Use of cloud radar observations for model evaluation: A probabilistic approach

Christian Jakob
Bureau of Meteorology Research Centre, Melbourne, Australia

Robert Pincus and Cécile Hannay
Climate Diagnostics Center, National Oceanic and Atmospheric Administration Cooperative Institute for Research in Environmental Sciences, Boulder, Colorado, USA

Kuan-Man Xu
NASA Langley Research Center, Hampton, Virginia, USA

Received 30 January 2003; revised 21 October 2003; accepted 17 November 2003; published 5 February 2004.

[1] The use of narrow-beam, ground-based active remote sensors (such as cloud radars and lidars) for long-term observations provides valuable new measurements of the vertical structure of cloud fields. These observations might be quite valuable as tests for numerical simulations, but the vastly different spatial and temporal scales of the observations and simulation must first be reconciled. Typically, the observations are averaged over time and those averages are claimed to be representative of a given model spatial scale, though the equivalence of temporal and spatial averages is known to be quite tenuous. This paper explores an alternative method of model evaluation based on the interpretation of model cloud predictions as probabilistic forecasts at the observation point. This approach requires no assumptions about statistical stationarity and allows the use of an existing, well-developed suite of analytic tools. Time-averaging and probabilistic evaluation techniques are contrasted, and their performance is explored using a set of "perfect" forecasts and observations extracted from a long cloud system model simulation of continental convection. This idealized example demonstrates that simple time averaging always obscures forecast skill regardless of model domain size. Reliability diagrams are more robust, though scalar scores derived from the diagrams are sensitive to the forecast probability distribution. Forecasts by cloud system and weather forecasting models then provide examples as to how probabilistic techniques might be used in a variety of contexts. *INDEX TERMS:* 3337 Meteorology and Atmospheric Dynamics: Numerical modeling and data assimilation; 3360 Meteorology and Atmospheric Dynamics: Remote sensing; 3394 Meteorology and Atmospheric Dynamics: Instruments and techniques; *KEYWORDS:* model evaluation, probabilistic verification, cloud radar

Citation: 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.

1. Comparing Observations at a Point to Model Forecasts

[2] The last decade has seen an enormous increase in the diversity and ubiquity of long-term active remote sensing sites. These observatories typically contain a suite of colocated passive and active remote sensors, including millimeter wavelength cloud radars and lidars, that can be combined to provide information on cloud microphysical properties [e.g., *Platt*, 1979; *Mace et al.*, 1998a, 2001; *Sassen and Mace*, 2002]. The data provided by these sites is unique: active sensors provide valuable information on the vertical structure of clouds and the atmosphere, while the long-term records available from some sites sample an enormous range of conditions.

[3] Long time-series from ground-based sites offer new opportunities for evaluating the treatment of clouds in numerical models of the atmosphere. Models range in scale and resolution from global models used for weather and climate prediction (with resolutions of tens to hundreds of kilometers) to cloud system models encompassing domains a few hundred kilometers across at resolutions of a few kilometers. Model predictions may be statistical (like climate projections, which are boundary value problems) or deterministic (like weather forecasts, which are initial condition problems), depending on the ways large-scale observations are used to initialize or force the model. Different kinds of predictions may require different techniques of evaluation, but no matter what kind of prediction is being

0148-0227/04/2003JD003473$09.00

tested there is an inherent mismatch between the spatial and temporal scales of models and ground-based observations. Many active ground-based instruments make frequent observations of a vertical pencil beam through the atmosphere at high vertical resolution [e.g., *Clothiaux et al.*, 2000]. Models, on the other hand, predict the distribution of cloud properties within much larger domains, be they a single grid column in a global model or the full domain of a cloud system model.

[4] How can we robustly compare measurements at a point to instantaneous predictions in a large domain? The usual answer is to leave the model predictions untouched and temporally average the observations, assuming that advection over time provides the same statistics as would be gathered from observing instantaneous spatial variability [e.g., *Barnett et al.*, 1998]. The averaging time is chosen to match the model domain size based on the cloud propagation speed. In principle, averaging intervals could vary according to the propagation speed with time, height, and the size of the model domain being evaluated, but this is rarely done in practice; most studies average over fixed time-intervals, even though the resulting statistics can depend significantly on which interval is chosen [e.g., *Hogan and Illingworth*, 2000].

[5] This paper explores a complementary route to model comparison with point-wise observations: leaving the observations untouched and interpreting model predictions as probabilistic forecasts at the observation point. This rethinking opens up new avenues for model evaluation, since a rich suite of techniques related to probabilistic forecasts already exists and is used extensively by the ensemble forecasting community (among others). Probabilistic verification is restricted to forecasts (predictions of specific events at specific times) of quantities for which statistical information is available within the domain but requires fewer ad hoc assumptions than does time averaging.

[6] This exploratory study contains two parts. The first (sections 2 and 3) contrasts time-averaging and probabilistic evaluation methods and examines their performance in an idealized context. The second (section 4) shows examples of using probabilistic techniques to test forecasts by various classes of models against a set of remote sensing observations. Issues resulting from these findings are discussed in section 5.

2. Time-Averaging and Probabilistic Techniques for Using Point Measurements

[7] The general evaluation problem at hand is comparing point observations taken frequently in time to model predictions of the same (or another, closely related) quantity. Our goal is to use the observations to quantitatively evaluate the model's forecasts so we can tell, for instance, if a change to the model has resulted in better predictions. As a concrete example we consider forecasts of total (vertically projected) cloud cover within some model domain, which we would like to evaluate using a time series of observations of the presence or absence of cloud above a point within the domain, as might be observed by cloud radars.

[8] Typically, such point observations of cloud occurrence are averaged in time to produce a time series of total cloud cover [e.g., *Hogan et al.*, 2000] varying from zero to one that is directly comparable to the model prediction. The model and observational time series can then be compared with statistical measures such as mean error, root-mean square error, and correlation. We refer to this as the time-averaging approach.

[9] Alternatively, model predictions of total cloud cover at each time may be treated as a probabilistic forecast. If we assume that clouds are distributed randomly throughout the model domain, we can interpret a cloud cover forecast as being the probability that, at any observation point within the domain, there will be cloud overhead at the forecast time. Forecasts are deemed successful if cloud is observed P% of the times when the forecasts of cloud cover is P%. The degree of success can be quantified using the methods usually applied to the evaluation of probabilistic forecasts and ensemble prediction systems [e.g., *Mullen and Buizza*, 2001; *Buizza et al.*, 1999; *Toth et al.*, 1997]. We refer to this as probabilistic model evaluation.

[10] We focus here on two specific measures: reliability diagrams and the relative operating characteristics (ROC) diagram, as well as the scalar scores associated with each diagram. To construct these diagrams, observations are sorted according to the model forecast of total cloud cover and the observations in each class are considered in aggregate. Skillful forecasts are those in which the correct value is predicted at the correct time; the ROC scores emphasize agreement in timing, while the reliability diagram scores account for mistakes in magnitude as well as timing. (A more complete introduction to reliability and ROC diagrams and their associated scalar scores can be found in Appendix A.)

[11] The central advantage to probabilistic evaluation methods is that the observations do not need to be averaged in any way, which is especially appealing when point observations are used. However, because the verifying observations are analyzed according to the forecast probability, probabilistic evaluation methods can only be applied to predictions that are statistical in nature, i.e., those in which the distribution of the forecast quantity is resolved to some degree. When evaluating contemporary large-scale models, point measurements are usually compared to single columns, in which case cloud cover and variables related to its vertical overlap (e.g., optical thickness) are the only quantities for which statistical information is currently available at each forecast time. (Cloud cover predictions are inherently statistical, even within a single grid column, in the sense that a single value defines a binary probability distribution function.) Probabilistic evaluation may prove most valuable for testing cloud system models: since point observations can be compared to the distribution of forecast quantities within the entire domain, any variable may be evaluated.

