A Comparative Verification of Localized Aviation Model Output Statistics Program (LAMP) and Numerical Weather Prediction (NWP) Model Forecasts of Ceiling Height and Visibility

DAVID E. RUDACK AND JUDY E. GHIRARDELLI

Meteorological Development Laboratory, Office of Science and Technology, NOAA/NWS, Silver Spring, Maryland

(Manuscript received 18 November 2009, in final form 3 March 2010)

ABSTRACT

In an effort to support aviation forecasting, the National Weather Service's Meteorological Development Laboratory (MDL) has recently redeveloped the Localized Aviation Model Output Statistics (MOS) Program (LAMP) system. LAMP is designed to run hourly in NWS operations and produce short-range aviation forecast guidance at 1-h projections out to 25 h. This paper compares and contrasts LAMP ceiling height and visibility forecasts with forecasts produced by the 20-km Rapid Update Cycle model (RUC20), the Weather Research and Forecasting Nonhydrostatic Mesoscale Model (WRF-NMM), and the Short-Range Ensemble Forecast system (SREF). RUC20 and WRF-NMM forecasts of continuous ceiling height and visibility were interpolated to stations and converted into categorical forecasts. These interpolated forecasts were also categorized into instrument flight rule (IFR) or lower conditions and verified against LAMP forecasts at stations in the contiguous United States. LAMP and SREF probabilistic forecasts of ceiling height and visibility from LAMP and the SREF system were also verified. This study demonstrates that for the 0000 and 1200 UTC cycles over the contiguous United States, LAMP station-based categorical forecasts of ceiling height, visibility, and IFR conditions or lower are more accurate than the RUC20 and WRF-NMM ceiling height and visibility forecasts interpolated to stations. Moreover, for the 0900 and 2100 UTC forecast cycles and verification periods studied here, LAMP ceiling height and visibility probabilities exhibit better reliability and skill than the SREF system.

1. Introduction

The meteorological and aviation communities have recently undertaken a coordinated effort to improve ceiling height (CIG) and visibility (VIS) forecasts. These forecasts are not only valuable in a societal context (e.g., flight delays and guarding against the loss of life) but are also instrumental in making economic decisions that routinely impact airline operations.

Terminal aerodrome forecasts (TAFs) are official aviation forecasts produced by the National Weather Service (NWS) four times daily, with amendments as necessary. TAFs consist of forecasts of critical weather elements, such as CIG and VIS, which are expected to impact an airport over a specific time period. This time period is usually 24 h, with selected airports requiring TAFs out to 30 h, the first 6 h being recognized as the critical TAF period (NWS 2008).

The NWS's Meteorological Development Laboratory (MDL) has been producing objective statistical guidance in the form of model output statistics (MOS) since the 1970s (Glahn and Lowry 1972). To provide guidance to the aviation forecaster preparing TAFs, MDL produces a short-term statistically based forecast guidance product termed the Localized Aviation MOS Program (LAMP). LAMP is designed to update the Global Forecast System (GFS) MOS forecast guidance on an hourly basis with hourly forecasts extending out to 25 h in advance (Ghirardelli 2005; Ghirardelli and Glahn 2010).[1] LAMP generates both categorical and probabilistic guidance for a variety of weather elements with special emphasis on those affecting the aviation community. The

Corresponding author address: David E. Rudack, Meteorological Development Laboratory, National Weather Service, 1325 East–West Hwy., Silver Spring, MD 20910.
E-mail: david.rudack@noaa.gov

[1] In this paper, the acronym LAMP refers to the GFS LAMP system that updates the GFS MOS and not to the Nested Grid Model (NGM) LAMP system, which has been discontinued.

DOI: 10.1175/2010WAF2222383.1

guidance includes, but is not limited to, probabilistic and categorical forecasts of CIG and horizontal VIS.

Verification of LAMP forecast guidance has shown improvements over the GFS MOS forecasts for VIS (Rudack 2005) during the 1–9-h projections and improvements throughout the 25-h forecast period for CIG (Weiss and Ghirardelli 2005). Moreover, LAMP displays better accuracy than persistence during the LAMP forecast period, even in the short term of 1–6 h. This is a time frame in which persistence forecasts are regarded as being highly competitive (Dallavalle and Dagostaro 1995).

Modeling the conditions and various meteorological processes that lead to low CIG and poor VIS is complex. For example, the independent or combined effects of terrain, water bodies, soil moisture, and radiation fluxes may in some instances produce poor VIS. Yet, in other similar meteorological situations, VIS may not be reduced at all. Despite these complicating factors, the Global Systems Division's (GSD) 20-km Rapid Update Cycle model (RUC; Benjamin et al. 1999; Benjamin et al. 2004; hereafter referred to as the RUC20), the National Centers for Environmental Prediction's (NCEP) Weather Research and Forecasting (WRF) Nonhydrostatic Mesoscale Model (NMM; Skamarock et al. 2005; hereafter referred to as the WRF-NMM), and the Short-Range Ensemble Forecasting (SREF) system (Zhou et al. 2004) have begun producing CIG and VIS forecasts. Very little has been published concerning the verification of these forecasts.

To quantify the overall utility of LAMP CIG and VIS guidance to the forecast process at TAF sites, we verified LAMP CIG and VIS forecasts along with forecasts produced by the RUC20, WRF-NMM, and SREF over the contiguous United States (CONUS). This verification study was conducted by pooling 1462 stations across the CONUS. While this approach yields a CONUS-wide average performance measure and may not be representative of specific sites, we believe that verifying the data in this manner provides useful information concerning the overall strengths and weaknesses of these forecasting systems.

The structure of this paper is as follows. Section 2 briefly discusses the models and types of data used in this verification study as well as how the dynamical model data were interpolated to stations for verification purposes. Verification results are presented in sections 3 and 4 followed by a summary and concluding remarks in section 5.

2. Model data and methodology

LAMP produces forecasts from multiple linear regression equations that update the GFS MOS guidance and provides forecasts at an hourly resolution out to 25 h in advance. Most forecast weather elements are available for 1591 stations located in the CONUS, Alaska, Hawaii, Puerto Rico, and the Virgin Islands.[2] The predictor sources that are used as input to the LAMP regression system include observations, GFS MOS forecasts, climatic variables (e.g., cosine of the day of the year), and forecasts generated by simple advective models. Forecast equations can be developed regionally; that is, a regression equation is developed with data obtained from several stations in a region. In this instance, guidance for all stations in that region is generated from that same set of regression equations. An alternative approach, which is more commonly used in forecasting temperature, dewpoint, wind speed, and wind direction, involves developing a regression equation that applies to a specific station. The regional approach is typically used in situations where the forecast events are rare and a larger sample is required to stabilize the regression analysis. This is the approach used in developing LAMP CIG and VIS forecast equations.

The RUC20 and WRF-NMM are dynamical models that generate continuous CIG and VIS forecasts through postprocessing algorithms. The models produce forecasts on a 20- and 12-km horizontal resolution Lambert Conformal grid, respectively. The SREF system is ensemble based and produces probabilistic forecasts that are interpolated onto a 40-km horizontal resolution Lambert Conformal grid. The SREF used in this study comprises 10 perturbations from the Eta Model, 5 perturbations from the Regional Spectral Model, 3 perturbations from the WRF-NMM, and 3 perturbations from the National Center for Atmospheric Research's version of the Weather WRF model (NCAR-WRF), totaling 21 members (Du et al. 2004). The RUC20, WRF-NMM, and SREF VIS algorithms are discussed in Stoelinga and Warner (1999), Smirnova et al. (2000), and Zhou et al. (2004).

With respect to the RUC20, we would have preferred to use the higher-resolution 13-km RUC model; however, an archive of that RUC data was not available. Although the coarser 20-km horizontal resolution RUC20 model does degrade the quality of CIG and VIS forecasts, the forecasts are still considered skillful (S. G. Benjamin 2007, personal communication).

a. Verification data

To evaluate the quality of categorical CIG and VIS forecasts, LAMP station-based CIG and VIS probabilistic and categorical forecasts were generated from 0000 and 1200 UTC initial conditions for the period of

[2] Real-time hourly updated GFS-based LAMP forecasts are available online (http://weather.gov/mdl/gfslamp/gfslamp.shtml).

October 2006–September 2007.[3] These weather forecasts are at an hourly resolution from 1 to 25 h in advance. Operational model data from the 0000 and 1200 UTC runs of the RUC20 and WRF-NMM were retrieved for the verification period. Specifically, the model analyses and forecasts of continuous CIG and VIS at an hourly resolution extending out to 12 h for the RUC20 and 25 h for the WRF-NMM were collected. To evaluate the performance of LAMP and SREF probability forecasts, the 0900 and 2100 UTC postprocessed SREF CIG and VIS probability forecasts were retrieved for the same verification period. The collected SREF probabilities are at 3-h intervals extending out to 24 h. For a specific element and category, the SREF probability forecasts of CIG represent a consensus relative frequency based on all 21 members. For example, if 10 of the 21 members predict a CIG of ≤3000 ft, the probability forecast for CIG of ≤3000 ft would be 48%. Also note that unlike LAMP and the verifying observations that account for low VIS caused by nonhydrometeors such as haze or blowing phenomena, the SREF VIS forecasts only account for hydrometers (Zhou et al. 2004).

RUC20 and WRF-NMM continuous forecasts were verified for CIG of ≤3000, <1000, and <500 ft, and VIS of <3, <1, and <½ mi. Since correctly forecasting instrument flight rule (IFR) conditions or lower is extremely important in aviation forecasting, these forecasts were also verified. IFR or lower conditions occur when either the lowest CIG is <1000 ft and/or the VIS is <3 mi (NWS 2008).

The verification periods for both categorical and probabilistic forecasts were stratified into two seasons. The cool season spanned October 2006 through March 2007, while the warm season stretched from April through September 2007. This is consistent with the seasonal stratification used in the development and verification of MDL's LAMP and MOS statistical products. All cross-model comparison verification scores are derived from matched samples.

