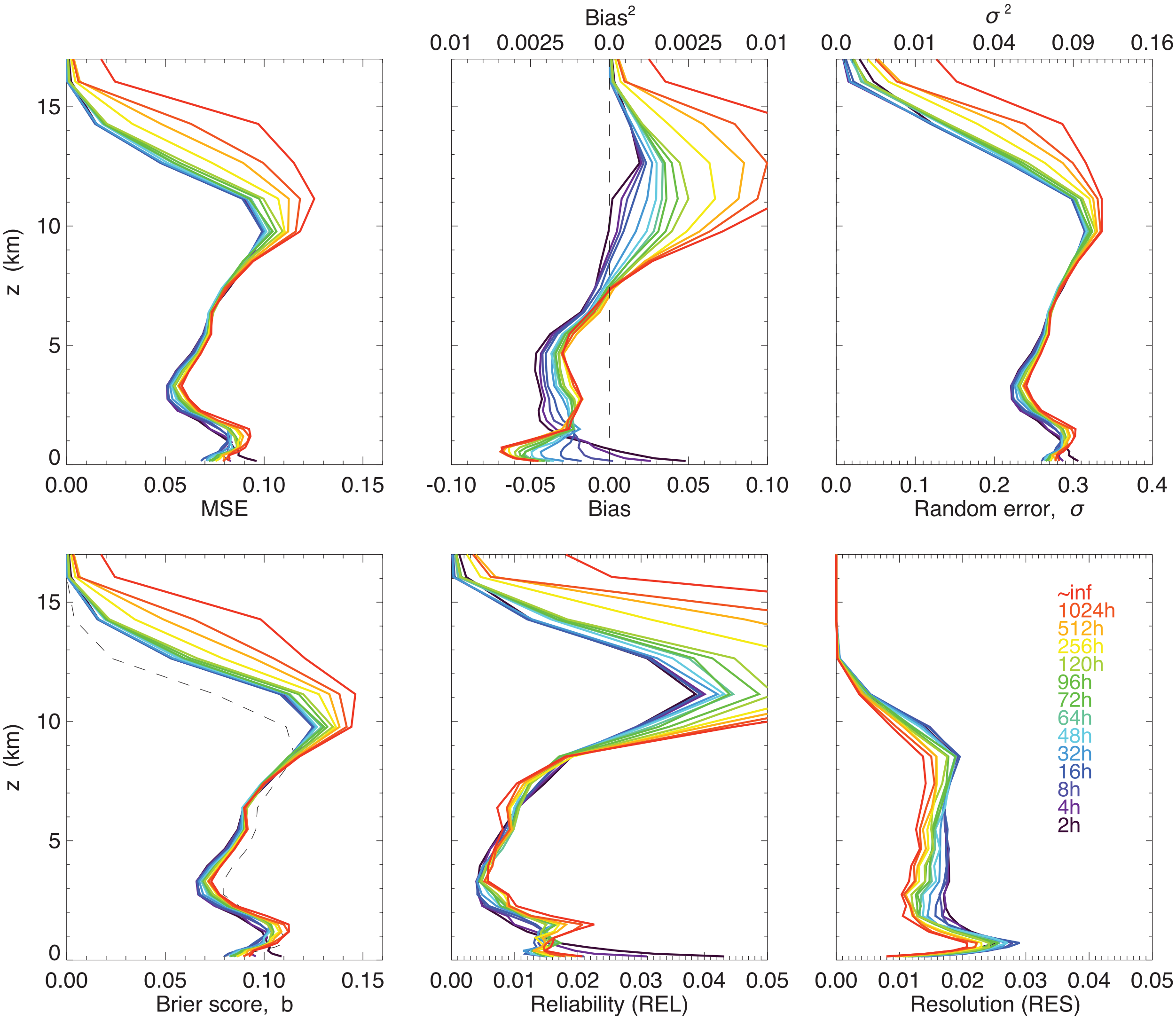

FIG. 2. Performance scores of 14 separate 3-yr runs at the SGP, each using different nudging periods, $\tau_T$ (note that inf denotes no nudging of $T$). Overall bias, $\sigma$, and MSE are with respect to observed temporal (hourly) cloud fraction. All other measures are with respect to observed instantaneous cloud occurrence. The dashed line in the plot of Brier score is UNC. Horizontal value ranges are fixed to allow direct comparison with the other figures and to be the same for REL and RES.