3. Optimal Behavior of Time-Averaging and Probabilistic Model Evaluation

[12] Model evaluation techniques are most useful if they can robustly distinguish between accurate and inaccurate forecasts without introducing their own biases or sensitivities. In this section we investigate the performance of time-averaging and probabilistic model evaluation techniques in a context in which the forecasts are known to be perfect. We construct an idealized pair of point observations and model

forecasts from a single data set, then apply the evaluation techniques. The ways in which evaluation results vary with spatial and temporal scale helps us understand the limits of the scoring techniques themselves.

3.1. Constructing "Perfect" Forecasts

[13] We generate a set of forecasts and pseudo-observations using a cloud system model (CSM) simulation of summertime deep convection over the Atmospheric Radiation Measurement (ARM) [*Stokes and Schwartz*, 1994] Southern Great Plains (SGP) site in Oklahoma. We use the UCLA/CSU cloud system model [*Krueger*, 1988; *Xu and Krueger*, 1991], configured as a two-dimensional 512 km domain with horizontal grid spacing of 2 km and 35 vertical levels on a stretched grid. The model is run continuously for 29 days while being driven by large-scale forcing observed during the Summer 1997 Intensive Observation Period from 19 June to 17 July [*Xu et al.*, 2002]. Snapshots of the model state are reported every 5 min. We compare model forecasts with instantaneous values drawn from a single CSM column near the center of the domain. Because the pseudo-observations are drawn directly from the population that defines the forecasts at each time step, the forecasts are ''perfect'' by definition. If evaluation techniques applied to this set of forecasts and observations indicate less than perfect agreement or vary with evaluation parameters, we have identified a weakness in the evaluation technique.

[14] The primary cloud-measuring instrument at the ARM SGP is the millimeter cloud radar, which is sensitive to precipitation (large particles) as well as cloud condensate [e.g., *Mace et al.*, 1998b; *Hogan et al.*, 2000]. To simulate observations by this instrument, we compute the radar reflectivity due to all forms of condensed water in each model grid cell following *Luo et al.* [2003]; cells with reflectivities greater than -40 dBZ are considered to contain hydrometeors. Microphysical calculations in the model are not performed when the mixing ratio is less than 10^{-6} kg/kg, so we ignore cells with lower values of mixing ratio. Pseudo-observations are obtained by extracting the hydrometeor occurrence profile at each time step from a CSM grid column. We construct forecasts of domain-averaged hydrometeor cover (HC; the fraction of points at each level within a time-space window with reflectivity greater than -40 dBZ) and total hydrometeor cover (THC; the fraction of columns containing hydrometeors anywhere in their vertical extent) by spatially averaging the CSM fields. Because precipitation only occurs under existing clouds, total hydrometeor cover is essentially identical to total cloud cover.

[15] We test the assumption that clouds are randomly distributed within the CSM domain by examining the dependence of mean THC over the 29 day period to time-averaging interval and domain size. The variation in mean THC in domains of five different sizes (32, 64, 128, 256, and 512 km) extracted from the CSM domain is small (in the range 0.50–0.515). Clouds are also well sampled in time: mean THC computed every hour using four time windows (5 and 10 min; 1 and 3 hours, with substantial overlap in the 3 hour calculation) is almost constant (0.504 to 0.507).

[16] Evaluation methods should reflect model performance and in particular should not depend on the size of the model domain being evaluated. We test for insensitivity to domain size in the sections below by examining model skill in five domains of varying size (32, 64, 128, 256, and 512 km, as above) extracted from the entire CSM simulation as a proxy for the domain of the model being tested (i.e., the grid resolution in a global model, the domain size of a CSM, etc.). Again, because we have constructed the pseudo-observations by sampling the forecasts and because the clouds are randomly distributed in space and time in the domain, the forecasts are perfect by definition, and any changes in forecast skill with domain size indicated by the techniques are spurious.

3.2. Sensitivity of Time-Averaging Evaluations to Domain Size and Averaging Time

[17] Given a set of point observations and a model of some specified domain size, time-averaging evaluations must first choose a time interval over which to average the observations. We explore the degree to which measures of agreement depend (artificially) on this value and on the model's domain size by constructing 20 pairs of hourly time-series by averaging the pseudo-observations of THC over one of the four time windows and comparing these to averages over the five domains sizes.

[18] The correlation and standard deviation of each pair of pseudo-observations and forecasts are compared in Figure 1 in a diagram devised by *Taylor* [2001]. Each time series of domain-averaged THC is used as a forecast vector and the time-averaged THC in the single CSM column is the reference (observation) vector. Forecasts and observations are said to agree perfectly when they have the same standard deviation and a correlation of 1; this is represented by a point at radius 1 on the abscissa. Unfortunately, the results of our model evaluation using time-averaging depend on the combination of domain size and averaging interval, and no combination used here indicates that the forecasts are indeed perfect, as these are by construction.

[19] One could, in principle, determine from Figure 1 either the optimal time-averaging interval (color) for a given model domain size (letter) or the optimal domain size for a given averaging interval. In large domains (512 km = A and 256 km = B) none of the averaging intervals are long enough to provide a good comparison, while the optimum averaging interval for the 128 km domain (C) lies somewhere between 1 and 3 hours. Equivalently, it appears that a 3 hour averaging interval (red) might be optimally compared to a domain size between 128 and 256 km. However, these ''best matches'' of domain size and time-averaging intervals are misleading. Not only does the space-time match depend on the meteorological conditions (e.g., wind speed, presence of convection), but these particular results may be a model artifact altogether. What Figure 1 underlines is that when comparing model forecasts to time-averaged observations, any constant, arbitrarily chosen time-averaging interval is almost certain to degrade measures of model skill by some unknowable amount.

3.3. Sensitivity of Probabilistic Evaluations to Domain Size

[20] Rather than averaging the point observations over time, probabilistic model evaluation begins by grouping the observations according to classes in the forecast probability.

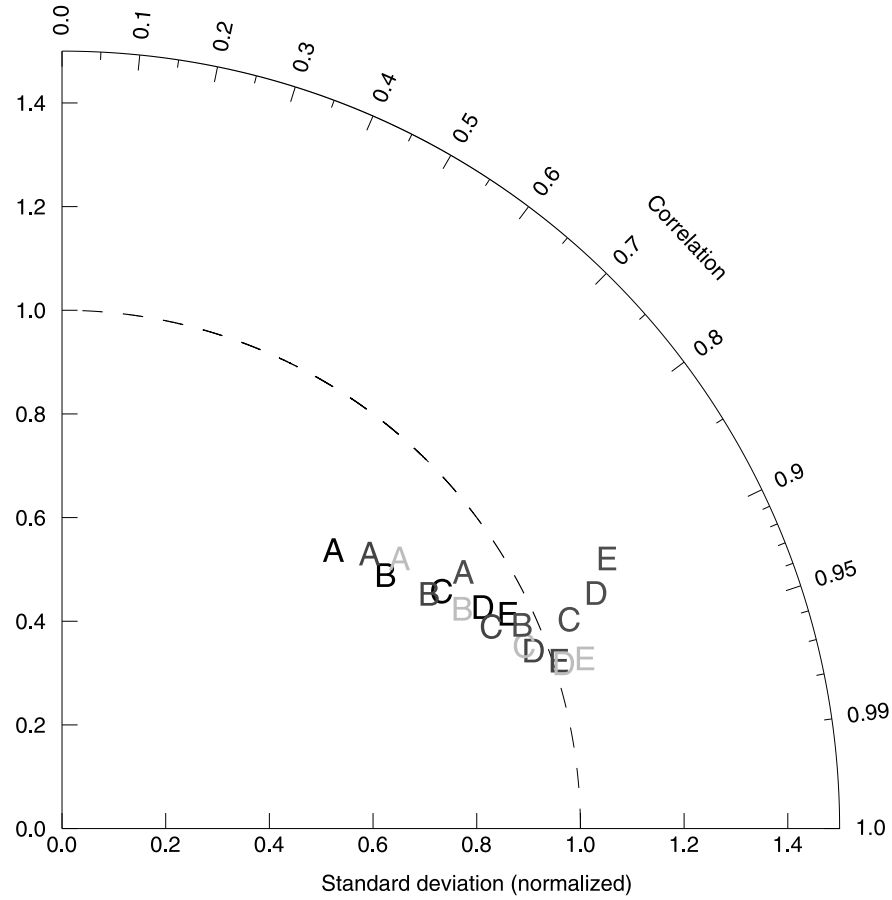

Figure 1. Taylor diagram comparing hourly CSM THC time series of all model domains (A = 512 km, B = 256 km, C = 128 km, D = 64 km, E = 32 km) with pseudo-observations drawn from the CSM (see text) using different time intervals (black = 600 s, blue = 1200 s, orange = 3600 s, red = 10800 s). Perfect agreement lies at radius 1 on the abscissa. The forecast skill determined using time-averaging depends strongly on the combination of averaging interval and domain size even when all the forecasts are known to have equal (and perfect) skill. See color version of this figure in the HTML.