Although the LAMP system covers both the CONUS and areas outside the CONUS, the domains of the RUC20, WRF-NMM, and SREF are primarily restricted to the CONUS. Thus, the verification domain was limited to the CONUS. A total of 1462 aviation routine weather report (METAR) stations in the CONUS are used in this verification study. Figure 1 depicts the system domains and the locations of the 1462 CONUS stations.

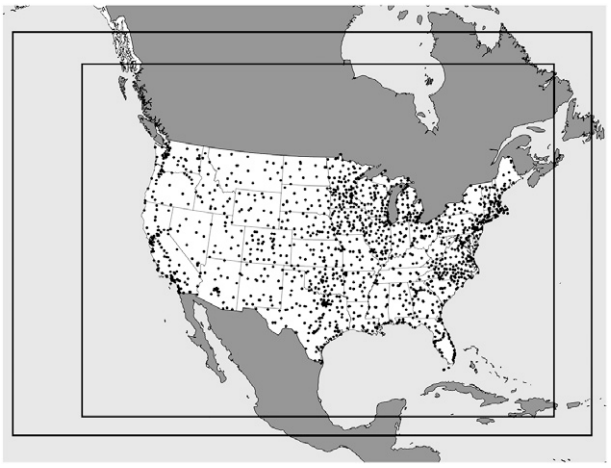

FIG. 1. Domains of the forecast systems investigated. The outer rectangle indicates the area of both the WRF-NMM and SREF domains. The inner rectangle indicates the area of the RUC20 domain. The dots indicate the locations of the 1462 LAMP stations in the CONUS, which coincide with the verification points.

b. Model data conversion

The verification in this study is intended to assess the quality of guidance available for TAFs valid at stations. Therefore, the model forecasts had to be interpolated to stations while accounting for the discontinuous nature of CIG and VIS. A nearest-neighbor matching technique was used; that is, the CIG or VIS value at the grid point closest to a station was assigned to that station. Generally, a more realistic representation of the discontinuous field at stations is preserved when applying this technique, rather than, for instance, using bilinear interpolation. To correctly treat the gridded RUC20 CIG values, which are given relative to sea level, the CIG forecasts were adjusted by subtracting the station elevation from the CIG forecast. This results in CIG forecasts relative to the ground. Since the WRF-NMM CIG forecasts are defined relative to the model surface, the station-based CIG forecasts were calculated by adding the differences between the gridpoint elevations and the station elevations to the nearest-neighbor matched CIG forecast.

Once the model-based CIG and VIS forecasts were properly matched to stations, the continuous forecast values were binned into the categories defined in the LAMP system (see Tables 1 and 2). Note that the RUC20 and WRF-NMM CIG and VIS data were converted from Système International (SI) to English units. CIG forecasts were then rounded to the nearest hundred feet. The

[3] The reader should note that prior to June 2007, the 0000 and 1200 UTC LAMP cycles were not yet implemented operationally. Consequently, to obtain forecasts for the verification period studied here, we retrospectively generated LAMP forecasts in a manner following the same techniques used in the operational setting. A separate study was performed (using the same metrics in this paper) comparing the LAMP operational CIG and VIS forecasts to the regenerated forecasts for the period of June–September 2007 and we found that the scores were virtually identical.

TABLE 1. LAMP categories of ceiling height.

Category	Ceiling height (ft)
1	<200
2	200–400
3	500–900
4	1000–1900
5	2000–3000
6	3100–6500
7	6600–12 000
8	>12 000 or unlimited ceiling

TABLE 2. LAMP categories of visibility.

Category	Visibility (mi)
1	<1/2
2	1/2 to <1
3	1 to <2
4	2 to <3
5	3 to 5
6	6
7	>6

LAMP CIG and VIS categories are consistent with significant aviation flight category levels (NWS 2008).

The processing of the 0900 and 2100 UTC SREF probabilistic forecasts was handled somewhat differently than the RUC20 and WRF-NMM CIG and VIS forecasts. To retain the spatial regularity that is generally observed for probabilistic forecasts, these SREF forecasts were interpolated to stations by using bilinear interpolation—a standard MDL practice for fields that are generally spatially well behaved. For verification purposes, each interpolated CIG and VIS forecast station value was converted from a percent to a probability ranging between 0 and 1, inclusively.

3. Verification results of categorical forecasts

Categorical and probabilistic forecasts require different metrics to evaluate their accuracy or skill. The threat score or critical success index (CSI) is often used to determine the accuracy of categorical forecasts (e.g., forecasts of CIG $<$ 1000 ft) (Wilks 2006). A perfect forecasting system has a CSI of one while a CSI value of zero represents the worst possible score. Since one metric does not adequately describe the quality of a forecasting system, we also evaluate CIG and VIS categorical forecasts by using the bias. Bias, in the context of categorical forecasts, is the ratio of the number of forecasts for a particular event (e.g., forecasts of CIG $<$ 1000 ft) divided by the number of observed occurrences of that event (Wilks 2006). A bias value of one (unit bias) means that the system forecasts the occurrence of the event with the same frequency that it is observed. A bias value greater (less) than one means that the system is overforecasting (underforecasting) the occurrence of that particular event.

a. Categorical ceiling height forecasts

Since the LAMP, RUC20, and WRF-NMM categorical CIG forecast CSI scores are very similar at both 0000 and 1200 UTC, only results from the 0000 UTC cycle will be presented unless otherwise noted. Figure 2 displays the CSI and bias values for categorical CIG forecasts issued from the 0000 UTC LAMP, WRF-NMM, and RUC20 for the 2006–07 cool season. (Recall that the RUC20 model forecasts beyond the 12-h projection were not available.) The CSI values in Figs. 2a–c exhibit the same general pattern of behavior. At every projection, LAMP forecasts are as accurate, or more so, than persistence, the RUC20, and the WRF-NMM. Improvement over persistence and other models in the very-short term (1–6 h) distinguishes LAMP. Prior to the 9-h projection, LAMP demonstrates marked improvement over the RUC20 and WRF-NMM. Although LAMP CSI scores tend to level off beginning at the 9-h projection, the CSI scores still remain at or above all three other forecast systems throughout the 12-h forecast period.[4] However, the accuracy of the RUC20 and WRF-NMM is comparable to LAMP during the 9–12-h projections for CIG $\leq$ 3000 ft. For projections beyond 12 h, CSI scores for WRF-NMM categorical CIG forecasts for all three categories remain noticeably lower than LAMP, with the exception of CIG forecasts of $\leq$3000 ft for projections of 12–15 h, when LAMP and WRF-NMM scores are very close.

Higher CSI values can reflect a system's tendency to overforecast the frequency of the event (i.e., the system has a bias greater than 1.0). To investigate this, we calculated the corresponding bias values for CIG (Fig. 2). During the 1–12-h projections for CIG $\leq$ 3000 ft, bias values are generally close to one. However, a noticeable increase in bias is detected for the RUC20 and WRF-NMM during this period for CIG $<$ 1000 and $<$500 ft. In contrast, the LAMP bias values for the same categories remain close to 1.0. For projections of 12–25 h, LAMP forecast bias values remain relatively constant. In contrast, WRF-NMM bias values deteriorate for projections 15 through 25 for all three CIG categories. Note that the bias of the WRF-NMM is as high as 4.5 for CIG $<$ 500 ft at the 23-h projection.

Though not shown here, the accuracy of the WRF-NMM forecasts for CIG $\leq$ 3000 ft appears to be closely

[4] Confidence testing was not done on any of the verification results discussed in this paper.

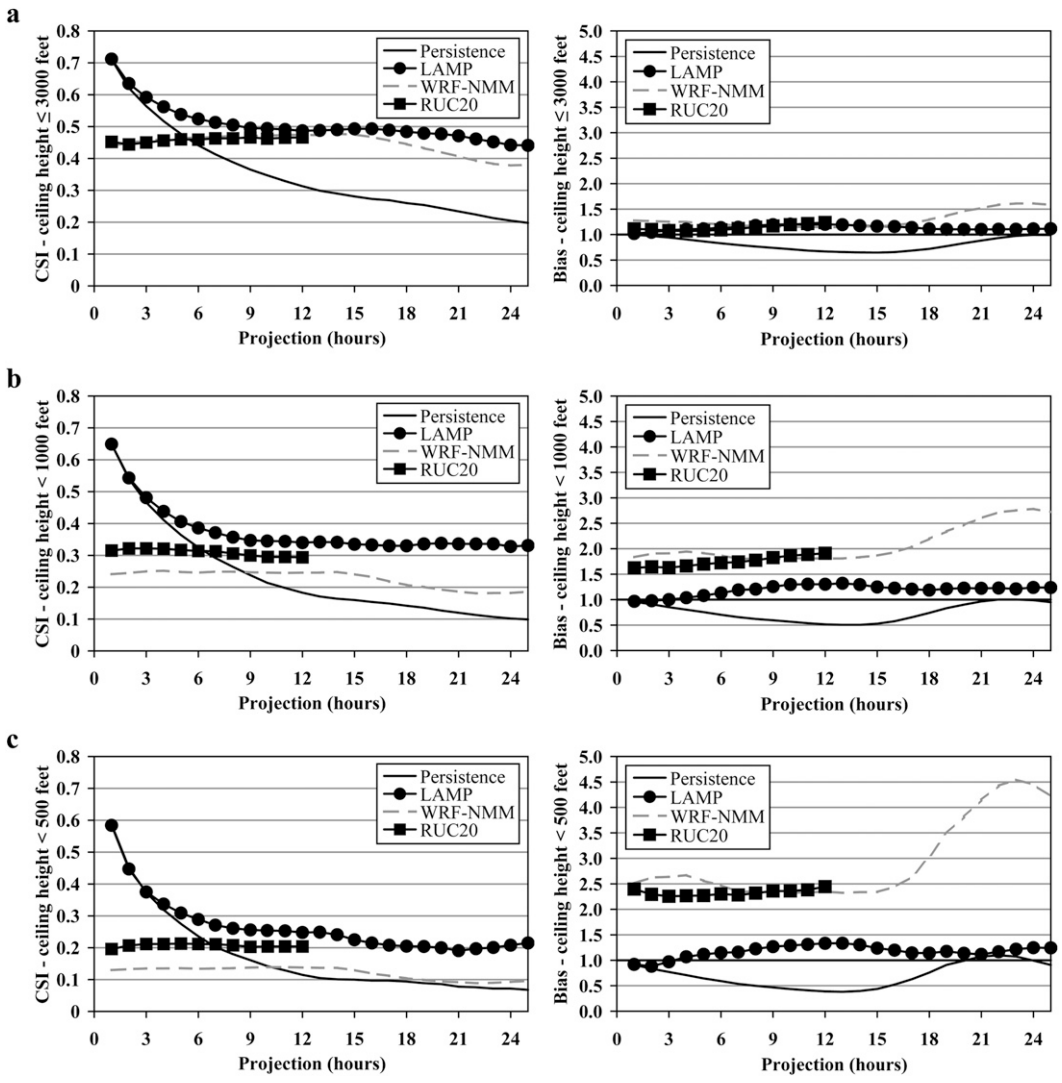

FIG. 2. (left) CSI and (right) bias for categorical forecasts of ceiling height (a) ≤3000, (b) <1000, and (c) <500 ft. Shown are persistence, LAMP, WRF-NMM, and RUC20 scores from the 0000 UTC cycle for the cool season of Oct 2006–Mar 2007.

