Verification of Cloud Forecasts over the Eastern Pacific Using Passive Satellite Retrievals

JASON E. NACHAMKIN, JEROME SCHMIDT, AND CRISTIAN MITRESCU

Naval Research Laboratory, Monterey, California

(Manuscript received 4 November 2008, in final form 3 March 2009)

ABSTRACT

Operational cloud forecasts generated by the Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS)[1] were verified over the eastern Pacific Ocean. The study focused on the accuracy of cloud forecasts associated with extratropical cyclone and convective activity during the late winter and spring of 2007. The condensed total water (liquid and solid) path was used as a proxy for cloud cover to compare the forecasts with retrievals from the Geostationary Operational Environmental Satellites (GOES). Analyses of the GOES retrievals indicate that deep cloud systems were generally well represented during daylight hours. Thus, the bulk of the verification focused on the general aspects of quality and timing of these deep systems. Multiple statistics were collected, ranging from simple correlations and histograms to more sophisticated fuzzy and composite statistics. The results show that synoptic-scale systems were generally well predicted to at least two days, with the primary error being an overestimation of deep cloud occurrence. Smaller subsynoptic-scale systems were subject to spatial and timing biases in that a number of the forecasts were lagged by 3–6 h. Despite the bias, 60%–70% of the forecasts of the mesoscale phenomena displayed useful skill.

1. Introduction

Many current technologies are increasingly sensitive to the effects of cloud cover. Cirrus often interferes with military applications that depend on a clear line of sight (Norquist 1999). Low ceilings impact both civilian and military aviation (Carter and Glahn 1976), and in this era of tight schedules even minor disruptions have significant consequences. Depending on the application, depictions of the cloud state are desired at many scales, and accurate forecasts at long lead times are highly desirable for planning purposes. However, cloud prediction has long been problematic.

The large range of scales (from microscopic to synoptic) impacting cloud development is a challenge to even the highest-resolution mesoscale models. However, bulk cloud parameterizations and microphysical schemes have recently shown the ability to produce useful cloud forecasts. Chaboureau and Pinty (2006) compared brightness temperature forecasts from the nonhydrostatic Méso-NH regional model with Meteosat Second Generation observations over Brazil. With a horizontal grid spacing of 30 km, the model successfully reproduced diurnal variations in convective cloudiness as well as lower-frequency variations produced by changing weather regimes. Other comparisons by Chaboureau et al. (2002) show good qualitative agreement between simulated cloud systems and retrieved ice and liquid water path observations. These case studies were conducted for both midlatitude and subtropical cloud systems using horizontal grid resolutions from 12 to 75 km. Additional studies by Söhne et al. (2006), Li et al. (2005), and Chevallier and Kelly (2002) have noted good quantitative agreement between cloud forecasts and the accompanying satellite observations.

As model resolution increases, cloud forecasts become increasingly realistic. Like precipitation forecasts, this realism adds value but does not necessarily lead to higher scores (Ebert and McBride 2000). Bieringer et al. (2006) compared 3-km fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5) ceiling forecasts with those from the parent 9- and 27-km domains, as well as forecasts from a separate

[1] COAMPS is a registered trademark of the Naval Research Laboratory.

Corresponding author address: Jason E. Nachamkin, Naval Research Laboratory, 7 Grace Hopper Ave., Monterey, CA 93943. E-mail: jason.nachamkin@nrlmry.navy.mil

DOI: 10.1175/2009MWR2853.1

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAR 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Verification of Cloud Forecasts over the Eastern Pacific Using Passive Satellite Retrievals	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory,7 Grace Hopper Ave,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**17**	

20-km Rapid Update Cycle (RUC) domain. Categorical statistics based on pointwise comparisons with automated observations showed comparable performance at all resolutions. However, spatial variability within the 3-km nest was a potential indicator of forecast error. Errors were significantly reduced by averaging the 3-km forecasts over a 54 km × 54 km region. Söhne et al. (2006) noted that high-resolution forecasts of North African convective clouds performed significantly better in a spatially averaged sense than their lower-resolution counterparts. Essentially these studies indicate that the small-scale variance is better simulated at high resolution. Even if the forecast weather features are not identically located in their observed positions, these gains can be very useful if they lead to improvements in probability forecasts of high impact weather.

In this paper, cloud total liquid + ice condensed water path (LWP) forecasts from the Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS; Hodur 1997) are compared with Geostationary Operational Environmental Satellite (GOES) retrievals. The main goal of the verification is to determine the viability of cloud forecasts over the eastern Pacific, where vast areas of sparse data present a uniquely challenging forecast environment. The fact that COAMPS produces useful forecasts in this region attests to the many advances in data assimilation and physics representation. Abundant cloudiness in this region poses a unique challenge to verification as cloud amount forecasts score deceptively well because of the high probability of a correct random forecast. Upper-tropospheric cloudiness is particularly interesting for military aviation, but it can be quite difficult to verify because of brightness temperature ambiguities. Thus, the focus will be on the deep, cloud-producing systems typically associated with synoptic fronts and subtropical convection. Since these systems are responsible for a large percentage of upper-tropospheric clouds, tracking the forecast accuracy of these systems is helpful. Verification statistics of many types will be used to track model performance, starting from some basic statistics and progressing to more sophisticated methods. New verification techniques have recently shown promise in accounting for the small-scale spatial errors. Fuzzy methods (Roberts and Lean 2008) and the composite method (Nachamkin 2004) will be employed here to evaluate forecast performance with respect to scale as well as event occurrence.

2. Data

a. Atmospheric forecast model

The operational nonhydrostatic COAMPS forecasts for the eastern Pacific conducted at the Fleet Numerical Meteorology and Oceanography Center (FNMOC) were selected for this study. The domain setup consists of two one-way nested grids with horizontal spacings of 81 and 27 km (Fig. 1), and 30 vertical sigma levels with the lowest at 10 m AGL and the highest near 35 km. Forecasts were initialized daily at 0000, 0600, 1200, and 1800 UTC, using the Naval Research Laboratory's Atmospheric Variational Data Assimilation System (NAVDAS; Daley and Barker 2001). The previous 6-h forecast acted as a first guess. Boundary conditions were supplied from the Navy Operational Global Atmospheric Prediction System (NOGAPS; Hogan et al. 2002) at 3-h intervals using a Davies (1976) scheme. Subgrid-scale convection was parameterized using the Kain–Fritsch scheme (Kain and Fritsch 1993), while the explicit microphysics were parameterized using a modified Rutledge and Hobbs (1983, 1984) scheme described by Schmidt (2001). These modifications include predictive equations for graupel and drizzle, as well as the Meyers et al. (1992) ice nucleation, homogeneous freezing, ice multiplication processes, and nonzero pristine ice fall speeds. For this study, the 27-km LWP forecasts valid from 0 to 48 h were collected for validation from 1 February through 31 May 2007. Notably, the LWP retrievals (described below) were restricted to daylight hours. Thus, simulations initialized at 0000 UTC were used to verify forecasts with lead times of 0–3, 18–27, and 42–48 h, while the runs initialized at 1200 UTC were used to verify the 6–15- and 30–39-h forecasts.[2] Some fluctuations occurred in the statistical time series, but the general trends remained consistent for both initialization times.

b. Satellite observations

GOES LWP observations were retrieved using optimal estimation techniques described by Mitrescu et al. (2006). The method involves prescribing forward radiative and microphysical models to retrieve cloud properties from passive multispectral observations. During daylight hours, GOES channels 1, 2, and 4 are used to retrieve cloud-top temperature, cloud-top effective radius, and cloud optical depth. LWP is in turn derived from these variables. In this study, only daytime retrievals were used as they have proven to be more reliable. The radiative forward model follows Nakajima and King (1990), Miller et al. (2000), and Heidinger (2003). The atmosphere, cloud, and land/ocean surface are represented as three separate layers. Cloud optical depth calculations rely on reflectance measurements, and a plane-parallel, homogeneous atmosphere is assumed.

[2] The 0600 and 1800 UTC interim runs were not verified.