### a. *Aggregate scores*

The long-term performance of all three configurations is similar as measured by any score (see Fig. 3). The bias is low in all runs, suggesting that cloud is reasonably well calibrated overall; consequently, almost all of the MSE consists of random error. At many altitudes, the Brier scores attained by all configurations suggest considerable skill, with better performance below 8 km where $b <$ UNC.

In all configurations, the greatest conditional biases (REL) are seen for high cloud, particularly so in the IPHOC and 3D runs between 10 and 13 km. The 3D run is the most conditionally biased in the PBL; however, in all runs, bias here is typically a factor of 3–4 less than for high cloud. Performance in timing (RES) has a pronounced maximum in the PBL and drops off rapidly above 9 km.

### b. *Seasonal scores*

Model performance during summer and winter is markedly different, regardless of model configuration. The relative differences between the runs are also seasonally dependent (see Fig. 4). In fact, differences in skill between seasons are typically greater than differences between configurations during the same season. Given the range of computational costs of the configurations, this is surprising.

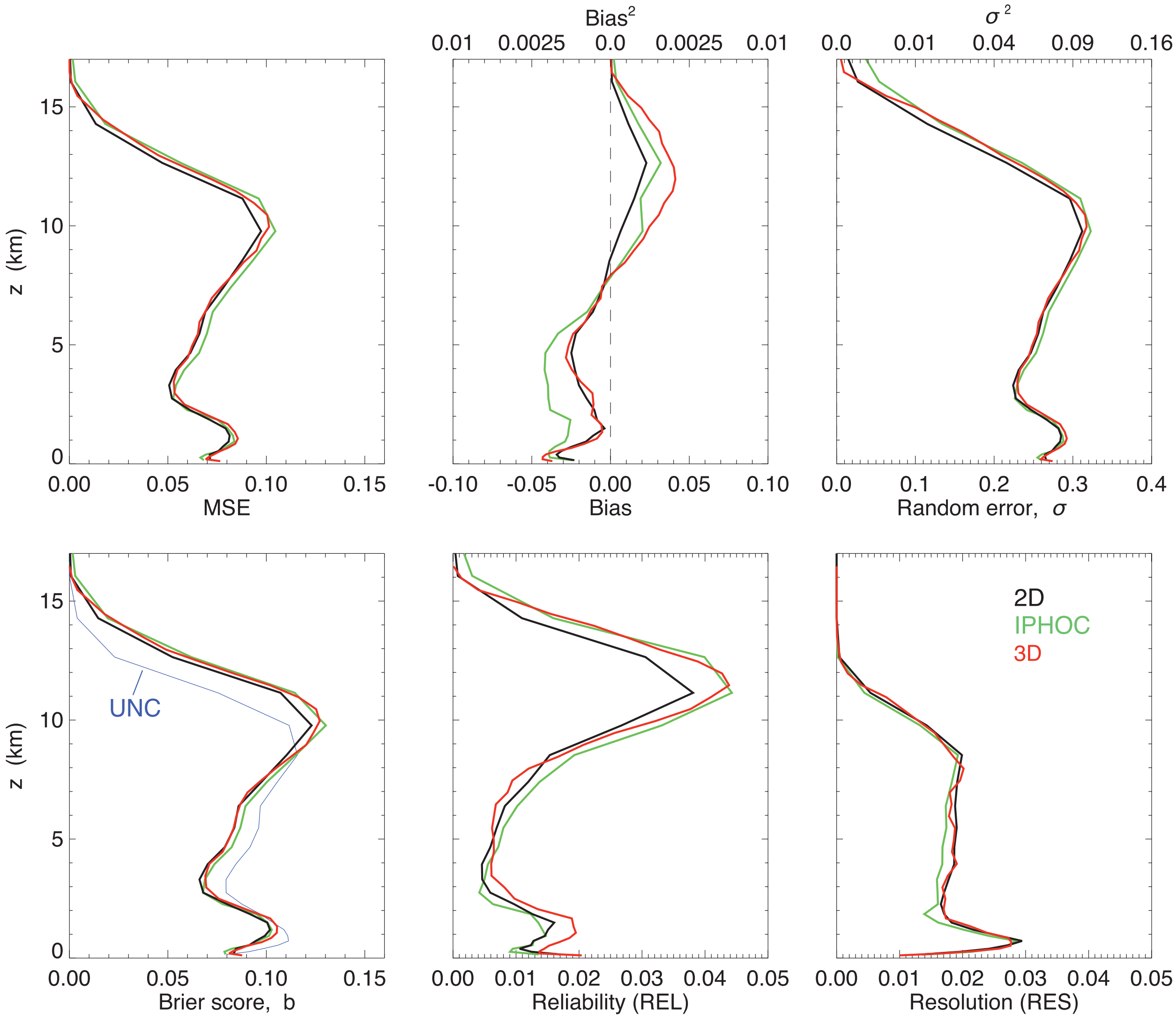

FIG. 3. Performance scores of 3-yr runs with different model configurations: (i) standard 2D domain, (ii) 2D + IPHOC, and (iii) higher-resolution 3D domain. All runs use $\tau_T = 24$ h.

Model errors are typically greater in winter than in summer; the most notable exception to this is worse timing below 9 km in summer. At some altitudes (e.g., PBL for the 3D run) the overall bias is comprised of opposing seasonal biases.

We now look in more detail at conditional bias and timing as a function of $p$ close to 1 km and 10 km. These altitudes are of interest for a number of reasons: (i) they are close to the maxima in observed cloud; (ii) large seasonal differences in performance occur here; (iii) relatively large differences exist here between the scores of different model configurations; and (iv) the variance of observed cloud (UNC) is approximately the same for the boundary layer and for high cloud, which means that the Brier scores at these altitudes are directly comparable.

By conditionally sampling the observations using the forecasts of cloud fraction (i.e., the probability of cloud, $p$), we can construct attribute diagrams for each altitude and season (Fig. 5). The diagrams plot observed frequencies of cloud occurrence against the corresponding forecast probabilities, as well as forecast distributions and observed climatologies. The diagrams are augmented with information about the relative contributions that each range of $p$ makes to the conditional bias (REL), which is ideally zero, and the skill in timing (RES), which is ideally large.

### c. *Skill as a function of forecast cloud fraction*

Extra wintertime conditional bias in the PBL and high cloud are two of the main seasonal differences