related to the time of day for which the forecasts are valid rather than the forecast projection itself. Figure 2a shows the 0000 UTC WRF-NMM CSI scores increasing through the 1–12-h projections and decreasing thereafter. In contrast, the 1200 UTC CSI scores (not shown) reach a minimum by the 12-h projection and increase thereafter. That is to say, the traces of the 0000 and 1200 UTC CSI scores are in phase with each other with respect to the verifying hour of the day.

b. Categorical visibility forecasts

The overall pattern of behavior of the verification scores for categorical VIS forecasts (Fig. 3) is similar to the CIG categorical forecasts. One difference, however, is the sharp decline in the LAMP CSI scores for VIS forecasts in the first 6 h. A second distinction is that the LAMP CSI scores become almost constant from the 6-h projection onward. This leveling-off period begins approximately 2 h earlier than in the CIG plots.

Of note, the RUC20 and WRF-NMM CSI scores for all three categories and all projections are low, with values hovering at or below 0.20. While the CSI scores do not vary much between the RUC20 and WRF-NMM during the 1–12-h forecast period, the bias values do behave quite differently (Fig. 3). The WRF-NMM bias values for each category generally remain uniform with only a slight modulation. However, the RUC20 bias values

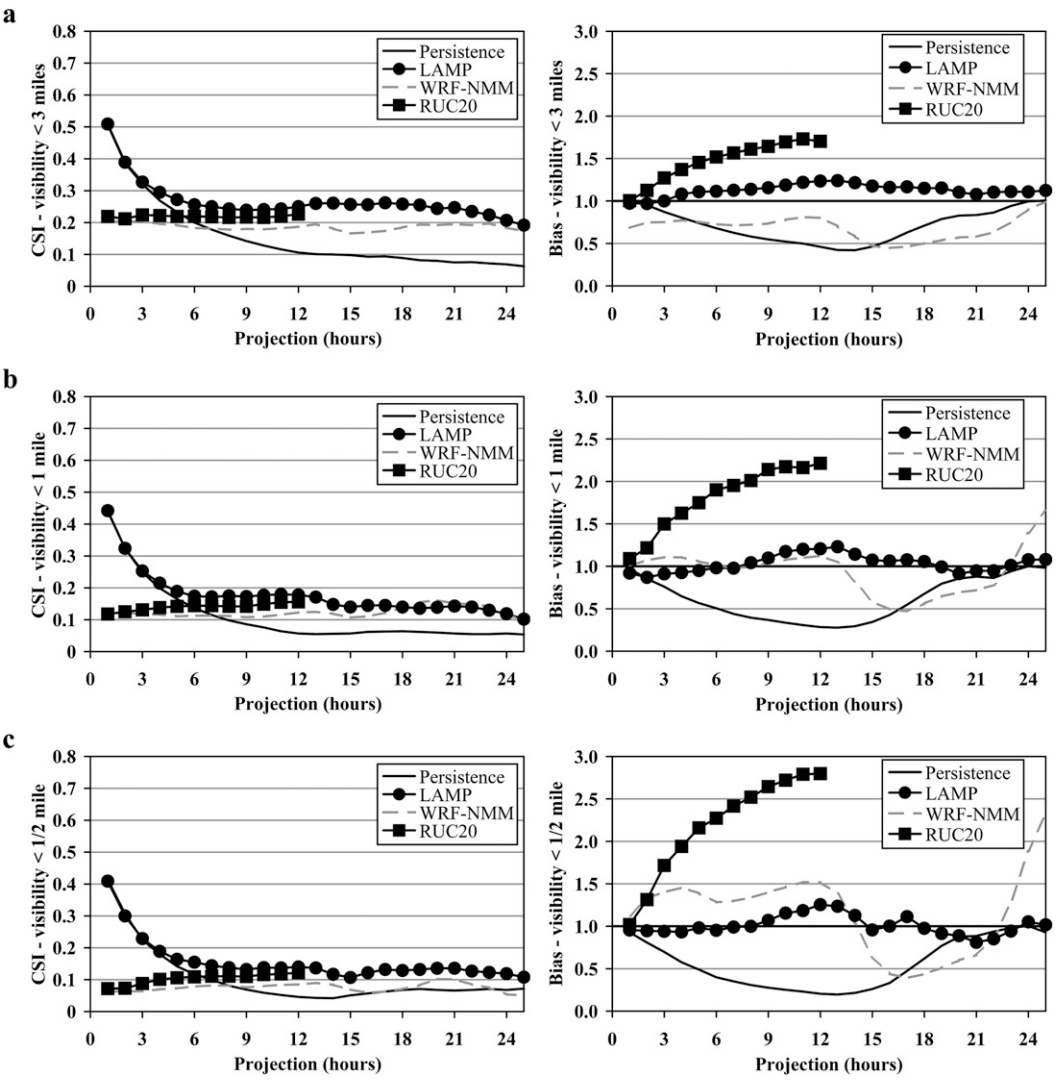

FIG. 3. (left) CSI and (right) bias for categorical forecasts of visibility (a) <3, (b) <1, and (c) <½ mi. Shown are persistence, LAMP, WRF-NMM, and RUC20 scores from the 0000 UTC cycle for the cool season of Oct 2006–Mar 2007.

steadily increase by projection when moving from VIS < 3 miles (mi) to VIS < ½ mi. We suspect that these bias differences between the RUC20 and WRF-NMM can be partially attributed to the different weights given to the low-level relative humidity (which is known to have an early morning high bias) used in the visibility algorithm (G. Manikin 2009, personal communication).

During the 13–25-h projections, LAMP VIS CSI scores generally remain higher than those of the WRF-NMM, with the exception of forecasts of <1 mi in the 19–23-h projection range, where the WRF-NMM scores are slightly higher. Unlike the initial 12 projections for the WRF-NMM, subsequent projections exhibit a pronounced bias oscillation (especially for VIS < 1 mi and VIS < ½ mi) (Fig. 3). Note, however, as in the earlier projections, LAMP bias scores do not deviate much from unit bias.

c. *IFR or lower forecasts*

Figure 4a shows the CSI and bias values for IFR conditions or lower for the 2006–07 cool season. Scores are shown for persistence, LAMP, RUC20, and the WRF-NMM. As expected, the same overall pattern of behavior exhibited in the categorical CIG and VIS verifications is evident here. As demonstrated by the CSI, LAMP is more accurate than all other systems at every projection. The performance of the RUC20 and WRF-NMM is generally constant through all 12 projections with CSI values of RUC20 forecasts generally higher than those from the WRF-NMM.

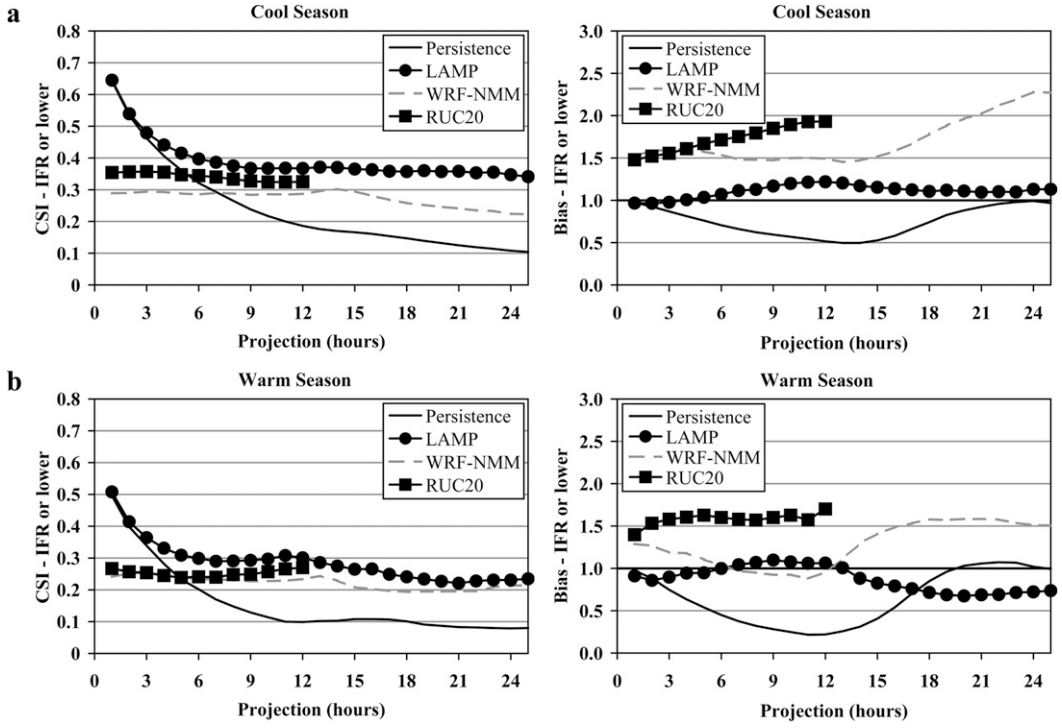

FIG. 4. (left) CSI and (right) bias for forecasts of IFR or lower conditions for the (a) cool season Oct 2006–Mar 2007 and (b) warm season Apr–Sep 2007. Shown are persistence, LAMP, WRF-NMM, and RUC20 scores from the 0000 UTC cycle.