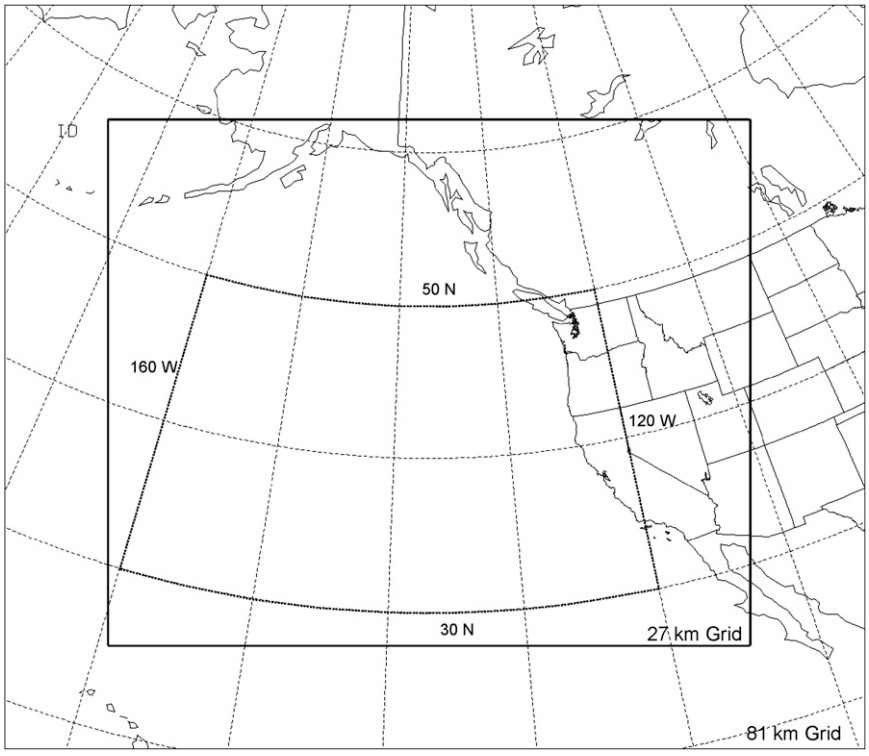

FIG. 1. The COAMPS 81- and 27-km eastern Pacific operational domains are indicated by the thin and thick rectangles, respectively. The GOES analysis subdomain extends from 30° to 50°N and 120° to 160°W as indicated by the dark dotted labeled lines.

Surface albedo is prescribed from the International Geosphere Biosphere Program (IGBP) dataset. Cloud particles are assumed to follow a Gamma size distribution with a constant width throughout the cloud (Mitrescu et al. 2005). Ice particles are assumed to have a spherical cross section with a fixed effective density. It should be noted that a Marshall–Palmer size distribution is assumed for the liquid species in the COAMPS microphysics scheme. Since mass mixing ratio is predicted, the assumed distribution primarily affects the partitioning of the particles and not the total mass. Thus, the difference in the distributions is not expected to produce major differences in the integrated LWP calculations.

The native footprint of the satellite retrieval was approximately 4 km. For comparison with the COAMPS output, these data were linearly averaged onto the 27-km forecast grid. Representativeness errors were alleviated as the satellite data themselves represent averages over an area. To minimize errors associated with land albedo, low zenith angle, and model boundary effects, a data window was defined from 160° to 120°W longitude and 30° to 50°N latitude (Fig. 1). All verification was conducted within the area defined by this window.

Nakajima and King (1990) and Miller et al. (2000) estimate optical depth and effective radius errors to be as high as 50%, thus LWP may also contain errors up to a factor of 2. For this study, the LWP output was carefully studied visually for some indication of the nature of these errors. At low solar zenith angles LWP was generally underestimated, and occasionally large discrepancies occurred when shadows developed in regions of textured clouds. Given the observational ambiguities it was decided to present the statistics in a bias-corrected format. Though quantitative bias and RMS errors of the forecasts will remain unknown, useful statistics can still be derived pertaining to the forecasts of the general positions of large cloud systems.

The bias was removed by comparing LWP measurements with the forecasts over bimonthly intervals in a series of scatter diagrams. Direct point-to-point comparisons were of little use because of the high variability and the presence of phase errors. Instead, the phase errors were statistically removed by sorting the forecasts and observations by magnitude at each realization and deriving regression relationships from the sorted distributions (Fig. 2). The resulting average LWP distributions exhibited considerable variability, especially

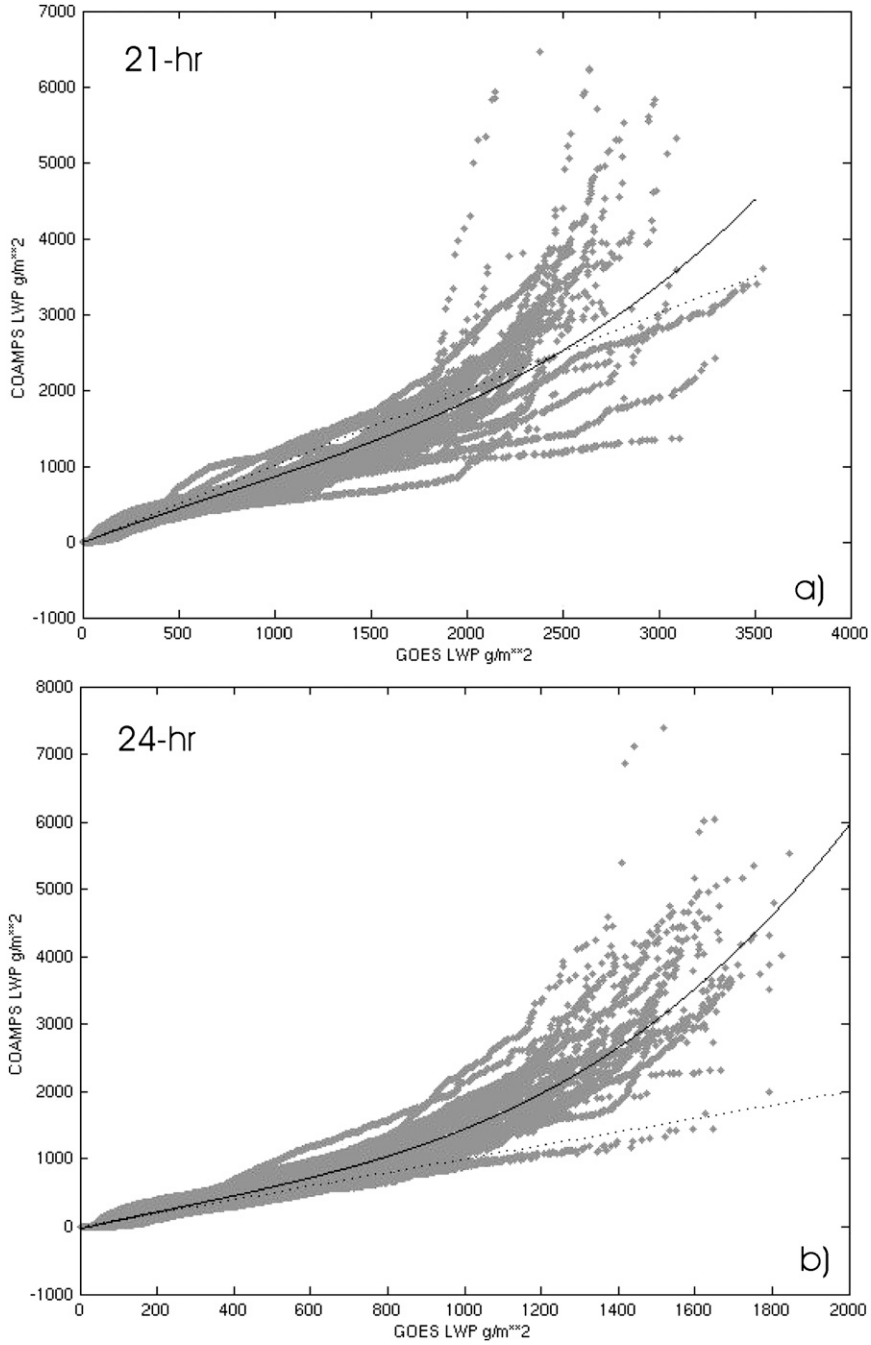

FIG. 2. Scatter diagrams of the magnitude-sorted LWP forecasts and observations for February and March 2007 for the (a) 21- and (b) 24-h forecasts. Third-order regression calculations are represented by solid curves while $x = y$ is represented by the light dotted line.

at large values. During the middle portions of the day (Fig. 2a), the scatter was relatively uniform about the $x = y$ line. However, by late afternoon (Fig. 2b) the zenith angle bias manifested itself as a quadratic error that was generally correctable using regression techniques. To minimize the effects of the bias, the majority of the statistics discussed herein will be limited to periods when most of the grid experienced relatively high zenith angle, namely, 1800–2100 UTC. As Fig. 2a indicates, the systematic bias was relatively small during these times.

Given the observational data caveats, it is not surprising that Chevallier and Kelly (2002) and others note

that brightness temperature comparisons are a more consistent means of comparing the model with the observations as retrieval errors are not incorporated. However, the derived products from both the model and the satellite are regularly used for day-to-day operations. Most user applications rely on cloud height and depth information that brightness temperature alone does not provide. The compromise entailed by presenting the results in a user-oriented space was worth the additional error, especially if both products are widely used. The results should thus be viewed as a general consistency check between the predictions and observations of large-scale deep cloud areas.

3. Results

a. Point-to-point comparisons

The 21-h forecast from 0000 UTC 1 February (Fig. 3) graphically demonstrates the difficulties associated with point-to-point verification of high-gradient fields. The most prominent features include a north–south frontal band through the central portions of the analysis grid (near 140°W), along with a region of scattered post-frontal clouds west of the front (near 153°W). While the general location of the cloud features was correctly predicted, the scatter diagram inset in Fig. 3b shows little correlation between the two LWP fields. A closer look reveals that the northern portions of the frontal band lag to the west of the observations, while the southern portions are too intense and narrow. The broken nature of the cloudiness in the central and southern portions of the band is also not well simulated. Cloud coverage to the west of the front is underrepresented, and the leading edge of the cloud field is 1°–2°W of its observed position. Although these errors are relatively small in scale, high LWP gradients significantly reduce the correlations. The traditional verification scores presented in this section measure the direct field correspondence, providing little allowance for small-scale errors in high-variance fields. Thus, the scores should be interpreted with caution.