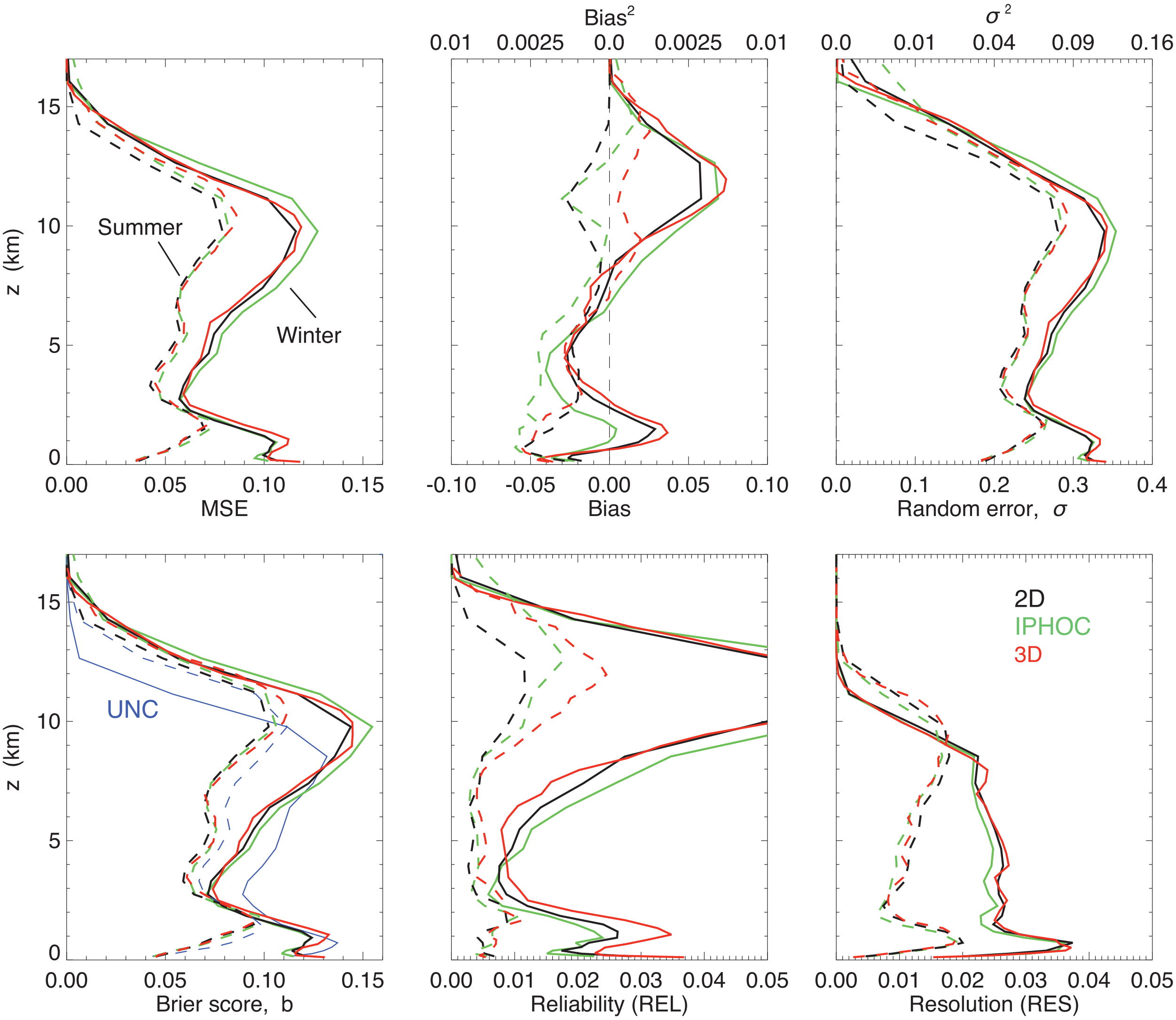

FIG. 4. As in Fig. 3, but restricted to events from April to September (dashed), covering boreal summer, and events from October to March (solid), covering boreal winter. See the comment on horizontal value ranges in Fig. 2.

identified by the scores; here, we investigate which parts of the forecast distribution are responsible. For high cloud, all configurations overforecast most $p$ (indicated by values lying below the 1:1 line of the attribute diagrams), particularly in winter. In the PBL, all configurations overforecast mid–high $p$, but also underforecast low $p$ (above the 1:1 line), particularly in summer.

In summer, high cloud is actually most conditionally biased in the 3D run. This is due to both the extra overforecasting of mid–high $p$ (i.e., greater REL for this range of $p$) and the greater population of this probability range, rather than to predicting more $p = 1$ events, which occur more often in the other model runs. Even though slightly more cloud is observed at 10 km in the summer, in all runs the number of $p = 1$ predictions at this altitude is an order of magnitude less than in winter.