IFR and lower bias scores for LAMP, RUC20, and WRF-NMM for the cool season generally reflect the composite bias scores shown in Figs. 2 and 3 for CIG $<$ 1000 ft and VIS $<$ 3 mi, respectively. LAMP biases generally remain close to 1.0 throughout the 25 projections while the WRF-NMM biases remain close to 1.5 through the 12-h projection and then steadily increase thereafter. The RUC20 bias values steadily increase through the 12-h projection due to the high bias associated with forecasting VIS $<$ 1 mi.

Figure 4b displays the warm season (April–September 2007) CSI and bias verification results for IFR conditions or lower. For the CSI, the verifications generally exhibit patterns of behaviors similar to their cool season counterparts with some notable exceptions. As expected, the CSI scores for IFR or lower conditions for all systems are lower when compared to the cool season. This result is consistent with the less frequent occurrence of this phenomenon in the warm season. The warm season RUC20 and WRF-NMM CSI score values shadow each other much more closely throughout the 12-h forecast period compared to the cool season CSI scores. The corresponding LAMP and WRF-NMM bias scores for the warm season display a diurnal signal and are completely out of phase with respect to each other (Fig. 4b).

4. Verification results of probabilistic forecasts

Recently, end users of aviation products have expressed interest in using probability forecasts of CIG and VIS. These probabilities can be incorporated into cost–loss models to make critical economic decisions (Keith and Leyton 2007).

In 2006, the National Research Council (NRC 2006) released a report characterizing and communicating uncertainty information. This report discusses the problems with deterministic forecasts and how they can be misleading without understanding the underlying uncertainties. It further charges that the NWS has a responsibility to provide products that communicate the underlying forecast uncertainty. LAMP provides probabilistic forecast guidance and, as such, provides information relating to forecast uncertainty.

The *P* score, which was proposed by Brier (1950), is often used to quantify the quality of a set of probabilistic forecasts. The *P* score represents the mean-squared error for probabilistic forecasts and ranges between 0 and 2, inclusively. (The results presented in this paper are the half-*P* score and range between values of 0 and 1, inclusively. We will hereafter refer to the half-*P* score as the Brier score.) The Brier score comprises three terms: reliability, resolution, and uncertainty (Murphy 1973; Wilks

2006). Reliability and resolution provide information on the bias and sharpness, respectively, of the probability forecasts within specific ranges and, thus, provide some indication of the quality of the probabilities. The uncertainty term is a measure of the complexity associated with forecasting the event and is quantified in terms of the climatology of the event.

Lower Brier scores indicate greater accuracy while higher scores indicate the opposite. However, one should not assume that the Brier score defines the usefulness of a forecasting system. For example, probabilities for rare events such as VIS $<$ ½ mi may result in a low Brier score but may not be any more accurate than a probability forecast equal to the climatic relative frequency of the event. Thus, to demonstrate skill, the forecasts must be compared to some baseline. When two systems are compared to each other, the Brier skill score (Wilks 2006) is typically used to evaluate the skill of one system over the other. We would have liked to compute the Brier skill score for each system using climatic relative frequencies as the reference system; however, a long-term climatology of CIG and VIS for all 1462 stations used in this verification study was not available. Instead, we have computed the Brier skill score for the LAMP forecasts by using the SREF forecasts as the reference system.

The reliability diagram is one method for visually assessing the bias behavior of a set of probabilistic forecasts. Probability forecasts for a particular event (e.g., VIS $<$ 1 mi) are generally partitioned across evenly spaced bins (0%–10%, 10%–20%, . . . , 90%–100%) and are compared to the observed relative frequency of the event within that bin. Probabilistic forecasts are deemed reliable when the average probability forecast and the average observed frequency of the event are about the same for all or a majority of bins.

During the verification period examined here, an error was present in the process used to generate the operational SREF probabilities for CIG (J. Du 2009, personal communication). For this paper, we have corrected this mistake and have plotted the reliabilities of both the operational and corrected CIG probabilities. Since, however, the incorrect probabilities are considered the official operational forecast, our verification discussion is restricted to those forecasts.

a. Probabilistic ceiling height forecasts

1) RELIABILITY

To compare the LAMP and SREF probabilities, reliability diagrams were generated for each of the categories discussed in sections 3a and 3b for the 3-, 6-, 9-, 12-, 18-, and 24-h projections for the stations and seasons noted earlier. Figure 5 displays the cool season reliability diagrams for CIG probability forecasts of ≤3000 ft issued at 0900 UTC. Note that LAMP exhibits better reliability for all projections; scores close to the diagonal line of perfect reliability are desirable. The SREF demonstrates comparable reliability to LAMP at the 3- and 6-h projections with only a slight tendency to underforecast in the lower probability bins. For the 12-, 18-, and 24-h projections the SREF becomes less reliable with a clear tendency to overforecast CIG. The 2100 UTC SREF reliability (Fig. 6) at the 3- and 6-h projections display a tendency to overforecast CIG ≤ 3000 ft. Similar behavior is noted for the same time of day in the 0900 UTC SREF forecasts for the 12- and 18-h projections. The behavior of the SREF 0900 and 2100 UTC reliabilities appears to be at least partly influenced by the time of day for which the forecasts are valid. This is akin to the behavior noted above for continuous CIG for the WRF-NMM (section 3a). The 0900 and 2100 UTC warm season reliabilities for CIG ≤ 3000 ft (not shown) display the same overall temporal behavior as is seen during the cool season.

The 0900 and 2100 UTC SREF forecasts for CIG $<$ 1000 ft (Figs. 7 and 8) exhibit a distinct overforecasting bias—even more than what was noted for CIG ≤ 3000 ft. This is true across the 3-, 6-, 12-, 18-, and 24-h projections and for both the cool and warm seasons (not shown). Unlike the SREF reliabilities for CIG ≤ 3000 ft, the reliabilities for CIG $<$ 1000 ft do not appear to vary as a function of the time of day. Note that LAMP displays better reliability than the SREF for all the projections examined here.

As a word of caution, reliability scores in the higher probability bins must be interpreted judiciously. The significant drop in the number of forecasts in these bins by both LAMP and the SREF (as shown in the histograms in Figs. 5–8) makes evaluating the reliability difficult. When the number of forecasts in these bins is small, chance plays a role in the calculation of the reliability score.

2) BRIER SKILL SCORE

Brier skill scores (using the SREF as the reference system) were calculated for the same cycles and verification periods described in the previous section. Figure 9a shows the Brier skill scores for CIG $<$ 1000 ft for the 0900 UTC 2006–07 cool season. LAMP demonstrates better skill as seen by the Brier skill score (between 43% and 32%) for all forecast projections. The maximum improvement occurs at the 3-h projection with a second relative maximum at the 15-h projection. The LAMP improvement over the SREF is maximized at the 3-h projection partly because LAMP uses the METAR observations (of CIG) valid at the initial cycle time as

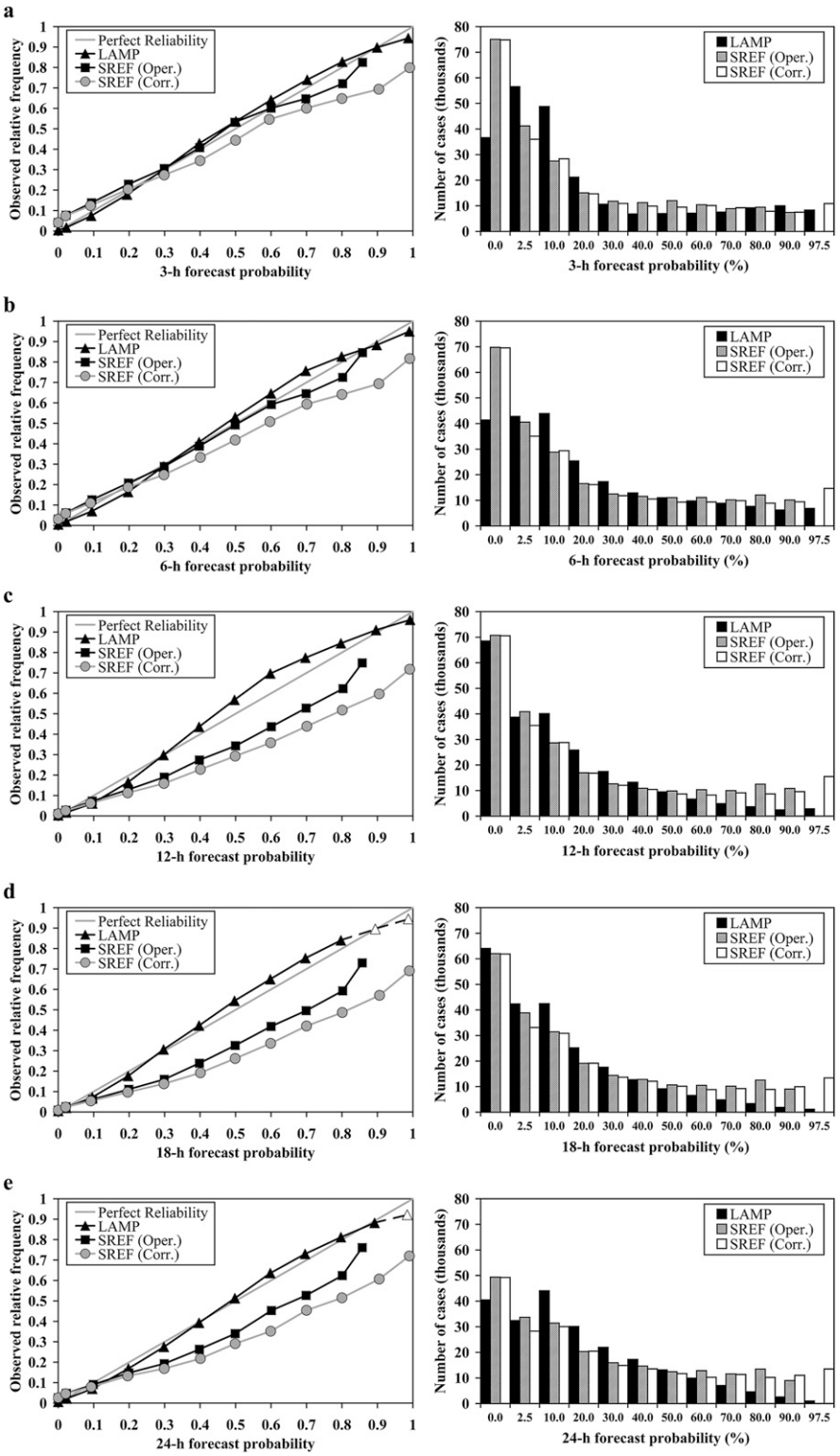

FIG. 5. (left) Reliability diagrams and (right) histograms for probabilities of ceiling height ≤3000 ft for the (a) 3-, (b) 6-, (c) 12-, (d) 18-, and (e) 24-h projections for LAMP and SREF. Verification is from the 0900 UTC cycle for the cool season of Oct 2006–Mar 2007. Reliability lines from both the operational (Oper.) and corrected (Corr.) SREF probability forecasts are included. Reliability values (left side) composing less than 1% of the total number of cases are indicated with a hollow marker and a dashed line to emphasize that these values should be interpreted judiciously.