The 4-month average correlations (Fig. 4) reflect some of the features mentioned in the example above. While many of the forecasts appeared to be viable on visual inspection, the correlations were generally low. Correlation coefficients start below 0.5 at the analysis time[3] and slowly decrease through the forecast. High variability at small scales essentially acts as noise (Fig. 3), reducing the overall strength of the correlations. The slow degradation of the correlations with time primarily reflects medium- to large-scale errors. Daily correlations at the synoptic scale (discussed later) were on the order of 0.8, and these large-scale correlations also dropped very slowly with lead time. The choppy nature of the descent is likely due to the variations in LWP magnitude with sun angle. GOES persistence correlations were calculated by comparing the satellite observations valid at the analysis time with the observations over the ensuing 48-h period. The persistence correlations drop rapidly with time, reflecting the complex structure and rapid evolution of the cloud field. GOES LWP persistence drops below COAMPS after 3 h, while the infrared (IR) cloud-top persistence forecast decays more slowly. The IR field is probably a better overall representation of the satellite persistence forecast decay rate as this field is not sensitive to solar zenith angle. Note also that LWP persistence was only calculated against forecasts initialized at 0000 UTC because darkness prevented the collection of LWP observations between 0600 and 1500 UTC.

Perhaps the most interesting aspect of Fig. 4 is reflected in the behavior of the lagged forecast correlations. These correlations were generated by comparing current observations with forecasts valid 3 h later than (fm3) and 3 h earlier than (fp3) the observation time. The fp3 and on time (f00) forecast correlations were about the same magnitude, while the fm3 correlations were considerably lower. Tests using a t distribution indicate that the differences between the f00 and fm3 correlations are significant at the 99% confidence level for all lead times prior to 42 h. Tests with 6-h lags (not shown) resulted in correlations at or below the fm3 levels. These results indicate that the model tends to be too slow with the progression of weather systems through the region. The near equality of the fp3 and f00 correlations suggests a peak somewhere between them, reflecting an average lag error in the 0–3-h range. The phase error is not universal, though, as indicated by the example in Fig. 3. Note the northern portions of the frontal band (located near 140°W) are lagged in the forecast, north of about 40°N, while farther south the frontal position is better simulated. Analysts at FNMOC have relayed similar anecdotal accounts of 3–6-h time lags in the precipitation forecasts (J. Lerner, FNMOC models and data section, 2007, personal communication).

Although much of the bias was removed from the data, average LWP histograms (Fig. 5) still reveal some interesting trends. The majority of the observed LWP values fell into the lowest threshold categories, with values below 100 g m^{-2} accounting for about 60% of the distribution. Most of the LWP values below 200 g m^{-2} derive from cirrus or boundary layer stratus clouds that

[3] LWP values were not assimilated.

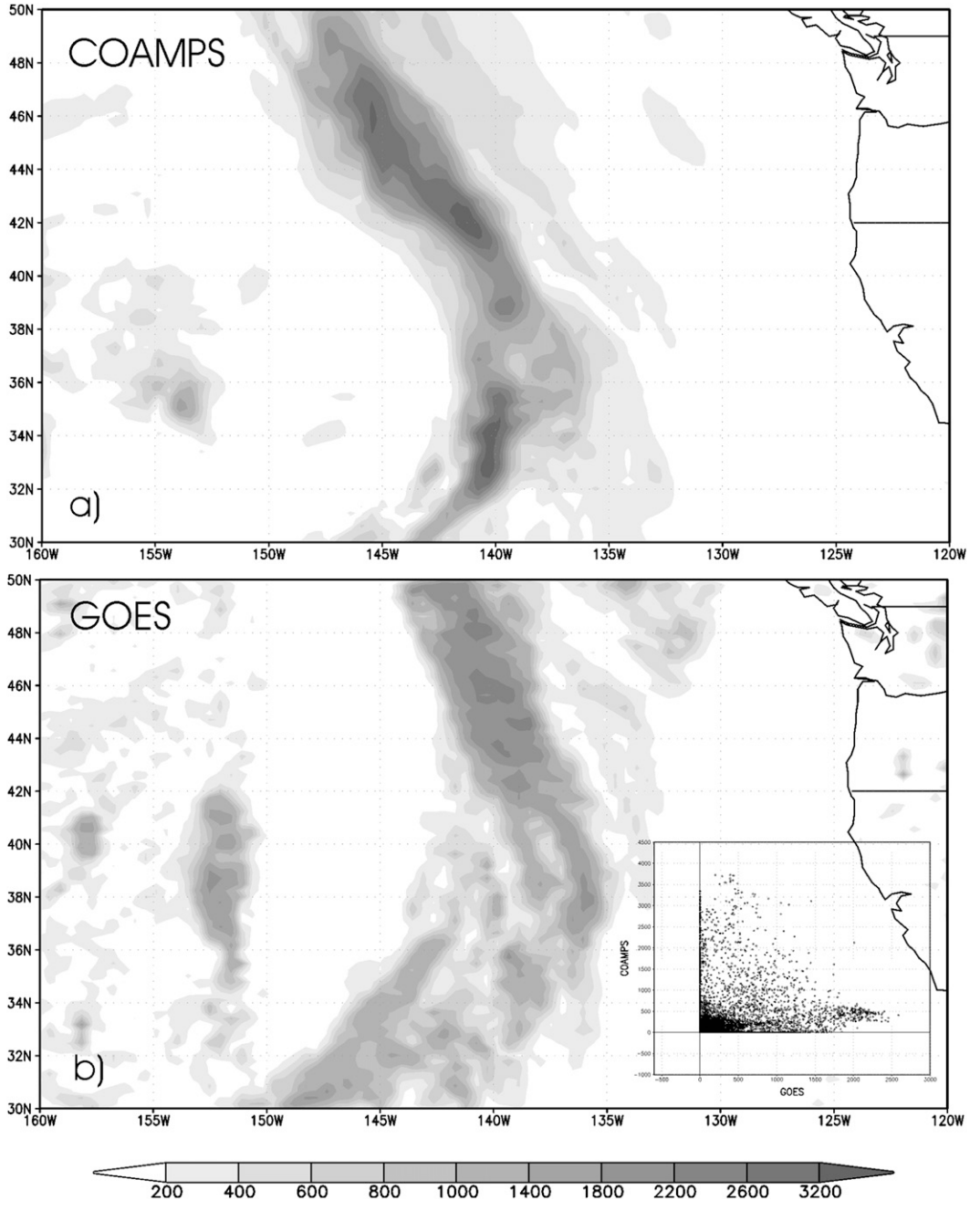

FIG. 3. (a) The COAMPS 21-h LWP forecast and (b) the verifying GOES analysis. The shading is in grams per meter squared as represented by the scale. The scatter diagram inset in (b) represents a point-to-point comparison between the predicted and observed fields.

often cover vast regions of the Pacific. Deeper precipitating clouds are generally associated with values above 500 g m^{-2}. In the model, precipitating convective clouds exhibited values over 3000 g m^{-2}, though the corresponding observed values rarely exceeded 2000 g m^{-2}. These clouds are highly three-dimensional and thus violate the plane-parallel and vertical homogeneity assumptions in the retrievals. Multidirectional scattering of outgoing radiation likely results in underestimates of LWP. A simple hydrostatic calculation reveals that even a modest average liquid mixing ratio of 0.5 g kg^{-1} over a 700-hPa depth from 1000 to 300 hPa results in an LWP of over 3500 g m^{-2}, which is well above any of the observed values.

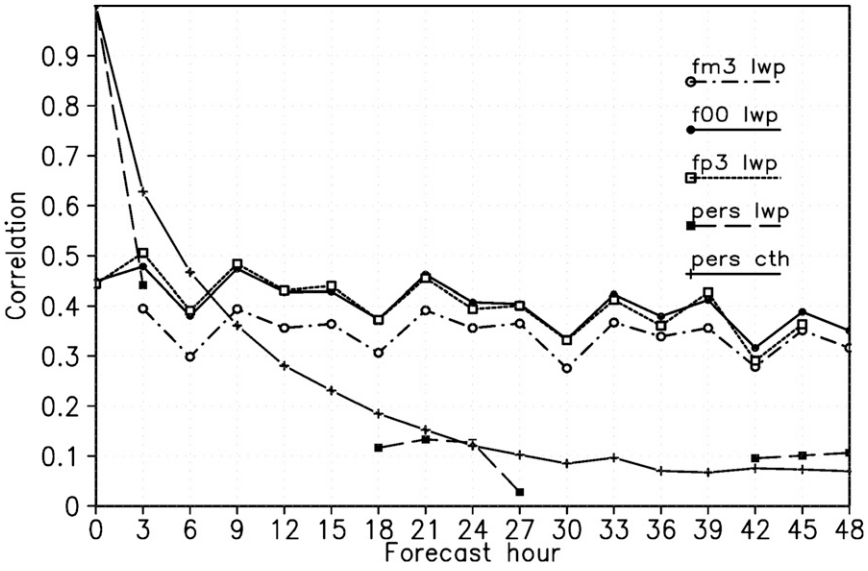

FIG. 4. Correlation coefficients between the predicted and observed variables with respect to forecast hour. GOES infrared cloud-top height (pers cth) and LWP (pers lwp) persistence forecasts are represented by the thick solid and dashed lines. The COAMPS forecasts valid at the observation time (f00), 3 h earlier than the observation time (fm3), and 3 h later than the observation time (fp3) are indicated by the light dotted, dashed–dotted, and short dashed lines, respectively.