In all runs, high cloud is substantially more conditionally biased (greater REL) in winter than in summer. No rapid increase in overforecasting occurs for high $p$ (a distinctive feature of the summertime attributes diagram) but the contributions from this range (high $p$) are responsible for the seasonal differences. The greatest contributions to REL are for $p = 1$ predictions because these are more numerous—IPHOC forecasts most and is therefore the most biased here. Although we do not show the attributes diagram for 7 km, the 3D run performs best here, and IPHOC performs worst (see Fig. 4), because of the extra number of $p = 1$ predictions in the

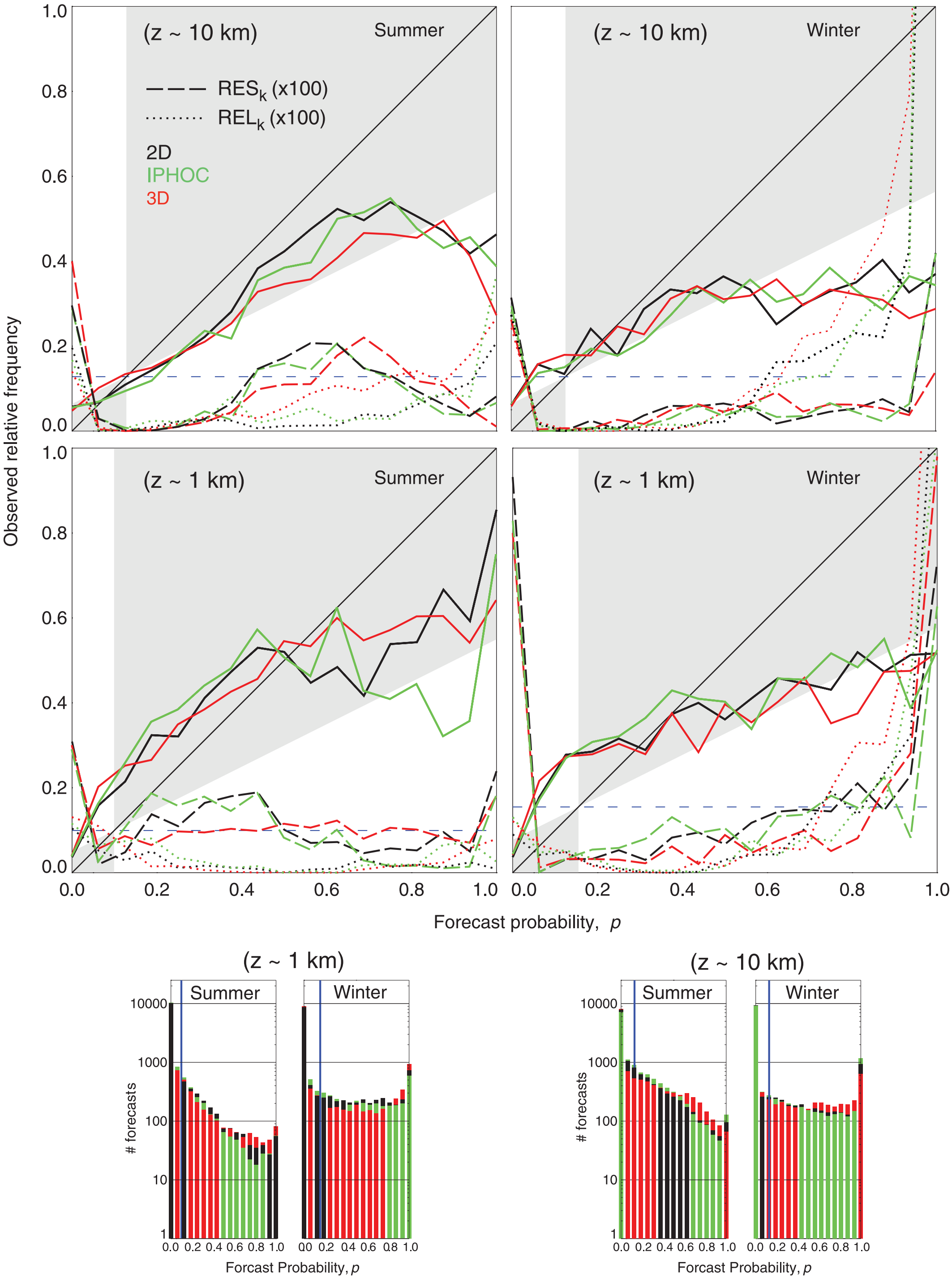

FIG. 5. Augmented attributes diagrams covering boreal summer and winter for high and low cloud. The distance of the solid lines from the 1:1 line indicates reliability (REL) and their distance from the observed climatologies (blue dashed line) indicates resolution (RES). Solid lines within the shaded regions demonstrate positive skill. Contributions from each of the $K$ probability bins were defined in Eq. (3). Ideally, REL$_k$ is 0 and RES$_k$ is large.