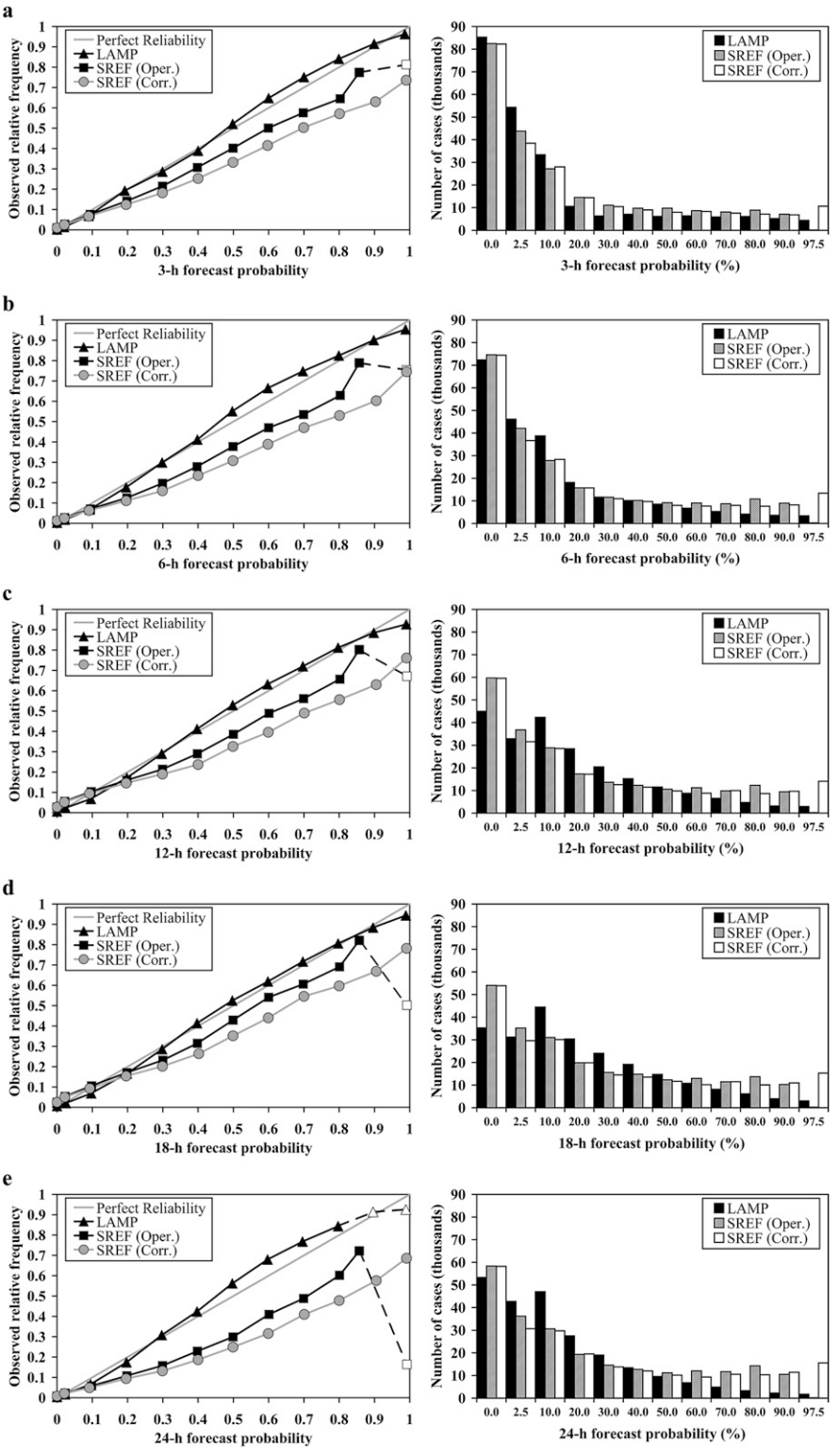

FIG. 6. As in Fig. 5, but for 2100 UTC.

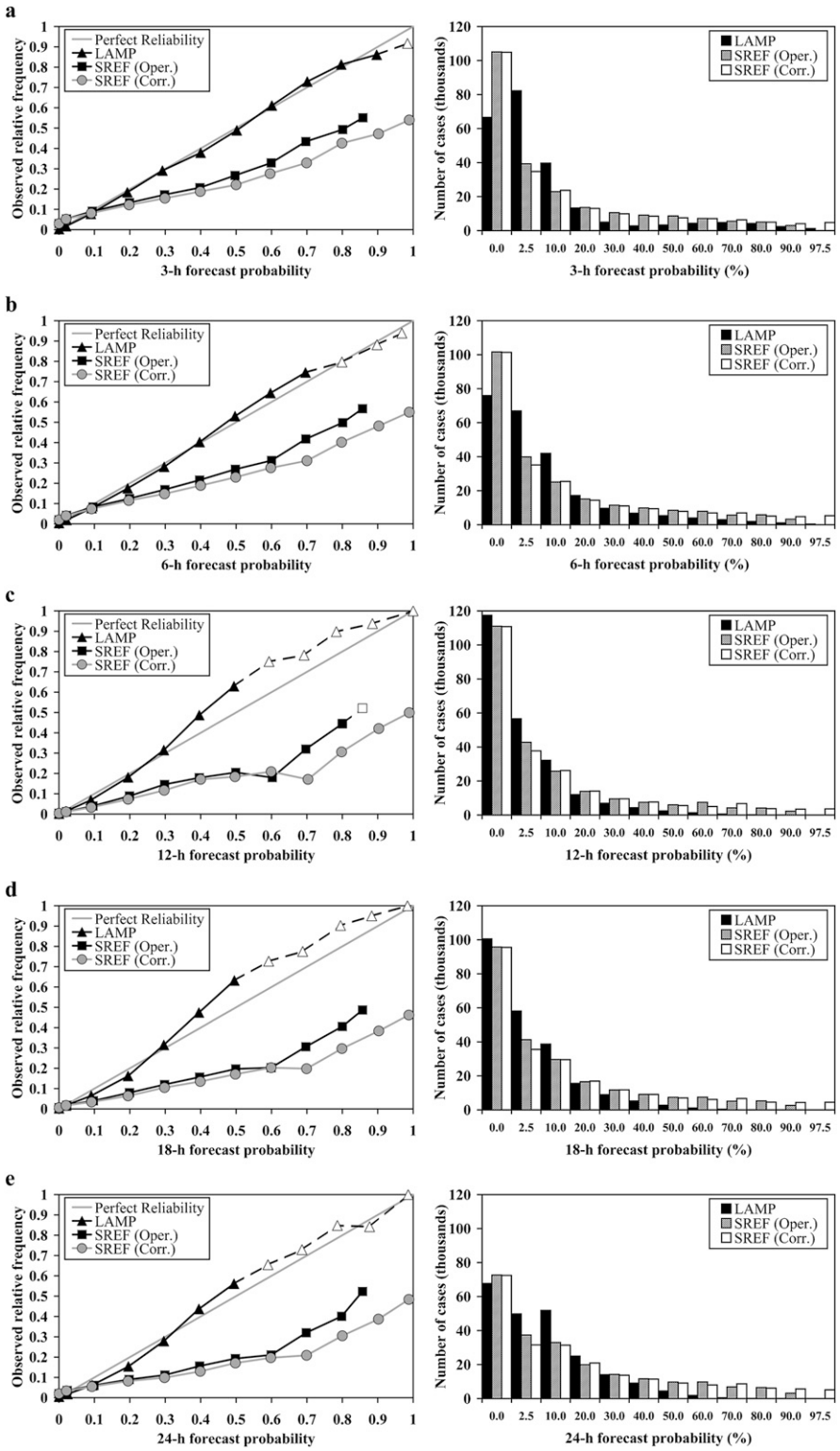

FIG. 7. As in Fig. 5, but for ceiling height <1000 ft.