The observed and predicted distributions deviate significantly from one another at values below 100 g m^{-2}. The regression curves (Fig. 2) did not remove these biases as the absolute differences in LWP were small and highly compressed at the low end of the scale. Corrections were primarily weighted by the very large biases at the high end of the LWP distribution. The low-end biases likely result from deficiencies in subgrid-scale cloud prediction. To demonstrate this, the observed values were reinterpolated to the model grid such that all grid areas with less than 50% cloud coverage were automatically set to a value of LWP = 0 g m^{-2} (clear). All grid points covering observed areas of scattered or partly cloudy conditions were thus eliminated. The resulting distribution, represented by the bold horizontal lines in Fig. 5, is much closer to the forecasts, especially for clear skies. A notable deviation from this trend was the tendency for the model to produce too much thin cirrus. Much of this overproduction occurred in areas that were otherwise covered by low stratus, thus leaving LWP statistics relatively unaffected.

The Hanssen–Kuipers discriminant (HK) and the equitable threat score (ETS) are presented in Fig. 6 for completeness, as these values are often presented in the verification literature. The HK discriminant ranges from −1 to 1, with 1 being a perfect score, 0 indicating no skill, and −1 indicating a perfect negative correlation. Ebert et al. (2004) note that HK is the false-alarm rate subtracted from the probability of detection. It reflects the ability of the forecast to discern between cloudy and clear areas. The ETS measures the number of correct forecasts in proportion to the total number of forecasts and observations of a given event, adjusted for the probability of a correct random forecast. Since cloud areas were often extensive, the random adjustment factor was sometimes as large as the ETS itself for lower LWP thresholds.

For consistency, the scores were calculated only at the time of the maximum average zenith angle (2100 UTC) using forecasts initialized at both 0000 and 1200 UTC. Since the scores are threshold based, a value of 500 g m^{-2} was chosen to represent the deeper, precipitating systems. These systems reflect the majority of the well-defined cloud entities that tracked across the region. Lower thresholds produced artificially high scores as both cirrus and stratus clouds could share the same LWP value. Given the poor point-to-point correlations the scores were generally low. However, both scores indicated some skill through much of the forecast, with little reduction with increasing lead time.

The sensitivity of the pointwise scores to small-scale errors is mitigated by considering averages over larger areas. For example, the general ability for the model to discern between synoptically disturbed and quiescent conditions is an important indication of overall performance, especially in data-sparse regions. Figure 7 depicts

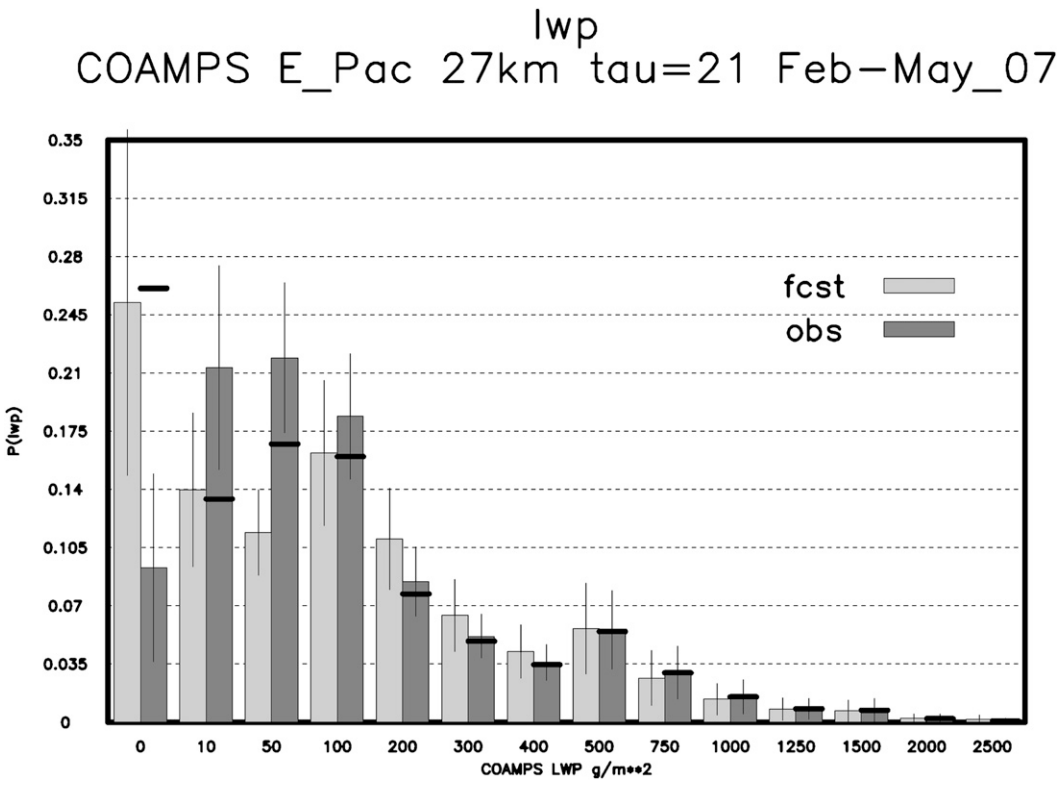

FIG. 5. The bias-corrected 21-h LWP forecasts (light bars) and observations (dark bars). Values along the vertical axis represent the average coverage over the analysis grid during the 4-month period. Horizontal axis values represent the lower bound of each bin. Standard deviations are depicted by the vertical solid lines. Thick horizontal lines represent the observations when all grid squares containing less than 50% cloud coverage are considered as clear.

the daily total coverage of LWP above 500 g m^{-2} as a fraction of the entire analysis region. The model was generally able to predict periods of disturbed weather, though coverage errors on the order of 5% were not uncommon on a given day. The correlation coefficients at the 9-, 21-, 33-, and 48-h lead times were 0.82, 0.87, 0.80, and 0.80, respectively. These values were considerably higher than the point-to-point correlations in Fig. 4, reflecting the improved performance at large scales. The rate of decrease with time was also slower than that in Fig. 4, again due to the large scale. The reduction in correlation with lead time was primarily due to increasing

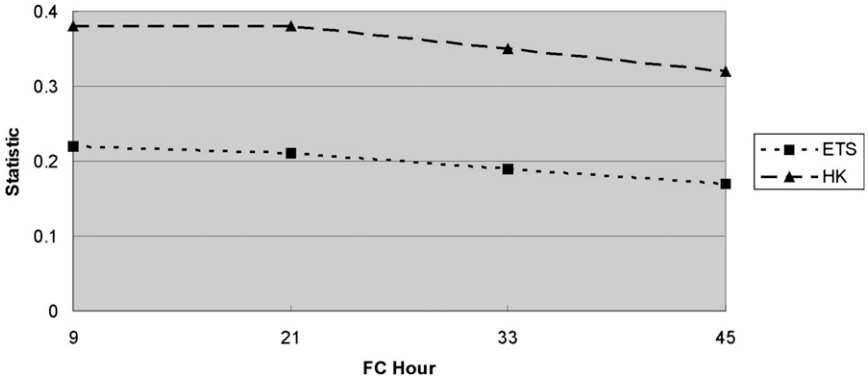

FIG. 6. The 4-month-average HK discriminant (solid line) and ETS (dashed line) as a function of forecast lead time.