Here we divide forecasts of THC into 10 classes of width 0.1 and evaluate the forecasts using reliability diagrams, the closely related scalar Brier score and its components, and the relative operating characteristics (ROC) diagram. (Readers unfamiliar with these techniques may see Appendix A for a brief introduction.) Reliability diagrams show the observed frequency of occurrence of an event (y-axis) for each class of forecast probabilities (x-axis). Forecast skill consists of making the correct prediction at the correct time, and reliability diagrams are useful in part because biases and errors in timing can be clearly identified. Unbiased forecasts have as many points (weighted by the forecast probability distribution) below the diagonal as above. Well-timed forecasts run parallel to the diagonal, while randomly timed unbiased forecasts track the horizontal line (the sample climate value, determined as the mean over the entire time period).

[21] Reliability diagrams and forecast probabilities for three proxy model domain sizes are shown in Figure 2. In this example the forecast probability is the predicted THC, while the observations are the frequency of occurrence of hydrometeors at the pseudo-observation point. In this case, forecasts are accurate if about P% of the observations made at times when the forecast is P% cloud cover do indeed contain cloud so that the curve tracks the diagonal. Points lying within the shaded area contribute positively to forecast skill.

[22] Forecast probabilities (right panels) in our example depend strongly on domain size, with larger domains exhibiting broad distributions of THC and forecasts in smaller domains tending to be near zero or one. (Were the domain reduced to a single CSM column, only values of THC of 0 or 1 would be possible.) Nonetheless, the reliability diagrams indicate that the forecasts are skillful at all domain sizes. Since the forecasts are perfect by construction, we conclude that the reliability diagram is robust to changes in the spatial scale at which it is applied.

[23] Scalar forecast scores derived from these diagrams, however, are not so forgiving. The Brier score (BS) can be decomposed into three components: reliability (REL), resolution (RES), and uncertainty. (Appendix A provides more details.) Perfect forecasts are traditionally defined as those for which BS = 0, REL = 0, with resolution large and equal to the uncertainty term so that the Brier skill score BSS = 1. These scores appear to be unachievable in practice, however.

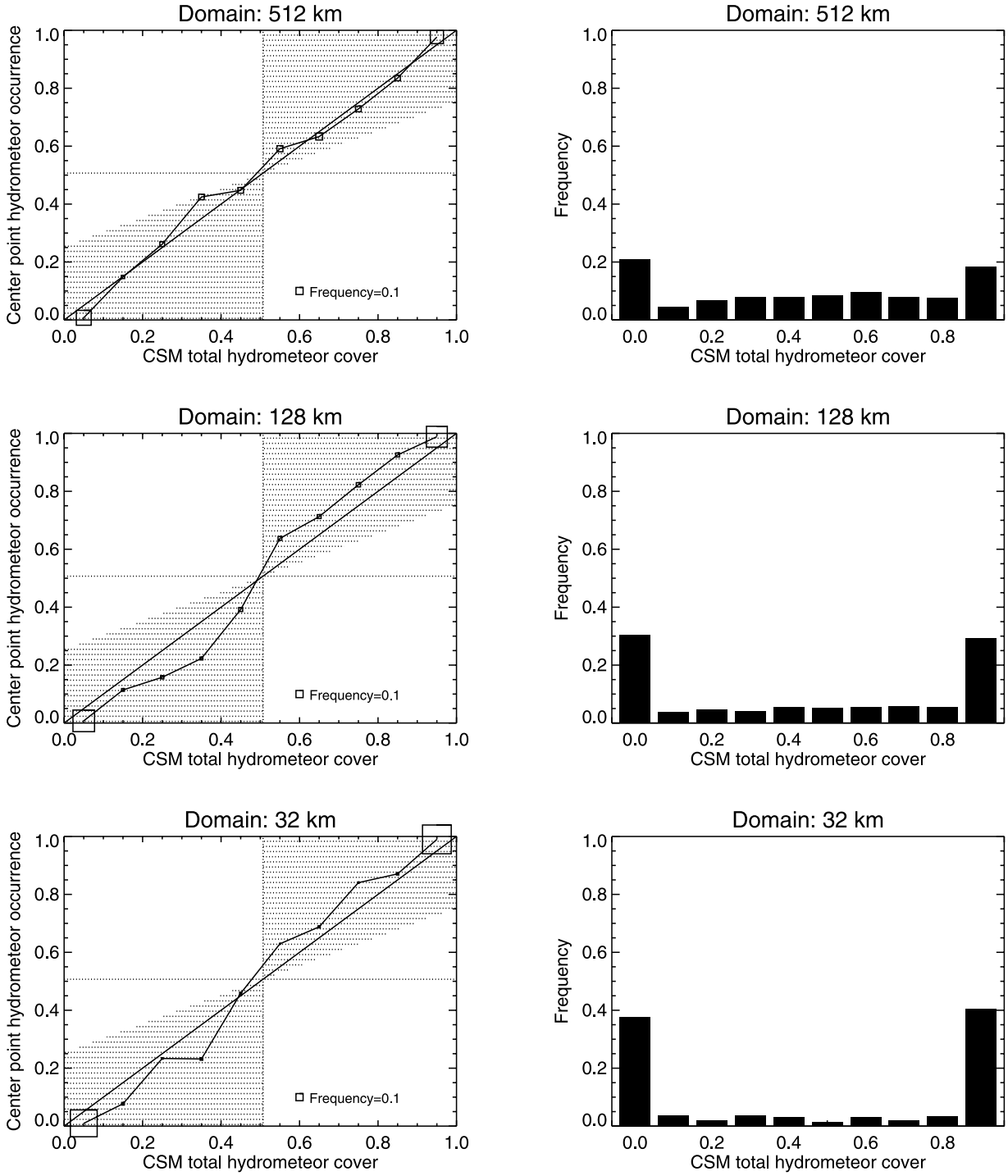

Figure 2. Reliability diagrams (left) and forecast probability frequency distribution (right) for perfect forecasts, constructed by sampling a single CSM column. Results are shown for three domains extracted from the CSM simulation (512, 128, and 32 km from top to bottom) as a proxy for the domain size of the model being evaluated. The stippled area in the reliability diagram indicates points that contribute positively to forecast skill. Sample size is indicated in the reliability diagrams by the size of the mark. Forecast probabilities change dramatically with the size of the domain, but reliability diagrams show good agreement at all sizes, indicating their robustness.

Table 1 shows the score values for our perfect forecasts made for five domain sizes. Both the Brier score and the Brier skill score improve as domain size decreases, because resolution (which indicates the sharpness of the forecasts) increases while reliability remains small. Table 1 also shows the ROC area (ROCA) for each domain size, as is derived from ROC curves (Figure 3 shows these for three domain sizes). ROC area for a perfect forecast is 1.0, so forecast skill as measured by ROCA also increases as the forecast domain gets smaller.