3D run, since the observed frequencies corresponding to these predictions are similar in both runs.

For any given season, the timing of cloud is remarkably similar in all runs; the smallest seasonal differences are at 10 km. Here, most contributions to good timing (high RES) come from forecasts of clear sky, intermediate cloud fractions in summer, and 100% cloud fractions in winter.

The PDFs of wintertime forecasts in the PBL are similar to those of high cloud, with the conditional bias again being determined by the number of high $p$ predictions; however, here the 3D run is worst because it has the greatest number of $p = 1$ forecasts. In the PBL, the greatest number of clear and least mid–high $p$ forecasts occur during summer, which consequently demonstrates the least conditional bias in all configurations.

Generally, more seasonal differences in timing are seen in the PBL than aloft, with greater contributions being made by predictions of clear and overcast skies ($p = 0, 1$) in winter than in summer. This gives rise to the better timing of wintertime cloud here, most of which is attributable to the increased number of predictions of high $p$. The same explanation also applies to the worse timing found in most cloud below 9 km in summer.

## 6. Discussion and conclusions

### a. How skilled is this cloud-resolving model at predicting cloud occurrence?

Figure 3 demonstrates that SAM has considerable skill in predicting cloud occurrence at the ARM SGP site, when the thermodynamic state is constrained to remain near the observations. Bias in modeled cloud occurrence is relatively small for all three model configurations, suggesting that this field is relatively well calibrated. The Brier scores attained by all configurations also suggest considerable model skill at many altitudes, when compared to those typically achieved by numerical weather models (Stanski et al. 1989). Verification of model skill across such a wide range of atmospheric states puts the use of CSRMs as benchmark calculations, or as components in multiscale models, on a firmer footing. However, this is only true to the extent that the model can be relied upon to maintain a realistic thermodynamic state in the absence of nudging. We are currently working to chart model skill as a function of forecast lead time in unconstrained versions of the model and to investigate how tightly this skill is coupled to errors in temperature and humidity.

The similar behavior of the MSE and $b$ for cloud occurrence suggests that the ergodic hypothesis holds for these simulations. This means that we can approximate the first few statistical moments of the instantaneous, spatial cloud fraction with pointlike, temporal cloud fraction and vice versa, thus validating the application of the traditional measures. We do not expect that this result is universally true.

Model skill varies substantially between seasons, as judged by comparison of the skill in a single configuration run across a large range of nudging time scales. This, together with the range of scores seen for different nudging time scales, suggests that the scores (particularly REL and RES) are sufficiently sensitive to be able to identify differences in model performance and increases our confidence in interpreting similar scores as real similarity in model performance rather than as a lack of precision of the measures.

### b. Model skill and computational cost

Model skill in each of the three configurations is remarkably similar, despite a wide range in the computational cost (reflecting the conceptual sophistication or spatial resolution) among the configurations. Although CSRM behavior can depend strongly on dimensionality in some circumstances (e.g., Petch and Gray 2001; Petch 2006), our results are consistent with multiweek studies of SAM (Khairoutdinov and Randall 2003). While this result is favorable for use of this CSRM in SP-CAM, it also suggests that the cloud model's deficiencies are deeper than can be ameliorated by simple changes.

One might expect a strong dependence on spatial resolution or dimensionality, since moving from 2D to 3D or increasing spatial resolution allows for a better representation of model dynamics, but this is not reflected in the skill scores for different model configurations. We hypothesize that the distribution of cloud is quite sensitive to the model thermodynamic state, which is constrained to be roughly the same in all three configurations.