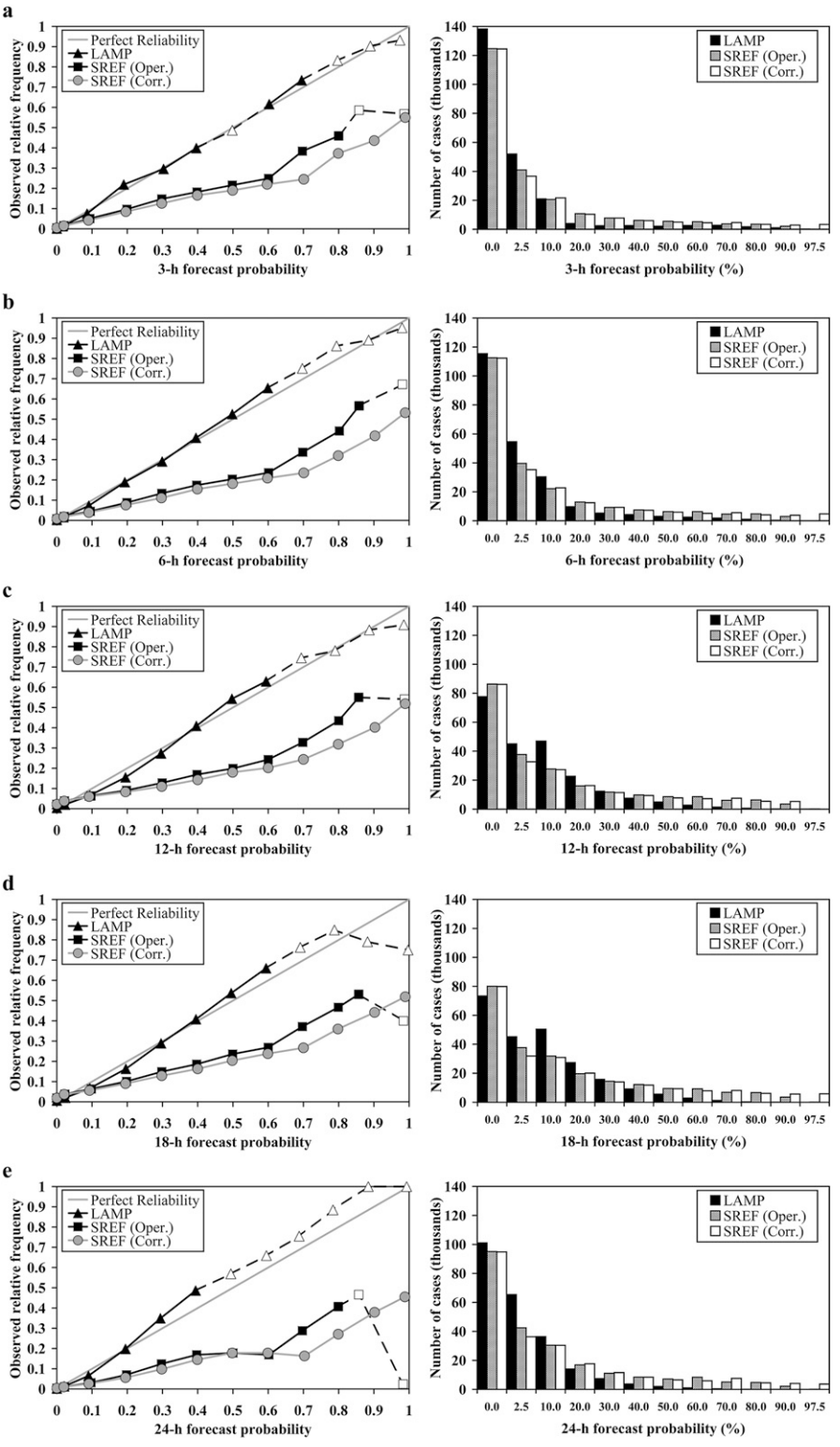

FIG. 8. As in Fig. 7, but for 2100 UTC.

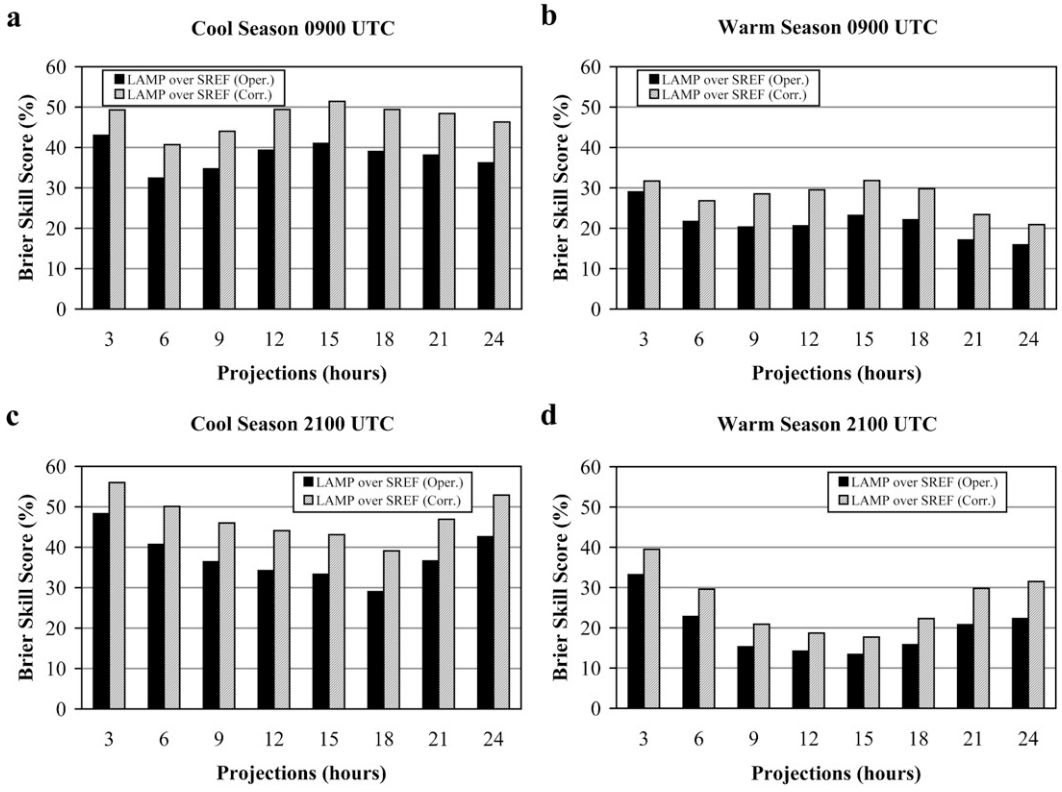

FIG. 9. Brier skill scores of LAMP over SREF for probability ceiling height forecasts of <1000 ft. Verification is from (a) the 0900 and (c) 2100 UTC cycles for the cool season of Oct 2006–Mar 2007 and (b) the 0900 and (d) 2100 UTC cycles for the warm season of Apr–Sep 2007. Brier skill scores from both the operational (Oper.) and corrected (Corr.) are included.

predictors. Observations, when used as persistence forecasts, are very difficult to improve upon in the very short term. By including the most recent observation as a predictor, LAMP consistently produces forecasts with comparable or better skill than persistence forecasts, even in the very short term. The 0900 UTC warm season scores (Fig. 9b) exhibit the same overall type of behavior as seen in the cool season except that the overall LAMP improvement is less. LAMP improvement over SREF is evident at every projection with a maximum occurring at the 3-h projection. The same overall improvement of LAMP over the SREF is also noted for the 2100 UTC cycle for both the cool and warm seasons (Figs. 9c and 9d) with the greatest improvement occurring at the 3-h projection.

b. Probabilistic visibility forecasts

1) RELIABILITY

The 0900 and 2100 UTC LAMP cool (Figs. 10 and 11, respectively) and warm season (not shown) probability forecasts of VIS $<$ 3 mi demonstrate better reliability than the SREF for all projections. LAMP forecasts for the 3- and 6-h projections exhibit no discernable bias, while the forecasts for 12 and 18 h tend to underforecast VIS at the higher probabilities. In contrast, the reliabilities for the SREF probability forecasts exhibit discernable projection-to-projection fluctuations. These vacillations are independent of season and cycle time and appear to be related to the time of day for which the forecast is valid. The VIS reliability results exhibit a diurnal rhythm similar to the CIG forecasts of $\leq$3000 ft, albeit more pronounced. This rhythm can possibly be attributed to the tendency of some of the ensemble models to oversaturate the lowest model levels during the late night and early morning hours. The sensitivity of the VIS algorithm to the overabundance of hydrometeors (or surface cloud water) could explain the distinct overforecasting of VIS $<$ 3 mi during this time period (G. Manikin 2009, personal communication). Similar diurnal fluctuations are not present in LAMP probabilistic forecasts.

The overall 0900 UTC cool season SREF probability forecasts for VIS $<$ 3 mi generally exhibit better reliability than the 2100 UTC cool season forecasts. While both cycles have a bias toward overforecasting, the 0900 and 2100 UTC SREF reliabilities are competitive with