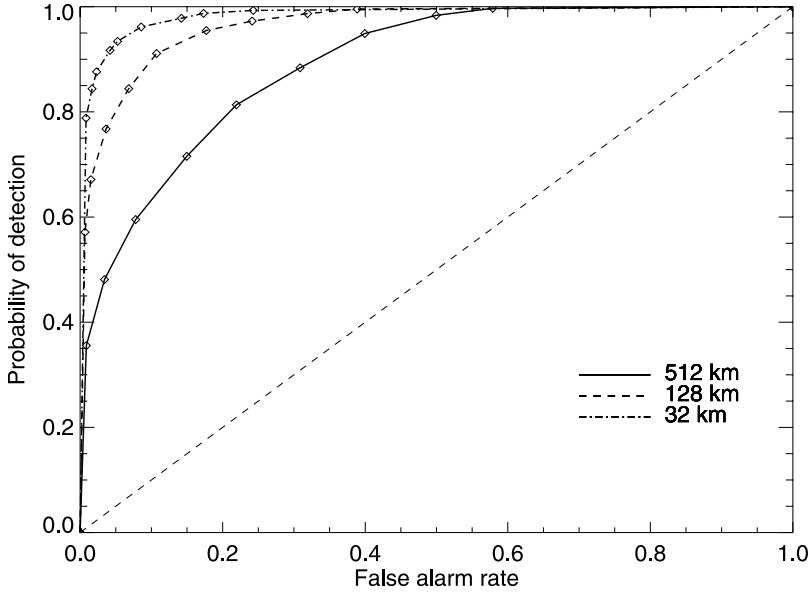

Figure 3. Relative Operating Characteristics for the comparison of the CSM forecasts of THC to the CSM pseudo-observations for model domain sizes of 512 km (solid), 128 km (dashed), and 32 km (dot-dashed).

[24] The forecasts in this example are known to be perfect, in the sense that they are drawn directly from the same population and with the same timing as the observations. Nonetheless, only the reliability score is near its theoretically perfect value, and many components of the Brier score depend on domain size. The explanation lies in the change of the forecast distribution of THC with domain size. The BS measures the distance between the forecast probability and the observation of cloud occurrence, aggregated over all forecasts. However, observed values can only take the values of zero or one, so the perfect Brier score can only be achieved if the forecast values are also restricted to zero and one. Similarly, ROCA measures the ability of the forecast to discriminate events from nonevents, regardless of the forecast value. For THC this distinction is again most clear if the forecasts themselves are restricted to values of zero and one. In our example, however, when the forecast value of THC = 0.5 the pseudo-observations will be about evenly split between one and zero, muddying the distinction between event and nonevent.

[25] If the difficulty with time-averaging observations comes in trying to match time intervals to spatial scales, probabilistic methods are weak when scalar scores derived from reliability diagrams are sensitive to the distribution of forecasts. As traditionally formulated, probabilistic measures show greater skill when the forecasts are unequivocable. Depending on the application, this sensitivity may translate into a dependence on domain size.

4. Examples of Model Evaluation Using Probabilistic Techniques

[26] We now show several examples of how models can be evaluated using probabilistic techniques. We compare predictions of cloud cover made by the CSM and by a weather forecasting model to observations from a collection of active remote sensors, then examine the skill of the CSM in predicting cloud liquid water path as compared to passive microwave radiometer measurements.

[27] Our main verification data set is the Active Remotely-Sensed Cloud Locations (ARSCL) [*Clothiaux et al.*, 2000] data stream, which combines observations made with high-sensitivity millimeter-wave cloud radar, micropulse lidar, and commercial laser ceilometers to determine the presence or absence of hydrometeors in a narrow vertical beam. The vertical grid of the ARSCL data set is very fine; we say that hydrometeors are observed in a model layer if ARSCL reports hydrometeors anywhere within the layer. We ignore the ceilometer-derived estimate of cloud base and consider hydrometeor occurrence throughout the column.

[28] We sample the ARSCL data stream at the temporal resolution of the model being tested (i.e., 5 min). The CSM is forced by imposed surface fluxes and large-scale advective tendencies derived from synoptic observations; because this forcing was determined at a spatial scale of about 500 km [*Zhang et al.*, 2001], we examine forecasts only at approximately commensurate scales (i.e., the full simulation domain).

4.1. Total Hydrometeor Cover

[29] In addition to the CSM forecasts, we examine forecasts of total cloud cover made with the version of the European Centre for Medium-Range Weather Forecasts

Table 1. Probabilistic Scores for the Idealized Comparison of THC for Several CSM Domain Sizes Versus the CSM Pseudo-observations

Domain Size, km	Brier Score	Brier Skill Score	Reliability	Resolution	ROC Area
512	0.133	0.476	0.001	0.118	0.89
256	0.100	0.601	0.004	0.154	0.94
128	0.075	0.698	0.004	0.178	0.96
64	0.057	0.772	0.003	0.196	0.98
32	0.046	0.816	0.002	0.206	0.98

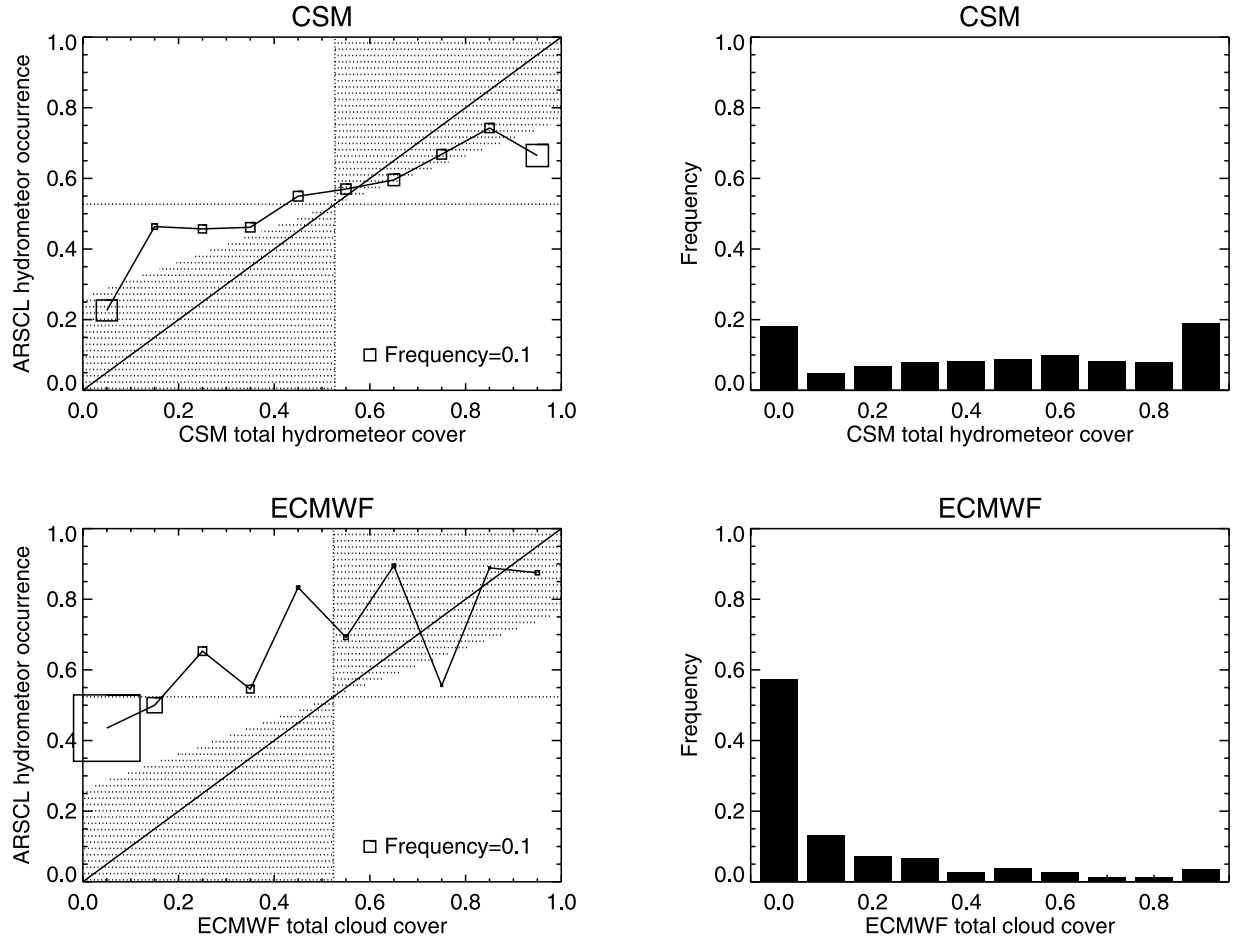

Figure 4. Reliability diagrams (left) and forecast probability frequency distribution (right) comparing forecasts of THC to the occurrence as derived from remote sensing observations. The top panel indicates results from the CSM; the bottom from a single column extracted from the the ECMWF's forecast model.