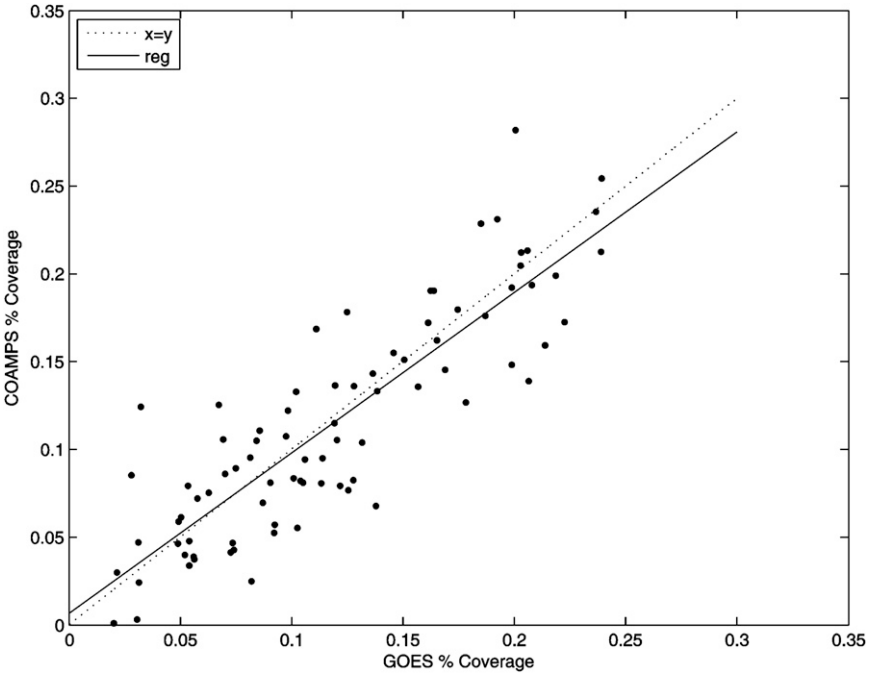

FIG. 7. Bias-corrected 21-h forecasts and observations of LWP values ≥500 g m^{-2} as a function of total coverage over the satellite analysis grid. The regression relation is represented by the solid line while $x = y$ is represented by the dotted line.

random scatter, and the slope of the regression lines remained relatively close to unity.

b. Multiscale statistics

A number of new verification methods have recently been developed to sample forecast accuracy over a range of length scales. Verifying at multiple scales provides a means to gauge the spatial forecast error. If the errors are small in scale the statistics will rapidly improve with increasing sample area. Ebert (2008) reviews a number of these methods, some examples include upscaling (Zepeda-Arce et al. 2000), wavelet decomposition (Casati et al. 2004), and the fuzzy neighborhood method (Roberts and Lean 2008). Since clouds often occur as fractional or amorphous elements that are difficult to characterize, the fuzzy neighborhood method was particularly appealing as accuracy is expressed in terms of event frequency (Roberts and Lean 2008; Ebert 2008). The method was also very simple to code and it ran very efficiently within the existing software.

The fuzzy neighborhood method works by calculating verification metrics over a series of squares, or neighborhoods, of increasing size centered at each grid point. Larger neighborhoods allow for displaced forecast values to be counted as correct. We follow Roberts and Lean (2008) in using the fractions skill score (FSS) as a skill metric. The FSS is defined as

$$\mathrm{FSS}_{(n)} = 1 - \frac{\mathrm{MSE}_{(n)}}{\mathrm{MSE}_{(n)_{\mathrm{ref}}}}, \tag{1}$$

where $\mathrm{MSE}_{(n)}$ refers to the mean-square error over a neighborhood of length n as given by

$$\mathrm{MSE}_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{(n)_{i,j}} - M_{(n)_{i,j}}]^2. \tag{2}$$

The MSE is calculated from the binary field I created by imposing a threshold on the data to be verified. Here, all LWP points exceeding 500 g m^{-2} are assigned a value of $I = 1$, with $I = 0$ assigned to all other points. The MSE is summed for each forecast over the entire analysis grid consisting of N_x by N_y points, where $N_x \times N_y = 9971$. The quantities $O_{(n)_{i,j}}$ and $M_{(n)_{i,j}}$ represent the sum over each neighborhood of the observed and predicted components of the binary field centered at each i, j grid point:

$$O_{(n)}(i,j) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} I_o \left[i + k - 1 - \frac{(n-1)}{2}, j + l - 1 - \frac{(n-1)}{2} \right] \quad \text{and} \tag{3}$$

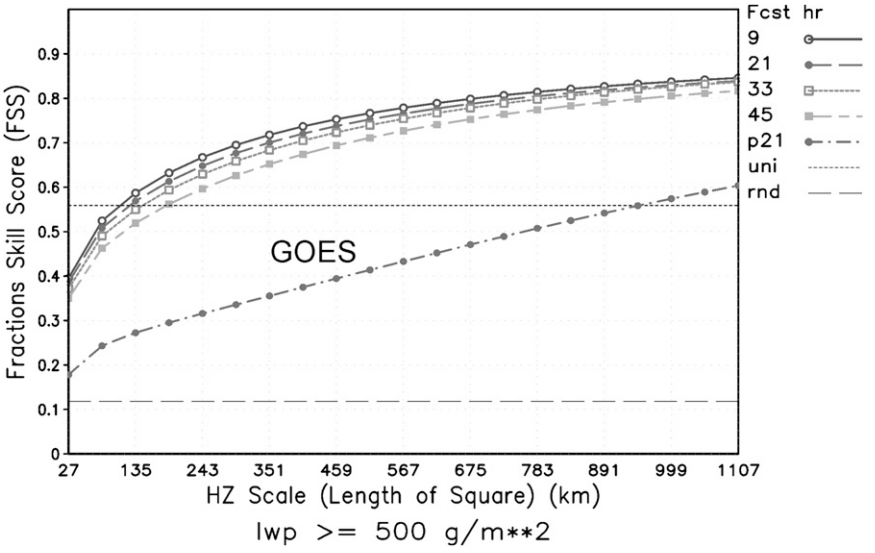

FIG. 8. The variation of the FSS with neighborhood length for the 9-, 21-, 33-, and 45-h forecast lead times. The FSS associated with the 21-h GOES LWP persistence forecast (p21) is indicated by the dashed–dotted line. The uniform and random FSS are displayed as horizontal dotted and dashed horizontal lines, respectively.

$$M_{(n)}(i,j) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} I_M\left[i + k - 1 - \frac{(n-1)}{2}, j + l - 1 - \frac{(n-1)}{2}\right]. \quad (4)$$

The MSE is zero for all neighborhoods where equal numbers of observed and predicted points exceed the threshold value, regardless of the position of the points within each neighborhood. To mitigate fluctuations in the MSE related to coverage, as well as to ensure that the FSS retains a value between 0 (no skill) and 1 (perfect skill) Roberts and Lean normalize the score with a reference $\mathrm{MSE}_{(n)_{\mathrm{ref}}}$ representing the largest possible MSE from a given set of observed and predicted fractions:

$$\mathrm{MSE}_{(n)_{\mathrm{ref}}} = \frac{1}{N_x N_y}\left[\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} O^2_{(n)_{i,j}} + \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} M^2_{(n)_{i,j}}\right]. \quad (5)$$

The above system was used to calculate the FSS over scales ranging from a single grid point (27 km) to a 41×41 point square (1107 km). Again to minimize bias issues, observations from 2100 UTC were used exclusively. The resulting curves in Fig. 8 show the progression of the FSS over the range of scales listed above. Point-scale FSS values are quite low, but the FSS rapidly increases with increasing neighborhood scale. This behavior is consistent with the example in Fig. 3 reflecting that the general patterns in the forecast are better than the point-scale values.

The neighborhood scale with the optimum combination of forecast accuracy with minimal loss in precision due to averaging is largely up to the individual user. Since the neighborhood method is relatively new few studies exist to offer comparative scores from other cloud forecasts. Söhne et al. (2006) reported FSS values near 0.8 at length scales of 150 km for a set of 6-h brightness temperature forecasts of anvil cloudiness for a single flash flood case. This case was clearly well simulated considering the horizontal grid spacing was 50 km in some of their sensitivity studies. Their score is likely on the high end of the forecast quality scale, especially for convection. The eastern Pacific FSS in our study is somewhat lower, though probably more representative because of the large sample size. Murphy and Epstein (1989) also noted that the FSS is sensitive to the grid coverage of the field subject to the threshold. Large coverage tends to score higher, thus the grid size, threshold, and event climatology must be similar for adequate comparison.

Roberts and Lean (2008) derived two simple measures of quality based on the FSS for random and uniform forecasts. The FSS for random forecasts with the same fractional coverage over the domain as the event to be verified is simply equal to the coverage, which was about 0.12 in this case (Fig. 8). Uniform forecasts were defined as a forecast at the gridpoint scale ($n = 1$) with the fraction/probability equal to the average fractional

coverage of the event to be verified. The uniform FSS is defined as being half-way between the random and perfect skill, which in this case was 0.56 (Fig. 8). Uniform forecasts possess a reasonable amount of skill, but have zero precision. Roberts and Lean contend that FSS scores above the uniform score represent the smallest scale over which the forecast contains useful information. Applying this standard here indicates that length scales between 135 and 189 km (5–7 grid points) should provide useful forecasts. This result is not unreasonable given well-known constraints on model resolution. Additional tests with the GOES persistence forecasts showed a large degradation by 21 h (Fig. 8). The persistence FSS remained below the uniform skill threshold for neighborhoods below 900 km on a side. Unfortunately, LWP persistence was not available for most other lead times prior to 21 h because of daylight constraints.

c. *Composite statistics*

The question of forecast utility and optimal length scale can be further investigated using composite methods developed by Nachamkin (2004). Composites provide visual and quantitative feedback on the average state of the forecasts and observations when specific events are expected. The basic premise of the composite method involves identifying events of interest in both the forecasts and the observations and creating composites based on the event occurrence. Accurate forecasts result in strong similarities between the observed and predicted spatial distributions, while systematic spatial errors manifest themselves as displacements in the composited fields. For this study moderate- to small-sized events were composited in an effort to gauge the mesoscale performance. All deep cloud events containing between 100 and 600 contiguous grid points with LWP values equaling or exceeding 500 g m^{-2} were composited. These typically represent convective cloud clusters, small developing cyclones, or cold cutoff portions of mature cyclones. The events in this composite reside in the low to midportions of the spatial range that the neighborhood method indicated would be viable forecasts (Fig. 8).