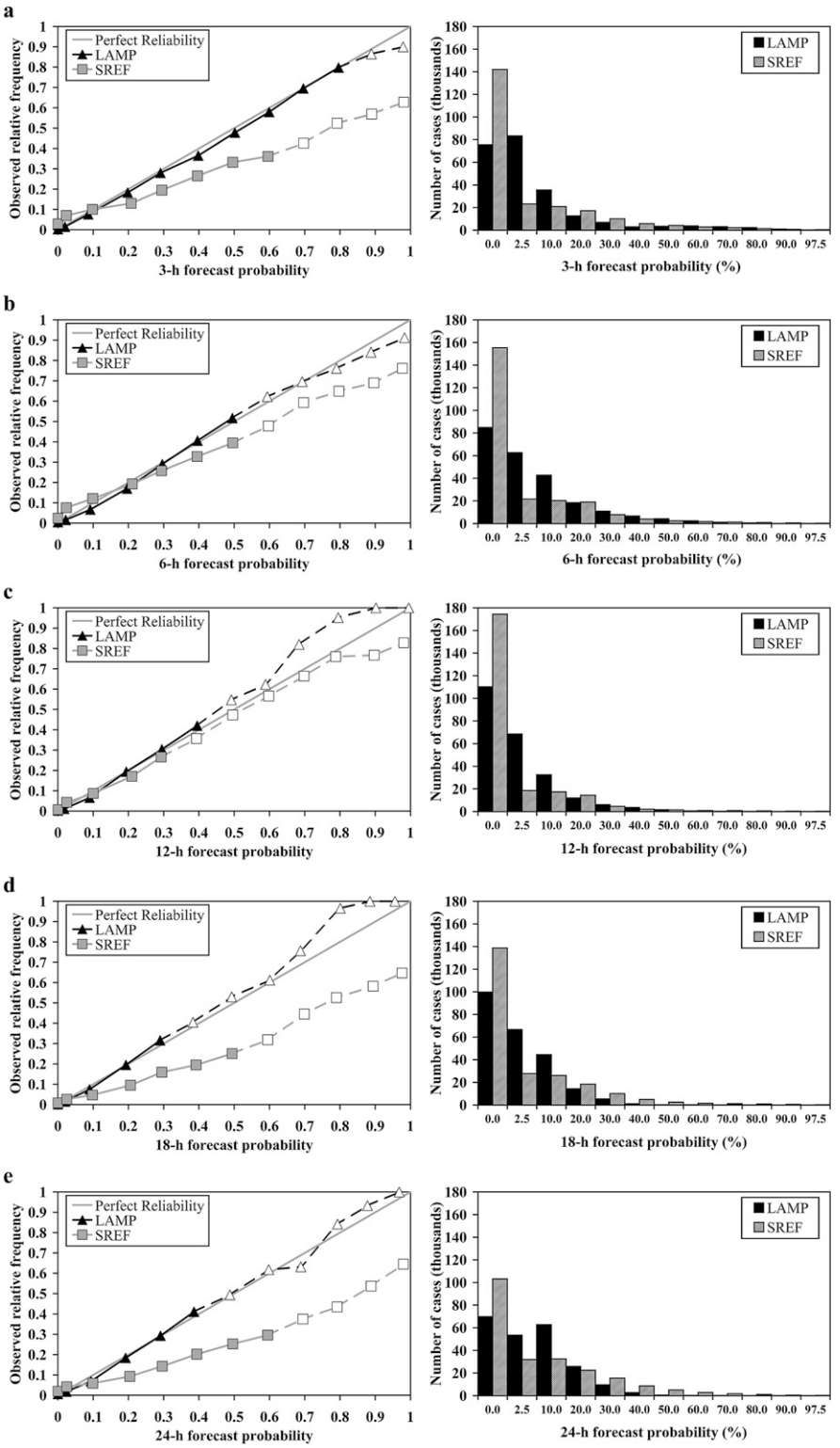

FIG. 10. (left) Reliability diagrams and (right) histograms for probabilities of visibility <3 mi for the (a) 3-, (b) 6-, (c) 12-, (d) 18-, and (e) 24-h projections for LAMP and SREF. Verification is from the 0900 UTC cycle for the cool season of Oct 2006–Mar 2007. Reliability values (left side) composing less than 1% of the total number of cases are indicated with a hollow marker and a dashed line to emphasize that these values should be interpreted judiciously.

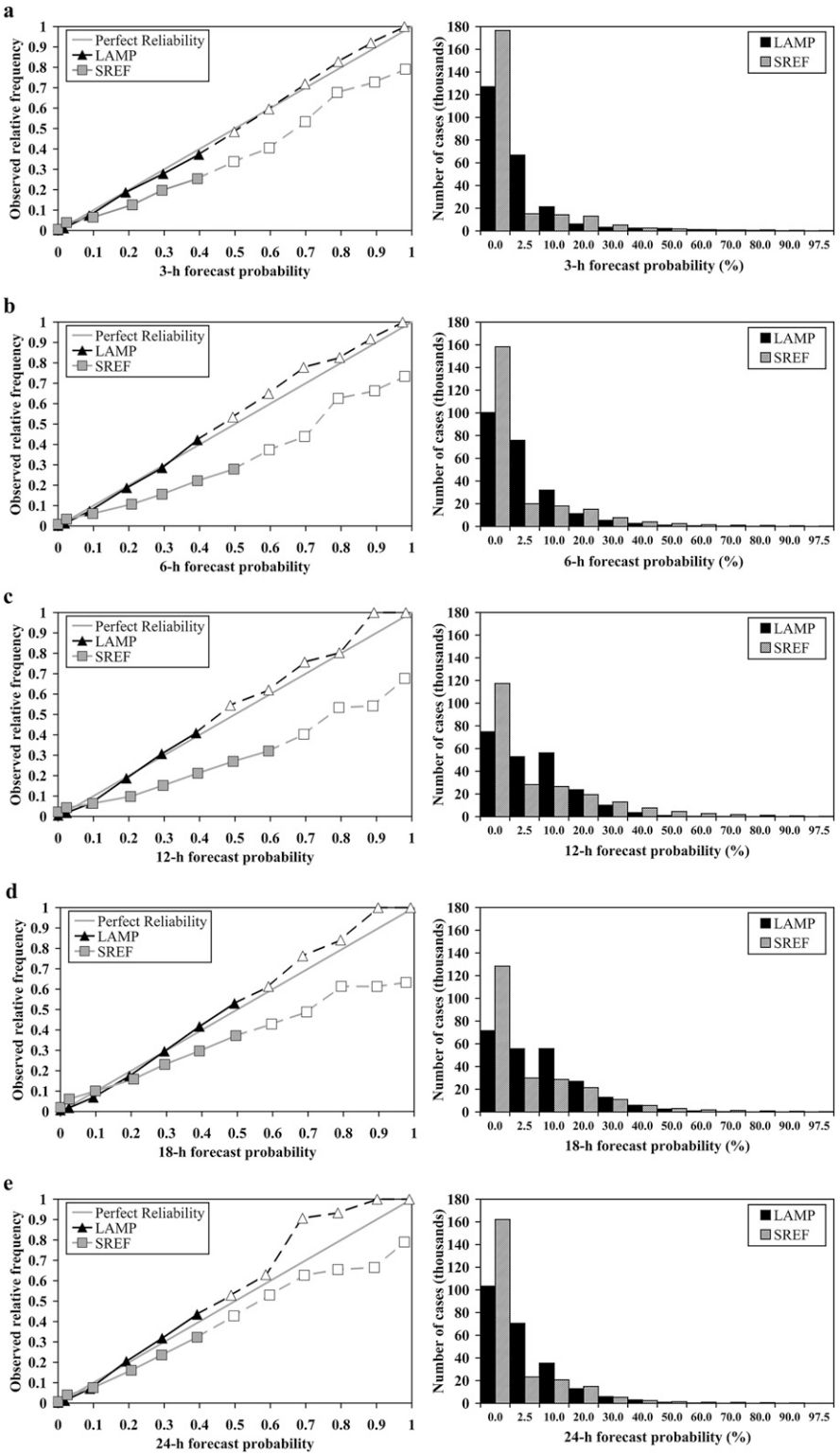

FIG. 11. As in Fig. 10, but for 2100 UTC.

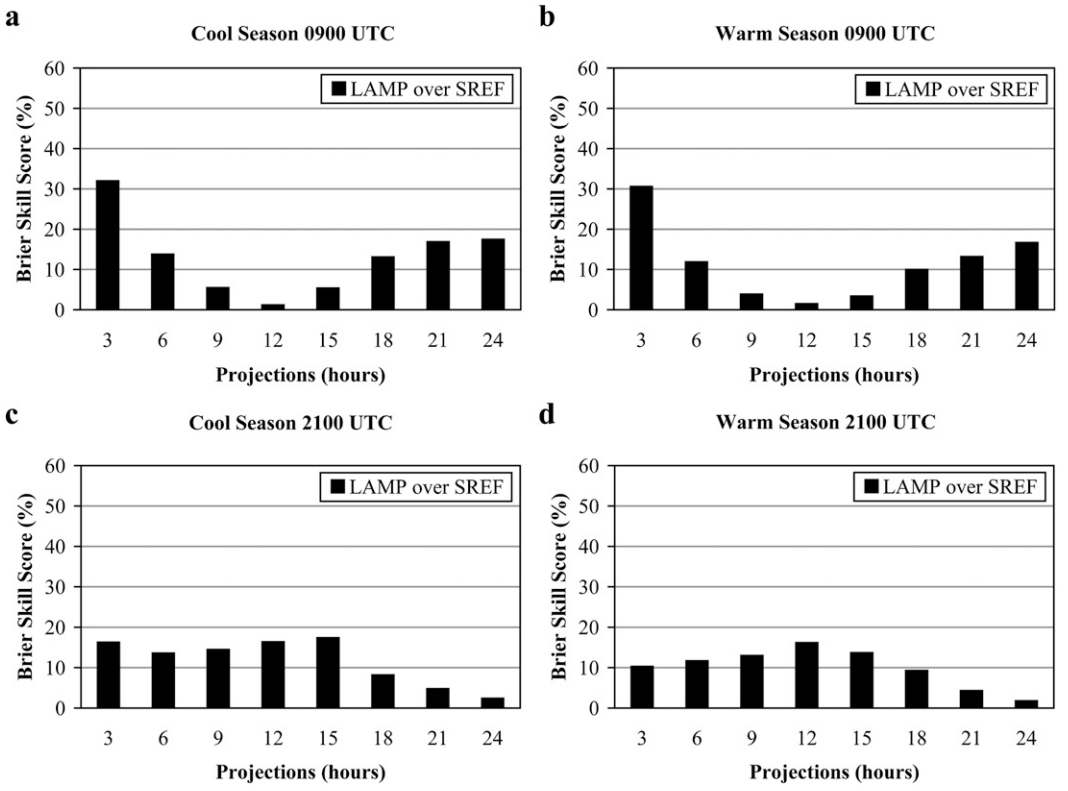

FIG. 12. Brier skill scores of LAMP over SREF for probability visibility forecasts of <3 mi. Verification is from (a) the 0900 and (c) 2100 UTC cycles for the cool season of Oct 2006–Mar 2007 and (b) 0900 and (d) 2100 UTC cycles for the warm season of Apr–Sep 2007.

LAMP at the 12- and 24-h projections, respectively (both valid at 2100 UTC). This behavior is evident for the warm season as well (not shown).

2) BRIER SKILL SCORE

Figure 12 shows the Brier skill scores for probabilistic forecasts of VIS < 3 mi. The same cycles and periods described in the previous section were verified. While LAMP probabilistic forecasts improve over SREF, the improvement is not as pronounced as for CIG < 1000 ft (Fig. 9). The greatest improvement in the LAMP visibility forecasts for the 0900 UTC cycle occurs at the 3-h projection for both the cool and warm seasons (32% for cool season and 31% for warm season). For the 2100 UTC cycle, the greatest improvement of LAMP over the SREF occurs at the 15- (17%) and 12-h (16%) projections for the cool and warm seasons, respectively. These valid times correspond to the forecast period when the SREF probabilities demonstrate their poorest reliability. At the 0900 UTC cycle for both cool and warm seasons, LAMP marginally improves over the SREF between the 9- and 15-h projections (generally under 5%). Similar relative improvement is also evident at the 2100 UTC cycle between the 18- and 24-h projections. These times correspond to the forecast period when the SREF probabilities demonstrate their best reliability.