(ECMWF) operational in summer 1997. The model has an equivalent horizontal resolution of 60 km and 31 levels in the vertical. A prognostic cloud parameterization [*Tiedtke*, 1993] predicts cloud cover at all model levels, which is converted to total cloud cover using a maximum-random overlap assumption [e.g., *Morcrette and Jakob*, 2000]. We compare the ARSCL observations to the single value of total cloud cover at the model grid point nearest the observation site.

[30] As we expect, forecasts by the CSM are substantially better than those from the weather forecasting model. CSM forecasts of THC are largely unbiased and show some skill in the timing of the THC prediction, as the top panel of Figure 4 shows. The scalar measures support this idea: the ROC area (shown in Table 2) is 0.67, as compared with the 0.89 obtained for this domain size in our idealized experiments and the value of 0.5 obtained when forecasts have the correct probability but random timing. The ECMWF forecasts, on the other hand, shows a substantial negative bias (as reflected in the upward displacement of the reliability curve from the diagonal), with far too many forecasts of TCC near 0. The model nonetheless exhibits some skill in distinguishing periods of higher than average from those of lower cloud cover: the ROC area value is 0.63, only slightly less than that of the CSM.

[31] The various scalar forecast skill scores in Table 2 emphasize different aspects of the models' skill in forecasting THC. The Brier skill score is negative for both models, indicating little or no skill, while the ROC area exceeds the no-skill value of 0.5. The difference arises because ROCA is sensitive to timing, while the Brier score weights success according to the difference between the forecast and the event. Thus it is possible for a forecasting system to provide forecasts that are worse than a climatological forecast as measured by the BS but better as measured by ROC. Interpretation of the Brier skill score is further complicated because BSS is computed assuming a perfect forecast delivers a BS of zero, and as we saw in section 3 this assumption does not hold in our application.

4.2. Vertical Cloud Structure

[32] Total hydrometeor cover provides a fairly loose constraint on forecasts. Almost all models compute cloud occurrence as a function of height in the atmosphere, and

Table 2. Probabilistic Scores for the Comparison of CSM THC and ECMWF TCC Forecasts With the ARSCL Derived Occurrence Observations

Model	Brier Score	Brier Skill Score	Reliability	Resolution	ROC Area
CSM	0.255	−0.022	0.032	0.027	0.67
ECMWF	0.353	−0.414	0.123	0.019	0.63

these more detailed predictions may also be tested by active ground-based remote sensors. Figure 5 shows vertical profiles of the mean relative frequency of hydrometeor occurrence as observed (solid) and simulated by the CSM (dashed). We exclude the lowest two model levels, which fall below the lowest range gate of the cloud radar. Below 5 km the CSM forecasts agree fairly well in the mean with the observations, while between 5 and 11 km the model strongly overpredicts hydrometeor occurrence, and cloud tops in the forecasts are lower than observed, leading to an underestimation of hydrometeor occurrence above 11 km. The model's vertical resolution in the upper troposphere is around 800 m, so these errors reflect shifts of cloud top by only one or two model levels.

[33] Reliability diagrams for the model levels indicated by horizontal lines in Figure 5 (shown in Figure 6) can enhance our understanding of the simulation's strengths and weaknesses. For example, although HC forecasts are fairly accurate in the mean at both the 2 and 11 km levels (c.f. Figure 5), the model's behavior is in fact quite different. Forecasts show some skill (BSS = 0.07, ROCA = 0.71) at 2 km, primarily because the simulation correctly identifies the many occurrences of HC < 0.1. At 12 km, however, BSS is less than zero due to errors in the lowest HC classes, and the ROC area value is also smaller (ROCA = 0.63), even though the forecast probability distribution is quite close to the distribution at 2 km.

[34] In contrast, model bias at 9 km is quite large because the simulation overpredicts all classes of HC > 0.2. However, the simulation is somewhat skilled at discriminating between times of greater and lesser cloudiness: much of the reliability curve runs parallel to the diagonal, and the value of ROCA = 0.75.

[35] Figure 7 shows probabilistic verification scores as a function of height. Comparing Figure 7 to Figure 5 shows that the variation of the REL with height largely reflects forecast bias. Forecasts are skillful as measured by the Brier score when REL < RES, as is marginally the case for a few model levels. The value of the ROC area, however, is not particularly sensitive to bias, and so indicates fairly good skill (values ranging from 0.7 and 0.8) between the surface and around 9 km.

[36] From these results we conclude that the simulation of clouds is fairly realistic in terms of timing and low cloud amount at 2 km, although the accuracy of forecasts of cloud cover greater than about 0.3 is weak. This skill in timing exists up to about the 9 km level, where it begins to degrade, while the model also begins to overpredict the amount of cloud when cloud is present at levels higher than 6 km. By 11 km, the simulation predicts fewer cloud-free times than are observed and overpredicts the amount of cloud when cloud is present and does a poor job of timing in both cases. This level of detail is likely to be more helpful as a diagnostic tool for model errors than even a Taylor diagram.

4.3. Liquid Water Path

[37] In principle, any quantity measured at a point whose spatial distribution is predicted by a model can be subject to probabilistic evaluation. We have focused on hydrometeor cover predictions (as a surrogate for cloud cover predictions) which might be considered zero-order measures of cloudiness. More stringent tests are also possible. As we have noted, such comparisons are restricted to models which predict the distribution of the test quantity within the domain. Thus, though general tests of single-column predictions (as are available from global models) are not currently possible, any quantity in the CSM can be evaluated.

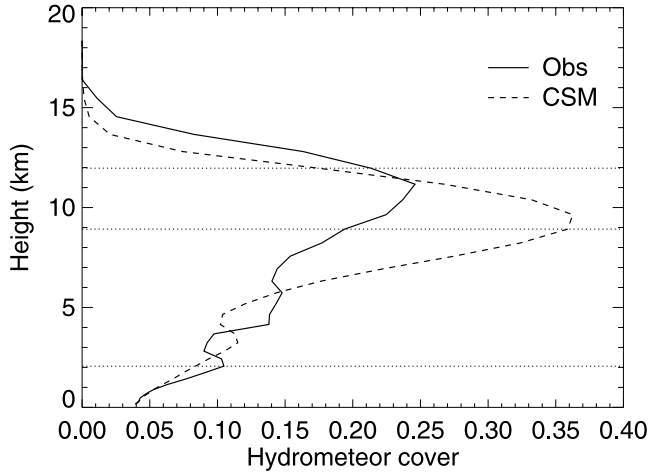

Figure 5. Mean observed (solid) and model (dashed) hydrometeor cover as a function of height for the entire analysis period (19 June to 17 July 1997). The horizontal lines mark the levels for which further analysis is carried out in Figure 6.