The composite contingent on the occurrence of a predicted event in the 21-h forecasts is displayed in Fig. 9. The sample comprised 86 events distributed over the analysis region. For a given event, all points on the 41×41 point (1107 km $\times$ 1107 km) sample grid that fell outside the satellite analysis region were ignored as bad data. Thus, the gradient in the number count in the northern portions of the composite (Fig. 9b) indicates that a number of events occurred near the northern boundary of the satellite analysis region (50°N). This result reflects the prevalence of cyclone activity north of

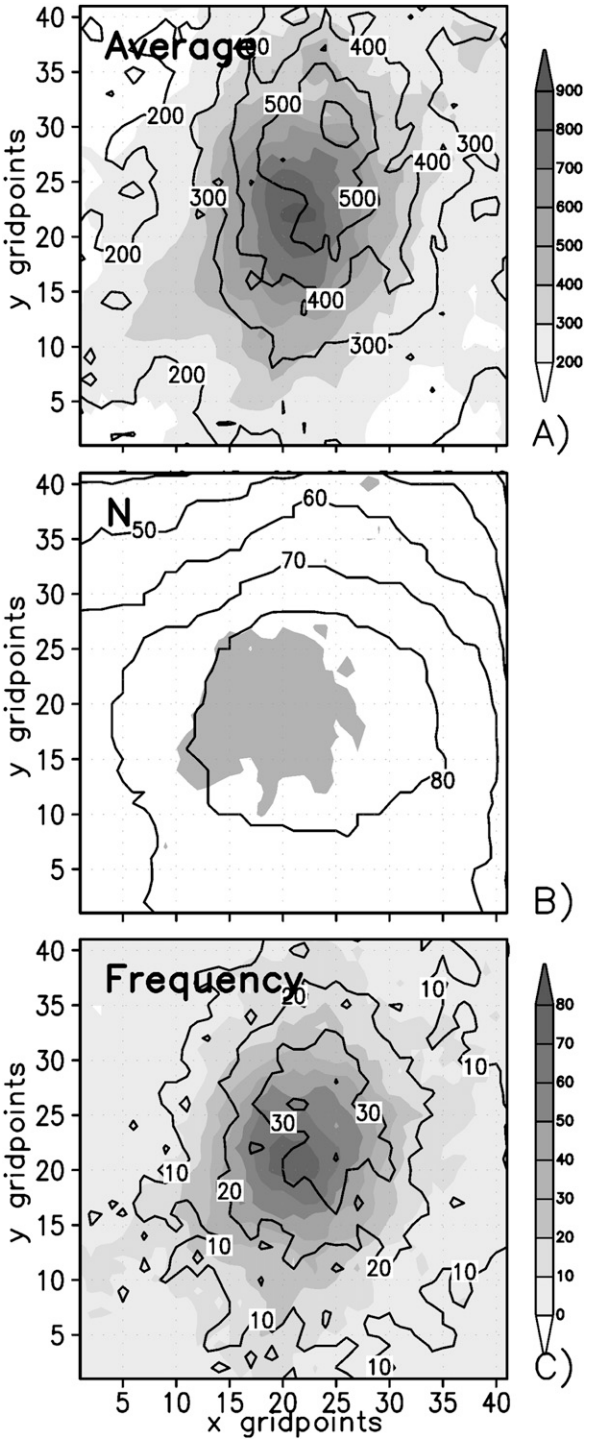

FIG. 9. Statistics from the composite based on the existence of a predicted mesoscale event in the 21-h forecasts. (a) The average observed (contoured) and predicted (shaded) LWP values are in grams per meter squared. (b) The number of valid data points is contoured and the region of LWP differences between the observed and predicted fields $\geq$50 g m^{-2} at the 95% confidence level is shaded. (c) The observed (contoured) and predicted (shaded) occurrence frequency of LWP values $\geq$500 g m^{-2}.

40°N. A distinct shift was evident in both the average LWP (Fig. 9a), and event frequency distributions (Fig. 9c) suggesting that the model forecasts were too far south and west of the observations. Given that the prevailing flow is often southwesterly during disturbed weather the temporal lag in Fig. 4 is consistent with this spatial phase shift. The statistical significance of the spatial shift was tested using a t distribution with a null hypothesis that the observed and predicted average LWP differed by at least 50 g m^{-2}. The region of statistical significance at the 95% confidence level, indicated by the shading in Fig. 10b, covers the center of the predicted events, but does not extend northeastward to the corresponding lobe of increased observed LWP (Fig. 9a). Reduced number counts and high standard deviations lowered the statistical confidence in this area.

The composite of the forecasts contingent on the occurrence of an observed event (Fig. 10) depicts a less distinct, though still apparent spatial shift. The region of statistical significance (Fig. 10b) extends farther northeast, due in part to increased number counts. A total of 124 events were sampled, indicating more small- to midsized events existed in the observations than the forecasts. The disparity likely results from the enhanced variability in the observations compared to the forecasts (Skamarock 2004). Although the observations were averaged to the model grid, a given cyclone often consisted of several closely associated mesoscale cloud areas. Note the broken nature of the observed frontal cloudiness in Fig. 3 compared to the forecast. As such the observation-based composite likely included a number of essentially synoptic-scale events. Cloud coverage constraints could serve to filter the observed events, though such efforts were not attempted here as both composites indicate the same basic trends.

A greater understanding of the nature of the forecast errors as well as the verification results can be achieved by combining concepts used in both the neighborhood and composite methods. Although the FSS is very useful for comparing forecasts with one another, applying the FSS to obtain an optimal length scale is somewhat abstract. How do the curves in Fig. 8 translate to forecast performance? In an attempt to address this issue, events in the composite sample above were recomposited using criteria based roughly on the FSS. For each event, a single sample FSS was defined using Eq. (1), where $MSE_{(n)}$ was the binary MSE over the $n \times n$ region centered on the event. The quantity $MSE_{(n)_{ref}}$ was defined individually for each event sample using Eq. (5). Event forecasts were grouped into high- and low-quality categories based on whether the single sample FSS met or exceed the average FSS at scale n as shown in Fig. 8. While the comparison is only an approximate analogy, it

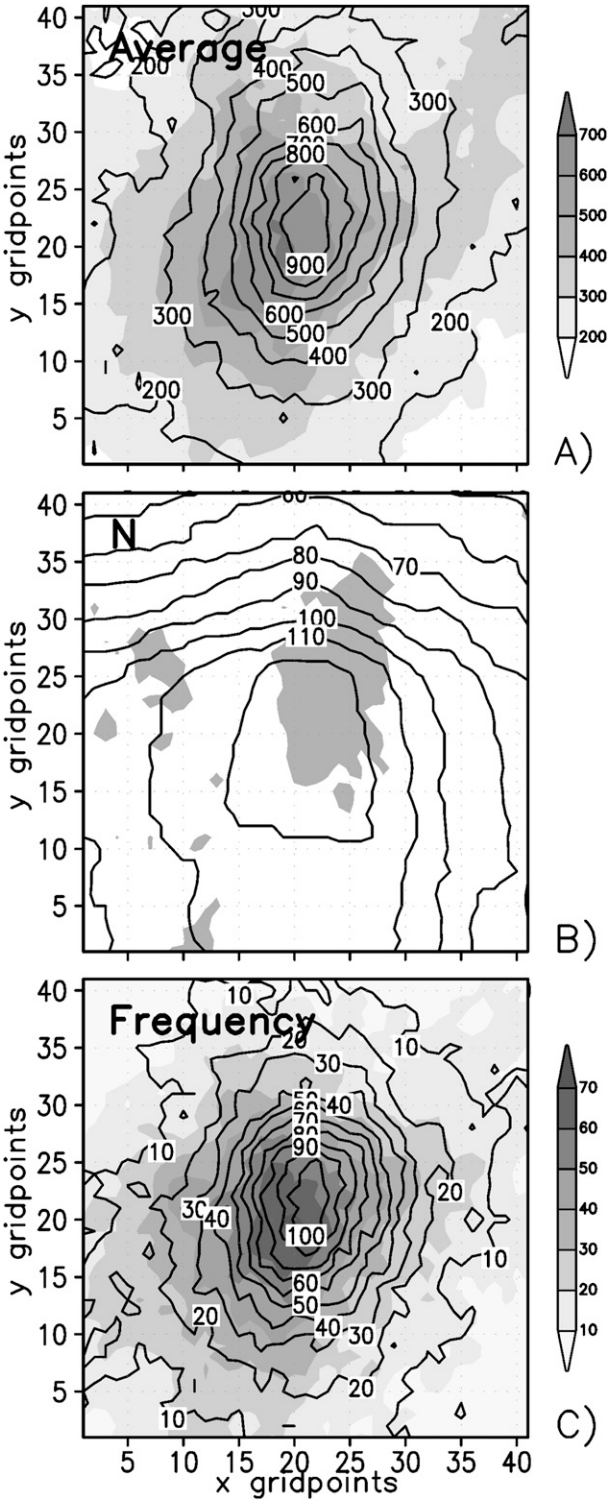

FIG. 10. As in Fig. 9, but for the composite based on the occurrence of an observed mesoscale event.