5. Summary and conclusions

This paper compares statistically based LAMP CIG and VIS forecasts over the CONUS with forecasts produced by the RUC20, WRF-NMM, and SREF. This verification study was conducted by pooling 1462 stations across the CONUS. While this approach yields a CONUS-wide average performance measure and may not be representative of specific sites, we believe that verifying the data in this manner provides useful information concerning the overall strengths and weaknesses of these forecasting systems.

We found that independent of season, the 0000 and 1200 UTC station-based LAMP CIG, VIS, and IFR or lower categorical forecasts are more accurate than RUC20 and WRF-NMM postprocessed forecasts when interpolated to stations and then categorized. In the early projections (1–6 h), LAMP forecasts are noticeably more accurate. In the 6–12-h projection period, the CSI scores for the RUC20, and to a lesser extent WRF-NMM, begin to approach the LAMP CSI scores. WRF-NMM

forecasts of CIG, VIS, and IFR or lower in the 13–25-h forecast period are generally less accurate than LAMP especially for the elements of CIG and IFR or lower. We also note that the warm season verification scores of the less common events (e.g., VIS $<$ 1 mi, CIG $<$ 500 ft, and IFR or lower) display less accuracy than the cool season verification scores across all systems.

By performing verifications at both 0000 and 1200 UTC for the WRF-NMM, we found that forecasts of VIS and IFR or lower, and to a lesser extent CIG, are more sensitive to the time of day for which the forecasts are valid than the forecast projection itself. This can be partially attributed to the tendency of the WRF-NMM to saturate the model level closest to the ground too frequently in clear conditions with light winds during the nighttime hours. WRF-NMM developers are aware of this issue and have been making incremental improvements over the years (G. Manikin 2009, personal communication).

For the 0900 and 2100 UTC forecast cycles and verification periods studied here, LAMP CIG ($<$1000 and $\leq$3000 ft) and VIS ($<$3 mi) forecast probabilities exhibit overall better reliability across all probability bins than the SREF probabilities. While the SREF probabilities are sometimes very reliable for both CIG $\leq$ 3000 ft and VIS $<$ 3 mi, the valid times generally occur during the daylight hours (independent of cycle and season). In contrast, LAMP probability forecasts of CIG and VIS generally display consistent projection-to-projection reliability independent of cycle time. Where LAMP CIG reliabilities do differ slightly from projection to projection (e.g., 0900 UTC cycle – cool season for CIG $\leq$ 3000 ft), the fluctuation does not appear to be diurnally driven and is possibly due to a small sample size.

The skill of LAMP CIG and VIS probability forecasts over the SREF is demonstrated through the Brier skill score. For both cycle times and seasons, LAMP forecasts of CIG $<$ 1000 ft show considerable skill over the SREF, especially at the 3-h projection when observations have a significant impact on LAMP forecasts. Although the SREF reliabilities for VIS $<$ 3 mi vary as a function of the time of day (rather than cycle issuance time), their overall relative stable behavior strongly contributes to a smaller LAMP Brier score percentage improvement. As a final thought, we note that the verification of both the operational and corrected SREF CIG probabilities indicates the potential benefits of calibrating SREF probability forecasts. Calibrating these probabilistic forecasts, that is, correctly populating the bins with forecast probability values closer to the observed frequency of events, would remove a portion of the bias exhibited by the SREF and in turn improve forecast reliability.

LAMP forecasts are being used by a variety of different customers in the decision support process for aviation forecasts. Currently, National Weather Service forecasters can use LAMP guidance for both generating and updating TAFs (Oberfield et al. 2008). The utility of LAMP in the TAF preparation process has been documented by Thompson and Baumgardt (2009). In their paper, they discuss the evolution of a particular snow event and its associated impacts on flight conditions around Rochester, Minnesota. Thompson and Baumgardt noted that 1) by integrating LAMP guidance into the aviation forecast process, TAF accuracy can be improved, and 2) LAMP CIG and VIS forecasts, in some instances, actually verify better than the official TAF.

The National Ceiling and Visibility (NCV) forecast system also benefits from LAMP CIG and VIS forecasts (Black et al. 2008). NCV ingests CIG and VIS forecasts from a host of objective weather forecasting systems and determines the most reasonable CIG and VIS forecasts. Statistical analyses in the NCV system have shown that LAMP performs very well and, in some instances, outperforms all other forecast system inputs.

A defining characteristic of the LAMP system is its ability to utilize station-based observations in a meaningful way to generate forecasts. The observations are not only integral to generating skillful and reliable probabilistic forecasts but are also used in the postprocessing step of producing threshold values for accurate categorical forecasts. The results presented in this paper show that RUC20 and WRF-NMM continuous CIG and VIS forecasts and SREF CIG and VIS probabilistic forecasts could benefit from additional postprocessing.

Acknowledgments. The authors thank Stan Benjamin, Geoff Manikin, and Jun Du for their valuable comments and assistance with the model data used in this verification study. We are also indebted and grateful to Kelly Malone, Tyler Fleming, and James Cipriani for their assistance in retrieving the necessary data and for producing many of the plots shown in this paper. We would also like to thank Mitch Weiss for the development of the LAMP CIG equations and his review of the manuscript. Finally, we wish to thank the anonymous reviewers of this manuscript for their insightful comments and helpful suggestions.

REFERENCES

Benjamin, S. G., J. M. Brown, K. J. Brundage, D. Kim, B. Schwartz, T. Smirnova, and T. L. Smith, 1999: Aviation forecasts from the RUC-2. Preprints, *Eighth Conf. on Aviation, Range, and Aerospace Meteorology,* Dallas, TX, Amer. Meteor. Soc., 486–490.

——, and Coauthors, 2004: An hourly assimilation–forecast cycle: The RUC. *Mon. Wea. Rev.,* **132,** 495–518.

Black, J. L., R. E. Bateman, P. H. Herzegh, G. Wiener, J. Cowie, and C. J. Kessinger, 2008: An automated national-scale ceiling

and visibility forecast system: Development progress. Preprints, *13th Conf. on Aviation, Range, and Aerospace Meteorology,* New Orleans, LA, Amer. Meteor. Soc., 7.2. [Available online at http://ams.confex.com/ams/pdfpapers/133936.pdf.]

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.,* **78,** 1–3.

Dallavalle, J. P., and V. J. Dagostaro, 1995: The accuracy of ceiling and visibility forecasts produced by the National Weather Service. Preprints, *Sixth Conf. on Aviation Weather Systems,* Dallas, TX, Amer. Meteor. Soc., 213–218.

Du, J., and Coauthors, 2004: The NOAA/NWS/NCEP Short-Range Ensemble Forecast (SREF) system: Evaluation of an initial condition vs. multi-model physics ensemble approach. Preprints, *16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., 21.3. [Available online at http://ams.confex.com/ams/pdfpapers/71107.pdf.]

Ghirardelli, J. E., 2005: An overview of the redeveloped Localized Aviation MOS Program (LAMP) for short-range forecasting. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., 13B.5. [Available online at http://ams.confex.com/ams/pdfpapers/95038.pdf.]

——, and B. Glahn, 2010: The Meteorological Development Laboratory's Aviation Weather Prediction System. *Wea. Forecasting,* in press.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Keith, R., and S. M. Leyton, 2007: An experiment to measure the value of statistical probability forecasts for airports. *Wea. Forecasting,* **22,** 928–935.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

NRC, 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. The National Academies Press, 124 pp.

NWS, 2008: Terminal Aerodrome Forecasts. National Weather Service Instruction 10-813, NOAA/NWS, 61 pp.

Oberfield, M. G., M. R. Peroutka, and C. Abelman, 2008: Using probabilistic forecast guidance and an update technique to generate terminal aerodrome forecasts. Preprints, *13th Conf. on Aviation, Range, and Aerospace Meteorology,* New Orleans, LA, Amer. Meteor. Soc., 5.4. [Available online at http://ams.confex.com/ams/pdfpapers/128927.pdf.]

Rudack, D. E., 2005: Improvements in the Localized Aviation MOS Program (LAMP) categorical visibility and obstruction to vision statistical guidance. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., P1.52. [Available online at http://ams.confex.com/ams/pdfpapers/95050.pdf.]

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Note NCAR/TN-468+STR, 88 pp.

Smirnova, T. G., S. G. Benjamin, and J. M. Brown, 2000: Case study verification of RUC/MAPS fog and visibility forecasts. Preprints, *Ninth Conf. on Aviation, Range, and Aerospace Meteorology,* Orlando, FL, Amer. Meteor. Soc., 31–36.

Stoelinga, M. T., and T. T. Warner, 1999: Nonhydrostatic, mesobeta-scale model simulations of cloud ceiling and visibility for an East Coast winter precipitation event. *J. Appl. Meteor.,* **38,** 385–404.

Thompson, S., and D. Baumgardt, 2009: Improving forecasts of Instrument Flight Rule conditions over the upper Mississippi valley and beyond. Preprints, *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction,* Omaha, NE, Amer. Meteor. Soc., JP3.9. [Available online at http://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154154.htm.]

Weiss, M., and J. E. Ghirardelli, 2005: A summary of ceiling height and total sky cover short-term statistical forecasts in the Localized Aviation MOS Program (LAMP). Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., 13B.6. [Available online at http://ams.confex.com/ams/pdfpapers/95121.pdf.]

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

Zhou, B., and Coauthors, 2004: An introduction to NCEP SREF aviation project. Preprints, *11th Conf. on Aviation, Range, and Aerospace Meteorology,* Hyannis, MA, Amer. Meteor. Soc., 9.15. [Available online at http://ams.confex.com/ams/pdfpapers/81314.pdf.]