[38] As an example we compare model predictions of liquid water path (LWP) against retrievals from measurements made with a single ground-based microwave radiometer (MWR) at the ARM SGP site [*Liljegren*, 1994]. CSM spatial distributions of LWP at each model time are converted into probabilistic forecasts by setting a threshold and counting the number of columns which satisfy an inequality. Figure 8 shows reliability diagrams for the conditions $LWP > 0.05$ kg $\cdot$ m^{-2} and $LWP > 0.1$ kg $\cdot$ m^{-2}. Forecasts of thin liquid water clouds are poor: the model predicts less than half of the observed cases and does so at the wrong times as indicated by the reliability diagram and a ROCA of 0.54. However, when the LWP threshold is doubled, the simulation bias decreases in magnitude and changes sign, and both BSS (0.004) and ROCA (0.64) indicate some model skill.

[39] These results most likely reflect the simulation's inability to generate boundary layer clouds including shallow cumulus. The CSM's subgrid scale cloud parameterization was inactive in these runs, and the model's spatial resolution is 2 km; such "coarse"-resolution CSMs are known to have difficulty simulating small clouds under these circumstances [e.g., *Petch et al.*, 2002].

5. Outstanding Issues

5.1. Utility of Scalar Scores

[40] Though the scalar probabilistic evaluation scores have formal definitions of perfection, these appear to be unachievable, based on section 3, due to sampling errors and the sensitivity to factors like domain size. This makes it hard to evaluate success in a general and quantitative way when the scores are computed for real clouds. It would be useful if we

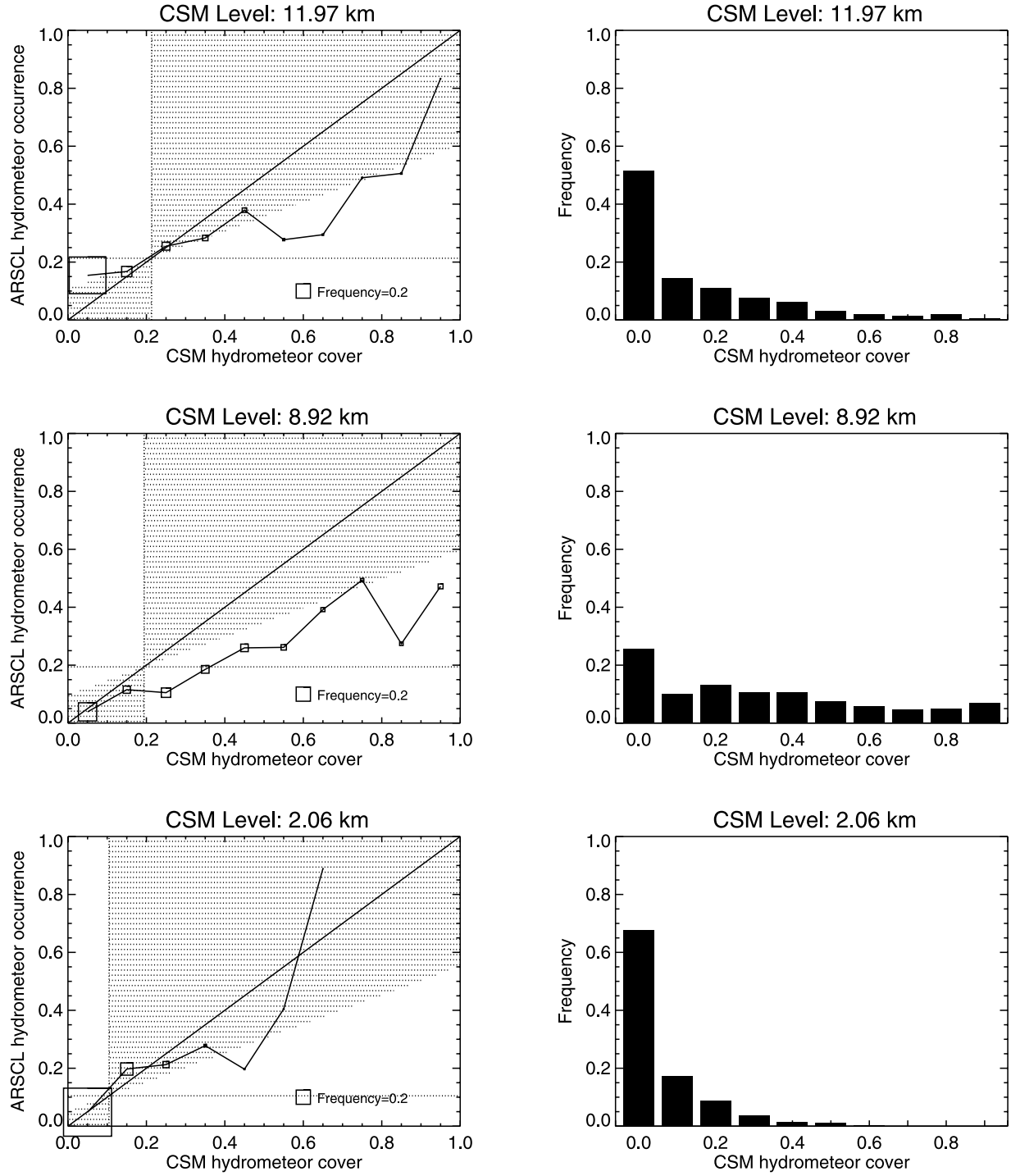

Figure 6. Reliability diagrams (left) and forecast probability frequency distribution (right) for the comparison of the CSM forecasts of hydrometeor cover to the ARSCL-derived occurrence observations at selected model levels. The model levels shown are near 12 km (top), 9 km (middle), and 2 km (bottom).

could provide a context-dependent estimate of the maximally achievable value for each of the score measures.

[41] One possibility is to introduce additional measurements. As an example we show total cloud cover (TCC) retrievals from the GOES-8 satellite [*Minnis et al.*, 1995] averaged over an area comparable to the model domain (5.5 × 5.5 deg) centered on the ARM SGP site. Reliability diagrams treating the satellite retrievals as "forecasts" and ARSCL point observations as "verification" are shown in Figure 9. The satellite sees slightly less cloud than does ARSCL, but overall agreement is quite good (BS = 0.17, BSS = 0.33, and ROCA = 0.85). When perfect forecasts are unavailable one could use these scores as a proxy, since undoubtedly one would be satisfied with a model prediction

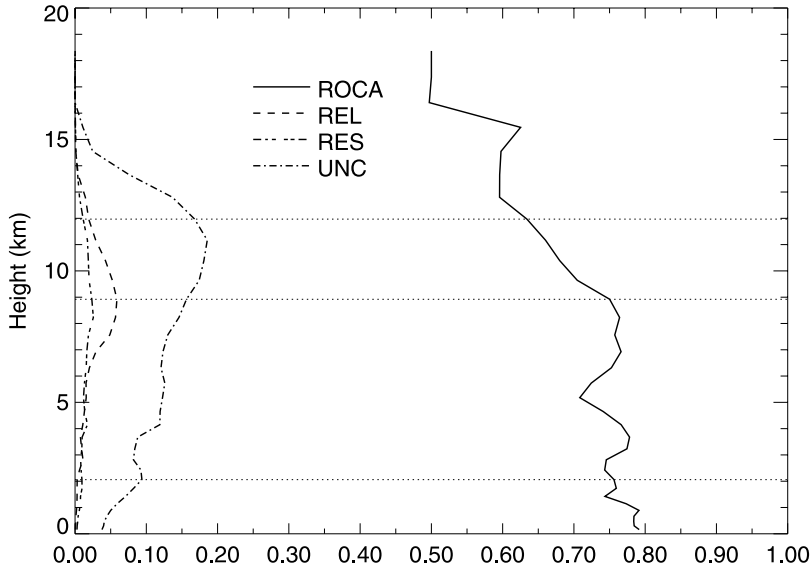

Figure 7. Probabilistic forecast scores for the comparison of CSM-predicted hydrometeor cover to the ARSCL-derived occurrence observations as a function of height. The measures shown are ROCA (solid), REL (dashed), RES (triple-dot-dashed), and UNC (dot-dashed).