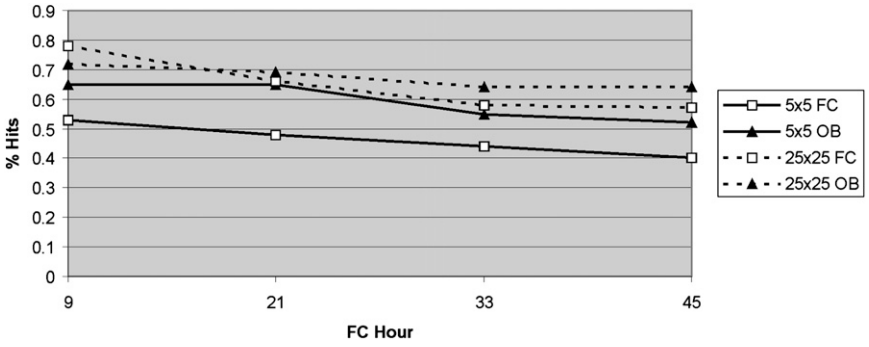

FIG. 11. The percentage of correct mesoscale event forecasts (hits) as defined by the FSS criteria for the 5-(solid lines) and 25-point (dashed lines) neighborhood length. Solid triangles represent the values derived from the observation-based composite while open squares represent those from the forecast-based composite.

does provide insight regarding the ability of the FSS to assess error as well as the nature of the error associated with the lower-quality forecasts.

Two FSS criteria at opposite ends of the spectrum, squares with sides of $n = 5$ and 25 grid points (135 and 675 km), were applied to the mesoscale events depicted in Figs. 9 and 10. For consistency, the FSS for the 21-h forecasts was applied to define the forecast quality at all lead times. Based on the calculations in Fig. 8 the critical FSS values for the 5 and 25 point areas were 0.57 and 0.79, respectively. As might be expected, the five-point criteria were quite restrictive (Fig. 11). The small box size required a large portion of very close or collocated positive forecasts and observations to exceed a given FSS. Only 40%–50% of the events in the forecast-based composite met or exceeded the critical FSS, while a somewhat greater 55%–65% of the observation-based events were accepted. Relaxing the criteria to 25 points pushed the acceptance rate above 70% for both composites during the early portions of the forecast, with decreased rates at longer lead times. The gains in the forecast-based composite were markedly higher than those in the observation-based composite, reflecting the smaller phase shift as well as the likelihood that some larger-scale events were incorporated in the sample.

The effect of the scale-based acceptance criteria can be illustrated by comparing composites of the forecasts that were defined to be of high and low quality. Note in the figures that the high-quality forecasts are loosely referred to as ''hits.'' The forecast-based event frequency distributions for the 21-h forecasts show strong agreement between the observations and predictions for the five-point criteria hits (Fig. 12a). As mentioned above, the strict criteria selected only those events where the forecasts agreed closely with the observations. In Fig. 12b, the composite of the five-point false alarms[4] displays considerably less agreement between the predicted and observed event frequencies. However, a coherent region of observed events is evident to the north and east of the forecasts. These observations result from a subset of shifted forecasts that were still useful despite having failed the small-scale quality criteria. Increasing the box scale to 25 points (Fig. 12c) shifted these forecasts to the higher-quality bin. The resulting frequency distributions are similar to the original composite (Fig. 10). In fact, observed event frequencies were close to zero over most of the low-quality composite (Fig. 12d), indicating that these forecasts were truly false alarms. The FSS frequency distributions for the observation-based composite (not shown) display similar trends to those depicted in Fig. 12. Although a greater number of these events met the five-point FSS criteria, a subset of phase-shifted forecasts was still evident in the five-point low-quality forecasts. Similarly, the 25-point low-quality forecasts depicted very little agreement between the observed and predicted frequency distributions. Forecasts that fail the 25-point criteria likely contain significant error.

A separate composite study using large events with sizes ranging from 601 to 3000 contiguous grid points with LWP values at or above the 500 g m^{-2} depicts much improved performance (Fig. 13). For both the 5- and 25-point FSS criteria, over 80% of the forecasts were binned as high quality at all lead times for the observation-based composite. The forecast-based rates were somewhat lower. The differences between the 5- and 25-point acceptance rates were less than those for the small events, primarily because of the increased

[4] Because this composite is based on the occurrence of a forecast event, a low-quality forecast will most likely be a false alarm.

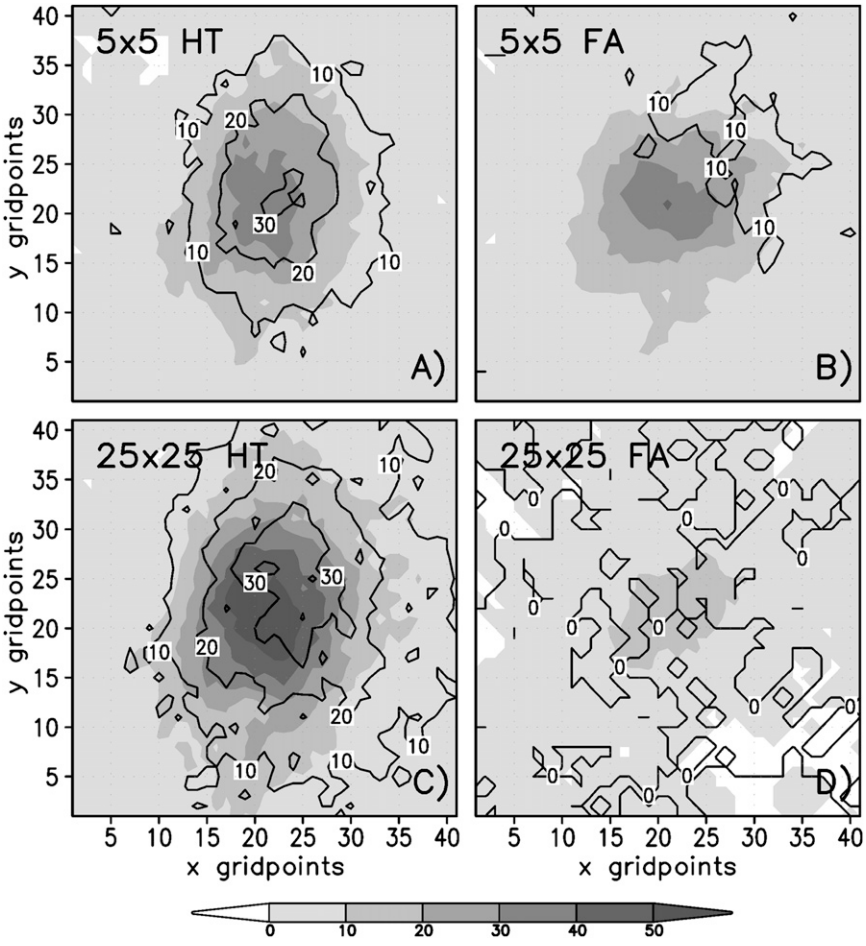

FIG. 12. Forecast-based composites of hits and false alarms as determined by the (a),(b) 5- and (c),(d) 25-point FSS criteria applied to the 21-h mesoscale event forecasts. Observed occurrence frequencies of LWP values $\geq$500 g m^{-2} are contoured while predicted frequencies are shaded. (left) Hits and (right) false alarms.

event size and attendant increase in coverage near the event center. Notably, displacement errors were much less evident, with no statistically significant trends. At this scale, the event shapes were quite complex, resulting in increased variability near the edges. Also, fewer events were sampled, with sample sizes ranging from 38 to 53 depending on the lead time. Large sample sizes may be necessary to determine if no systematic shifts truly exist at this scale. Given these caveats, the primary source of error for the large-scale events was a tendency for the model to predict too much cloud cover. The resulting false alarm tendency was responsible for the reduced quality in the forecast-based composite (Fig. 13).

4. Conclusions

In this study, the performance of the COAMPS deep cloud forecasts was evaluated using observed GOES LWP retrievals as ground truth. Manual inspections of the LWP observations in conjunction with visible and infrared satellite imagery indicate a strong likelihood that deep cloud systems were well represented by the retrievals. However, some care should be taken when interpreting forecast quality from these results as the satellite observations are subject to considerable variability that is difficult to estimate. The greatest errors resulted from an underestimate of cloud depth at values above 800 g m^{-2}. Presenting the results in a bias-corrected form helped alleviate these issues.