Some of the similarities found could also be attributable to the effects of nudging. However, low correlations ($<0.2$) between pointwise thermodynamic and cloud errors (not shown) suggests that cloud errors are not significantly influenced by nudging temperature.

It is also possible that all configurations show approximately equal skill because we are only looking at (binary) cloud occurrence and that differences may exist in the structure of the (continuous) hydrometeor fields. However, the mean total condensate is very similar in all three cases (not shown), suggesting that this is not the case.

Errors in the forcing data could also have such a large influence on thermodynamics that they reduce scores enough to hide comparatively smaller interconfigurational differences. Sources of error include the use of area-averaged surface precipitation, which has the most influence over advective tendencies, and spatial-scalealiasing in fields with large subgrid-scale variability, such as water vapor and winds, and severe weather. This might be explored by constructing multiple physically consistent sets of forcing data, spanning the range of uncertainty in the observed fields (Zhang et al. 2001), which could be used to drive a CSRM ensemble to study sensitivities to observational uncertainty (Hume and Jakob 2007).

Skill scores such as MSE and the Brier score are metrics without much diagnostic utility and so are mute on the subject of what changes might improve model performance. Model developers must find other sources of inspiration in their efforts to improve the model. Once candidate improvements have been identified, however, skill scores provide an objective, context-independent way of evaluating any improvement in performance.

In the future, probabilistic metrics could also be applied to other CSRM fields, both binary (e.g., types of precipitation) and continuous (e.g., cloud liquid and ice content).

### c. Model skill and seasonality

In these simulations the model shows greater differences in performance between seasons than between model configurations during the same season, despite substantial differences between the computational costs of the configurations. In particular, there is greater conditional bias but better timing of PBL cloud in winter and substantially less conditional bias in high cloud during summer.

During winter, there is more stratiform cloud in the PBL and large-scale frontal cloud systems, whereas in summer, there is more small-scale shallow convection, deep convection, and anvil cirrus. The local and more intermittent nature of summertime cloud may be the main reason for worse timing in the summer season. The contribution that clear-sky predictions make to the timing of low cloud is greater in winter than in summer because the mean observed cloud fraction is higher in this season. Conditional sampling on variables relevant to the predominant cloud type may shed some light on the mechanisms responsible.

Little cloud is observed above 10 km in winter, but more occurs in summer; it is interesting that the greatest interconfigurational differences in conditional bias are found here, such that more sophisticated configurations are most biased. Inspection of the attribute diagrams for these altitudes (not shown) confirms that this is due to the extra number of 100% cloud fraction predictions, compared to those of the standard 2D configuration—more such predictions are actually made in winter, however, more so in all runs. Identification of the circumstances under which the 3D and IPHOC runs produce extra very high cloud may provide further insight.

*Acknowledgments.* We are most grateful to Marat Khairoutdinov for providing and supporting SAM; Anning Cheng and Kuan-Man Xu for implementing IPHOC; NOAA-HPCS for use of Jet and Christopher Harrop for advice; Roger Marchand for help with the radar simulator and Patrick Hofmann for helping build a QuickBeam interface; and Eugene Clothiaux, Christian Jacob, and Tom Hamill for discussion and insight. This research was supported by the Office of Science (BER), U.S. Department of Energy, under Grants DE-FG03-06ER64182 and DE-FG02-03ER6356.

## REFERENCES

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.,* **131,** 1509–1523.

Bodas-Salcedo, A., M. J. Webb, M. E. Brooks, M. A. Ringer, K. D. Williams, S. F. Milton, and D. R. Wilson, 2008: Evaluating cloud systems in the Met Office global forecast model using simulated CloudSat radar reflectivities. *J. Geophys. Res.,* **113,** D00A13, doi:10.1029/2007JD009620.

Cheng, A., and K.-M. Xu, 2008: Simulation of boundary-layer cumulus and stratocumulus clouds using a cloud-resolving model with low- and third-order turbulence closures. *J. Meteor. Soc. Japan,* **86A,** 67–86.