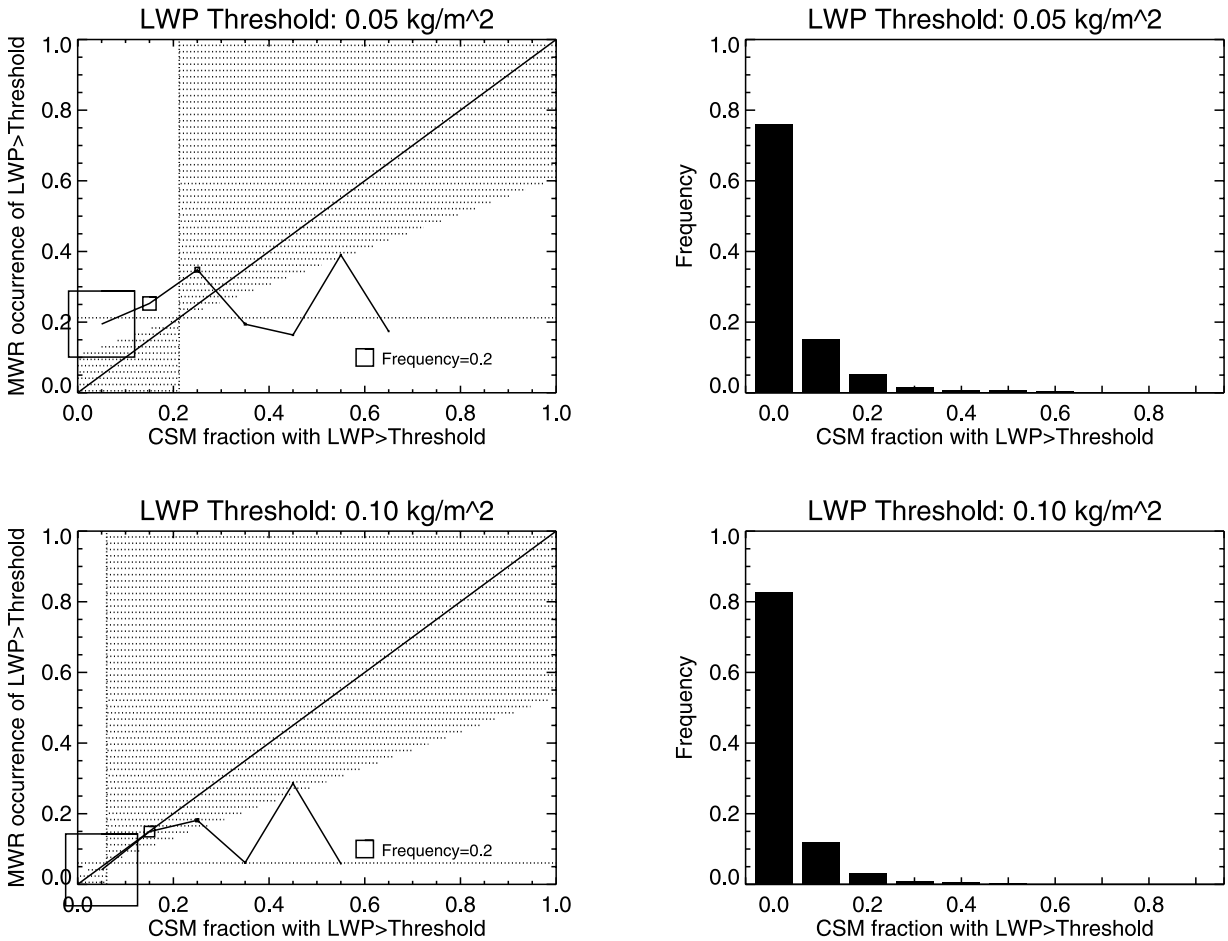

Figure 8. Reliability diagram (left) and forecast probability frequency distribution (right) for the comparison of CSM LWP to MWR derived observations.

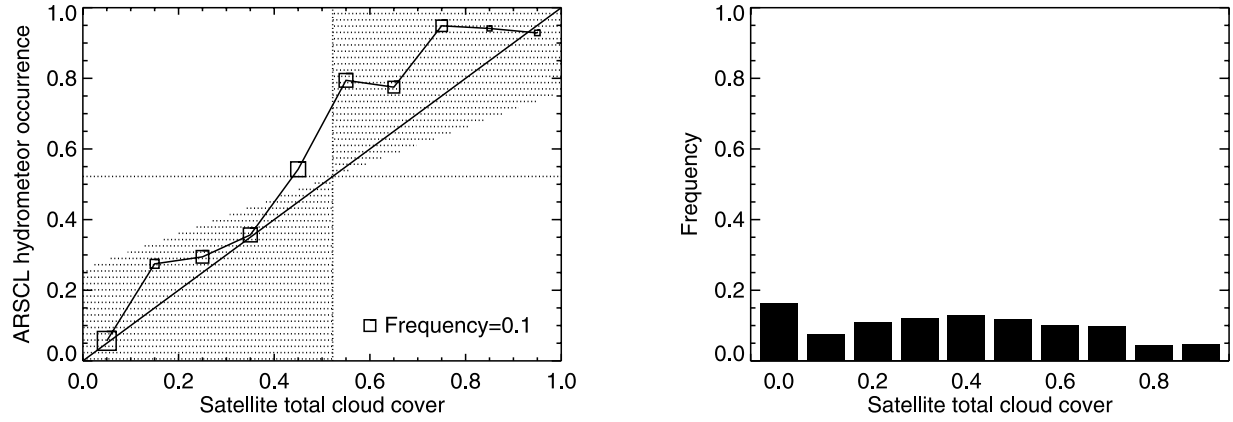

Figure 9. Reliability diagram (left) and forecast probability frequency distribution (right) for the comparison of satellite-derived total cloud cover to the ARSCL-derived occurrence observations. The averaging area is 5.5 × 5.5 deg, similar to the largest CSM domain size.

that achieves this degree of skill. On the other hand, it is not clear from this comparison what skill scores might be expected in perfect layer-by-layer comparisons. (Ironically, forecasts of THC can be tested directly against satellite data, obviating the need for probabilistic evaluation techniques.) However, probabilistic evaluation is useful in a relative sense even in the absence of exact targets, since it provides quantitative guidance when comparing models to each other or when assessing the impact of changes to a model's formulation. Probabilistic evaluation might also be useful in comparing ground-based and satellite measurements.

5.2. Prospects

[42] Though we have concentrated on their use for model evaluation, probabilistic techniques may also be valuable when comparing different measurements of the same quantity. For instance, comparing satellite retrieved cloud and precipitation fields to point observations on the ground is notoriously difficult due to both navigation and representativeness problems, and in data-rich situations these may be avoided using probabilistic evaluation.

[43] Probabilistic evaluation techniques provide useful quantitative information about the performance of cloud system models. Many issues remain, including the derivation of scalar scores whose performance is better understood and which do not depend on the details of the model configuration. The potential of probabilistic techniques is in the evaluation of a wide range of model predictions, and it is in those applications that further strengths and weaknesses of the method will emerge.

Appendix A

A1. Brier Score

[44] The Brier score [e.g, *Brier*, 1950; *Wilks*, 1995] is defined as

$$BS = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2, \tag{A1}$$

where f_i is the forecast probability (i.e., total hydrometeor cover) and o_i indicates the occurrence of the forecast event (i.e., $o_i = 0$ in case of no hydrometeor at the radar site and $o_i = 1$ otherwise). The summation is over all forecasts issued (i.e., over all hourly CSM values). The Brier score of a perfect set of forecasts is 0.