The deep cloud forecasts evaluated during this period displayed sufficient accuracy with respect to the satellite observations to be considered quite useful. The majority of the synoptic-scale systems ($\geq$1000 km) were well simulated with the primary error being an over abundance of deep cloudiness in the forecasts. At smaller scales, the model failed to capture the variability

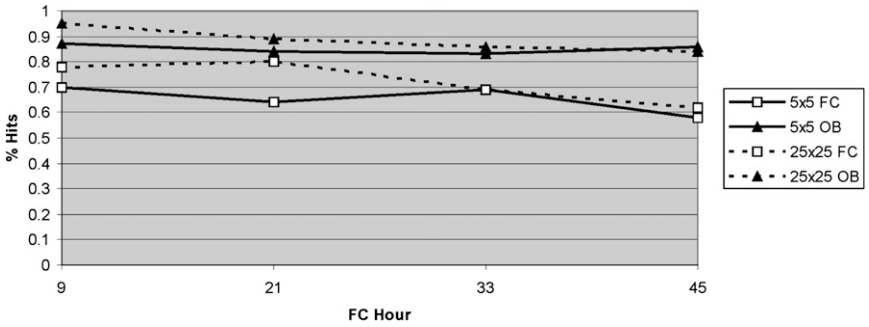

FIG. 13. As in Fig. 11, but for synoptic-scale events with sizes ranging from 601 to 3000 contiguous grid points with LWP $\geq$ 500 g m^{-2}.

depicted in the observations. Observed synoptic systems often consisted of several closely associated meso-alpha-scale cloud masses that the model tended to depict as contiguous areas. About 40%–50% of the subsynoptic event predictions exceeded the 5 $\times$ 5 point FSS-based quality standard, displaying little or no phase error. Another 20%–30% of the subsynoptic forecasts contained more moderate errors, including phase errors, but still passed a 25 $\times$ 25 point FSS quality standard. The remaining forecasts contained significant errors. Lag correlation calculations indicate the spatial errors translate to a slow timing bias of approximately 3 h. In general, the model was able to discern between quiescent and cloudy periods over the scale of the entire region with correlation coefficients of 0.82, 0.87, 0.80, and 0.80 at lead times of 9, 21, 33, and 45 h. Clear-sky coverage was overestimated by the model, though many of the overestimates occurred in regions that were partly cloudy in the observations. When all observed cells containing less than 50% clouds counted were assigned clear values, the underestimate was alleviated. Such a discrepancy likely results from the lack of subgrid-scale cloudiness. The forecast skill as defined by the correlations with the observations indicates that the model beat persistence at lead times beyond and including 9 h. Considering that these forecasts were located in an oceanic region with relatively few observations, the results of this evaluation are quite encouraging.

Future work in this area will focus on improving the cloud forecasts. Additional studies are planned to determine the effects of horizontal and vertical grid resolution as well as adjustments to the microphysics and turbulence schemes. Verification of cloud layer and cloud-top characteristics are also planned. New instruments such as *CloudSat* (more information is available online at http://cloudsat.atmos.colostate.edu), the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO, more information is available online at http://www-calipso.larc.nasa.gov), and the Atmospheric Infrared Sounder (AIRS, more information is available online at http://airs.jpl.nasa.gov) offer additional information regarding cloud depth and cloud height that can be used to refine current estimates. Research is currently being conducted toward this end.

Acknowledgments. This research is supported by the Office of Naval Research (ONR) through Program Element N0001408WX21169. The COAMPS and GOES data archival and processing were supported in part by a grant of high performance computing (HPC) time from the Department of Defense Major Shared Resource Center, Stennis Space Center, Mississippi. The work was performed on a Sun F12000 and an IBM P575+ computer. Computing time was also supported by an HPC grant from FNMOC as part of their Distributed Center for computing. The work was performed on an SGI Origin supercomputer. The GOES data were supplied by the Naval Research Laboratory. Kim Richardson (NRL) and Steve Miller (Cooperative Institute for Research in the Atmosphere, Colorado State University) also provided considerable assistance with the GOES retrievals.

REFERENCES

Bieringer, P., M. Donovan, F. Robasky, D. Clark, and J. Hurst, 2006: A characterization of NWP ceiling and visibility forecasts for the terminal airspace. Preprints, *12th Conf. on Aviation Range and Aerospace Meteorology,* Atlanta, GA, Amer. Meteor. Soc., P3.6. [Available online at http://ams.confex.com/ams/Annual2006/techprogram/paper_103720.htm.]

Carter, G. M., and H. R. Glahn, 1976: Objective prediction of cloud amount based on model output statistics. *Mon. Wea. Rev.,* **104,** 1565–1572.

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.,* **11,** 141–154.

Chaboureau, J.-P., and J.-P. Pinty, 2006: Validation of a cirrus parameterization with Meteosat Second Generation observations. *Geophys. Res. Lett.,* **33,** L03815, doi:10.1029/2005GL024725.

——, J.-P. Cammas, P. Mascart, J.-P. Pinty, and J.-P. Lafore, 2002: Mesoscale cloud scheme assessment using satellite observations. *J. Geophys. Res.,* **107,** 4301, doi:10.1029/2001/JD000,714.

Chevallier, F., and G. Kelly, 2002: Model clouds as seen from space: Comparison with geostationary imagery in the 11-μm window channel. *Mon. Wea. Rev.,* **130,** 712–722.

Daley, R., and E. Barker, 2001: NAVDAS source book 2001: NRL atmospheric variational data assimilation system. NRL/PU/7530-01-441, Naval Research Laboratory, Monterey, CA, 163 pp.

Davies, H. C., 1976: A lateral boundary formulation for multi-level prediction models. *Quart. J. Roy. Meteor. Soc.,* **102,** 405–418.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.,* **15,** 51–64.

——, and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

——, L. J. Wilson, B. G. Brown, P. Nurmi, H. E. Brooks, J. Bally, and M. Jaeneke, 2004: Verification of nowcasts from the WWRP Sydney 2000 Forecast Demonstration Project. *Wea. Forecasting,* **19,** 73–96.

Heidinger, A., 2003: Rapid daytime estimation of cloud properties over a large area from radiance distributions. *J. Atmos. Oceanic Technol.,* **20,** 1237–1250.

Hodur, R. M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Wea. Rev.,* **125,** 1414–1430.

Hogan, T. F., M. S. Peng, J. A. Ridout, and W. M. Clune, 2002: A description of the impact of changes to NOGAPS convection parameterization and the increase in resolution to T239L30. NRL Memo. Rep. NRL/MR/7530-02-52, Naval Research Laboratory, Monterey, CA, 10 pp.

Kain, J. S., and J. M. Fritsch, 1993: Convective parameterization for mesoscale models: The Kain–Fritsch scheme. *The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr.,* No. 46, Amer. Meteor. Soc, 165–170.

Li, J.-L., and Coauthors, 2005: Comparisons of EOS MLS cloud ice measurements with ECMWF analyses and GCM simulations: Initial results. *Geophys. Res. Lett.,* **32,** L18710, doi:10.1029/2005GL023788.

Meyers, M. P., P. J. DeMott, and W. R. Cotton, 1992: New primary ice-nucleation parameterizations in an explicit cloud model. *J. Appl. Meteor.,* **31,** 708–721.

Miller, S. D., G. L. Stephens, C. K. Drummond, A. K. Heidinger, and P. T. Partain, 2000: A multisensor diagnostic satellite cloud property retrieval scheme. *J. Geophys. Res.,* **105,** 19 955–19 971.

Mitrescu, C., J. M. Haynes, G. L. Stephens, S. D. Miller, G. M. Heymsfield, and M. J. McGill, 2005: Cirrus cloud optical, microphysical and radiative properties observed during CRYSTAL-FACE experiment: A lidar-radar retrieval system. *J. Geophys. Res.,* **110,** D09209, doi:10.1029/2004JD005605.

——, S. Miller, and R. Wade, 2006: Cloud optical and microphysical properties derived from satellite data. Preprints, *14th Conf. on Satellite Meteorology and Oceanography,* Atlanta, GA, P1.7. [Available online at http://ams.confex.com/ams/Annual2006/techprogram/paper_100515.htm.]

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572–581.

Nachamkin, J. E., 2004: Mesoscale verification using meteorological composites. *Mon. Wea. Rev.,* **132,** 941–955.

Nakajima, T., and M. D. King, 1990: Determination of the optical thickness and effective particle radius of clouds from reflected solar radiation measurements. Part I: Theory. *J. Atmos. Sci.,* **47,** 1878–1893.

Norquist, D. C., 1999: Cloud predictions diagnosed from mesoscale weather model forecasts. *Mon. Wea. Rev.,* **127,** 2465–2483.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.,* **136,** 78–97.

Rutledge, S. A., and P. V. Hobbs, 1983: The mesoscale and microscale structure and organization of clouds and precipitation in midlatitude cyclones. VIII: A model for the ''seeder-feeder'' process in warm-frontal rainbands. *J. Atmos. Sci.,* **40,** 1185–1206.

——, and ——, 1984: The mesoscale and microscale structure and organization of clouds and precipitation in midlatitude cyclones. XII: A diagnostic modeling study of precipitation development in narrow cold-frontal rainbands. *J. Atmos. Sci.,* **41,** 2949–2972.

Schmidt, J. M., 2001: Moist physics development for the Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *BACIMO Conf.,* Fort Collins, CO, Army Research Laboratory, CD-ROM.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.,* **132,** 3019–3032.

Söhne, N., J.-P. Chaboureau, S. Argence, D. Lambert, and E. Richard, 2006: Objective evaluation of mesoscale simulations of the Algiers 2001 flash flood by the model-to-satellite approach. *Adv. Geosci.,* **7,** 247–250.

Zepeda-Arce, J., E. Foufoula-Georgiou, and K. K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.,* **105** (D8), 10 129–10 146.