Clothiaux, E., T. P. Ackerman, G. G. Mace, K. P. Moran, R. T. Marchand, M. A. Miller, and B. E. Martner, 2000: Objective determination of cloud heights and radar reflectivities using a combination of active remote sensors at the ARM CART sites. *J. Appl. Meteor.,* **39,** 645–665.

Ghan, S. J., L. R. Leung, and J. McCaa, 1999: A comparison of three different modeling strategies for evaluating cloud and radiation parameterizations. *Mon. Wea. Rev.,* **127,** 1967–1984.

Haynes, J., R. Marchand, Z. L. A. Bodas-Salcedo, and G. Stephens, 2007: A multi-purpose radar simulation package: QuickBeam. *Bull. Amer. Meteor. Soc.,* **88,** 1723–1727.

Hume, T., and C. Jakob, 2007: Ensemble single column model validation in the tropical western Pacific. *J. Geophys. Res.,* **112,** D10206, doi:10.1029/2006JD008018.

Jakob, C., R. Pincus, C. Hannay, and K.-M. Xu, 2004: Use of cloud radar observations for model evaluation: A probabilistic approach. *J. Geophys. Res.,* **109,** D03202, doi:10.1029/2003JD003473.

Khairoutdinov, M., and D. Randall, 2003: Cloud resolving modeling of the ARM summer 1997 IOP: Model formulation, results, uncertainties, and sensitivities. *J. Atmos. Sci.,* **60,** 607–625.

——, ——, and C. DeMott, 2005: Simulations of the atmospheric general circulation using a cloud-resolving model as a superparameterization of physical processes. *J. Atmos. Sci.,* **62,** 2136–2154.

Klein, S. A., and C. Jakob, 1999: Validation and sensitivities of frontal clouds simulated by the ECMWF model. *Mon. Wea. Rev.,* **127,** 2514–2531.

Lock, A. P., A. R. Brown, M. R. Bush, G. M. Martin, and R. N. B. Smith, 2000: A new boundary layer mixing scheme. Part I: Scheme description and single-column model tests. *Mon. Wea. Rev.,* **128,** 3187–3199.

Marchand, R., J. Haynes, G. G. Mace, T. Ackerman, and G. Stephens, 2009: A comparison of simulated cloud radar output from the multiscale modeling framework global climate model with CloudSat cloud radar observations. *J. Geophys. Res.,* **114,** D00A20, doi:10.1029/2008JD009790.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, 1996: General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. *Wea. Forecasting,* **124,** 2353–2369.

Petch, J. C., 2006: Sensitivity studies of developing convection in a cloud-resolving model. *Quart. J. Roy. Meteor. Soc.,* **132,** 345–358.

——, and M. E. B. Gray, 2001: Sensitivity studies using a cloud-resolving model simulation of the tropical west Pacific. *Quart. J. Roy. Meteor. Soc.,* **127,** 2287–2306.

Räisänen, P., H. W. Barker, M. F. Khairoutdinov, J. N. Li, and D. A. Randall, 2004: Stochastic generation of subgrid-scale cloudy columns for large-scale models. *Quart. J. Roy. Meteor. Soc.,* **130,** 2047–2067.

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

——, and Coauthors, 2003: Confronting models with data: The GEWEX cloud system study. *Bull. Amer. Meteor. Soc.,* **84,** 455–469.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: A survey of common verification methods in meteorology. WMO Tech. Rep. 8, 358 pp.

Toth, Z., 2003: Probability and ensemble forecasting. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley & Sons, 137–163.

Xie, S., R. T. Cederwall, and M. Zhang, 2004: Developing long-term single-column model/cloud system–resolving model forcing data using numerical weather prediction product constrained by surface and top of the atmosphere observations. *J. Geophys. Res.,* **109,** D01104, doi:10.1029/2003JD004045.

Zhang, M. H., J. L. Lin, R. T. Cederwall, J. J. Yio, and S. C. Xie, 2001: Objective analysis of ARM IOP data: Method and sensitivity. *Mon. Wea. Rev.,* **129,** 295–311.