[45] The Brier score may be decomposed into three components [*Wilks*, 1995] as

$$BS = \frac{1}{n}\sum_{i=1}^{C} N_i(f_i - \overline{o_i})^2 - \frac{1}{n}\sum_{i=1}^{C} N_i(\overline{o_i} - \overline{o})^2 + \overline{o}(1 - \overline{o}). \tag{A2}$$

The three terms of this equations are called reliability (REL), resolution (RES), and uncertainty (UNC). The summation in the first two terms of equation (A2) is over the number of forecast classes (C). N_i is the number of forecasts falling into each class, $\overline{o_i}$ is the relative frequency of occurrence of the event in forecast class i, and $\overline{o}$ the overall observed relative frequency of occurrence of the event. Better forecasts have smaller values of BS, which may arise though lower values of reliability or increases in resolution. The uncertainty term in equation (A2) is a function of the observations only.

[46] The decomposed BS is closely related to the reliability diagram shown in Figure 2, in that REL measures the distance of the reliability curve from the diagonal, while RES measures the distance from the horizontal line indicating the sample climate. Note that distances are calculated as averages weighted by the number of forecasts falling into each forecast class, N_i. Hence similar looking reliability curves may lead to different values of REL and RES and hence BS, if the distribution of forecast values changes, as in Figure 2.

[47] The Brier skill score BSS is defined as

$$BSS = \frac{BS - BS_{climate}}{0 - BS_{climate}}, \tag{A3}$$

where $BS_{climate}$ is the Brier score of a constant forecast using the sample mean value (climate), and the zero appears in the denominator as the perfect BS. The perfect value for this score is BSS = 1.

A2. ROC

[48] Another common measure used in the evaluation of probabilistic forecast is the Relative Operating Character-

istics (ROC) curve [e.g., *Mason*, 1982; *Harvey et al.*, 1992]. Here the probability of detection (POD) and false alarm rate (FAR) or probability of false detection for categorical forecasts are compared to each other. A number of these categorical forecasts are derived from the forecast probabilities by cycling through all forecast categories, C, and considering forecasts with a higher probability than the lower boundary of forecast class C_i as forecasts of the event and all those below as forecasts of the nonevent. This leads to $C + 1$ pairs of POD and FAR, which are then plotted as a curve. The curves for the three domain sizes used before are shown in Figure 3. A scalar measure indicating the forecast quality can be derived from those curves, by integrating over the area underneath the curve, often referred to as the ROC area (ROCA). It can be shown [e.g., *Mason*, 1982] that in this measure skillful forecasts are identified by $ROCA > 0.5$, while the perfect forecast has $ROCA = 1$.

[49] **Acknowledgments.** We thank Beth Ebert and Peter May (BMRC) and Jeffrey S. Whitaker (CDC) for fruitful discussions. We greatly appreciate the insights of Steve Krueger and an anonymous referee, which helped improve the manuscript. We are grateful for support from the U. S. Department of Energy under grants DE-FG03-01ER63124 and LANL-23662-001-013T as part of the Atmospheric Radiation Measurement Program.

References

Barnett, T. P., J. Ritchie, J. Gloat, and G. Stokes (1998), On the space-time scales of the surface solar radiation field, *J. Clim.*, *11*, 88–96.

Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, *78*, 1–3.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli (1999), Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System, *Weather Forecasting*, *14*, 168–189.

Clothiaux, E. 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. Meteorol.*, *39*, 645–665.

Harvey, L. O., K. R. Hammond, C. M. Lusk, and E. F. Mross (1992), The application of signal detection theory to weather forecasting behavior, *Mon. Weather Rev.*, *120*, 863–883.

Hogan, R. J., and A. J. Illingworth (2000), Deriving cloud overlap statistics from radar, *Q. J. R. Meteorol. Soc.*, *126*, 2903–2909.

Hogan, R. J., C. Jakob, and A. J. Illingworth (2000), Comparison of ECMWF winter-season cloud fraction with radar derived values, *J. Appl. Meteorol.*, *40*, 513–525.

Krueger, S. K. (1988), Numerical-simulation of tropical cumulus clouds and their interaction with the subcloud layer, *J. Atmos. Sci.*, *45*, 2221–2250.

Liljegren, J. C. (1994), Two-channel microwave radiometer for observations of total column precipitable water vapor and cloud liquid water path, paper presented at Fifth Symposium on Global Change Studies, Am. Meteorol. Soc., Nashville, Tenn.

Luo, Y., S. K. Krueger, G. G. Mace, and K.-M. Xu (2003), Cirrus cloud properties from a cloud-resolving model simulation compared to cloud radar observations, *J. Atmos. Sci.*, *60*, 510–525.

Mace, G. G., T. P. Ackerman, P. Minnis, and D. F. Young (1998a), Cirrus layer microphysical properties derived from surface-based millimeter radar and infrared interferometer data, *J. Geophys. Res.*, *103*, 23,207–23,216.

Mace, G. G., C. Jakob, and K. P. Moran (1998b), Validation of hydrometeor occurrence predicted by the ECMWF model using millimeter wave radar data, *Geophys. Res. Let.*, *25*, 1645–1648.

Mace, G. G., E. E. Clothiaux, and T. P. Ackerman (2001), The composite characteristics of cirrus clouds: Bulk properties revealed by one year of continuous cloud radar data, *J. Clim.*, *14*, 2185–2203.

Mason, I. (1982), A model for assessment of weather forecasts, *Aust. Meteorol. Magn.*, *30*, 291–303.

Minnis, P., W. L. Smith Jr., D. P. Garber, J. K. Ayers, and D. R. Doelling (1995), *Cloud Properties Derived From GOES-7 for Spring 1994 ARM Intensive Observing Period Using Version 1.0.0 of ARM Satellite Data Analysis Program, NASA Ref. Publ. 1366*, NASA Langley Res. Cent., Hampton, Va.

Morcrette, J.-J., and C. Jakob (2000), The response of the ECMWF model to changes in cloud overlap assumption, *Mon. Weather Rev.*, *128*, 1707–1732.

Mullen, S. L., and R. Buizza (2001), Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System, *Mon. Weather Rev.*, *129*, 638–663.

Petch, J. C., A. R. Brown, and M. E. B. Gray (2002), The impact of horizontal resolution on the simulation of convective development over land, *Q. J. R. Meteorol. Soc.*, *128*, 2031–2044.

Platt, C. M. R. (1979), Remote sensing of high clouds: I: Visible and infrared optical properties from lidar and radiometer measurements, *J. Appl. Meteorol.*, *18*, 1130–1143.

Sassen, K., and G. G. Mace (2002), Ground-based remote sensing of cirrus clouds, in *Cirrus*, edited by D. K. Lynch et al., pp. 168–196, Oxford Univ. Press, New York.

Stokes, G. M., and S. E. Schwartz (1994), The Atmospheric Radiation Measurement (ARM) Program: Programmatic background and design of the Cloud and Radiation Test Bed, *Bull. Am. Meteorol. Soc.*, *75*, 1201–1221.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.*, *106*, 7183–7192.

Tiedtke, M. (1993), Representation of clouds in large-scale models, *Mon. Weather Rev.*, *121*, 3040–3061.

Toth, Z., E. Kalnay, S. Tracton, R. Wobus, and J. Irwin (1997), A synoptic evaluation of the NCEP ensemble, *Weather Forecasting*, *12*, 140–153.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*, 467 pp., Academic, San Diego, Calif.

Xu, K.-M., and S. K. Krueger (1991), Evaluation of cloudiness parameterizations using a cumulus ensemble model, *Mon. Weather Rev.*, *119*, 342–367.

Xu, K.-M., et al. (2002), An intercomparison of cloud-resolving models with the ARM summer 1997 IOP data, *Q. J. R. Meteorol. Soc.*, *128*, 593–624.

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. Weather Rev.*, *129*, 295–311.

C. Hannay and R. Pincus, NOAA-CIRES Climate Diagnostic Center, 325 Broadway, R/CDC1, Boulder, CO 80305, USA. (cecile.hannay@noaa.gov; robert.pincus@colorado.edu)

C. Jakob, BMRC, GPO Box 1289K, 150 Lonsdale Street, Melbourne 3001, Australia. (c.jakob@bom.gov.au)

K.-M. Xu, NASA Langley Research Center, Hampton, VA 23681-0001, USA. (k.m.xu@larc.nasa.gov)